Recombination and purifying and balancing selection determine the evolution of major antigenic protein 1 (*map 1*) family genes in *Ehrlichia ruminantium*

Bashir Salim[1,2], Mutaz Amin[3], Manabu Igarashi[4], Kimihito Ito[5], Frans Jongejan[6,7], Ken Katakura[2], Chihiro Sugimoto[8,] Ryo Nakao[2*]

[1]Department of Parasitology, Faculty of Veterinary Medicine, University of Khartoum, P.O Box 32 Khartoum-North, Sudan.

[2]Laboratory of Parasitology, Department of Disease Control, Graduate School of Veterinary Medicine, Hokkaido University, Sapporo, Japan.

[3]Faculty of Medicine, University of Khartoum, Qasr Street, 11111 Khartoum, Sudan.

[4]Division of Global Epidemiology, Hokkaido University Research Center for Zoonosis Control, Sapporo, Japan.

[5]Division of Bioinformatics, Hokkaido University Research Center for Zoonosis Control, Sapporo, Japan.

[6]Utrecht Centre for Tick-borne Diseases (UCTD), FAO Reference Centre for Ticks and Tick-borne Diseases, Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands.

[7]Department of Veterinary Tropical Diseases, Faculty of Veterinary Science, University of Pretoria, Onderstepoort, South Africa.

[8]Division of Collaboration and Education, Hokkaido University Research Center for Zoonosis Control, Sapporo, Japan.

Keywords: *Ehrlichia ruminantium*; heartwater; recombination; negative and balance selection; *map1*

Running head: Evolution of *map1* family genes of *Ehrlichia ruminantium*

BS: bashirsalim@gmail.com

MA: mtz88@hotmail.co.uk

MI: igarashi@czc.hokudai.ac.jp

KI: itok@czc.hokudai.ac.jp

FJ: F.Jongejan@uu.nl

KK: kenkata@vetmed.hokudai.ac.jp

CS: sugimoto@czc.hokudai.ac.jp

RN: ryo.nakao@vetmed.hokudai.ac.jp

***Corresponding author:**

Dr. Ryo Nakao

Laboratory of Parasitology, Department of Disease Control, Graduate School of Veterinary Medicine, Hokkaido University, Sapporo, Japan.

Kita 20, Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0020, Japan

Tel: +81 11 706 5196

Fax: +81 11 706 5196

E-mail: ryo.nakao@vetmed.hokudai.ac.jp

1. ABSTRACT

Heartwater is an economically important disease of ruminants caused by the tick-borne bacterium *Ehrlichia ruminantium*. The disease is present throughout sub-Saharan Africa as well as on several islands in the Caribbean, where it poses a risk of spreading onto the American mainland. The dominant immune response of infected animals is directed against the variable outer membrane proteins of *E. ruminantium* encoded by polymorphic multigene families. Here, we examined the full-length sequence of the major antigenic protein 1 (*map1)* family genes in multiple *E. ruminantium* isolates from different African countries and the Caribbean, collected at different time points to infer the possible role of recombination breakpoint and natural selection. A high level of recombination was found particularly in *map1* and *map1-2*. Evidence of strong negative purifying selection in *map1* and balancing selection to maintain genetic variation across these samples from geographically distinct countries suggests host–pathogen co-evolution. This co-evolution between the host and pathogen results in balancing selection by maintaining genetic diversity that could be explained by the demographic history of long-term pathogen pressure. This signifies the adaptive role and the molecular evolutionary forces underpinning *E. ruminantium map1* multigene family antigenicity.

2. INTRODUCTION

Ehrlichia ruminantium is an obligate intracellular Gram-negative bacterium causing heartwater in wild and domestic ruminants. This pathogen is transmitted by ticks of the genus *Amblyomma* and occurs throughout sub-Saharan Africa and on several islands in the Caribbean, where it poses a threat of spreading to the American mainland (Burridge et al., 2002). The disease has a significant economic impact on livestock production in endemic countries as reported previously (Mukhebi et al., 1999).

Outer membrane proteins (OMPs) of Gram-negative bacteria are known to play important roles in interaction with the host (Lin et al., 2002). Multigene families coding for OMPs have been characterized in several bacterial species of the genus *Ehrlichia*, including the outer membrane protein p28 (*omp-1*) in *Ehrlichia chaffeensis* (Ohashi et al., 1998b; Yu et al., 1999), the p30 outer membrane protein (*p30*) in *Ehrlichia canis* (McBride et al., 1999; Ohashi et al., 1998a), and the major antigenic protein 1 (*map1*) in *E. ruminantium* (Sulsona et al., 1999; van Heerden et al., 2004). As the molecules encoded by these genes are recognized by the host immune system (Cheng et al., 2003; Jongejan and Thielemans, 1989; Li et al., 2002; McBride et al., 1999), they have been exploited as potential targets for the development of vaccines and serodiagnostic tests (Crocquet-Valdes et al., 2011; Feburay et al., 2017; McBride et al., 1999; Nyika et al., 2002; Ohashi et al., 1998a; Peter et al., 2001). Several lines of evidence suggest that host immune evasion could result from the differential expression of individual genes within the paralogs (Ge and Rikihisa, 2007; McBride et al., 2003; Ohashi et al., 2001; Reddy and Streck, 2000). Hence, for designing effective vaccines

and diagnostic tools, it is important to understand the diversity and evolutionary mechanisms of these multigene families.

Homologous recombination (HR) is a housekeeping mechanism associated with the maintenance of chromosome integrity and generation of genetic variability. HR was initially defined as the result of the sexual process in prokaryotes as in eukaryotes and was later acknowledged as a major DNA repair process. Both genetic and biochemical studies have revealed the crucial role that HR plays in all organisms in the repair of a variety of DNA damages of exogenous and endogenous origin (Kuzminov, 1999; Michel et al., 2004). Additionally, HR is essential for the bacterial genome by allowing integration of homologous alien DNA arising from transformation or conjugation (Smith, 1991; Lorenz and Wackernagel, 1994). It also helps adaptive mutations and removal of the deleterious mutations hitchhiking with them (Otto and Michalakis, 1998), thus allowing allelic recombination between closely related strains (Feil, 2004). Recombination between homologous segments in genomes leads to chromosomal instability (Hughes, 1999; Rocha, 2004), and among bacteria, the rate of chromosomal rearrangements correlates with the number of repeated sequences in genomes (Rocha, 2003). Further, intrachromosomal HR between large repeated regions is often adaptive, allowing the generation of genotypic diversity in pathogens (Finlay and Falkow, 1997; Mehr and Seifert, 1998; Rocha and Blanchard, 2002).

The *map1* multigene family of *E. ruminantium* comprises 16 paralogs tandemly arranged in the genome (van Heerden et al., 2004). All the paralogs are maintained in the same order in the genomes of the *E. ruminantium* strains sequenced so far (Collins et al., 2005; Frutos et al., 2006; Nakao et al., 2016). Transcriptional analysis revealed that all the paralogs were

119 transcriptionally active when cultured in bovine endothelial cells (Bekker et al., 2005; van
120 Heerden et al., 2004), while this was not the case in ticks and tick cell lines (Bekker et al.,
121 2005; Postigo et al., 2007). Differential expression of MAP1-family proteins was also
122 reported using different analytical approaches (Marcelino et al., 2012; Postigo et al., 2008).
123 These findings may support the possible involvement of *map1* paralogs in adaptation to
124 different host environments.

125 The diversity of *map1* paralogs has been analysed only for a limited number of strains so far
126 (Barbet et al., 2009). The present study extended the comparison by including *E.*
127 *ruminantium* strains from wide geographic origins to understand the evolutionary
128 mechanisms of these polymorphic genes. The results indicated that recombination and
129 purifying and balancing selection play a significant role in the evolution of *map1* paralogs.

130

131 **3. MATERIALS AND METHODS**

132 **3.1 Bacterial strains**

133 The *E. ruminantium* strains used in this study are listed in Table 1. All *E. ruminantium*
134 strains were grown in bovine aorta endothelium cells and subjected to DNA extraction as
135 described previously (Nakao et al., 2010). We also included two *Amblyomma variegatum*
136 samples which were positive for *E. ruminantium* infection as reported previously (Nakao et
137 al., 2011). Sequences encoding the locus of all 16 *map1* paralogs were either retrieved from
138 the database Welgevonden (Erwe and Erwo), Gardel, Crystal Springs, Kerr Seringe,
139 Pokoase 417 and Sankat 430) or determined by long PCRs followed by Sanger sequencing
140 as described below.

3.2 Long-range PCR and Sanger sequencing of *map1* paralogs

Four long-range PCR assays were developed using primers amplifying approximately 20 kb of the genomic region spanning all 16 *map1* paralogs (Fig. 1). The primer sequences used for long-range PCR assays are listed in Table S1. PCR was carried out using high-fidelity PrimeSTAR GXL DNA polymerase (Takara Bio, Shiga, Japan) according to the manufacturer's instructions. The PCR products were purified using ExoSAP-IT (USB Corporation, Cleveland, OH) and sequenced using primers listed in Table S1. The sequencing reactions were conducted using the BigDye Terminator version 3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA) and the products were analysed on an ABI Prism 3130x genetic analyser (Applied Biosystems). The obtained sequences were submitted to the DNA Data Bank of Japan (DDBJ) (http://www.ddbj.nig.ac.jp) under accession nos. AB818931 to AB818946.

Downstream analysis

To detect the possible signatures of positive, negative, or neutral selection across *map1* family genes as well as those of HR we carried out the following analysis:

3.4 Sequence alignment and nucleotide diversity

Multiple sequence alignment was performed using MUSCLE implemented in MEGA7 (Kumar et al., 2016). To estimate genetic variability across the gene family, genetic diversity (nucleotide and haplotype diversity and the mean number of nucleotide differences) and their standard deviations were calculated for the entire dataset and for each gene. The number of haplotypes were determined using DNaSP v5 (Librado and Rozas, 2009). To infer the population demographic history and the dynamics of these ehrlichial

163 pathogens we measured the haplotype mismatch distribution patterns (Rogers and
164 Harpending, 1992) for each family gene. Departures of the observed sum of squares
165 differences (SSD) from the simulated model of expansion were tested with the chi-square
166 test of the goodness of fit statistic and Harpending's raggedness index 'r' (Harpending,
167 1994) following 1000 coalescent simulations. Analysis of mismatch distribution patterns
168 was supported by two coalescent-based estimators of neutrality; Tajima's D (Tajima, 1989)
169 and Fu's F (Fu, 1997) statistics. The expected value for Tajima's D and Fu's F is zero, and
170 both positive and negative deviations are informative about the distinct demographic and/or
171 selective events. The significance of these two statistics was tested with 1000 coalescent
172 simulations in Arlequin v3.5 (Excoffier and Lischer, 2010).

173 **3.5 Detection of recombination breakpoints**

174 From abundant algorithms and software tools designed to detect and analyse recombination,
175 we used the Genetic Algorithm Recombination Detection (GARD) implemented in the
176 Datamonkey 2.0 modern web application (Weaver et al., 2018) to analyse multiple-
177 sequence alignments for recombination to estimate the number and location of breakpoints
178 and segment-specific phylogenetic trees. The method searches for all possible breakpoints
179 in the sequence alignment, infers phylogenies for each putative non-recombinant fragment,
180 and assesses goodness of fit using Akaike Information Criterion (AIC) (Sugiura, 1978), an
181 information-based criterion derived from a maximum likelihood model fit for each segment
182 augmented by a genetic algorithm (GA) heuristic to quickly explore a large-state space
183 (Kosakovsky Pond et al., 2006).

184 **3.6 Detection sites under positive/negative selection**

185 Single-Likelihood Ancestor Counting (SLAC) implemented in the Datamonkey web-server
186 was executed to identify the sites evolving under the influence of positive, neutral, and/or
187 negative selection. This method depends on a combination of maximum-likelihood (ML)
188 and counting approaches by inferring the nonsynonymous (dN) and synonymous (dS)
189 substitution rates on a per-site basis for a given coding alignment and corresponding
190 phylogeny. The method assumes that the selection pressure for each site is constant along
191 the entire phylogeny. The default significance cut-offs used are $P = 0.1$ for SLAC and
192 posterior probability = 0.9. Additionally, dN and dS were also calculated using MEGA7.

193 **4. RESULTS**

194

195 **1.4 Diversity and population demography**

196 A large number of haplotypes and a high level of haplotype diversity was observed across
197 the dataset (Table 2; Figs. S1 & S2, with *map1*, *map1-4*, *map1-5* and *map1-14* scored 11
198 haplotypes in each, and the same scored high haplotype diversity = 0.97). The mismatch
199 distribution pairwise differences pattern under sudden or spatial expansion revealed bi- and
200 multimodal shapes with significant deviation of the goodness-of-fit, Sum of Square
201 Deviation, and Harpending's raggedness index (Table S2 & Fig. S3A-O). The positive
202 value of Tajima's D is the result of an excess of intermediate frequency alleles that
203 indicates either population bottlenecks or structure and/or balancing selection, of which we
204 assume the latter (see the section 4.4 and Fig. 2). Negative values of Tajima's D for *map1-4*
205 and *map1-3* designate an excess of low frequency alleles indicating positive selection of
206 these two genes (Fig. 2).

207

208 **4.2 Homologous recombination**

209 HR was detected in seven genes of the family of *map1*. Among them, *map1* scored the
210 highest recombination breakpoints (r = 262) followed by *map1-2* (r = 208), and this
211 dropped dramatically to 106, 101, and 100 in *map1-6*, *map1-5,* and *map1-14*, respectively
212 (Fig. 3). The least recombination breakpoints detected were in *map1-10* (r = 30) and *map1-*
213 *1* (r = 29). The actual location of the recombination breakpoints is illustrated in Fig. 4, for
214 the four highly recombining genes. For instance, *map1* has shown three locations of
215 breakpoints around 145, 220, and 280, whereas, *map1-2* has shown four locations of
216 breakpoints around 95-100, 330-350, 500, and 590.

217

218 **4.3 Negative diversifying selection**

219 We detected negative purifying selection in 12 out of 16 genes and no single positive
220 selection was detected (Fig. 5). Again, *map1* scored the highest negative selection with the
221 selection site being 48 compared to the following *map1-2* that only scored 8 selection sites
222 (Fig. 5). Five genes (*map1-1*, *map1-5*, *map1-10*, *map1-12*, and hypothetical gene (HG)) did
223 not show either negative or positive selection, indicating neutral selection. This result was
224 supported by separate analysis of the dN and dS nucleotide substitution that also displayed
225 the same order (Fig. S4).

226

227 **4.4 Balancing selection**

228 We assumed a balancing selection for the following reasons: first, the excess of dN
229 polymorphisms represents either balancing or purifying selection (Table 2) (Fijarczyk et al.,
230 2016); second, a positive Tajima's *D* value for all the gene family except two genes, *map1-*
231 *3* and *map1-4*, that are under directional positive selection (Fig. 2) (Fijarczyk et al., 2016);

third, an excess of polymorphisms and high frequencies of segregating sites, as indicated clearly in Table 2. The purifying selection detected can be explained by the purifying balancing selection as shown in Fig. 5, where *map1* also scored the highest in purifying balancing selection. In contrast, *map1-3* and *map1-4* were found to experience purifying directional selection.

The patterns of genetic variation are shaped by two stochastic events that are the history of coalescent events and the history of mutational events. Under neutrality, mutations are uniformly distributed along the branches of a genealogy, and therefore, the number of mutations occurring on a branch is proportional to its length (see Fig. 6A-D), the time to coalescence is longer in (A), and moderate in (B) or short in (C & D), depending on the time required to reach this balancing selection for A and B. Considering recombination events, different regions of the genome can have distinct gene genealogies.

DISCUSSION

Negative/purifying or background selection plays a substantial role in maintaining the long-term stability of *E. ruminantium* populations across different countries by removing deleterious mutations. Therefore, this type of selection is more prevailing for the success of evolution in optimizing the functions of an organism. Purifying selection ensures that deleterious mutations cannot take over a population and that any improved structures when fixed in a population are maintained as long as they are required. Host-parasite interactions are a renowned example of this type of situation. In this regard, the host immune system evolves to recognize a special structure on the parasite and allows its removal. This in turn

induces negative selection on the current form of the parasite while leading to positive selection of variants that cannot be recognized by the host (Charlesworth et al., 1995).

In this study, a strong purifying selection was illustrated in *map1* gene compared to other family genes, indicating that this gene plays an essential functioning role in the immune system. On the other hand, other members of the gene family demonstrated varied levels from moderate to weak purifying selection, which could be explained as a signature of functional orchestration that results in each family gene playing parts with varying degrees. This negative/purifying selection is the main evolutionary force that has shaped most of the outer membrane gene families and could be explained by the demographic history of longstanding pathogen pressure.

We also detected balancing selection on this *map 1* gene family, a type of selection that is known to maintain the genetic variation in immunity genes. Balancing of host–parasite coevolution is well documented in literature (Charlesworth and Charlesworth 2010; Eizaguirre et al., 2012; Phillips et al., 2018). This coevolution between hosts and pathogens results in balancing selection by maintaining the genetic diversity at immunity genes such as the *map1* family. In other words, *E. ruminantium* overcomes the host immune system by maintaining extensive and strong genetic diversity in *map1* gene among the 16 genes in the family, with moderate to weak diversity in other family genes in infected cattle hosts, thus trying to escape recognition by the immune system of the cattle host.

It is well known that balancing selection is expected to operate and maintain high variation in some immune genes through mechanisms of overdominance, negative frequency-dependent selection, or temporally and spatially fluctuating selection. All these are expected to work on *map1* family genes, operating with various degrees in these paralogous

277 genes that play a major role in balancing selection plus negative purifying selection in the
278 evolution of these immunity-related genes.

279 Selection here tends to predominantly maintain *map1* and to some extent *map1-2* and
280 *map1-6* (Fig. 4). The genetic signature of balancing selection is an excess of polymorphic
281 sites at intermediate (balanced) frequencies, relative to expectations under neutrality
282 (Weedall and Conway, 2010). This balancing selection seems to be involved in maintaining
283 diversity at *map1* that coordinates recognition between the self and non-self (Richman and
284 Kohn, 1999), which can be considered as a good immuno-diagnostic or vaccine candidate
285 gene.

286 It has been previously documented that *E. ruminantium map1* variants are not
287 geographically constrained and show no evidence of having evolved under positive
288 selection pressure (Allsopp et al., 2001). Additionally, Hughes and French (2007) reported
289 evidence for HR in the *map1* gene, which supports our current findings. Besides the above,
290 we also demonstrate a greater HR rate at *map1* (Figs. 3 and 5) indicating a higher genomic
291 adaptation in this region. This along with ongoing balanced selection forces that are purged
292 by "purifying" any lower fitness "mutants" produced, prevents them from accumulation. In
293 contrast, wherever recombination is low there is a greater density of selective variants that
294 do not segregate freely, lowering the efficiency of selection and consequently the
295 adaptation rate. It appears that recombination increases proportionally with the intensity of
296 selection at the loci/gene subject to recombination. Many reports revealed a positive
297 correlation between recombination and adaptation. For instance, Levin and Cornejo (2009)
298 simulated mutation, recombination, selection and inter-population competition to explore
299 the conditions under which: i) recombination augments the rates of evolution in bacterial

300 populations and, ii) when the capacity for HR is favoured in competition with non-
301 recombining populations. They demonstrated that under broad conditions, HR occurring at
302 rates in a range estimated for *Escherichia coli, Haemophilus influenza, Streptococcus*
303 *pneumoniae,* and *Bacillus subtilis* can increase the rate of adaptive evolution in bacterial
304 populations.

305 It appears that recombination increases the rate at which populations adapt to their
306 environment, whereby the capacity for shuffling homologous genes within a population
307 provides an advantage to the recombining strain when competing with populations without
308 this capacity. Commonly, recombination in bacteria is a rare event and not a part of the
309 reproductive process. Nonetheless, recombination is broadly defined to include the
310 acquisition of genes from external sources via horizontal gene transfer (from other
311 populations of bacteria of the same and different species, as well as from eukaryotes and
312 archaea) and plays a central role as a source of variation for adaptive evolution in many
313 bacterial species. In this manner, bacteria can expand their ecological niches by expressing
314 the genes and genetic elements obtained from other bacterial populations of the same and
315 different species, as well as from eukaryotes and archaea. This scenario may be ideal
316 especially in vector-borne bacteria where diverse populations of microbiota and other
317 organisms may be encountered.

318 Recombination in the form of the receipt and incorporation of genes and genetic elements
319 from other strains and species of bacteria (Mazodier and Davies, 1991) as well as archaea
320 and eukaryotes (Nelson et al., 1999; Brown, 2003) plays a prominent role as a source of
321 variation for the adaptive evolution of many bacterial species (Lawrence and Ochman,
322 1998; Ochman et al., 2000; Gal-Mor and Finlay, 2006). However, Bekker et al. (2005)

found that a recombination between two *map1* genes, namely at *map1-3* and *map1-2*, had occurred in one subpopulation with deletion of one entire gene.

As previously noted, the *E. ruminantium map1* multigene family is regulated differently in the host and tick cell environments subject to recombination causing an altered gene arrangement and different transcriptional activities (Bekker et al., 2005). This has also been demonstrated in malaria parasites, where the stage of the parasite and the period of exposure to the immune system determine the pattern of selection (Weedall and Conway, 2010). These observations support our results in which a similar pattern of evolution in the *E. ruminantium map1* gene family is observed across different countries and time points suggesting the coevolution of these parasites in their host and the vector.

CONCLUSION

Here, we identified negative and balanced polymorphism selections in the *map1* multigene family, a selective pressure that has shaped *E. ruminantium* in response to environmental changes over time and space, most noticeably observed in *map1* followed by *map1-2* and to a lesser extent in the other gene family members.

Acknowledgements

This work was supported in part by JSPS KAKENHI grant numbers 25850195, 15K14850, 15H05633, and 16H06431.

345 **TABLES**

346 **Table 1. *Ehrlichia ruminantium* strains used in this study.**

Strain	Geographical origin	Year of isolation	Reference
Crystal Springs	Zimbabwe	1990	(Byrom et al., 1991)
Gardel	Guadelope, Caribbean	1982	(Uilenberg et al., 1985)
Ifé Nigeria	Nigeria	1983	(Ilemobade et al., 1978)
Kerr Serigne	Gambia	2001	(Faburay et al., 2005)
Kwanyanga	South Africa	NR[a]	(Mackenzie and van Rooyen, 1981)
Lutale	Zambia	1986	(Jongejan et al., 1988)
Pokoase 417	Ghana	1996	(Bell-Sakyi et al., 1997)
Sankat 430	Ghana	1996	(Bell-Sakyi et al., 1997)
Um Banein	Sudan	1981	(Jongejan et al., 1984)
Welgevonden (Erwo)	South Africa	1985	(Plessis, 1985)
Welgevonden (Erwe)	South Africa	1985	(Plessis, 1985)
Zeerust	South Africa	1979	(Jongejan et al., 1980)
Uganda P016[b]	Uganda	2008-2009[c]	(Nakao et al., 2011)
Uganda T020[b]	Uganda	2008-2009[c]	(Nakao et al., 2011)

347 [a]NR, not recorded. [b]Tick ID. [c]Year of tick collection.

348

349

350

351

352

353

354

355

356 **Table 2: Genetic diversity indices of *map1* family genes.**

Gene	L	N	H	H. D	S	(π)	π (SS)	π (NS)
HG	716	14	7	0.8132	30	0.017(0.009)	27	8
map1	972	14	11	0.9670	117	0.124(0.063)	182	83
map1-1	858	14	9	0.9011	28	0.011(0.006)	28	1
map1+1	849	14	8	0.9451	33	0.022(0.011)	34	6
map1-2	966	14	9	0.9231	89	0.107(0.054)	115	44
map1-3	972	14	8	0.8901	48	0.019(0.010)	57	25
map1-4	906	14	11	0.9560	50	0.016(0.008)	50	14
map1-5	648	14	11	0.9560	57	0.064(0.033)	65	23
map1-6	900	14	9	0.9121	62	0.052(0.027)	74	25
map1-7	922	14	8	0.8242	37	0.017(0.009)	35	10
map1-8	859	14	9	0.8791	29	0.016(0.009)	30	12
map1-9	872	14	10	0.9451	31	0.013(0.007)	30	3
map1-10	781	14	6	0.8352	22	0.012(0.007)	21	9
map1-11	886	14	9	0.9231	26	0.012(0.007)	25	7
map1-12	831	14	8	0.8901	31	0.013(0.007)	31	3
map1-13	895	14	8	0.9011	57	0.031(0.016)	62	11
map1-14	984	14	11	0.9670	31	0.039(0.020)	34	16

357

358 L, length (bp). N, number of analysed sequences. H, number of haplotypes. HG,
359 hypothetical gene. S, number of segregating sites. π, pairwise sequence diversity (Jukes
360 and Cantor). SS, number of synonymous substitutions. NS, number of nonsynonymous
361 substitutions.

RFERENCES

Allsopp, M.T., Dorfling, C.M., Maillard, J.C., Bensaid, A., Haydon, D.T., van Heerden, H., Allsopp, B.A., 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 (11), 4200-4203.

Barbet, A.F., Byrom, B., Mahan, S.M., 2009. Diversity of *Ehrlichia ruminantium* major antigenic protein 1-2 in field isolates and infected sheep. Infect. Immun. 77 (6), 2304–2310.

Bekker, C.P.J., Postigo, M., Taoufik, A., Bell-Sakyi, L., Ferraz, C., Martinez, D., Jongejan, F., 2005. Transcription analysis of the major antigenic protein 1 multigene family of three in vitro-cultured *Ehrlichia ruminantium* isolates. J. Bacteriol. 187 (14), 4782–4791.

Bell-Sakyi, L., Koney, E.B.M., Dogbey, O., Abbam, J.A., Aning, K.G., 1997. Isolation and *in vitro* cultivation in Ghana of *Cowdria ruminantium*, the causative agent of heartwater., in: Presented at the Proceedings of the W.A.C.V.AG.V.M.A. Conference, Ministry of Food and Agriculture, Accra, pp. 46–51.

Brown, J.R., 2003. Ancient horizontal gene transfer. Nat. Rev. Genet. 4 (2), 121–132.

Burridge, M.J., Simmons, L.A., Peter, T.F., Mahan, S.M., 2002. Increasing risks of introduction of heartwater onto the American mainland associated with animal movements. Ann. N. Y. Acad. Sci. 969, 269–274.

Byrom, B., Yunker, C.E., Donovan, P.L., Smith, G.E., 1991. *In vitro* isolation of *Cowdria ruminantium* from plasma of infected ruminants. Vet. Microbiol. 26 (3), 263–268.

Charlesworth, B., Charlesworth, D., 2010. Elements of Evolutionary Genetics. Roberts & Company, Greenwood Village.

Charlesworth, D., Charlesworth, B., Morgan, M.T., 1995. The pattern of neutral molecular variation under the background selection model. Genetics. 141 (4), 1619-1632.

Cheng, C., Paddock, C.D., Reddy Ganta, R., 2003. Molecular heterogeneity of *Ehrlichia chaffeensis* isolates determined by sequence analysis of the 28-kilodalton outer membrane protein genes and other regions of the genome. Infect. Immun. 71 (1), 187–195.

Collins, N.E., Liebenberg, J., de Villiers, E.P., Brayton, K.A., Louw, E., Pretorius, A., Faber, F.E., van Heerden, H., Josemans, A., van Kleef, M., Steyn, H.C., van Strijp, M.F., Zweygarth, E., Jongejan, F., Maillard, J.C., Berthier, D., Botha, M., Joubert, F., Corton, C.H., Thomson, N.R., Allsopp, M.T., Allsopp, B.A., 2005. The genome of the heartwater agent *Ehrlichia ruminantium* contains multiple tandem repeats of actively variable copy number. Proc. Natl. Acad. Sci. U.S.A. 102 (3), 838-843.

396 Crocquet-Valdes, P.A., Thirumalapura, N.R., Ismail, N., Yu, X., Saito, T.B., Stevenson, H.L.,
397 Pietzsch, C.A., Thomas, S., Walker, D.H., 2011. Immunization with *Ehrlichia* P28 outer
398 membrane proteins confers protection in a mouse model of ehrlichiosis. Clin. Vaccine
399 Immunol. 18 (12), 2018–2025.
400 Eizaguirre, C., Lenz, T.L., Kalbe, M., Milinski, M., 2012. Rapid and adaptive evolution of
401 MHC genes under parasite selection in experimental vertebrate populations. Nat.
402 Commun. 3, 621.
403 Excoffier, L., Lischer, H.E.L., 2010. Arlequin suite ver 3.5: a new series of programs to
404 perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10
405 (3), 564-567.
406 Faburay, B., McGill, J., Jongejan, F., 2017. A glycosylated recombinant subunit candidate
407 vaccine consisting of *Ehrlichia ruminantium* major antigenic protein1 induces specific
408 humoral and Th1 type cell responses in sheep. PLoS One. 12 (9), e0185495.
409 Faburay, B., Munstermann, S., Geysen, D., Bell-Sakyi, L., Ceesay, A., Bodaan, C., Jongejan,
410 F., 2005. Point seroprevalence survey of *Ehrlichia ruminantium* infection in small
411 ruminants in The Gambia. Clin. Diagn. Lab. Immunol. 12 (4), 508–512.
412 Feil, E.J., 2004. Small change: Keeping pace with microevolution. Nat. Rev. Microbiol 2 (6),
413 483–495.
414 Fijarczyk A., Dudek, K., Babik, W., 2016. Selective landscapes in newt immune genes
415 inferred from patterns of nucleotide variation. Genome Biol. Evol. 8 (11), 3417-3432.
416 Finlay, B.B., Falkow, S., 1997. Common themes in microbial pathogenicity revisited.
417 Microbiol. Mol. Biol. Rev. 61 (2), 136–169.
418 Frutos, R., Viari, A., Ferraz, C., Morgat, A., Eychenié, S., Kandassamy, Y., Chantal, I.,
419 Bensaid, A., Coissac, E., Vachiery, N., Demaille, J., Martinez, D., 2006. Comparative
420 genomic analysis of three strains of *Ehrlichia ruminantium* reveals an active process of
421 genome size plasticity. J. Bacteriol. 188 (7), 2533–2542.
422 Fu YX. 1997. Statistical tests of neutrality of mutations against population growth,
423 hitchhiking and background selection. Genetics. 147 (2), 915-925.
424 Gal-Mor, O., Finlay, B.B., 2006. Pathogenicity islands: a molecular toolbox for bacterial
425 virulence. Cell Microbiol. 8 (11), 1707–1719.
426 Ge, Y., Rikihisa, Y., 2007. Surface-exposed proteins of *Ehrlichia chaffeensis*. Infect. Immun.
427 75 (8), 3833–3841.
428 Harpending, H.C., 1994. Signature of ancient population growth in a low-resolution
429 mitochondrial DNA mismatch distribution. Hum. Biol. 66 (4), 591-600.
430 Hughes, A.L., French, J.O., 2007. Homologous recombination and the pattern of nucleotide
431 substitution in *Ehrlichia ruminantium*. Gene. 387(1-2), 31-37.

Hughes, D. 1999. Impact of homologous recombination on genome organization and stability. In: Charlebois RL, editor. Organization of the prokaryotic genome. Washington (DC): ASM Press. pp. 109–128

Ilemobade, A.A., Blotkamp, C., 1978. Heartwater in Nigeria. II. The isolation of *Cowdria ruminantium* from live and dead animals and the importance of routes of inoculation. Trop. Anim. Health Prod. 10 (1), 39-44.

Jongejan, F., Morzaria, S.P., Omer, A.S., Hashim, M.A., 1984. Isolation and transmission of *Cowdria ruminantium* (causal agent of heartwater disease) in Blue Nile Province, Sudan. Vet. Res. Commun. 8 (2), 141-145.

Jongejan, F., Thielemans, M.J., 1989. Identification of an immunodominant antigenically conserved 32-kilodalton protein from *Cowdria ruminantium*. Infect. Immun. 57 (10), 3243–3246.

Jongejan, F., Uilenberg, G., Franssen, F.F., Gueye, A., Nieuwenhuijs, J., 1988. Antigenic differences between stocks of *Cowdria ruminantium*. Res. Vet. Sci. 44 (2), 186–189.

Jongejan, F., Van Winkelhoff, A.J., Uilenberg, G., 1980. *Cowdria ruminantium* (Rickettsiales) in primary goat kidney cell cultures. Res. Vet. Sci. 29 (3), 392–393.

Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., Frost, S.D., 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Mol. Biol. Evol. 23 (10), 1891–1901.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0. for bigger datasets. Mol. Biol. Evol. 33 (7), 1870-1874.

Kuzminov, A., 1999. Recombinational repair of DNA damage in Escherichia coli and bacteriophage lambda. Microbiol. Mol. Biol. Rev. 63 (4), 751–813.

Lawrence, J.G., Ochman, H., 1998. Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. U. S. A. 95 (16), 9413–9417.

Levin, B.R., Cornejo, O.E., 2009. The population and evolutionary dynamics of homologous gene recombination in bacterial populations. PLoS Genet. 5 (8), e1000601.

Li, J.S.-Y., Chu, F., Reilly, A., Winslow, G.M., 2002. Antibodies highly effective in SCID mice during infection by the intracellular bacterium *Ehrlichia chaffeensis* are of picomolar affinity and exhibit preferential epitope and isotype utilization. J. Immunol. 169 (3), 1419–1425.

Librado, P., Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 25 (11), 1451-1452.

Lin, J., Huang, S., Zhang, Q., 2002. Outer membrane proteins: key players for bacterial adaptation in host niches. Microbes Infect. 4 (3), 325–331.

Lorenz, M.G., Wackernagel, W. 1994. Bacterial gene transfer by natural genetic transformation in the environment. Microbiol. Rev. 58 (3), 563–602.

Mackenzie, P.K.I., van Rooyen, R.E., 1981. Isolation and culture of *Cowdria ruminantium* in albino mice. Tick biology and control: proceedings of an International Conference January 27-29, 1981. Edited by G.B. Whitehead and J.D. Gibson.

Marcelino, I., de Almeida, A.M., Brito, C., Meyer, D.F., Barreto, M., Sheikboudou, C., Franco, C.F., Martinez, D., Lefrançois, T., Vachiery, N., Carrondo, M.J.T., Coelho, A.V., Alves, P.M., 2012. Proteomic analyses of *Ehrlichia ruminantium* highlight differential expression of MAP1-family proteins. Vet. Microbiol. 156 (3-4), 305–314.

Mazodier, P., Davies, J. 1991. Gene transfer between distantly related bacteria. Annu. Rev. Genet. 25, 147–171.

McBride, J.W., Corstvet, R.E., Gaunt, S.D., Boudreaux, C., Guedry, T., Walker, D.H., 2003. Kinetics of antibody response to *Ehrlichia canis* immunoreactive proteins. Infect. Immun. 71 (5), 2516-2524.

McBride, J.W., Yu, X.J., Walker, D.H., 1999. Molecular cloning of the gene for a conserved major immunoreactive 28-kilodalton protein of *Ehrlichia canis*: a potential serodiagnostic antigen. Clin. Diagn. Lab. Immunol. 6 (3), 392–399.

Mehr, I.J., Seifert, H.S., 1998. Differential roles of homologous recombination pathways in Neisseria gonorrhoeae pilin antigenic variation, DNA transformation, and DNA repair. Mol. Microbiol. 30 (4), 697–710.

Michel, B., Grompone, G., Flore`s, M.J., Bidnenko, V., 2004. Multiple pathways process stalled replication forks. Proc. Natl. Acad. Sci. U. S. A. 101 (35), 12783–12788.

Mukhebi, A.W., Chamboko, T., O'Callaghan, C.J., Peter, T.F., Kruska, R.L., Medley, G.F., Mahan, S.M., Perry, B.D., 1999. An assessment of the economic impact of heartwater (*Cowdria ruminantium* infection) and its control in Zimbabwe. Prev. Vet. Med. 39 (3), 173–189.

Nakao, R., Jongejan, F., Sugimoto, C., 2016. Draft genome sequences of three strains of *Ehrlichia ruminantium*, a tick-borne pathogen of ruminants, isolated from Zimbabwe, The Gambia, and Ghana. Genome Announc. 4, e00453–16.

Nakao, R., Magona, J.W., Zhou, L., Jongejan, F., Sugimoto, C., 2011. Multi-locus sequence typing of *Ehrlichia ruminantium* strains from geographically diverse origins and collected in *Amblyomma variegatum* from Uganda. Parasit. Vectors 4, 137.

Nakao, R., Stromdahl, E.Y., Magona, J.W., Faburay, B., Namangala, B., Malele, I., Inoue, N., Geysen, D., Kajino, K., Jongejan, F., Sugimoto, C., 2010. Development of Loop-Mediated Isothermal Amplification (LAMP) Assays for Rapid Detection of *Ehrlichia ruminantium*. BMC Microbiol. 10, 296.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn. M.L., Dodson, R.J., et al. 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399 (6734), 323–329.

Nyika, A., Barbet, A.F., Burridge, M.J., Mahan, S.M., 2002. DNA vaccination with *map1* gene followed by protein boost augments protection against challenge with *Cowdria ruminantium*, the agent of heartwater. Vaccine 20 (7-8), 1215–1225.

Ochman, H., Lawrence, J.G., Groisman, E.A., 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405 (6784), 299–304.

Ohashi, N., Rikihisa, Y., Unver, A., 2001. Analysis of transcriptionally active gene clusters of major outer membrane protein multigene family in *Ehrlichia canis* and *E. chaffeensis*. Infect. Immun. 69 (4), 2083–2091.

Ohashi, N., Unver, A., Zhi, N., Rikihisa, Y., 1998a. Cloning and characterization of multigenes encoding the immunodominant 30-kilodalton major outer membrane proteins *of Ehrlichia canis* and application of the recombinant protein for serodiagnosis. J. Clin. Microbiol. 36 (9), 2671–2680.

Ohashi, N., Zhi, N., Zhang, Y., Rikihisa, Y., 1998b. Immunodominant major outer membrane proteins of *Ehrlichia chaffeensis* are encoded by a polymorphic multigene family. Infect. Immun. 66 (1), 132–139.

Otto, S.P., Michalakis, Y., 1998. The evolution of recombination in changing environments. Trends Ecol. Evol. 13 (4), 145–151.

Peter, T.F., O'Callaghan, C.J., Medley, G.F., Perry, B.D., Semu, S.M., Maha, S.M., 2001. Population-based evaluation of the *Ehrlichia ruminantium* MAP 1B indirect ELISA. Exp. Appl. Acarol. 25 (10-11), 881–897.

Phillips, K.P., Cable, J., Mohammed, R.S., Herdegen-Radwan, M., Raubic, J., Przesmycka, K.J., van Oosterhout, C., Radwan, J., 2018. Immunogenetic novelty confers a selective advantage in host-pathogen coevolution. Proc. Natl. Acad. Sci. U S A. 115 (7):1552-1557.

Plessis, Du, 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 (2), 55-61.

Postigo, M., Taoufik, A., Bell-Sakyi, L., Bekker, C.P.J., de Vries, E., Morrison, W.I., Jongejan, F., 2008. Host cell-specific protein expression in vitro in *Ehrlichia ruminantium*. Vet. Microbiol. 128 (1-2), 136–147.

Postigo, M., Taoufik, A., Bell-Sakyi, L., de Vries, E., Morrison, W.I., Jongejan, F., 2007. Differential transcription of the major antigenic protein 1 multigene family of *Ehrlichia ruminantium* in *Amblyomma variegatum* ticks. Vet. Microbiol. 122 (3-4), 298–305.

538 Reddy, G.R., Streck, C.P., 2000. Variability in the 28-kDa surface antigen protein multigene
539 locus of isolates of the emerging disease agent *Ehrlichia chaffeensis* suggests that it plays
540 a role in immune evasion. Mol. Cell Biol. Res. Commun. 1 (3), 167-175.
541 Richman, A.D., Kohn, J. R., 1999. Self-incompatibility alleles from Physalis: implications for
542 historical inference from balanced genetic polymorphisms. Proc. Natl Acad. Sci. U. S. A.
543 96 (1), 168–172.
544 Rocha E.P., 2003. DNA repeats lead to the accelerated loss of gene order in bacteria. Trends
545 Genet. 19 (11), 600–604.
546 Rocha, E.P, 2004. Order and disorder in bacterial genomes. Curr. Opin. Microbiol. 7 (5),
547 519–527.
548 Rocha, E.P.C., Blanchard, A., 2002. Genomic repeats, genome plasticity and the dynamics of
549 *Mycoplasma* evolution. Nucleic Acids Res. 30 (9), 2031–2042.
550 Rogers, A.R., Harpending, H., 1992. Population growth makes waves in the distribution of
551 pairwise genetic differences. Mol. Biol. Evol. 9 (3):552-69.
552 Smith, G.R., 1991. Conjugational recombination in *E. coli*: Myths and mechanisms. Cell 64
553 (1), 19–27.
554 Sugiura, N., 1978. Further analysis of the data by Akaike's information criterion and the
555 finite corrections. Commun. Statist. A7, 13–26.
556 Sulsona, C.R., Mahan, S.M., Barbet, A.F., 1999. The *map1* gene of *Cowdria ruminantium* is a
557 member of a multigene family containing both conserved and variable genes. Biochem.
558 Biophys. Res. Commun. 257 (2), 300–305.
559 Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA
560 polymorphism. Genetics 123 (3), 585–595.
561 Uilenberg, G., Camus, E., Barre, N., 1985. A strain of *Cowdria ruminantium* isolated in
562 Guadeloupe (French West Indies). Rev. Elev. Med. Vet. Pays. Trop. 38 (1), 34–42.
563 van Heerden, H., Collins, N.E., Brayton, K.A., Rademeyer, C., Allsopp, B.A., 2004.
564 Characterization of a major outer membrane protein multigene family in *Ehrlichia*
565 *ruminantium*. Gene 330, 159–168.
566 Weaver, S., Shank, S.D., Spielman, S.J., Li, M., Muse, S.V., Kosakovsky Pond S.L., 2018.
567 Datamonkey 2.0: a modern web application for characterizing selective and other
568 evolutionary processes. Mol. Biol. Evol. In press.
569 Weedall, G.D., Conway, D.J., 2010. Detecting signatures of balancing selection to identify
570 targets of anti-parasite immunity. Trends Parasitol. 26 (7), 363-369.
571 Yu, X.J., McBride, J.W., Walker, D.H., 1999. Genetic diversity of the 28-kilodalton outer
572 membrane protein gene in human isolates of *Ehrlichia chaffeensis*. J. Clin. Microbiol. 37
573 (4), 1137–1143.

574 **FIGURE LEGENDS**

575 **Figure 1. A schematic representation of *map1* family genes.**

576 **Figure 2. Tajima's D neutrality test scenario for *map1* family genes.** Positive and
577 negative values indicate balancing and directional selections, respectively.

578 **Figure 3**. **The number of homologous recombination points (potential breakpoints) in**
579 ***map1* family genes.** It illustrates that *map1* and *map1*-2 are experiencing the highest
580 recombination.

581 **Figure 4. Actual locations of recombination breakpoints in the four highly**
582 **recombining genes.** Three positions of breakpoints in *map1* and *map1-2* were observed at
583 locations 150, 225, 275-280, and 95, 340-350, 505, respectively. Only two positions of the
584 breakpoints in *map1-6* and *map1-14* were observed at locations 490 and 495, and 190 and
585 200, respectively.

586 **Figure 5. Detection of negative/purifying selection in *map1* family genes.** It indicates
587 that *map1* was subjected to the highest amount of purifying selection compared to the
588 following gene *map1-2* and the rest of the gene family.

589 **Figure 6. Genealogies and the effect of balancing selection.** The trees show the effect
590 balancing selection on tree topologies, giving rise to long internal branches proportional to
591 the amount of selection in *map1* (A) and *map1-2* (B). No balancing selection was reflected
592 in the short internal branches as seen in *map1-10* (C) and *map1-12* (D).

Fig. 1

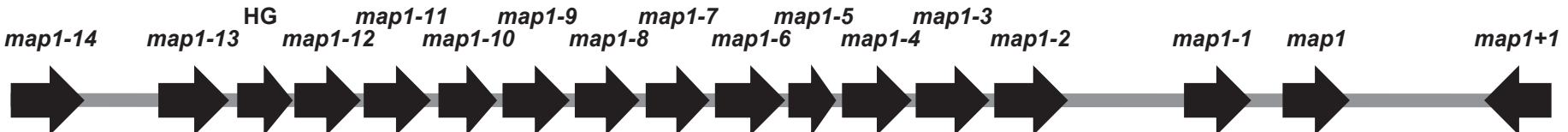

map1-14
map1-13
HG
map1-12
map1-11
map1-10
map1-9
map1-8
map1-7
map1-6
map1-5
map1-4
map1-3
map1-2
map1-1
map1
map1+1

Fig. 2

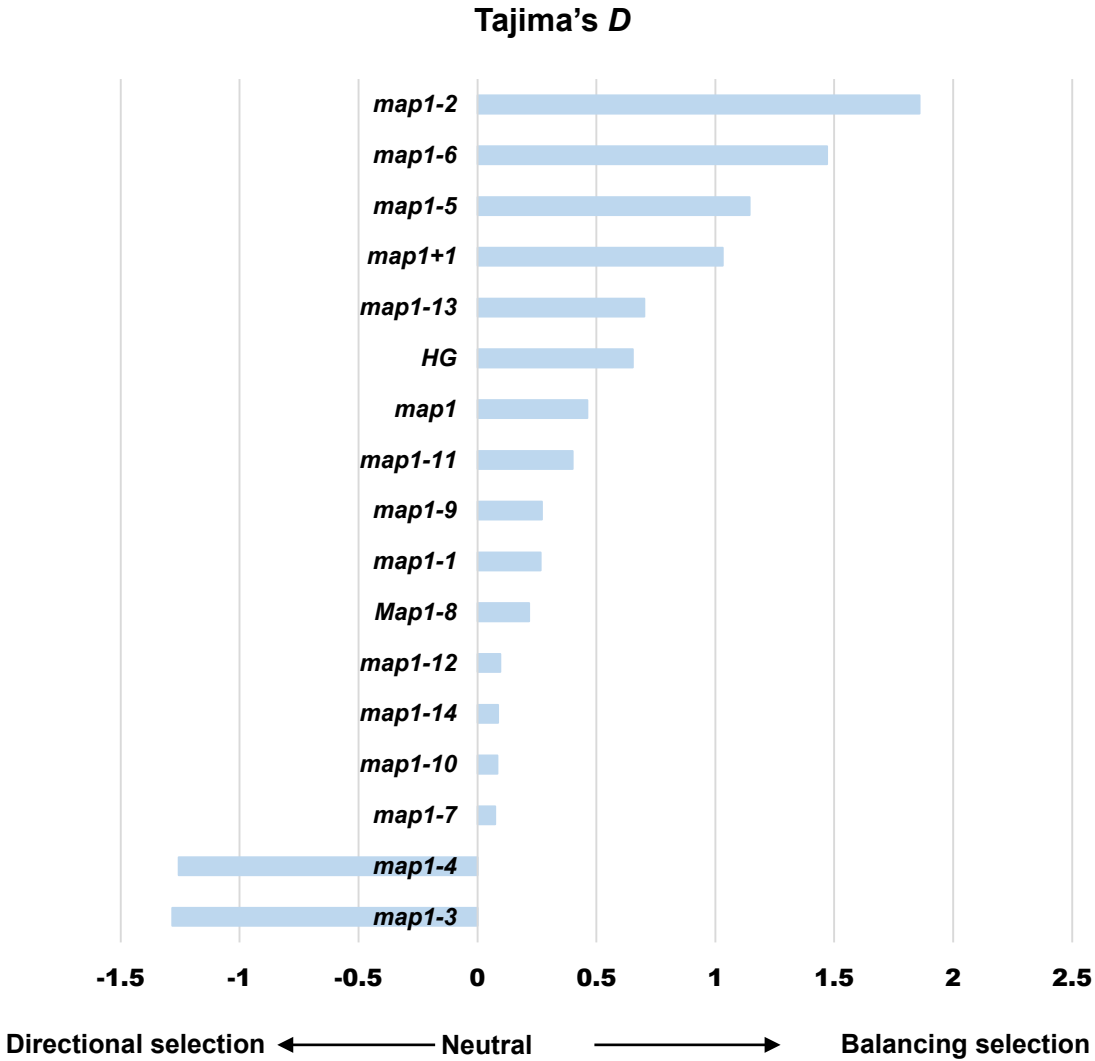

Tajima's D
map1-2
map1-6
map1-5
map1+1
map1-13
HG
map1
map1-11
map1-9
map1-1
Map1-8
map1-12
map1-14
map1-10
map1-7
map1-4
map1-3
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
Directional selection
Neutral
Balancing selection

Fig. 3

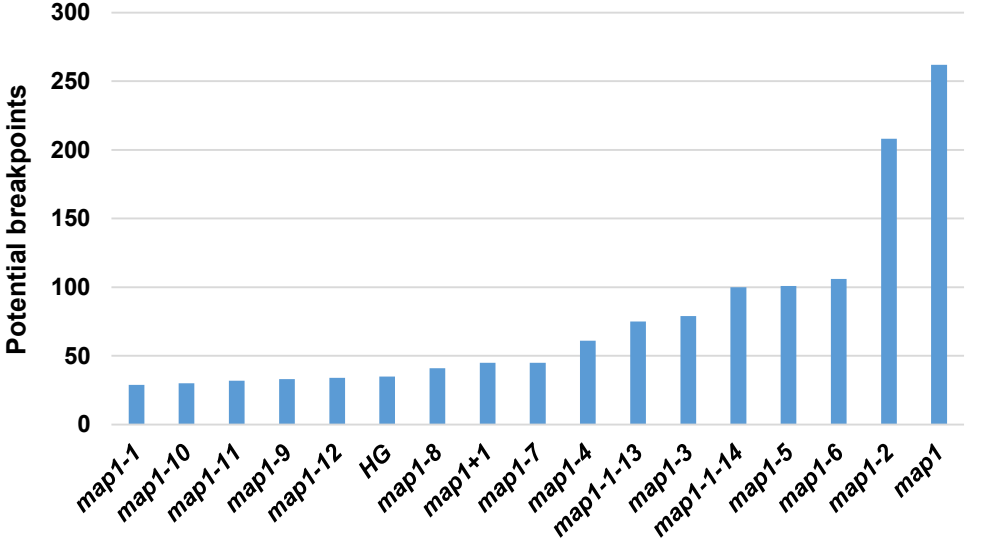

Potential breakpoints
300
250
200
150
100
50
0
map1-1
map1-10
map1-11
map1-9
map1-12
HG
map1-8
map1+1
map1-7
map1-4
map1-1-13
map1-3
map1-1-14
map1-5
map1-6
map1-2
map1

Fig. 4

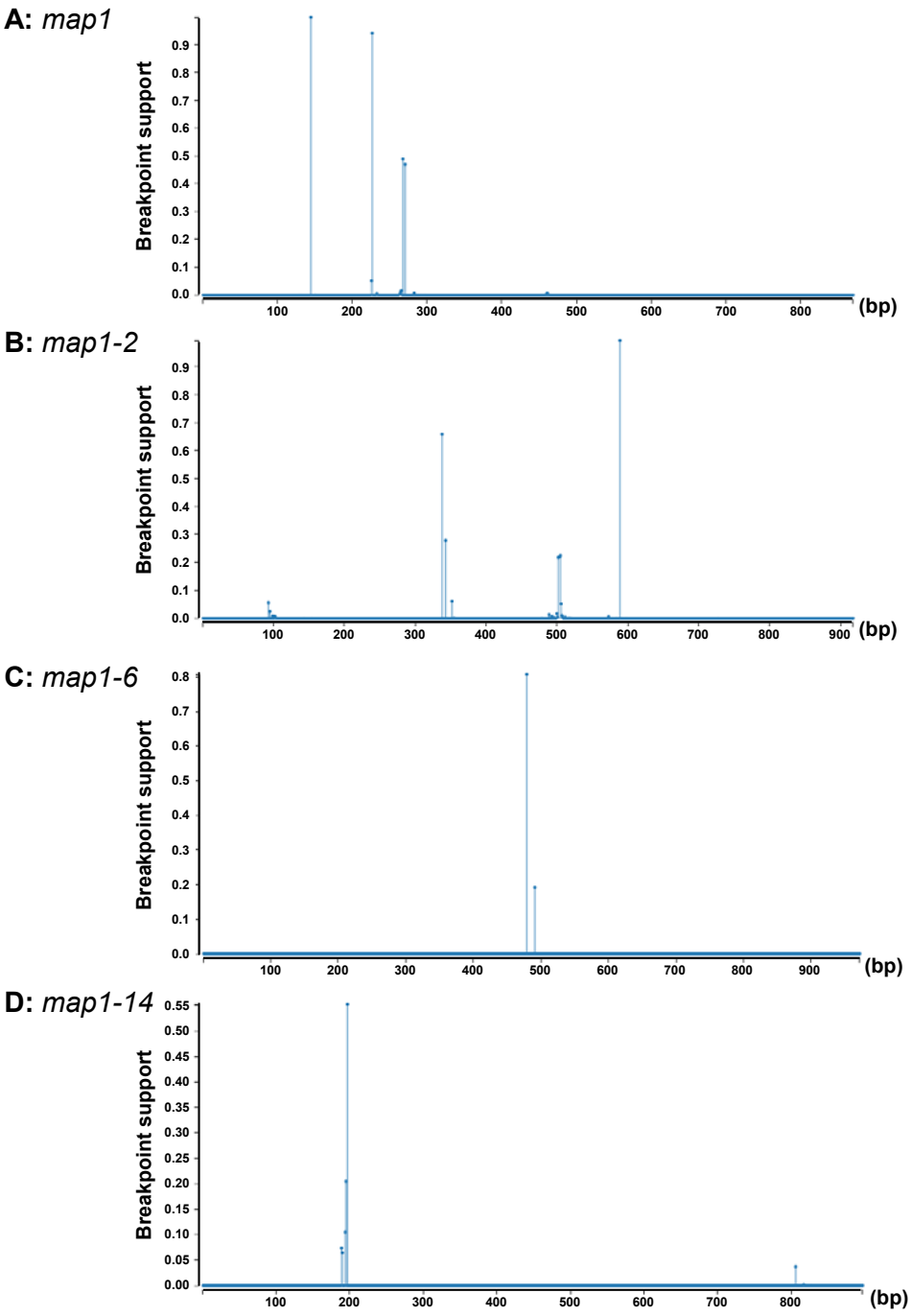

A: map1
Breakpoint support
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
100
200
300
400
500
600
700
800
(bp)
B: map1-2
Breakpoint support
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
100
200
300
400
500
600
700
800
900
(bp)
C: map1-6
Breakpoint support
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
100
200
300
400
500
600
700
800
900
(bp)
D: map1-14
Breakpoint support
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
100
200
300
400
500
600
700
800
(bp)

Fig. 5

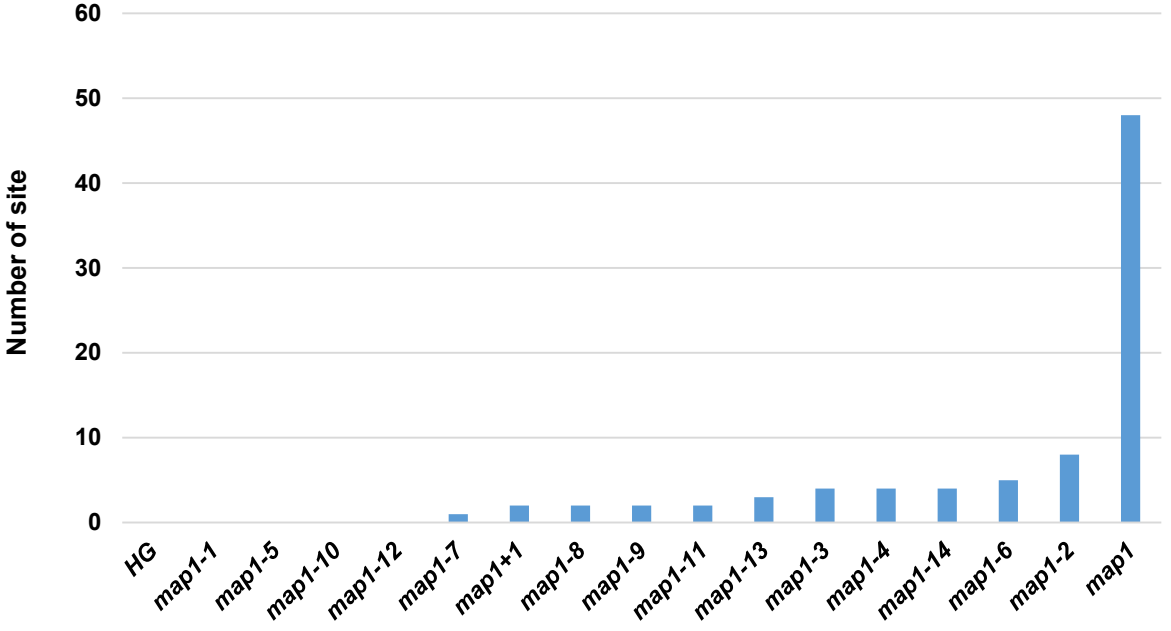
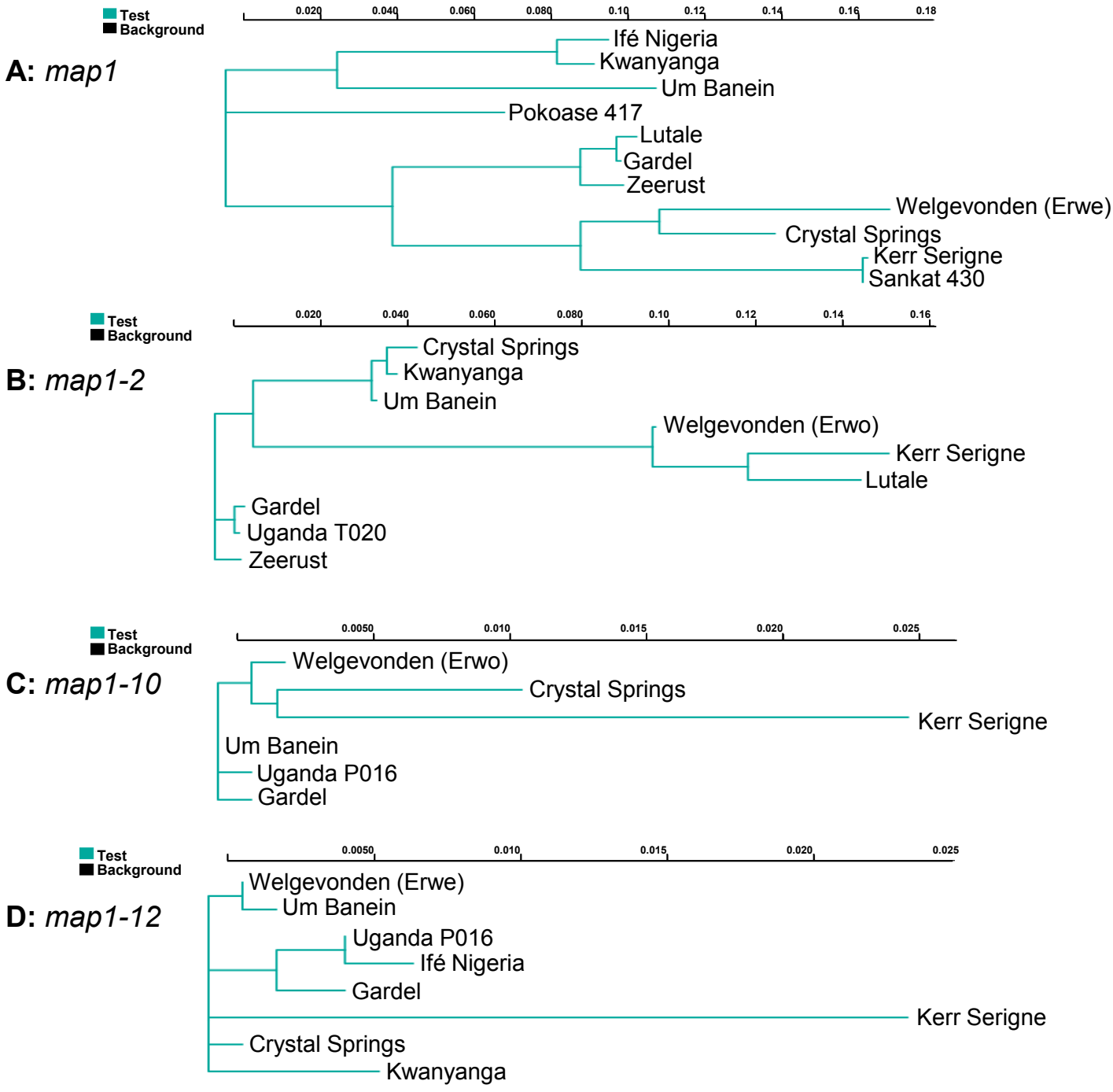

Number of site
60
50
40
30
20
10
0
HG
map1-1
map1-5
map1-10
map1-12
map1-7
map1+1
map1-8
map1-9
map1-11
map1-13
map1-3
map1-4
map1-14
map1-6
map1-2
map1

Fig. 6


Test
Background
0.020
0.040
0.060
0.080
0.10
0.12
0.14
0.16
0.18
A: map1
Ifé Nigeria
Kwanyanga
Um Banein
Pokoase 417
Lutale
Gardel
Zeerust
Welgevonden (Erwe)
Crystal Springs
Kerr Serigne
Sankat 430
Test
Background
0.020
0.040
0.060
0.080
0.10
0.12
0.14
0.16
B: map1-2
Crystal Springs
Kwanyanga
Um Banein
Welgevonden (Erwo)
Kerr Serigne
Lutale
Gardel
Uganda T020
Zeerust
Test
Background
0.0050
0.010
0.015
0.020
0.025
C: map1-10
Welgevonden (Erwo)
Crystal Springs
Kerr Serigne
Um Banein
Uganda P016
Gardel
Test
Background
0.0050
0.010
0.015
0.020
0.025
D: map1-12
Welgevonden (Erwe)
Um Banein
Uganda P016
Ifé Nigeria
Gardel
Kerr Serigne
Crystal Springs
Kwanyanga