DNA Research Advance Access published February 12, 2013

DNA Research pp. 1–12, (2013) doi:10.1093/dnares/dst003

Whole-Genome Sequencing of *Theileria parva* Strains Provides Insight into Parasite Migration and Diversification in the African Continent

Kyoko Hayashida[1], Takashi Abe[2], William Weir[3], Ryo Nakao[1,4], Kimihito Ito[4], Kiichi Kajino[1], Yutaka Suzuki[5], Frans Jongejan[6,7], Dirk Geysen[8], and Chihiro Sugimoto[1,*]

Division of Collaboration and Education, Research Center for Zoonosis Control, Hokkaido University, Sapporo-shi, Hokkaido 001-0020, Japan[1]; Information Engineering, Niigata University, Niigata-shi, Niigata 950-2181, Japan[2]; Institute of Comparative Medicine, Glasgow University Veterinary School, Glasgow G61 1QH, UK[3]; Division of Bioinformatics, Research Center for Zoonosis Control, Hokkaido University, Sapporo-shi, Hokkaido 001-0020, Japan[4]; Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa-shi, Chiba 277-8568, Japan[5]; Department of Infectious Diseases and Immunology, Faculty of Veterinary Medicine, Utrecht Centre for Tick-borne Diseases (UCTD), Utrecht University, Yalelaan 1, Utrecht 3584CL, The Netherlands[6]; Department of Veterinary Tropical Diseases, Faculty of Veterinary Science, University of Pretoria, Private Bag X04, Onderstepoort 0110, South Africa[7] and Department of Animal Health, Institute of Tropical Medicine, Nationalestraat 155, Antwerp 2000, Belgium[8]

*To whom correspondence should be addressed. Tel. +81 11-706-5297. Fax. +81 11-706-7370.
Email: sugimoto@czc.hokudai.ac.jp

Edited by Dr Takao Sekiya
(Received 1 October 2012; accepted 21 January 2013)

Abstract

The disease caused by the apicomplexan protozoan parasite *Theileria parva*, known as East Coast fever or Corridor disease, is one of the most serious cattle diseases in Eastern, Central, and Southern Africa. We performed whole-genome sequencing of nine *T. parva* strains, including one of the vaccine strains (Kiambu 5), field isolates from Zambia, Uganda, Tanzania, or Rwanda, and two buffalo-derived strains. Comparison with the reference Muguga genome sequence revealed 34 814–121 545 single nucleotide polymorphisms (SNPs) that were more abundant in buffalo-derived strains. High-resolution phylogenetic trees were constructed with selected informative SNPs that allowed the investigation of possible complex recombination events among ancestors of the extant strains. We further analysed the dN/dS ratio (non-synonymous substitutions per non-synonymous site divided by synonymous substitutions per synonymous site) for 4011 coding genes to estimate potential selective pressure. Genes under possible positive selection were identified that may, in turn, assist in the identification of immunogenic proteins or vaccine candidates. This study elucidated the phylogeny of *T. parva* strains based on genome-wide SNPs analysis with prediction of possible past recombination events, providing insight into the migration, diversification, and evolution of this parasite species in the African continent.

Key words: *Theileria parva*; genome sequence; SNPs; recombination; dN/dS

1. Introduction

Theileria parva is a tick-borne protozoan parasite belonging to the phylum Apicomplexa. Infection of *T. parva* in cattle causes a severe disease known as East Coast fever (ECF) or Corridor disease.[1–3] The disease is endemic in East African countries, where it has caused a serious economical problem to the livestock industry. Although the mortality in cattle may reach 100%, especially in exotic breeds, the Cape buffalo (*Syncerus caffer*) shows no clinical signs and is considered to be the main natural host. Although clinical differences have been documented,[4] ECF and Corridor disease have similar presentations. However,

a major epidemiological difference is that, whereas ECF spreads from cattle to cattle, Corridor disease is believed to be transmitted solely from buffalo to cattle. The parasites causing ECF and Corridor disease were designated as *T. p. parva* and *T. p. lawrencei*, respectively.[3]

Vaccination against ECF is based on an infection and treatment method that involves inoculation of live sporozoite-stage parasites and simultaneous treatment with long-acting tetracycline.[5] The Muguga cocktail, consisting of the three strains of Muguga, Serengeti-transformed, and Kiambu 5, is the most widely used vaccine in East Africa. Importantly, there is an extensive debate concerning the risk of vaccination with live non-attenuated sporozoites such as the Muguga cocktail vaccine, as the vaccination may introduce parasites with an exotic genetic background into the local parasite population.[6–9] This was proven to be a real risk when Oura et al.[7] demonstrated the transmission of a strain of vaccine constituent to unvaccinated cattle under field conditions in Uganda. In addition, the presence of the vaccine component strain (Muguga or Serengeti-transformed) was confirmed in clinical cases of ECF in the Southern Province of Zambia,[6] following deployment of the Muguga Cocktail over a 7-year period, ranging from 1986 to 1992. Therefore, two indigenous Zambian strains (Katete and Chitongo) have been used as a vaccine in the Eastern and Southern Provinces of Zambia,[10] although the consequences of this vaccination have not been analysed.

Given that *Theileria* parasites could recombine between divergent strains during the sexual stage in ticks, vaccine-derived 'exotic' and 'local' strains could exchange genetic information, resulting in parasites with genetic mosaics and diversity. In addition to the problems with the current vaccine, quality control of the cocktail vaccine in terms of the composition of each component is difficult. This may be related to recombination and selection during the maintenance and passage of the stabilates through ticks.[11] Thus, precise and reliable methods for parasite genotyping or phenotyping during vaccine production and its field application are required.

Genetic diversity between different *T. parva* strains has been assessed using various approaches, including polymerase chain reaction (PCR) or PCR-restriction fragment length polymorphism (RFLP) of polymorphic antigen-encoding genes,[6,12] or the indirect immunofluorescence assay (IFA) using monoclonal antibodies against the surface protein, the polymorphic immunodominant molecule (PIM).[13] A panel of micro- and mini-satellite markers has also been developed[14,15] that is widely used in the genetic analysis of field populations[7,8] and has also been used to characterize vaccine stabilates[11] and genetic recombination analysis.[16–18] However, the resolution of genetic differentiation in these studies is limited because of the relatively low marker density.

In this study, we carried out the whole-genome sequencing of nine *T. parva* strains, comprising seven cattle-derived and two buffalo-derived strains, using next-generation sequencing technology. Genome-wide comparison of strains revealed genetic polymorphisms on a fine scale and was used to infer phylogenetic relationships among the parasites. The analysis enabled us to determine potential immune selective pressures against parasite genes, which may prove useful in identifying potential antigens. Moreover, the allelic diversity pattern among strains gave us insight into the evolution, diversification, and migration of this parasite in the African continent.

2. Materials and methods

2.1. Parasite strains

In total, nine strains of *T. parva*, mainly isolated in the 1980s, were used in this study. The place and the year isolated are shown in Table 1. These strains were originally isolated in ticks from infected cows and cultured as schizont-infected bovine lymphocyte cell lines. ChitongoZ2and KateteB2 have been used as sporozoite stabilate vaccines in the Eastern and Southern Provinces of Zambia.[10] Kiambu 5[19] is one of the Muguga cocktail vaccine components, and KiambuZ464/C12 is a strain that has been cloned out from Kiambu 5 (Kenya, stabilate 68). Zambian strains KateteB2, ChitongoZ2, and MandaliZ22H10 were isolated before the introduction of the Muguga cocktail into Zambia, thus representing ECF epidemiology in Zambia, excluding human-induced genetic contamination. In addition, the analysis included two buffalo-derived isolates, LAWR and Z5E5. Z5E5 is a buffalo-type isolate obtained from a bovine, whereas LAWR is a buffalo-type isolate obtained from a buffalo. KiambuZ464/C12, MandaliZ22H10, and Z5E5 were cloned by limiting dilution. These *Theileria*-infected cell lines did not undergo extensive passages (<30 passage) and were stored in liquid nitrogen until use. Cultures were maintained in Roswell Park Memorial Institute (RPMI) -1640 culture medium containing 10 or 20% heat-inactivated fetal bovine serum, 50 μM 2-mercaptoethanol, 50 units/ml penicillin, and 50 mg/ml streptomycin.

2.2. Parasite purification and genomic DNA preparation

Schizont-enriched material was prepared from the infected lymphocytes by a density-gradient separation method as previously described,[20–22] with some modifications. The cells were treated with

Table 1. *T. parva* strains sequenced in this study with the summary of Solexa sequence results

Strain name	Place isolated	Isolated year	Total reads obtained	Reference genome mapped reads	Mapped read (%)	Average coverage	Genome covered (%)	SNP number			SNP density (per 1 kb)		
								Overall	Coding	Non-coding	Overall	Coding	Non-coding
ChitongoZ2	Zambia	1982	14 405 285	11 225 629	77.9	49.1	97.4	46 366	31 753	14 613	5.63	5.48	5.99
KateteB2	Zambia	1989	16 558 765	4 954 291	29.9	21.3	97.3	43 873	31 533	12 340	5.33	5.44	5.06
Kiambu Z464/C12	Kenya	1972	15 848 447	6 278 932	39.6	27.4	97.2	46 435	33 021	13 414	5.64	5.70	5.50
MandaliZ22H10	Zambia	1985	16 362 287	3 904 897	23.9	17.1	97	38 498	28 270	10 228	4.67	4.88	4.19
Entebbe	Uganda	1980	10 171 312	3 547 208	34.9	15.5	95.2	34 814	27 195	7619	4.23	4.69	3.12
Nyakizu	Rwanda	1979	29 366 782	5 710 634	19.4	25	97	51 790	34 700	17 090	6.29	5.99	7.01
Katumba	Tanzania	1981	35 406 725	4 089 736	11.6	17.9	97.1	46 441	32 321	14 120	5.64	5.58	5.79
Buffalo LAWR	Kenya	1990	17 072 360	6 155 888	36.1	26.9	94.7	121 545	77 472	44 073	14.76	13.37	18.07
Buffalo Z5E5	Zambia	1982	14 821 054	5 119 542	34.5	22.4	95.3	103 880	68 454	35 426	12.61	11.81	14.52

3 μM nocodazole for 18 h, and then harvested cells were lyzed for 30–60 min at room temperature with a Gram-negative bacterium, *Aeromonas hydrophila* (AH-1)-derived haemolysin, in a suspension of HEPES-$CaCl_2$ (10 mM-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 150 mM NaCl, 20 mM KCl, and 1 mM $CaCl_2$, pH 7.4) to obtain a cell concentration of 4×10^7 cells/ml ($0.5–2 \times 10^8$ cells in total). Crude AH-1 haemolysin was prepared by bacterial culture supernatant according to a previously described method[23] and was added to the cell suspension at a final concentration of 100 U/ml. Lysis of infected lymphocytes was observed under a microscope. If complete cell lysis was not observed after 15 min, then the incubation period was prolonged until almost 100% of cells were lyzed, whereas schizonts remained intact. Because the sensitivity of schizont-infected cells varied significantly between cell lines, the maximum incubation time was 120 min. After lysis, the suspension was washed with HEPES-$CaCl_2$ and re-suspended in 3 ml of HEPES-ethylenediamine tetraacetic acid (EDTA) (10 mM HEPES, 150 mM NaCl, 20 mM KCl, and 5 mM $CaCl_2$, pH 7.4). Four layers of Percoll solution comprising 10, 10, 5, and 5 ml of 65, 40, 30, and 20% Percoll in HEPES-EDTA, respectively, were prepared in an ultracentrifuge tube. The cell lysate was overlaid on top of the Percoll solution and ultracentrifuged at 87 000 g for 30 min at 4°C, using a SW41 rotor (Beckman, USA). The schizont layer that formed at the interface between 40 and 65% Percoll solutions was carefully collected with a Pasteur pipette and then washed in phosphate-buffered saline (PBS) to remove the Percoll. A sample of each schizont preparation was stained with Giemsa, and preparations with negligible amounts of contamination with host-cell components were subjected to DNA isolation.

2.3. *DNA preparation, whole-genome amplification, and Illumina genome analyzer II (GAII) sequencing*

Genomic DNA was prepared from the purified schizonts using the NucleoSpin Tissue XS protocol (Machery-Nagel, Duren, Germany). Whole-genome amplification was performed on 10 ng of the total template DNA using an Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare) following the manufacturer's instructions.[24,25] The obtained DNA was purified by ethanol precipitation and subjected to sequence analysis. A 36 nucleotide, single-end sequence run was performed on the Illumina GAII Analyzer following the manufacturer's protocols (Illumina, San Diego, GA, USA).

The obtained reads, as listed in Table 1, were mapped on the 8 235 476 bp sequence of the *T. parva* Muguga strain (AAGK01000001, AAGK01000002, AAGK01000005,

AAGK01000006, and AAGK01000004) using the CLC Genomics Workbench (CLC bio, Aarhus, Denmark, Version 4.0.2). The ungapped alignment algorithm was used for all alignments, keeping the default parameters for mismatch and deletion costs (mismatch cost = 2, deletion cost = 1). Files containing these short sequence reads were submitted to the DDBJ Sequence Read Archive (accession number DRA000613).

2.4. *Single nucleotide polymorphisms (SNPs) analysis*

Three sets of single nucleotide polymorphisms (SNPs) were defined (Supplementary Fig. S1). SNPs were identified by comparing each re-sequenced genome with the reference Muguga strain.[26] SNP detection was performed using the SNP detection tool in the CLC Genomic Workbench with the default parameters (window length = 11, maximum number of gaps and mismatch = 2, minimum average quality surrounding bases = 15, and minimum quality of central base = 20),[27] except for the minimum coverage that was set at five reads, and the list was manually curated to include only SNPs, where all reads within a single sample agreed (SNP dataset I). The extracted SNPs data were exported and analysed by Microsoft Office Excel 2010. SNP dataset I was used for creating a SNPs density map and for dN/dS analysis. From SNP dataset I, SNPs identified among the eight bovine *T. parva* strains were extracted. To avoid calling block substitutions as SNPs, SNPs were only selected, if they did not exist within 100 bp of another SNP, and this provided SNP dataset II. Allelic data from each strain were extracted, and this information was used for the allelic combination and recombination analysis. SNP dataset III was created using the eight cattle-derived and two buffalo-derived strains, and again SNP positions were required to have at least 100 bp intervals. Thus, the high stringency dataset encompassing all 10 *Theileria* strains (including the reference strain), SNP dataset III, was used for phylogenetic analysis. Plots of the allele combination pattern for each chromosome were generated using freeware and open-source R software version 2.11.1 (R Development Core Team, 2010; http://www.R-project.org). Genes under selection pressure were estimated by calculating the dN/dS between strains with the SNP dataset I by the method of Yang et al.,[28] implemented in the PAML package.[29] Signal sequences for all the annotated genes of the Muguga strain were predicted using SignalP v4.0.[30]

2.5. *Phylogenetic tree and recombination detection*

To identify the relationship between the sequences of the nine strains and the Muguga reference strain, an unrooted neighbour-net tree[31] was constructed based on the concatenated SNP dataset III using Split tree version 4.11.3.[32] The Recombination Detection Program version 3.44 (RDP3) was used to detect possible recombination regions.[33] This software incorporates several recombination prediction methods. As the reliability of each method has not been fully evaluated, it is anticipated that some of the recombination events predicted may be artifactual. We manually curated the results choosing Geneconv[34] and maximum Chi-square[35] as the selection priority, as the accuracy of these tests is relatively well defined.[36] Predicted recombination events were considered valid, if at least one additional program supported the findings, i.e. ($P \leq 0.001$) for that event from RDP,[37] Boot scanning,[38] 3 Seq method,[39] or the sister-scanning method.[40] Predictions that did not meet these criteria were removed. For phylogenetic analysis of p150 and p104, we used mapping sequence information for each strain, and unmapped or unreliable regions were filled by manual Sanger sequencing. The sequences obtained in this study were submitted to GenBank under accession no. AB739676–AB739693.

3. Results

3.1. *Genome sequencing of nine* T. parva *strains using Illumina technology*

Single runs of Illumina produced over 10 million reads for each sample, and this provided coverage of 94.7–97.5% for genomes of individual strains against 8.3 M of the reference Muguga genome, with an average coverage between ×17 and ×49 (Table 1). Depending on the purity of the preparations, 11.6–77.9% of the total reads for any one strain were successfully mapped, whereas unmapped reads were considered to be derived from host genomic DNA. All four chromosomes of each stock were evenly covered in general, except for ChitongoZ2 (Fig. 1). As the concentration of extracted DNA from purified schizonts in ChitongoZ2 strain was lowest, we suspect that the whole-genome amplification procedure for this strain caused biased amplification, resulting in an uneven distribution of the coverage; however, this did not affect SNPs detection.

3.2. *SNPs detection*

Stringent conditions for SNPs detection were used, i.e. more than five high-quality reads covering the SNPs and 100% concordance in position. If multiple allele variants calling was allowed, 5216 loci had complex SNPs in at least one strain (0.0633% of the reference Muguga genome). As the genome of *Theileria* at the schizont stage is haploid, only a single allele is expected at each locus, and complex

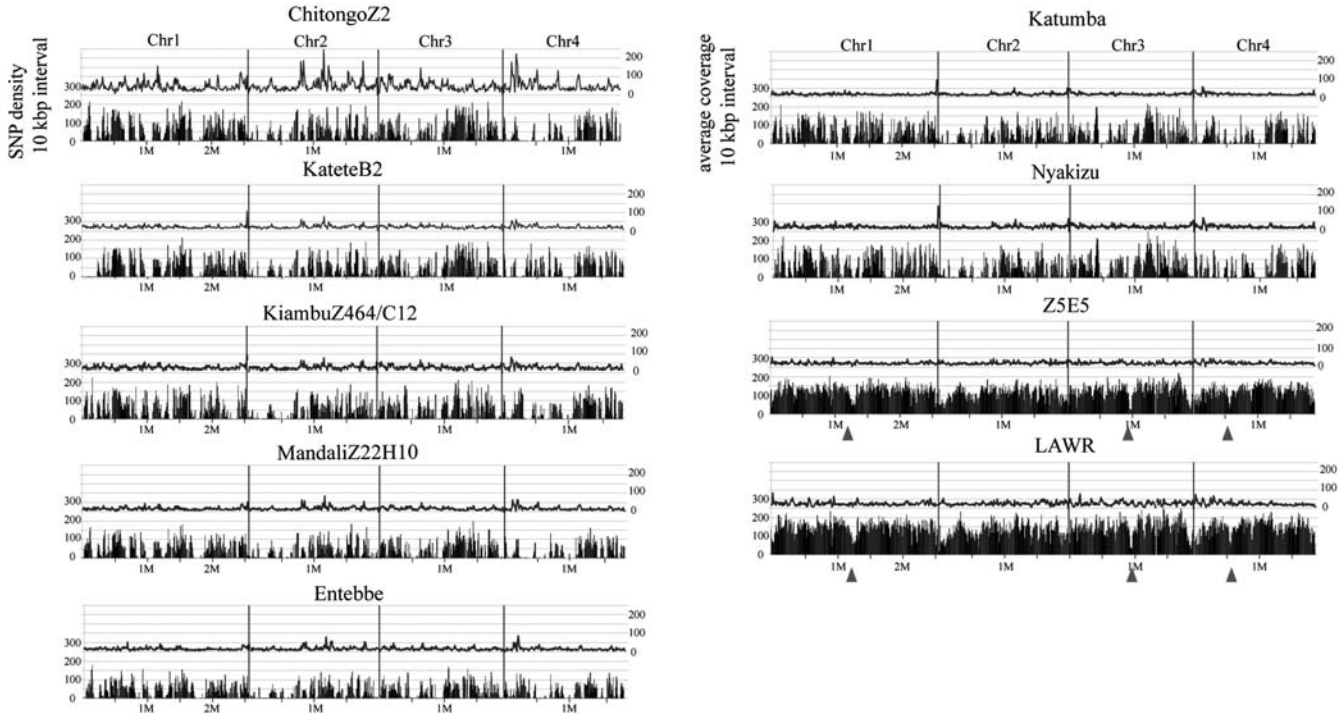


Figure 1. SNPs distribution across the *Theileria* genome. SNPs in individual strains were detected after mapping to the reference genome Muguga strain. The entire datasets of 34 814–121 545 SNPs (SNP dataset I) were plotted as SNP densities (per 10 kb intervals) alongside chromosome 1–4. The *x*-axis shows the chromosomal position, and the left *y*-axis shows the number of SNPs (black bars) per 10 kb interval. Average short read coverage is also shown on the right *y*-axis (above line). Arrowheads indicate the possible location of the centromere.

SNPs are unexpected, if the sample contains a clonal population. The appearance of these multi-allelic SNPs could represent base-calling or mapping errors (due to repetitive sequence or paralogous genes). Because other possibilities that these SNPs were generated during *in vitro* passages after cloning by the limited dilution and that minor populations in the original materials obtained from host animals remained in the analysed samples cannot be excluded, such questionable SNPs were excluded in further analysis. Although it is likely that some genuine SNPs may be overlooked, a high stringency SNPs calling protocol was utilized to avoid false SNPs calls.

The number of SNPs identified in bovine-derived strains when compared with the Muguga strain ranged from 34 814 in the Entebbe strain to 51 790 in the Nyakizu strain. Additionally, 121 545 and 103 880 SNPs were identified in buffalo-derived LAWR and Z5E5 strains, respectively (Table 1). The densities of the SNPs in each chromosome tended to be higher in chromosomes 1 and 3 than in chromosomes 2 and 4 in most of the strains (Fig. 2). Out of a total of 533 642 SNPs identified in 9 strains (Table 1), 364 719 were present in coding regions (cSNP) and 168 923 were present in non-coding regions (ncSNP), although the SNP density (calculated per 1 kp) of cSNPs and ncSNPs were similar (Table 1). The numbers of SNPs ranged from 34 814 (Entebbe) to 121 545 for the buffalo-derived LAWR strain, and more than 2-fold SNPs were identified in 2 of the buffalo-derived strains when compared with the cattle-derived strains (Table 1), suggesting a degree

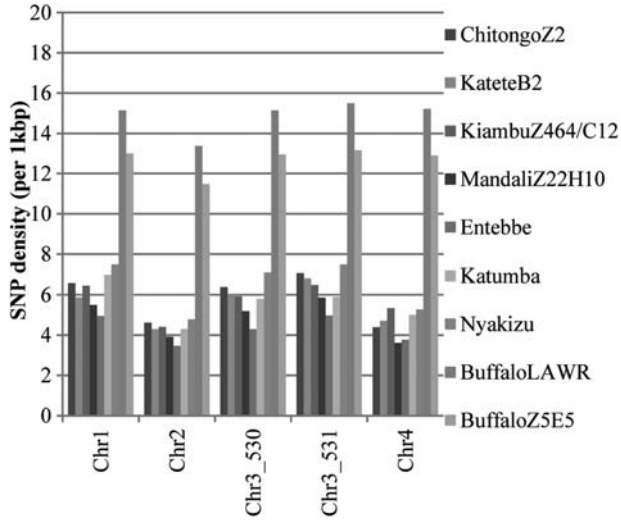


Figure 2. SNP density in each chromosome (SNP dataset I). Average SNP densities per 1 kb interval were calculated for each chromosome in nine *T. parva* strains with reference to the Muguga genome strain. In the published full genome sequence of *T. parva*, there is a large gap in the assembly of chromosome 3, due to the repetitive Tpr locus. The large contig AAGK01000005 and smaller contig AAGK 01000006 are shown as Chr3_530 and Chr3_531, respectively.

of genetic differentiation between these types of *Theileria*. As shown in Fig. 1, clustered distribution of SNPs was observed (black bars in each panel). The uneven distribution of SNPs was not found to correlate with the sequence coverage distribution (line); thus, the effect of low SNPs calling efficiency in particular regions can be excluded. In addition, lower SNPs densities were observed within defined regions on chromosomes 1, 3, and 4, which was most evident in buffalo-derived *Theileria* strains (Fig. 1, arrowhead). These regions correspond to the putative centromeres with an extremely AT-rich composition.

3.3. *dN/dS analysis*

The ratio of the number of non-synonymous substitutions per non-synonymous site (dN) to synonymous substitutions per synonymous site (dS) both at the inter- and intra-species level has been used to estimate the potential selective pressure acting on the genes.[41] A dN/dS ratio lower than one suggests negative or purifying selection, whereas a ratio higher than one suggests positive selection or diversification. Estimation of dN/dS ratios can potentially identify genes encoding immunogenic proteins and, thus, putative vaccine candidates.[42] Therefore, we calculated dN/dS ratios for individual genes using SNP dataset I for seven bovine *Theileria* strains with the yn00 program of the PAML package.[29] Overall, the dN/dS ratios calculated between cattle *T. parva* strains were average values of 0.0894–0.0993 when pair-wise comparisons were performed against the Muguga strain, with similar values to those observed in the comparison between *T. parva* versus *Theileria annulata* (average dN/dS = 0.097).[43] Among a total of 4011 genes annotated on the Muguga genome, 263 genes showed elevated levels of dN/dS values (average + 3SD) in at least 1 strain (Supplementary Table S1). We further narrowed the list down to 71 genes by selecting only those genes that have a signal sequence for targeting to the endoplasmic reticulum. Those selected genes may be potential targets of the host's immune system. The final list of these possible antigenic, and therefore vaccine target, genes is shown in Table 2, and the orthologous groups were also assigned according to our previous study.[44] Most of the other genes listed here are currently annotated as hypothetical proteins without any predicted functional domain. However, some of them are known to be recognized by host humoral immunity. For example, p32 (TP01_1056)[45] and 23 kDa piroplasm surface protein (TP02_0551)[46] are erythrocytic piroplasm stage antigens, and strong antibody response in infected cattle has been reported.

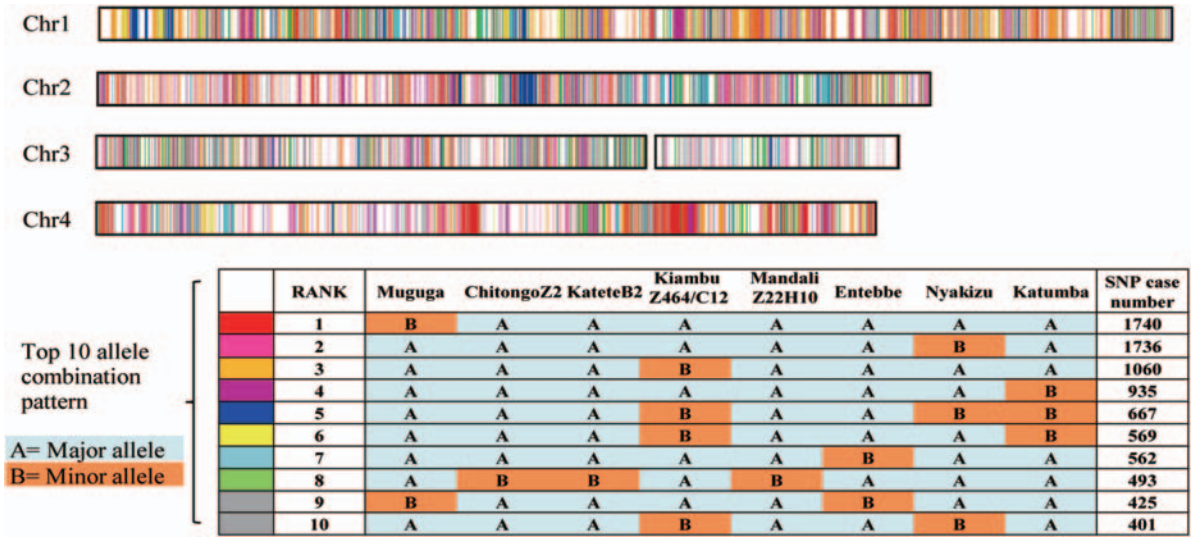
3.4. *Phylogenetic relationship among 10* T. parva *strains and evidence of recombination*

The allele frequency or combination of the bovine *Theileria* strain alleles collected in SNP dataset II was determined. By scoring biallelic positions only, 127 allelic combinations were identified among 8 bovine *Theileria*. Each of the 15 901 SNPs was assigned 1 of the 127 combinations. When the rank order of these combinations was calculated, the allele pattern unique to the Muguga strain came first, followed by Nyakizu-, KiambuZ464/C12-, and Katumba-unique allele combinations (Fig. 3). Because Muguga strains were used as the reference sequence, ranking 'Muguga strain-unique allele pattern' as the first event seems reasonable, as it incorporates a minor allele that is present in the Muguga strain. The distribution of frequencies among the 127 events was uneven because 54% of all SNPs were assigned to these top 10 allelic combinations. When the list was extended to cover the top 20 or 25 combinations, this ratio increased to 73 and 80%, respectively, indicating that most of the SNPs alleles were represented by a limited number of combinations. The distribution of these different SNPs patterns is represented on a schematic diagram of the chromosomes, and different combination events are colour coded (Fig. 3). As shown in Fig. 3, allelic combinations among the strains are distributed throughout every chromosome. A major observation was that SNPs with particular allelic combinations tend to cluster into defined loci, giving rise to a rough, large-scale mosaic pattern of allelic combinations. If the evolution of these strains had taken place completely independently, i.e. without interaction between strains, this clustering of allelic combinations would not be expected.

The relationships among the 10 *T. parva* strains were analysed by creating a phylogenetic tree (Fig. 4). The allelic combinations are well correlated with the phylogenetic relationship among these strains, and the top 10 allelic combination events represented major nodes in the tree. Neighbour net is a phylogenetic network construction method that combines aspects of the neighbour joining and Split tree. In this neighbour-net analysis, the appearance of the reticulated branches indicates the recombination events. Considered together with the mosaic allelic combination patterns as described above (Fig. 3), we speculate that recombination events are responsible for the interrelationships between strains. To verify this hypothesis, we carried out further recombination event estimations with the RDP programs. The concatenated SNP dataset II was subjected to six recombination detection tests, namely Geneconv, maximum Chi-square, RDP, Boot scanning, 3 Seq., and sister-scanning methods. This

Table 2. List of genes with high dN/dS ratios and a secretion signal peptide 71 genes were listed from 263 genes (higher dN/dS ratios), by selecting secretion signal peptide-predicted genes

GeneID	Description	Ortholog group	Signal	GeneID	Description	Ortholog group	Signal
TP01_0144	Hypothetical protein	PiroF0002444	Y	TP03_0003	Hypothetical (SVSP)	PiroF0100037	Y
TP01_0178	Hypothetical protein	PiroF0002919	Y	TP03_0039	Hypothetical protein	Not assigned	Y
TP01_0180	40S ribosomal protein S11	PiroF0000589	Y	TP03_0040	Hypothetical protein	PiroF0003613	Y
TP01_0291	Hypothetical protein	PiroF0002390	Y	TP03_0123	Hypothetical protein	PiroF0002851	Y
TP01_0367	Hypothetical protein	PiroF0000012	Y	TP03_0217	Hypothetical protein	PiroF0000012	Y
TP01_0378	Hypothetical protein	PiroF0003402	Y	TP03_0297	Hypothetical (FAINT superfamily)	PiroF0100056	Y
TP01_0380	Hypothetical protein	PiroF0003404	Y	TP03_0298	Hypothetical (FAINT superfamily)	PiroF0000056	Y
TP01_0610	Hypothetical (Tash family)	PiroF0100038	Y	TP03_0319	Hypothetical protein	PiroF0000012	Y
TP01_0619	Hypothetical (Tash family)	PiroF0100038	Y	TP03_0368	Hypothetical (FAINT superfamily)	PiroF0100056	Y
TP01_0621	Hypothetical (Tash family)	PiroF0100038	Y	TP03_0405	Hypothetical protein	PiroF0002425	Y
TP01_0914	Hypothetical protein	PiroF0002316	Y	TP03_0498	Hypothetical (SVSP)	PiroF0100037	Y
TP01_0955	Hypothetical protein	PiroF0003569	Y	TP03_0520	Hypothetical protein	PiroF0000012	Y
TP01_0987	Hypothetical protein	PiroF0002967	Y	TP03_0530	Hypothetical protein		Y
TP01_1011	Hypothetical protein	PiroF0100045	Y	TP03_0664	Hypothetical protein	PiroF0000012	Y
TP01_1044	Hypothetical protein	Not assigned	Y	TP03_0780	Hypothetical protein	PiroF0002660	Y
TP01_1056	32 kDa surface antigen	PiroF0002963	Y	TP03_0810	Hypothetical protein	PiroF0002675	Y
TP01_1109	Hypothetical protein	PiroF0000207	Y	TP03_0886	Hypothetical (SVSP)	PiroF0100037	Y
TP01_1227	Hypothetical (SVSP)	PiroF0100037	Y	TP03_0893	Hypothetical (SVSP)	PiroF0100037	Y
TP02_0004	Hypothetical (SVSP)	PiroF0100037	Y	TP04_0009	Hypothetical (SVSP)	PiroF0100037	Y
TP02_0006	Hypothetical (SVSP)	PiroF0100037	Y	TP04_0012	Hypothetical (FAINT superfamily)	PiroF0100056	Y
TP02_0010	Hypothetical (SVSP)	PiroF0100037	Y	TP04_0013	Hypothetical (SVSP)	PiroF0100037	Y
TP02_0018	Hypothetical protein	PiroF0100055	Y	TP04_0096	Hypothetical (FAINT superfamily)	PiroF0100056	Y
TP02_0239	Hypothetical protein	PiroF0002609	Y	TP04_0097	Hypothetical (FAINT superfamily)	PiroF0100056	Y
TP02_0327	Hypothetical protein	PiroF0000012	Y	TP04_0101	Hypothetical (FAINT superfamily)	PiroF0100056	Y
TP02_0331	Ubiquitin-activating enzyme, putative	PiroF0002575	Y	TP04_0104	Hypothetical (FAINT superfamily)	PiroF0100056	Y
TP02_0551	23 kDa piroplasm surface protein	PiroF0003021	Y	TP04_0110	Hypothetical protein	PiroF0001224	Y
TP02_0575	Hypothetical protein	PiroF0003017	Y	TP04_0116	Hypothetical protein	PiroF0003546	Y
TP02_0819	Hypothetical (FAINT superfamily)	PiroF0100056	Y	TP04_0150	Hypothetical (SVSP)	PiroF0000037	Y
TP02_0856	Hypothetical (FAINT superfamily)	PiroF0100056	Y	TP04_0328	Hypothetical protein	PiroF0002219	Y
TP02_0875	Hypothetical protein	PiroF0002985	Y	TP04_0411	Hypothetical protein	PiroF0003185	Y
TP02_0952	Hypothetical protein	PiroF0003456	Y	TP04_0437	104 kDa antigen	PiroF0003088	Y
TP02_0954	Hypothetical (SVSP)	PiroF0100037	Y	TP04_0558	Hypothetical protein	PiroF0001517	Y
TP02_0956	Hypothetical (SVSP)	PiroF0100037	Y	TP04_0919	Hypothetical (SVSP)	PiroF0100037	Y
TP03_0001	Hypothetical (SVSP)	PiroF0100037	Y	TP04_0920	Hypothetical (SVSP)	PiroF0100037	Y
TP03_0002	Hypothetical protein (SVSP)	PiroF0100037	Y	TP04_0921	Hypothetical protein (FAINT superfamily)	PiroF0000056	Y



RANK	Muguga	ChitongoZ2	KateteB2	Kiambu Z464/C12	Mandali Z22H10	Entebbe	Nyakizu	Katumba	SNP case number
1	B	A	A	A	A	A	A	A	1740
2	A	A	A	A	A	A	B	A	1736
3	A	A	A	B	A	A	A	A	1060
4	A	A	A	A	A	A	A	B	935
5	A	A	A	B	A	A	B	B	667
6	A	A	A	B	A	A	A	B	569
7	A	A	A	A	A	B	A	A	562
8	A	B	B	A	B	A	A	A	493
9	B	A	A	A	A	B	A	A	425
10	A	A	A	B	A	A	B	A	401

Figure 3. Mosaic pattern of SNPs in *T. parva* strains. The frequency of each of the 127 possible allelic combinations for the 8 cattle-derived *T. parva* strains was calculated using the SNP dataset II. The 10 top-ranking combinations were plotted onto schematic chromosomes in the assigned colours. Each line within a chromosome represents a single SNP marker position.

resulted in a minimum of 133 recombination events being predicted as shown in Supplementary Fig. S2. A snapshot of the alignment of a concatenated version of the SNP dataset II is also shown in Supplementary Fig. S3. An RDP analysis was also carried out using the SNP dataset III, but no significant evidence for recombination was detected between cattle- and buffalo-derived strains (Z5E5 and LAWR, data not shown).

As polymorphic antigens such as p104 or p150 have been used for the genotyping of *T. parva,*[6] we compared results of genotyping based on p104 or p150 with those obtained by SNPs analysis. As shown in Supplementary Fig. S4, there was no congruency in tree shapes. The most likely explanation for this inconsistency is that the recombination event between the ancestral strains involved these loci, as is evident in Supplementary Fig. S2. RDP3 program predicts recombination events within those two loci. In p104 loci, KateteB2 and Katumba are predicted to be recombinant from unknown parent or Entebbe strains. And this is true for Muguga, KiambuZ464/C12, and the possible donor, Nyakizu, at the p150 locus as marked in Supplementary Fig. S2.

4. Discussion

Comparison of whole-genome sequencing data of several *Theileria* strains, using short reads sequencing and mapping on the reference genome sequence, revealed genome-wide nucleotide-based polymorphisms in this species. SNPs density plots evaluate clustered SNPs distribution across the genome and identify SNP-poor and SNP-rich regions. Such a clustering of SNPs has been also reported in mammalian genome, although the forces responsible (e.g. mutation hot spot, recombination, or balancing selection) remain poorly understood.[47,48] For apicomplexan parasite, reports for the genome-wide SNPs analysis are limited, but similar SNPs distribution pattern was observed in *Plasmodium*[49] suggesting existence of the same underlying mechanisms between parasite and mammalian genomes for these SNPs clustering.

Our SNPs analysis clarified the phylogenetic relationships among 10 *Theileria* strains on a genome-wide scale. When these *Theileria* strains were further analysed using neighbour-net analysis, clusters were formed in accordance with both host species and geographical origin. For example, three Zambian strains (ChitongoZ2, MandaliZ22H10, and KateteB2) were clustered together in the same node, inferring that they are closely related genotypes, but distant from strains isolated in Eastern Africa. In addition, there is a clear demarcation between the bovine- and buffalo-derived strains (Z5E5 and LAWR). Genetic difference between Z5E5 and LAWR was also confirmed as high numbers of SNPs were not shared between Z5E5 and LAWR, as is shown in Supplementary Fig. S5. However, reticulated patterns between strains belonging to different clusters are evident, as shown in Fig. 4, which suggests genetic recombination between ancestors of the strains that are currently geographically separated. The evidence for recombination among the analysed strains was further supported by the presence of a mosaic pattern of allelic combinations, together with the statistical analysis of recombination. This result is intuitive when one considers the fact that the parasite has an obligate

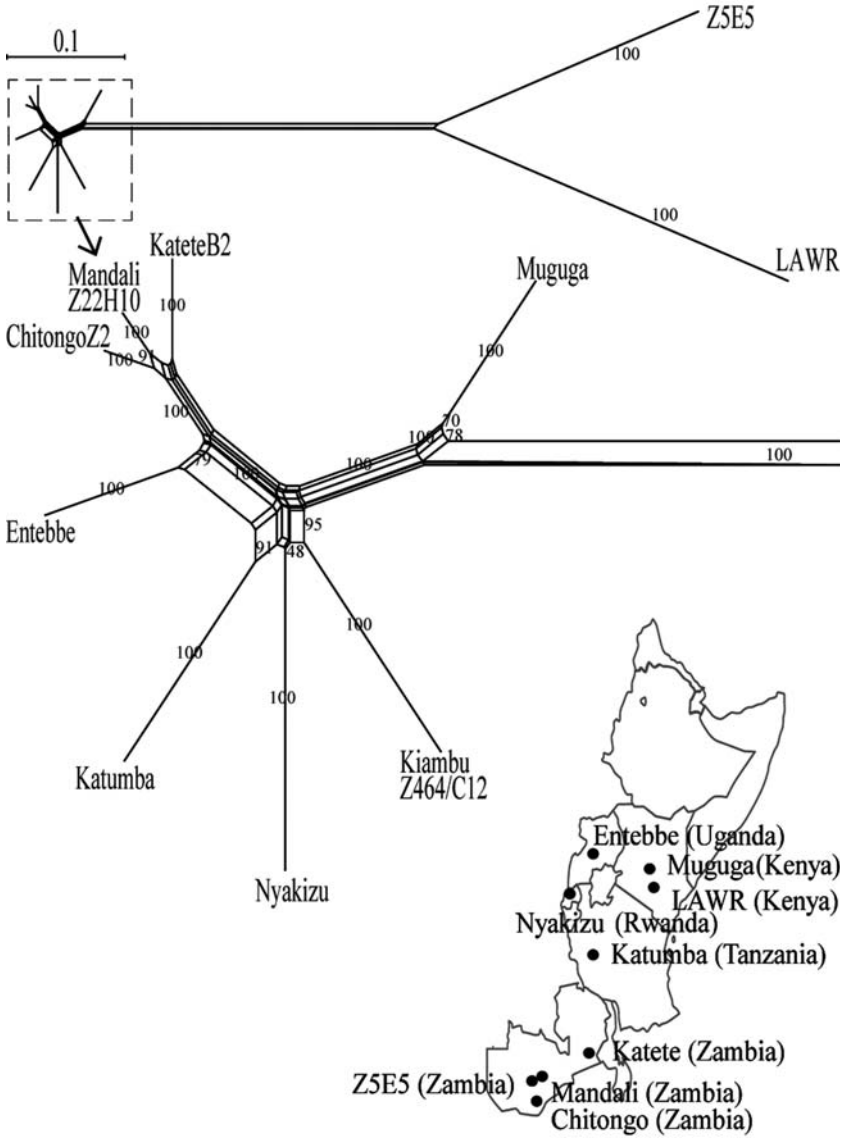


Figure 4. Neighbour-net network analysis of 10 *T. parva* strains. Neighbour-net network analysis was performed with the concatenated SNP allele sequence data from SNP dataset III. Bootstrap values are based on 100 replicates and were near 100 at most of the nodes.

sexual cycle and that analysis of field populations suggests that recombination in the tick vector is commonplace.[14] There are two possibilities of ticks being infected with parasites with different genotypes: infestation on a single bovine host infected with genotypically mixed parasite populations or multiple infestations on different animals infected with different parasite genotypes that are possible for two or three host tick species. However, the latter is less likely, as synchronization of the sexual stages (micro- and macro gamete) between two parasites is difficult, if they are picked up by ticks at different feeding times.

We hypothesize that genetic recombination occurred in the ancestral bovine *Theileria* populations in the distant past, and parasites had evolved independently after geographical isolation. The origin of *T. p. parva* in cattle is unknown, but it is considered likely to have originated in the African buffalo.[50] Evidently, *T. parva* populations in buffalo are considered to be more diverse than in cattle[6,13,51] and cause severe disease in cattle. Historically, domestic cattle were believed to have been introduced to the African continent thousands of years ago, possibly into Sub-Saharan Africa from the Mideast.[52–54] After the introduction of cattle, a subset of the buffalo *Theileria* population may have been transmitted (at that stage, it would be called *T. p. lawrencei*, as it could not infect other cattle), adapted, and co-evolved within cattle, resulting in the emergence of *T. p. parva* that can spread within cattle.

It should be emphasized that the phylogenetic tree obtained from two polymorphic antigens (p104 and

p150) showed a different topology from that based on genome-wide SNPs. Thus, the interpretation of the phylogenetic relationship, analysed by a limited number of loci, must be made carefully in the case of pathogens that acquire genetic diversity by recombination, rather than by accumulation of mutations. This is due to the fact that each locus can become chimaeric by crossing over between genotypes that have different evolutional histories. Therefore, a number of independent loci should be included to estimate the real relationship between isolates such as multilocus sequencing typing, but genome-wide SNPs analysis is the ultimate solution in this context.

Two buffalo-derived strains (Z5E5 and LAWR) were genetically distant from cattle-derived *Theileria* strains and between the two strains, as expected from earlier studies.[6,13,51] It has been proposed that genetic exchange between buffalo-derived *Theileria* and cattle-derived *Theileria* still occurs through sexual recombination, based on evidence that *T. p. lawrencei* and *T. p. parva* showed a mosaic sequence pattern in the ITS region.[55] However, in our recombination analysis using the RDP program, no recombination events were detected between bovine and buffalo *Theileria* strains. It might be hypothesized that cattle-infecting strains were originated from a subset of buffalo-infecting *T. parva* population that has been circulating in Africa for a long time and now have evolved a genetic barrier to recombination. Further analysis with a greater number of buffalo-derived samples and denser SNPs coverage would be needed to clarify the genetic relationship between buffalo and cattle *Theileria* more precisely.

Estimation of dN/dS values can potentially identify immunogenic genes under possible selective pressure and, thus, possible vaccine candidate molecules. The selected candidate 71 antigen list (Table 2) covers most of the known genes for antigenic or host-interacting proteins, which confirms the effectiveness of this genome-wide approach. p23[46] and p32[45] are surface or secreted antigens recognized by humoral immunity in infected animals. The diversification of these genes may be related to immune evasion of the *Theileria* parasite.

On the other hand, although several genes with CTL targets have been identified as being under possible immune pressure,[56] only *Tp1* (TP03_0849) showed a higher dN/dS value in this study, whereas other genes for CTL targets (*Tp2-9*) showed relatively low dN/dS values (Supplementary Table S2). Relative conservation of the sequences of these CTL antigen genes among the different parasite strains has already been reported.[56,57] Considering that the CTL response is a function of the host MHC type/TCR repertoire and antigenic types of parasites, the positive selective pressure acting on a particular gene may be too weak to be detected. In addition, CTL recognizes short peptides presented by MHC class I molecules, and, therefore, immune-based selective pressure is likely to be focused on a limited region of the targeted genes that dN/dS analysis is not sensitive enough to detect.

The selected 71 antigen list also contained several genes from 3 large gene families, namely the *Tash* gene family (Ortholog group number: PiroF0100038), the *SVSP* gene family (PiroF0100037), and FAINT super family (PiroF0100056, also called as SfiI-subtelomeric fragment related protein family member). The *Tash* gene family has been characterized extensively in *T. annulata*.[58] Some of the *Tash* and *SVSP* gene products have been predicted or demonstrated to be translocated in the host nucleus, and most of the *Tash* and *SVSP* genes are expressed predominantly in the schizont stage.[58,59] This entails that the potential selective pressure will not be humoral, although the possibility that these proteins are exposed to the humoral immune response when infected cells are lysed cannot be excluded. A previous comparison between *T. annulata* and *T. parva* genomes also revealed high inter-species dN/dS ratios for the *Tash* and the *SVSP* family,[60] consistent with our analysis. It was argued that gene expansion and divergence of *Tash* and *SVSP* family genes were associated with different functionality in each species.

In conclusion, this study highlighted the phylogenetic relationship of 10 *T. parva* strains based on full genome sequences with prediction of possible past recombination events. The high-density SNPs map developed in this study is now applicable for genotyping or linkage analysis of the parasite. Practically, SNPs-based genotyping can discriminate vaccine and field strains of *T. parva*. Recent methodological advances in high-throughput technologies such as Taq man-real-time PCR and Golden gate technologies[61] for SNPs genotyping will likely facilitate future genotyping studies. Further phylogenetic analysis in combination with phenotypic data will assist in the investigation of the virulence and evolution of bovine theilerias after their diversification from buffalo. Importantly, the putative antigen-encoding genes listed in this study should be further investigated to assess their candidacy as *Theileria* subunit vaccine components.

Supplementary data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported in part by the Grants-in-Aid for Scientific Research and Asia-Africa S & T Strategic Cooperation Promotion Program by the Special Coordination Funds for Promoting Science & Technology, from the Ministry of Education, Culture,

Sports, Science and Technology of Japan (MEXT) to C.S. K.H. was supported by the Program of Founding Research Centers for Emerging and Reemerging Infectious Diseases, MEXT.

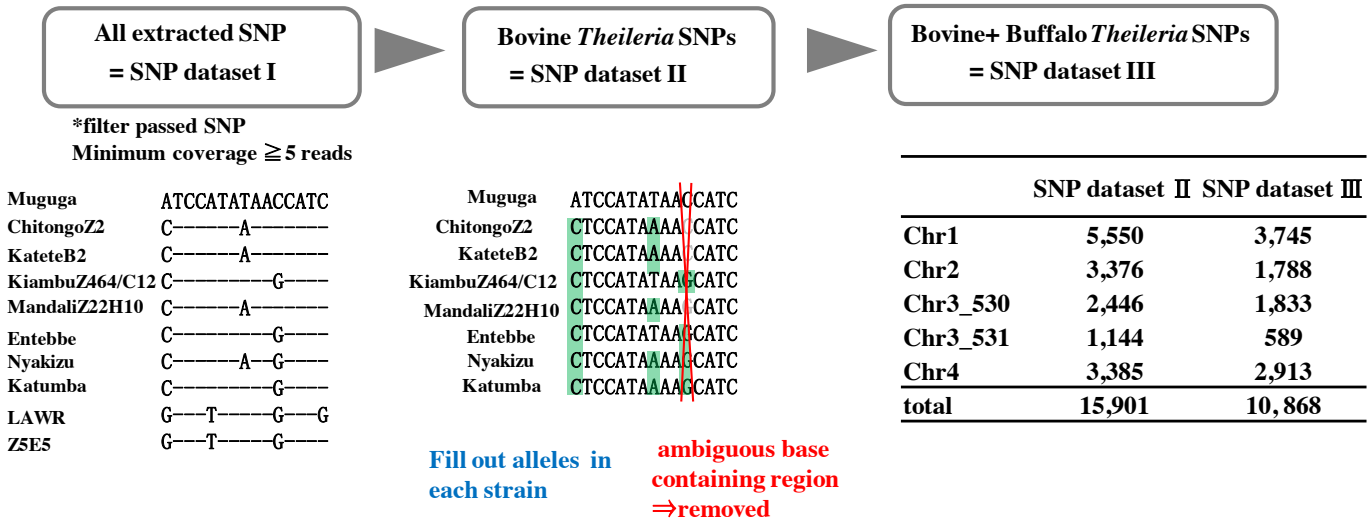
References

1. Brown, C.G., Stagg, D.A., Purnell, R.E., Kanhai, G.K. and Payne, R.C. 1973, Letter: infection and transformation of bovine lymphoid cells in vitro by infective particles of *Theileria parva*, *Nature*, **245**, 101–3.
2. Lawrence, J.A. and Irvin, A.D. 1994, Theilerioses, In: Coetzer, J.A.W., Thomson, G.R. and Tustin, R.C. (eds.), *Infectious Diseases of Livestock*, Oxford University Press: New York, pp. 307–41.
3. Uilenberg, G. 1981, Theileria species of domestic livestock. In: Irvun, A.D., Cunningham, M.P. and Young, A.S. (eds.), *Advances in the Control of Theileriosis*, Martinus Nijhoff Publishers: The Hague, pp. 4–37.
4. Young, A.S. and Purnell, R.E. 1973, Transmission of *Theileria lawrencei* (Serengeti) by the ixodid tick, *Rhipicephalus appendiculatus*, *Trop. Anim. Health Prod.*, **5**, 146–52.
5. Brown, C.G., Radley, D.E., Burridge, M.J. and Cunningham, M.P. 1977, The use of tetracyclines on the chemotherapy of experimental East Coast Fever (*theileria parva* infection of cattle), *Tropenmed. Parasitol.*, **28**, 513–20.
6. Geysen, D., Bishop, R., Skilton, R., Dolan, T.T. and Morzaria, S. 1999, Molecular epidemiology of *Theileria parva* in the field, *Trop. Med. Int. Health*, **4**, A21–7.
7. Oura, C.A., Bishop, R., Wampande, E.M., Lubega, G.W. and Tait, A. 2004, The persistence of component *Theileria parva* stocks in cattle immunized with the 'Muguga cocktail' live vaccine against East Coast fever in Uganda, *Parasitology*, **129**, 27–42.
8. Oura, C.A., Bishop, R., Asiimwe, B.B., Spooner, P., Lubega, G.W. and Tait, A. 2007, *Theileria parva* live vaccination: parasite transmission, persistence and heterologous challenge in the field, *Parasitology*, **134**, 1205–13.
9. McKeever, D.J. 2007, Live immunisation against *Theileria parva*: containing or spreading the disease? *Trends Parasitol.*, **23**, 565–8.
10. Berkvens, D.L. 1991, Re-assessment of tick control after immunization against East Coast fever in the Eastern Province of Zambia, *Ann. Soc. Belg. Med. Trop.*, **71**, 87–94.
11. Patel, E.H., Lubembe, D.M., Gachanja, J., Mwaura, S., Spooner, P. and Toye, P. 2011, Molecular characterization of live *Theileria parva* sporozoite vaccine stabilates reveals extensive genotypic diversity, *Vet. Parasitol.*, **179**, 62–8.
12. Bishop, R., Geysen, D., Spooner, P., et al. 2001, Molecular and immunological characterisation of *Theileria parva* stocks which are components of the 'Muguga cocktail' used for vaccination against East Coast fever in cattle, *Vet. Parasitol.*, **94**, 227–37.
13. Minami, T., Spooner, P.R., Irvin, A.D., Ocama, J.G., Dobbelaere, D.A. and Fujinaga, T. 1983, Characterisation of stocks of *Theileria parva* by monoclonal antibody profiles, *Res. Vet. Sci.*, **35**, 334–40.
14. Oura, C.A., Odongo, D.O., Lubega, G.W., Spooner, P.R., Tait, A. and Bishop, R.P. 2003, A panel of microsatellite and minisatellite markers for the characterisation of field isolates of *Theileria parva*, *Int. J. Parasitol.*, **33**, 1641–53.
15. Oura, C.A., Asiimwe, B.B., Weir, W., Lubega, G.W. and Tait, A. 2005, Population genetic analysis and substructuring of *Theileria parva* in Uganda, *Mol. Biochem. Parasitol.*, **140**, 229–39.
16. Katzer, F., Ngugi, D., Oura, C., et al. 2006, Extensive genotypic diversity in a recombining population of the apicomplexan parasite *Theileria parva*, *Infect. Immun.*, **74**, 5456–64.
17. Katzer, F., Ngugi, D., Walker, A.R. and McKeever, D.J. 2010, Genotypic diversity, a survival strategy for the apicomplexan parasite *Theileria parva*, *Vet. Parasitol.*, **167**, 236–43.
18. Katzer, F., Lizundia, R., Ngugi, D., Blake, D. and McKeever, D. 2011, Construction of a genetic map for *Theileria parva*: identification of hotspots of recombination, *Int. J. Parasitol.*, **41**, 669–75.
19. Radley, D.E., Brown, C.G., Cunnigham, M.P., et al. 1975, East Coast fever: challenge if immunised cattle by prolonged exposure to infected ticks, *Vet. Rec.*, **96**, 525–7.
20. Sugimoto, C., Conrad, P.A., Ito, S., Brown, W.C. and Grab, D.J. 1988, Isolation of *Theileria parva* schizonts from infected lymphoblastoid cells, *Acta Trop.*, **45**, 203–16.
21. Goddeeris, B.M., Dunlap, S., Innes, E.A. and McKeever, D.J. 1991, A simple and efficient method for purifying and quantifying schizonts from *Theileria parva*-infected cells, *Parasitol. Res.*, **77**, 482–4.
22. Baumgartner, M., Tardieux, I., Ohayon, H., Gounon, P. and Langsley, G. 1999, The use of nocodazole in cell cycle analysis and parasite purification from *Theileria parva*-infected B cells, *Microbes Infect.*, **1**, 1181–8.
23. Asao, T., Kinoshita, Y., Kozaki, S., Uemura, T. and Sakaguchi, G. 1984, Purification and some properties of *Aeromonas hydrophila* hemolysin, *Infect. Immun.*, **46**, 122–7.
24. Hosono, S., Faruqi, A.F., Dean, F.B., et al. 2003, Unbiased whole-genome amplification directly from clinical samples, *Genome Res.*, **13**, 954–64.
25. Silander, K. and Saarela, J. 2008, Whole genome amplification with Phi29 DNA polymerase to enable genetic or genomic analysis of samples of low DNA yield, *Methods Mol. Biol.*, **439**, 1–18.
26. Gardner, M.J., Bishop, R., Shah, T., et al. 2005, Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes, *Science*, **309**, 134–7.
27. Brockman, W., Alvarez, P., Young, S., et al. 2008, Quality scores and SNP detection in sequencing-by-synthesis systems, *Genome Res.*, **18**, 763–70.
28. Yang, Z., Nielsen, R. and Goldman, N. 2009, In defense of statistical methods for detecting positive selection, *Proc. Natl. Acad. Sci. USA*, **106**, E95.
29. Yang, Z. 2007, PAML 4: phylogenetic analysis by maximum likelihood, *Mol. Biol. Evol.*, **24**, 1586–91.

30. Petersen, T.N., Brunak, S., von Heijne, G. and Nielsen, H. 2011, SignalP 4.0: discriminating signal peptides from transmembrane regions, *Nat. Methods*, **8**, 785–6.
31. Bryant, D. and Moulton, V. 2004, Neighbor-net: an agglomerative method for the construction of phylogenetic networks, *Mol. Biol. Evol.*, **21**, 255–65.
32. Huson, D.H. and Bryant, D. 2006, Application of phylogenetic networks in evolutionary studies, *Mol. Biol. Evol.*, **23**, 254–67.
33. Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D. and Lefeuvre, P. 2010, RDP3: a flexible and fast computer program for analyzing recombination, *Bioinformatics*, **26**, 2462–3.
34. Padidam, M., Sawyer, S. and Fauquet, C.M. 1999, Possible emergence of new geminiviruses by frequent recombination, *Virology*, **265**, 218–25.
35. Smith, J.M. 1992, Analyzing the mosaic structure of genes, *J. Mol. Evol.*, **34**, 126–9.
36. Posada, D. 2002, Evaluation of methods for detecting recombination from DNA sequences: empirical data, *Mol. Biol. Evol.*, **19**, 708–17.
37. Martin, D. and Rybicki, E. 2000, RDP: detection of recombination amongst aligned sequences, *Bioinformatics*, **16**, 562–3.
38. Martin, D.P., Posada, D., Crandall, K.A. and Williamson, C. 2005, A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints, *AIDS Res. Hum. Retroviruses*, **21**, 98–102.
39. Boni, M.F., Posada, D. and Feldman, M.W. 2007, An exact nonparametric method for inferring mosaic structure in sequence triplets, *Genetics*, **176**, 1035–47.
40. Gibbs, M.J., Armstrong, J.S. and Gibbs, A.J. 2000, Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences, *Bioinformatics*, **16**, 573–82.
41. Kimura, M. 1991, Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics, *Proc. Natl. Acad. Sci. USA*, **88**, 5969–73.
42. Endo, T., Ikeo, K. and Gojobori, T. 1996, Large-scale search for genes on which positive selection may operate, *Mol. Biol. Evol.*, **13**, 685–90.
43. Pain, A., Renauld, H., Berriman, M., et al. 2005, Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*, *Science*, **309**, 131–3.
44. Hayashida, K., Hara, Y., Abe, T., et al. 2012, Comparative genome analysis of three eukaryotic parasites with differing abilities to transform leukocytes reveals key mediators of Theileria-induced leukocyte transformation, *MBio.*, **3**, e00204–12.
45. Skilton, R.A., Musoke, A.J., Wells, C.W., et al. 2000, A 32 kDa surface antigen of *Theileria parva*: characterization and immunization studies, *Parasitology*, **120**, 553–64.
46. Sako, Y., Asada, M., Kubota, S., Sugimoto, C. and Onuma, M. 1999, Molecular cloning and characterisation of 23-kDa piroplasm surface proteins of *Theileria sergenti* and *Theileria buffeli*, *Int. J. Parasitol.*, **29**, 593–9.
47. Hellmann, I., Prufer, K., Ji, H., Zody, M.C., Paabo, S. and Ptak, S.E. 2005, Why do human diversity levels vary at a megabase scale? *Genome Res.*, **15**, 1222–31.
48. Michelizzi, V.N., Wu, X., Dodson, M.V., et al. 2010, A global view of 54,001 single nucleotide polymorphisms (SNPs) on the Illumina BovineSNP50 BeadChip and their transferability to water buffalo, *Int. J. Biol. Sci.*, **7**, 18–27.
49. Chan, E.R., Menard, D., David, P.H., et al. 2012, Whole genome sequencing of field isolates provides robust characterization of genetic diversity in *Plasmodium vivax*, *PLoS. Negl. Trop. Dis.*, **6**, e1811.
50. Young, A.S., Brown, C.G., Burridge, M.J., et al. 1978, The incidence of theilerial parasites in East African buffalo (*Syncerus caffer*), *Tropenmed. Parasitol.*, **29**, 281–8.
51. Conrad, P.A., Stagg, D.A., Grootenhuis, J.G., et al. 1987, Isolation of Theileria parasites from African buffalo (*Syncerus caffer*) and characterization with anti-schizont monoclonal antibodies, *Parasitology*, **94**, 413–23.
52. Bradley, D.G., MacHugh, D.E., Cunningham, P. and Loftus, R.T. 1996, Mitochondrial diversity and the origins of African and European cattle, *Proc. Natl. Acad. Sci. USA*, **93**, 5131–5.
53. Hanotte, O., Tawah, C.L., Bradley, D.G., et al. 2000, Geographic distribution and frequency of a taurine *Bos taurus* and an indicine *Bos indicus* Y specific allele amongst sub-saharan African cattle breeds, *Mol. Ecol.*, **9**, 387–96.
54. Hanotte, O., Bradley, D.G., Ochieng, J.W., Verjee, Y., Hill, E.W. and Rege, J.E. 2002, African pastoralism: genetic imprints of origins and migrations, *Science*, **296**, 336–9.
55. Collins, N.E. and Allsopp, B.A. 1999, *Theileria parva* ribosomal internal transcribed spacer sequences exhibit extensive polymorphism and mosaic evolution: application to the characterization of parasites from cattle and buffalo, *Parasitology*, **118**, 541–51.
56. MacHugh, N.D., Weir, W., Burrells, A., et al. 2011, Extensive polymorphism and evidence of immune selection in a highly dominant antigen recognized by bovine CD8 T cells specific for *Theileria annulata*, *Infect. Immun.*, **79**, 2059–69.
57. MacHugh, N.D., Connelley, T., Graham, S.P., et al. 2009, CD8+ T-cell responses to *Theileria parva* are preferentially directed to a single dominant antigen: implications for parasite strain-specific immunity, *Eur. J. Immunol.*, **39**, 2459–69.
58. Swan, D.G., Phillips, K., Tait, A. and Shiels, B.R. 1999, Evidence for localisation of a Theileria parasite AT hook DNA-binding protein to the nucleus of immortalised bovine host cells, *Mol. Biochem. Parasitol.*, **101**, 117–29.
59. Schmuckli-Maurer, J., Casanova, C., Schmied, S., et al. 2009, Expression analysis of the *Theileria parva* subtelomere-encoded variable secreted protein gene family, *PLoS One*, **4**, e4839.
60. Weir, W., Sunter, J., Chaussepied, M., et al. 2009, Highly syntenic and yet divergent: a tale of two Theilerias, *Infect. Genet. Evol.*, **9**, 453–61.
61. Ragoussis, J. 2009, Genotyping technologies for genetic research, *Annu. Rev. Genomics Hum. Genet.*, **10**, 117–33.

Supplementary Figure S1 Schematic diagram of SNP dataset construction

SNP dataset I represents all SNPs that fulfilled our quality control criteria in each strain. SNP dataset II represents the SNPs identified in all the bovine-derived *Theileria* strains, while SNP dataset III represents the combined SNPs for bovine and buffalo *Theileria* strains. SNP dataset II and SNP dataset III are SNP alleles useful for comparison between strains. At some reference positions, in which SNPs were detected at least in one strain, alleles for other strains were acquired. Alleles at these positions may not pass the quality control; therefore, these positions were discarded in SNP datasets II and III.



	SNP dataset Ⅱ	SNP dataset Ⅲ
Chr1	5,550	3,745
Chr2	3,376	1,788
Chr3_530	2,446	1,833
Chr3_531	1,144	589
Chr4	3,385	2,913
total	15,901	10,868

I All filter passed SNPs in each strains.

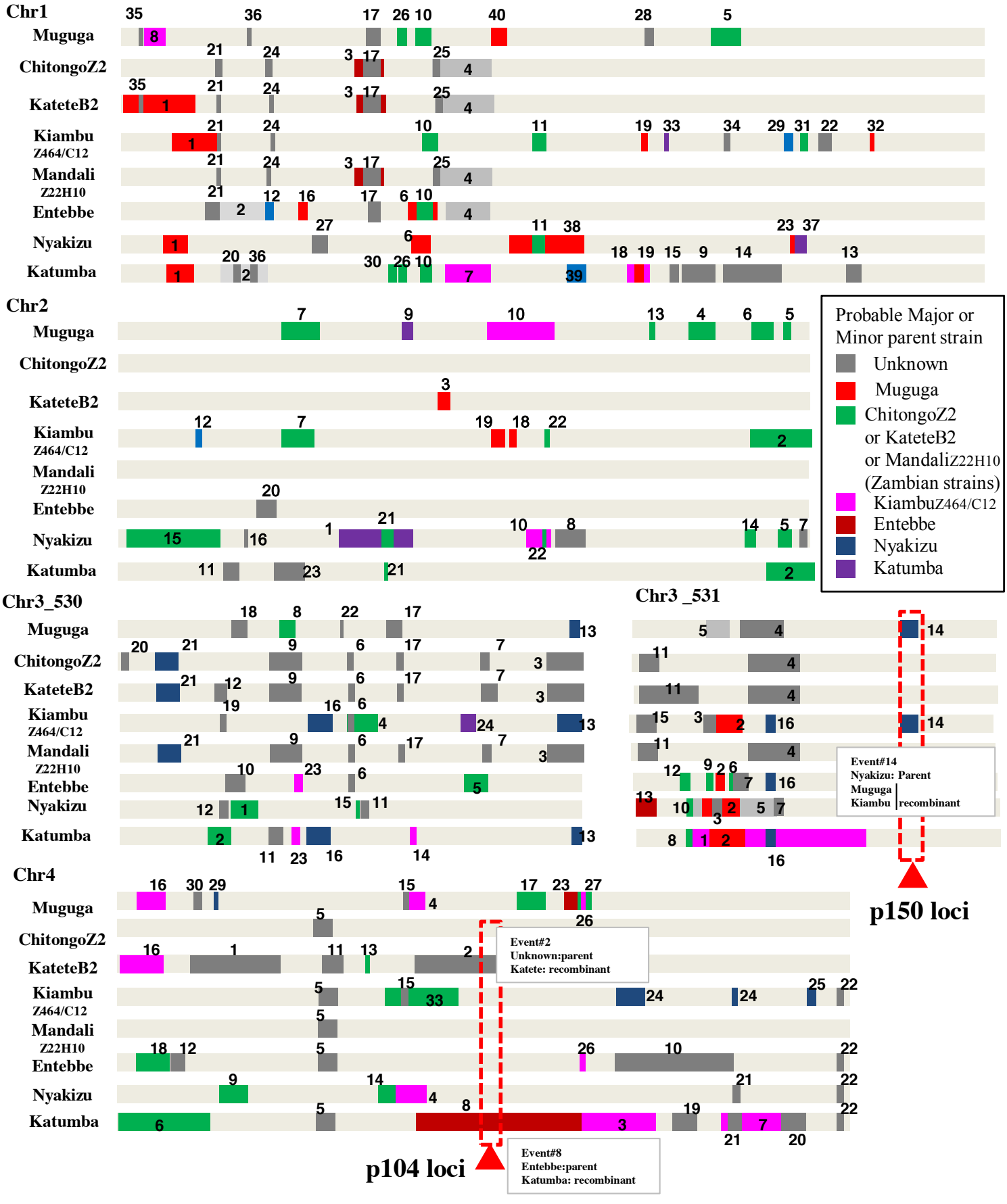
II From the SNP dataset I, 8 bovine *T. parva* data were taken together, and the only SNP position where every strain have passed quality criteria were extracted.

III Two buffalo-derived *Theileria* strains were included, again these position were required to have passed quality criteria for all strains.

Supplementary Figure S2

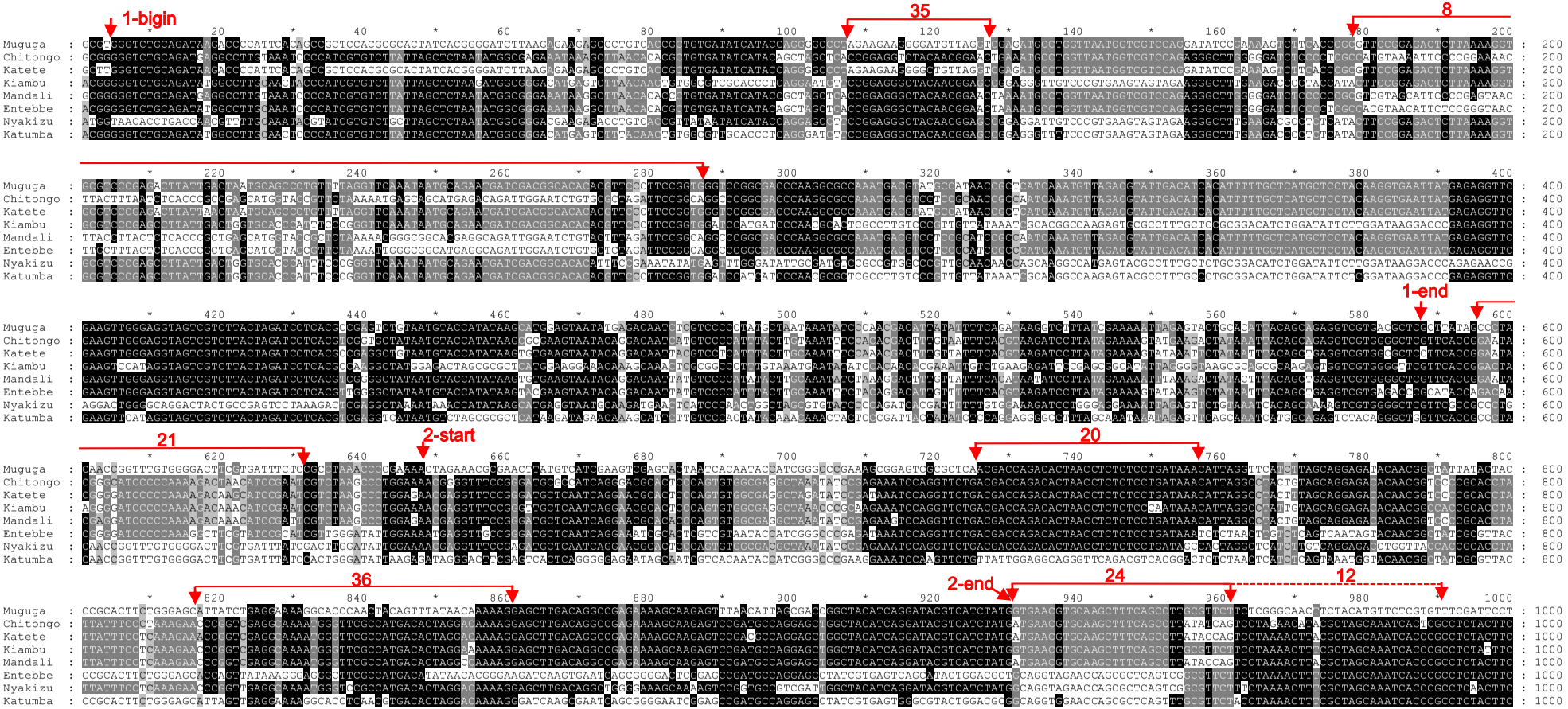
Schematic representation of putative recombination events in bovine *Theileria* strains

A total of 133 recombination events were predicted using the concatenated SNP dataset II using RDP software. All the recombination events detected by RDP3 are numbered in order, and origins of possible major or minor parent strains are colour-matched.



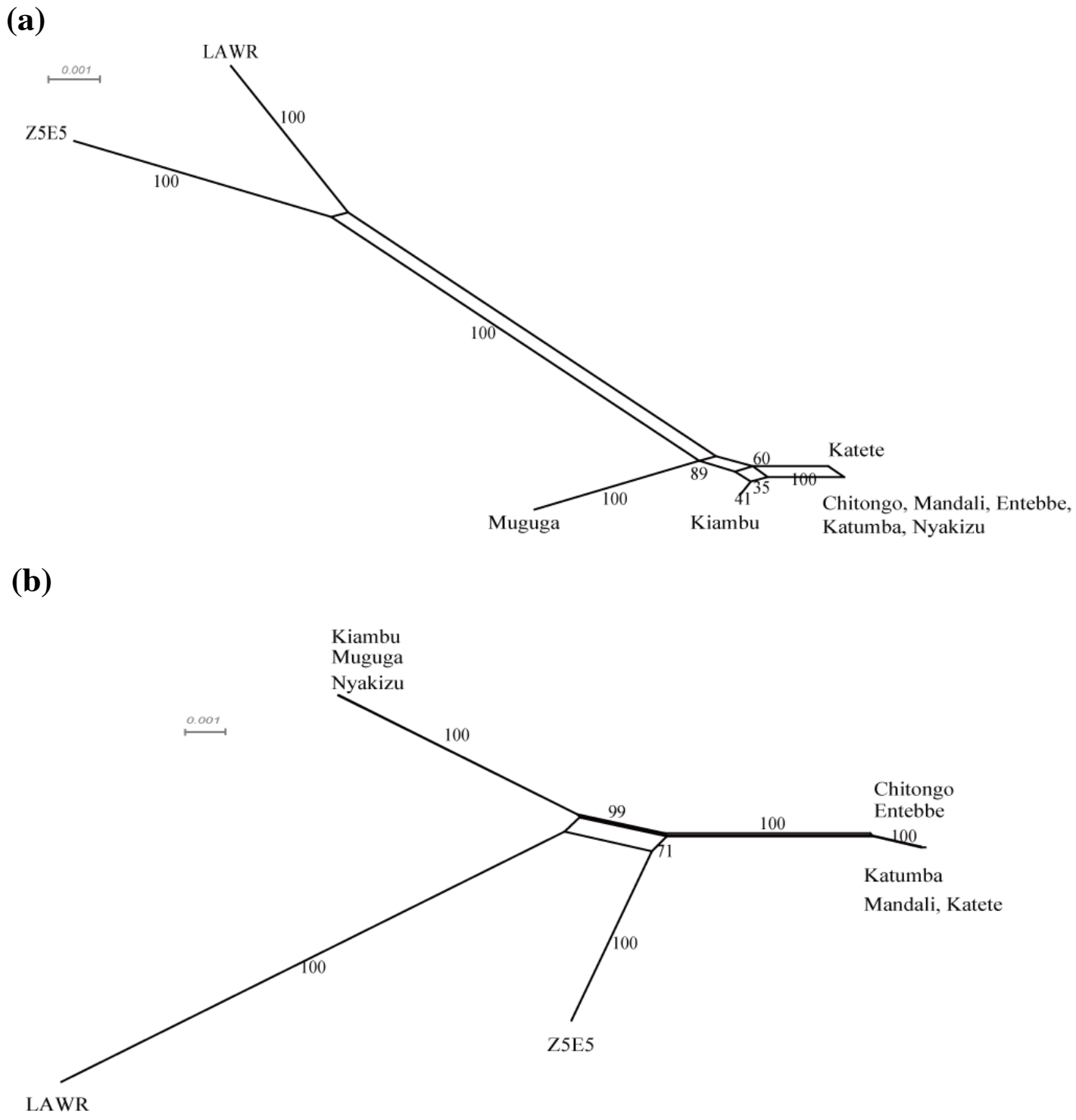
Supplementary Figure S3 : Alignment of an extracted portion of SNP dataset II

The allelic data for SNP positions #1–#1000 from dataset II were concatenated and aligned. These SNPs lie within a ~0.5 Mb region of chromosome 1 and were selected to provide a snapshot representing one chromosomal region. Black, dark grey, and light grey shading indicate sequence conservation among seven, six or five, or four alleles, respectively. An asterisk (*) or number indicate every 10th residue. The start and end-points of predicted recombination events are shown as arrows. Recombination event numbers correspond with the numbers used in Supplementary Fig.S2.

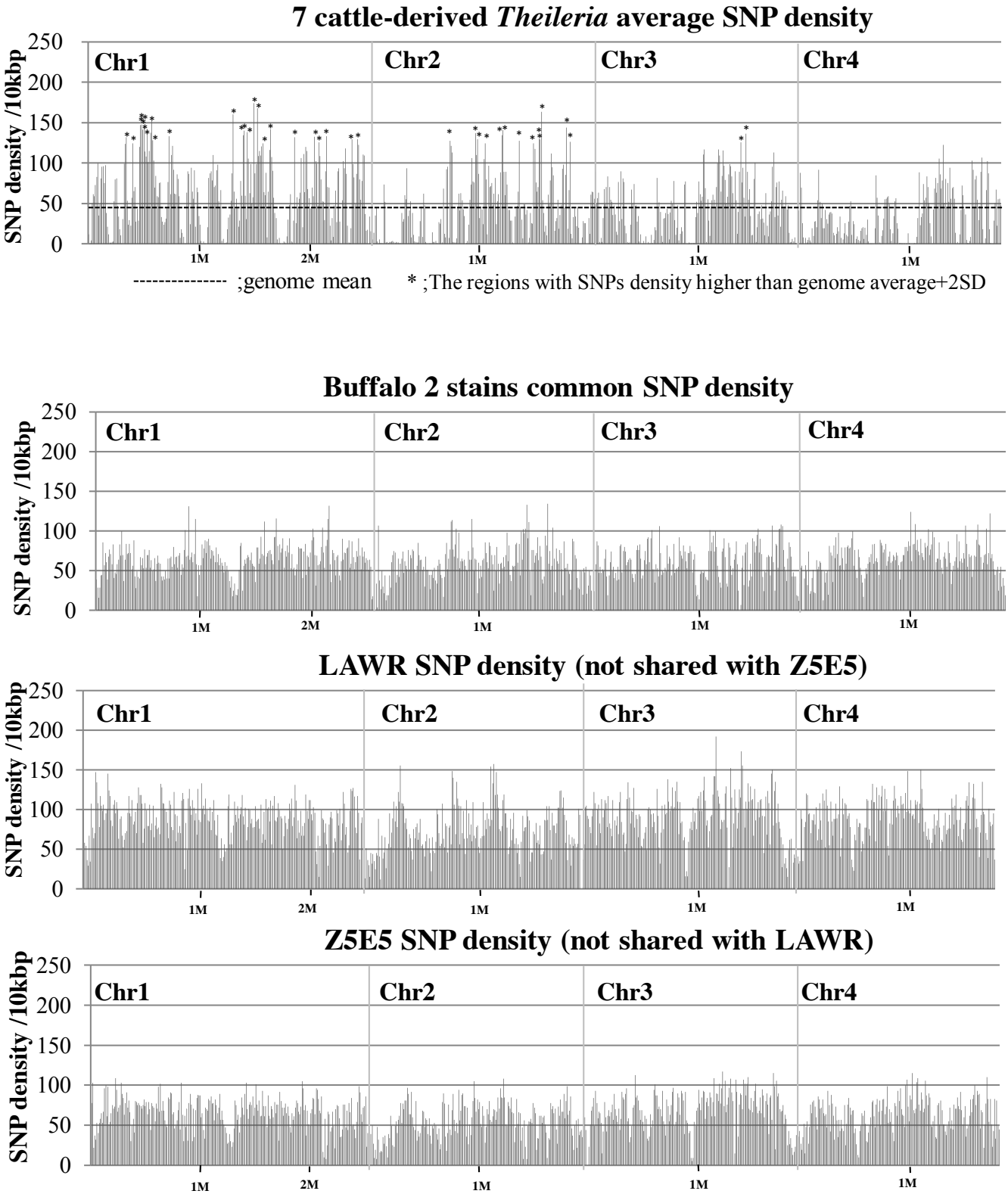


Supplementary Figure S4: Neighbor-net analysis of p104 (TP04_0437) sequences (a) and p150 (TP03_0861) sequences (b) between 10 *T. p. parva* strains

Neighbor-net trees base on the sequence of p104 and p150 was constructed using Split tree version 4.11.3. Bootstrap values shown close to branches are based on 100 bootstrap replicates.



Supplementary Figure S5: SNP distribution across the genome.
Average SNP densities were plotted alongside chromosome1-4. The x-axis shows the chromosomal position, and the left y-axis shows the number of SNPs (black bars) per 10kbp interval. Average in 7 cattle-derived *Theileria* strains was presented in the topmost graph. The mark (*) indicate the region where SNP density was above the genome average (+2SD). The SNPs of two buffalo-derived *Theileria* strains (LAWR and Z5E5) were plotted in the following three ways; SNPs shared between two strain (2nd graph), LAWR SNPs which were not shared with Z5E5 (3rd graph) and Z5E5 SNPs which were not shared with LAWR (4th graph).



Supplementary Table S1: high dN/dS list

signalP positive and high dNdS			Ortholog Group	ChitongoZ2	kateteB2	kiambuZ46 4/C12	MandaliZ2 2H10	Entebbe	Nyakizu	Katumba	LAWR	Z5E5	cattle-derived *Theileria* average	mpss* tpm	signalP4.0	GPI-SOM (C & N-terminal)	NMT - The MYR Predictor	Myristoylator	CSS-P a l m	TMHMM
	TP01_0144	hypothetical protein	PiroF0002444	1.0425	1.0425	1.0425	1.0425	0.4125	1.4384	0.8239	0.3419	0.3784	0.98	0	Y	N	N	N	Y	1
	TP01_0178	hypothetical protein	PiroF0002919	1.4366	0.7181	0.7181	0.7181	0.7181	0	0.8047	0.727	0.9519	0.73	9	Y	N	N	N	Y	0
	TP01_0180	40S ribosomal protein S11	PiroF0000589	0	0	0	0	0	1.129	0.6359	1.1235	infinite(1/1)	0.25	8	Y	N	N	N	Y	0
	TP01_0291	hypothetical protein	PiroF0002390	0	0.159	0	0.159	0.4	0.7675	0	0.1207	0.944	0.21	103	Y	N	N	N	N	2
	TP01_0367	hypothetical protein	PiroF0000012	1.1195	0.4305	0.4354	infinite(2/2)	0	0.2193	0.2151	0.1087	0.0755	0.35	0	Y	Y	N	N	N	0
	TP01_0378	hypothetical protein	PiroF0003402	0.941	1.122	0.9912	1.0407	1.6578	1.3164	1.9015	0.8666	1.2875	1.28	15	Y	N	N	N	Y	0
	TP01_0380	hypothetical protein	PiroF0003404	0.8229	0.7905	infinite(2/2)	0.6723	1.0161	0.957	0.64	0.8427	0.9716	0.70	1602	Y	N	N	N	N	0
	TP01_0610	hypothetical protein (Tash family)	PiroF0100038	0.5569	1.4743	0.7309	0.8145	0.558	0.4879	0.3724	0.5377	0	0.71	0	Y	N	N	N	N	0
	TP01_0619	hypothetical protein (Tash family)	PiroF0100038	0.3799	#N/A	0.5792	1.0094	0.2759	0	0	1.4195	0	0.32	0	Y	N	N	N	N	0
	TP01_0621	hypothetical protein (Tash family)	PiroF0100038	0.8235	0.6882	0.5493	0.5493	0.537	0.435	0.5967	0.702	0.3722	0.60	0	Y	N	N	N	Y	0
	TP01_0914	hypothetical protein	PiroF0002316	1.681	0.8478	0.8143	0.8478	0.8478	1.6347	0.8143	0.2797	1.4132	1.07	607	Y	N	N	N	N	0
	TP01_0955	hypothetical protein	PiroF0003569	0.6447	0.9696	0.9432	0.5637	0.8068	0.6447	0	1.405	0.9096	0.65	8	Y	N	N	N	N	0
	TP01_0987	hypothetical protein	PiroF0002967	infinite(5/5)	infinite(6/6)	infinite(9/9)	infinite(5/5)	infinite(5/5)	infinite(8/8)	infinite(6/6)	0.2514	0.1878	0.00	380	Y	N	N	N	N	0
	TP01_1011	hypothetical protein	PiroF0100045	0	0	0.4718	0	0	1.2933	0	0.6669	0.9052	0.25	0	Y	N	N	N	Y	0
	TP01_1044	hypothetical protein		0.3524	0.281	1.2797	0.3871	0.3303	0.3202	0.322	0.6671	0.5223	0.47	0	Y	N	N	N	Y	0
	TP01_1056	**32 kDa surface antigen**	PiroF0002963	0	0	0.6515	0	0	0.7972	infinite(1/1)	0.4721	0.3471	0.21	351	Y	Y	N	N	Y	0
	TP01_1109	hypothetical protein	PiroF0000207	0.638	0.6513	infinite(1/1)	0.957	0.5033	1.2198	0.638	0.1574	0.2517	0.66	103	Y	N	N	N	N	0
	TP01_1227	hypothetical protein (SVSP)	PiroF0100037	0.6739	infinite(5/5)	infinite(3/3)	1.3293	0.39	1.5254	infinite(4/4)	0.8692	0.638	0.56	33	Y	N	N	N	Y	0
	TP02_0004	hypothetical protein (SVSP)	PiroF0100037	0	infinite(1/1)	infinite(3/3)	infinite(1/1)	0	0.3813	1.7092	infinite(7/7)	infinite(3/3)	0.30	138	Y	N	N	N	Y	0
	TP02_0006	hypothetical protein (SVSP)	PiroF0100037	0.2188	0.2188	infinite(2/2)	0	0.2188	0.2188	infinite(4/4)	0.4308	0	0.13	128	Y	N	N	N	Y	0
	TP02_0010	hypothetical protein (SVSP)	PiroF0100037	1.5173	nfinite(10/10	infinite(2/2)	infinite(6/6)	1.7777	1.7777	1.7777	0	0.2393	0.98	7	Y	N	N	N	Y	0
	TP02_0018	hypothetical protein	PiroF0100055	infinite(3/3)	infinite(2/2)	0.946	infinite(3/3)	infinite(2/2)	infinite(2/2)	infinite(2/2)	0.3897	0.2724	0.14	46	Y	N	N	N	Y	0
	TP02_0239	hypothetical protein	PiroF0002609	infinite(4/4)	infinite(4/4)	infinite(1/1)	infinite(1/1)	infinite(1/1)	infinite(4/4)	infinite(3/3)	0.3607	0.1927	0.00	44	Y	N	N	N	Y	0
	TP02_0327	hypothetical protein	PiroF0000012	infinite(1/1)	infinite(1/1)	0.9598	infinite(1/1)	infinite(1/1)	0	infinite(1/1)	0	0	0.14	0	Y	N	N	N	N	0
	TP02_0331	ubiquitin activating enzyme, putatuve	PiroF0002575	0	0	0	0	0	0	0.918	0.239	0.2307	0.13	0	Y	N	N	N	Y	0
	TP02_0551	**23 kDa piroplasm surface protein**	PiroF0003021	infinite(2/2)	infinite(3/3)	infinite(4/4)	infinite(3/3)	infinite(3/3)	infinite(4/4)	infinite(2/2)	infinite(1/1)	0.3618	0.00	15	Y	Y	N	N	Y	2
	TP02_0575	hypothetical protein	PiroF0003017	0.3936	0	0.7631	0.3936	0.3936	0.7631	0	0.2651	0.8733	0.39	0	Y	N	N	N	N	0
	TP02_0819	hypothetical protein (FAINT superfamily)	PiroF0100056	0.849	1.0187	0	0.7722	0.978	0	0	0.3982	0.6176	0.52	0	Y	N	N	N	Y	0
	TP02_0856	hypothetical protein (FAINT superfamily)	PiroF0100056	0	0	0.3118	0	0	1.0399	0.308	0.2811	0.2973	0.24	0	Y	N	N	N	Y	0
	TP02_0875	hypothetical protein	PiroF0002985	0.5799	0.3531	0.7446	0.3531	0.2551	0.6999	0.666	0.6078	0.4638	0.52	0	Y	Y	N	N	Y	0
	TP02_0952	hypothetical protein	PiroF0003456	infinite(2/2)	infinite(2/2)	infinite(1/1)	infinite(2/2)	infinite(1/1)	1.4371	infinite(1/1)	0.453	1.4835	0.21	941	Y	N	N	N	Y	0
	TP02_0954	hypothetical protein (SVSP)	PiroF0100037	0	0	infinite(2/2)	0	infinite(2/2)	0.9263	0.4587	0.9604	0.5277	0.20	34	Y	N	N	N	Y	0
	TP02_0956	hypothetical protein (SVSP)	PiroF0100037	infinite(9/9)	infinite(9/9)	infinite(1/1)	infinite(8/8)	infinite(1/1)	0.5683	infinite(1/1)	1.5146	0.8902	0.08	0	Y	N	N	N	Y	0
	TP03_0001	hypothetical protein (SVSP)	PiroF0100037	0.5611	1.036	0.7568	0.8434	0.6743	0.3951	0.6496	0.3026	1.6111	0.70	0	Y	N	N	N	Y	0
	TP03_0002	hypothetical protein (SVSP)	PiroF0100037	1.2121	0.3745	0.8113	0.4633	0.6457	0.3544	0.3468	0.6059	0.2642	0.60	606	Y	N	N	N	Y	0
	TP03_0003	hypothetical protein (SVSP)	PiroF0100037	0.3337	0.6074	0.3541	0.5915	1.4207	0.2029	0.53	0.2692	0.3135	0.58	31	Y	N	N	N	Y	0
	TP03_0039	hypothetical protein		0.5023	0.5645	0.3754	0.5023	1.1749	0.1658	0.5023	0.9774	0.4684	0.54	284	Y	N	N	N	Y	0
	TP03_0040	hypothetical protein	PiroF0003613	0.6408	0.825	0.6408	0.4219	0.315	0.305	0.6408	0.2459	0.3014	0.54	0	Y	N	N	N	Y	0
	TP03_0123	hypothetical protein	PiroF0002851	0.43	0.4006	0.3831	0.6309	1.7277	0.4352	0.4882	0.6048	0.3307	0.64	68	Y	N	N	N	N	0
	TP03_0217	hypothetical protein	PiroF0000012	0.1731	0.0685	1.2578	0	0	infinite(2/2)	infinite(1/1)	0.6254	0.1482	0.21	0	Y	N	N	N	N	0
	TP03_0297	hypothetical protein (FAINT superfamily)	PiroF0100056	0	0	1.0335	0	infinite(1/1)	1.0335	2.0641	0.591	0.4818	0.59	0	Y	N	N	N	Y	0
	TP03_0298	hypothetical protein	PiroF0000056	0	0	0.8134	0	0	0.6661	1.3555	0.1691	0.1375	0.41	0	Y	N	N	N	Y	0
	TP03_0319	hypothetical protein	PiroF0000012	0	infinite(1/1)	0	0	0.2461	0	0.8452	0.4104	1.2054	0.16	0	Y	N	N	N	N	0
	TP03_0368	hypothetical protein, conserved	PiroF0100056	0.3131	0.2154	0.3131	0.1892	0	0.3131	0.8192	0.3967	0.7685	0.31	0	Y	N	N	N	Y	0
	TP03_0405	hypothetical protein	PiroF0002425	0	0	1.0376	0	0.3176	0	0.4144	0.2575	0.2218	0.25	0	Y	Y	N	N	Y	0
	TP03_0498	hypothetical protein (SVSP)	PiroF0100037	1.1282	0.9685	1.2624	1.2642	1.066	1.2618	1.4023	1.3029	1.026	1.19	98	Y	N	N	N	Y	0
	TP03_0520	hypothetical protein	PiroF0000012	0.7471	0.4219	0	1.2794	0.1612	0.2818	0	0.4933	0	0.41	18	Y	N	N	N	N	0
	TP03_0530	hypothetical protein		1.0054	0.8016	2.1445	infinite(3/3)	0	2.5138	0	infinite(7/7)	0.9878	0.92	0	Y	N	N	N	N	0
	TP03_0664	hypothetical protein	PiroF0000012	infinite(1/1)	infinite(2/2)	0.9966	infinite(2/2)	infinite(5/5)	infinite(7/7)	0.7354	0.2009	0.4886	0.25	0	Y	N	N	N	N	0
	TP03_0780	hypothetical protein	PiroF0002660	0	0	0.0578	0	0.151	0.1206	0.0578	0.3103	infinite(6/6)	0.06	0	Y	N	N	N	Y	0
	TP03_0810	hypothetical protein	PiroF0002675	0.2652	0.2435	0.1468	0.2433	0.7497	0.6237	0.1517	0.3391	0.2131	0.35	4	Y	N	N	N	Y	0
	TP03_0886	hypothetical protein (SVSP)	PiroF0100037	0.425	0.429	0.4257	0.7522	0.751	0.7522	0.4969	0.8823	0.3741	0.58	66	Y	N	N	N	Y	0
	TP03_0893	hypothetical protein (SVSP)	PiroF0100037	0.2352	0.9296	0	infinite(4/4)	infinite(6/6)	0.2126	0	0.3021	0.1502	0.20	0	Y	N	N	N	Y	0
	TP04_0009	hypothetical protein (SVSP)	PiroF0100037	0.7886	infinite(9/9)	0.2207	0.885	0.3386	0.3812	0.2207	1.2752	0.2814	0.40	300	Y	N	N	N	Y	0
	TP04_0012	hypothetical protein (FAINT superfamily)	PiroF0100056	1.4742	0.7032	#N/A	1.6609	0.5854	0.5143	#N/A	0.965	1.7088	0.71	0	Y	N	N	N	N	0
	TP04_0013	hypothetical protein (SVSP)	PiroF0100037	1.3413	0.851	0.8804	1.0932	0.486	0.6102	1.0322	1.1612	0.6202	0.90	13	Y	N	N	N	Y	0
	TP04_0096	hypothetical protein (FAINT superfamily)	PiroF0100056	0.4255	0.0761	0.9973	0.83	0.5441	0	0.5554	0.2299	0	0.49	0	Y	N	N	N	Y	0
	TP04_0097	hypothetical protein (FAINT superfamily)	PiroF0100056	0.5611	0.108	0.6407	0.2408	0.3557	0.8515	0.3211	#N/A	0	0.44	0	Y	N	N	N	Y	0
	TP04_0101	hypothetical protein (FAINT superfamily)	PiroF0100056	0	1.4059	0.5301	0	0	0.4799	0	0.5046	0	0.35	0	Y	N	N	N	Y	0
	TP04_0104	hypothetical protein (FAINT superfamily)	PiroF0100056	0.6437	0	0.5902	0.4796	0.3407	2.4726	0.5902	0.609	0.4003	0.73	0	Y	N	N	N	N	0
	TP04_0110	hypothetical protein, conserved	PiroF0001224	0.2023	0	1.2025	0.2608	0.1924	0.1413	0.1419	0.2186	0.1358	0.31	0	Y	N	N	N	N	0
	TP04_0116	hypothetical protein	PiroF0003546	0.1808	0	0.2833	0.1332	2.3416	0.6762	0.089	0.2182	0.4104	0.53	45	Y	N	N	N	Y	1
	TP04_0150	hypothetical protein	PiroF0000037	0	0	2.0623	0	1.3929	0	2.0902	0.4012	0.873	0.79	0	Y	Y	N	N	Y	1
	TP04_0328	hypothetical protein	PiroF0002219	0	0	0	0	0	0	0.7583	0.4713	0.6122	0.11	0	Y	N	N	N	Y	0
	TP04_0411	hypothetical protein	PiroF0003185	2.2732	infinite(1/1)	2.2732	2.2732	infinite(3/3)	2.2732	2.2732	1.0763	1.4158	1.62	0	Y	N	N	N	N	0
	TP04_0437	**104 kDa antigen**	PiroF0003088	0.9455	0.685	0.7699	1.7544	0.9459	0.7699	1.0901	1.0587	1.2416	0.99	1294	Y	Y	N	N	Y	1
	TP04_0558	hypothetical protein, conserved	PiroF0001517	0	0	0.9234	0	infinite(2/2)	infinite(3/3)	0.9234	0.1602	0.1128	0.26	4	Y	N	N	N	Y	0
	TP04_0919	hypothetical protein (SVSP)	PiroF0100037	0.556	0.4163	0.7825	0.3154	0.5553	0.6747	0.6057	0.1986	0.2115	0.56	96	Y	N	N	N	Y	0
	TP04_0920	hypothetical protein (SVSP)	PiroF0100037	0.5272	0.4468	0.7808	0.3235	0.5795	0.5742	0.6699	0.5895	0.4561	0.56	22	Y	N	N	N	Y	0
	TP04_0921	hypothetical protein (FAINT superfamily)	PiroF0000056	2.2105	1.6782	0.5092	0.7833	0.3794	0.4488	0.8927	0.4682	0.287	0.99	0	Y	N	N	N	Y	0
signalP negative and high dNdS			Ortholog Group	ChitongoZ2	kateteB2	kiambuZ46 4/C12	MandaliZ2 2H10	Entebbe	Nyakizu	Katumba	LAWR	Z5E5	cattle-derived *Theileria* average	mpss* tpm	signalP4.0	GPI-SOM (C & N-terminal)	NMT - The MYR Predictor	Myristoylator	CSS-P a l m	TMHMM
	TP01_0033	hypothetical protein		infinite(4/4)	0	0.3423	infinite(4/4)	infinite(4/4)	0.3423	0.295	0.2663	0.3256	0.14	0	N	N	N	N	Y	0
	TP01_0046	hypothetical protein	PiroF0002912	0.7657	0	0.3454	0.4921	0.1678	0.3454	0.3454	0.458	0.6889	0.35	0	N	N	N	N	N	0
	TP01_0047	hypothetical protein	PiroF0001900	1.0587	0	0	0.5722	1.0587	infinite(1/1)	0	0.7533	0	0.38	0	N	N	N	N	N	0
	TP01_0073	orotate phosphoribosyltransferase	PiroF0000754	0.8446	0	0.1581	0.8446	0.4214	0.1405	0.0974	0.0323	0.1317	0.36	26	N	N	N	N	N	0
	TP01_0120	hypothetical protein	PiroF0000089	0	0.6937	0	0.4303	0.6937	0	0	0.1677	0.0438	0.26	49	N	N	N	N	N	0
	TP01_0122	hypothetical protein	PiroF0000089	0.075	0.101	0.1182	0.0852	0.074	0.0564	1.2799	0.1661	0.0802	0.26	0	N	N	N	N	Y	0
	TP01_0134	hypothetical protein		infinite(1/1)	infinite(1/1)	0.6289	0	infinite(1/1)	infinite(1/1)	0.9374	0.3336	0.9571	0.22	0	N	N	N	N	Y	0
	TP01_0145	histidyl-tRNA synthetase	PiroF0000725	0.3739	0.7478	0.7478	0.3737	0.7478	infinite(1/1)	0	0.0599	0.0774	0.43	130	N	N	N	N	Y	0
	TP01_0183	hypothetical protein	PiroF0002812	0.8482	0.8482	0.6132	0.8482	1.1365	0.4343	0.1645	0.6945	1.2363	0.70	68	N	N	N	N	N	0
	TP01_0205	hypothetical protein		0.442	0.3309	0.418	0.839	0.669	0.418	0	0.3197	0.2633	0.45	116	N	N	N	N	Y	0
	TP01_0253	hypothetical protein	PiroF0000129	infinite(3/3)	infinite(3/3)	infinite(4/4)	infinite(4/4)	infinite(3/3)	0.1939	infinite(4/4)	0.283	0.4569	0.03	815	N	N	N	N	Y	2
	TP01_0270	hypothetical protein	PiroF0000554	0.5464	0.3601	0.3654	1.1085	0	0.7701	1.0194	0.6065	0.3003	0.60	0	N	N	N	N	Y	0
	TP01_0288	hypothetical protein	PiroF0100056	0.9399	2.7428	0.7902	0.9784	0	0.8336	0.2134	0.6086	1.0821	0.93	0	N	N	N	N	Y	0
	TP01_0309	hypothetical protein	PiroF0000569	0.377	0	0.7533	0.377	0.377	0.377	0.0976	0.6376	0.1888	0.34	68	N	N	N	N	N	0
	TP01_0337	hypothetical protein	PiroF0003399	0	0	1.4066	0	0.3865	0	0.6438	0.6145	0.8538	0.35	0	N	N	N	N	N	0
	TP01_0344	hypothetical protein	PiroF0000583	0	0	infinite(2/2)	0	0	0.7825	0.7825	0.0603	#N/A	0.22	44	N	N	N	N	Y	0
	TP01_0358	hypothetical protein	PiroF0002364	0.2592	0.2484	1.9501	0.2615	0.2367	0.2742	0.2266	0.2265	0.3425	0.49	150	N	N	TWILIGHT ZONE	Medium Confidence	Y	0
	TP01_0365	hypothetical protein	PiroF0002362	0.7746	0.7746	0	infinite(1/1)	infinite(1/1)	0.7746	0.7746	0.128	0.1079	0.44	75	N	N	N	N	N	1
	TP01_0395	hypothetical protein	PiroF0000499	#N/A	infinite(3/3)	0	infinite(4/4)	0	0	0.3857	1.4015	0.2387	0.06	0	N	N	N	N	N	0
	TP01_0405	hypothetical protein	PiroF0000493	1.0398	1.0398	0.605	0.3462	0.693	0.2977	0.6128	0.6148	0.2756	0.66	0	N	N	N	N	Y	0
	TP01_0415	hypothetical protein	PiroF0000125	1.0721	1.5952	1.047	1.347	0.6497	1.0721	0.7868	0.2662	0.2441	1.08	0	N	N	N	N	Y	0
	TP01_0462	hypothetical protein		2.2693	1.6445	1.6445	1.6445	0.9652	infinite(5/5)	1.6445	0	0	1.40	0	N	N	N	N	Y	0
	TP01_0463	hypothetical protein	PiroF0100056	0.2625	0.2593	0.2827	0.3234	1.545	0.2969	0.2689	0.3262	0.3845	0.46	16	N	N	N	N	N	0
	TP01_0532	hypothetical protein	PiroF0002901	0.4785	0.9208	1.4705	1.2213	0.4866	0.2424	0.95	0.4842	0.4522	0.82	84	N	N	N	N	N	0
	TP01_0551	hypothetical protein	PiroF0100055	0.5548	0.7236	0.8459	1.0883	1.088	0.3042	0.1787	0.3337	0.2572	0.68	0	N	N	N	N	Y	0
	TP01_0562	hypothetical protein		#N/A	#N/A	#N/A	#N/A	#N/A	1.5488	1.2408	0.9419	0.3916	0.40	0	N	N	N	N	Y	0
	TP01_0563	hypothetical protein	PiroF0100029	0.3608	0.2384	0.1857	0.2071	0.2732	0.149	0.8789	0.2709	0.2937	0.33	20	N	N	N	N	N	8
	TP01_0567	hypothetical protein	PiroF0001836	0.4268	0.3626	0	0.321	0.3015	0.168	1.5872	0.6969	0.2004	0.45	124	N	N	N	N	N	0
	TP01_0599	hypothetical protein	PiroF0000174	0.7143	0.7143	0.7143	0.7143	0.7143	0.3103	0.7143	0.1544	0.1075	0.66	0	N	N	N	N	N	0
	TP01_0613	hypothetical protein (Tash family)	PiroF0100037	1.8764	infinite(2/2)	infinite(1/1)	0	infinite(1/1)	0.5355	0	0.6943	1.3081	0.34	75	N	N	N	N	N	0
	TP01_0615	hypothetical protein (Tash family)	PiroF0100037	0.5052	0.4213	0.7713	0.3188	infinite(1/1)	0.7713	0.7645	0.663	0.5503	0.51	3578	N	N	N	N	Y	0
	TP01_0617	hypothetical protein (Tash family)	PiroF0100037	1.1608	0.9289	1.1608	0.8647	1.1608	0.3816	1.5251	0.5076	0.7607	1.03	11	N	N	N	N	Y	1
	TP01_0626	hypothetical protein	PiroF0001966	0.5033	0.5033	0.5033	1.0134	0.5033	0.3352	1.0158	#N/A	#N/A	0.63	0	N	N	N	N	N	0
	TP01_0627	hypothetical protein		0.7799	0.7799	0.42	1.1726	0.7799	0.3122	0.3905	0.8261	0.781	0.66	70	N	N	N	N	Y	0
	TP01_0668	hypothetical protein		1.2891	0.8689	0.98	0.8689	0	0	0.98	0.8837	0.4309	0.71	148	N	N	N	N	Y	0
	TP01_0735	hypothetical protein	PiroF0002019	0.668	0.3874	0.668	0.7878	0.6031	0.668	0.5004	0.1405	0.4129	0.61	0	N	N	N	N	Y	0
	TP01_0742	hypothetical protein	PiroF0003566	0.6971	0.6971	0.6971	0.6971	1.66	0.8279	0	0.5119	0.4909	0.75	17	N	N	N	N	N	0
	TP01_0745	hypothetical protein	PiroF0000080	1.386	infinite(3/3)	1.386	1.386	0.4569	0.5358	0.457	0.2289	0.1075	0.80	0	N	N	N	N	N	1
	TP01_0755	hypothetical protein	PiroF0002270	0.536	0.5084	0.4657	0.8041	0.4217	0.4657	0	0.223	0.3442	0.46	38	N	N	N	N	Y	0
	TP01_0867	hypothetical protein	PiroF0003207	infinite(1/1)	infinite(2/2)	0.5392	0	infinite(2/2)	0.222	0.7853	0.4021	0.6078	0.22	0	N	N	N	N	Y	0
	TP01_0905	hypothetical protein	PiroF0000373	1.0013	0.626	1.0013	0.7508	0.626	0	0.626	0.1391	0.2356	0.66	7	N	N	N	N	N	0
	TP01_0953	hypothetical protein	PiroF0000398	0.2845	0.38	0.3127	0.38	0.8921	0.3392	infinite(1/1)	0.2533	0.4243	0.37	0	N	N	N	N	Y	0
	TP01_1027	hypothetical protein	PiroF0000591	0.235	0.2737	0.4086	0.2915	0.2509	0.166	0.7982	0.1324	0.1251	0.35	7	N	N	N	N	Y	0
	TP01_1038	hypothetical protein		0	0	0	0	0	1.8082	0	1.4967	1.6452	0.26	0	N	N	N	N	Y	0
	TP01_1054	hypothetical protein	PiroF0000776	infinite(2/2)	infinite(2/2)	0.7814	infinite(2/2)	infinite(2/2)	0.1987	0.2591	0.1833	0.1561	0.18	18	N	N	N	N	Y	0
	TP01_1055	hypothetical protein	PiroF0002964	0	0	0.5068	0	0	1.1754	0.5767	0.6846	0.3316	0.32	111	N	N	N	N	Y	0
	TP01_1062	replication factor C subunit 5	PiroF0000621	0.7659	0.7659	0	infinite(1/1)	0.3821	0	infinite(2/2)	0.0584	0.0889	0.27	0	N	N	N	N	Y	0
	TP01_1106	hypothetical protein	PiroF0000206	1.7622	0.529	0	0.4679	0.7062	0.4229	0.7055	0.1943	0.1845	0.66	0	N	N	N	N	N	0
	TP01_1108	hypothetical protein	PiroF0002251	0.3166	0.2988	0.841	0.3258	0.2556	0.3166	0.2988	0.1866	0.1715	0.38	0	N	N	N	N	Y	9
	TP01_1128	hypothetical protein	PiroF0002257	3.4777	1.8274	0	1.6621	1.3082	2.2922	1.8163	0.9009	0.6919	1.77	0	N	N	N	N	N	0
	TP01_1133	orotidine 5-phosphate decarboxylase	PiroF0000218	0.3601	0.3866	0	0.3542	0.6879	0.4034	0.2897	0.4392	0.3905	0.35	20	N	N	N	N	Y	0
	TP01_1191	hypothetical protein	PiroF0001969	0.406	0.2919	0.3078	0.2451	1.1103	0.1302	0.72	0.1637	0.1025	0.46	411	N	N	N	N	Y	0
	TP01_1208	hypothetical protein	PiroF0002347	0.2854	0.2253	infinite(2/2)	0.2253	0.7181	0.8236	0.7181	0.1217	0.2314	0.43	0	N	N	N	N	Y	0
	TP01_1211	hypothetical protein	PiroF0000473	0.3368	0.4229	0.532	0.3755	0.8431	0.4229	0.8431	0.4907	0.3162	0.54	0	N	N	N	N	N	0
	TP01_1224	hypothetical protein	PiroF0003346	0	0	0.3811	0	infinite(1/1)	0.3811	0.7627	infinite(1/1)	0	0.22	0	N	N	N	N	N	0
	TP02_0001	hypothetical protein	PiroF0100055	1.0059	1.2604	0.7015	0.6683	0.698	1.2604	0.6598	0.4312	0.5158	0.89	0	Y	N	N	N	Y	0
	TP02_0025	ubiquitin-protein ligase	PiroF0002483	1.0275	1.0275	1.0275	1.0275	0.8345	1.0275	1.0275	0.5288	0.3082	1.00	81	N	N	N	N	Y	0
	TP02_0057	N-myristoyltransferase	PiroF0000823	0.09	0.3627	0.1018	0.2409	0.7256	0.18	0.18	0.1225	0.1796	0.27	5	N	N	N	N	N	0
	TP02_0179	hypothetical protein	PiroF0002143	0	0	0.7336	0	0	0	0	0.1759	0.1481	0.10	20	N	N	N	N	Y	0
	TP02_0182	hypothetical protein	PiroF0001376	1.0602	0.5303	0	0.6198	0.835	0.806	0.6164	0.3294	0.4529	0.64	505	N	N	N	N	Y	4
	TP02_0219	hypothetical protein	PiroF0002612	1.2935	1.2935	1.2935	infinite(3/3)	0	1.2935	0.6452	0.6781	0.5988	0.83	0	N	N	N	N	Y	0
	TP02_0226	hypothetical protein	PiroF0002614	0.8725	0.8725	0	0	0.8725	0.8725	0.8725	0.3637	0.6124	0.62	0	N	N	N	N	Y	0
	TP02_0227	hypothetical protein		0.7468	0.7468	0	0.7468	0.7468	0.7468	0.7468	0	0.3683	0.64	98	N	N	N	N	Y	1
	TP02_0285	hypothetical protein	PiroF0001498	1.3102	infinite(2/2)	infinite(1/1)	1.3102	1.3102	1.3102	0.1646	0.2841	0.2152	0.77	6	N	N	N	N	Y	5
	TP02_0295	hypothetical protein	PiroF0001506	0	infinite(1/1)	0	0	0	0	0.7218	0.158	0.0529	0.10	0	N	N	N	N	N	0
	TP02_0303	hypothetical protein	PiroF0001049	infinite(1/1)	0	1.1795	infinite(1/1)	infinite(1/1)	infinite(1/1)	0	0.4165	0.3462	0.17	81	N	N	N	N	Y	0
	TP02_0366	hypothetical protein	PiroF0001068	0	0	1.7831	0	0	0	0	0.1279	0.1089	0.25	0	N	N	N	N	N	0
	TP02_0406	hypothetical protein	PiroF0001112	0.8735	0.7153	0.8735	0.8735	0.5946	0.519	0.8735	0.3167	0.2679	0.76	62	N	N	N	N	Y	0
	TP02_0465	hypothetical protein	PiroF0003264	0.126	0.2099	0	0.7685	0.2099	0.1056	0	0	0.4493	0.20	5	N	N	N	N	N	0
	TP02_0519	hypothetical protein	PiroF0003023	0.1599	0.1599	0.2325	0.1599	0.2006	0.7525	0.1912	0.1571	0.1571	0.27	0	N	N	N	N	Y	0
	TP02_0523	hypothetical protein		0.7848	0.7848	0.7848	infinite(2/2)	0.7848	0.7848	0.7848	0.7731	0	0.67	0	N	N	N	N	Y	1
	TP02_0526	hypothetical protein	PiroF0000198	1.519	1.5102	1.2567	1.7044	0.9423	1.8692	1.271	0.7804	0.6624	1.44	648	N	N	N	N	Y	0
	TP02_0527	hypothetical protein	PiroF0100040	1.0185	0.9096	0.9117	1.0067	0.949	0.6731	0.9628	0.5274	0.9032	0.92	0	N	N	N	N	N	0
	TP02_0536	hypothetical protein	PiroF0002156	0.8418	0.4119	0	0.523	0.4119	0	0	0.1966	0	0.31	74	N	N	N	N	N	0
	TP02_0560	proteasome precursor	PiroF0001437	0.8201	0.3959	0	0.1969	0.3959	0	0	0.0207	0.0441	0.26	570	N	N	N	N	Y	0
	TP02_0564	hypothetical protein	PiroF0100022	0.9215	0	0.5526	0.7379	1.4729	0	0	0.8388	0.7838	0.53	0	N	N	N	N	N	0

TP02_0594	hypothetical protein	PiroF0002702	0.1891	0.1229	0.8195	0.1891	0.3895	0.6415	0.2992	0.3961	0.1304	0.38	91	N	N	N	N	Y	0
TP02_0646	hypothetical protein	PiroF0003482	0.8726	0.5844	#N/A	0.9331	0.4059	0.8865	#N/A	0.987	0.7828	0.53	0	N	N	N	N	Y	0
TP02_0668	hypothetical protein	PiroF0001394	0.4098	0.8219	0.272	infinite(2/2)	infinite(1/1)	0	0.272	0.3077	0.1857	0.25	877	N	N	N	N	Y	1
TP02_0705	hypothetical protein	PiroF0100034	infinite(3/3)	1.5262	0.0479	1.5262	0.7615	0.3768	0.0661	0.0651	0	0.61	0	N	N	N	N	N	1
TP02_0788	hypothetical protein	PiroF0100056	0	0.1668	0.7333	0.3345	0.3345	0.3358	0.2647	0.3861	0.8758	0.31	0	N	N	N	N	Y	0
TP02_0802	hypothetical protein	PiroF0100056	1.2816	0.9475	0	2.7701	1.0969	0.5373	0	0.437	0.3631	0.95	5	N	N	N	N	Y	0
TP02_0808	hypothetical protein	PiroF0003471	1.4859	1.3462	0	1.8337	1.247	0	0	0.8138	0.6008	0.84	0	N	N	N	N	N	0
TP02_0810	farnesyltransferase subunit beta	PiroF0001153	0.8419	0.802	infinite(7/7)	0.9082	0.6423	infinite(1/1)	infinite(7/7)	0.4484	0.2981	0.46	0	N	N	N	N	Y	0
TP02_0811	hypothetical protein	PiroF0003470	1.409	1.409	infinite(4/4)	1.409	0.9666	0	infinite(4/4)	0.6507	0.3096	0.74	71	N	N	N	N	Y	0
TP02_0816	hypothetical protein	PiroF0002993	0.4598	0.5863	0.5918	0.6357	0.8221	0	0.5853	2.5159	0.4639	0.53	0	N	N	N	N	Y	0
TP02_0824	hypothetical protein	PiroF0002992	0.2632	0.4562	infinite(5/5)	0.4603	0.4915	0	infinite(4/4)	0.167	0.1397	0.24	40	N	N	N	N	Y	0
TP02_0844	hypothetical protein	PiroF0002991	0	0	infinite(1/1)	0	0	1.7207	infinite(1/1)	0.3949	0.7791	0.25	0	N	N	N	N	Y	0
TP02_0870	hypothetical protein	PiroF0001907	0.8758	0.6161	0.3511	0.6764	0.5181	0.2374	0.7251	0.1934	0.4039	0.57	99	N	N	N	N	N	0
TP02_0873	hypothetical protein	PiroF0001908	0.5669	0.513	1.2254	0.9896	0.2095	0.5471	0.2174	0.0698	0.2594	0.61	0	N	N	N	N	N	0
TP02_0888	hypothetical protein	PiroF0003462	3.6097	4.1491	0.3748	infinite(6/6)	infinite(2/2)	0	0	0.335	0.6543	1.16	221	N	N	N	N	N	0
TP02_0889	hypothetical protein		0	0	0.9052	0	0	0	0	0.3834	0.3985	0.13	99	N	N	N	N	N	0
TP02_0896	hypothetical protein	PiroF0100041	0.5844	2.7388	0.5379	1.2618	0.222	0.7147	0.7563	0.511	1.0312	0.97	43	N	N	N	N	N	0
TP03_0004	hypothetical protein (SVSP)	PiroF0100037	0	1.6445	0.096	0.8345	0.8345	0.2914	0.1443	0.868	0.6312	0.55	69	N	N	N	N	N	0
TP03_0012	hypothetical protein	PiroF0001177	infinite(1/1)	1.1081	0.1325	1.1081	0.1478	0.0989	0.0802	0.0403	0.0635	0.38	113	N	N	N	N	Y	0
TP03_0066	hypothetical protein	PiroF0002198	0.4374	0.4992	0.6226	0.5692	infinite(8/8)	0.6226	0.7821	0.851	0.4792	0.50	227	N	N	N	N	Y	0
TP03_0079	hypothetical protein	PiroF0001844	0.0827	0.0571	0.1046	0.0672	0.7217	0.0842	0.143	0.0944	0.0487	0.18	0	N	N	N	N	Y	10
TP03_0086	hypothetical protein	PiroF0100056	1.4929	1.6638	1.3972	1.0393	0	0.6879	0.2651	infinite(1/1)	0.554	0.94	0	N	N	N	N	Y	0
TP03_0095	hypothetical protein	PiroF0002200	0	0	0	0	infinite(1/1)	0.8936	0	0.2093	0.3656	0.13	230	N	N	N	N	Y	0
TP03_0096	hypothetical protein		infinite(1/1)	infinite(1/1)	0.8403	infinite(1/1)	infinite(1/1)	0	infinite(1/1)	0.0822	0.1357	0.12	175	N	N	N	N	N	0
TP03_0097	rhoGAP protein, putative	PiroF0001855	0.243	0.2428	0.2015	0.2119	0.839	0.0828	0.2481	0.1001	0.1861	0.30	58	N	N	N	N	Y	0
TP03_0098	hypothetical protein	PiroF0003350	0.1519	0	0.4606	0.4606	1.4278	1.4278	0.4606	0.6027	0.5179	0.63	0	N	N	N	N	N	1
TP03_0099	hypothetical protein	PiroF0100056	0.7286	0	infinite(2/2)	0.7286	0.728	0	0	0.2465	infinite(2/2)	0.31	0	N	N	N	N	Y	0
TP03_0114	hypothetical protein	PiroF0100056	0.5644	0.5522	0.2603	1.1289	0.9662	0.2698	0.5321	0.2903	0.2577	0.61	0	N	N	N	N	Y	0
TP03_0119	hypothetical protein	PiroF0001864	0	0	0.9674	0	0.9674	0	0	0.1956	0.1827	0.28	0	N	N	N	N	Y	0
TP03_0127	hypothetical protein	PiroF0003583	0.4481	0.9025	0.47	0.4481	1.1838	infinite(1/1)	0.4481	0.4532	0.909	0.56	0	N	N	N	N	N	0
TP03_0188	hypothetical protein	PiroF0002830	infinite(1/1)	infinite(1/1)	infinite(1/1)	infinite(1/1)	infinite(1/1)	0.8407	infinite(1/1)	0.3498	1.2061	0.12	0	N	N	N	N	N	0
TP03_0198	hypothetical protein		0.4069	0.3024	0.4069	0.2019	0.2039	0.4069	0.8236	0.1264	0.6961	0.39	0	N	N	N	N	Y	1
TP03_0211	hypothetical protein, conserved	PiroF0002038	0	0	0.9681	0	0.773	0.3781	0.6435	0.5437	1.5721	0.39	0	N	N	N	N	N	0
TP03_0213	hypothetical protein, conserved	PiroF0100056	0.8718	0.3592	0.3027	infinite(4/4)	0.2457	0.4395	0.4685	0.2013	0.5398	0.38	0	N	N	N	N	Y	0
TP03_0303	hypothetical protein		1.2516	1.2516	0.2647	infinite(3/3)	0.8254	infinite(2/2)	0.8254	0.2003	0.813	0.63	157	N	N	N	N	N	0
TP03_0307	hypothetical protein		0	0	0	0	infinite(4/4)	infinite(2/2)	#N/A	0.2456	0.4169	0.00	0	N	N	N	N	N	1
TP03_0314	hypothetical protein	PiroF0002626	0.8001	0.637	0.2433	0.5286	0.2433	0.2791	0.2427	0.181	0.1849	0.42	178	N	N	N	N	N	0
TP03_0320	hypothetical protein	PiroF0003274	0.4993	0.2864	0.3844	1.057	0.5111	0.2749	0	0.0842	0.164	0.43	249	N	N	N	N	Y	0
TP03_0347	hypothetical protein	PiroF0001728	0.39	0.4344	2.4422	0.7283	1.2896	2.4422	0.1381	0.6034	0.1607	1.12	139	N	N	N	N	Y	0
TP03_0389	hypothetical protein	PiroF0003594	0.3351	0.3351	0.3351	0.3351	infinite(2/2)	1.0121	0.3351	0.5594	0.2481	0.38	0	N	N	N	N	Y	0
TP03_0390	hypothetical protein	PiroF0003151	0	0	0	0	1.8948	1.8948	0	0.529	0.2229	0.54	0	N	N	N	N	N	0
TP03_0391	hypothetical protein	PiroF0000648	0	0	0	0	1.2015	1.2015	0	0.0645	0	0.34	45	N	N	N	N	N	0
TP03_0393	hypothetical protein	PiroF0002422	infinite(2/2)	0.7999	0.1603	0.7999	0.0324	0.1097	0.111	0.1287	0.0859	0.29	87	N	N	N	N	Y	0
TP03_0399	hypothetical protein	PiroF0002884	0	0	1.5625	0	0.3331	0	0	0.4478	0.7813	0.27	0	N	N	N	N	Y	0
TP03_0423	hypothetical protein	PiroF0000227	0.8625	0.4328	0.8625	0.4863	0	0	0.4863	0.2865	0.2897	0.45	0	N	N	N	N	Y	0
TP03_0467	hypothetical protein, conserved	PiroF0000690	0	0	2.358	0	0.3189	0.4208	infinite(1/1)	0.2199	0.134	0.44	262	N	N	N	N	Y	0
TP03_0468	hypothetical protein	PiroF0002430	infinite(4/4)	0.8602	1.4192	0.8865	0.4705	0.1116	0	0.436	#N/A	0.54	154	N	N	N	N	Y	0
TP03_0471	hypothetical protein	PiroF0000198	0.8406	0.2922	1.2497	infinite(2/2)	1.0475	2.5007	infinite(3/3)	2.2518	0.752	0.85	631	N	N	N	N	Y	0
TP03_0482	hypothetical protein	PiroF0000026	infinite(2/2)	infinite(1/1)	infinite(5/5)	infinite(1/1)	infinite(4/4)	infinite(2/2)	infinite(7/7)	1.154	0.9285	0.00	0	N	N	N	N	N	1
TP03_0485	hypothetical protein	PiroF0100056	0.3738	1.1233	0.3	1.1233	0.0733	0.3566	1.1233	0.335	0.4226	0.64	0	N	N	N	N	N	0
TP03_0517	hypothetical protein	PiroF0001739	0.2452	0.4214	0.1606	0.3845	1.1287	0	0.1606	0.3703	0.6059	0.36	84	N	N	N	N	Y	0
TP03_0525	hypothetical protein	PiroF0003142	#N/A	#N/A	0	0.775	0.7789	0	0	0.3541	#N/A	0.22	0	N	N	N	N	Y	0
TP03_0562	hypothetical protein (Tpr family)	PiroF0100022	0	0	0	0	0	0.8538	1.2309	0.613	0	0.30	21	N	N	N	N	Y	11
TP03_0585	hypothetical protein, conserved	PiroF0001254	infinite(4/4)	infinite(4/4)	0	infinite(3/3)	0	infinite(3/3)	infinite(3/3)	#N/A	infinite(2/2)	0.00	376	N	N	N	N	N	0
TP03_0590	hypothetical protein		0.587	0.6593	infinite(1/1)	0.7784	0	0.5275	0.5881	0.5239	0.4898	0.45	4	N	N	N	N	Y	1
TP03_0595	hypothetical protein	PiroF0001248	0.5665	0.8527	1	0.6777	1.1818	1.1597	1.7917	0.8242	1.4618	1.03	0	N	N	N	N	N	0
TP03_0597	hypothetical protein	PiroF0002642	0.8423	1.0026	0.5137	1.0019	0.6791	0.5463	0.6598	0.48	0.3403	0.75	6	N	N	N	N	Y	0
TP03_0598	hypothetical protein	PiroF0001264	0.6084	0.9739	0	0.9391	1.7115	0	0	1.1055	1.1626	0.60	0	N	N	N	N	Y	0
TP03_0605	hypothetical protein	PiroF0001268	0.7331	0.4886	0	0.3196	1.0228	0	0	infinite(4/4)	0.8953	0.37	235	N	N	N	N	N	0
TP03_0607	hypothetical protein	PiroF0002645	0.3871	0.543	0	0.7415	0.5004	0	0	0.4794	0.3286	0.31	202	N	N	N	N	Y	0
TP03_0611	hypothetical protein	PiroF0002129	#N/A	#N/A	#N/A	#N/A	infinite(4/4)	#N/A	0	1.9102	0.5407	0.00	79	N	N	N	N	N	1
TP03_0615	hypothetical protein	PiroF0100022	0	0	1.4934	0	0	0.4913	0.9871	0	0	0.42	7	N	N	N	N	Y	5
TP03_0616	hypothetical protein	PiroF0100022	0	0	0.3256	0	0	0.7603	0.3269	0	0	0.20	7	N	N	N	N	Y	5
TP03_0619	hypothetical protein	PiroF0001263	0.2058	0.0661	0	0.1718	1.107	0.2137	0.129	0.1138	0.2541	0.27	85	N	N	N	N	N	0
TP03_0630	hypothetical protein	PiroF0001951	1.3351	1.3351	0.398	2.2511	0	0	0	0.2839	0.4391	0.76	112	N	N	N	N	N	0
TP03_0648	hypothetical protein	PiroF0003417	0.8161	1.6789	infinite(3/3)	1.2538	0	infinite(3/3)	0	0.3992	infinite(1/1)	0.54	189	N	N	N	N	N	0
TP03_0652	hypothetical protein	PiroF0002923	0.419	0.3097	0.3229	0.3713	0.6892	0.3803	0.4682	0.2447	0.2956	0.42	0	N	N	N	N	N	0
TP03_0658	hypothetical protein	PiroF0002647	1.7336	1.3227	1.3259	0.632	0.7554	1.1263	0.5184	0.3122	0.2614	1.06	0	N	N	N	N	Y	0
TP03_0660	hypothetical protein	PiroF0100056	0.4769	0.6422	0.7108	0.5218	0.9764	0.6275	0.6102	0.3988	0.4159	0.65	0	N	N	N	N	N	0
TP03_0700	hypothetical protein	PiroF0002656	0	0	0	0	0.7211	0	0	0.1189	0.0874	0.10	0	N	N	N	N	N	0
TP03_0708	hypothetical protein	PiroF0002943	infinite(4/4)	infinite(4/4)	infinite(3/3)	infinite(3/3)	0	0	infinite(4/4)	infinite(3/3)	infinite(1/1)	0.00	17	N	N	N	N	Y	0
TP03_0719	hypothetical protein	PiroF0000103	0	0	1.108	0	0	0	infinite(5/5)	0.4733	0.372	0.16	19	N	N	N	N	Y	1
TP03_0767	hypothetical protein	PiroF0002237	0	0	0.7108	0	0.8676	0.7273	0.5451	0.6203	0.142	0.41	0	N	N	N	N	Y	0
TP03_0783	hypothetical protein		0	infinite(1/1)	infinite(4/4)	infinite(1/1)	infinite(3/3)	infinite(2/2)	0.6042	0.2485	0.8952	0.09	0	N	N	N	N	Y	0
TP03_0793	cytidine deaminase	PiroF0001333	0.1574	0.1574	infinite(4/4)	0.0911	infinite(2/2)	0	1.3024	0.2762	0.2137	0.24	7	N	N	N	N	Y	0
TP03_0812	hypothetical protein		0.9042	1.181	1.181	1.181	0	0	0.7803	1.5898	2.787	0.75	0	N	N	N	N	N	0
TP03_0814	hypothetical protein		0	0	0	0	0.9692	0	0	0.3179	0.2381	0.14	0	N	N	N	N	N	0
TP03_0829	hypothetical protein	PiroF0003426	0.4363	0.3626	infinite(4/4)	0.4363	0.5461	infinite(1/1)	0	0.0839	0.1159	0.25	5	N	N	N	N	Y	0
TP03_0844	hypothetical protein		infinite(1/1)	infinite(1/1)	0.2004	infinite(1/1)	infinite(1/1)	1.4101	1.4101	0.1125	0.4516	0.43	400	N	N	N	N	N	1
TP03_0848	hypothetical protein	PiroF0001930	0.324	0.324	0	0.3462	0.4197	0.2969	0.9809	0.1146	0.1935	0.38	42	N	N	N	N	Y	0
TP03_0849	hypothetical protein	PiroF0002945	0.493	0.4827	0	1.1693	0.8773	0	0.262	0.4096	0.3811	0.47	261	N	N	N	N	N	1
TP03_0851	hypothetical protein	PiroF0002009	1.2107	infinite(3/3)	0	1.1086	0.6503	1.5032	0	0.9552	0.2049	0.64	0	N	N	N	N	N	0
TP03_0877	hypothetical protein (SVSP)	PiroF0100037	infinite(6/6)	1.5152	0.681	infinite(3/3)	0	1.5158	3.0307	2.6249	2.7456	0.96	24	N	N	N	N	Y	0
TP04_0056	hypothetical protein, conserved	PiroF0002720	0	0	0	0	0.8974	0.2895	0	0.7412	infinite(1/1)	0.17	15	N	N	N	N	N	1
TP04_0060	hypothetical protein	PiroF0002345	0	0	0	0	0	1.6451	0	0.0954	0	0.24	0	N	N	N	N	Y	0
TP04_0073	hypothetical protein	PiroF0003215	0	0	0.8378	0	0.8376	0	0.4064	0.4854	0.5826	0.30	73	N	N	N	N	Y	0
TP04_0080	hypothetical protein	PiroF0002628	0.433	0	0.3318	0.3341	0.8411	0.3318	0.433	0.1875	0.1429	0.39	30	N	N	N	N	Y	0
TP04_0081	hypothetical protein	PiroF0003100	0.2492	0	0.3711	0.2555	0.5457	0.7448	0.2157	0.1807	0.2411	0.34	155	N	N	N	N	Y	7
TP04_0086	hypothetical protein	PiroF0000026	0	0	0.8547	0	0.7115	0.423	0	0.3174	0.14	0.28	0	N	N	N	N	N	0
TP04_0127	hypothetical protein	PiroF0000198	0	0	0.39	0	1.4807	0.4438	0	0.6622	0.1999	0.33	86	N	N	N	N	N	0
TP04_0135	hypothetical protein, conserved	PiroF0000072	0	0	0.1874	0	0.0962	0	0.751	0	0.048	0.15	4	N	N	N	N	N	0
TP04_0191	hypothetical protein	PiroF0100042	0	0.4645	0	0	0	0.7395	0	0.4829	0.7833	0.17	8	N	N	N	N	Y	0
TP04_0197	hypothetical protein	PiroF0000168	0	0.6859	0	0	0.6859	0.6859	0.6859	0.6031	0.3352	0.39	0	N	N	N	N	N	0
TP04_0246	hypothetical protein	PiroF0100056	1.2139	0.1749	0.5227	0.9217	1.0567	0.9184	0.7134	0.1787	0.9779	0.79	0	N	N	N	N	N	0
TP04_0252	hypothetical protein	PiroF0000198	0	5.9252	0	0	0	1.4277	0	0.3682	0.2875	1.05	10	N	N	N	N	Y	0
TP04_0280	hypothetical protein	PiroF0002176	0	0	0.556	0	0	0.4118	0.8503	0.3147	0.2642	0.26	38	N	N	N	N	Y	1
TP04_0301	hypothetical protein, conserved	PiroF0002178	0	0.6012	0.3942	0	0	0	0.7916	0.095	0.264	0.26	0	N	N	N	N	Y	1
TP04_0412	hypothetical protein	PiroF0003308	0.426	1.2829	0.4254	1.2829	0	0.4254	0.0668	0.1666	0.309	0.56	33	N	N	N	N	Y	0
TP04_0420	hypothetical protein	PiroF0003092	2.4157	1.0159	0.0768	1.2077	0.6117	0	1.2077	0.2918	0.2718	0.93	0	N	N	N	N	Y	1
TP04_0445	hypothetical protein		1.4134	1.0086	1.2103	1.0086	1.0086	1.2101	1.0086	0.2994	0.6994	1.12	0	N	N	N	N	Y	0
TP04_0483	hypothetical protein	PiroF0001695	0	0	0	0	0	0	1.9656	0.2811	0.3022	0.28	9	N	N	N	N	Y	0
TP04_0589	hypothetical protein	PiroF0002564	0.4238	0.8616	0.1902	0.4238	0.1902	0.878	0.878	infinite(2/2)	0.4238	0.55	45	N	N	N	N	N	0
TP04_0591	hypothetical protein		1.23	2.7487	0	3.4509	0	0	infinite(1/1)	1.6072	1.0423	1.06	0	N	N	N	N	Y	1
TP04_0643	hypothetical protein	PiroF0002551	1.6149	1.6149	1.2105	1.2082	1.2105	0.267	1.6149	0.8051	0.8051	1.25	209	N	N	N	N	Y	0
TP04_0644	hypothetical protein		2.1118	1.8293	#N/A	1.2021	infinite(3/3)	0.6173	2.3314	1.1536	1.3847	1.16	4	N	N	N	N	Y	0
TP04_0654	hypothetical protein	PiroF0100023	0.3624	0.5429	0.3957	0.3624	1.3793	0.3624	0.4352	0.2903	0.5588	0.55	0	N	N	N	N	Y	6
TP04_0671	hypothetical protein	PiroF0003063	0.8437	0.8437	0.8437	0	0	0.8437	0.7768	0.7162	0.3875	0.59	93	N	N	N	N	Y	2
TP04_0684	hypothetical protein		2.2574	0	2.2574	2.2574	0	2.2574	0	0	0	1.29	0	N	N	N	N	N	0
TP04_0685	hypothetical protein	PiroF0000145	0.657	0.5223	0.657	0.51	0.7799	0.657	0.657	0.1256	0.161	0.63	0	N	N	N	N	Y	0
TP04_0701	hypothetical protein	PiroF0003064	infinite(3/3)	infinite(4/4)	0.6681	infinite(3/3)	0	infinite(3/3)	infinite(1/1)	0.8539	infinite(3/3)	0.10	0	N	N	N	N	Y	0
TP04_0720	hypothetical protein	PiroF0000980	0	0	0.5047	0	0	0.9517	0	0.2791	0.4158	0.21	52	N	N	N	N	Y	0
TP04_0756	hypothetical protein	PiroF0000960	0.4133	0.3699	0.414	0.3536	1.4072	0.3757	0.2859	0.2546	0.1714	0.52	143	N	N	N	N	Y	0
TP04_0839	hypothetical protein	PiroF0003517	0.5124	0.5124	0.5124	0.4852	0.7774	0.4565	0.5124	0.5164	0.2079	0.54	0	N	N	N	N	N	0
TP04_0842	hypothetical protein	PiroF0003518	#N/A	0.7076	#N/A	#N/A	infinite(2/2)	#N/A	0.8952	#N/A	0.7087	0.23	0	N	N	N	N	N	0
TP04_0879	mannose-6-phosphate isomerase, putative	PiroF0000893	0.7481	0.7481	0.7481	0.7481	0.7362	0.7481	0.2252	0.2523	0.211	0.67	16	N	N	N	N	Y	0
TP04_0905	hypothetical protein	PiroF0003337	0.8428	0.8351	0.6363	0.8335	0.7569	0.9086	1.8194	0.4177	0.6288	0.95	4	N	N	N	N	Y	0
TP04_0907	hypothetical protein	PiroF0100023	0.4468	0.3178	0.7076	0.3447	0.3293	0.2788	0.3478	0.3146	0.3004	0.40	152	N	N	N	N	Y	18

*ortholog group The ortholog group number was shown (Hayashida et al., 2012)[44]
*mpss: The massively parallel signature sequencing of the *T. parva* schizont stage was shown (Bishop et al., 2005)
*#N/A: stop codon mutation

Prediction algorithm: Secretion signal sequence was predicted by SignalP 4.0 server (Petersen et al., 2011)[30].
TMHMM Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) was used to predict transmembrane domains.
GPI-SOM (Fankhauser and Mäser, 2005) was used to identify potential GPI-anchor proteins.
Myristoylator (Bologna et al., 2004) and NMT-The MYR Predictor (mendel.imp.ac.at/myristate/) were used to predict myristoylation.
CSS-Palm 2.0 (Ren et al., 2008) was used to predict palmitoylation sites.

References
Bishop, R., Shah, T., Pelle, R. et al., 2005, Analysis of the transcriptome of the protozoan *Theileria parva* using MPSS reveals that the majority of genes are transcriptionally active in the schizont stage. *Nucleic Acids Res*. **25**. 5503-5511.
Fankhauser, N. and Mäser, P. 2005. Identification of GPI anchor attachment signals by a Kohonen self-organizing map. *Bioinformatics* . **1**. 1846-1852.
Bologna, G., Yvon, C., Duvaud, S. and Veuthey, A. L. 2004, N-Terminal myristoylation predictions by ensembles of neural networks. Proteomics, **4**, 1626-1632.
Ren, J., Wen, L., Gao, X., Jin, C., Xue Y., and Yao, X. 2008. CSS-Palm 2.0: an updated software for palmitoylation sites prediction. *Protein Eng. Des. Sel.* **21**. 639-644

Suppremental Table S2: dNdS of Tp1-Tp9

	T. parva ID	*T. annulata* ID	discription	secretion signal	Chitongo	kateteB2	kiambu	Mandali	Uganda	Nyakizu	Katumba	LAWR	Z5E5
Tp1	TP03_0849	TA17450	hypothetical	Yes	0.493	0.483	0	1.169	0.877	0	0.262	0.41	0.381
Tp2	TP01_0056	TA19865	hypothetical (surface protein D_ T. sp China)	Yes	0.2692	0	infinite(2/2)	0	0	infinite(2/2)	infinite(2/2)	infinite(1/1)	0.2683
Tp4	TP03_0210	TA03370	T-complex protein 1 subunit eta	No	0.086	0.1151	0.0367	0	0.0453	0.0175	0.0347	0.0272	0.0263
Tp5	TP02_0767	TA14970	translation initiation factor eIF-1A	No	0	0	0	0	0	0	0	0	0
Tp7	TP02_0244	TA12105	hsp90	No	0.0301	0	0	0.0352	0.0351	0.0301	0.0677	0.0165	0.0277
Tp8	TP02_0140	TA11565	cysteine proteinase	Yes	0	0	infinite(1/1)	0	0	0	infinite(1/1)	0.0975	0.1975
Tp9	TP02_0895	TA15705	hypothetical	Yes	0.2576	0.2066	0	0.256	0.0646	0.1611	0.2491	infinite(3/3)	0