

Research article

Open Access

Properties of non-coding DNA and identification of putative *cis*-regulatory elements in *Theileria parva*

Xiang Guo*^{1,2} and Joana C Silva*^{1,3,4}

Address: ¹The Institute for Genomic Research/J. Craig Venter Institute, 9712 Medical Center Drive, Rockville, MD 20850, USA, ²Advanced Biomedical Computing Center, NCI-Frederick/SAIC-Frederick Inc., Frederick, MD 21702, USA, ³Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA and ⁴Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD 21201, USA

Email: Xiang Guo* - guo424@gmail.com; Joana C Silva* - jcsilva@som.umaryland.edu

* Corresponding authors

Published: 3 December 2008

Received: 4 July 2008

BMC Genomics 2008, 9:582 doi:10.1186/1471-2164-9-582

Accepted: 3 December 2008

This article is available from: <http://www.biomedcentral.com/1471-2164/9/582>

© 2008 Guo and Silva; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Parasites in the genus *Theileria* cause lymphoproliferative diseases in cattle, resulting in enormous socio-economic losses. The availability of the genome sequences and annotation for *T. parva* and *T. annulata* has facilitated the study of parasite biology and their relationship with host cell transformation and tropism. However, the mechanism of transcriptional regulation in this genus, which may be key to understanding fundamental aspects of its parasitology, remains poorly understood. In this study, we analyze the evolution of non-coding sequences in the *Theileria* genome and identify conserved sequence elements that may be involved in gene regulation of these parasitic species.

Results: Intergenic regions and introns in *Theileria* are short, and their length distributions are considerably right-skewed. Intergenic regions flanked by genes in 5'-5' orientation tend to be longer and slightly more AT-rich than those flanked by two stop codons; intergenic regions flanked by genes in 3'-5' orientation have intermediate values of length and AT composition. Intron position is negatively correlated with intron length, and positively correlated with GC content. Using stringent criteria, we identified a set of high-quality orthologous non-coding sequences between *T. parva* and *T. annulata*, and determined the distribution of selective constraints across regions, which are shown to be higher close to translation start sites. A positive correlation between constraint and length in both intergenic regions and introns suggests a tight control over length expansion of non-coding regions. Genome-wide searches for functional elements revealed several conserved motifs in intergenic regions of *Theileria* genomes. Two such motifs are preferentially located within the first 60 base pairs upstream of transcription start sites in *T. parva*, are preferentially associated with specific protein functional categories, and have significant similarity to known regulatory motifs in other species. These results suggest that these two motifs are likely to represent transcription factor binding sites in *Theileria*.

Conclusion: *Theileria* genomes are highly compact, with selection seemingly favoring short introns and intergenic regions. Three over-represented sequence motifs were independently identified in intergenic regions of both *Theileria* species, and the evidence suggests that at least two of them play a role in transcriptional control in *T. parva*. These are prime candidates for experimental validation of transcription factor binding sites in this single-celled eukaryotic parasite. Sequences similar to two of these *Theileria* motifs are conserved in *Plasmodium* hinting at the possibility of common regulatory machinery across the phylum *Apicomplexa*.

Background

Species in the genus *Theileria* are the causative agents of lymphoproliferative diseases in cattle. The high mortality and morbidity in cattle associated with these diseases cause devastating socio-economic losses [1]. The species that cause the largest economic burden are *T. parva*, which causes East Coast fever, and *T. annulata*, which causes tropical theileriosis. Both species are transmitted by ticks and have complex intracellular life-cycles [2]. Two intracellular stages follow the introduction of sporozoite stage into mammalian hosts, the intra-lymphocytic schizont stage and the intra-erythrocytic piroplasm stage. The schizont stage has a unique ability to cause infected host cells to proliferate indefinitely, resulting in a leukemia-like phenotype [3]. Stage differentiation takes place at the time of parasite transmission to and from hosts, as well as during their establishment and amplification within the host. It has been shown that the molecular regulation of stage differentiation could be targeted for the development of new disease control strategies [4]. One such target are the regulators of gene expression, since they may determine the progression through life-cycle stages. However, little is known about the extent of, and the elements involved in, transcriptional regulation in these protozoan parasites.

Theileria species are among several thousand unicellular eukaryotic parasites in the phylum *Apicomplexa*, whose members include *Plasmodium falciparum*, the most deadly causative agent of malaria, and *Toxoplasma gondii*, one of the most successful parasites, infecting more than 30% of the human population and imposing a severe threat to immuno-compromised individuals. Apicomplexans are thought to have reduced transcriptional machinery relative to multicellular eukaryotes. The highly biased nucleotide composition of the apicomplexan genomes sequenced to date hamper the detection of *bona fide* regulatory elements [5], thus making it difficult to assess the extent of that reduction. Classical eukaryotic promoter elements such as the TATA-box and the CAAT-box appear to be absent in *T. gondii* and *P. falciparum* [6]. General transcription factors such as TFIIA and most of the TATA-binding protein associated factors are not found in apicomplexan genomes [7]. Instead, non-canonical regulatory motifs are correlated with gene expression in *T. gondii* [8]. However, recent studies that take into account the genome composition and/or the timing of gene expression are uncovering novel conserved sequence motifs, leading to a considerable expansion of the repertoire of known and putative transcription regulators in this phylum [9-14]. In *T. parva*, a massively parallel signature sequencing (MPSS) study indicates that the majority of genes are transcriptionally active in the schizont stage, but their transcription may not be stringently regulated and,

instead, vary stochastically between different host cells or time points. Active regulation of transcription was observed only at certain classes of loci [15]. In *T. annulata*, mobility shift assays have identified an element upstream of the *TamS1* gene that is a site of DNA-protein interactions during differentiation to the merozoite [16].

The complex life-cycle of *Theileria* poses a serious challenge to the development of a workable laboratory system, thus making it difficult to study the regulatory mechanism and developmental biology of this organism. Recently, the genome sequences of *T. parva* and *T. annulata* were published [17,18], enabling the study of the unique biological characteristics of *Theileria* parasites using systematic and global bioinformatics approaches. The genomes of *T. parva* and *T. annulata* are ~8 Mb long, with approximately 4000 genes each, identified through both automated and manual annotation. The gene density in these two genomes is fairly high, with the annotated coding sequences comprising ~70% of the nuclear DNA, introns ~10% and the remaining 20% consisting of intergenic regions (IGRs) [17,18]. The short length of *Theileria* IGRs, which average ~400 base pairs (bp), again suggests a different mechanism of transcriptional regulation from those found in multi-cellular eukaryotes. More importantly, these compact IGRs allow us to assess the feasibility of finding regulatory sequences with an exhaustive scan of the non-coding segment of the genome, and in the absence of comprehensive gene expression data.

In the current work, we try to understand the evolutionary forces that determine the characteristic of non-coding regions in *Theileria*, such as length, nucleotide composition and degree of conservation. Selective constraint is calculated based on alignment of orthologous IGRs and introns of *T. parva* and *T. annulata*. In particular, this study aimed to test whether sequence conservation in general, and conserved sequence motifs in particular, are most common in IGRs flanked by genes in head-to-head orientation, as expected if transcriptional regulation is an important component in the regulation of gene expression in *Theileria*. We apply a *de novo* motif discovery algorithm to identify putative *cis*-regulatory elements in IGRs and compare their conservation in two species. Candidate motifs are then characterized based on their location relative to transcription start sites and the function of neighboring genes in *T. parva*, their similarity to known transcription factor binding sites, and their distribution pattern in different non-coding genomic regions. In this first comprehensive study of non-coding sequence evolution and motif discovery in *Theileria*, we have demonstrated that selection favors short introns and IGRs, and identified conserved sequence motifs whose role in gene regulation can now be tested by experimental approaches.

Results

Characterization of non-coding DNA in *T. parva* and *T. annulata*

The length and "guanine and cytosine" (GC) content of introns and IGRs are similarly distributed in *T. parva* (intron number = 10408; IGR number = 3982) and *T. annulata* (intron number = 10816; IGR number = 3738).

IGR length fits a lognormal distribution with estimated modal value of 105 bp (Figure 1). The positive skew in the length distribution suggests that selective pressure maintains IGRs small, possibly to minimize genome size. Alternatively, such a distribution could result from the sporadic accumulation of large insertions in IGRs that are gradually eroded by more frequent deletions. The distri-

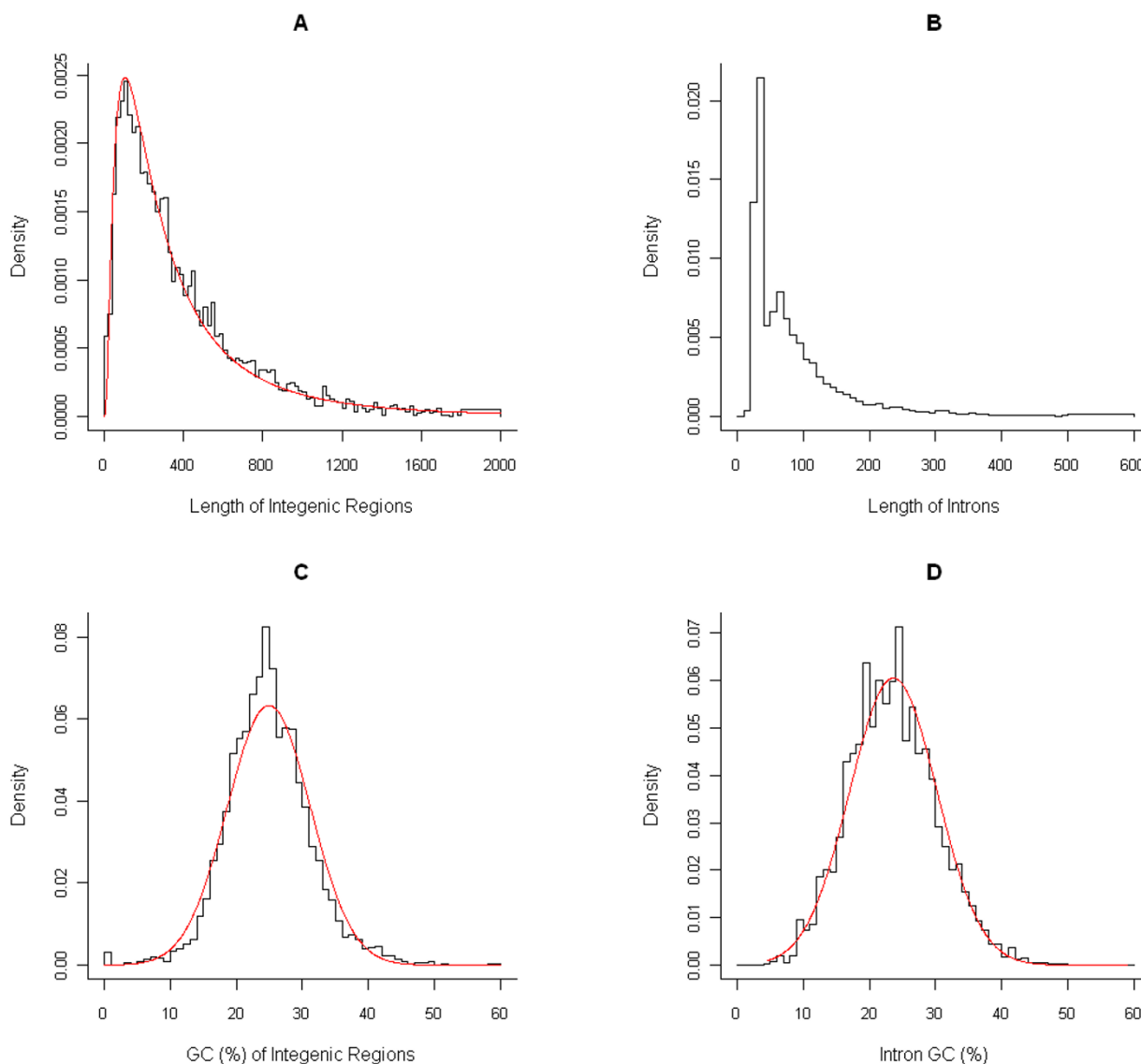


Figure 1

Distribution of length and GC content of non-coding DNA in *T. parva*. Histograms of intergenic region length (A) and intron length (B), and GC content (%) of intergenic regions (C) and introns (D). The length of intergenic regions fits a lognormal distribution (A), while GC content of both types of non-coding DNA is normally distributed (C, D). The histogram of intron length is biomodal, which probably results from the overlay of two independent distributions of small (mode = 55) and large (mode = 80) introns, much like what had been documented for other organism [19]. Similar results are found in *T. annulata* (not shown).

bution of intron length is also right-skewed, but it is bimodal, similar to what is observed in other eukaryotes [19]. This suggests the presence of two distinct classes of introns: short introns, which make up the majority of the sample, with a modal value at 55 bp, and long introns, with a modal value of 80 bp. The distribution of GC content is approximately normal for both introns and IGRs (Figure 1). The average GC content across all IGRs is $26.2 \pm 6.4\%$ and $24.0 \pm 6.5\%$ in *T. parva* and *T. annulata*, respectively, while the average in introns is slightly lower ($24.5 \pm 6.6\%$ GC in *T. parva* and $22.4 \pm 6.9\%$ GC in *T. annulata*). The GC content in non-coding regions is $\sim 10\%$ lower than the overall genome GC content in these species, which is approximately 34.1% in *T. parva* and 32.5% in *T. annulata* [17,18]. This difference suggests a mutational bias towards adenine and thymine (AT) and/or the accumulation of AT-rich regulatory motifs in non-coding regions of the genome.

In order to study the evolution of non-coding regions, homologous sequence segments between species are required. Establishing homology between non-coding regions is often difficult due to the lack of conserved landmarks that results from the accumulation of insertions, deletions and substitutions. In addition, the exact location of IGRs and introns is uncertain due to the probabilistic nature of the gene models generated through automated annotation. For this study, we identified a high-quality subset of introns and IGRs for which homology could be inferred with high probability (see Methods). The remaining analyses in this section were performed on this data subset, unless otherwise noted. The GC content in this subset of IGRs and introns (Table 1) is lower than the overall average, suggesting that coding regions that have not yet been identified are present in regions currently identified as non-coding.

IGRs can be classified into three classes based on the orientation of the genes that flank them: 5'-5', 3'-3' or 5'-3', when surrounded respectively by two start codons, two termination codons or a start and a stop codon of successive genes. The length and GC content distributions differ significantly among IGR classes ($p < 0.001$, Kruskal-Wallis test). IGR length increases with increasing number of start

codons (0, 1, or 2) flanking it (Figure 2A). This could be due to errors in determining the correct start codon, with the resulting gene models being consistently shorter than the true genes. This hypothesis seems unlikely, as it would result in IGRs enriched with coding segments, which in turn should increase GC content. In fact, the opposite trend in GC content is observed. In IGRs, GC content decreases with the number of flanking start codons, such that 5'-5' IGRs have the lowest %GC, 3'-3' IGRs have the highest %GC, and %GC of 3'-5' IGRs falls in between those of the other two classes (Figure 2B). A more likely explanation for the length trend among the three IGR classes is the preferential accumulation of functional motifs upstream of genes, causing 5'-5' IGR to be longest and 3'-3' IGRs to be the shortest. This hypothesis is supported by the significant difference in selective constraint across IGR classes (see below).

Intron length has been shown to be negatively correlated to intron ordinal number in *Arabidopsis* and in murids [20,21], a pattern that can be attributed to the uneven distribution of functional DNA elements across introns, with larger fractions present towards the 5' end of genes. A similar relationship is observed in *Theileria*, with the average length for first introns being 108 bp, and that value gradually decreasing to 94 bp, 92 bp, 87 bp and 84 bp for intron ordinal number 2, 3, 4 and ≥ 5 , respectively (Figure 3A). In addition, GC content is positively related with intron position (Figure 3B). Both intron length and GC content differ significantly among different classes of introns ($p < 0.001$, Kruskal-Wallis test).

The degree of selective constraint was estimated for each high-quality, orthologous intron pair ($n = 1487$) and IGR pair ($n = 990$), based on individual global alignments. Our results show the degree of conservation to be higher in IGRs than in introns (Table 1). Interestingly, selective constraint in IGRs of *T. parva* increases with the number of start codons that flank it (Figure 2C), and decreases with intron ordinal number (Figure 3C). The difference is statistically significant among different classes of IGRs and among introns ($p < 0.001$, Kruskal-Wallis test). These results point to the accumulation of functional motifs upstream of genes and also in introns that are closer to the

Table 1: Property of orthologous intergenic regions (IGRs) and introns

	Count	Length (bp)	GC (%)	Constraint (%)
<i>T. parva</i> IGRs	990	356.6 \pm 9.8	23.3 \pm 0.17	35.6 \pm 0.59
<i>T. annulata</i> IGRs	990	336.2 \pm 9.1	21.9 \pm 0.19	36.8 \pm 0.57
<i>T. parva</i> Introns	1487	75.6 \pm 1.83	21.9 \pm 0.16	28.4 \pm 0.52
<i>T. annulata</i> Introns	1487	63.8 \pm 1.21	21.1 \pm 0.16	30.3 \pm 0.52

Each value of length, GC content and selective constraint is the mean \pm SEM.

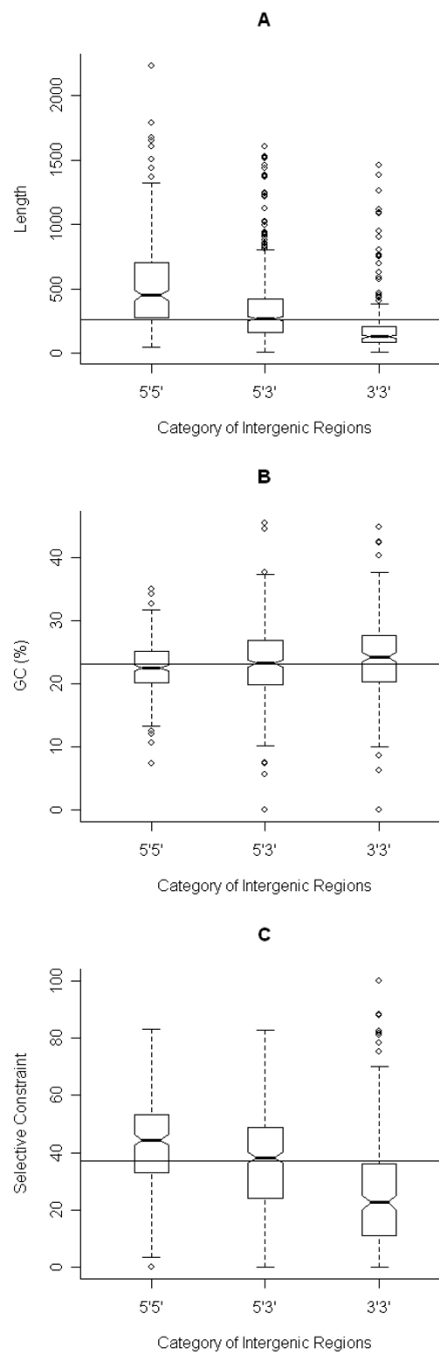


Figure 2

Length, GC content and selective constraint distributions in three classes of intergenic regions in *T. parva*. Distribution of length (A), GC content (B) and selective constraint (C) per class of intergenic region (IGR) are depicted by box-plots. The three IGR classes (5'-5', 3'-5', 3'-3') are named according to the orientation of the flanking genes. Each box (interquartile range, IQR) contains the sample's 25% to 75% range (quartiles Q1 to Q3, respectively), with the bottleneck placed at the sample median. Horizontal tick marks show the range of all elements within $Q1 - 1.5 * IQR$ and $Q3 + 1.5 * IQR$. Open circles mark data points outside this range, which are considered outliers. The width of the bottleneck (i.e the length of the V-shaped notch) is an indication of the confidence of the median; a lack of overlap of the bottleneck between samples implies that the medians are significantly different at ~95% confidence level. Similar results are found in *T. annulata* (not shown).

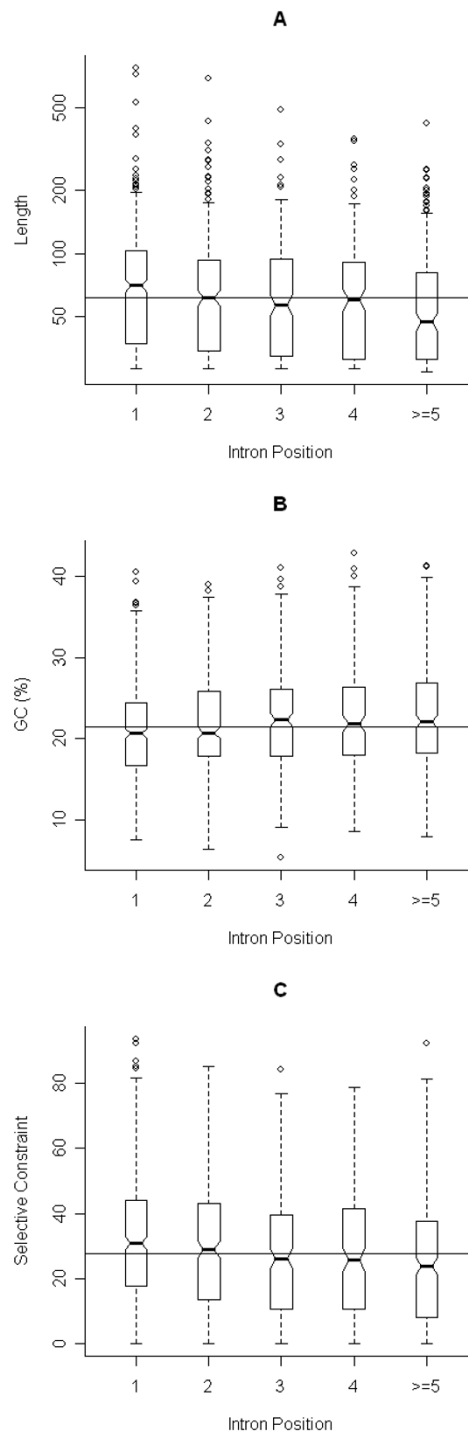


Figure 3
Length, GC content and selective constraint distributions across intron positions in *T. parva*. Distribution of length (A), GC content (B) and selective constraint (C) in introns of different ordinal numbers are depicted by boxplots. The last class averages across introns of position equal to or larger than 5. Graph description as in legend of Figure 2. In the boxplot of length distribution, a log scale is used for the vertical axis since intron lengths span a large range. Similar results are found in *T. annulata* (not shown).

5' end of genes, and are consistent with the assertion of a direct relationship between the length of non-coding regions and the frequency of functional motifs.

Conserved Sequence Motifs and their Biological Relevance

Using MEME, we mined the 300 nucleotides proximal to the 5' end of all genes in either *T. parva* or *T. annulata* for enriched motifs. Given the average IGR length of ~400 bp, this length restriction used in the initial motif search should adequately capture relevant regulatory regions, while at the same time minimizing the probability of motifs affecting neighboring genes. Table 2 lists the top five statistically significant motifs found in each species. Motifs 1, 2, 3 in *T. parva* and motifs 1, 2, 4 in *T. annulata* are found in hundreds of copies per genome in the upstream region of genes, while the frequency of other two motifs is an order of magnitude lower (Table 2). Motif comparison indicates that top three motifs in *T. parva* are almost identical to motifs 1, 2 and 4 in *T. annulata* (Figure 4). We further investigated the biological relevance of these three conserved motifs using transcription start site (TSS) and gene functional annotation information.

First, we looked for an association between putative motifs and the transcriptional process by determining whether they had a fixed position relative to the TSS [22]. Putative TSSs have been determined for 650 *T. parva* genes based on EST mapping, and on average map 138 nucleotides upstream of the start codon. *T. parva* motifs 1, 2 and 3 were present in IGRs next to, respectively, 132, 168 and 67 of the 650 genes with a defined UTR. Figure 5 shows the distribution of three motifs relative to the TSS. The distribution of motif 1 is remarkably restricted, with motifs almost exclusively located within 60 bases upstream of the TSS. Motif 3 is also enriched in this region, with more than 50% of the sequences located

within this 60 bp range. However, the distribution of motif 2 is relatively flat, with ~70% of the sequences located upstream of the TSS and the remaining 30% located between the TSS and the initiation codon (Figure 5).

Next, we investigated the biological relevance for each motif in the context of the function of adjacent genes. A hypergeometric test demonstrated that certain functional categories are significantly enriched among genes downstream of motifs 1 and 3 ($p < 0.01$), but not of motif 2. Motif 1 is associated with genes involved in protein synthesis, with telomeric ORFs, and with proteins containing signal peptides, while motif 3 is associated with genes related to protein fate (Additional file 1). Using STAMP [23], we identified known motifs that are most similar to each of the three *Theileria* motifs. Motif 1 has similarity to a DNA consensus binding site for myeloid zinc finger protein 1 (MZF1), a C2H2 zinc finger transcription factor involved in granulopoiesis, cellular proliferation and oncogenesis [24] (Figure 4). In addition, motif 1 is identical to the motif ATGGGGC, which has been identified independently in different studies in *Plasmodium* [11,14,25] and which may be preferentially associated with metabolic genes that are highly expressed during the trophozoite stage [14]. For motif 2, STAMP detected a highly similar sequence element that is known to interact with a nuclear protein in a plant species [26] (Figure 4). It is also nearly identical with a *P. falciparum* conserved motif TGTGT(G/A)(A/T) which, much like motif 2, has a widespread genome distribution [14]. Motif 3 has significant similarity to the binding site for NF- κ B, a family of transcription factors whose activation has been shown to be associated with host invasion in various pathogens, including *T. parva* [27] (Figure 4).

Table 2: Top five motifs in 5' intergenic regions of *T. parva* and *T. annulata*

	Consensus	Width	Bits	E-value	Sites
<i>T. parva</i> Motif 1	HDWTYCCCATVVR	14	16.0	5.7e-472	639
2	WAATGTGTARR	11	13.6	5.2e-414	1199
3	BDRGATTCCAY	11	16.0	8.5e-108	298
4	CCBCBRSARGGAGCY	15	24.2	3.7e-048	39
5	GCTCCCKCYWGGSG	14	26.2	1.2e-019	19
<i>T. annulata</i> Motif 1	WTTCCCATNND	12	15.5	4.1e-388	615
2	WAATGTGTA AW	11	13.3	1.1e-348	1173
3	TGTTTGTCTGATGG	15	27.2	7.1e-140	56
4	BDRGATTCCAY	11	15.8	8.8e-103	302
5	GTAATATTTCTGTAA	15	23.7	5.0e-096	62

Motifs are identified by MEME and their consensus sequences are shown here using IUPAC (International union of pure and applied chemistry) nucleic acid codes. Information content in bits is calculated based on the first-order Markov background model. The E-value estimates the number of motifs with the same width and number of occurrence that would have equal or higher likelihood in the same number of random sequences generated by the background model. Sites refer to the number of sequences that MEME used to build each motif model.

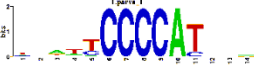

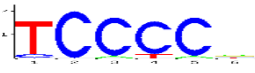






<i>T. parva</i> Motifs	Similar Motifs in <i>T. annulata</i>	Name (Source)	Best Matching Known Motifs Logo	E-value	Functional Enrichment
		MZF1_1-4 (JASPAR)		3.4e-04	Telomeric ORF Signal Peptide Protein Synthesis
		NAPINMOTIF (PLACE)		7.1e-10	/
		NF-kappaB M00051 (TRANSFAC)		4.2e-05	Protein Fate

Figure 4
Conserved motifs and their best matches in databases of known motifs. Pictogram representation of the top three MEME-derived motifs in *T. parva* and their most similar motifs in *T. annulata* (two left columns). The best matches to conserved motifs in *T. parva* among known motifs are shown (center), including name, database source, sequence logo, and STAMP E-value, which is a relative measure of similarity between two motifs based on simulated position specific score matrix models. Functional and structural annotations enriched in downstream genes of each *T. parva* motif are shown on the right.

Finally, we investigated the distribution pattern of three motifs across the entire *T. parva* genome. MAST was used to detect occurrence of each motif in coding regions (CDS), introns, and the three types of IGRs, at different probability thresholds. All three motifs were found more often in 5'-5' and 5'-3' IGRs than in the other regions (Figure 6). Together with the higher selective constraints observed in IGRs bordered by 5' ends of genes, these results are consistent with the presence of regulatory motifs in the upstream region of genes. Although the average length of CDS (1407 bp) is much longer than that of IGRs (407 bp), the frequency of the three motifs is higher in 5'-5' and 3'-5' IGRs than in CDS (Figure 6). This distribution pattern provides an additional evidence for the biological significance of our putative motifs. In addition to its presence in IGRs flanked by at least one start codon, motif 2 is also found in 3'-3' IGRs and in introns, revealing a widespread distribution throughout non-coding regions of the *T. parva* genome.

Discussion

The fraction of non-coding DNA in the majority of bacterial and archaeal genomes is between 6% and 14%, but close to 90% in multicellular eukaryotes [28]. Single-celled eukaryotes have a higher proportion of non-coding DNA than prokaryotes, but a much more compact genome than multicellular eukaryotes. The smallest known nuclear genome, that of a chlorarachniophyte nucleomorph, contains 22% intergenic DNA [29], and the single-celled model organism *Saccharomyces cerevisiae*, the budding yeast, contains 30% intergenic DNA [30]. Intron length and number varies among taxa, and dramatic difference can be seen across related species. Approximately 40% of genes contain introns in the fission yeast, *Schizosaccharomyces pombe*, while only 4% of genes

have introns in *S. cerevisiae* [30]. The non-protein coding regions of multi-cellular eukaryotes include the remnants of transposable elements that have lost functionality, ribosomal genes, motifs involved in gene regulation and chromosomal structure and possibly other unknown functions. Only a few eukaryotic unicellular parasites have so far been shown to contain transposable elements [31-34], and many of these organisms are known to lack transposable elements all together [17,35,36], which probably explains much of the observed difference in the amount of non-coding DNA between these organisms and other eukaryotes. However, non-coding regions of these small parasitic genomes remain remarkably understudied, and little is known about the forces that shape them.

The genome of *Theileria* species is highly compact. Non-coding sequences make up ~30% of the genome, as the average length of both introns and IGRs is smaller than that of many eukaryote genomes, and no transposable elements have so far been found. Our results show that a large fraction of non-coding DNA is kept constant due to purifying selection (Table 1). This high conservation rate confirms the functional importance of non-coding sequences in *Theileria*, which goes beyond a role as passive intergenic spacers. This assertion is further supported by the higher degree of sequence conservation in IGRs that border the 5' end of genes relative to what is observed in IGRs flanked by termination codons, since IGR sequence conservation between species in regions upstream of genes is associated with the presence of regulatory elements [37]. The presence of functionally important motifs at higher frequencies in 5'-5' IGRs relative to 3'-3' IGRs limits the fixation of deletion events in the former regions, which in turn remain longer.

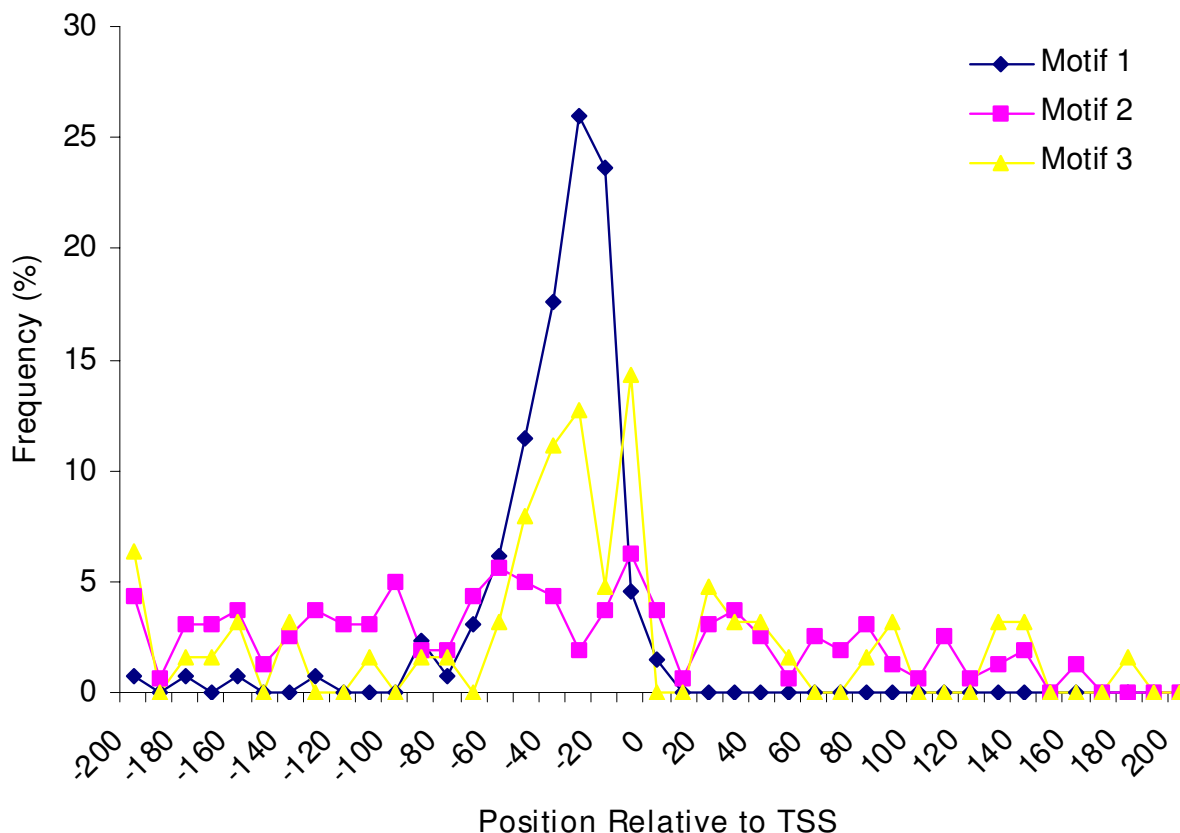


Figure 5

Distribution of conserved sequence motifs with respect to putative transcription start sites in *T. parva*. The distances between the first base of each motif and the TSS were determined for all genes for which both types of information were available. Distances are binned in 10-bp intervals. The frequency is determined based on 132 sites for motif 1, 168 sites for motif 2, and 67 sites for motif 3.

Introns play an important role in gene transcription regulation and mRNA processing, and functional elements are often found in first introns [38,39]. First introns tend to be longer and more conserved than introns of higher ordinal number, a pattern observed both in mammals [21] and in invertebrates [40]. Our analyses yielded a similar result in *Theileria*, extending this pattern to unicellular eukaryotes for the first time. These results also suggest the accumulation of functional elements towards the 5' end of *Theileria* genes.

The study of transcriptional regulation in apicomplexan parasites has identified some unique features of this large group of protists. Canonical elements seem to be absent from promoter regions, while non-canonical regulatory elements in upstream regions have been found to be involved in the regulation of gene expression. Transcriptome analysis has demonstrated active regulation of transcription in *T. parva* [15], but no regulatory elements or transcription factors have been identified so far.

In this study, we found three putative motifs that are present in hundreds of copies throughout the genome. Two of them, motifs 1 and 3 in *T. parva*, are preferentially located in the 60 nucleotides upstream of TSS, suggesting that they may be transcription factor binding sites in this species. Motif 1 appears to be enriched near telomere-associated ORFs and signal peptide-containing proteins. While the function of telomeric ORFs in *Theileria*, so called due to their extreme proximity to telomeres, remains unknown they encode hyper-polymorphic gene families [41]. In *Plasmodium*, gene families with these characteristics are known to be important to pathogenesis and antigenic variation [42,43]. In fact, a large fraction of *T. parva*'s telomeric ORFs expressed in the schizont stage contain predicted signal sequences, consistent with their involvement in host-parasite interaction [15]. The current discovery of a putative regulatory element preferentially located in upstream regions of telomeric ORFs may help the functional study and design of molecular tools to manipulate this important group of proteins in *Theileria*. Based on a MIPS classification of *T. parva* proteins [15], we

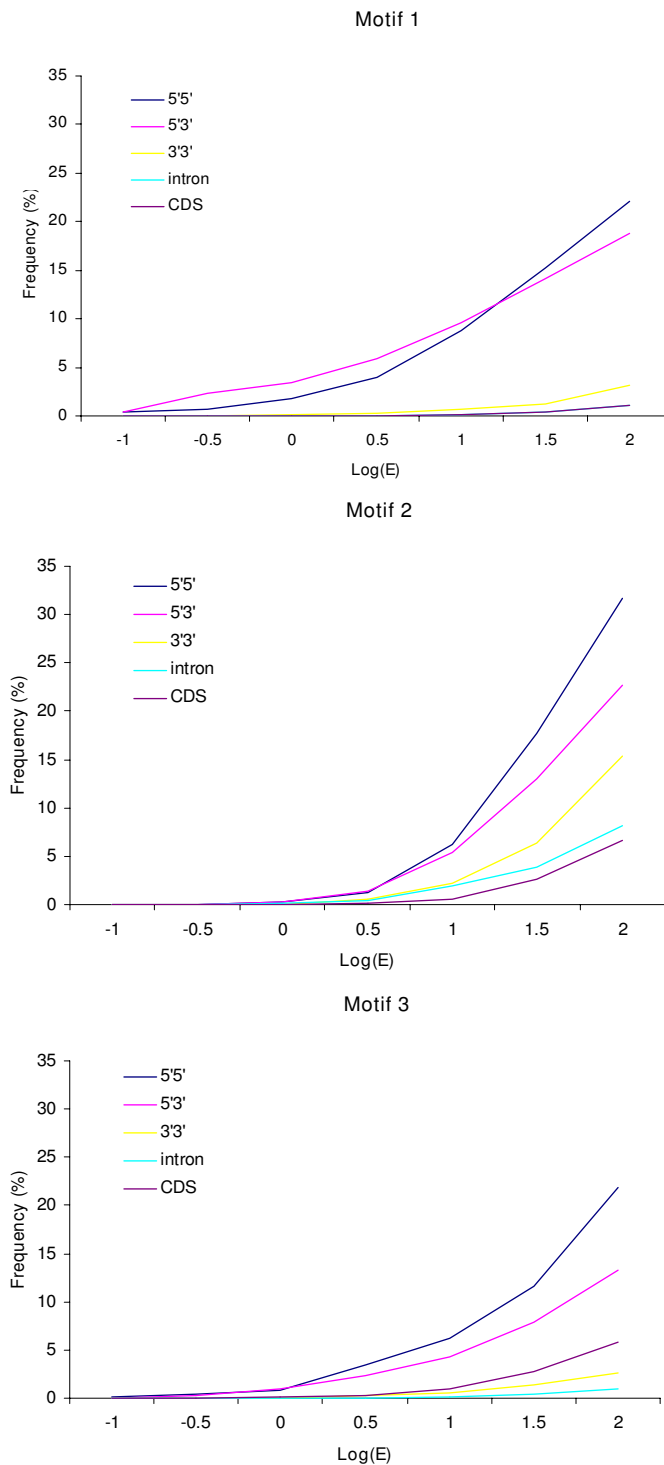


Figure 6
Distribution of three conserved motifs in different partition of the *T. parva* genome. The distribution of the three highest-scoring MEME-derived motifs in *T. parva* was determined in coding regions (CDS), 5'-5', 5'-3' and 3'-3' intergenic regions, and introns using the MAST algorithm. Relative frequency of each type of sequence with at least one occurrence of a motif is plotted as a function of the MAST E-value cutoff.

detected an association of this motif with genes involved in protein synthesis, and database searches indicated that it is similar to a DNA consensus binding site for myeloid zinc finger protein 1 (MZF1), a C2H2 zinc finger transcription factor involved in cellular proliferation and oncogenesis [24]. The finding of an identical motif in *Plasmodium* associated with genes that are highly expressed during phases of rapid cellular multiplication makes it tempting to speculate whether motif 1 may be the binding site for an unidentified transcription factor in *Theileria* associated with protein synthesis and/or cell division.

Motif 3 appears to be associated with genes that are involved in protein modification, stabilization, degradation, targeting, sorting, translocation, and other protein fate-related functions. It is similar to the binding site for NF- κ B transcription factors, which has been shown to be involved in host cell transformation mechanisms in various pathogens, including *Theileria* [27]. The schizont stage of *T. parva* induces sustained activation of NF- κ B, which regulates the expression of genes involved in immune and inflammatory responses [44], proliferation [45], and survival [46]. Pathogens may also utilize the NF- κ B system to enhance their own replication, survival, and dissemination within the host [47]. For example, NF- κ B binding sites have been found in the enhancer region of the long terminal repeat [48] and 5' UTR [49] of human immunodeficiency virus (HIV), through which host NF- κ B promotes viral replication and survival. Although no homolog to human NF- κ B has been identified in *Theileria*, it is conceivable that a transcription factor with a similar binding motif to that used by NF- κ B might exist in these species. To that effect we searched the *T. parva* protein-coding genes for peptides with homology to the IPT domain of NF- κ B, which is involved in DNA binding. Our BLASTP analysis uncovered a significant match ($E = 3.7e-5$) to a hypothetical protein, TP02-0125, which has 33% identity and 55% similarity to the DNA-binding domain of NF- κ B. A support vector machine-based algorithm predicted TP02-0125 to be a DNA-binding protein [50]. This result, together with the distribution characteristics of motif 3, suggests that attempting the experimental validation of this motif is warranted.

In contrast with the previous two motifs, motif 2 is found throughout non-coding regions, without a localized distribution relative to TSS or specific protein functional classes. However, an almost identical motif has been demonstrated to interact with a nuclear protein in developing rape seeds [18], suggesting the role of our motif as a binding site for regulatory proteins other than transcription factors. Interestingly, our motif is highly similar to a conserved motif in *P. falciparum* that also has a widespread genome distribution [14,51]. While it is possible that this sequence pattern is a characteristic feature of apicomplexan genomes, its function remains unknown.

As more *Theileria* genome sequences become available, the search for conserved motifs in non-coding sequences will have added power. A comprehensive list of conserved elements may be derived by combining phylogenetic footprinting and *de novo* pattern matching algorithms. Various experimental approaches are available for the verification of putative regulatory elements, including *in vitro* protein binding experiments, *in vivo* DNA structure assays, and reverse genetics methods. In particular electrophoretic mobility shift assays have been documented to work well in apicomplexan systems [9,11,16]. Revealing the functional potential of these conserved elements will advance the study of gene regulation in *Theileria* and possibly lead to the improved control and therapeutics for East Coast Fever and tropical theileriosis in cattle.

Conclusion

The highly compact genome of *Theileria* seems to result from selection pressure for small introns and IGRs. While, much like in other apicomplexan genomes, classical eukaryotic promoter elements have not been found in *Theileria*, genome-wide *de novo* searches identified several conserved sequence motifs in IGRs. Two putative *T. parva* motifs have localized distribution relative to transcriptional start sites and are preferentially associated with specific protein functions, which is consistent with the hypothesis that they participate in transcriptional regulation in this eukaryotic parasite. The fact that conserved motifs with similar sequence are found in *Plasmodium* hints at the possibility of common regulatory mechanisms across the phylum *Apicomplexa*.

Methods

Sequence Data

An in-house database was developed to store and analyze genome sequences and annotations of two *Theileria* species. There are 4011 annotated genes in *T. parva*, with 10408 introns. 5' and 3' UTR information is available for 650 and 545 genes, respectively. The total number of IGRs, defined as complete DNA sequences between start and/or stop codons of two consecutive annotated genes, is 3982 in *T. parva*. *T. annulata* has 3784 annotated genes, with 10816 introns and a total of 3738 IGRs. Transfer and ribosomal RNA genes were also excluded from IGRs.

Homology Identification

Orthologous clusters (OCs) were created using the publicly available gene annotations of six apicomplexan genomes, namely *T. parva*, *T. annulata*, *Babesia bovis*, *Plasmodium falciparum* (version released with PlasmoDB 5.0 Beta), *Plasmodium yoelii*, and *Cryptosporidium parvum*. Except for *P. falciparum*, the original genome annotation release was used for each species. Jaccard-filtered OC analysis [52] was used to construct the final ortholog set, and resulted in 3137 OCs containing at least one gene from each of the two *Theileria* species (Jaccard filter cutoff at

0.6). OCs with paralogs in *Theileria* were excluded, and 2904 OCs were retained that include exactly one gene from each species. OCs were also excluded if the ratio of the shortest to the longest of the two *Theileria* genes was smaller than 0.9. Genes eliminated by this criterion are likely to have an incorrect structural gene annotation in one or both species, or to contain repeats that render sequence alignment questionable. The final set of high-quality OCs contains 1956 genes, with one sequence each in *T. parva* and in *T. annulata*.

A high-quality (HQ) set of orthologous non-coding regions was then defined as follows: HQ orthologous IGRs are flanked by HQ OCs at both ends. HQ orthologous introns are those from HQ OCs with the same number of introns in both species; in addition, HQ orthologous introns are flanked by exon regions in which amino acid similarity is $\geq 75\%$ between species. This cut-off was determined empirically and resulted in the elimination of 25% of introns. The final dataset contains 990 pairs of HQ orthologous IGRs and 1487 pairs of HQ orthologous introns.

Selective Constraint

Orthologous non-coding sequences were aligned using *Owen* [53], to generate global alignments consisting of segments with significant sequence similarity ($p < 0.001$), also called hits, interspersed with segments for which sequence similarity was below that significance threshold. Selective constraint, c , in IGRs and introns was estimated according to *Shabalina* and *Kondrashov* [54]. It is defined as the fraction of invariant nucleotides within a sequence segment. Given a similarity (s) within a hit defined as the number of matches divided by the length of the shorter sequence (l_{short}), and the probability r that a site is identical between two sequences due to non-deterministic reasons, selective constraint within a hit is estimated as

$$c_{short} = \frac{s-r}{1-r} \text{ for shorter sequence and } c_{long} = \frac{l_{short}}{l_{long}} * \frac{s-r}{1-r}$$

for longer sequence. In this study, we estimated r from the similarity in third codon positions of 4-fold degenerate amino acids in HQ OCs between the two *Theileria* species; these should correspond mostly to sites that are identical due to chance or because not enough time has elapsed since the two species split for substitutions to occur. For an IGR or intron having n hits, the selective constraint is defined as the sum, over all n hits, of the total number of constrained nucleotides divided by the length L of non-coding sequence in either species, $\frac{\sum l_{hit} c_{hit}}{L}$ [55].

Motif Discovery

Although computational motif discovery approaches have traditionally been used to find over-represented patterns among co-regulated genes, they may also be applied to large sets of unrelated promoter regions [56]. In the current study, we retrieved up to 300 bases of IGRs flanking the translation start sites of all *T. parva* and *T. annulata* genes and used MEME (Multiple EM for Motif Elicitation) to search for over-represented motifs on both DNA strands. The search was done separately for each species. MEME looks for conserved ungapped blocks in a group of sequences using an iterative expectation-maximization algorithm [57]. A background model is used by MEME to calculate the log likelihood ratio and statistical significance of a motif. The model used in our search is a first-order Markov chain derived from all intergenic sequences in *T. parva* and *T. annulata* respectively. We assume the presence of either zero or one motif per sequence, with motif width between 5–15 bp. Putative motifs found in two species were compared by the Smith-Waterman local alignment method with similarity score defined by Pearson's correlation coefficient [23].

Biological Relevance of Putative Motifs

Putative motifs were further investigated in terms of their location distribution in relation to TSS, functional characterization of downstream genes, similarity to known transcription factor binding sites, and occurrence pattern in different non-coding genomic regions. These four analyses were performed in *T. parva* only. First, we retrieved all *T. parva* genes with putative TSSs information available and looked at the intersection between these genes and our set of putative motifs. Distance between the first base of a motif and TSS was calculated and binned in 10-bp intervals. Localized distribution of motifs in relation to TSS will provide evidence for the biological relevance of putative motifs.

Secondly, we investigated if any specific gene function(s) is over-represented in the region downstream of each motif. *T. parva* genes have been previously annotated based on the Munich Information Centre for Protein Sequences (MIPS) catalog and Signal P software had been used to predict the existence of signal peptides/signal anchors in expressed genes [15]. We tested functional over-representation in terms of 9 MIPS categories, telomere association and signal peptide/signal anchor information associated with the group of genes whose 5' IGRs contain a particular motif. The statistical significance is measured by a test for hypergeometric distribution (Fisher's exact test). If R genes have been annotated with an annotation category and the total number of unique genes in the genome is N , then the p -value of finding at least t genes with the annotation category among m genes with the motif of interest in their upstream regions is:

$$P(t) = \sum_{i=t}^{\min(R,m)} \frac{\binom{R}{i} \binom{N-R}{m-i}}{\binom{N}{m}}$$

As multiple functional categories were tested simultaneously, *p*-values are corrected for multiple hypotheses testing using a Bonferroni approach. The calculation was done using the function *phyper* in the statistical language R.

In the third analysis, we identified known motifs with significant similarity to our putative motifs using the web tool STAMP [23]. *T. parva* motifs were searched against a collection of databases including two comprehensive eukaryotic motif databases JASPAR [58] and TRANSFAC [59], drosophila DNase I footprint database FlyReg [60], plant motif databases PLACE [61] and AthaMap [62], prokaryotic motif databases RegTransBase [63] and DPInteract [64], as well as yeast motifs predicted by Harbison et al [65] and MacIsaac et al [66]. Motif similarity was estimated by the Smith-Waterman local alignment method with similarity score defined by Pearson's correlation coefficient. The significance is estimated based on simulated position specific scoring matrix (PSSM) models [67].

Lastly, we used MAST (Motif Alignment and Search Tool) to find all occurrences of each putative motif in *T. parva* non-coding regions [68]. The E-value of a match in an IGR or intron to a motif is based on a random sequence model derived from all non-coding sequences in *T. parva*. Various E-value cutoffs are used to retrieve matches with different statistical significance.

Authors' contributions

JS conceived the project. JS and XG devised computational approaches, XG performed the analyses. XG and JS wrote the manuscript.

Additional material

Additional file 1

Enrichment analysis of functional and structural annotations in *T. parva*. The data provided represent the analysis of functional and structural enrichment of genes downstream of each putative motif in *T. parva*. Click here for file
<http://www.biomedcentral.com/content/supplementary/1471-2164-9-582-S1.doc>

References

- Norval RAI, Perry BD, Young AS: *The epidemiology of Theileriosis in Africa* London: Academic Press; 1992.
- Dobbelaere DA, Kuenzi P: **The strategies of the Theileria parasite: a new twist in host-pathogen interactions.** *Curr Opin Immunol* 2004, **16(4)**:524-530.
- Brown CGD, Stagg DA, Purnell RE, Kanhai GK, Payne RC: **Infection and transformation of bovine lymphoid cells in vitro by infective particles of Theileria parva.** *Nature* 1973, **245(5420)**:101-103.
- Shiels B, Swan D, McKellar S, Aslam N, Dando C, Fox M, Ben-Miled L, Kinnaird J: **Directing differentiation in Theileria annulata: old methods and new possibilities for control of apicomplexan parasites.** *Int J Parasitol* 1998, **28(11)**:1659-1670.
- van Noort V, Huynen MA: **Combinatorial gene regulation in Plasmodium falciparum.** *Trends Genet* 2006, **22(2)**:73-78.
- Militello KT, Dodge M, Bethke L, Wirth DF: **Identification of regulatory elements in the Plasmodium falciparum genome.** *Mol Biochem Parasitol* 2005, **134(1)**:75-88.
- Meissner M, Soldati D: **The transcription machinery and the molecular toolbox to control gene expression in Toxoplasma gondii and other protozoan parasites.** *Microbes Infect* 2005, **7(13)**:376-1384.
- Kibe MK, Coppin A, Dendouga N, Oria G, Meurice E, Mortuaire M, Madec E, Tomavo S: **Transcriptional regulation of two stage-specifically expressed genes in the protozoan parasite Toxoplasma gondii.** *Nucleic Acid Res* 2005, **33(5)**:1722-1736.
- Behnke MS, Radke JB, Smith AT, Sullivan WJ Jr, White MW: **The transcription of bradyzoite genes in Toxoplasma gondii is controlled by autonomous promoter elements.** *Mol Microbiol* 2008, **68(6)**:1502-1518.
- Hackney JA, Ehrenkauf GM, Singh U: **Identification of putative transcriptional regulatory networks in Entamoeba histolytica using Bayesian inference.** *Nucleic Acid Res* 2007, **35(7)**:2141-2152.
- Sunil S, Chauhan VS, Malhotra P: **Distinct and stage specific nuclear factors regulate the expression of falcipains, Plasmodium falciparum cysteine proteases.** *BMC Mol Biol* 2008, **9**:47.
- De Silva EK, Gehrke AR, Olszewski K, León I, Chahal JS, Bulyk ML, Linás M: **Specific DNA-binding by apicomplexan AP2 transcription factors.** *Proc Natl Acad Sci USA* 2008, **105(24)**:8393-8398.
- Mullapudi N, Lancto CA, Abrahamsen MS, Kissinger JC: **Identification of putative cis-regulatory elements in Cryptosporidium parvum by de novo pattern finding.** *BMC Genomics* 2007, **8**:13.
- Young JA, Johnson JR, Benner C, Yan SF, Chen K, Le Roch KG, Zhou Y, Winzeler EA: **In silico discovery of transcription regulatory elements in Plasmodium falciparum.** *BMC Genomics* 2008, **9**:70.
- Bishop R, Shah T, Pelle R, Hoyle D, Pearson T, Haines L, Brass A, Hulme H, Graham SP, Taracha ELN, Kanga S, Lu C, Hass B, Wortman J, White O, Gardner MJ, Nene V, Villiers EP: **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 Acid Res* 2005, **33(17)**:5503-5511.
- Shiels B, Fox M, McKellar S, Kinnaird J, Swan D: **An upstream element of the TamSI gene is a site of DNA-protein interactions during differentiation to the merozoite in Theileria annulata.** *J Cell Sci* 2000, **113(Pt 12)**:2243-2252.
- Gardner M, Bishop R, Shah T, de Villiers EP, Carlton JM, Hall N, Ren Q, Paulsen IT, Pain A, Berriman M, Wilson RJ, Sato S, Ralph SA, Mann DJ, Xiong Z, Shallom SJ, Weidman J, Jiang L, Lynn J, Weaver B, Shoaibi A, Domingo AR, Wasawo D, Crabtree J, Wortman JR, Haas B, Angiuoli SV, Creasy TH, Lu C, Suh B, Silva JC, Utterback TR, Feldblyum TV, Perteza M, Allen J, Nierman WC, Taracha EL, Salzberg SL, White OR, Fitzhugh HA, Morzaria S, Venter JC, Fraser CM, Nene V: **Genome sequence of Theileria parva, a Bovine Pathogen that transforms lymphocytes.** *Science* 2005, **309(5731)**:134-137.
- Pain A, Renaud H, Berriman M, Murphy L, Yeats CA, Weir W, Kershonou A, Aslett M, Bishop R, Bouchier C, Cochet M, Coulson RM, Cronin A, de Villiers EP, Fraser A, Fosker N, Gardner M, Goble A, Griffiths-Jones S, Harris DE, Katzer F, Larke N, Lord A, Maser P, McKellar S, Mooney P, Morton F, Nene V, O'Neil S, Price C, Quail MA, Rabinowitsch E, Rawlings ND, Rutter S, Saunders D, Seeger K, Shah T, Squares R, Squares S, Tivey A, Walker AR, Woodward J, Dobbelaere DA, Langsley G, Rajandream MA, McKeever D, Shiels B, Tait A, Barrell B, Hall N: **Genome of the host-cell transforming parasite Theileria annulata compared with T. parva.** *Science* 2005, **309(5731)**:131-133.

19. Lim LP, Burge CB: **A computational analysis of sequence features involved in recognition of short introns.** *Proc Natl Acad Sci USA* 2001, **98(20)**:11193-11198.
20. Seoighe C, Gehring C, Hurst LD: **Gametophytic selection in *Ara-bidopsis thaliana* supports the selective model of intron length reduction.** *PLoS Genet* 2006, **1(2)**:e13.
21. Gaffney DJ, Keightley PD: **Genomic selective constraints in murid noncoding DNA.** *PLoS Genet* 2006, **2(11)**:e204.
22. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM: **Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome.** *Genome Res* 2006, **16(1)**:1-10.
23. Mahony S, Benos PV: **STAMP: a web tool for exploring DNA-binding motif similarities.** *Nucleic Acids Res* 2007, **35**:W253-258.
24. Morris JF, Hromas R, Rauscher FJ 3rd: **Characterization of the DNA-binding properties of the myeloid zinc finger protein MZF1: two independent DNA-binding domains recognize two DNA consensus sequences with a common G-rich core.** *Mol Cell Biol* 1994, **14(3)**:1786-1795.
25. Militello KT, Dodge M, Bethke L, Wirth DF: **Identification of regulatory elements in the Plasmodium falciparum genome.** *Mol Biochem Parasitol* 2004, **134(1)**:75-88.
26. Ericson ML, Muren E, Gustavsson HO, Josefsson LG, Rask L: **Analysis of the promoter region of napin genes from *Brassica napus* demonstrates binding of nuclear protein in vitro to a conserved sequence motif.** *Eur J Biochem* 1991, **197(3)**:741-746.
27. Heussler VT, Rottenberg S, Schwab R, Kuenzi P, Fernandez PC, McKellar S, Shiels B, Chen ZJ, Orth K, Wallach D, Dobbelaere DA: **Hijacking of host cell IKK signalosomes by the transforming parasite *Theileria*.** *Science* 2002, **298(5595)**:1033-1036.
28. Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, Yin J, Koonin EV: **Congruent evolution of different classes of non-coding DNA in prokaryotic genomes.** *Nucleic Acid Res* 2002, **30(19)**:4264-4271.
29. Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI: **Complete nucleotide sequence of the chlorarachniophyte nucleomorph: Nature's smallest nucleus.** *Proc Natl Acad Sci USA* 2006, **103(25)**:9566-9571.
30. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 Genes.** *Science* 1996, **274(5287)**:546-567.
31. Bhattacharya S, Bakre A, Bhattacharya A: **Mobile genetic elements in protozoan parasites.** *J Genet* 2002, **81(2)**:73-86.
32. Silva JC, Bastida F, Bidwell SL, Johnson PJ, Carlton JM: **A potentially functional mariner transposable element in the protist *Trichomonas vaginalis*.** *Mol Biol Evol* 2005, **22(1)**:126-134.
33. Pritham EJ, Putliwala T, Feschotte C: **Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses.** *Gene* 2007, **390**:1-2.
34. Souza RT, Santos MR, Lima FM, El-Sayed NM, Myler PJ, Ruiz JC, da Silveira JF: **New *Trypanosoma cruzi* repeated element that shows site specificity for insertion.** *Eukaryot Cell* 2007, **6(7)**:1228-38.
35. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shal-lom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B: **Genome sequence of the human malaria parasite *Plasmodium falciparum*.** *Nature* 2002, **419(6906)**:498-511.
36. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, Peterson JD, Pop M, Kosack DS, Shumway MF, Bidwell SL, Shallom SJ, van Aken SE, Riedmuller SB, Feldblyum TV, Cho JK, Quackenbush J, Sedegah M, Shoabi A, Cummings LM, Florens L, Yates JR, Raine JD, Sinden RE, Harris MA, Cunn-ingham DA, Preiser PR, Bergman LW, Vaidya AB, van Lin LH, Janse CJ, Waters AP, Smith HO, White OR, Salzberg SL, Venter JC, Fraser CM, Hoffman SL, Gardner MJ, Carucci DJ: **Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*.** *Nature* 2002, **419(6906)**:512-519.
37. Castillo-Davis C, Hartl DL, Achaz G: **Cis-regulatory and protein evolution in orthologous and duplicate genes.** *Genome Res* **14(8)**:1530-1536.
38. Rohrer J, Conley ME: **Transcriptional regulatory elements within the first intron of Broton's tyrosine kinase.** *Blood* 1998, **91(1)**:214-221.
39. Chan RY, Boudreau-Lariviere C, Angus LM, Mankal FA, Jasmin BJ: **An intronic enhancer containing an N-box motif is required for synapse- and tissue-specific expression of the acetylcholinesterase gene in skeletal muscle fibers.** *Proc Natl Acad Sci USA* 1999, **96(8)**:4627-4632.
40. Marais G, Nouvellet P, Keightley PD, Charlesworth B: **Intron size and exon evolution in *Drosophila*.** *Genetics* 2005, **170(1)**:481-485.
41. Bishop R, Gobrigh E, Nene V, Morzaria S, Musoke A, Sohanpal B: **Polymorphic open reading frames encoding secretory proteins are located less than 3 kilobases from *Theileria parva* telomeres.** *Mol Biochem Parasitol* 2000, **110(2)**:359-371.
42. Mancio-Silva L, Rojas-Meza AP, Vargas M, Scherf A, Hernandez-Rivas R: **Differential association of Orcl and Sir2 proteins to telomeric domains in *Plasmodium falciparum*.** *J Cell Sci* 2008, **121(Pt 12)**:2046-2053.
43. Freitas-Junior LH, Hernandez-Rivas R, Ralph SA, Montiel-Condado D, Ruvalcaba-Salazar OK, Rojas-Meza AP, Mancio-Silva L, Leal-Silvestre RJ, Gontijo AM, Shorte S, Scherf A: **Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites.** *Cell* 2005, **121(1)**:25-36.
44. Karin M, Delhase M: **The I κ B kinase (IKK) and NF- κ B: key elements of proinflammatory signaling.** *Semin Immunol* 2000, **12(1)**:85-98.
45. Joyce D, Albanese C, Steer J, Fu M, Bouzazah B, Pestell RG: **NF- κ B and cell-cycle regulation: the cyclin connection.** *Cytokine Growth Factor Rev* 2001, **12(1)**:73-90.
46. Kucharczak J, Simmons MJ, Fan Y, Gelinas C: **To be, or not to be: NF- κ B is the answer-role of Rel/NF- κ B in the regulation of apoptosis.** *Oncogene* 2003, **22(56)**:8961-8982.
47. Tato CM, Hunter CA: **Host-pathogen interactions: subversion and utilization of the NF- κ B pathway during infection.** *Infect Immun* 2002, **70(7)**:3311-3317.
48. Roulston A, Lin R, Beauparlant P, Wainberg MA, Hiscott J: **Regulation of human immunodeficiency virus type 1 and cytokine gene expression in myeloid cells by NF- κ B/Rel transcription factors.** *Microbiol Rev* 1995, **59(3)**:481-505.
49. Al-Harthi A, Roebuck KA: **Human immunodeficiency virus type-1 transcription: role of the 5'-untranslated leader region.** *Int J Mol Med* 1998, **1(5)**:875-881.
50. Kumar M, Gromiha MM, Raghava GPS: **Identification of DNA-binding proteins using support vector machines and evolutionary profiles.** *BMC Bioinformatics* 2007, **8**:463.
51. Imamura H, Persampieri JH, Chuang JH: **Sequences conserved by selection across mouse and human malaria species.** *BMC Genomics* 2007, **8**:372.
52. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu DC, Haas BJ, Tran AN, Wortman JR, Alsmark UC, Angiuoli S, Anupama A, Badger J, Bringaud F, Cadag E, Carlton JM, Cerqueira GC, Creasy T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC, Kummerfeld SK, Pereira-Leal JB, Nilsson D, Peterson J, Salzberg SL, Shallom J, Silva JC, Sundaram J, Westerberger S, White O, Melville SE, Donelson JE, Andersson B, Stuart KD, Hall N: **Comparative genomics of *Trypanosomatid* parasitic protozoa.** *Science* 2005, **309(5733)**:404-409.
53. Ogurtsov AY, Roytberg MA, Shabalina SA, Kondrashov AS: **OWEN: aligning long collinear regions of genomes.** *Bioinformatics* 2002, **18(12)**:1703-1704.
54. Shabalina SA, Kondrashov AS: **Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes.** *Genet Res* 1999, **74(1)**:23-30.
55. Shabalina SA, Ogurtsov AY, Kondrashov VA, Kondrashov AS: **Selective constraint in intergenic regions of human and mouse genomes.** *Trends Genet* 2001, **17(7)**:373-376.
56. Ohler U, Liao G, Niemann H, Rubin G: **Computational analysis of core promoters in the *Drosophila* genome.** *Genome Biol* 2002, **3(12)**:1-0087.
57. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** In *Proc Sec*

and Int Conf Intelligent Systems for Mol Bio: August 1994; Menlo Park, CA AAAI Press; 1994:28-36.

58. Sandelin A, Wasserman WW: **Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics.** *J Mol Biol* 2004, **338(2)**:207-215.
59. Matys V, Fricke E, Geffers R, Gösling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31(1)**:374-378.
60. Bergman CM, Carlson JW, Celniker SE: **Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster.** *Bioinformatics* 2005, **21(8)**:1747-1749.
61. Higo K, Ugawa Y, Iwamoto M, Korenaga T: **Plant cis-acting regulatory DNA elements (PLACE) database.** *Nucleic Acids Res* 1999, **27(1)**:297-300.
62. Galuschka C, Schindler M, Bülow L, Hehl R: **AthaMap web tools for the analysis and identification of co-regulated genes.** *Nucleic Acids Res* 2007, **35**:D857-862.
63. Kazakov AE, Cipriano MJ, Novichkov PS, Minovitsky S, Vinogradov DV, Arkin A, Mironov AA, Gelfand MS, Dubchak I: **RegTransBase—a database of regulatory sequences and interactions in a wide range of prokaryotic genomes.** *Nucleic Acids Res* 2007, **35**:D407-D412.
64. Robison K, McGuire AM, Church GM: **A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome.** *J Mol Biol* 1998, **284(2)**:241-254.
65. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431(7004)**:99-104.
66. MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E: **An improved map of conserved regulatory sites for Saccharomyces cerevisiae.** *BMC Bioinformatics* 2006, **7**:113.
67. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles.** *Nucleic Acid Res* 2004, **32**:D91-D94.
68. Bailey TL, Gribskov M: **Combining evidence using p-values: application to sequence homology searches.** *Bioinformatics* 1998, **14(1)**:48-54.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

