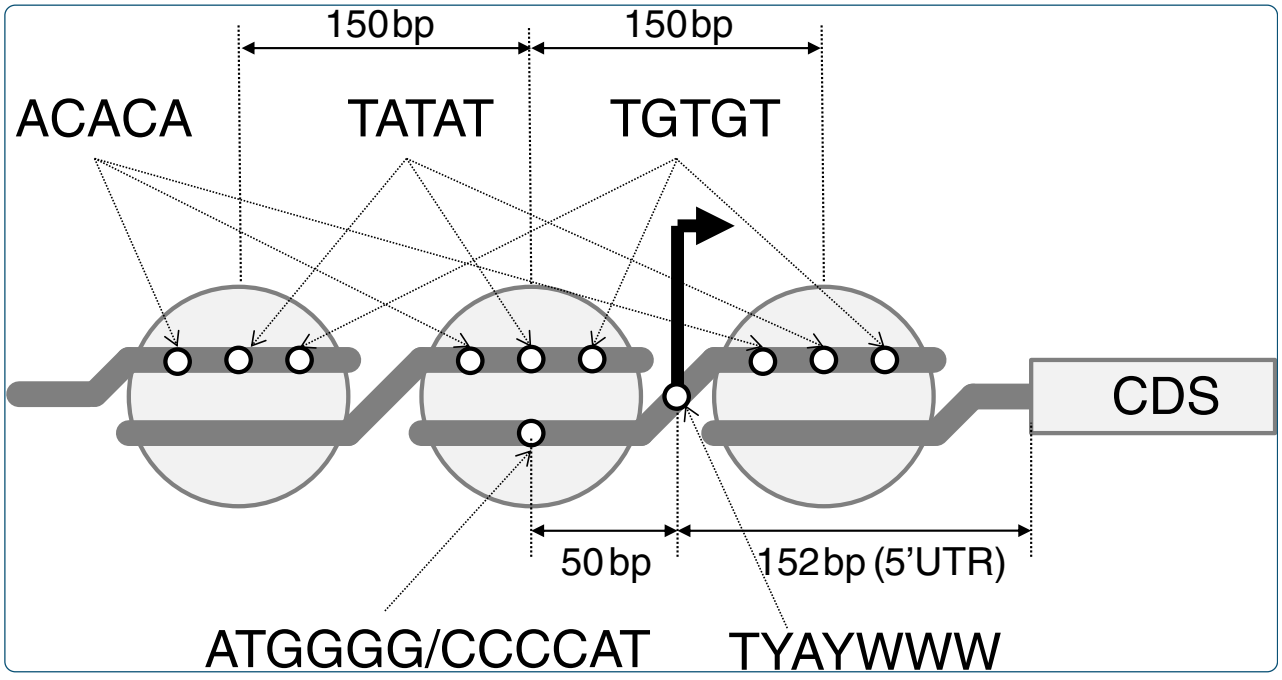
BMC Genomics



The *Babesia bovis* gene and promoter model: an update from full-length EST analysis

Yamagishi *et al.*

BioMed Central

Yamagishi *et al. BMC Genomics* 2014, **15**:678
http://www.biomedcentral.com/1471-2164/15/678

BMC Genomics

RESEARCH ARTICLE Open Access

The *Babesia bovis* gene and promoter model: an update from full-length EST analysis

Junya Yamagishi[1,2], Hiroyuki Wakaguri[3], Naoaki Yokoyama[2], Riu Yamashita[1], Yutaka Suzuki[3], Xuenan Xuan[2] and Ikuo Igarashi[2*]

Abstract

Background: *Babesia bovis* is an apicomplexan parasite that causes babesiosis in infected cattle. Genomes of pathogens contain promising information that can facilitate the development of methods for controlling infections. Although the genome of *B. bovis* is publically available, annotated gene models are not highly reliable prior to experimental validation. Therefore, we validated a preproposed gene model of *B. bovis* and extended the associated annotations on the basis of experimentally obtained full-length expressed sequence tags (ESTs).

Results: From *in vitro* cultured merozoites, 12,286 clones harboring full-length cDNAs were sequenced from both ends using the Sanger method, and 6,787 full-length cDNAs were assembled. These were then clustered, and a nonredundant referential data set of 2,115 full-length cDNA sequences was constructed. The comparison of the preproposed gene model with our data set identified 310 identical genes, 342 almost identical genes, 1,054 genes with potential structural inconsistencies, and 409 novel genes. The median length of 5′ untranslated regions (UTRs) was 152 nt. Subsequently, we identified 4,086 transcription start sites (TSSs) and 2,023 transcriptionally active regions (TARs) by examining 5′ ESTs. We identified ATGGGG and CCCCAT sites as consensus motifs in TARs that were distributed around −50 bp from TSSs. In addition, we found ACACA, TGTGT, and TATAT sites, which were distributed periodically around TSSs in cycles of approximately 150 bp. Moreover, related periodical distributions were not observed in mammalian promoter regions.

Conclusions: The observations in this study indicate the utility of integrated bioinformatics and experimental data for improving genome annotations. In particular, full-length cDNAs with one-base resolution for TSSs enabled the identification of consensus motifs in promoter sequences and demonstrated clear distributions of identified motifs. These observations allowed the illustration of a model promoter composition, which supports the differences in transcriptional regulation frameworks between apicomplexan parasites and mammals.

Keywords: Babesia bovis, Expressed sequence tags, Full-length cDNA, Transcription start sites, Cis-elements

Background

Bovine babesiosis is a parasitic infection caused by a protozoan of the genus *Babesia*, order Piroplasmida, phylum Apicomplexa. *Babesia bovis* and *Babesia bigemina* are major species that impose a considerable economic burden on cattle industries because of their wide geographical distribution and pathogenicity [1]. The clinical symptoms of *B. bovis* are more serious than those of *B. bigemina*, including fever, extensive erythrocyte lysis leading to anemia, icterus, hemoglobinuria, and death. Although antiparasitic drugs such as imidocarb successfully control these symptoms [2], they have severe side effects and may promote the emergence of resistant strains and residual chemicals. Therefore, safer chemical agents and vaccinations are required.

In general, the genome is an excellent tool for understanding all life forms. Unique genes and pathways that are elucidated from genomes are often recognized as targets for chemical or vaccine development. Because the genome sequence of *B. bovis* is publically available [3], it may offer promising information for the development of novel approaches for controlling parasitic infections. According to a previous bioinformatics study, the *B. bovis*

* Correspondence: igarcpmi@obihiro.ac.jp
[2]National Research Center for Protozoan Diseases, Obihiro University of Agriculture and Veterinary Medicine, Inada-cho west 2-13, Obihiro, Hokkaido 080-8555, Japan
Full list of author information is available at the end of the article

genome encodes 3,671 nuclear protein-coding genes. However, estimated gene models based on bioinformatics lack accuracy in nonmodel organisms. Inconsistencies in gene models have been reported between bioinformatics estimates and experimental observations of apicomplexan parasites [4,5]. Therefore, to improve reliability, gene models require verification with experimental evidence.

The acquisition of mRNA sequences is one of the most straightforward strategies for verifying gene models. Specifically, full-length cDNA libraries facilitate the identification of transcription start sites (TSSs), exon and intron structures, 5′ and 3′ untranslated regions (UTRs), and polyadenylation sites. Moreover, massive sets of TSSs can be used to identify transcriptionally active regions (TARs), which are closely related to promoter regions [6,7]. Therefore, the determination of full-length cDNA transcriptomes is critical for revisions of gene models, and for elucidation of transcriptional mechanisms.

In this study, we collected 5′ and 3′ expressed sequence tags (ESTs) from full-length cDNAs of *B. bovis* that were synthesized using the oligo-capping method [8]. In brief, cap structures at 5′ ends of mRNA were replaced with synthetic linker RNA sequences using the oligo-capping method. Subsequently, chimeric RNA was used to synthesize cDNA with fixed 5′ transcript sequences. This cDNA was then sequenced, and the data was entered into an updated gene model to identify novel genes. In addition, consensus sequences around TSSs and putative DNA cis-elements for transcriptional control were identified by comparison with promoter regions identified in genome-wide analyses.

Results and discussion

Construction and analysis of full-length cDNA

A total of 12,286 clones were randomly selected for plasmid extraction (Table 1). Subsequent one-pass sequencing from 5′ and 3′ ends using the Sanger method produced 9,573 and 10,956 sequences, respectively (DDBJ: HX874250-HX894778). After assembly of paired 5′ and 3′ ESTs using Cap3, 7,797 sequences were successfully united into one sequence, and one-pass sequences with poor quality and genes with long transcripts were excluded by miss assembly. Finally, 6,787 sequences passed the filter for coding capacity and were selected. These were annotated and redundancy was eliminated, resulting in 2,115 full-length cDNA sequences (DDBJ: AK440354–AK442468), including 1,706 cDNAs that corresponded with preproposed gene models in PiroplasmaDB, and 409 newly annotated genes (Table 1 and Additional file 1: Table S1). Among the 409 newly annotated genes, 134 showed sufficient homology to genes of other apicomplexan parasites (Additional file 1: Table S1B). In addition, features of these 134 cDNA sequences were sufficiently similar to those of the other gene sets (Additional file 2: Table S2), indicating that they may be newly identified protein coding transcripts. Among these, numbers of the genes with multiple exons and average exon numbers per gene were higher than those in other gene models (Additional file 2: Table S2), indicating that genes with multiple exons are relatively difficult to predict from genome sequences and result in miss annotation. In contrast, 273 cDNA sequences with little homology showed unique features. Specifically, the median coding sequence (CDS) length was shorter, as indicated by the smaller numbers of genes with multiple exons and longer median exon lengths than those in other gene sets (Additional file 2: Table S2). These observations suggest that certain parts of the transcripts identified in this EST analysis are noncoding RNA, or were derived from genomic DNA as artifacts. Nonetheless, promising protein coding cDNA sequences with large CDS lengths and multiple exons such as XBBk025260.contig, XBBk029358.contig, and XBBk014264.contig remained in this gene set. These

Table 1 Summary of ESTs and contigs

	Number	Accession number
Total number of isolated clones	12286	
5′ one-pass sequence	9573	DDBJ: HX874250-HX894778
3′ one-pass sequence	10956	
Contig sequence	6787	
Non-redundant contig sequence[1)]	2115	DDBJ: AK440354-AK442468
Identical[2)]	310	
Amino acid variant[3)]	342	
Structural variant[4)]	1054	
Assigned in this study[5)]	409	

1) Nonredundant contig sequences were selected from the contig sequence. Identical, amino acid, structural, and assigned variants were subsets of nonredundant contig sequences. 2) Contig sequences with identical coding sequences to the preproposed gene model (ppgm); 3) Contig sequences with almost identical coding sequence but amino acid variant(s) derived from single nucleotide variant(s); 4) Contig sequences with structural differences to that of the ppgm assigned in this study; 5) Contig sequences not described in the ppgm.

B. bovis-specific novel genes may have *B. bovis*-specific functions in proliferation and host–parasite interactions. In general, gene finding algorithms such as GlimmerHMM [9] require training data sets for better prediction. Although training data sets for model organisms have been constructed using experimental data, available *Babesia* spp. training data sets are limited, potentially reflecting the observed discrepancies between experimentally observed cDNAs and preproposed gene models. Because a degree of consistency was observed between the 1,706 full-length cDNA sequences and preproposed annotations, we performed genome and amino acid alignments of these sequences (Table 1). In these analyses, 310 sequences were identical to preproposed genes, whereas 342 were almost identical but with amino acid substitutions that probably originated from sequencing errors or polymorphisms among strains. The remaining 1,054 sequences had partial homology to existing annotations, although they had structural inconsistencies that may reflect the alternative usage of start codons and/or splicing.

The 5′ UTRs that lie between TSSs and first in-frame initiation codons are known to play crucial roles in post-transcriptional regulation by modulating translational efficiency and mRNA stability through the actions of IRES and riboswitches [10,11]. This mechanism is observed in a wide variety of organisms, including humans, plants, and yeast [12-14], suggesting that apicomplexan parasites have similar functions. However, these functions have been poorly investigated. Therefore, to elucidate the functions of the 5′ UTRs of *B. bovis*, we constructed a genome-wide 5′ UTR sequence data set using full-length cDNA sequences and demonstrated that the median length of the 5′ UTRs of *B. bovis* is 152 nts. The average 5′ UTRs are 210.2 nts in humans, 186.3 nts in rodents, 221.9 nts in invertebrates, 103.0 nts in viridiplantae, and 134.0 nts in fungi [15] and the mode length is approximately 130 nts in *Toxoplasma gondii* [16]. These lengths agree with our observations in *B. bovis*. Similarly, the median length of the 3′ UTRs of *B. bovis* is 116 nts (Additional file 2: Table S2).

Gene expression frequencies are also indicated in EST data. Therefore, we examined the 9,573 5′ ESTs data set and selected 9,546 sequences following successful mapping onto the *B. bovis* genome. To estimate expression frequencies, these were then mapped onto preproposed CDSs with novel sequences identified in this study (Additional file 3: Table S3 and Additional file 4: Figure S1). The resulting ranking was not identical to that in a previous study of ESTs [17], although it showed similar tendencies. These discrepancies may reflect differences in culture conditions and parasite strains or sampling errors associated with small data sets. Logarithmic plots of expression levels and ranks of each gene resembled the power law (Additional file 4: Figure S1) and indicated similar transcriptome distributions to those observed in previous studies [18,19].

B. bovis promoter components and typical structure

Transcription is controlled by the coordinated binding of promoter sequences by transactivators. In humans and model organisms, promoter structures have been intensively examined in a genome-wide manner [20-22] and have been shown to play pivotal roles in gene and phenotype expression. However, the promoter structure of *Babesia* spp. remains unknown. Therefore, we characterized the promoter structure of *B. bovis* using high resolution TSS information derived from a full-length cDNA data set.

Genome-wide TSS distributions were examined by mapping 5′ ends of 5′ ESTs. In briefly, 9,412 reliable 5′ end sequences of 36 nts were selected from 9,573 5′ ESTs. Of these, 7,111 were successfully mapped onto the *B. bovis* genome sequence, 4,086 locations were assigned as TSSs after considering redundantly mapped sequences, and 2,023 TARs were identified.

We selected motifs in the −10 to +10 regions of TSSs from the TAR data set and examined these using MEME [23]. The estimated consensus sequence TYAYWWW was found in 801 of the 2,023 TARs, with p values of <0.05 (Table 2 and Figure 1). We also examined the positional distribution of this motif around TSSs. Examinations of sequences around TSSs (−100 to +100 region) showed that the motif was distributed only on TSSs (Figure 2). Moreover, adenine residues at TSSs and cytosine residues at the −1 position were clearly conserved and +3 to +5 positions tended to be thymidine, as shown in *T. gondii* [7]. This CA motif was also conserved in initiator consensus sequences from vertebrates [24] and dicotyledonous plants [25], despite differences in the methods for identifying consensus and diversity of subject species. Data sets for *B. bovis* and *T. gondii* were collected from single organisms, whereas the data sets from vertebrates and plants were collected from multiple organisms. According to molecular recognition analyses, the initiators TAF1 and TAF2 play pivotal roles [26-28]. In *Plasmodium falciparum*, PFL1645w and MAL7P1.134

Table 2 FWM for *B. bovis* initiator-like motifs from 801 TSSs

Position from TSSs	−2	−1	0*	+1	+2	+3	+4
A	142	0	801	109	270	308	234
T	411	245	0	326	480	255	325
G	104	0	0	133	0	127	137
C	144	556	0	233	51	111	105
consensus	T	Y	A	Y	W	W	W

*position of TSS.

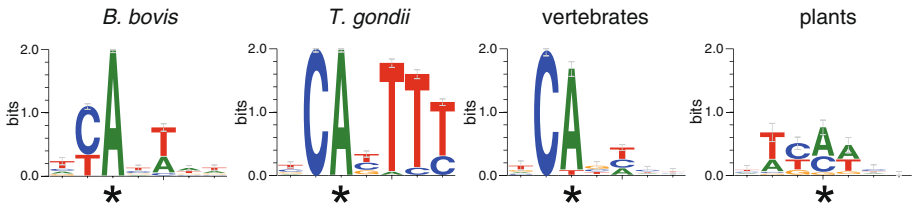


Figure 1 Consensus sequences on TSSs. Asterisks indicate TSSs.

are promising functional homologues of TAF1 and TAF2, respectively, as predicted using bioinformatics methods [29]. Moreover, their corresponding genes BBOV_IV004260 and BBOV_II003570 were annotated in the *B. bovis* genome, implying that initiator recognition and TSSs have evolved with closely related molecular mechanisms across taxonomic kingdoms.

In subsequent analyses, we identified a cis-element that is involved in transcriptional control. To generate a putative promoter set, −1000 to +1000 regions from typical TSSs of the 2,023 TARs were selected and examined using CisFinder [30], and −100–0 regions were examined using MEME [23]. These analyses showed frequent distribution of ATGGGG and ACACA within promoter regions.

To validate the ATGGGG motif, we examined positional distributions of these candidates around TSSs and found a clear peak at 50 nts upstream (Figure 3A). Further investigations of the reciprocal sequence CCCCAT showed equivalent distribution to that of ATGGGG (Figure 3B), implying that the motif may be functional regardless of its direction. The CCCCAT motif has been identified in *Theileria parva* and *Theileria annulata* using encyclopedic promoter analyses. Although the reciprocal motif ATGGGG was not examined in these species, its peak was found at −20 nts from TSSs, differing slightly from our observations [31]. In further investigations, we examined functional enrichments of genes carrying these promoter motifs, and identified genes corresponding to the 2,023 TARs by calculating relative distances. Subsequently, 1,315 TARs were found with candidate initiation codons. Among these, 222 TARs had the ATGGGG or CCCCAT motifs in the −80 to −20 region from TSSs. Subsequent enrichment analyses using gene ontology terms from GOstat [32] indicated significant enrichment in “structural constituent of ribosome” (GO:0003735) and “translation” (GO:0009058) categories, with E-values of $3.43e^{-08}$ and $2.06e^{-06}$, respectively. Enrichments of protein synthesis have also been reported for the CCCCAT motif in *T. parva* and *T. annulata* [31], suggesting that the motif may be conserved in piroplasms as a transcriptional regulator of genes involved in protein synthesis.

To validate the ACACA motif, we examined the positional distribution of these candidates and found periodical distribution around TSSs (Figure 3C). The reciprocal sequence TGTGT was also periodically distributed, but its phase was shifted (Figure 3C). Based on these observations, we examined all 5-mer repeat motifs comprising two nucleotides and found periodical distribution of TATAT as an additional motif (Figure 3C), with similar cycles but differing phases (Figure 3C). The related motifs CACAC, GTGTG, and ATATA also showed similar distributions (data not shown). The only other combinations that showed distinguishing distributions were GGGGG, CCCCC, GCGCG, and CGCGC, with peaks around −50 nts (Figure 3D). GGGGG and CCCCC motifs are closely related to ATGGGG and CCCCAT motifs, respectively. However, GCGCG and CGCGC motifs may be functional and gene ontology enrichment analyses showed frequent

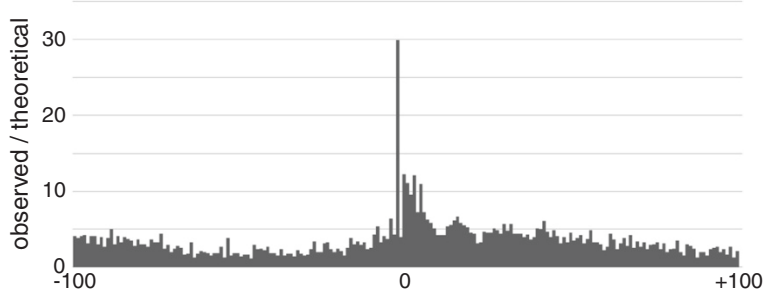


Figure 2 Distribution of the TYAYWWW motif around TSSs. The horizontal axis represents sequence areas from −100 to +100 around TSSs with 1-nt resolution. Position 0 represents TSS. A peak was observed at the −2 position. The vertical axis represents the ratio of the motif frequency and the theoretical frequency (see Methods).

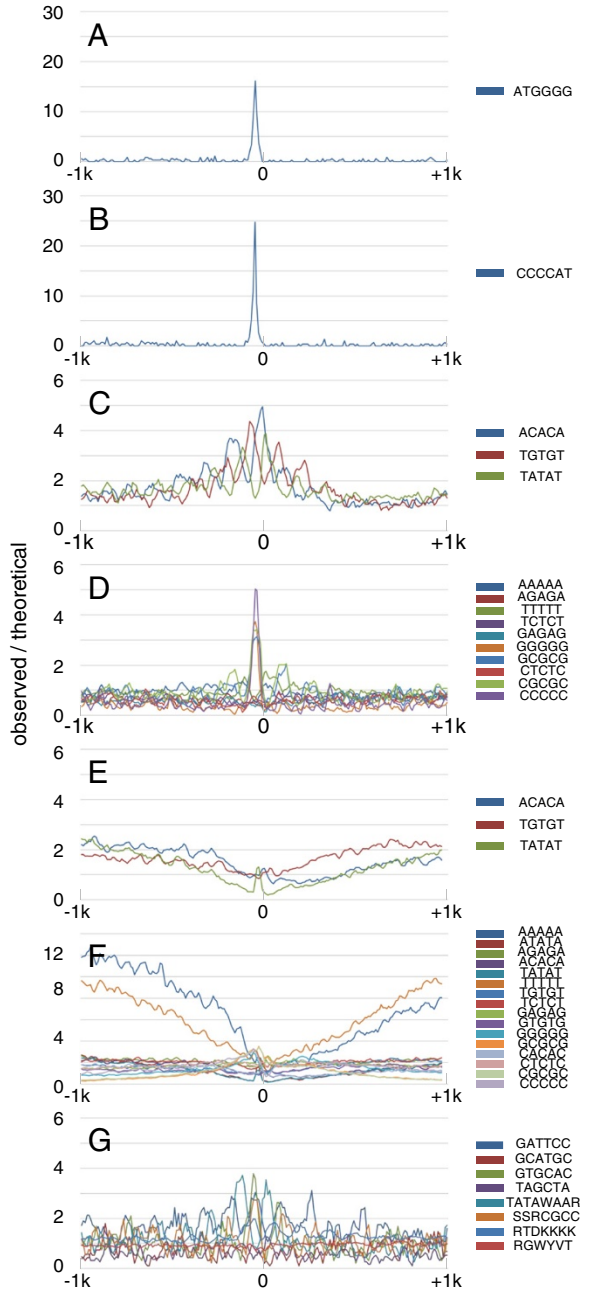


Figure 3 Distributions of motifs around TSSs. Horizontal axes represent sequence areas from −1000 to +1000 from TSSs. Vertical axes represent the ratio of motif frequencies and theoretical frequencies (see Methods). **(A)** Distributions of the ATGGGG motif; **(B)** Distribution of the CCCCAT motif; **(C)** Distributions of the ACACA (blue), TGTGT (red), and TATAT (green) motifs; **(D)** Distributions of the other 5-mer motifs; **(E)** Distributions of the ACACA (blue), TGTGT (red), and TATAT (green) motifs in humans; **(F)** Distributions of all 5-mer repeat combinations in humans; **(G)** Distributions of the known apicomplexan motifs GATTCC, GCATGC, GTGCAC, and TAGCTA and the core mammalian promoter motifs TATA, TATAWAAR; BRE^{u}, SSRCGCC; BRE^{d}, RTDKKKK; and the DPE, RGWYVT. In **A**, **B**, and **G**, the scanning window was 10 bp. In **C–F** the scanning window was 30 bp.

but insignificant presence of these in upstream promoter regions of ribonucleoprotein complex biogenesis (GO:0022613) genes (p = 0.078). To confirm the specificity of these motifs for apicomplexan parasites, we examined periodical distributions in the TSS database DBTSS, which contains precise positions of TSSs in the genomes of various organisms [33]. Promoter regions from −1000 to +1000 of human and mouse TSSs were obtained and the distribution of ACACA, TGTGT, and TATAT motifs were examined as in *B. bovis*. However, no periodical distributions were found in human (Figure 3E) or mouse (data not shown) databases, and no related periodical distributions of other combinations were observed as in *B. bovis* (Figure 3F). In contrast, the ACACA motif was reportedly observed in *T. parva* and *P. falciparum* [31,34], although periodical distributions have not been reported. Rather than reflecting the differences in species, these discrepancies may have been caused by differences in the precision of TSS identification. Nonetheless, these observations imply that the motif is common among some apicomplexan parasites, and the present periodical patterns had interval lengths of 140–150 nts. Minimum units of nucleosome repeat lengths comprise 147-bp DNA sequences around core histone octamers and 20-bp DNA linkers and are much longer than our observations. However, previous studies demonstrate that the minimum observed nucleosome repeat length is much closer to our observation of approximately 155 bp in *Schizosaccharomyces pombe* and *Aspergillus nidulans* [35-37], and *P. falciparum* [38,39]. On the other hand, these discrepancies may reflect the involvement of unconventional nucleosome structures. The conventional histone octomer comprises two H2A, H2B dimers and H3, H4 tetramers. In contrast, unconventional histones comprising variants such as H2B.Z, H2A.Z, and CENP-A have specific functions that are distinguishable from the conventional one. Crystal structure analysis of human centromeric nucleosomes containing CENP-A suggests that only 121-bp DNA fragments tightly bind to nucleosomes, unlike conventional H3 nucleosomes [40]. In *P. falciparum*, it was demonstrated that the nucleosome with H2A.Z specifically localizes to intergenic regions [41,42]. Moreover, no homologue to the linker histone H1 has been identified in apicomplexan parasites [43]. FAIRE-seq and MAINE-seq analyses in *P. falciparum* demonstrated that nucleosome binding to TSSs is associated with gene expression [44] and there are preferred DNA motifs for nucleosome assembly [45,46]. These collateral data warrant the assumption that the observed periodical patterns in this study are involved in chromatin structure and regulate gene expression via chromatin remodeling processes.

In further analyses, we applied this scanning method to known apicomplexan and mammalian core promoter

motifs. A previous study showed the distribution of GATTCC in *T. parva* and *T. annulata* at regions that are −20 nts from TSS [31]. Moreover, GCATGC was identified as a PF14_0633-binding target in *P. falciparum* [47], an AP2-Sp-binding target in *Plasmodium berghei* [48], and a Toxoplasma Ribosomal Protein (TRP)-2-binding target in *T. gondii* [49]. GTGCAC is known as a subtelomeric variant gene promoter element (SPE)-2 [49] and a binding target of PFF0200c_DLD and PfSIP2 [47,50,51]. TAGCTA is also reportedly a binding target of Pb.AP2-O [52]. Therefore, although the GTGCAC motif was moderately concentrated in the −50-nt area, the other motifs did not show distinguishing distributions in comparison with the ATGGGG/CCCCAT motif (Figure 3G). In particular, GATTCC was not specifically distributed around TSSs, as observed in *T. parva* and *T. annulata,* indicating that the motif is specific to *Theileria* spp. and may be involved in specific biological phenotypes, such as infectivity in lymphocytes. According to mammalian motifs, we examined TATA boxes, upstream TFIIB-recognition elements (BRE^u), downstream BRE (BRE^d), and downstream promoter elements (DPE), containing the consensus sequences TATAWAAR, SSRCGCC, RTDKKKK, and RGWYVT, respectively [20]. In these analyses, TATA boxes showed periodical-like patterns (Figure 3G). In contrast, TATA boxes are known to be distributed around −30 nts from TSSs (Figure 3E), and TATA box consensus sequences are closely related to TATAT and ATATA motifs. These are also periodically distributed, suggesting that the observed pattern for TATA boxes was residual and no functional motifs correspond with the TATA box in *B. bovis*. This observation also indicates that other mammalian motifs are nonfunctional (Figure 3G).

Collectively, we speculate model promoter structures and transcriptional mechanisms in *B. bovis* that explain our observations (Figure 4). Primarily, we identified the TSS initiator-like motif TYAYWWW. In other taxonomic kingdoms, this initiator works as a binding site for the general transcription factors TAF1 and TAF2 [20], and previous *in silico* analyses demonstrate that apicomplexan parasites express homologs of TAF1 and TAF2 [29]. Therefore, *B. bovis* may also use this molecular mechanism at the final step of transcriptional initiation, as described previously in *P. falciparum* [53]. Similar to *T. gondii* and majority of other organisms, the average length of 5′ UTRs was 150 nts, suggesting similar involvement in the regulation of gene expression, similar to that in other organisms. Periodical distributions of ACACA, TGTGT, and TATAT were observed around TSSs. However, this profile was not observed in human and mouse (Figure 3), and previous studies indicate that transcriptional mechanisms differ between apicomplexan parasites and other eukaryotes to a certain degree [43,53,54]. In particular, we assumed that the periodical distributions are involved in tight assembly of nucleosome structures and control transcription, although discrepancies of nucleosome repeat lengths remains to be clarified by additional experimental evidences. On the other hand, we observed clear peak distributions of ATGGGG and CCCCAT at −50-bp regions from TSSs. Although it remains unclear how this motif functions regardless of orientation, chromatin remodeling factors may be recruited to loosen nucleosome structures. Therefore, the scheme shown in Figure 4 proposes transcriptional arrest by histones and subsequent activation by putative chromatin remodeling factors that interact with ATGGGG or CCCCAT elements.

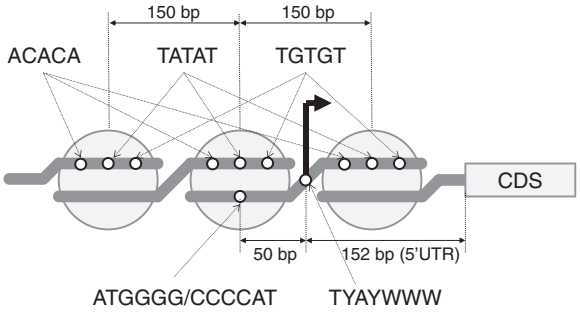


Figure 4 Schematic representation of the composition and speculative nucleosome structure of a model *B. bovis* promoter. Angled arrows, thick lines, big circles, and boxes represent TSSs, DNA, histones, and coding sequences, respectively. Motifs in promoter regions are represented by small circles and the corresponding sequences are indicated by dashed arrows. The TYAYWWW motif over TSS is shown in Figures 1 and 2, and Table 2. The median length from TSS to the 5′ end of CDS (5′ UTR) was 152 bp (Additional file 2: Table S2). The ATGGGG motif and its reciprocal CCCCAT are distributed around −50 nt from TSS (Figure 3A and B). The ACACA, TGTGT, and TATAT motifs appear every 150 bp (Figure 3C). Positional relationships among the motifs and histones are arbitrarily described in this illustration.

Previous investigations of *Plasmodium* and *Toxoplasma* demonstrate promoter structures [43,53-55], putative DNA cis-elements [34,44,50,54,56-58], and the involvement of chromatin structures in transcription [44,53,55,59]. The present analyses of *Babesia* parasites were almost consistent with these studies and warrant the expansion of the concepts related to *Babesia* species. Nonetheless, the use of fine TSS mapping is a critical distinction between the present and previous studies and allowed more specific and sensitive assessment of the distribution of examined motifs, particularly for ACACA, TGTGT, and TATAT motifs that lack definition in previous studies [34,50]. Therefore, the present analyses indicate that the distance from TSSs may be a critical factor for functionality of DNA cis-elements in apicomplexsan parasites.

Conclusions

The full-length cDNAs dataset enable us to revise previous gene model derived from the genome. In parallel, location-specific consensus motifs in promoter sequences were discovered by virtue of TSSs identification with one-base resolution of the method. These observations 1) indicate the utility of integrated bioinformatics and experimental data for improving genome annotations and 2) allowed the illustration of a model promoter composition, which supports the differences in transcriptional regulation frameworks between apicomplexan parasites and mammals.

Methods

Preparation of parasite RNA, and synthesis and sequencing of cDNA

The Texas strain of *B. bovis* was maintained in bovine erythrocytes cultured in GIT medium (WAKO, Osaka, Japan) using a microaerophilic stationary-phase culture system [60]. Total RNA was extracted from *B. bovis*-infected erythrocytes using TRIzol (Invitrogen), and cDNA was synthesized using a previously described oligo-capping method [61]. In briefly, 200 μg of purified total RNA was dephosphorylated using bacterial alkaline phosphatase and was ligated using the oligo-RNA 5′-AGCAU CGAGUCGGCCUUGUUGGCCUACUGG-3′ and T4 RNA ligase. Subsequently, cDNA was synthesized using the oligo-dT fusion primer 5′-GCGGCTGAAGACGGC CTATGTGGCCTTTTTTTTTTTTTTTT-3′ with SuperScript® II (Invitrogen). The cDNA library was amplified using PCR with the primers 5′-AGCATCGAGTCGGC CTTGTTG-3′ and 5′-GCGGCTGAAGACGGCCTATG T-3′. Amplified fragments were then digested using SfiI and were ligated into a DraIII-digested pME18SFL3 plasmid vector in an orientation-defined manner. ESTs of 5′ and 3′ ends were obtained using the Sanger method with ABI 3730 sequencers following standard protocols for sequencing analysis.

Assembly, clustering, and annotation of ESTs

To obtain full-length cDNA sequences, 5′ and 3′ ESTs were assembled using a CAP3 [62] with default parameters. The overlapping nucleic acid length cutoff was 40 and the overlapping identity cutoff was 90% for 5′ and 3′ ESTs, respectively. Putative CDSs of full-length cDNA were examined and intact CDSs of >50 amino acids were selected. Amino acid homology with preproposed genes (BbovisT2BoAnnotatedProteins_PiroplasmaDB-1.1.fasta) was examined using BLAST and homology was considered significant when E-values were 10^{-10}. Full-length cDNA sequences were also mapped onto the genome of *B. bovis* (BbovisT2BoGenomic_PiroplasmaDB-1.1.fasta) using BLAT [63] with default parameters. The categories of full-length cDNA sequences (tier 1; identical, tier 2; amino acid variant, tier 3; structural variant, and tier 4; novel) were assigned according to BLAST and BLAT results using an in-house script. Copy DNAs that were mapped to positions of identical nucleic acid sequences of preproposed genes were assigned as “identical”, and those mapped to identical positions but with amino acid substitutions were assigned as “amino acid variants”. Copy DNAs that were mapped to similar but nonidentical positions to homologous preproposed genes were assigned as “structural variants”, and those with poor homology were assigned as “novel”. Full-length cDNAs were clustered using CAP3 with default parameters, and the subset of cDNAs in the same cluster were integrated into the highest tier. Among full-length cDNAs with differing sequences and the same tier status, cDNAs with fewer mutations and longer amino acids or nucleic acids were selected, and novel cDNAs were annotated using Blast2Go [64]. To estimate gene expression, 5′ ESTs were examined. Initially, these were filtered by BLAST using the *B. bovis* genome database (BbovisT2BoGenomic). Subsequently, the filtered sequences were mapped onto both CDS (BbovisT2BoAnnotatedCDS) and novel sequences using BLAST, and the frequencies of mapped ESTs were determined.

Identification of promoter regions and prevalent motifs

To identify TSSs for each 5′ EST, 36 nts were clipped from 5′ ends and removed if they contained ambiguous nucleotides such as N. Selected sequence sets were then mapped onto the genome sequence of the *B. bovis* T2Bo strain [3] using Bowtie [65] with the acceptance of two mismatches. To assign TARs, mapped positions corresponding with TSSs were clustered if two TSSs were positioned within 20 nts. Gross mapped counts for each position and TAR were tallied, and the most frequently mapped TSS in each TAR was assigned as a representative TSS, as defined in previous reports [33,66]. To identify motifs on TSSs, −10 to +10 regions from representative TSSs were selected and examined using MEME [23]. A frequency weight matrix (FWM) was calculated on the basis of sequence motifs with p values of <0.05, and a sequence logo was generated on the basis of FWM using WebLogo [67]. As a putative promoter region, sequences comprising −1000 to +1000 regions from representative TSSs of each TAR were selected. Human and mouse promoter regions were obtained from DBTSS [33]. Candidate DNA motifs were then estimated using CisFinder [30] with default parameters and −1000 to +1000 regions from representative TSSs. These were also estimated using MEME with n sites of 2023, maximum DNA sizes of 250000, and maxw of 8 as parameters in −100–0 regions of representative TSSs. The distributions of identified motifs around peak TSSs were examined by scanning the motif over the promoter using an in-house script. The

positions of each motif were scanned in exact match condition and summed for every 10 bases. Observed frequencies were divided by the theoretical frequency based on nucleotide biases that were estimated from the nucleotide composition of the genome. The distributions of 5-mer motif candidates were smoothed by averaging the surrounding 30 bases. For functional enrichment analyses, we selected TARs with motifs from the initial TAR set of 2,023-nt sequences. Genes and TARs were linked if TARs were present in the −500 to +200 region from the 5′ end of the CDS, as defined in the BbovisT2BoAnnotatedCDS. Finally, gene ontology terms for *B. bovis* genes were annotated using Blast2Go [64].

Availability of supporting data

Supporting sequence data are available in the DDBJ (http://www.ddbj.nig.ac.jp/index-e.html) under accession numbers HX874250–HX894778 and AK440354–AK442468.

Additional files

Additional file 1: Table S1. Analyzed sequences. (A) Nonredundant full-length cDNA sequences with corresponding *B. bovis* genes; 5′ UTR lengths, isolated sequences, and coding sequences (amino acid and nucleotide sequences) that corresponding with genes; (B) Nonredundant full-length cDNA sequences that were newly annotated in this study; Results from a BLAST search and InterProScan were added.

Additional file 2: Table S2. Model comparison of pre-proposed genes with those identified in this study.

Additional file 3: Table S3. Frequency of gene expression in *B. bovis*.

Additional file 4: Figure S1. Statistical profile of ESTs in *B. bovis*. Counts of ESTs that encode the same genes were converted to a logarithm and were plotted on the horizontal axis. The ranks of EST counts were converted to a logarithm and were plotted on the vertical axis.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JY performed sequence alignments and statistical analyses and drafted the manuscript. HW participated in data management. NY supplied the materials. RY suggested and evaluated motif distributions in human and mouse. YS organized library constructions and sequencing. XX participated in the design and coordination of the project. Il coordinated the project. All authors read and approved the final manuscript.

Authors' information

This paper is dedicated to the memory of Dr. Junichi Watanabe.

Acknowledgements

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) Core-to-Core Program "Research Exchange in Genome Cohort Studies for Field Malaria Parasites and Vector Insects," a Grant-in-aid for Scientific Research on Innovative Areas "Genome Science" (221S0002) from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT), and a Cooperative Research Grant (26-joint-12) of National Research Center for Protozoan Diseases, Obihiro University of Agriculture and Veterinary Medicine.

Author details

[1]Tohoku Medical Megabank Organization, Tohoku University, 6-3-09, aza Aoba, Sendai, Miyagi 980-8579, Japan. [2]National Research Center for Protozoan Diseases, Obihiro University of Agriculture and Veterinary Medicine, Inada-cho west 2-13, Obihiro, Hokkaido 080-8555, Japan. [3]Department of Medical Genome Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan.

Received: 24 May 2013 Accepted: 8 July 2014
Published: 13 August 2014

References

1. Bock R, Jackson L, De Vos A, Jorgensen W: **Babesiosis of cattle.** *Parasitology* 2004, **129**(Suppl):S247–S269.
2. Vial HJ, Gorenflot A: **Chemotherapy against babesiosis.** *Vet Parasitol* 2006, **138**(1–2):147–160.
3. Brayton KA, Lau AO, Herndon DR, Hannick L, Kappmeyer LS, Berens SJ, Bidwell SL, Brown WC, Crabtree J, Fadrosh D, Feldblum T, Forberger HA, Haas BJ, Howell JM, Khouri H, Koo H, Mann DJ, Norimine J, Paulsen IT, Radune D, Ren Q, Smith RK Jr, Suarez CE, White O, Wortman JR, Knowles DP Jr, McElwain TF, Nene VM: **Genome sequence of Babesia bovis and comparative analysis of apicomplexan hemoprotozoa.** *PLoS Pathog* 2007, **3**(10):1401–1413.
4. Wakaguri H, Suzuki Y, Sasaki M, Sugano S, Watanabe J: **Inconsistencies of genome annotations in apicomplexan parasites revealed by 5′-end-one-pass and full-length sequences of oligo-capped cDNAs.** *BMC Genomics* 2009, **10:**312.
5. Yamagishi J, Wakaguri H, Sugano S, Kawano S, Fujisaki K, Sugimoto C, Watanabe J, Suzuki Y, Kimata I, Xuan X: **Construction and analysis of full-length cDNA library of Cryptosporidium parvum.** *Parasitol Int* 2011, **60**(2):199–202.
6. Tsuchihara K, Suzuki Y, Wakaguri H, Irie T, Tanimoto K, Hashimoto S, Matsushima K, Mizushima-Sugano J, Yamashita R, Nakai K, Bentley D, Esumi H, Sugano S: **Massive transcriptional start site analysis of human genes in hypoxia cells.** *Nucleic Acids Res* 2009, **37**(7):2249–2263.
7. Yamagishi J, Wakaguri H, Ueno A, Goo YK, Tolba M, Igarashi M, Nishikawa Y, Sugimoto C, Sugano S, Suzuki Y, Watanabe J, Xuan X: **High-resolution characterization of Toxoplasma gondii transcriptome with a massive parallel sequencing method.** *DNA Res* 2010, **17**(4):233–243.
8. Suzuki Y, Sugano S: **Construction of full-length-enriched cDNA libraries. The oligo-capping method.** *Methods Mol Biol* 2001, **175:**143–153.
9. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.** *Bioinformatics* 2004, **20**(16):2878–2879.
10. Komar AA, Mazumder B, Merrick WC: **A new framework for understanding IRES-mediated translation.** *Gene* 2012, **502**(2):75–86.
11. Serganov A, Nudler E: **A decade of riboswitches.** *Cell* 2013, **152**(1–2):17–24.
12. Mignone F, Gissi C, Liuni S, Pesole G: **Untranslated regions of mRNAs.** *Genome Biol* 2002, **3**(3):REVIEWS0004.
13. Creancier L, Morello D, Mercier P, Prats AC: **Fibroblast growth factor 2 internal ribosome entry site (IRES) activity ex vivo and in transgenic mice reveals a stringent tissue-specific regulation.** *J Cell Biol* 2000, **150**(1):275–281.
14. Lawless C, Pearson RD, Selley JN, Smirnova JB, Grant CM, Ashe MP, Pavitt GD, Hubbard SJ: **Upstream sequence elements direct post-transcriptional regulation of gene expression under stress conditions in yeast.** *BMC Genomics* 2009, **10:**7.
15. Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, Liuni S: **Structural and functional features of eukaryotic mRNA untranslated regions.** *Gene* 2001, **276**(1–2):73–81.
16. Yamagishi J, Watanabe J, Goo YK, Masatani T, Suzuki Y, Xuan X: **Characterization of Toxoplasma gondii 5′ UTR with encyclopedic TSS information.** *J Parasitol* 2012, **98**(2):445–447.
17. De Vries E, Corton C, Harris B, Cornelissen AW, Berriman M: **Expressed sequence tag (EST) analysis of the erythrocytic stages of Babesia bovis.** *Vet Parasitol* 2006, **138**(1–2):61–74.
18. Ogasawara O, Kawamoto S, Okubo K: **Zipf's law and human transcriptomes: an explanation with an evolutionary model.** *C R Biol* 2003, **326**(10–11):1097–1101.

19. Ueda HR, Hayashi S, Matsuyama S, Yomo T, Hashimoto S, Kay SA, Hogenesch JB, Iino M: **Universality and flexibility in gene expression from bacteria to human.** *Proc Natl Acad Sci U S A* 2004, **101**(11):3765–3769.
20. Butler JE, Kadonaga JT: **The RNA polymerase II core promoter: a key component in the regulation of gene expression.** *Genes Dev* 2002, **16**(20):2583–2592.
21. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34**(Database issue):D108–D110.
22. ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee BK, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, *et al*: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**(7414):57–74.
23. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings/International Conference on Intelligent Systems for Molecular Biology; ISMB International Conference on Intelligent Systems for.** *Mol Biol* 1994, **2:**28–36.
24. Bucher P: **Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences.** *J Mol Biol* 1990, **212**(4):563–578.
25. Shahmuradov IA, Gammerman AJ, Hancock JM, Bramley PM, Solovyev VV: **PlantProm: a database of plant promoter sequences.** *Nucleic Acids Res* 2003, **31**(1):114–117.
26. Verrijzer CP, Yokomori K, Chen JL, Tjian R: **Drosophila TAFII150: similarity to yeast gene TSM-1 and specific binding to core promoter DNA.** *Science* 1994, **264**(5161):933–941.
27. Chalkley GE, Verrijzer CP: **DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator.** *EMBO J* 1999, **18**(17):4835–4845.
28. Tora L: **A unified nomenclature for TATA box binding protein (TBP)-associated factors (TAFs) involved in RNA polymerase II transcription.** *Genes Dev* 2002, **16**(6):673–675.
29. Callebaut I, Prat K, Meurice E, Mornon JP, Tomavo S: **Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium falciparum: conserved features and differences relative to other eukaryotes.** *BMC Genomics* 2005, **6:**100.
30. Sharov AA, Ko MS: **Exhaustive search for over-represented DNA sequence motifs with CisFinder.** *DNA Res* 2009, **16**(5):261–273.
31. Guo X, Silva JC: **Properties of non-coding DNA and identification of putative cis-regulatory elements in Theileria parva.** *BMC Genomics* 2008, **9:**582.
32. Beissbarth T, Speed TP: **GOstat: find statistically overrepresented Gene Ontologies within a group of genes.** *Bioinformatics* 2004, **20**(9):1464–1465.
33. Yamashita R, Sugano S, Suzuki Y, Nakai K: **DBTSS: DataBase of Transcriptional Start Sites progress report in 2012.** *Nucleic Acids Res* 2012, **40**(Database issue):D150–D154.
34. 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.
35. Godde JS, Widom J: **Chromatin structure of Schizosaccharomyces pombe. A nucleosome repeat length that is shorter than the chromatosomal DNA length.** *J Mol Biol* 1992, **226**(4):1009–1025.
36. Morris NR: **Nucleosome structure in Aspergillus nidulans.** *Cell* 1976, **8**(3):357–363.
37. Lantermann AB, Straub T, Stralfors A, Yuan GC, Ekwall K, Korber P: **Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of Saccharomyces cerevisiae.** *Nat Struct Mol Biol* 2010, **17**(2):251–257.
38. Lanzer M, De Bruin D, Wertheimer SP, Ravetch JV: **Transcriptional and nucleosomal characterization of a subtelomeric gene cluster flanking a site of chromosomal rearrangements in Plasmodium falciparum.** *Nucleic Acids Res* 1994, **22**(20):4176–4182.
39. Lanzer M, Wertheimer SP, De Bruin D, Ravetch JV: **Chromatin structure determines the sites of chromosome breakages in Plasmodium falciparum.** *Nucleic Acids Res* 1994, **22**(15):3099–3103.
40. Tachiwana H, Kagawa W, Shiga T, Osakabe A, Miya Y, Saito K, Hayashi-Takanaka Y, Oda T, Sato M, Park SY, Kimura H, Kurumizaka H: **Crystal structure of the human centromeric nucleosome containing CENP-A.** *Nature* 2011, **476**(7359):232–235.
41. Bartfai R, Hoeijmakers WA, Salcedo-Amaya AM, Smits AH, Janssen-Megens E, Kaan A, Treeck M, Gilberger TW, Francoijs KJ, Stunnenberg HG: **H2A.Z demarcates intergenic regions of the plasmodium falciparum epigenome that are dynamically marked by H3K9ac and H3K4me3.** *PLoS Pathog* 2010, **6**(12):e1001223.
42. Hoeijmakers WA, Salcedo-Amaya AM, Smits AH, Francoijs KJ, Treeck M, Gilberger TW, Stunnenberg HG, Bartfai R: **H2A.Z/H2B.Z double-variant nucleosomes inhabit the AT-rich promoter regions of the Plasmodium falciparum genome.** *Mol Microbiol* 2013, **87**(5):1061–1073.
43. 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):1376–1384.
44. Ponts N, Harris EY, Lonardi S, Le Roch KG: **Nucleosome occupancy at transcription start sites in the human malaria parasite: a hard-wired evolution of virulence?** *Infect Genet Evol* 2011, **11**(4):716–724.
45. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J: **A genomic code for nucleosome positioning.** *Nature* 2006, **442**(7104):772–778.
46. Ichikawa Y, Morohashi N, Nishimura Y, Kurumizaka H, Shimizu M: **Telomeric repeats act as nucleosome-disfavouring sequences in vivo.** *Nucleic Acids Res* 2014, **42**(3):1541–1552.
47. De Silva EK, Gehrke AR, Olszewski K, Leon I, Chahal JS, Bulyk ML, Llinas M: **Specific DNA-binding by apicomplexan AP2 transcription factors.** *Proc Natl Acad Sci U S A* 2008, **105**(24):8393–8398.
48. Yuda M, Iwanaga S, Shigenobu S, Kato T, Kaneko I: **Transcription factor AP2-Sp and its target genes in malarial sporozoites.** *Mol Microbiol* 2010, **75**(4):854–863.
49. Van Poppel NF, Welagen J, Vermeulen AN, Schaap D: **The complete set of Toxoplasma gondii ribosomal protein genes contains two conserved promoter elements.** *Parasitology* 2006, **133**(Pt 1):19–31.
50. Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinas M: **Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite.** *PLoS Pathog* 2010, **6**(10):e1001165.
51. Flueck C, Bartfai R, Niederwieser I, Witmer K, Alako BT, Moes S, Bozdech Z, Jenoe P, Stunnenberg HG, Voss TS: **A major role for the Plasmodium falciparum ApiAP2 protein PfSIP2 in chromosome end biology.** *PLoS Pathog* 2010, **6**(2):e1000784.
52. Yuda M, Iwanaga S, Shigenobu S, Mair GR, Janse CJ, Waters AP, Kato T, Kaneko I: **Identification of a transcription factor in the mosquito-invasive stage of malaria parasites.** *Mol Microbiol* 2009, **71**(6):1402–1414.
53. Horrocks P, Wong E, Russell K, Emes RD: **Control of gene expression in Plasmodium falciparum - ten years on.** *Mol Biochem Parasitol* 2009, **164**(1):9–25.
54. 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.
55. Horrocks P, Dechering K, Lanzer M: **Control of gene expression in Plasmodium falciparum.** *Mol Biochem Parasitol* 1998, **95**(2):171–181.
56. Wu J, Sieglaff DH, Gervin J, Xie XS: **Discovering regulatory motifs in the Plasmodium genome using comparative genomics.** *Bioinformatics* 2008, **24**(17):1843–1849.
57. Mullapudi N, Joseph SJ, Kissinger JC: **Identification and functional characterization of cis-regulatory elements in the apicomplexan parasite Toxoplasma gondii.** *Genome Biol* 2009, **10**(4):R34.
58. Harris EY, Ponts N, Le Roch KG, Lonardi S: **Chromatin-driven de novo discovery of DNA binding motifs in the human malaria parasite.** *BMC Genomics* 2011, **12:**601.
59. Westenberger SJ, Cui L, Dharia N, Winzeler E, Cui L: **Genome-wide nucleosome mapping of Plasmodium falciparum reveals histone-rich coding and histone-poor intergenic regions and chromatin remodeling of core and subtelomeric genes.** *BMC Genomics* 2009, **10:**610.
60. Bork S, Okamura M, Matsuo T, Kumar S, Yokoyama N, Igarashi I: **Host serum modifies the drug susceptibility of Babesia bovis in vitro.** *Parasitology* 2005, **130**(Pt 5):489–492.

61. Suzuki Y, Sugano S: **Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method.** *Methods Mol Biol* 2003, **221**:73–91.
62. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9**(9):868–877.
63. Kent WJ: **BLAT–the BLAST-like alignment tool.** *Genome Res* 2002, **12**(4):656–664.
64. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21**(18):3674–3676.
65. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**(3):R25.
66. Yamashita R, Suzuki Y, Sugano S, Nakai K: **Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity.** *Gene* 2005, **350**(2):129–136.
67. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator.** *Genome Res* 2004, **14**(6):1188–1190.

doi:10.1186/1471-2164-15-678
Cite this article as: Yamagishi *et al.*: The *Babesia bovis* gene and promoter model: an update from full-length EST analysis. *BMC Genomics* 2014 **15**:678.