Nucleic Acids Research, 2005, Vol. 33, No. 17 ***5503–5511***
doi:10.1093/nar/gki818

Analysis of the transcriptome of the protozoan *Theileria parva* using MPSS reveals that the majority of genes are transcriptionally active in the schizont stage

Richard Bishop*, Trushar Shah, Roger Pelle, David Hoyle[1], Terry Pearson[2], Lee Haines[2], Andrew Brass[3], Helen Hulme[3], Simon P. Graham, Evans L. N. Taracha, Simon Kanga[4], Charles Lu[4], Brian Hass[4], Jennifer Wortman[4], Owen White[4], Malcolm J. Gardner[4], Vishvanath Nene[4] and Etienne P. de Villiers

The International Livestock Research Institute (ILRI), PO Box 30709, Nairobi, Kenya, [1]Department of Computer Science, University of Exeter, North Park Road, Exeter EX4 4QF, UK, [2]Department of Biochemistry and Microbiology, Petch Building, Ring Road, University of Victoria, Victoria BC V8W 3P6, Canada, [3]Department of Computer Science, University of Manchester, Oxford Road, Manchester M13 9PL, UK and [4]The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA

Received June 2, 2005; Revised July 19, 2005; Accepted August 19, 2005

ABSTRACT

Massively parallel signature sequencing (MPSS) was used to analyze the transcriptome of the intracellular protozoan *Theileria parva*. In total 1 095 000, 20 bp sequences representing 4371 different signatures were generated from *T.parva* schizonts. Reproducible signatures were identified within 73% of potentially detectable predicted genes and 83% had signatures in at least one MPSS cycle. A predicted leader peptide was detected on 405 expressed genes. The quantitative range of signatures was 4–52 256 transcripts per million (t.p.m.). Rare transcripts (<50 t.p.m.) were detected from 36% of genes. Sequence signatures approximated a lognormal distribution, as in microarray. Transcripts were widely distributed throughout the genome, although only 47% of 138 telomere-associated open reading frames exhibited signatures. Antisense signatures comprised 13.8% of the total, comparable with *Plasmodium*. Eighty five predicted genes with antisense signatures lacked a sense signature. Antisense transcripts were independently amplified from schizont cDNA and verified by sequencing. The MPSS transcripts per million for seven genes encoding schizont antigens recognized by bovine CD8 T cells varied 1000-fold. There was concordance between transcription and protein expression for heat shock proteins that were very highly expressed according to MPSS and proteomics. The data suggests a low level of baseline transcription from the majority of protein-coding genes.

INTRODUCTION

Genome-wide transcription data are important for understanding organism biology in a systems context and when interfaced with complete genome sequences enable analysis of transcription in relation to genome organization. Several techniques are routinely used for analysis of transcriptomes. These include microarrays based on hybridization (1,2) and serial analysis of gene expression (SAGE), which provides quantitative information from relatively abundant transcripts, based on 3′ signatures from cDNA (3). A powerful new high-throughput method for transcriptome analysis is massively parallel signature sequencing (MPSS), a technique that improves on the SAGE concept by using novel amplification and sequencing technologies to increase the level of sequence

*To whom correspondence should be addressed. Tel: +254 20 4223002; Email: r.bishop@cgiar.org coverage. MPSS allows detection of transcripts expressed at very low levels (4,5). We employed MPSS to analyze the transcriptome of *Theileria parva* and annotated signatures derived from RNA of the schizont stage using the recently determined genome sequence (6). Previously, MPSS has mainly been applied to multicellular organisms, most comprehensively to *Arabidopsis thaliana* (7,8). Among microbial eukaryotes, the transcriptome of *Saccharomcyes cerevisiae* has been analyzed using SAGE (9) and that of the related apicomplexan protozoan *Plasmodium falciparum*, using both high-density oligonucleotide microarrays (10,11) and SAGE (12,13).

The *T.parva* schizont reversibly immortalizes bovine lymphocytes resulting in a leukemia-like phenotype and is responsible for the majority of the pathology resulting from *T.parva* infections of cattle. Infected bovine lymphocytes can be propagated like tumors *in vitro* (14), allowing schizont RNA to be obtained in sufficient quantity for analysis. We used a non-synchronized population of *T.parva* schizont-infected lymphocytes, cultured directly from a cattle lymph node biopsy for MPSS analysis of RNA. Previous transcriptome analyses in protozoan parasites have primarily focused on identification of stage-specific transcripts using microarray. We describe a high-resolution analysis of the genome-wide pattern of transcription within a single parasite life-cycle stage using a quantifiable technique based on sequence signatures whose origins in the genome are verifiable.

MATERIALS AND METHODS

Parasite material and culture conditions

A cloned *T.parva* Muguga sporozoite stabilate (15) was inoculated into an animal and parasite-transformed lymphocytes (from lymph node biopsies) were established in culture to provide the schizont-infected lymphocytes used for schizont purification for the MPSS experiment. The culture procedures were according to published methods (14). The schizonts were enriched to minimize bovine lymphocyte transcript contamination.

cDNA library construction and MPSS analysis and annotation

A cDNA library was generated at Lynx Therapeutics (Hayward, CA) from poly(A)$^+$ mRNA isolated from *T.parva* schizonts purified from bovine lymphocyte cultures. The cDNA library was amplified and loaded onto microbeads. Signature sequences comprising 20 bases 3′ to the poly(A) proximal DpnII site were determined by Lynx Therapeutics using serial enzymatic reactions as described (4). MPSS signature sequences were compared with the draft genome sequence and schizont expressed sequence tag (EST) data using established methods (16). All possible signatures were extracted from the *T.parva* genome sequence database (6) and schizont cDNA sequences generated at the Institute for Genomic Research, Rockville, MA. Each signature was ranked, based on position and orientation in the original sequence. The number of signatures obtained does not equate to the number of genes expressed for a number of technical reasons. Some upstream DpnII signatures may be captured owing to partial digestion of the cDNA and/or chimerisms. Cleavage at upstream DpnII sites can also generate more than one tag from a single transcript. Signatures were converted to transcripts per million (t.p.m.) for the purpose of comparison with other MPSS datasets. The transcripts per million counts are not directly comparable with MPSS data obtained previously from other organisms, since ∼870 000 contaminating bovine transcripts are incorporated into the transcripts per million calculations. The *T.parva* paralogous multicopy families were clustered using the TRIBE-MCL algorithm (17) and signal peptides were predicted using SignalP (18). Data presented in Figures 1 and 2 represent MPSS signature expression levels mapped onto the genome according to gene locus numbers but exclude 15 gene models that were annotated but not allocated a locus number.

Independent verification of antisense transcripts using RT–PCR

Total RNA (100 μg) isolated from a *T.parva* (Muguga) schizont-infected lymphocyte cell line was treated at 37°C for 1 h with 10 U of RNase free DNase I (Promega), extracted with phenol-chloroform and precipitated with ethanol. The RNA pellet was dissolved in 50 μl of sterile distilled water and 2.5 μl of RNA was reverse transcribed into ss-cDNA at 42°C using RNAse H-reverse transcriptase (Invitrogen, Paisley) in the presence of a specific reverse primer. The reaction mixture was heated at 75°C for 15 min and treated with 1 U of RNase H at 37°C for 1 h. Purified and precipitated cDNA was redissolved in 50 μl of TE, pH 7.5. The cDNA was PCR amplified using AmpliTaq Gold (Roche) in the presence of 20 pairs of specific forward and reverse primers. Thermal cycles were as follows: 95°C for 10 min; 95°C for 1 min; 60°C for 45 s; and 72°C for 1 min (35 cycles) with an extension time of 72°C for 10 min. PCR products (5 μl) were analyzed on an ethidium bromide-stained agarose gel (Supplementary Figure 6). For sequencing, 10 μl of the PCR products were treated with 10 U of exonuclease I [United States Biochemicals (USB), OH] and 1 U of shrimp alkaline phosphatase (USB) at 37°C for 15 min, followed by 15 min incubation at 80°C. Treated PCR products were sequenced directly using specific primers. To ascertain whether RNA transcripts contained a poly(A) stretch at their 3′ end, recombinant plasmid DNA was purified from a directional schizont cDNA library in pcDNA3 and used as template for PCR. The reverse primers of the RNA transcripts were replaced by the SP6 reverse primer of the pcDNA3 plasmid vector. Aliquots (10 μl) were treated with exonuclease I and shrimp alkaline phosphatase (USB) and sequenced on an Applied Biosystems 377 automated sequencer.

RESULTS

Summary of MPSS analysis of *T.parva* schizonts

A total of 2.35 million 20 bp signature sequences were generated from a *T.parva* schizont RNA preparation, using MPSS as described (4,5). Signatures detected in more than one iteration of MPSS, with a normalized count of >4 t.p.m., were regarded as having maximum reliability. There were 2132 additional annotated signatures, detected in a single MPSS

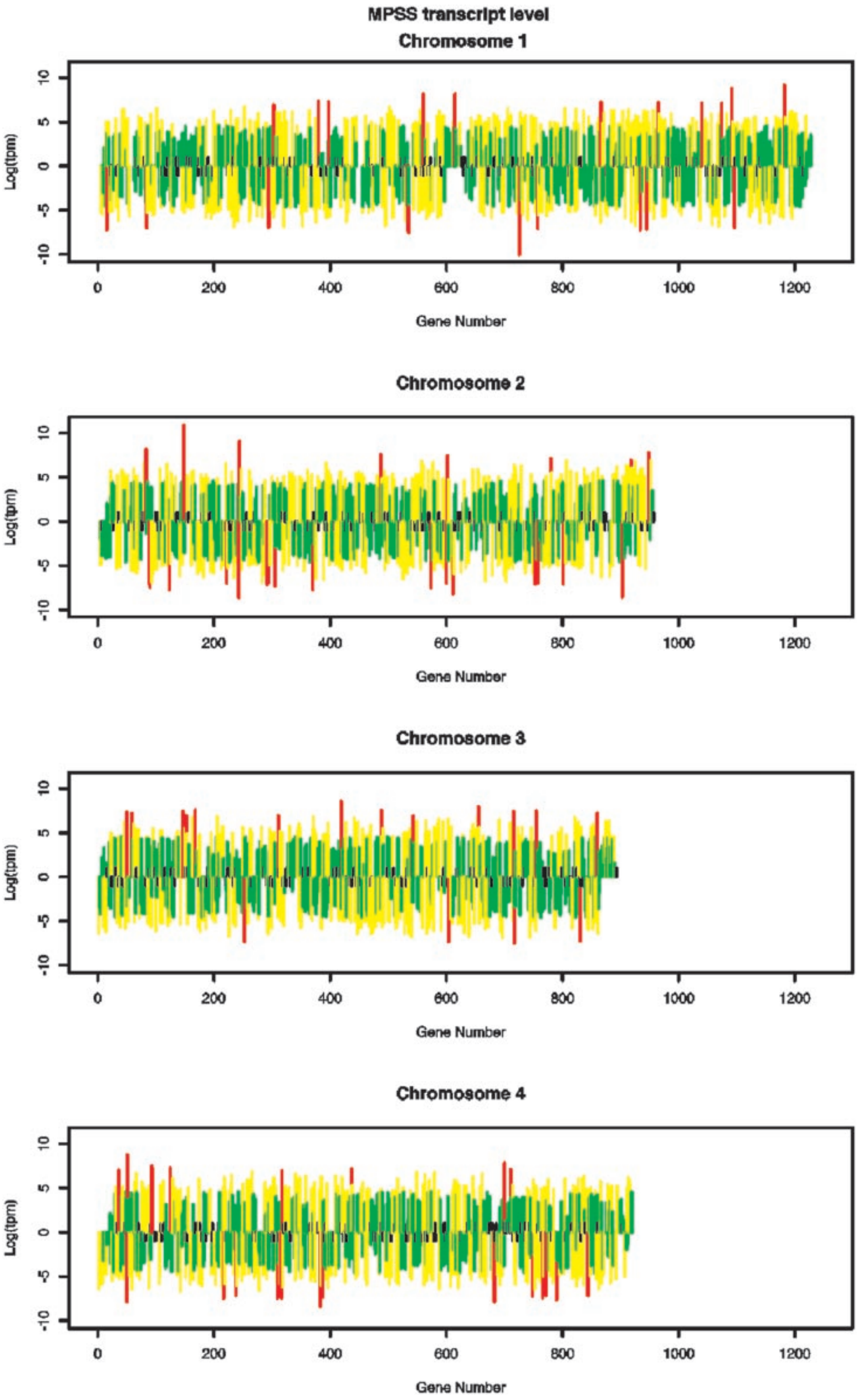

Figure 1. Distribution of sense transcripts within *T.parva* chromosomes 1–4. MPSS signatures >4 t.p.m. that were consistently positive over four cycles of MPSS are illustrated diagrammatically for each gene model. The bars indicate the log transcripts per million for each annotated gene as follows: green (4–99 t.p.m.), yellow (100–999 t.p.m.) and red (>1000 t.p.m.). Gene models lacking DpnII sites are indicated in black. Chromosomes 1–4 are represented in descending order on the page.

cycle. In total, 4571 interpretable signatures (including those with multiple genome hits) representing 4371 different 20mer sequence tags were mapped to the *T.parva* genome. These were derived from 1 095 000 *T.parva*-annotated tags with a dynamic range of >10 000, from 4–52 686 t.p.m. The median value for sense signatures was 83 t.p.m. In order to provide an independent assessment of transcript quantification by MPSS, we examined 20 genes with a wide range of MPSS values

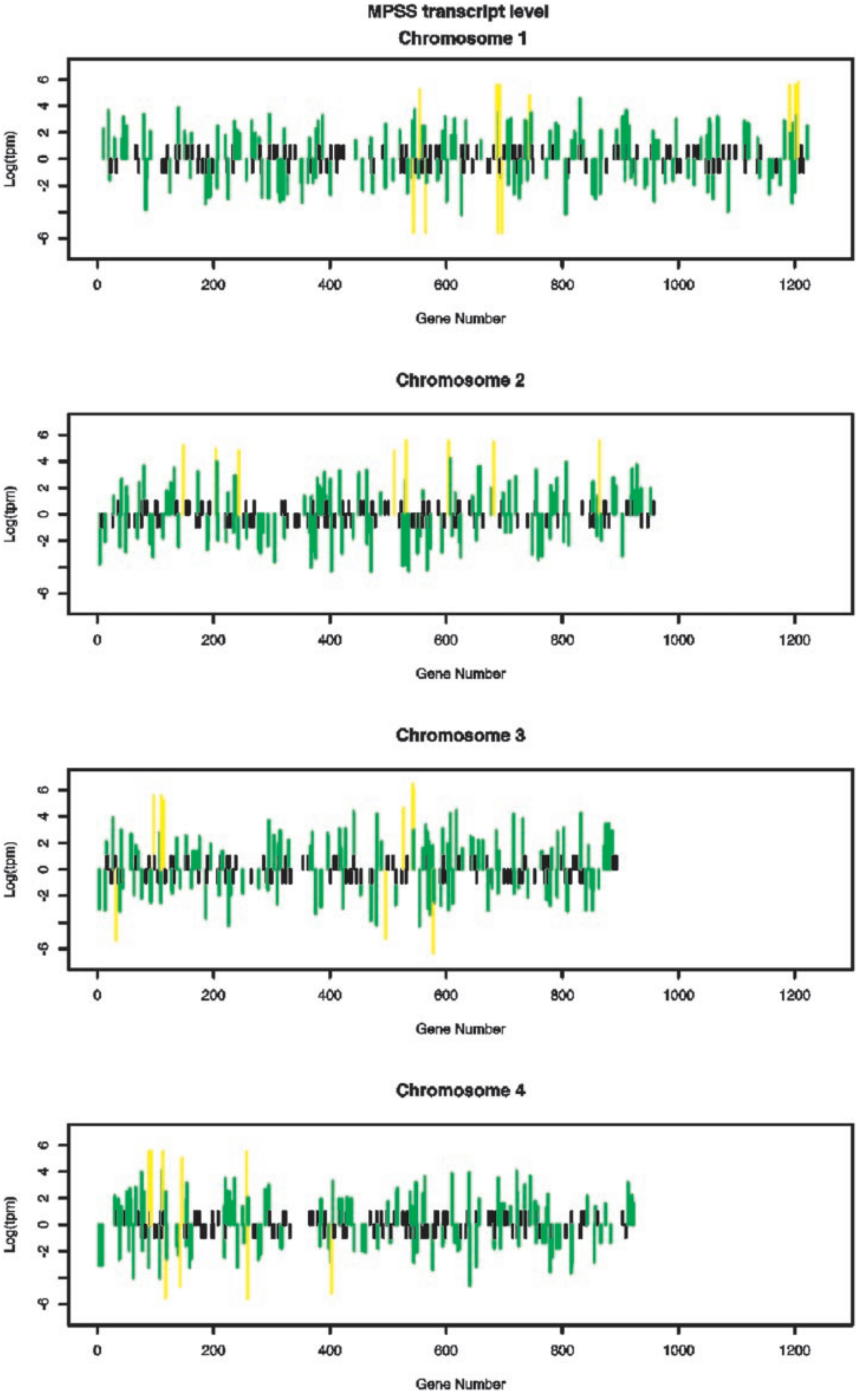

Figure 2. Distribution of antisense transcripts within *T.parva* chromosomes 1–4. MPSS signatures >4 t.p.m. consistently positive over four cycles, derived from the antisense strand of predicted genes are illustrated diagrammatically for each gene model. The bars indicate the log transcripts per million for each annotated gene as follows; green (4–99 t.p.m.) and yellow (100–999 t.p.m.). The distribution of gene models lacking DpnII sites is indicated in black. Chromosomes 1–4 are shown in descending order.

using quantitative PCR (Q-PCR) (19). The results that are shown in the form of a graph in Supplementary Figure 6 and the raw data is provided in Supplementary Table 5 revealed an R^2 of 0.6 in a graphical comparison of Q_PCR Ct versus log MPSS, when a single outlying gene among the 20 analyzed was excluded from the analysis. Although an R^2 value of 0.6 indicates only moderate correlation, it is known from making standard curves against different templates that the sensitivity and amplification efficiency of Q-PCR can vary. Thus, given the inherent difficulties in optimizing PCR conditions across different sequences our Q-PCR data would suggest that there is a reasonable correlation between Q-PCR and MPSS and across several orders of magnitude.

Further analysis showed that the 4371 signatures were derived from the 'sense' strand of 2533 of the 4034 predicted genes within the *T.parva* genome (6). For the purposes of this analysis, a gene was defined as the coding region, together with sequences 50 bp 5′ and 3′ that had a high probability of being located within transcribed, but untranslated regions. The 2533 genes contained 4429 signatures that were 100% identical to genome sequences. These were combined with an additional 142 signatures that matched the coding sequence after removal of putative introns. A detailed analysis of the location of the modeled genes and the MPSS signatures is shown in Supplementary Table 2. Since 573 of the predicted genes lacked a DpnII restriction site, theoretically expression of 86% of the predicted genes could potentially be detected by MPSS. The results revealed that 73% of detectable predicted genes had reproducible MPSS signatures that were detectable in more than one independent cycle of MPSS. Low-level signatures, 1098 between 4 and 10 t.p.m., and 1566 between 11 and 50 t.p.m. were evident at high frequency (Supplementary Table 3). Recent data from *Homo sapiens* suggest that a substantial percentage of such 'orphan' tags represent bona fide low abundance transcripts (20). Annotation against a bovine UniGene Cluster (release 46) database identified 7534 signature tags of apparent bovine origin. Many unassigned signatures will probably prove to be of bovine origin when the data is annotated against the complete bovine genome. A map of the quantitative distribution of transcriptional signatures across the four *T.parva* chromosomes is shown in Figure 1. The *T.parva* MPSS data exhibit an unimodal distribution of logged abundances, similar to a normal distribution (Figure 3). Such distributions are typical of microarray datasets (21).

Antisense MPSS signatures

Examination of the polarity of MPSS signatures relative to modeled genes revealed 637 signatures (13.8%) located on the antisense strand of a modeled gene. This is comparable with the 17% of antisense transcripts detected in *P.falciparum* in two SAGE libraries (12,13), but less than the 31.6% antisense transcripts identified in *A.thaliana*, using MPSS to sample a variety of tissues (8). The average level of antisense transcripts (44 t.p.m.) was considerably lower than for sense transcripts (232 t.p.m.). The dynamic range for the antisense signatures is ~100-fold from 4 to 624 t.p.m., considerably lower than for the sense signatures. The antisense median value is 12. The antisense signatures were derived from 541 different genes with sense signatures and 84 gene models that lacked sense signatures. As in *P.falciparum* (13), antisense transcripts are widely distributed (Figure 2). The presence of 11 of the putative antisense transcripts from the opposite strand of predicted open reading frames (ORFs) was confirmed by RT–PCR using purified schizont cDNA, primed with an 'antisense' oligonucleotide, (for primer details see Supplementary Figure 7 and Supplementary Table 4). The eight schizont cDNA PCR products for which sequences were determined corresponded to the predicted antisense sequence. Seven cloned sequences were also amplified from a directional cDNA library using a 5′ antisense transcript-specific primer and a vector primer. The PCR products derived from the library were also verified by nucleotide sequencing. None of the antisense sequences contained long ORFs with in-frame ATG codons.

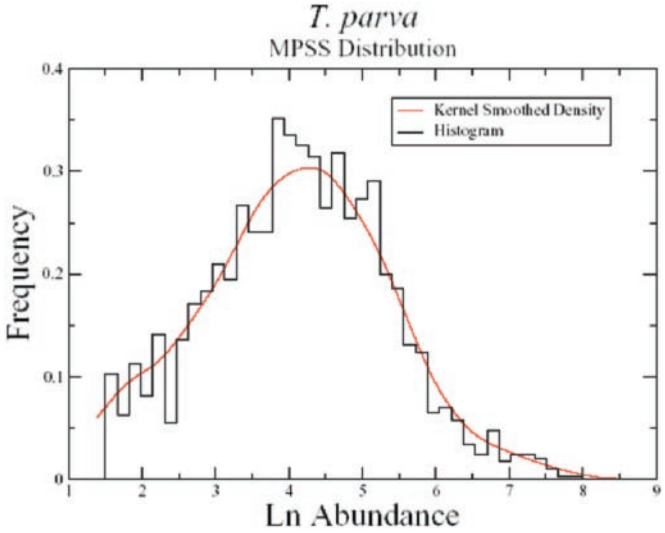

Figure 3. Distribution of signature abundances for *T.parva* signatures after logarithmic transformation of the data. Data are shown only for the subset of replicated and significant signatures annotated to a single locus in the *T.parva* genome, resulting in the truncation at the left-hand tail. Frequencies on the y-axis correspond to probability density. The solid black line shows a simple histogram of logged signature abundances whilst the solid red line shows a more reliable fixed width tri-cube kernel smoothed estimate of the probability density.

Transcription of telomere-associated ORFs

Genes located directly adjacent to telomeres are subject to position effects that modulate their expression, as described for *S.cerevisiae* (22) and several pathogenic microorganisms (23). In *T.parva*, protein-encoding genes are located very close to the telomeres. On average, the first protein-coding gene is 2.5 kb centromere-proximal to the telomeric repeats and genes encoding members of *Theileria*-specific multigene families are located in the subtelomeric regions. Signatures from 56 of the 85 members of the glutamine-proline rich subtelomeric variable secreted protein family and 13 of the 61 members of the SfiI-fragment-related family had low to medium level MPSS signatures (4–606 t.p.m.), indicating active transcription (Figure 4). There was a signature from only 1 of the 7 potentially detectable telomere-proximal ORFs, located on chromosome 1 (1 of the 8 ORFs adjacent to a telomere did not contain a DpnII site). As in the case of transcription of the *var* genes in *P.falciparum* erythrocytes, assessed using microarray analysis (10), several signatures for internal copies of the two subtelomeric protein families were detected. Among the 56 expressed telomeric ORFs, 33 (59%) have signal sequences, consistent with a role in pathogen-host dinteraction.

Functional classification of genes with MPSS signatures

T.parva proteins were sorted into functional classes based on the Munich Information Centre for Protein Sequences (MIPS)

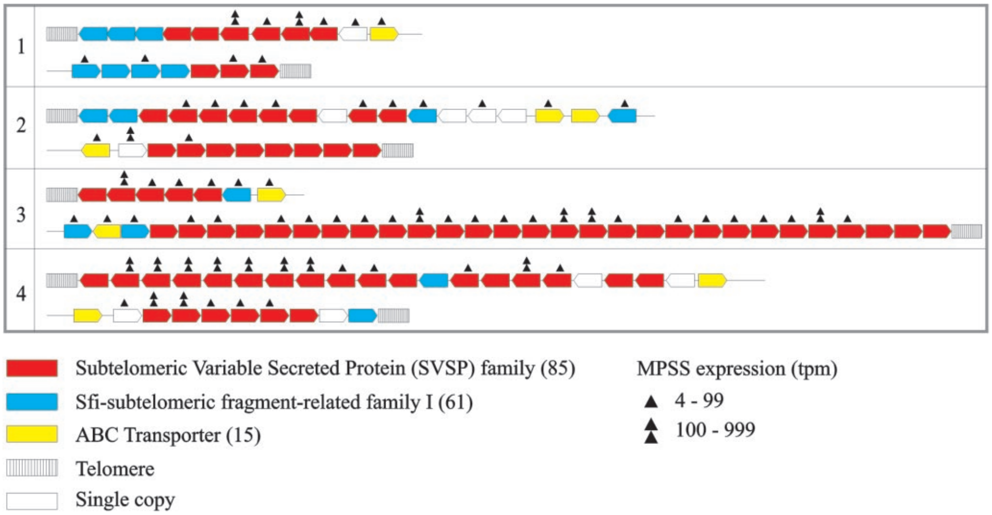
1
2
3
4
Subtelomeric Variable Secreted Protein (SVSP) family (85)
Sfi-subtelomeric fragment-related family I (61)
ABC Transporter (15)
Telomere
Single copy
MPSS expression (tpm)
4 - 99
100 - 999

Figure 4. Expression of telomere-associated genes in *T.parva* clustered using the TRIBE-MCL algorithm. Telomere-associated regions of each of the four *T.parva* chromosomes are illustrated showing the relative position of the ORFs on the forward (top) and reverse (bottom) strands with the multigene families color-coded. The left telomere is shown above the right. Genes within the same gene family identified using TRIBE-MCL are color-coded, with the copy number in the genome indicated. Arrows above each ORF indicate the MPSS expression level in two quantitative categories; 4–99, 100–999.

catalog (24). The different categories expressed are illustrated diagrammatically in Figure 5. Sixty nine percent of genes encoded hypothetical proteins (Supplementary Table 2). The 20 most highly expressed genes included 5 located within the mitochondrion (Supplementary Table 5A). We identified 405 (16%) expressed genes that contain sequences encoding a signal peptide predicted by Signal P, among which the majority (75%) have no significant identity to genes in the public databases and were annotated as encoding hypothetical or conserved-hypothetical proteins. These represent novel *T.parva* proteins, a proportion of which may be accessible to the host MHC class I antigen processing and presentation pathway and may represent potential vaccine targets.

Genes encoding four heat shock proteins were among the 20 most highly expressed schizont genes according to MPSS (Supplementary Table 5A). Hsp90 is important for normal growth and development in eukaryotes and together with Hsp70 helps newly synthesized proteins to fold. They also regulate activities of transcription factors and protein kinases. In *P.falciparum*, Hsp90 has been shown to be essential for survival in erythrocytes Hsp (25). In *T.parva*-infected bovine lymphocytes Hsp90 accesses the bovine cytosol and represents a target of CTL responses against the schizont (S. P. Graham, personal communication).

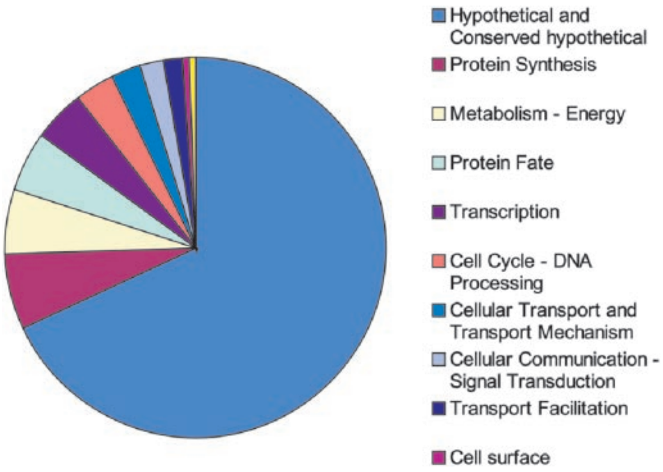

Figure 5. Functional profiles of expressed proteins plotted according to their classification as defined by MIPS catalog (24). Only one class was assigned per protein to avoid redundancy. A complete list of the properties of the signatures, including functional categorization of predicted proteins is provided in Supplementary Table 2.

Comparative analysis of the transcriptome and proteome

There appeared to be a strong correlation between MPSS transcripts per million levels (Supplementary Table 5A) and the 12 most abundant (based on staining with colloidal Coomassie Brilliant blue G-250) schizont proteins separated by two-dimensional gel electrophoresis and identified using tandem mass spectrometry (Supplementary Table 5B). Ten of the 12 abundant *T.parva* proteins had corresponding transcript levels of >500 t.p.m., and were among the most highly transcribed proteins identified by MPSS analysis. This was particularly evident for heat shock proteins. These represented 4 of the 10 most highly expressed proteins identified by protein mass spectrometry, including an Hsp70 homolog with the highest MPSS signature abundance (52 686 t.p.m.). The polymorphic immunodominant molecule, a highly expressed protein in schizonts (26), was also among the 20 most abundant signatures in the MPSS dataset (Supplementary Table 5A). Thus there was strong evidence for a degree of concordance between high levels of gene transcription and protein expression. In contrast there was a 1000-fold range, from 6 to 5974 t.p.m., (Table 1) in transcription as indicated by MPSS analysis for seven antigens that are recognized by $CD8^+$ MHC class I-restricted cytotoxic T cells, (S. P. Graham and E. L. Taracha, unpublished data) induced by live vaccination

Table 1. MPSS signatures of genes encoding antigens, Tp1–Tp5 and Tp7–Tp8 that are recognized by bovine CD8+ class I-restricted cytotoxic T cells

Antigen	Gene length (bp)	Protein size (amino acids)	Molecular weight (predicted)	PI (predicted)	MPSS t.p.m.
Tp1	1771	543	61434.6	6.31	261
Tp2	1019	174	19142.3	8.29	1069
Tp3	827	265	28685.4	6.54	44
Tp4	2595	579	63357.5	5.91	6
Tp5	765	155	17806.7	4.49	163
Tp7	2996	721	83667.7	4.77	8803
Tp8	1581	440	50026.8	7.37	64

Tp6 did not contain a DpnII site and was therefore not detectable by MPSS.

with *T.parva* stabilates. In the case of Tp4 (Table 1) the signature (6 t.p.m.) was detected in only one out of four MPSS cycles.

DISCUSSION

Transcriptome analysis using MPSS

For technical reasons, MPSS cannot provide a completely quantitative description of a transcriptome. Genes lacking DpnII sites, including 573 (14%) of the 4034 predicted *T.parva* genes, are not detectable using this technique, although use of additional enzymes would increase coverage. It has also been demonstrated that certain sequence signatures are underrepresented in *A.thaliana* MPSS datasets derived from multiple tissues (8). However, the observation that 1612 *T.parva* schizont EST sequences determined at TIGR (M. Gardner and V. Nene, unpublished data) had corresponding MPSS signatures, and the similarity of the distribution of the *T.parva* MPSS data to alternative methods of genome-wide transcriptome analysis in other organisms (21) strongly suggests that MPSS is broadly representative of the overall transcriptional profile. Recent data indicates that the accuracy of microarrays, the mostly widely used method for transcriptome analysis to date, is dependent on sequence verification of the probes, which is not currently routinely performed (27,28). Given the complex experimental protocols involved in assessing expression activity using microarrays, it is encouraging that the MPSS data also displays a lognormal distribution since, unlike microarray analysis, transcript abundance is assessed directly using MPSS, rather than indirectly by inference from hybridization.

Antisense transcripts originating from the opposite strand of ORFs have been detected in many prokaryotic organisms, and increasingly in eukaryotes, and are speculated to perform a variety of functional roles (29), including down-regulation of gene expression. We observed 88 antisense signatures originating from 85 genes, three of which contained two antisense signatures at different locations. The absence of a corresponding sense signature in these genes may be the result of regulatory silencing by the antisense transcript. Among these 85 genes, 56 were hypothetical, 8 were conserved hypothetical and 21 had a diverse range of putative functions based on sequence identity. Thus there is no clear indication of a common functional role among this set of genes and alternative explanations for the lack of a sense MPSS signature are also possible.

Posttranscriptional control of expression has been demonstrated in *T.parva* (30) and in certain life-cycle stages of *P.falciparum* (31). Thus factors affecting the composition of the proteome are complex and regulated at multiple levels in apicomplexan parasites. The relative importance of transcriptional and posttranscriptional processes in contributing to the *T.parva* proteome remains to be determined. Our data indicates a close correlation between high levels of transcription and protein expression in certain classes of gene, particularly heat shock proteins. By contrast the recognition of *T.parva* antigens by bovine cytotoxic T cells is consistent with the hypothesis that even genes with low and inconsistent levels of transcription, e.g. Tp4 (Table 1) from which a signature was detected only in a single MPSS cycle can be expressed sufficiently for peptides to be presented and recognized by components of the immune system. One explanation for this could be high affinity of binding of peptides derived from weakly expressed proteins to specific bovine class I MHC proteins.

As mentioned previously, recent observations indicate that microarray hybridization signals may require validation by probe sequence verification (27). Sequence signature techniques, such as MPSS, may therefore be a more accurate method for quantification of transcripts throughout the genome.

Implications of MPSS data for transcriptional processes in unicellular eukaryotes

The high depth of coverage of MPSS (5) theoretically enables the detection of only a few RNA molecules and in *T.parva* a significant proportion of transcripts appeared to be expressed at low levels. Inclusion of MPSS signatures present in only a single cycle increased the total number of unique signatures annotated to the *T.parva* genome to 6503 and the number of predicted genes with schizont MPSS signatures by 355. Overall, signatures were therefore detected from 83% of potentially detectable genes in at least one MPSS experiment, in RNA from a single parasite life-cycle stage. The biological significance of this observation is unclear, but inconsistent MPSS signatures present in only a single run have been observed from genes in other organisms (32). Such differences may be stochastic rather than stable and the expression profile could differ at another time point, or in different cell lines, as observed previously with *P.falciparum* (33). Additional data relating to expression of *T.parva* telomere-associated ORFs provide support for this hypothesis. For example cDNA clones corresponding to the 5th ORF centromeric to the $3'$ telomere of chromosome 1 (34) and the 7th ORF centromeric to the $5'$ telomere of chromosome 2 (35) have been isolated previously by screening cDNA libraries suggesting that they are not rare in all infections. The latter was also detected using northern blotting. However, these two genes had relatively low MPSS signatures in the current study, 18 and 39 t.p.m., respectively. The telomeres of chromosome 4 were the most transcriptionally active in the particular infected lymphocyte population that we analyzed that was derived from a lymph node biopsy from an animal infected with *T.parva in vivo*. ORFs 2–10 were transcribed at the 5′ end of chromosome 4 and 4–9 at the 3′ end, 9 of these at levels of >100 t.p.m. It will be of interest to see if bias in expression to these sub-telomeres is consistent in different *T.parva* infections of cattle.

Transcription in apicomplexan protozoa, which appear to lack specific transcription factors with homology to those of other eukaryotes (31,36,37), is of particular interest. The average intergenic distance in *T.parva* is only 408 bp (6), and it is unknown whether this length of sequence is sufficient for binding of a transcriptional regulatory complex. Distances between ORFs can be extremely short in *T.parva*; e.g. the gene encoding the p67 antigen and a downstream hypothetical coding sequence are separated by only 93 bp. The two genes have MPSS transcripts per million of 22 and 34, respectively, suggesting that their transcription may not be independent. The similarity of the *T.parva* MPSS signature distribution to a lognormal distribution suggests that the observed signatures result from several processes, indicating that regulation of gene expression at the transcriptional level is complex. The striking concordance of proteome and transcriptome data for heat shock proteins suggests active regulation of transcription at certain classes of loci. However, our data are also consistent with the hypothesis that there is a baseline of low level, non-stringently regulated transcription of the majority of genes, that may vary stochastically between different *T.parva* infected lymphocyte cell lines or at different time points within the same infection.

SUPPLEMENTARY DATA

Supplementary Data is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Paul Spooner, Luka Juma and Edward Okoth for culturing *T.parva* schizont-infected lymphocytes and Francis Chuma for schizont purification. We are grateful to Lynx Therapeutics, Hayward California and the UVIC/Genome BC Proteomics Centre, Victoria, British Columbia for their efficient service. We also thank Rose Ndegwa (ILRI) and Michael Brown (TIGR) for developing research agreements. We appreciate the contribution of ILRI scientific support staff, particularly the Sequencing and Genotyping Unit. The work was partially supported by funding from the Animal Health Programme of the Department for International Development (DFID), UK and by the Natural Sciences and Engineering Research Council of Canada. This is ILRI publication number 200501. Funding to pay the Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Duggan,D.J., Bittner,M., Chen,Y., Meltzer,P. and Trent,J.M. (1999) Expression profiling using cDNA microarrays. *Nature Genet*., **21**, 10–14.
2. Hacia,J.G. (1999) Resequencing and mutational analysis using oligonucleotide microarrays. *Nature Genet*., **21**, 42–47.
3. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
4. Brenner,S., Johnson,M., Bridgham,J., Golda,G., Lloyd,D.H., Johnson,D., Luo,S., McCurdy,S., Foy,M., Ewan,M. *et al*. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol*., **18**, 630–634.
5. Reinartz,J., Bruyns,E., Lin,J.Z., Burcham,T., Brenner,S., Bowen,B., Kramer,M. and Woychik,R. (2002) Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Brief Funct. Genomic. Proteomic*., **1**, 95–104.
6. Gardner,M.J., Bishop,R., Shah,T., de Villiers,E.P., Carlton,J.M., Hall,N., Ren,Q., Paulsen,I.T., Pain,A., Berriman,M. *et al*. (2005) Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science*, **309**, 134–137.
7. Meyers,B.C., Vu,T.H., Tej,S.S., Ghazal,H., Matvienko,M., Agrawal,V., Ning,J. and Haudenschild,C.D. (2004) Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing. *Nat. Biotechnol*., **22**, 1006–1011.
8. Meyers,B.C., Singh,T.S., Vu,T.H., Haudenschild,C.D., Agrawal,V., Edberg,S.B., Ghazal,H. and Decola,S. (2004) The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis*. *Genome Res*., **14**, 1614–1653.
9. Velculescu,V.E., Zhang,L., Zhou,W., Vogelstein,J., Basrai,M.A., Bassett,D.E.,Jr, Hieter,P., Vogelstein,B. and Kinzler,K.W. (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243–251.
10. Le Roch,K.G., Zhou,Y., Blair,P.L., Grainger,M., Moch,J.K., Haynes,J.D., De La Vega,P., Holder,A.A., Batalov,S., Carucci,D.J. *et al*. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, **301**, 1503–1508.
11. Bozdech,Z., Llinas,M., Pulliam,B.L., Wong,E.D., Zhu,J. and DeRisi,J.L. (2003) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS. Biol*., **1**, E5.
12. Patankar,S., Munasinghe,A., Shoaibi,A., Cummings,L.M. and Wirth,D.F. (2001) Serial analysis of gene expression in *Plasmodium falciparum* reveals the global expression profile of erythrocytic stages and the presence of anti-sense transcripts in the malarial parasite. *Mol. Biol. Cell*, **12**, 3114–3125.
13. Gunasekera,A.M., Patankar,S., Schug,J., Eisen,G., Kissinger,J., Roos,D. and Wirth,D.F. (2004) Widespread distribution of antisense transcripts in the *Plasmodium falciparum* genome. *Mol. Biochem. Parasitol*., **136**, 35–42.
14. Brown,C.G., Stagg,D.A., Purnell,R.E., Kanhai,G.K. and Payne,R.C. (1973) Letter: infection and transformation of bovine lymphoid cells *in vitro* by infective particles of *Theileria parva*. *Nature*, **245**, 101–103.
15. Morzaria,S.P., Dolan,T.T., Norval,R.A., Bishop,R.P. and Spooner,P.R. (1995) Generation and characterization of cloned *Theileria parva* parasites. *Parasitology*, **111**, 39–49.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
17. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*., **30**, 1575–1584.
18. Nielsen,H., Englebrecht,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*., **10**, 1–6.
19. Bustin,S.A. (2000) Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J. Mol. Endocrinol*., **25**, 169–193.
20. Silva,A.P., Chen,J., Carraro,D.M., Wang,S. and M., Camargo,A.A. (2004) Generation of longer 3′ cDNA fragments from massively parallel signature sequencing tags. *Nucleic Acids Res*., **32**, e94.
21. Hoyle,D.C., Rattray,M., Jupp,R. and Brass,A. (2002) Making sense of microarray data distributions. *Bioinformatics*, **18**, 576–584.

22. Vega-Palas,M.A., Martin-Figueroa,E. and Florencio,F.J. (2000) Telomeric silencing of a natural subtelomeric gene. *Mol. Gen. Genet.*, **263**, 287–291.
23. Barry,J.D., Ginger,M.L., Burton,P. and McCulloch,R. (2003) Why are parasite contingency genes often associated with telomeres? *Int. J. Parasitol.*, **33**, 29–45.
24. Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.
25. Banumathy,G., Singh,V., Pavithra,S.R. and Tatu,U. (2003) Heat shock protein 90 function is essential for *Plasmodium falciparum* growth in human erythrocytes. *J. Biol. Chem.*, **278**, 18336–18345.
26. Toye,P.G., Goddeeris,B.M., Iams,K., Musoke,A.J. and Morrison,W.I. (1991) Characterization of a polymorphic immunodominant molecule in sporozoites and schizonts of *Theileria parva*. *Parasite Immunol.*, **13**, 49–62.
27. Mecham,B.H., Wetmore,D.Z., Szallasi,Z., Sadovsky,Y., Kohane,I. and Mariani,T.J. (2004) Increased measurement accuracy for sequence-verified microarray probes. *Physiol. Genomics*, **18**, 308–315.
28. Tan,P.K., Downey,T.J., Spitznagel,E.L.,Jr, Xu,P., Fu,D., Dimitrov,D.S., Lempicki,R.A., Raaka,B.M. and Cam,M.C. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.*, **31**, 5676–5684.
29. Vanhee-Brossollet,C. and Vaquero,C. (1998) Do natural antisense transcripts make sense in eukaryotes? *Gene*, **211**, 1–9.
30. Nene,V., Bishop,R., Morzaria,S., Gardner,M.J., Sugimoto,C., ole-MoiYoi,O.K., Fraser,C.M. and Irvin,A. (2000) *Theileria parva* genomics reveals an atypical apicomplexan genome. *Int. J. Parasitol.*, **30**, 465–474.
31. Wirth,D.F. (2002) Biological revelations. *Nature*, **419**, 495–496.
32. Stolovitzky,G.A., Kundaje,A., Held,G.A., Duggar,K.H., Haudenschild,C.D., Zhou,D., Vasicek,T.J., Smith,K.D., Aderem,A. and Roach,J.C. (2005) Statistical analysis of MPSS measurements: application to the study of LPS-activated macrophage gene expression. *Proc. Natl Acad. Sci. USA*, **102**, 1402–1407.
33. Ganesan,K., Jiang,L. and Rathod,P.K. (2002) Stochastic versus stable transcriptional differences on *Plasmodium falciparum* DNA microarrays. *Int. J. Parasitol.*, **32**, 1543–1550.
34. Bishop,R., Gobright,E., Nene,V., Morzaria,S., Musoke,A. and Sohanpal,B. (2000) Polymorphic open reading frames encoding secretory proteins are located less than 3 kilobases from *Theileria parva* telomeres. *Mol. Biochem. Parasitol.*, **110**, 359–371.
35. Bishop,R., Geysen,D., Skilton,R., Odongo,D., Nene,V., Allsopp,B., Mbogo,S., Spooner,P. and Morzaria,S. (2002) Genomic polymorphism, sexual recombination and molecular epidemiology of *Theileria parva*. In Dobbelaere,D.A.E. and Mckeever,D.J. (eds), *Theileria*. Kluwer Academic Press, Boston, pp. 23–40.
36. Aravind,L., Iyer,L.M., Wellems,T.E. and Miller,L.H. (2003) *Plasmodium* biology: genomic gleanings. *Cell*, **115**, 771–785.
37. Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S. *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum.* *Nature*, **419**, 498–511.