J. Parasitol., 98(2), 2012, pp. 445–447

Characterization of *Toxoplasma gondii* 5′ UTR with Encyclopedic TSS Information

J. Yamagishi, J. Watanabe*, Y. K. Goo, T. Masatani, Y. Suzuki†, and X. Xuan‡, National Research Center for Protozoan Diseases, Obihiro University of Agriculture and Veterinary Medicine, Obihiro, Japan 080-8555; *Department of Parasitology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan 108-8639; †Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan 277-8583; ‡To whom correspondence should be addressed. *e-mail: gen@obihiro.ac.jp*

ABSTRACT: The 5′ UTR is widely involved in gene expression via post-transcriptional regulation. However, a detailed profile of the 5′ UTR for *Toxoplasma gondii* has not yet been demonstrated. To investigate the issue, we compared the predicted open reading frames (ORFs) and transcription start sites (TSSs) of *T. gondii* obtained by TSS-seq, a method that enables analysis of encyclopedic TSSs with next-generation sequencers. As a result, it was demonstrated that the mode length of the 5′ UTR is between 120 and 140 nucleotides (nts) when a subset of genes with predicted signal peptides was examined. However, when genes without the signal peptide were examined, the length was extended to approximately 600 nts. Because additional information on the predicted signal peptide generates increased reliability to the 5′ end estimation of each ORF, we believe that the former value was more reliable as a representative of the 5′ UTR length of *T. gondii*. The discrepancy suggests that current predictions of the 5′ end of the ORF were less accurate and considerably more discordant with the natural status. The 5′ untranslated region (5′ UTR) is defined as that between the 5′ end of the transcripts and just in front of a start codon of an ORF. Therefore, the 5′ UTR does not contain any information for a protein sequence; however, it is involved in the control of protein expression via the modulation of translational efficiency (Kozak, 1991b; Hughes, 2006).

Toxoplasma gondii is an obligate apicomplexan parasite that can invade a wide variety of mammalian cells, including those of humans and many economically important domestic animals (Dubey, 2009). However, its 5′ UTR is currently poorly understood. A comparative analysis between the 5′ EST tag and predicted ORFs in *T. gondii* demonstrated that the average length of the 5′ UTR is 288 nucleotides (nts) (Wakaguri et al., 2009). It is known that the average is 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 (Pesole et al., 2001). In addition, the 5′ UTR analysis for a limited number of individual *T. gondii* genes also showed discordance against the estimation with the 5′ EST analysis. For example, the estimated lengths are 95 nts for *SAG1* (Burg et al., 1988) and approximately 100 nts for *GRA1*, *GRA2*, *GRA5*, and *GRA6* (Mercier et al., 1996). Therefore, we implemented a detailed and global characterization of the *T. gondii* 5′ UTR lengths in the present study.

For this purpose, both accurate information of transcription stat sites (TSSs) and dependable annotation for open reading frame (ORF) were indispensable. For the former, a published data set for TSSs demonstrated by TSS-seq (Tsuchihara et al., 2009; Yamagishi et al., 2010), a currently established encyclopedic method to identify TSSs with 1-nt resolution, was sufficiently feasible and reliable. However, for the latter, the predicted 5′ end of each ORF was not sufficiently reliable because they were created using an algorithm calibrated for mammals or plants that belong to an entirely different group of taxa than *T. gondii*. In reality, full-length cDNAs and predicted ORFs have been reported to have 40% inconsistency for *T. gondii* (Wakaguri et al., 2009). To overcome this limitation, we generated a subset of annotated genes possessing a signal peptide for translocation to the secretory pathway to increase the reliability of the prediction. The signal peptide is an amino acid sequence present at the N- terminus of proteins, which is approximately 20 amino acids (aa) long (von Heijne, 1990). Peptides with this motif interact with signal recognition molecules, and the complexes are recruited to the endoplasmic reticulum (Egea et al., 2005). In general, this peptide has a specific, obvious pattern; therefore, several algorithms permit the prediction of the existence of the motif with high reliability (Choo et al., 2009). This means that the 5′ end positions of predicted ORFs with a predicted signal peptide are more reliable than those without the signal peptide.

In practice, we referred to TgondiiME49AnnotatedProteins_ToxoDB-5.3.fasta (Gajria et al., 2008) as an assigned protein data set. After that, 970 of 7,993 sequences were predicted to have a signal peptide after examination of the signal P (Bendtsen et al., 2004) using a default parameter. A score with more than 0.95 for the signal peptide or signal anchor probability was applied as a positive threshold. Here, the 5′ UTR lengths were defined as lengths between TSSs and the 5′ end positions of each predicted ORF. The 5′ end positions of each predicted ORF were then selected from TgondiiME49_ToxoDB-5.3.gff, a general feature format (GFF) description of *T. gondii* (Gajria et al., 2008). A TSS data set for the *T. gondii* RH strain propagated in mice (Yamagishi et al., 2010) was mapped on sequences corresponding to regions from −10,000 to +5,000 from the 5′ end positions of each ORF (Fig. 1A). The total mapped tag number for each gene was normalized (Fig. 1B) because the transcriptional frequency follows a power law, such as the distribution (Yamagishi et al., 2010) and, without normalization, the distribution of the 5′ UTR lengths will be biased by a few very strongly expressed genes. On the other hand, genes with few TSSs would disturb the ideal distribution of 5′ UTR lengths. Therefore, we selected 3,778 genes with more than 10-ppm tags against the total mapped tags. Among them, 499 genes had a signal peptide and 3,279 genes did not have it. Finally, the normalized TSS tags were integrated into 2 groups (Fig. 1C).

As a result, the mode values of the 5′ UTR were between 120 and 140 nts for the data set and with a signal sequence and between 580 and 600 nts for that without a signal sequence (Fig. 2A). This obvious discordance implies that genes with a signal peptide have specific profiles in their 5′ UTR, or that predicted ORFs without a signal peptide were less reliable than those with a signal peptide, with the failed prediction resulting in an inaccurate elongation of the 5′ UTR. To scrutinize these probabilities, we examined experimentally characterized 5′ UTRs. When *ENO1*, *ENO2*,

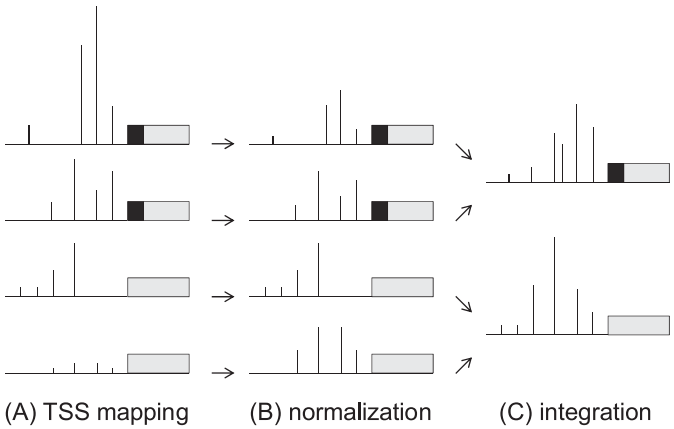

FIGURE 1. Schematic representation of 5′ UTR analysis. (**A**) The distribution of the 5′ UTR length was examined by comparison between the position and frequency of mapped TSSs and the 5′ end position of the coding region of annotated genes. (**B**) The TSS occurrence was normalized by division by total TSS occurrence gene by gene. (**C**) The normalized TSS occurrence was integrated depending on the signal peptide. The vertical lines represent the position and frequency of the mapped TSSs. The light-gray squares represent coding regions. The dark-gray squares represent signal peptides.

DOI: 10.1645/GE-2864.1

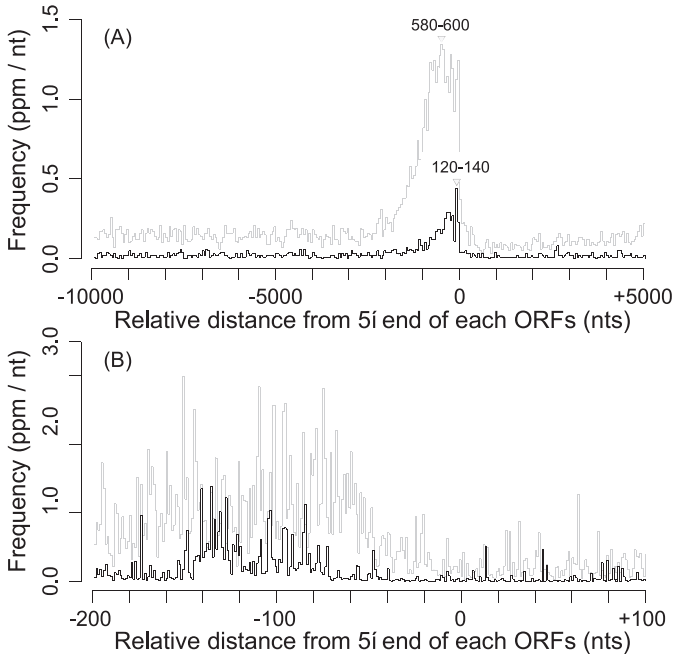

FIGURE 2. Distribution of the 5′ UTR length. (**A**) Distribution between −10,000 and +5,000 from the 5′ end position of an annotated coding region. (**B**) Expanded graph between −200 and +100. The black line represents the frequency distribution for those with a signal peptide. The gray line represents those without a signal peptide. The arrowheads represent mode positions.

DHFR-TS, and *ACT1* were investigated, the average length was 201.3 nts (Dobrowolski et al., 1997; Kibe et al., 2005). Because the number of subjected genes was limited, the average could not represent the overall trend, but it was clearly distinguished from the estimated average for the ORFs without a signal sequence. Moreover, it has also been reported that the ORF prediction in *T. gondii* has a considerable amount of incorrect annotation (Wakaguri et al., 2009). Indeed, the mode length for the group with signal sequences showed good accordance with the length observed in other organisms (Pesole et al., 2001). Consequently, it is suggested that the typical 5′ UTR length of *T. gondii* distributes between 120 and 140 nts, and the longer estimation was derived from a less-accurate ORF prediction.

Many distributions follow the normal distribution pattern. However, the observed distribution of the 5′ UTR length does not (Fig. 2A). In particular, the frequency was dramatically reduced when the region was shorter than 80 nts (Fig. 2B). On the contrary, it was gradually reduced in regions longer than the mode length (Fig. 2A). This suggests that a shorter 5′ UTR is preferred, but there is a minimum length that enables translation. In a related observation, the efficiency translation in vitro is proportional to a 5′ UTR length in the range of 17 to approximately 80 nts (Kozak, 1991a). However, this hypothesis remains preliminary, and more evidence from similar investigations of other living organisms is necessary.

Finally, the findings of our study may be useful for the identification of undiscovered genes of *T. gondii*. That is, the N-terminus signal peptide can be predicted from the genome. Experimentally obtained encyclopedic TSSs information is available (Tuda et al., 2010); typical 5′ UTR length was approximately 130 nts, as was demonstrated. If ORFs with predicted signal peptides are located close to 130 nts from TSSs, they can be transcribed and translated with high probability. Therefore, it is possible to consider that they are candidates of new genes if they have not yet been annotated. To validate this idea, all unassigned theoretical ORFs with more than 30 aa were examined to determine whether they had both the signal sequence at their 5′ end and proximal, transcriptionally active regions within 250 nts as well. In practice, 360,270 theoretical ORFs in *T. gondii* genome sequences were compared with 10,508 transcriptionally active regions identified in the previous study (Yamagishi et al., 2010). Results indicate that there are 146 ORFs that fit the criteria and were designated as unassigned gene candidates (Table I).

TABLE I. Identified unassigned gene candidates*.

Location of N-terminal†	TSS‡	Tag number§
X + 6943586	6943384	69968
VIII − 6167113	6167136	31728
XII + 595186	595087	29400
IX − 264010	264026	13530
VIII + 4420295	4420162	7514
VIIa − 3760364	3760478	3786
IV − 2405013	2405052	3406
VIII + 4753185	4753051	3267
IX + 5140576	5140511	3160
XI − 1007228	1007321	2869

* Ten candidates with the most TSS tags were selected and shown.
† Chromosome ID, orientation, and position, respectively.
‡ Position of representative transcription start site (TSS) in proximal transcriptionally active region.
§ Number of expression tags for proximal transcriptionally active region.

As suggested previously, computational estimation of a gene model in *T. gondii* is not highly reliable thus far; therefore, coordination among dry and wet analyses is vital to increase the reliability (Wakaguri et al., 2009). The findings of our study are considered as experimental evidence that would imbue increased reliability on the predicted gene model or improve them by providing an extra judgmental standard based on UTR.

We thank the Ministry of Education, Culture, Sports, Science, and Technology of Japan for financial support in KAKENHI (22780258) and the Global COE Program and the ''Asia-Africa Science and Technology Strategic Cooperation Promotion Program'' for Special Coordination Funds for the Promotion of Science and Technology schemes.

LITERATURE CITED

BENDTSEN, J. D., H. NIELSEN, G. VON HEIJNE, AND S. BRUNAK. 2004. Improved prediction of signal peptides: SignalP 3.0. Journal of Molecular Biology **340:** 783–795.

BURG, J. L., D. PERELMAN, L. H. KASPER, P. L. WARE, AND J. C. BOOTHROYD. 1988. Molecular analysis of the gene encoding the major surface antigen of *Toxoplasma gondii.* Journal of Immunology **141:** 3584–3591.

CHOO, K. H., T. W. TAN, AND S. RANGANATHAN. 2009. A comprehensive assessment of N-terminal signal peptides prediction methods. BMC Bioinformatics **10**(Suppl. 15)**:** S2.

DOBROWOLSKI, J. M., I. R. NIESMAN, AND L. D. SIBLEY. 1997. Actin in the parasite *Toxoplasma gondii* is encoded by a single copy gene, ACT1 and exists primarily in a globular form. Cell Motil Cytoskeleton **37:** 253–262.

DUBEY, J. P. 2009. Toxoplasmosis of animals and humans. 2nd ed. CRC Press, Boca Raton, Florida, p. 1–71.

EGEA, P. F., R. M. STROUD, AND P. WALTER. 2005. Targeting proteins to membranes: Structure of the signal recognition particle. Current Opinion in Structural Biology **15:** 213–220.

GAJRIA, B., A. BAHL, J. BRESTELLI, J. DOMMER, S. FISCHER, X. GAO, M. HEIGES, J. IODICE, J. C. KISSINGER, A. J. MACKEY ET AL. 2008. ToxoDB: An integrated *Toxoplasma gondii* database resource. Nucleic Acids Research **36:** D553–556.

HUGHES, T. A. 2006. Regulation of gene expression by alternative untranslated regions. Trends in Genetics **22:** 119–122.

KIBE, M. K., A. COPPIN, N. DENDOUGA, G. ORIA, E. MEURICE, M. MORTUAIRE, E. MADEC, AND S. TOMAVO. 2005. Transcriptional regulation of two stage-specifically expressed genes in the protozoan parasite *Toxoplasma gondii*. Nucleic Acids Research **33:** 1722–1736.

KOZAK, M. 1991a. Effects of long 5′ leader sequences on initiation by eukaryotic ribosomes in vitro. Gene Expression **1:** 117–125.

———. 1991b. Structural features in eukaryotic mRNAs that modulate the initiation of translation. Journal of Biological Chemistry **266:** 19867–19870.

MERCIER, C., S. LEFEBVRE-VAN HENDE, G. E. GARBER, L. LECORDIER, A. CAPRON, AND M. F. CESBRON-DELAUW. 1996. Common cis-acting elements critical for the expression of several genes of *Toxoplasma gondii*. Molecular Microbiology **21:** 421–428.

Pesole, G., F. Mignone, C. Gissi, G. Grillo, F. Licciulli, and S. Liuni. 2001. Structural and functional features of eukaryotic mRNA untranslated regions. Gene **276:** 73–81.

Tsuchihara, K., Y. Suzuki, H. Wakaguri, T. Irie, K. Tanimoto, S. Hashimoto, K. Matsushima, J. Mizushima-Sugano, R. Yamashita, K. Nakai et al. 2009. Massive transcriptional start site analysis of human genes in hypoxia cells. Nucleic Acids Research **37:** 2249–2263.

Tuda, J., A. E. Mongan, M. E. Tolba, M. Imada, J. Yamagishi, X. Xuan, H. Wakaguri, S. Sugano, C. Sugimoto, and Y. Suzuki. 2010. Full-parasites: Database of full-length cDNAs of Apicomplexa parasites, 2010 update. Nucleic Acids Research **39:** D625–631.

von Heijne, G. 1990. The signal peptide. Journal of Membrane Biology **115:** 195–201.

Wakaguri, H., Y. Suzuki, M. Sasaki, S. Sugano, and J. Watanabe. 2009. Inconsistencies of genome annotations in apicomplexan parasites revealed by 5′-end-one-pass and full-length sequences of oligo-capped cDNAs. BMC Genomics **10:** 312.

Yamagishi, J., H. Wakaguri, A. Ueno, Y. K. Goo, M. Tolba, M. Igarashi, Y. Nishikawa, C. Sugimoto, S. Sugano, Y. Suzuki et al. 2010. High-resolution characterization of *Toxoplasma gondii* transcriptome with a massive parallel sequencing method. DNA Research **17:** 233–243.

Date of Publication
Volume 98, No. 2, was mailed 20 April 2012