International Journal for Parasitology 45 (2015) 305–318

Contents lists available at ScienceDirect

International Journal for Parasitology

journal homepage: www.elsevier.com/locate/ijpara

Improving the gene structure annotation of the apicomplexan parasite *Neospora caninum* fulfils a vital requirement towards an in silico-derived vaccine

Stephen J. Goodswen [a,*], Joel L.N. Barratt [a], Paul J. Kennedy [b], John T. Ellis [a]

[a] School of Medical and Molecular Sciences, University of Technology Sydney (UTS), 15 Broadway, Ultimo, NSW 2007, Australia
[b] School of Software, Faculty of Engineering and Information Technology and the Centre for Quantum Computation and Intelligent Systems at the University of Technology Sydney (UTS), 15 Broadway, Ultimo, NSW 2007, Australia

ARTICLE INFO

Article history:
Received 24 September 2014
Received in revised form 12 January 2015
Accepted 12 January 2015
Available online 4 March 2015

Keywords:
Neospora caninum
Toxoplasma gondii
In silico vaccine discovery
Gene annotation
RNA-Seq
Comparative genomics

ABSTRACT

Neospora caninum is an apicomplexan parasite which can cause abortion in cattle, instigating major economic burden. Vaccination has been proposed as the most cost-effective control measure to alleviate this burden. Consequently the overriding aspiration for *N. caninum* research is the identification and subsequent evaluation of vaccine candidates in animal models. To save time, cost and effort, it is now feasible to use an in silico approach for vaccine candidate prediction. Precise protein sequences, derived from the correct open reading frame, are paramount and arguably the most important factor determining the success or failure of this approach. The challenge is that publicly available *N. caninum* sequences are mostly derived from gene predictions. Annotated inaccuracies can lead to erroneously predicted vaccine candidates by bioinformatics programs. This study evaluates the current *N. caninum* annotation for potential inaccuracies. Comparisons with annotation from a closely related pathogen, *Toxoplasma gondii*, are also made to distinguish patterns of inconsistency. More importantly, a mRNA sequencing (RNA-Seq) experiment is used to validate the annotation. Potential discrepancies originating from a questionable start codon context and exon boundaries were identified in 1943 protein coding sequences. We conclude, where experimental data were available, that the majority of *N. caninum* gene sequences were reliably predicted. Nevertheless, almost 28% of genes were identified as questionable. Given the limitations of RNA-Seq, the intention of this study was not to replace the existing annotation but to support or oppose particular aspects of it. Ideally, many studies aimed at improving the annotation are required to build a consensus. We believe this study, in providing a new resource on gene structure and annotation, is a worthy contributor to this endeavour.

1. Introduction

Neospora caninum is an obligate intracellular coccidian parasite of the phylum Apicomplexa (Dubey et al., 1988). Infection by *N. caninum* can cause the clinical disease neosporosis, characterised by neurological symptoms in dogs and abortion in cattle. Worldwide, the economic loss to both the dairy and beef industries due to outbreaks of abortion is estimated to exceed US$1 billion annually (Reichel et al., 2013). Results from economic analyses (Reichel and Ellis, 2006, 2008) propose that the most cost-effective approach to control neosporosis is a vaccine.

Toxoplasma gondii is also an apicomplexan pathogen and is responsible for birth defects in humans (Montoya and Liesenfeld, 2004). Moreover, it is an important model system for the phylum Apicomplexa (Kim and Weiss, 2004; Roos, 2005; Che et al., 2010). *Neospora caninum* is so morphologically and developmentally similar to *T. gondii* that, prior to 1988, infections by *N. caninum* were often wrongly identified as *T. gondii* (Dubey et al., 1988; Bjerkas and Dubey, 1991). Current opinion is that *N. caninum* and *T. gondii* have very similar genomes, with largely conserved gene content, with an almost one-to-one correspondence between most protein coding genes on the same chromosome i.e. they have shared synteny (Reid et al., 2012). This close phylogenetic relationship means that *T. gondii* genomic data can assist in the annotation of the *N. caninum* genome.

Researchers are beginning to capitalise on the vast potential of 'omics' data (e.g. genomes, transcriptomes and proteomes) to

* Corresponding author at: PO Box 123, Broadway, NSW 2007, Australia. Tel.: +61 2 9514 4161.
E-mail address: StephenJames.Goodswen@student.uts.edu.au (S.J. Goodswen).

http://dx.doi.org/10.1016/j.ijpara.2015.01.006
0020-7519/

Table 1
Type of evidence that supports the existence of *Neospora caninum* Liverpool strain and *Toxoplasma gondii* ME49 strain proteins. source: UniProtKB, October 2014.

Evidence for existence[a]	*N. caninum*[e]	*T. gondii*
Evidence at transcript level	0	1
Evidence at protein level[b]	2	6
Inferred from homology[c]	754	827
Predicted[d]	6447	7481
Total	7111	8315

[a] The 'protein existence' evidence does not give information on the accuracy or correctness of the sequence(s) displayed.
[b] There is experimental evidence for its existence e.g. partial or complete Edman sequencing, identification by MS, X-ray or NMR structure.
[c] The existence of a protein is probable because clear orthologues exist in closely related species.
[d] Used by default if a protein is without evidence at protein, transcript or homology levels.
[e] Proteins (4346) are annotated as 'Putative uncharacterized protein'. A further 1387 have 'Putative' in the annotation description and 57 protein sequences are incomplete (i.e. fragments).

further our understanding of *N. caninum* but especially to identify vaccine candidates via an in silico-derived approach (Goodswen et al., 2013a,b). Protein sequences are essential data for such an approach. The presence of signal peptides, transmembrane domains, epitopes and subcellular location can all be predicted, given accurate protein sequences (Krogh et al., 2001; Kall et al., 2004; Emanuelsson et al., 2007; Horton et al., 2007; Kim et al., 2011, 2012; Petersen et al., 2011). These predicted protein characteristics represent potential evidence from which a researcher can make an informed decision as to the suitability of a protein as a vaccine candidate (Leuzzi et al., 2006; Mora et al., 2006; Serino et al., 2006; Vivona et al., 2008).

Studies comparing isolates suggest that the *N. caninum* species consists of many diverse heterogeneous strains around the world (Al-Qassab et al., 2010), but almost all publicly available 'omics' data specific to *N. caninum* is for the Liverpool (NC-Liverpool) strain. Protein sequences for the NC-Liverpool strain are available both from the Universal Protein Resource knowledgebase (UniProtKB) (Consortium, 2014) and ToxoDB (Gajria et al., 2008). Table 1 shows the type of evidence annotated in UniProtKB that supports the existence of NC-Liverpool and *T. gondii* ME49 strain proteins. This illustrates that the vast majority of known protein sequences are derived from the translation of gene sequences i.e. the protein sequences are predicted and not experimentally derived. The primary source for the protein coding genes is from the Pathogen Sequencing Unit (PSU) at the Wellcome Trust Sanger Institute, UK (Reid et al., 2012). The PSU constructed a set of 14 pseudo-chromosomes for the NC-Liverpool genome by aligning 242 of 960 supercontigs to 14 publicly available *T. gondii* ME49 chromosomes based on predicted protein sequence similarity (Gajria et al., 2008). That is, approximately 90.4% of *N. caninum* chromosome sequences aligned successfully to *T. gondii* chromosomes. Of the remaining 718 contigs, any contigs that were of poor quality or <1 kb were abandoned. Only 343 of the 718 contigs were kept and grouped as unassigned contigs (Reid et al., 2012). To improve the genome annotation, the PSU sequenced the transcriptome of the invasive stage (tachyzoite) using mRNA sequencing (RNA-Seq). By combining ab initio gene predictors and RNA-Seq evidence, 7121 protein coding genes were identified. The current publicly available annotation that originated in 2011 from the Reid et al. study (referred to henceforth as the original study; Reid et al., 2012) has since remained essentially stagnant and unverified.

It is argued that there is currently no cost-effective, high-throughput laboratory technique to precisely sequence proteins. The two main direct methods, MS and Edman degradation, have limitations (Lubec and Afjehi-Sadat, 2007). RNA-Seq provides an indirect method, via mRNA transcripts, to obtain protein sequences i.e. the mRNA protein coding sequence (CDS) is translated to amino acids. One challenge, from a vaccine discovery perspective, is that the expression of *N. caninum* proteins is different at various stages of its life cycle and under altered environmental conditions such as interactions with a host during infection. Experimentally creating all relevant host-pathogen conditions to capture the entire complement of 'likely' vaccine candidates is beyond current technology. Nevertheless, RNA-Seq is a useful present day tool for predicting protein sequences from mRNAs, although it is not considered a perfect solution. Biases and errors can be introduced during RNA fragmentation and cDNA synthesis (Kassahn et al., 2011), and short length reads have considerable sequencing error rates (Haas and Zody, 2010). Furthermore, assembly is known to have computational challenges (Wang et al., 2009; Grabherr et al., 2011) increasing the potential for annotation errors.

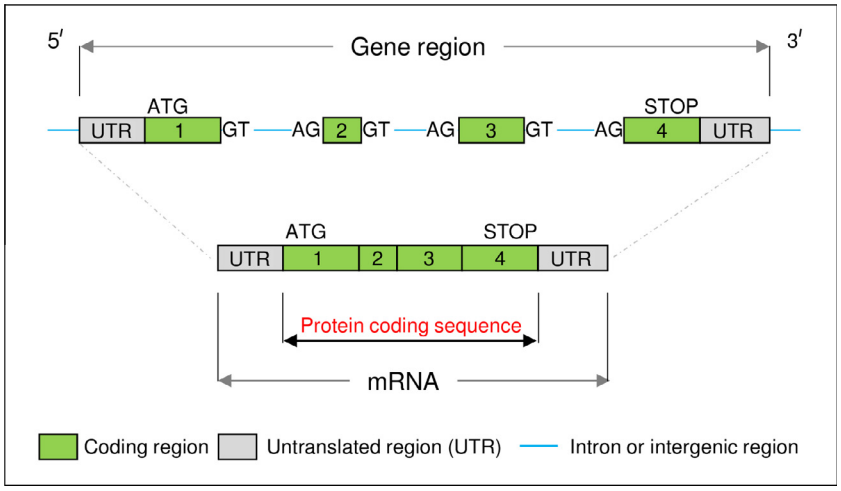

Fig. 1. Typical gene and mRNA structure in a eukaryote. UTR, untranslated region (start of 5' UTR for the gene region is the transcription start site); ATG, start codon (AUG is the correct codon for mRNA but ATG is commonly used in annotated mRNA sequences) – a start codon defines the start of the protein coding sequence i.e. the translation start site; GT, typical donor splice site; AG, typical acceptor splice site; STOP, one of three possible stop codons (TAA, TAG or TGA). Canonical intron splice sites observed in eukaryotic organisms are GT-AG, GC-AG and AT-AC. Note that an untranslated region of a gene can contain an intron(s) denoted with the same canonical intron splice sites.