DOI 10.1002/pmic.201400553 *Proteomics* 2015, *15*, 2618–2628

RESEARCH ARTICLE

A large-scale proteogenomics study of apicomplexan pathogens—*Toxoplasma gondii* and *Neospora caninum*

Ritesh Krishna[1,2], Dong Xia[2], Sanya Sanderson[2], Achchuthan Shanmugasundram[1,2], Sarah Vermont[2], Axel Bernal[3], Gianluca Daniel-Naguib[1], Fawaz Ghali[1], Brian P. Brunk[3], David S. Roos[3], Jonathan M. Wastling[2] and Andrew R. Jones[1]

[1] Institute of Integrative Biology, University of Liverpool, Liverpool, Merseyside, UK
[2] Institute of Infection and Global Health, University of Liverpool, Liverpool, Merseyside, UK
[3] Department of Biology, University of Pennsylvania, Philadelphia, PA, USA

Proteomics data can supplement genome annotation efforts, for example being used to confirm gene models or correct gene annotation errors. Here, we present a large-scale proteogenomics study of two important apicomplexan pathogens: *Toxoplasma gondii* and *Neospora caninum*. We queried proteomics data against a panel of official and alternate gene models generated directly from RNASeq data, using several newly generated and some previously published MS datasets for this meta-analysis. We identified a total of 201 996 and 39 953 peptide-spectrum matches for *T. gondii* and *N. caninum*, respectively, at a 1% peptide FDR threshold. This equated to the identification of 30 494 distinct peptide sequences and 2921 proteins (matches to official gene models) for *T. gondii*, and 8911 peptides/1273 proteins for *N. caninum* following stringent protein-level thresholding. We have also identified 289 and 140 loci for *T. gondii* and *N. caninum*, respectively, which mapped to RNA-Seq-derived gene models used in our analysis and apparently absent from the official annotation (release 10 from EuPathDB) of these species. We present several examples in our study where the RNA-Seq evidence can help in correction of the current gene model and can help in discovery of potential new genes. The findings of this study have been integrated into the EuPathDB. The data have been deposited to the ProteomeXchange with identifiers PXD000297and PXD000298.

Received: November 26, 2014
Revised: February 9, 2015
Accepted: April 9, 2015

Keywords:
Gene annotation / Microbiolgy / MS/MS / *N. Caninum* / Proteogenomics / *T. gondii*

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

Apicomplexa are obligate intracellular parasites that are of great medical and veterinary importance. Key members in the phylum include *Plasmodium falciparum*, the causative agent of malaria; *Toxoplasma gondii*, responsible for severe congenital defects and causing fatalities in immunocompromised patients; *Cryptosporidium*, causing waterborne diarrheal disease in humans; *Babesia*, tick transmitted hemoprotozoan parasites that cause disease in cattle, horses, dogs, and humans; and finally, *Eimeria* and *Theileria* that account for severe diseases of food-producing animals in the United Kingdom and Europe. Apicomplexans are unified by an apical complex consisting of a cluster of apical secretory organelles such as rhoptries and micronemes, an apical polar ring and in some species a polarized microtubule organizing center called the conoid [1]. Host cell invasion is a key event for

Correspondence: Dr. Andrew R. Jones, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom
E-mail: andrew.jones@liv.ac.uk

Abbreviations: 1DE SFIF, 1DE of soluble and insoluble fractions; **CSV**, comma separated values; **EST**, expressed sequence tags; **EuPathDB**, Eukaryotic Pathogen Database; **GFF**, generic file format; **MudPIT**, multidimensional protein identification technology

Colour Online: See the article online to view Figs. 1–3 in colour.

survival and replication of these parasites, which is conducted in a multistep process [2].

The genome information for these apicomplexan is maintained by the Eukaryotic Pathogen Database Resource Centre (EuPathDB: http://eupathdb.org) [3]. EuPathDB is an umbrella portal providing unified access to 11 specialized resources such as ToxoDB, PlasmoDB, etc. EuPathDB maintains genome sequences and annotations for these species along with a variety of pathogen genomics and functional genomics data and is an important resource for the apicomplexa research community. Among the first resources available at EuPathDB (formerly known as ApiDB back in 2005) was the genome of *Plasmodium falciparum*, which was the first apicomplexan genome to be sequenced [4]. This was followed by the sequencing efforts of many more apicomplexan species, including several other *Plasmodium* species [5–10], several strains of *T. gondii* [11], *Neospora caninum* [12], *Eimeria tenella* [10, 11], *Babesia bovis* [13], *Cryptosporidium* [12, 14, 15], and *Theileria* species [16, 17]. The efforts to sequence the *T. gondii* genome began in 2001 and ToxoDB release 2.0 had genome data for the ME49 and RH strains from a BAC clone-end sequencing project, an 8x random shotgun genome sequence, and expressed sequence tag (EST) assemblies [18]. The high-quality draft sequences and annotations for ME49, GT1, and VEG strains were completed by 2008 [11]. There are considerable manual curation efforts undertaken at EuPathDB and by the wider *Toxoplasma* community to produce high-quality gene models, as well as semi-automated reannotation based on experimental data. The high-quality genome of the closely related *Coccidian* parasite, *N. caninum* [12], has also been released recently, and the current *N. caninum* gene models were assembled using large-scale RNA-Seq datasets as a framework.

In this manuscript, we focus our attention on the current annotations of *T. gondii* and *N. caninum* only. We aim to study if proteomic evidence when used in conjunction with RNA-Seq evidence can help improve the current annotation of these species. The current gene model of *T. gondii* includes evidences from a number of previous MS-based studies performed to identify the proteome of *T. gondii*. These studies were typically based on the RH strain, which grows well and gives a high yield, and therefore is a well-accepted model for experimental studies. The first proteome studies either concentrated on studying a small number of tachyzoite proteins in detail by 2DE separation [19, 20] or on studying subproteomes of rhoptry organelles [21], the apical complex [22], and the excreted-secreted proteome [23] of *T. gondii*. The multiplatform (2D electrophoresis, gel-LC linked MS/MS, and MudPIT) analysis of *T. gondii* tachyzoite proteome by Xia et al. [24] was the first global proteome analysis of *T. gondii*. The peptide evidence from this study was instrumental in the refinement of annotation of the genome including correct assignment of exon–intron boundaries [24]. A proteomic analysis of cytosolic and membrane fractions of *T. gondii* tachyzoites and the validation of gene models (TigrScan, TwinScan, Glimmer, and ToxoDB Release 4, NCBI nonredundant protein database) with EST and peptide evidence data by Dybas et al. showed 31–42% false negative rate for various gene models [25]. These were followed by two independent analyses of the proteome of oocysts and sporozoites of *T. gondii* [26, 27]. All these data amount to a coverage of about 68% of predicted proteome of *T. gondii* [28]. The only proteome data publicly available to date for *N. caninum* include a gel electrophoresis based study identifying 26 differentially expressed proteins during tachyzoite to bradyzoite differentiation [29] and the analysis of the subproteome of rhoptry organelles [30, 31]. The RNA-Seq evidences on the other hand have recently become available and have been included in the current gene models of *T. gondii* and *N. caninum* at ToxoDB. There has been availability of some published [32, 33] and unpublished RNA-Seq datasets (http://toxodb.org/toxo/getDataset.do?display=detail) for both *T. gondii* and *N. caninum* that were used in the annotation.

Proteomics and RNA-Seq data provide complimentary evidence sets that can be exploited together for improvement in genome annotation. Next-generation sequencing techniques are already playing an increasingly important role in genome annotation. However, it is still difficult to unequivocally determine if a predicted gene actually produces a protein, or whether a predicted alternative splice of RNA is translated into protein. MS-derived proteomics data can play a direct role in genome annotation (proteogenomics), providing evidence that a given "official gene model" encodes a protein product, that a noncanonical (alternative) gene prediction is more likely to be correct, or that peptide evidence supports the discovery of new ORFs. Proteogenomics-based annotations utilize high-throughput MS/MS-based proteomic techniques to identify proteins present in a sample [34–36]. There are essentially four different ways of identifying peptide-spectrum matches from MS/MS: sequence database search [37, 38], de novo sequencing [39, 40], tag search [41, 42], and spectral library search [43, 44]. De novo sequencing does not require a sequence database to search against and thus could in theory be used to identify new peptides/proteins, but the error rates are generally too high for practical use in proteogenomics. The remaining three techniques are fundamentally dependent on the quality of the underlying database for identifying peptide sequences, with the sequence database search technique being the most common among them. The protein sequence databases are derived from gene annotations and a peptide sequence can only be identified by these methods if the corresponding gene sequence has been predicted correctly. The construction of a search database requires careful planning and one can simultaneously test a panel of gene models by combining evidences from independent computational and experimental sources. Gene models can differ from each other at a given locus in various ways, including prediction of start codon, presence–absence of coding regions, or precise intron–exon boundaries.

This manuscript takes a proteogenomics approach for studying the existing annotations of *T. gondii* and *N. caninum*, where we queried MS/MS datasets from eight

different experiments using sequence databases comprising official gene models and high-quality RNA-Seq-assisted predictions. For this study, we reanalyzed a number of published datasets for *T. gondii* [24] and generated new high-throughput MS/MS datasets for both *T. gondii* and *N. caninum*. Apart from the protein identifications, we show that how this proteogenomics approach can provide a useful contribution in improving genome annotations of these species and can lead to discovery of novel genes. The "official" gene models for *T. gondii* and *N. caninum* were obtained from the latest releases available at EuPathDB and the RNA-Seq predictions were produced by the data available at the same repository. The proteomic identifications against the official gene models provide evidence for genes that result into a protein product, establish the validity of putative genes, and confirm predicted splicing events. Identifications against the RNA-Seq-based models on the other hand identify putative novel genes that have RNA-Seq evidence to support them but are missing from the official release, differ in the splicing of exons, or show alternate splicing from the canonical prediction. For both the species, we identified a large number of proteins and several targets for improvement in the current available annotation. The datasets and results have been made available at ProteomeXchange [45] and EuPathDB.

2 Materials and methods

2.1 Sample preparation and datasets

2.1.1 *Toxoplasma gondii* datasets

Datasets of RH strain of *T. gondii* 1DE, 1DE of soluble and insoluble fractions (1DE SFIF), 2DE, and multidimensional protein identification technology (MudPIT) were created for a previously published *T. gondii* proteomics study [24]. Briefly, for 1DE, 1DE SFIF, and 2DE, *T. gondii* RH tachyzoites were separated either by 1D SDS-PAGE on a 12% v/v acrylamide gel or 2DE using pH 4–7 linear gradient and pH 3–10 nonlinear gradient strips. In total, 129 contiguous gel slices from the1DE experiment, 50 contiguous gel slices from the 1DE SFIF experiment, and 1217 gel spots from the 2DE experiments were collected and digested with trypsin. The peptide mixtures were then analyzed on an LC-MS/MS platform—an LTQ ion trap mass spectrometer (Thermo-Electron, Hemel Hempstead, UK) coupled online to a Dionex Ultimate 3000 (Dionex Company, Amsterdam, The Netherlands) HPLC system. For the *T. gondii* MudPIT experiment, five tris-soluble replicates and four tris-insoluble samples of *T. gondii* RH tachyzoite were each subjected to MudPIT analysis using a quaternary Agilent 1100 series HPLC coupled to an LTQ-ion trap mass spectrometer (Thermo, San Jose, CA, USA) with a nano-LC ESI source. Datasets of *T. gondii* "Orbitrap whole cell lysate" and "Orbitrap 1DE" were newly created for this study using previously published sample preparation and MS protocols [46]. These datasets were generated from RH parasites in order to be consistent with the previously published datasets. In the "Orbitrap whole cell lysate" experiment, *T. gondii* RH tachyzoite proteins were solubilized and tryptically digested. In the Orbitrap 1DE experiment, *T. gondii* RH tachyzoites were separated by 1D SDS-PAGE on a 12% v/v acrylamide gel, from which 16 gel bands were excised and digested with trypsin. The digests were then pooled into eight samples for LC-MS/MS analysis. Peptide mixtures from both experiments were analyzed by online nanoflow LC using the nanoACQUITY-nLC system (Waters MS technologies, Manchester, UK) coupled to an LTQ-Orbitrap Velos (ThermoFisher Scientific, Bremen, Germany) mass spectrometer equipped with the manufacturer's nanospray ion source.

2.1.2 *Neospora caninum* datasets

Datasets of *N. caninum* Orbitrap 1DE and Orbitrap 1DE-2 were collected from *N. caninum* LIV strain tachyzoites using the similar protocol described in *T. gondii* datasets. Briefly, 8 and 14 gel slices were, respectively, excised from Orbitrap 1DE and Orbitrap 1DE-2 experiments. Samples were then tryptically digested and analyzed by online nanoflow LC using the nanoACQUITY-nLC system (Waters MS technologies) coupled to an LTQ-Orbitrap Velos (ThermoFisher Scientific) mass spectrometer equipped with the manufacturer's nanospray ion source.

It is important to note that all datasets were generated from single life cycle stages of the parasites, where not all proteins are expected to be expressed.

2.2 Proteogenomic analysis pipeline

The proteogenomics analysis of the above dataset was performed by an automated software pipeline, the ProteoAnnotator (http://www.proteoannotator.org/), developed in our group. ProteoAnnotator is an open-source, platform-independent, toolkit for proteogenomics, which is compliant with formats from the Proteomics Standards Initiative [47, 48]. The pipeline starts with genomic coordinates for gene models (GFF3 (generic file)) format, and mass spectra in MGF (MASCOT Generic) format. ProteoAnnotator embeds open-source search engines OMSSA [49] and X!Tandem [50], which are accessed via the SearchGUI wrapper [51]. Postprocessing is performed to control for FDR at the peptide level, followed by a bespoke protein inference algorithm/scoring approach, which is able to score the strength of evidence for improvements to gene models at a given locus and assemble peptides into protein groups. For general identifications of protein groups—1% FDR is implemented. For (protein group) identifications that suggest improvements to the current gene models, 5% FDR threshold is applied—as discussed in the ProteoAnnotator publication [47].

2.3 MS/MS search parameters

2.3.1 Search parameters

The search parameters for all the datasets were fixed modification of carbamidomethylation of cysteine, variable modification of acetylation of the protein n-terminus, pyro-glu from n-terminal E, pyro-glu from n-terminal Q, and oxidation of methionine. A double missed trypsin cleavage was allowed. The product tolerance was set as ±0.5 Da and the precursor tolerance as 5 ppm and ±0.5 Da for Orbitrap and LTQ, respectively.

2.3.2 Gene models

The MS data for *T. gondii* and *N. caninum* were searched against the protein database assembled from two different sources: the official gene models and predicted gene models supported by RNA-Seq evidences. The official gene models (release 10 - January 30, 2014) were obtained from EuPathDB. The gene model for *T. gondii* was obtained from http://toxodb.org/common/downloads/release-10.0/TgondiiME49/gff/data/ and the *N. caninum* official model was obtained from http://toxodb.org/common/downloads/release-10.0/NcaninumLIV/gff/data/. At the time of submission of this manuscript, the latest official gene models are available as release 12 (September 10, 2014). However, releases 10 and 12 are identical in terms of protein counts and sequences. The RNA-Seq-assisted alternate gene model was generated using a version of CRAIG (default parameters) [52] that integrates RNA-Seq data available at http://toxodb.org/. The RNA-Seq data were encoded as features derived from the genomic mapping of RNA-Seq read libraries as performed by GSNAP version 2013-08-19 (default parameters) [53]. Following the same basic strategy as in [52], we obtained a set of gene predictions for each existing library. We sought completeness by forcing CRAIG to predict at least one gene model for each *biologically significant* junction. Junctions with conflicting evidence, that is junctions that overlap other junctions, are deemed biologically significant when their support is at least 20% of the highest support found in any overlapping junction. Junctions with no conflicts are biologically significant when they are supported by at least three reads mapping to the region. The RNA-Seq compliant version of CRAIG is unpublished at the time of preparing this manuscript, but the binaries for the current version are available at http://roos8core.bio.upenn.edu/~abernal/craig/craig-2.1.tar.gz. The search database was prepared by concatenating these gene models to the official gene models. The database was further appended by decoy sequences with a target–decoy ratio of 1:1 to create a final search database for performing the MS/MS searches. We have provided these databases in the ProteomeXchange submissions.

2.3.3 Functional analysis

Sequences from identified RNA-Seq-derived gene model with additional peptide evidence as well as overlapping official gene models were analyzed using InterProScan 5 [54]. Examples where additional domains and functional signatures identified in the RNA-Seq-derived gene model were further analyzed and annotated using Clustal Omega [55] and ToxoDB.

3 Results and discussion

Table 1 presents the composition of each gene model in terms of number of predicted gene sets and the total amino acid count. It also presents the total number of proteins identified by our analysis of these datasets for each set of gene models, summarized across all experiments. In order to count the total number of proteins, we selected only a single *representative protein* from grouped isoforms, where ambiguity was present in peptide to protein inference as per the ProteoGrouper algorithm in [47]. To our knowledge, the total number of official proteins identified for both *T. gondii* and *N. caninum* is the largest set of proteins with MS/MS evidence reported in any single study. Our analysis identifies ~35% of the total protein coding genes and ~50% of the total proteins with proteomic evidence reported for ME49 (http://toxodb.org/toxo/showRecord.do?name=OrganismRecordClasses.OrganismRecordClass&source_id=tgonME49&project_id=ToxoDB). There is no proteomics evidence available for *N. caninum*, so this is the first study where we report 1273 proteins with MS/MS evidence (http://toxodb.org/toxo/showRecord.do?name=OrganismRecordClasses.OrganismRecordClass&source_id=ncanLIV&project_id=ToxoDB). Our analysis also suggests that there are 289 and 140 potential loci for *T. gondii* and *N. caninum*, respectively, that have support from RNA-Seq and proteomics data but have incorrect gene models or are missing from the current annotation. When further stringent thresholding (5% FDR) is applied, there are 191 and 101 proteins identified from the RNA-Seq-derived gene models for *T. gondii* and *N. caninum*, respectively, as discussed in "Identifications From the RNA-Seq models."

The complete list of identified proteins, along with corresponding peptide hits and locations, is provided in the result files available at ProteomeXchange. Additionally, all results have been loaded into EuPathDB, aligned against the most recent genome release. An example region on chromosome VIII can be viewed at: http://toxodb.org/cgi-bin/gbrowse/toxodb/?start=5924028;stop=5974027;ref=TGME49_chrVIII;width=800;version=100;flip=0;grid=1;id=d6546393407f30e0e323ffd34a77b61a;l=peptide_ToxoDB_18db2e_http_www.proteoannotator.org_datasets_Toxo10_official_peptides.gff_1%1EGene—where datasets can be viewed alongside RNA-Seq and other large-scale datasets (On GBrowse—Click "Select Tracks," "Supplementary

Table 1. The composition of gene models and the number of representative proteins identified in total across our full data collection

Species	Gene model	Total database entries	Total amino acid count	Representative proteins identified as group leaders[a)]	Alternate loci with q-value < 0.05
T. gondii	Official	8322	6 669 204	2921	0
	RNA-Seq	86 699	37 847 722	289	191
N. caninum	Official	7122	6 054 032	1273	0
	RNA-Seq	13 777	8 158 875	140	101

[a)] A "representative protein" can encompasses more than one record (protein) from the protein database, incorporating the set of proteins that share the same set or subset of peptide identifications to avoid double counting of proteins with no independent evidence.

Data Provided by Users"—All On). Alternatively, the results can be loaded to any desktop-based genome browser by loading the GFF3 file from http://www.proteoannotator.org/datasets/Toxo10_official_peptides.gff.

3.1 Identifications from the RNA-Seq models

The conflict where a PSM identified a peptide that belonged to both the official and RNA-Seq models was resolved by using the ProteoGrouper algorithm in ProteoAnnotator, in which identifications are assembled into protein groups. The protein groups are assigned two types of score (see [47]): (i) the overall score for the (protein ambiguity) group (*PAG Score*) and (ii) the strength of evidence supporting identified loci not from the official gene models (noncanonical gene model score). In the case of a tie between an official model and an identification from the alternative (RNA-Seq) models (same set of peptides), the protein group leader is by default assigned as the official model and the *noncanonical gene model score* = 0. The sum of peptide scores uniquely assigned to the group is used to generate the *PAG Score*. The protein group list is ordered by *PAG Score* (including targets and decoys), and a 1% FDR (protein group-level) threshold is applied. This basis was used to generate the figures in Table 1 for official gene models. For any groups where peptides have uniquely been assigned to an RNA-Seq-derived protein, the *noncanonical gene model score* is the sum of all peptide scores within the group assigned to models other than the official gene models. All "decoy" peptides are by default not assigned to an official gene model, and thus contribute to the *noncanonical gene model score* for the corresponding decoy protein group list. As such, all decoy peptides/proteins have *noncanonical gene model score* >0. The protein group list is separately ordered by *noncanonical gene model score* and a 5% FDR (q-value) threshold is applied (since this is conservatively calculated from all decoy peptides/proteins, this step ensured that the weak identifications from the alternate models are filtered out—e.g. those supported by only a single weak peptide identification matched to one RNA-Seq model), giving a list of significant alternate loci supporting evidence for improvement in the genome annotation. The q-value threshold at 5% resulted in identification of 191 and 101 proteins from the RNA-Seq-derived gene model for *T. gondii* and *N. caninum*, respectively. There are number of proteins identified with high *noncanonical gene model scores* that have not been assigned a chromosome, for example tgondii-rna.Reid_tachy.day3.gene1112 (5 noncanonical peptides), tgondii-rna.Reid_tachy.day4.gene797 (17 noncanonical peptides), tgondii-rna.Saeij_Jeroen_strains.COUGAR.gene1934 (15 noncanonical peptides), etc. These identifications are examples of those cases that are missing from the official annotation and are potential candidates for new genes.

Figures 1–3 present three different cases selected from the list of alternate identifications that suggest various improvements for the official gene model, including a different start codon, an extended gene model, and a different splicing site.

Figure 1 presents a case of different start site picked by our analysis from two different datasets where two genes, TGME49_269442 and TGME49_269438, on chromosome VIII are in the immediate proximity of each other. The last exon of TGME49_269442 is within 116 base pairs of the first exon of TGME49_269438. The predicted gene model suggests a single transcript—tgondii-rna.Saeij_Jeroen_strains.CEPdelta.gene1305—that spans across the boundaries of both the genes. The presence of peptide evidence for this transcript suggests that these gene boundaries may be wrong. The official gene is annotated as putative calmodulin, a calcium-binding messenger protein. Calcium plays a critical role in several parasite-specific functions including host cell invasion and egress [56]. As shown in Fig. 1C, domain analysis suggested that, with a different starting site, the RNA-Seq-derived model possesses a complete EF-hand domain pair where the official model only includes a single EF-hand unit. Pairing of EF hands is thought to stabilize the protein and increase the affinity toward calcium [57], the RNA-Seq-derived model here may present a biologically more complete calmodulin protein with better calcium-binding efficiency.

A case of an extended gene model is shown in Fig. 2 where the official gene model for TGME49_324800 has no domain detected in domain analysis, but has been annotated as tryptophanyl-tRNA synthetase. The RNA-Seq-derived model has both Threonyl/alanyl tRNA synthetase, class II-like, putative editing domain (in red, position 12–139, IPR018163), and a Rossmann-like alpha/beta/alpha

Figure 1. Peptide evidence indicating an alternate start codon. RNA-Seq-derived gene model tgondii-rna.Saeij_Jeroen_strains.CEPdelta.gene1305 was identified in our analysis, which has a different starting site to the official gene model TGME49_269442. (A) GBrowse screenshot of the alignment of the official gene model, RNA-Seq-derived gene model, and peptides identified. (B) Sequence alignment of official and RNA-Seq-derived gene model. (C) Comparison of InterProScan results between official and RNA-Seq-derived models where a complete EF-hand domain pair was identified in the RNA-Seq-derived model.

sandwich fold (IPR014729) detected. The official model lies between the end of the first domain and beginning of the second domain, which might be the reason for the current product name annotation. With an extended gene structure, the RNA-Seq-derived model has two full domains to support the product name and to make the annotation more complete.

Our analysis has also detected cases for different splice sites, an example for gene TGVEG_295125 is shown in Fig. 3. TGVEG_295125 is annotated as rhoptry protein ROP4 with two exons (see Fig. 3A). However, sequence alignment has shown that an ORF (ToxoDB: KI544509-5-1423859-1422096) can be found to expand the full length of the gene, including the intron region. In fact, 17 nonofficial peptides were identified from five datasets that mapped to the intron region of the official gene model (Fig. 3B), which suggested that RNA-Seq-derived model tgondii-rna.Saeij_Jeroen_strains.VEG.gene444 represents the actual protein expressed in this region. Interestingly, tgondii-rna.Saeij_Jeroen_strains.VEG.gene444 aligns perfectly with ROP4 gene sequence deposited in GenBank (version: AAU87405.1; GI: 52788873), which was independently identified by cross-reacting mAbs [58]. Together, our analysis provides evidence that TGVEG_295125 should be annotated as a single exon gene with no splice site in the middle.

Figure 2. Peptide evidence indicating a suggested extension to the official gene model. RNA-Seq-derived gene model tgondii-rna.Saeij_Jeroen_strains.TgCATBr5.gene3708 was identified in our analysis that has a different starting site to the official gene model TGME49_324800. (A) Sequence alignment of official and RNA-Seq-derived gene model where peptides identified in both models are colored in blue and peptides identified only in RNA-Seq-derived model are colored in red. (B) Results from InterProScan indicate relevant domains detected in the RNA-Seq-derived gene model that are missing in the official gene model that has been annotated as tryptophanyl-tRNA synthetase.

The expression evidence of alternate gene structures provided by our analysis allowed us to identify new functional features in a gene. For example, a TRAM/LAG1/CLN8 homology domain was identified in RNA-Deq derived model tgondii-rna.Gregory_VEG_mRNA.hour16.gene5326, suggesting additional functions of the official gene TGME49_295080, which is annotated as a hypothetical protein. Signal peptides were also identified in RNA-Seq-derived models, which were missing in the equivalent official gene models, such as tgondii-rna.Saeij_Jeroen_strains.BOF.gene2384 versus TGME49_214220 and tgondii-rna.Gregory_VEG_mRNA .hour16.gene1072 versus TGME49_254470, both of which were annotated as hypothetical proteins. These additional features identified in our study would provide valuable information for functional annotation of the genome.

3.2 Description of available result files

The MS/MS datasets and result files are available at the ProteomeXchange Consortium with the dataset identifiers PXD000297 and PXD000298. There are six types of data files submitted at ProteomeXchange for both *T. gondii* and *N. caninum* datasets. These files are of types—raw MSMS files (.raw), peak list files (.mgf), search results in native OMSSA and X!Tandem formats (.omx and .xml), search results in mzIdentML format with various scores (.mzid), annotation files with integrated peptide evidences (.gff), and summary files (.csv). Each file name starts with a prefix representing the dataset it belongs to. The linkage is also available via ProteomeXchange interface. The datasets for *T. gondii* are denoted by self-explanatory prefixes—1DE, 2DE, MPIT, Orb-1DE, and Orb-WL. The datasets for *N. caninum* are prefixed by NORB-1 and NORB-2 for Orbitrap 1DE and Orbitrap 1DE-2 datasets, respectively. The GFF and CSV (comma separated values) files are of direct relevance to annotators. The GFF files exist in two forms, one for the peptides mapped on the official model, and another for the peptides mapped on the RNA-Seq model. The GFFs that correspond to the official model are named with a suffix "_gff_A," whereas the GFFs corresponding to the RNA-Seq

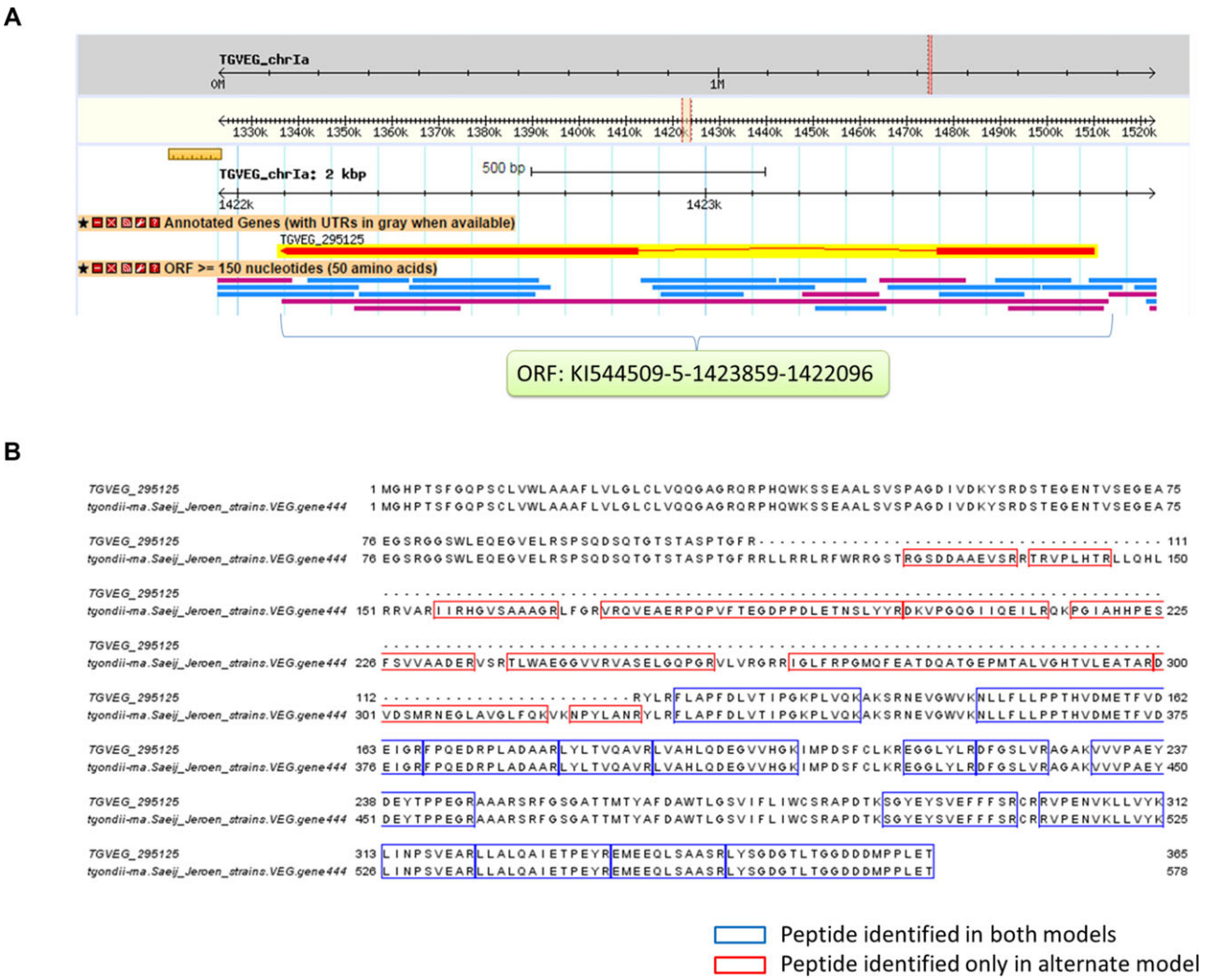

Figure 3. Peptide evidence indicating a different splicing site to the official gene model. RNA-Seq-derived gene model tgondii-rna.Saeij_Jeroen_strains.VEG.gene444 was identified in our analysis that has a different starting site to the official gene model TGVEG_295125. (A) GBrowse screenshot showing ORF KI544509-5-1423859-1422096 expands the intron region of the official gene model. (B) Sequence alignment of official and RNA-Seq-derived gene model where peptides identified in both models are colored in blue and peptides identified in the intron region of official gene model are colored in red.

model are named with a suffix "gff_B," in same manner as our search database was prepared. As an example, the GFFs produced by searching the 1DE dataset for *T. gondii* are named as 1DE_mapped_gff_A.gff (for official model) and 1DE_mapped_gff_B.gff (for RNA-Seq model). These GFFs can be directly loaded in a genome browser of choice, or can be uploaded at the GBrowse custom track interface of ToxoDB (http://toxodb.org/cgi-bin/gbrowse/toxodb/). There are five CSV files for each dataset and are prefixed in the same manner as above to identify the dataset they belong to. The type of information in the CSV files can be identified by the particular suffixes the files have in their names as described in [47]. A list of representative proteins from each experiment and the final counts are provided in the Supporting Information Files 1 and 2 for *T. gondii* and *N. caninum*, respectively. The identifications from the RNA-Seq-derived model can be loaded at the GBrowse interface of ToxoDB using Supporting Information File 3. The file contains the predicted transcripts and identified peptides. The InterProScan results for the RNA-Seq-derived identifications are listed in Supporting Information File 4. A genome-wide distribution plot of suggested correction sites for *T.gondii* can be seen in Supporting Information File 5.

4 Concluding remarks

MS-based proteomics has an important role to play in successful annotation of gene models. Our analysis of *T. gondii* and *N. caninum* not only identifies many known

protein-coding genes but also provides a framework for validation of hypothetical and putative proteins. Our study provides evidence that further improvements to the existing gene annotation models for these species are possible, and helps in pin-pointing the locations in the genome, which may require a revision. The inclusion of RNA-Seq predictions in our proteogenomics approach enables us to harness the combined power of genomics and proteomics data for discovery of new genes that are not part of the current annotation. We anticipate that the inclusion of this evidence into the routine curation process will assist in the generation of accurate gene models released at the genome databases for these important species.

An important aspect of our approach is the reproducibility of the entire process facilitated by the availability of the ProteoAnnotator software toolkit. The tools and results developed for this study have been installed locally at the EuPathDB database and will shortly be used as part of a standard build pipeline for adding proteomics data to the various databases held within different subsites. The findings of our study for both *T. gondii* and *N. caninum* genomes are also integrated on the ToxoDB and are part of standard query interface. The full dataset and result files are available in public domain at ProteomeXchange.

A.R.J. and J.M.W. would like to acknowledge funding from BBSRC (BB/G010781/1 to A.R.J and J.M.W. and BB/H024654/1, BB/L024128/1 to A.R.J.). A.S., G.N.-D. and S.V. are funded by BBSRC DTG studentships. Haiming Wang facilitated the GBrowse display of data at EuPathDB. The MS proteomics data in this paper have been deposited in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository [59]: dataset identifiers PXD000297 and PXD000298.

The authors have declared no conflict of interest.

5 References

[1] Morrissette, N. S., Sibley, L. D., Cytoskeleton of apicomplexan parasites. *Microbiol. Mol. Biol. Rev* 2002, *66*, 21–38.

[2] Sharma, P., Chitnis, C. E., Key molecular events during host cell invasion by Apicomplexan pathogens. *Curr. Opin. Microbiol.* 2013, *16*, 432–437.

[3] Aurrecoechea, C., Barreto, A., Brestelli, J., Brunk, B. P. et al., EuPathDB: the eukaryotic pathogen database. *Nucleic Acids Res.* 2013, *41*, D684–D691.

[4] Gardner, M. J., Hall, N., Fung, E., White, O. et al., Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 2002, *419*, 498–511.

[5] Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T. W. et al., Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* 2002, *419*, 512–519.

[6] Carlton, J. M., Adams, J. H., Silva, J. C., Bidwell, S. L. et al., Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* 2008, *455*, 757–763.

[7] Carlton, J., Silva, J., Hall, N., The genome of model malaria parasites, and comparative genomics. *Curr. Issues Mol. Biol.* 2005, *7*, 23–37.

[8] Pain, A., Bohme, U., Berry, A. E., Mungall, K. et al., The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* 2008, *455*, 799–803.

[9] Aurrecoechea, C., Brestelli, J., Brunk, B. P., Dommer, J. et al., PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.* 2009, *37*(Suppl 1), D539–D543.

[10] Logan-Klumpler, F. J., DeSilva, N., Boehme, U., Rogers, M. B. et al., GeneDB—an annotation database for pathogens. *Nucleic Acids Res.* 2011, *40*(Database issue): D98–D108.

[11] Gajria, B., Bahl, A., Brestelli, J., Dommer, J. et al., ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Res.* 2008, *36*(Suppl 1), D553–D556.

[12] Reid, A. J., Vermont, S. J., Cotton, J. A., Harris, D. et al., Comparative genomics of the apicomplexan parasites *Toxoplasma gondii* and *Neospora caninum*: coccidia differing in host range and transmission strategy. *PLoS Pathog.* 2012, *8*, e1002567.

[13] Brayton, K. A., Lau, A. O. T., Herndon, D. R., Hannick, L. et al., Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. *PLoS Pathog.* 2007, *3*, e148.

[14] Abrahamsen, M. S., Templeton, T. J., Enomoto, S., Abrahante, J. E. et al., Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* 2004, *304*, 441–445.

[15] Heiges, M., Wang, H., Robinson, E., Aurrecoechea, C. et al., CryptoDB: a Cryptosporidium bioinformatics resource update. *Nucleic Acids Res.*, *34*(Suppl 1), D419–D422.

[16] Gardner, M. J., Bishop, R., Shah, T., deVilliers, E. P. et al., Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science* 2005, *309*, 134–137.

[17] Pain, A., Renauld, H., Berriman, M., Murphy, L. et al., Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. *Science* 2005, *309*, 131–133.

[18] Kissinger, J. C., Gajria, B., Li, L., Paulsen, I. T., Roos, D. S., ToxoDB: accessing the *Toxoplasma gondii* genome. *Nucleic Acids Res.* 2003, *31*, 234–236.

[19] Cohen, A. M., Rumpel, K., Coombs, G. H., Wastling, J. M., Characterisation of global protein expression by two-dimensional electrophoresis and mass spectrometry: proteomics of *Toxoplasma gondii*. *Int. J. Parasitol.* 2002, *32*, 39–51.

[20] Lee, E. G., Kim, J. H., Shin, Y. S., Shin, G. W. et al., Application of proteomics for comparison of proteome of *Neospora caninum* and *Toxoplasma gondii* tachyzoites. *J. Chromatogr. B* 2005, *815*, 305–314.

[21] Bradley, P. J., Ward, C., Cheng, S. J., Alexander, D. L. et al., Proteomic analysis of rhoptry organelles reveals many novel constituents for host-parasite interactions in *Toxoplasma gondii*. *J. Biol. Chem.* 2005, *280*, 34245–34258.

[22] Hu, K., Johnson, J., Florens, L., Fraunholz, M. et al., Cytoskeletal components of an invasion machine—the apical complex of *Toxoplasma gondii*. *PLoS Pathog.* 2006, *2*, e13.

[23] Zhou, X. W., Kafsack, B. F. C., Cole, R. N., Beckett, P. et al., The opportunistic pathogen *Toxoplasma gondii* deploys a diverse legion of invasion and survival proteins. *J. Biol. Chem.* 2005, *280*, 34233–34244.

[24] Xia, D., Sanderson, S. J., Jones, A. R., Prieto, J. H. et al., The proteome of *Toxoplasma gondii*: integration with the genome provides novel insights into gene expression and annotation. *Genome Biol.* 2008, *9*, R116.

[25] Dybas, J. M., Madrid-Aliste, C. J., Che, F.-Y., Nieves, E. et al., Computational analysis and experimental validation of gene predictions in *Toxoplasma gondii*. *PLoS One* 2008, *3*, e3899.

[26] Fritz, H. M., Bowyer, P. W., Bogyo, M., Conrad, P. A., Boothroyd, J. C., Proteomic analysis of fractionated Toxoplasma oocysts reveals clues to their environmental resistance. *PloS One* 2012, *7*, e29955.

[27] Possenti, A., Fratini, F., Fantozzi, L., Pozio, E. et al., Global proteomic analysis of the oocyst/sporozoite of *Toxoplasma gondii* reveals commitment to a host-independent lifestyle. *BMC Genomics* 2013, *14*, 183.

[28] Wastling, J. M., Armstrong, S. D., Krishna, R., Xia, D., Parasites, proteomes and systems: has Descartes' clock run out of time? *Parasitology* 2012, *139*, 1103–1118.

[29] Marugán-Hernández, V., Álvarez-García, G., Risco-Castillo, V., Regidor-Cerrillo, J., Ortega-Mora, L. M., Identification of *Neospora caninum* proteins regulated during the differentiation process from tachyzoite to bradyzoite stage by DIGE. *Proteomics* 2010, *10*, 1740–1750.

[30] Marugán-Hernández, V., Álvarez-García, G., Tomley, F., Hemphill, A. et al., Identification of novel rhoptry proteins in *Neospora caninum* by LC/MS-MS analysis of subcellular fractions. *J. Proteomics* 2011, *74*, 629–642.

[31] Sohn, C. S., Cheng, T. T., Drummond, M. L., Peng, E. D. et al., Identification of novel proteins in *Neospora caninum* using an organelle purification and monoclonal antibody approach. *PLoS One* 2011, *6*, e18383.

[32] Reid, A. J., Vermont, S. J., Cotton, J. A., Harris, D. et al., Comparative genomics of the apicomplexan parasites *Toxoplasma gondii* and *Neospora caninum*: coccidia differing in host range and transmission strategy. *PLoS Pathog.* 2012, *8*, e1002567.

[33] Buchholz, K. R., Fritz, H. M., Chen, X., Durbin-Johnson, B. et al., Identification of tissue cyst wall components by transcriptome analysis of in vivo and in vitro *Toxoplasma gondii* bradyzoites. *Eukaryot. Cell* 2011, *10*, 1637–1647.

[34] Yates, J. R., 3rd, Eng, J. K., McCormack, A. L., Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal. Chem.* 1995, *67*, 3202–3210.

[35] Yates, J. R., 3rd, McCormack, A. L., Eng, J., Mining genomes with MS. *Anal. Chem.* 1996, *68*, 534A–540A.

[36] Armengaud, J., A perfect genome annotation is within reach with the proteomics and genomics alliance. *Curr. Opin. Microbiol.* 2009, *12*, 292–300.

[37] Nesvizhskii, A. I., Protein identification by tandem mass spectrometry and sequence database searching. *Methods Mol. Biol.* 2007, *367*, 87–119.

[38] Kapp, E., Schutz, F., Overview of tandem mass spectrometry (MS/MS) database search algorithms. *Curr. Protoc. Protein Sci.* 2007, *49*, 25.2.1–25.2.19.

[39] Allmer, J., Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Exp. Rev. Proteomics* 2011, *8*, 645–657.

[40] Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E., Pevzner, P. A., De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 1999, *6*, 327–342.

[41] Frank, A., Tanner, S., Bafna, V., Pevzner, P., Peptide sequence tags for fast database search in mass-spectrometry. *J. Proteome Res.* 2005, *4*, 1287–1295.

[42] Tanner, S., Shu, H., Frank, A., Wang, L. C. et al., InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* 2005, *77*, 4626–4639.

[43] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K. et al., Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* 2008, *5*, 873–875.

[44] Mann, M., Kelleher, N. L., Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. USA* 2008, *105*, 18132–18138.

[45] Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A. et al., ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 2014, *32*, 223–226.

[46] Jackson, A. P., Otto, T. D., Darby, A., Ramaprasad, A. et al., The evolutionary dynamics of variant antigen genes in Babesia reveal a history of genomic innovation underlying host-parasite interaction. *Nucleic Acids Res.* 2014, *42*, 7113–7131.

[47] Ghali, F., Krishna, R., Perkins, S., Collins, A. et al., ProteoAnnotator—open source proteogenomics annotation software supporting PSI standards. *Proteomics* 2014, *14*, 2731–2741.

[48] Ghali, F., Krishna, R., Lukasse, P., Martinez-Bartolome, S. et al., A toolkit for the mzIdentML standard: the ProteoIDViewer, the mzidLibrary and the mzidValidator. *Mol. Cell Proteomics* 2013, *12*, 3026–3035.

[49] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L. et al., Open mass spectrometry search algorithm. *J. Proteome Res.* 2004, *3*, 958–964.

[50] Craig, R., Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, *20*, 1466–1467.

[51] Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A., Martens, L., SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* 2011, *11*, 996–999.

[52] Bernal, A., Crammer, K., Hatzigeorgiou, A., Pereira, F., Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Comput. Biol.* 2007, *3*, e54.

[53] Wu, T. D., Nacu, S., Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010, *26*, 873–881.

[54] Jones, P., Binns, D., Chang, H. Y., Fraser, M. et al., InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014, *30*, 1236–1240.

[55] Sievers, F., Wilm, A., Dineen, D., Gibson, T. J. et al., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 2011, *7*, 539.

[56] Arrizabalaga, G., Boothroyd, J. C., Role of calcium during *Toxoplasma gondii* invasion and egress. *Int. J. Parasitol.* 2004, *34*, 361–368.

[57] Seamon, K. B., Kretsinger, R. H., Calcium-modulated proteins. *Met. Ions Biol.* 1983, *6*, 1–51.

[58] Carey, K. L., Jongco, A. M., Kim, K., Ward, G. E., The *Toxoplasma gondii* rhoptry protein ROP4 is secreted into the parasitophorous vacuole and becomes phosphorylated in infected cells. *Eukaryot. Cell* 2004, *3*, 1320–1330.

[59] Vizcaíno, J. A., Côté, R. G., Csordas, A., Dianes, J. A. et al., The *PRoteomics IDEntifications (PRIDE)* database and associated tools: status in 2013. *Nucleic Acids Res.* 2013, *41*, D1063–D1069.

