Published in final edited form as:

Expert Rev Proteomics. 2009 June ; 6(3): 303–313. doi:10.1586/epr.09.16.

***Toxoplasma gondii* proteomics**

Louis M Weiss, MD, MPH[†],

Department of Medicine (Division of Infectious Diseases) and Department of Pathology (Division of Parasitology and Tropical Medicine), Albert Einstein College of Medicine, 1300 Morris Park Avenue, Forchheimer 504, 1300 Morris Park Avenue, Bronx, NY 10461, USA Tel.: +1 718 430 2142 Fax: +1 718 430 8543 lmweiss@aecom.yu.edu

Andras Fiser, PhD,

Department Systems and Computational Biology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Forchheimer 504, 1300 Morris Park Avenue, Bronx, NY 10461, USA

Ruth Hogue Angeletti, PhD, and

Department Cell Biology and Laboratory for Macromolecular Analysis and Proteomics, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Forchheimer 504, 1300 Morris Park Avenue, Bronx, NY 10461, USA

Kami Kim, PhD

Department of Medicine (Division of Infectious Diseases) and Department of Microbiology and Immunology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Forchheimer 504, 1300 Morris Park Avenue, Bronx, NY 10461, USA

Abstract

Toxoplasma gondii is a ubiquitous, Apicomplexan parasite that, in humans, can cause several clinical syndromes, including encephalitis, chorioretinitis and congenital infection. *T. gondii* was described a little over 100 years ago in the tissues of the gundi (*Ctenodoactylus gundi*). There are a large number of applicable experimental techniques available for this pathogen and it has become a model organism for the study of intracellular pathogens. With the completion of the genomes for a type I (GT-1), type II (ME49) and type III (VEG) strains, proteomic studies on this organism have been greatly facilitated. Several subcellular proteomic studies have been completed on this pathogen. These studies have helped elucidate specialized invasion organelles and their composition, as well as proteins associated with the cytoskeleton. Global proteomic studies are leading to improved strategies for genome annotation in this organism and an improved understanding of protein regulation in this pathogen. Web-based resources, such as EPIC-DB and ToxoDB, provide proteomic data and support for studies on *T. gondii*. This review will summarize the current status of proteomic research on *T. gondii*.

Keywords

Apicomplexa; cell biology; genome; proteomic; *Toxoplasma gondii*

Biology of *Toxoplasma gondii*

Nicolle and Manceaux (in 1908) identified *Toxoplasma gondii* tachyzoites in tissues of the gundi, a North African rodent [1], and Splendore (also in 1908) identified the same parasite in the tissue of a rabbit [2]. Nicolle and Manceaux named the genus *Toxoplasma* for its bow-like shape (from Greek *toxos*, meaning bow or arc and *plasma*, meaning creature). *T. gondii* is able to infect all warm-blooded animals and it is estimated that a third of all humans have been infected. In humans and other animals, infection is frequently associated with congenital infection and abortion. *T. gondii* is an opportunistic pathogen that is associated with encephalitis in immunocompromised hosts, such as individuals with AIDS.

T. gondii is a coccidian parasite and has several life stages, including a rapidly growing tachyzoite stage (responsible for dissemination during acute infection), a slow-growing bradyzoite stage (which forms tissue cysts and is responsible for transmission by carnivorousness), and the sexual lifecycle oocyst stage that develops in felids [3], responsible for its transmission by water or food. As illustrated by an outbreak in Victoria, Canada [4] and a better understanding of the epidemiology of toxoplasmosis in South America [5,6], oocysts transmitted via water or other environmental sources are a significant source of *T. gondii* infection. The association of *T. gondii* with waterborne outbreaks has led to its classification as a National Institute of Allergy and Infectious Diseases (NIAID) Category B priority agent.

T. gondii has both an asexual and sexual lifecycle. Bradyzoites and tachyzoites can inter-convert, allowing it to expand asexually without having to go through the oocyst lifecycle. This has allowed the development of distinct clonal lineages in this organism. Analysis using either enzyme zymodemes or single-nucleotide polymorphisms (SNPs) has demonstrated that most *T. gondii* isolates from North America and Europe can be grouped into one of three genotypes [7,8]. These lineages are type I, typified by RH or GT-1 strains; type II, typified by the ME49 strain; and type III, typified by the VEG strain. It is believed that the expansion of these lineages 10,000 years ago is related to the domestication of animals by humans [9]. Type I strains grow rapidly in tissue culture and are highly virulent in mice. These strains seem to be frequently associated with ocular toxoplasmosis and acute outbreaks [10]. Type II and III isolates are significantly less virulent in mice and readily form cysts *in vitro*. Type II strains are most commonly isolated from clinical cases of toxoplasmosis. It is now recognized that, in other parts of the world (particularly in South America), other genotypes predominate [11]. These 'exotic' genotypes most likely represent isolates that are more common in nondomesticated animal species, and their degree of genetic variation suggests that, although clonal expansion occurs in these isolates [12], the sexual lifecycle has been more important in their evolution [5,13].

T. gondii research has been revolutionized by the available genome sequences and expressed sequence tags (ESTs) from different lifecycle stages, which are publicly accessible at ToxoDB [14,15,101]. Genome sequence data is available for ME49 (a type II strain), GT-1 (a type I strain able to complete the entire lifecycle) and VEG (a type III strain also able to complete the lifecycle). Alignments of the sequences of these three strains are available as genetic maps, as are SNP analyses of these three major genotypes. Expression data and proteomic data are also available on this site, as is sequence data for chromosome Ia and Ib from the RH strain, a commonly used type I laboratory strain [14,15,16]. ToxoDB is part of the Apicomplexan database (ApiDB) and Eukaryotic Pathogens Database (EuPathDB), gathering data for several Apicomplexa, including *T. gondii*, *Plasmodium* species and *Cryptosporidium*, as well as data from other sources with the goal of facilitating comparative research [15,17]. A separate proteomics resource is available at EPIC-DB [18,102], which provides a comprehensive theoretical *T. gondii* proteome based on the existing gene models, as well as a comprehensive proteomics dataset on *T. gondii*.

***T. gondii* proteomics**

Proteomic experiments entail experimental design, sample preparation (biological and protein), data acquisition, data processing and database searching, interpretation and validation of results. While the mass spectrometry (MS) data acquisition may represent the most challenging aspect to a biologist, biology is what drives good proteomic experiments. The identification and characterization of the function of proteins in *T. gondii* has been a focus of many research groups since the development of techniques that enabled the study of individual genes and molecules. Among these active research topics have been studies of the mechanism(s) of host cell invasion, the structure and composition of the apical organelles, stage conversion and the organization of the cytoskeleton. The availability of the *T. gondii* genome has allowed proteomic studies to be undertaken to investigate these research topics.

Proteomic identification is entirely dependent on the accuracy of the associated gene models, against which MS data is searched. Unmatched peptides contain information that may assist in correcting gene models and provide key input for training gene-finding algorithms. Proteomic analysis is, therefore, also essential to fully realize the biological value of the sequenced genome. By directly measuring peptides arising from expressed proteins, it is possible to validate coding regions, to confirm alternative splice variants in the intron-rich genomes of *T. gondii* and to identify missed protein-coding genes.

A confounding issue in the interpretation of proteomics experiments in *T. gondii* is the state of gene-prediction databases at the time of the analysis. The implications of this are that one cannot be certain that a given protein will be present in the database, nor that a high-quality mass spectrum of peptide(s) will identify the originating protein. For example, if intron–exon junctions are predicted incorrectly such that a protein is predicted to be two gene products instead of one, two peptides identified might reside one in each prediction. Since current criteria for publishing MS data require at least two peptide hits per polypeptide, the data in this situation would be lost. It is useful to be able to use the highest resolution mass spectrometer available. This permits the use of ‘one-hit wonders’ and facilitates *de novo* sequencing of peptides, which is not dependent upon databases.

Several bioinformatic tools exist to evaluate the enormous data-sets generated by proteomic studies; however, improved tools are needed to integrate this data into gene annotation systems [19–24]. A criticism of proteomic approaches has been that it does not follow a classical ‘hypothesis-driven’ research approach. Instead, proteomic experiments are designed not to prejudge the outcome of an analysis by, in an unbiased way, identifying genes and proteins that are associated with specific biological events. These proteomic experiments are involved in hypothesis formation rather than hypothesis testing. Proteomic experiments are, therefore, part of an iterative hypothesis development and testing process that involves classical experimentation to test the hypothesis developed by proteomic studies.

Apicomplexan proteomics has benefited from a range of advances, such as improved subfractionation of complex protein mixtures, separation and analysis of Apicomplexan sub proteomes and the improvement of bioinformatics resources, both for computational analysis and, specifically, for Apicomplexan proteins. There are many excellent generic reviews on proteomics in the literature [25–29] that can be referred to for a more thorough introduction to this subject. Large-scale proteomic approaches have been used to analyze genomes of various organisms, such as *Saccharomyces cerevisiae* [30], *Mycoplasma mobile* [31], *Corynebacterium parvum* [32], *T. gondii* [33,34] and *Streptomyces luteogriseus* [35]. Targeted studies of *T. gondii* rhoptry (ROP) [36], secretory [37] and micronemal [38] proteins highlight the value of applying proteomics to explore important subproteomes. Furthermore, proteomics

can be used to elucidate the role of post-translational modifications, such as N-glycosylation, in the function of important proteins [39].

Wastling *et al.* examined the available proteomic data for various Apicomplexa, including *T. gondii*, on EuPathDB (which includes ToxoDB) and how these data correlated with transcriptomic data, including microarray, EST, serial analysis of gene expression (SAGE) and massively parallel signature-sequencing tags [40]. Overall, what was striking was the presence of proteomic data for genetic loci with absent transcriptional evidence. In a global study of *T. gondii* proteins by Xia *et al.* of the 2252 proteins identified, only 626 had EST data, 1131 had microarray data and 72 had no demonstrated transcripts [34]. This suggests that proteomic analysis can identify expressed genes which are ‘missed’ by conventional transcriptional analysis. Similar observations, demonstrating proteomic evidence without detectable mRNA transcripts have been seen in mammalian cells [41]. Differences in the expressed genes identified by proteomic and transcriptional datasets may be due to post-translational control mechanisms, as well as the ‘stock-and-go’ hypothesis described in *Plasmodium* [42].

Proteomic experiments may involve gel-based analysis of proteins (1D or 2D electrophoresis) followed by MS analyses of bands or spots or they may involve a global shotgun approach (i.e., MudPIT), where proteins are digested and the peptides then separated by serial chromatography followed by MS analysis of the total protein profile. Each method has advantages and should be viewed as complementary strategies for obtaining a complete protein profile of the proteome of interest. For example, gel electrophoresis provides more-detailed protein data, such as size, potential post-translational modifications and abundance prior to MS, whereas global MudPIT avoids some of the common problems relating to gel resolution of proteins with high hydrophobicity (e.g., membrane proteins) or extreme mass or pI. It should also be appreciated that different instruments, such as MALDI-TOF/TOF and various electrospray ionization instruments (e.g., double quadrupole TOFs and ion traps) also produce overlapping, but not identical, protein lists.

Data processing and database searches are the link between MS and biology. It is important to understand that the parameters that are set in the primary data processing (e.g., conversion to *data* or *mgf* files) can affect the results, as can the settings used in the search engines. Furthermore, each search engine (e.g., Sequest, Mascot, X!Tandem and SpectrumMill) provides overlapping but different protein lists. All are valuable and all require validation. There are software programs available that can evaluate results obtained from different search protocols. Another tool used to evaluate the quality of the MS peptide hit is the false-discovery rate. Data can be searched automatically against the gene prediction database and against a scrambled version of the database. The false-discovery rate is a measure of the percentage of MS spectra, which generate hits in the scrambled database and are, thus, suspect.

Sample preparation

Sample preparation is a critical step in proteomic analysis. To date, the tachyzoite stage of *T. gondii* has been used for proteomic studies and no significant data has been reported on other lifecycle stages. As the aim of most proteomic research is to provide a comprehensive inventory of proteins present in a structure or an intact organism, it is critical that a highly enriched fraction or purified lifecycle stage is utilized for the experiments. Most of the proteomic studies have utilized RH tachyzoites, since this type I strain has almost no bradyzoite differentiation *in vitro* under standard culture conditions (pH 7.1). Further large quantities can be produced in tissue culture from infected host cells after either spontaneous egress or mechanical disruption of infected cells using a narrow-gauge needle. Filtration of *T. gondii* tachyzoites through a 3- μ M pore-filter membrane produces a preparation that provides high-quality data on proteome analysis [43].

The proteome of any cell consists of a heterogeneous mixture of both soluble and hydrophobic proteins that are present in a large dynamic range. To overcome this challenge of complexity it is useful to simplify the protein mixture prior to analysis. Methods have been developed to purify organelles from tachyzoites or produce subcellular fractions. Rupture of parasites using a French press technique yields high-quality subproteomes that have been useful for studies on soluble cytosolic proteins, membrane-associated proteins, cytoskeletal proteins and the invasion organelles (ROPs and micronemes). Another subproteome that has been characterized is the excretory–secretory products of *T. gondii* (or ESA), which are released following treatment of isolated tachyzoites with various chemical agents (e.g., 1% ethanol) [37].

To obtain a ROP proteome, Bradley *et al.* disrupted tachyzoites by a French press [36] and these were fractionated the cell lysates on a Percoll gradient to isolate a fraction enriched for ROPs [44]. As this fraction still contained dense granules, mitochondria and plastids, the fraction was further subjected to sucrose-gradient flotation to improve enrichment for ROPs [36]. Analysis of this fraction by immunoblots employing antibodies to known proteins from the ROPs, dense granules and mitochondria, demonstrated that this technique was able to enrich for the ROPs while effectively removing the vast majority of dense granules and contaminating mitochondria. Using carbonate extraction followed by immunoblot analysis, it was determined that the majority of ROP proteins of interest [36], were found within the membrane fraction and not amenable to analysis by 2D electrophoresis (2DE). To this end, an alternative to 2DE proteome analysis was used separating the proteins by 1D sodium dodecyl sulfate (SDS) polyacrylamide gel electrophoresis (PAGE) followed by the excision of 51 contiguous gel slices, each of which was subjected to in-gel trypsin digestion and then MS-MS to obtain peptide fragmentation data suitable for proteomic database searching. Identification of known ROP proteins within the subproteome validated this approach and localization of candidate ROP proteins was verified by immunolocalization.

Liu *et al.* have described a three-layer sandwich gel electrophoresis method that improved proteomic analysis by allowing removal of salt with near quantitative recovery of low concentrations of proteins from harsh solubilization solutions [45]. This procedure combines agarose gels, which serve as the matrix to immobilize proteins of interest, with low- and high-concentration polyacrylamide gels, which serve as the concentration and sealing layers respectively. This technique has been successfully applied to proteomic studies of the parasitophorous vacuole membrane (SINAI A, PERS. COMM.) [45]. This technique was used to concentrate proteins from a *T. gondii* lysate, mimicking conditions used for immuno-precipitation. A total of 627 peptides, mapping to 257 proteins, were then identified by 2D liquid chromatography (LC) MS-MS.

Martin *et al.* describe a method to improve the MS of immune precipitations using biotinylated antisera [46]. Caprylic acid is used to prepare pure antibody free of serum proteins, which is then biotinylated and crosslinked to protein A. *T. gondii* lysate was applied to these beads and then eluted using standard techniques. On MS the presence of eluted IgG masked the *T. gondii* peptides so prior to MS streptavidin beads were used to capture any free IgG. This resulted in improved detection of low-abundance *T. gondii* proteins in the sample. This method has been used to examine proteins purified from a total *T. gondii* lysate by an antibody raised to *T. gondii* parasitophorous vacuole membrane. The identified proteins were serine threonine phosphatase 2C, p30 (SAG1) precursor, tetra A protein, actin and a hypothetical protein [46].

It should be appreciated that the interpretation of any data from a subproteome needs to take into account the degree of enrichment obtained for the organelle or subproteome in question. For most preparations, some degree of contamination from proteins from other sources is inevitable. Many molecules in the cytoplasm interact with organelle proteomes and may be

associated with these proteins, although not strictly derived from the organelle itself. In addition, the proteins within an organelle may be in flux and, therefore, the modifications and proteins identified may not be the same under all conditions. It is, therefore, critical that such studies utilize reproducible sample-preparation techniques that are replicated during analysis and that the localization of identified proteins should be verified by other techniques, such as immune fluorescence microscopy and/or bioinformatic approaches.

***T. gondii* subproteomes**

As an obligate intracellular parasite, *T. gondii* must successfully invade host cells and create a hospitable environment in which it can acquire nutrients, yet, avoid killing by its host cell. The micronemes, ROPs and dense granules are specialized secretory organelles for invading and remodeling host cells. The micronemes secrete a collection of adhesion proteins, termed microneme proteins (MICs), that mediate host cell entry [47–49]. ROP proteins and lipids, with the assistance of microneme protein AMA1, form the moving junction during invasion [50], eventually resulting in a parasitophorous vacuole that enables efficient procurement of nutrients and evasion of host immune defenses [36,51]. The parasitophorous vacuole membrane is modified extensively by the parasite and contains multiple proteins that interact with host cell organelles, including the mitochondria and endoplasmic reticulum. Dense granule proteins appear to facilitate formation of specialized tubules that enable nutrient acquisition by the parasite [52], and some of these proteins, most notably GRA7, are also secreted into host cells [52]. ROPs are also injected into the host cell, resulting in extensive modification of host gene expression and signaling pathways [53].

As discussed, Bradley *et al.* developed a procedure to purify ROP organelles [36]. Owing to the hydrophobic nature of the ROP proteins, a 1D LC-MS-MS approach was used for analysis of 51 gel slices. A total of 38 previously unidentified candidate ROP proteins were detected after proteomic analysis. A combination of approaches was used to determine the localization of the identified proteins including epitope tagging and the production of antibodies against peptides and recombinant proteins. Of the 13 proteins for which antisera were produced, 12 localized to the ROPs. Several proteins were identified in the ROP fraction by Bradley *et al.*: ROP1, 2, 4, 5, 8, 9, 10, 11, 12, 13, 14, 15 and 16, RON1, R2, R3, R4, Toxofilin, and Rab11 [36]. A major discovery from this proteomic analysis was the recognition of ROP neck proteins (RONs) that were later demonstrated to be critical in the formation of the moving junction during host cell invasion. El Hajj *et al.* used immunoprecipitation coupled with MS analysis to identify and characterize the ROP2 family proteins consisting of ROP2, 4, 5 and 7 [54]. By basic local alignment search tool (BLAST) analysis of the *T. gondii* genome, they also identified ROP8 (ROP2L1), ROP11 (ROP2L2), ROP2L3, ROP2L4, ROP2L5, ROP2L6 and ROP16 as members of this family of proteins.

T. gondii secretes proteins from ROPs, micronemes and dense granules during invasion. The ESA protein profile consists of microneme and dense granule proteins. This repertoire of secreted proteins is surprisingly complex, with evidence for multiple redundant adhesive complexes whose components undergo extensive proteolysis during organelle maturation and host cell invasion [37,38]. Bioinformatic analysis of the *T. gondii* genome suggest that at least 800 genes encode proteins with putative secretory signal peptides [37]. Zhou *et al.* [37,38] and Fauguenoy *et al.* [39] characterized proteins released during host cell invasion, including ESAs, using 2DE and multidimensional protein identification technology (MudPIT) techniques. Approximately 100 spots could be identified by 2DE and most were successfully identified by protein microsequencing or MALDI MS analysis. Many proteins were present in multiple spots, consistent with the presence of post-translational modifications and or post-translational processing of these proteins. MudPIT yielded additional novel adhesion proteins and hypothetical secretory proteins similar to proteins identified in *Plasmodia*. This analysis

was done using a database of the EST gene predictions, *T. gondii* TwinScan gene predictions, and NCBI (NR) gene predictions. Many of the identified proteins were surface antigens, and dense-granule proteins, demonstrating that the ESA preparation was not limited to micronemes.

In similar studies, Kawase *et al.* used A23187 to stimulate calcium-mediated excretion of proteins from *T. gondii*. A total of 213 protein spots were identified by 2DE and, on comparison of spots identified from A23187- and DMSO-treated parasites, a total of eight spots were increased [57]. These proteins were identified as MIC2, MIC4, TgSPATR, AMA1, ROP9, MIC6 and MIC10.

Zhou *et al.* characterized a subset of the novel ESA proteins by expressing them as fusion proteins with yellow fluorescence protein [37]. This screen revealed shared and distinct localizations within the secretory compartments of *T. gondii* tachyzoites and significantly broadened our understanding of these proteins. Only one of the 38 ROP proteins identified by Bradley *et al.* [36] was found in the ESA fraction of Zhou *et al.* [37], illustrating that these two subproteomes were distinct. Purified microneme proteins have also been analyzed using proteomic techniques [55,56]. These studies have expanded our knowledge of the proteins contained in the micronemes, as well as proteolytic processing (KIM K, UNPUBLISHED DATA) and other post-translational modifications of these proteins.

Hu *et al.* purified the conoid/apical complex of *T. gondii* and were able to identify 286 proteins using a MudPIT approach employing SEQUEST software screening a database consisting of six frame translations of all open-reading frames (ORFs) over 15 amino acids of the *T. gondii* ME49 genome and all ORFs over 50 amino acids from the EST sequences (available at [101]) [58]. These authors analyzed 10–12 replicate samples of conoid-enriched and conoid-depleted fractions resulting in the identification of 1157 proteins. Of these, approximately 30% were unique to each fraction and 35% were easily identified as contaminants from other subcellular organelles (e.g., ribosomal proteins, mitochondrial proteins and microneme proteins). Validation of seven candidates was performed by either the production of antibodies to purified peptides or recombinant proteins or by *T. gondii* tachyzoite transfection of tagged genes, identified by proteomics. Of these proteins, TgMORN1 is localized to the end of daughter cells and TgCAM1 and CAM2 to the conoid region.

The Albert Einstein Biodefense Proteomics Center (EPIC-DB) has analyzed 279 replicate samples of cytoskeletal material prepared from detergent-treated French Press-disrupted parasites [102] purified according to a modification of the method of Bradley *et al.* [36]. At least 1000 proteins have been identified in this proteomic survey. Analysis by 2DE demonstrated that isoforms and post-translational modifications of proteins were a common feature of proteins expressed in this *T. gondii* cytoskeletal fraction. Complex modifications were present in both α - and β -tubulin (Hui X, El Bissati K, Verdier-Pinard P *et al.*, Unpublished Data). Plessmann *et al.* demonstrated that acetylation and glutamylation occur on the α -tubulin of *T. gondii* [59].

***T. gondii* post-translational modifications**

Analysis by 2DE and other techniques of *T. gondii* proteins has demonstrated the presence of complex isoforms for many proteins, suggesting that post-translational modifications are common in this parasite. For example, lectin-binding studies coupled with proteomic analysis have identified that glycoproteins are present in *T. gondii* [102]. Since post-translational modifications are likely to be functionally important, determining when and why they occur will be essential to understanding the biology of the parasite.

Fauquenoy *et al.* examined SDS soluble material from RH and 76K strain *T. gondii* for the presence of proteins that could be isolated by ConA lectin-affinity chromatography and then

identified by nano-LC-MS-MS [39]. In addition, they characterized the *N*-glycan modifications present in *T. gondii*. Using MALDI-TOF they identified three major and four minor ions that suggested that the major *N*-glycans of *T. gondii* have compositions with high mannose type structures (Man 6–9(GlcNAc)₂) and a minor component have paucimannose type structures of (Man3–4[GlcNAc]₂). A total of 26 proteins were identified by ConA-affinity chromatography, including MyoA, MyoB/C, IMC2A, actin, tubulin β -chain, membrane anchor for myosin XIV precursor (GAP50), RON1, RON2, ROP18, ROP7, apical membrane antigen 1 homolog, perforin-like protein 1, sortilin, DNJ domain protein and PfSec61. Based on these data, they treated parasite cultures with tunicamycin, an inhibitor of *N*-glycosylation, and were able to demonstrate effects on both gliding motility and host cell invasion [39].

Braun *et al.* used a proteomics approach to define the small ubiquitin-like modifier (SUMO) conjugation system in *T. gondii* [60]. SUMOylation is associated with increased protein stability. A parasite expressing a HA FLAG epitope-tagged *T. gondii* SUMO was utilized along with FLAG-affinity chromatography to purify a subset of proteins that had been SUMOylated. A nontagged parasite was used as a negative control. A total of 120 candidate proteins were identified by at least two peptides. These proteins fell into several categories, including chromatin and transcriptional machinery, ribosomal biogenesis, translation-related proteins, ROP proteins and stress-related proteins. SUMOylated proteins were also present in the bradyzoite parasitophorous vacuole membrane, both *in vitro* and *in vivo*.

Data suggest that selenoproteins are probably present in the Apicomplexa, including *T. gondii* [61]. Selenoproteins contain a rare amino acid, selenocysteine (Sec), encoded by the codon UGA. In eukaryotes, incorporation of Sec requires a Sec insertion sequence (SECIS) element, a stem-loop structure located in the 3'-untranslated regions of selenoprotein mRNAs. *In silico* analysis has demonstrated a noncanonical SECIS element with a GGGA sequence in the SBP2-binding site in place of AUGA as classical AUGA SECIS elements in *T. gondii* [61]. No proteomic data have been obtained on these potential selenoproteins.

***T. gondii* global proteomics**

Global proteomic experiments provide extensive lists of protein 'hits' without a biological reference, except the tissue or organism. To identify all proteins in a cell or tissue, whatever the level of abundance (representation), requires extensive fractionation of proteins or peptides, since the levels of expression may range across many orders of magnitude. Several rationales can justify a global proteomics effort. As discussed previously, gene prediction algorithms are still imperfect, particularly in organisms with many introns. Global proteomics can help validate gene predictions. In order to accomplish this, it can be helpful to have information about the expressed genes, such as molecular weight. It should be noted that it is possible for a particular protein to be completely missing from a gene prediction database. Therefore, the MS data may need to be interpreted *de novo* and orthogonal validation carried out. Finally, exhaustive global proteomics experiments provide a method to quantitate changes in protein expression between cells, tissues or a developmental state, such as tachyzoites and bradyzoites.

An early global proteomics effort was that of Cohen *et al.*, who were able to resolve over 1000 polypeptides by 2DE [43]. They successfully defined a protocol for a reproducible 2DE map of RH tachyzoites. MS was used to analyze 71 of these proteins by MALDI MS and MALDI post-source decay analysis [43]. This study showed that, in many cases, several protein spots were encoded by the same gene, indicating that post-translational modification and/or alternative splicing events occurred. As expected from the genome predictions, MS-MS was much more useful than mass fingerprint data for the identification of genes. Approximately 30 tachyzoite proteins were identified with these being the more highly expressed ROP, dense-granule and structural proteins. This early experiment demonstrated the viability of proteomic

analyses for *T. gondii*, even in the absence of complete genome sequence. Another early proteomics effort was that of Dlugonska *et al.* who identified 13 excretory antigens from tachyzoites on a standardized 2DE map [62].

Using a multiplatform proteomics approach to characterize the tachyzoite stage of the parasite, Xia *et al.* identified 2252 proteins, an estimated 30% of the predicted proteins in *T. gondii* [34]. This analysis identified 2477 intron-spanning peptides, providing evidence for correct splice-site annotation. At least 15% of the identified peptides matched more convincingly to alternative gene models. 2DE was used to identify 1217 spots using electrospray MS. A total of 616 nonredundant proteins were identified, of which 547 corresponded to *T. gondii* Release4-gene annotations and 69 to alternative gene models or ORFs [101]. LC-MS-MS was used to examine SDS-soluble RH tachyzoite proteins and a total of 2778 proteins, which, when redundancy was removed, collapsed to a dataset of 1012 gene products (939 Release4 and 73 alternative gene models). MudPIT was used to analyze both Tris-soluble and -insoluble fractions from RH tachyzoites. When combined and redundancy was eliminated, these experiments resulted in a dataset of 2409 proteins (2121 Release4 and 288 alternative models).

Dybas *et al.* used MS techniques (1D LC-MS-MS) to identify 2477 gene-coding regions, with 6438 possible alternative gene predictions from the tachyzoite stage of this parasite, representing approximately a third of the expected genome [33]. Analysis demonstrated that commonly used gene-prediction algorithms produced very disparate sets of protein sequences with overlaps ranging from 1.4 to 12%. Overall, current prediction methods had observed false-negative rates of 31–43%. A proteomics database was created (EPIC-DB), which combined current experimental (National Center for Biotechnology information [NCBI]) and predicted *T. gondii* genes. The amino acid sequences of the hypothetical proteome of *T. gondii* were searched against the complete NCBI nonredundant (NR) database Apicomplexa proteins and human proteins to identify unique and conserved proteins. In general, 67% of the hypothetical *T. gondii* proteome had a homologous sequence in NR and 64% an Apicomplexan ortholog. Approximately 52% of *T. gondii* sequences were unique compared with the human genome. Of the experimentally identified proteins, 3838 (60%) have been annotated as ‘hypothetical’, ‘putative’ or ‘predicted’ in the NCBI NR database. Additionally, 609 proteins identified by MS in *T. gondii* were unique compared with other known organisms, providing an important subset of validated proteins that can be investigated further for functional studies or as candidates for drug targeting.

Bioinformatics overview of the hypothetical *T. gondii* proteome

Interpretation of MS-MS peptide spectra is performed either by a database-search approach, by comparing measured masses against a set of theoretical masses of possible proteolytic peptides or by *de novo* sequencing [63,64]. These approaches may limit the discovery space of the study to a subset of the genome that is composed of predicted genes. However, it is also possible to translate the genomic sequence in all possible six reading frames and the search MS-MS data against all possible identified ORFs. While this latter approach circumvents the bias imposed by gene-prediction methods, it has limitations. In *T. gondii*, the number of possible ORFs obtained from all six framed transition is approximately 0.77 million. Such a database makes MS-MS data searches much more time-consuming compared with only matching data against the expected 6000–8000 predicted genes. Furthermore, a peptide match in a given ORF may not be fully conclusive if the corresponding full gene structure is not obvious. A *T. gondii* gene is composed on average of 3.6 ORFs, making accurate predictions more difficult. Scanning peptide masses is also more complicated as significance scores in typical search programs are empirically calibrated to a smaller number of average-sized proteins instead of a very large number of short ORFs. The average length of ORFs in *T.*

gondii corresponds to approximately 100 residues. Finally, many MS peptide hits that bridge ORF boundaries will not match any data.

Recently, two large-scale proteomics studies were completed [33,34] providing new insights into the gene structure of *T. gondii*. Database-oriented MS-MS analysis requires the existence of a hypothetical compilation of *T. gondii* proteome. For this purpose Dybas *et al.* collected the gene predictions from four gene prediction algorithms, TigrScan [65], TwinScan [66], GlimmerHMM [65] and Release4 (currently available as version 5) [15] into a hypothetical proteome. These predictions were developed by the ToxoDB team [15] to identify genes in the ME49 strain of *T. gondii*. The hypothetical proteome, comprised of 30,197 amino acid sequences, combined predicted proteins from the TigrScan (8336 sequences), TwinScan (7588 sequences), Glimmer (4954 sequences) and Release4 (7793 sequences) data-sets, and the available *T. gondii* sequences from the NCBI nonredundant protein database (1526 [NR]). Comparisons of the computational gene-prediction methods produce protein sequence sets that are quite different. Identical or nearly identical proteins in the hypothetical proteome were grouped to explore the overlap and differences among the computational gene prediction methods. Surprisingly, any two prediction methods shared less than 12% (and as low as 1.4%) identical predicted genes in a head-to-head pairwise comparison. As many as 75% of gene predictions were unique to one of the methods and only 101 sequences agreed among all four gene-predictions sets.

Discrepancies in protein sequences that are transcribed from similar genomic locations, whether the result of prediction differences or splicing variability or tandemly arrayed genes of similar sequence, probably account for most of the disparities among the protein datasets. In order to assess the fraction of sequences that are not identical, yet share a common part and, thus, may have been derived from the same genomic location, the hypothetical proteome of 30,197 sequences was clustered with a 90% sequence identity threshold requirement, allowing very large gaps in the alignment.

The hypothetical proteome of 30,197 *T. gondii* sequences collapses to 14,983 nonredundant clusters with an average size of 2.02 sequences. The majority of the clusters (55%) are composed of a single sequence, strengthening the observation that there are many unique sequences that do not share even a common subsequence with other predicted proteins and that the predicted protein datasets provide remarkably different alternatives. The percentages of the clusters that contain a TigrScan, TwinScan or Release4 sequence are similar (53, 47 and 49%, respectively), while the percentage that contains a Glimmer sequence is lower (33%). The small fractions of common protein predictions suggest that any single prediction dataset either covers a small portion of the theoretical *T. gondii* genome (~50% or less) and/or produces a very large percentage of false-positive results.

Clustering proteins by sequence similarity may group together paralogs from different genomic locations that emerged through gene-duplication or gene-shuffling mechanisms. Therefore, the homogeneity of sequence clusters was explored in terms of genomic localization. Genomic mappings of the proteins that share more than 90% in their overlapping part showed that approximately 1% of these sequences maps to different genomic locations, which suggests a relatively low level of gene-duplication events. There are 3120 and 2969 genomic regions that are comprised of either an individual cluster or a colocalized cluster group (clusters where at least one common sequence overlap), respectively and, thus, it is estimated that there are at least 6089 potential protein-coding regions in the *T. gondii* genome. In general, according to the various prediction algorithms, the *T. gondii* genome is expected to encode approximately 7800 genes [26,27], of which, 18% are expected to be lifestage specific (~1400) [67].

Dybas *et al.* explored the proteome of *T. gondii* tachyzoites with high-throughput proteomics experiments and by comparison to publicly available cDNA sequence data [33]. The average peptide coverage of the MS-MS-supported sequences was 13.5%. The percentages of sequences with an assigned peptide and the overall peptide coverages were similar for the TigrScan, TwinScan, Glimmer and Release4 gene-prediction datasets (17, 19, 27 and 20%, and 10.7, 13.4, 8.6 and 12.5%, respectively). The observation that each protein-prediction method has a similar degree of experimental validation, along with the disparate nature of the sequences predicted by each method, indicates that each method predicts the proteome with a similar level of accuracy. There are 2477 predicted protein-sequence clusters (17% of the total) that contain at least one sequence with an assigned MS-MS peptide. The portions of the MS-MS supported clusters that contain a TigrScan, TwinScan, or Release4 sequence are similar (61, 66 and 68%, respectively) while the portion that contains a Glimmer sequence is slightly lower (57%). These data show that none of the full-genome predictions manages to identify all of the proteins that are supported by MS-MS data; each prediction method is missing 31–43% of the experimentally validated part of the proteome. Overall, the Dybas *et al.* high-throughput MS-MS experiments explored at least a third of the genome of *T. gondii* tachyzoites (2477 clusters) [33]. This will underestimate the total numbers of expressed proteins if a significant fraction of clusters represent two or more gene products.

Analysis of EST mapping of the hypothetical *T. gondii* proteome offers similar conclusions to those generated from the MS-MS analysis [33]. It was possible to validate 20,123 sequences (67% of the combined dataset) with an aligned EST sequence, which corresponds to 9242 sequence clusters (62% of the total clusters). The TigrScan and TwinScan datasets have approximately the same percentages of sequences that are supported with an EST alignment (58 and 62%, respectively) while the Glimmer and Release4 methods have similar and higher percentages of sequences with an EST alignment (75 and 72%, respectively). The enrichment of Glimmer sequences with an EST alignment is probably a result of the Glimmer sequences being, on average, much longer than the TigrScan, TwinScan or NR sequences. Thus, there is a greater chance that a Glimmer sequence will cover part of the genome that is also sequenced by an EST. The high percentage of Release4 sequences with an EST alignment is possibly the result of ESTs (and other experimental data) being included in the integrated data that was used to derive the Release4 sequences [15].

The MS-MS proteomics data were crossreferenced with the EST genomics data [33]. There are 5881 sequences (19.5% of the original combined dataset) that are experimentally supported by both an assigned MS-MS peptide and an EST alignment and, therefore, should be considered the most highly validated proteins in the compiled proteome. These sequences correspond to 2275 clusters (15.2% of the total clusters). The average MS-MS peptide coverage of sequences that are supported by MS and EST data or exclusively by MS data alone are 14.3 and 5.3%, respectively. The difference in peptide coverage may be a result of the proteins with EST data being more highly expressed in the *T. gondii* proteome compared with those without EST alignments.

Sequence clusters that are supported by both MS data and EST data have a comparable portion of Glimmer, TigrScan, TwinScan and Release4 sequences (58, 61, 66 and 69%, respectively) [33]. Therefore, one can conservatively estimate that the current gene-prediction algorithms exhibit a false-negative rate of 31–42% as defined as the portion of experimentally validated coding regions (clusters; MS and EST) that are completely missed by individual prediction methods. While it is not possible to precisely identify the false-positive predictions from these data, it is clear that the proportion of false-positive predictions must be substantial for each prediction method.

Functional annotation of the *T. gondii* proteome

Dybas *et al.* functionally annotated the predicted *T. gondii* proteome [33]. There are 13,430 sequences (45% of the combined dataset) with at least one Pfam domain annotation compared with 20,326 sequences (67%) that are annotated with orthologs predicted from pairwise BLAST alignments with sequences from the NCBI NR database [68]. The percentage of proteins that have Pfam domains compares well with the published results of Pfam 22.0, which reports that 49% of 213 *T. gondii* UniProtKB/TrEMBL deposited proteins have at least one Pfam domain [69]. Sequences with Pfam domains correspond to 5741 clusters (38% of the total clusters).

Transmembrane proteins of *T. gondii* are of particular interest because of their role in the interaction and adhesion of the parasite to the host cell and their link to the virulence of the parasite. These critical functions make some subsets of membrane proteins good candidates for chemotherapeutic targets. In total, 6927 sequences (23% [3730 clusters (25%)]) have predicted transmembrane domains, which is similar to the proportion of membrane proteins that have been predicted in other genomes (20–35%) [70]. Sequences with signal peptides generally refer to proteins that are secreted from the cell. In the case of *T. gondii*, secreted proteins include those of the micronemes, ROPs and dense granules, which are integral actors in the unique process of interaction, invasion and infection of the host cell by *T. gondii* [71]. Dybas *et al.* found 4330 sequences (14%) that have predicted signal peptides, which correspond to 2601 clusters (17%). The PredGPI program identified 138 highly probable and 391 probable GPI-anchored proteins that can be clustered into 56 and 180 groups of proteins [72]. Xia *et al.* also explored the possible fraction of membrane and signal peptide proteins in *T. gondii* and found that 11% of proteins contain a signal peptide and 18% contain transmembrane domains [34].

Bioinformatics, proteomics resources for *T. gondii*

All MS data and analysis by Dybas *et al.* is publicly accessible at the EPIC-DB [18,102,103] and summaries have been deposited in the NIH Resource Center for Biodefense Proteomics Research [104], as well as ToxoDB [101].

ToxoDB [14,15] is a functional genomic database for *T. gondii* that incorporates sequence and annotation data and is integrated with other genomic-scale data, including community annotation; ESTs and gene-expression data. It is a component site of the Apicomplexan Bioinformatics Resource Center, which provides a common research platform to facilitate data access among this important group of organisms and is part of the EuPathdB [105], which provides a relational database on eukaryotic pathogens. ToxoDB contains proteomic data from the research community, including the datasets from Dybas [33] and Xia [34]. Proteomic data is displayed via a GBrowse interface, allowing rapid correlation of MS data to genome predictions, expression data and epi genetic data that is also archived at ToxoDB. ToxoDB also supports researching of MS-MS data in a variety of common outputs, such as Mascot [106], Global Proteome Machine [107,108] and MALDI peptide mass fingerprint data searches using emowse [109].

The EPIC-DB is a publicly accessible, queryable, relational database that organizes and displays experimental, high-throughput proteomics data for *T. gondii* (and *Cryptosporidium parvum*) [18]. It incorporates many features that allow for the analysis of the entire proteomes and/or annotation of specific protein sequences. This site provides downloadable MS data along with detailed information on MS experiments. Raw data are available from this site on all experiments. While raw data is unlikely to be accessed widely by the user community, its retention preserves the possibility of requerying the original data at a later time. The database also provides antibody experimental results and analysis of functional annotations,

comparative genomics, and aligned EST and genomic ORF sequences. The database contains all available alternative gene-prediction sets for each organism, comprising a complete theoretical proteome. All experimental data is referenced to these sequences. The database is structured around clusters of protein sequences (as discussed previously), which allows for the evaluation of redundancy, protein-prediction discrepancies and possible splice variants. EPIC-DB is complementary to other genomics-oriented databases of these organisms by offering MS analysis on a comprehensive set of predicted protein sequences.

Expert commentary

The availability of high-throughput technologies for sequencing, proteomics, transcriptomics, epigenomics and metabolomics has changed the nature of research approaches from hypothesis driven to hypothesis generating. The available proteomics data on *T. gondii* is providing a key resource to develop techniques to utilize proteomic data to improve genome annotation techniques. The difficulty of accurate gene prediction demonstrated by the global proteomic data generated on *T. gondii* is not unique to this organism and underscores the need for fundamental improvements in the *in silico* methods used to predict protein datasets from genome sequence data. This problem is particularly acute in organisms with multiple splice sites in each protein.

T. gondii is readily amenable to analysis using the techniques at the forefront of proteomics. Improved methods to purify sub proteomes, such as the parasitophorous vacuole membrane, present major challenges but should revolutionize our understanding of the structures formed by this organism. Proteomics technology provides powerful tools to further elucidate aspects of *T. gondii*'s relationship to its host cell. Moreover, the application of selective labeling techniques (i.e., isobaric tag for relative and absolute quantitation [iTRAQ] and stable isotopic labeling using amino acids in cell culture [SILAC]) in conjunction with genetic approaches will accelerate identification of the functional targets of proteins that *T. gondii* utilizes to modulate its host cells.

To use proteomics as a link between genome data and its function, then proteomics must move forward to provide more detailed and complete functional data on the proteins that it describes. This will involve the development of methods to perform a high-throughput analysis of post-translational modifications, as well as techniques to use proteomic technology to map protein interactions. All of this must be coupled with new bioinformatics platforms to handle the complex datasets that are generated and with standards for community-wide deposition of data that will facilitate repeat analysis of data as new tools and insights are developed. Integration of the evolving genomic, proteomic and metabolomic data promises fascinating insights into how *T. gondii* functions as a successful eukaryotic pathogen.

Five-year view

Proteomics has proven to be a powerful approach for cataloguing the repertoire of proteins expressed in *T. gondii* tachyzoites, as well as identifying proteins in important subproteomes. These datasets have led to identification of candidate proteins that are hypothesized to mediate critical biological processes, such as cell adhesion, motility and host cell invasion. While our understanding of specific organelles has advanced, little is known about how *T. gondii* transitions from different lifecycle forms. MS datasets for other lifecycle stages, including bradyzoites and sporozoites, will enhance our understanding of how each of these lifecycle stages differ. Attention has also shifted toward identification of post-translational modifications of proteins and protein complexes that are involved in signaling environmental cues and effecting changes in gene expression. Understanding the host signaling complexes or transcriptional complexes that interact with *T. gondii* secreted signaling molecules will be critical to understanding *T. gondii* virulence and disease pathogenesis. Initial proteomics

experiments focused upon cataloguing gene products, but future studies will need to examine the timing and nature of protein interactions in protein complexes.

Key issues

- Global proteomic studies suggest that current gene-prediction algorithms exhibit a false-negative rate of 31–42% in eukaryotic genomes, such as *Toxoplasma gondii*.
- The integration of proteomic data with other experimental data improves genome annotation.
- Purification of subcellular fractions of *T. gondii* has resulted in the definition of new proteins belonging to rhoptries, micronemes and the cytoskeleton.
- The purity of subcellular fractions and the ability to concentrate such samples is a limiting factor in their analysis.
- Proteomic studies have extended our understanding of the glycoproteins that exist in *T. gondii*.
- Proteomic studies suggest novel post-translational modifications in *T. gondii*.

Financial & competing interests disclosure

This work was supported by the Biodefense Proteomic Research Program, Contract HHSN266200400054C-02, NIH-NIAID-DMID and by an NIH-NIAID grant AI39454 (LMW). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as:

- of interest
- of considerable interest

1. Nicolle C, Manceaux L. Sur une infection a coprs de Leshman (ou organisms voisins) due gondi. CR Acad. Sci 1908;147:763.
2. Splendore A. Un nuovo parassita deconigli incontrato nelle lesioni anatomiche d'une malattia che ricorda in molti punti il Kala-azar. Rev. Soc. Sci. Sao Paulo 1908;3:109–112.
3. Frenkel JK, Dubey JP, Miller NL. *Toxoplasma gondii* in cats: fecal stages identified as coccidian oocysts. Science 1970;167(3919):893–896. [PubMed: 4903651]
4. Bowie WR, King AS, Werker DH, et al. Outbreak of toxoplasmosis associated with municipal drinking water. The BC Toxoplasma Investigation Team. Lancet 1997;350(9072):173–177. [PubMed: 9250185]
5. Belfort-Neto R, Nussenblatt V, Rizzo L, et al. High prevalence of unusual genotypes of *Toxoplasma gondii* infection in pork meat samples from Erechim, Southern Brazil. An. Acad. Bras. Ciênc 2007;79(1):111–114.
6. Bahia-Oliveira LM, Jones JL, Azevedo-Silva J, Alves CC, Orefice F, Addiss DG. Highly endemic, waterborne toxoplasmosis in north Rio de Janeiro state, Brazil. Emerg. Infect. Dis 2003;9(1):55–62. [PubMed: 12533282]
7. Howe DK, Sibley LD. *Toxoplasma gondii* comprises three clonal lineages: correlation of parasite genotype with human disease. J. Infect. Dis 1995;172(6):1561–1566. [PubMed: 7594717]

8. Sibley LD, Boothroyd JC. Virulent strains of *Toxoplasma gondii* comprise a single clonal lineage. *Nature* 1992;359(6390):82–85. [PubMed: 1355855]
9. Su C, Evans D, Cole RH, Kissinger JC, Ajioka JW, Sibley LD. Recent expansion of *Toxoplasma* through enhanced oral transmission. *Science* 2003;299(5605):414–416. [PubMed: 12532022]
10. Grigg ME, Ganatra J, Boothroyd JC, Margolis TP. Unusual abundance of atypical strains associated with human ocular toxoplasmosis. *J. Infect. Dis* 2001;184(5):633–639. [PubMed: 11474426]
11. Lehmann T, Marcet PL, Graham DH, Dahl ER, Dubey JP. Globalization and the population structure of *Toxoplasma gondii*. *Proc. Natl Acad. Sci. USA* 2006;103(30):11423–11428. [PubMed: 16849431]
12. Khan A, Fux B, Su C, et al. Recent transcontinental sweep of *Toxoplasma gondii* driven by a single monomorphic chromosome. *Proc. Natl Acad. Sci. USA* 2007;104(37):14872–14877. [PubMed: 17804804]
13. Pena HF, Gennari SM, Dubey JP, Su C. Population structure and mouse-virulence of *Toxoplasma gondii* in Brazil. *Int. J. Parasitol* 2008;38(5):561–569. [PubMed: 17963770]
14. Kissinger JC, Gajria B, Li L, Paulsen IT, Roos DS. ToxoDB: accessing the *Toxoplasma gondii* genome. *Nucleic Acids Res* 2003;31(1):234–236. [PubMed: 12519989]
15. Gajria B, Bahl A, Brestelli J, et al. ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Res* 2008;36:D553–D556. [PubMed: 18003657] • of interest Provides a description of a ToxoDB, a key depository for data on *Toxoplasma gondii*
16. Khan A, Bohme U, Kelly KA, et al. Common inheritance of chromosome Ia associated with clonal expansion of *Toxoplasma gondii*. *Genome Res* 2006;16(9):1119–1125. [PubMed: 16902086]
17. Aurrecochea C, Heiges M, Wang H, et al. ApiDB: integrated resources for the apicomplexan bioinformatics resource center. *Nucleic Acids Res* 2007;35:D427–D430. [PubMed: 17098930]
18. Madrid-Aliste CJ, Dybas JM, Hogue Angeletti R, et al. EPIC-DB: a proteomics database for studying Apicomplexan organisms. *BMC Genomics* 2009;10(1):38. [PubMed: 19159464] • of interest Provides a description of the Experimental Proteomics DataBase providing an integrated theoretical proteome for mass spectrometry research on *T. gondii*
19. Kremer A, Schneider R, Terstappen GC. A bioinformatics perspective on proteomics: data storage, analysis, and integration. *Biosci. Rep* 2005;25(1–2):95–106. [PubMed: 16222422]
20. Souchelnytskyi S. Bridging proteomics and systems biology: what are the roads to be traveled? *Proteomics* 2005;5(16):4123–4137. [PubMed: 16196099]
21. Hamady M, Cheung TH, Resing K, Cios KJ, Knight R. Key challenges in proteomics and proteoinformatics. *Progress in proteins. IEEE Eng. Med. Biol. Mag* 2005;24(3):34–40. [PubMed: 15971839]
22. Haley-Vicente D, Edwards DJ. Proteomic informatics: in silico methods lead to data management challenges. *Curr. Opin. Drug Discov. Devel* 2003;6(3):322–332.
23. Jones A, Hunt E, Wastling JM, Pizarro A, Stoeckert CJ Jr. An object model and database for functional genomics. *Bioinformatics* 2004;20(10):1583–1590. [PubMed: 15145818]
24. Taylor CF, Paton NW, Garwood KL, et al. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat. Biotechnol* 2003;21(3):247–254. [PubMed: 12610571]
25. Bradshaw RA, Burlingame AL. From proteins to proteomics. *IUBMB Life* 2005;57(4–5):267–272. [PubMed: 16036609]
26. Phillips CI, Bogoy M. Proteomics meets microbiology: technical advances in the global mapping of protein expression and function. *Cell. Microbiol* 2005;7(8):1061–1076. [PubMed: 16008574]
27. Lane CS. Mass spectrometry-based proteomics in the life sciences. *Cell. Mol. Life Sci* 2005;62(7–8):848–869. [PubMed: 15868409]
28. Yates JR 3rd. Mass spectral analysis in proteomics. *Annu. Rev. Biophys. Biomol. Struct* 2004;33:297–316. [PubMed: 15139815]
29. Johnson JR, Florens L, Carucci DJ, Yates JR 3rd. Proteomics in malaria. *J. Proteome Res* 2004;3(2):296–306. [PubMed: 15113107]
30. Washburn MP, Wolters D, Yates JR 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol* 2001;19(3):242–247. [PubMed: 11231557]

31. Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* 2004;4(1):59–77. [PubMed: 14730672]
32. Sanderson SJ, Xia D, Prieto H, et al. Determining the protein repertoire of *Cryptosporidium parvum* sporozoites. *Proteomics* 2008;8(7):1398–1414. [PubMed: 18306179]
- 33••. Dybas JM, Madrid-Aliste CJ, Che FY, et al. Computational analysis and experimental validation of gene predictions in *Toxoplasma gondii*. *PLoS ONE* 2008;3(12):e3899. [PubMed: 19065262]•• of considerable interestAnalysis of global proteomic data of *T. gondii* and the performance characteristics of current gene prediction algorithms
- 34••. Xia D, Sanderson SJ, Jones AR, et al. The proteome of *Toxoplasma gondii*: integration with the genome provides novel insights into gene expression and annotation. *Genome Biol* 2008;9(7):R116. [PubMed: 18644147]•• of considerable interestAnalysis of global proteomic data from *T. gondii* providing mass spectrometry evidence for intron splice regions in current gene predictions
35. Wang R, Prince JT, Marcotte EM. Mass spectrometry of the *M. smegmatis* proteome: protein expression levels correlate with function, operons, and codon bias. *Genome Res* 2005;15(8):1118–1126. [PubMed: 16077011]
- 36••. Bradley PJ, Ward C, Cheng SJ, et al. Proteomic analysis of rhoptry organelles reveals many novel constituents for host–parasite interactions in *Toxoplasma gondii*. *J. Biol. Chem* 2005;280(40):34245–34258. [PubMed: 16002398]•• of considerable interestProvides critical data on the proteome of rhoptries. Demonstrates a clear integrated approach to the characterization of subproteomes of this organism
- 37••. Zhou XW, Kafsack BF, Cole RN, Beckett P, Shen RF, Carruthers VB. The opportunistic pathogen *Toxoplasma gondii* deploys a diverse legion of invasion and survival proteins. *J. Biol. Chem* 2005;280(40):34233–34244. [PubMed: 16002397]•• of considerable interestProvides a proteome of excretory secreted proteins related to invasion. Many of these proteins are components of the micronemes
38. Zhou XW, Blackman MJ, Howell SA, Carruthers VB. Proteomic analysis of cleavage events reveals a dynamic two-step mechanism for proteolysis of a key parasite adhesive complex. *Mol. Cell Proteomics* 2004;3(6):565–576. [PubMed: 14982962]
- 39•. Fauquenoy S, Morelle W, Hovasse A, et al. Proteomics and glycomics analyses of N-glycosylated structures involved in *Toxoplasma gondii* – host cell interactions. *Mol Cell Proteomics* 2008;7(5):891–910. [PubMed: 18187410]• of interestProvides evidence of the existence of glycoproteins in *T. gondii* and how this post-translational modification is important for protein function
- 40•. Wastling JM, Xia D, Sohal A, Chaussepied M, Pain A, Langsley G. Proteomes and transcriptomes of the Apicomplexa – where's the message? *Int. J. Parasitol* 2009;39(2):135–143. [PubMed: 18996390]• of interestEvaluates the relationship of current transcriptome and proteomic data on the Apicomplexa, providing interesting insights into the biology of these organisms
41. Cox J, Mann M. Is proteomics the new genomics? *Cell* 2007;130(3):395–398. [PubMed: 17693247]
42. Mair GR, Braks JA, Garver LS, et al. Regulation of sexual development of *Plasmodium* by translational repression. *Science* 2006;313(5787):667–669. [PubMed: 16888139]
43. Cohen AM, Rumpel K, Coombs GH, Wastling JM. Characterisation of global protein expression by two-dimensional electrophoresis and mass spectrometry: proteomics of *Toxoplasma gondii*. *Int. J. Parasitol* 2002;32(1):39–51. [PubMed: 11796121]
44. Leriche MA, Dubremetz JF. Characterization of the protein contents of rhoptries and dense granules of *Toxoplasma gondii* tachyzoites by subcellular fractionation and monoclonal antibodies. *Mol. Biochem. Parasitol* 1991;45(2):249–259. [PubMed: 2038358]
45. Liu T, Martin AM, Sinai AP, Lynn BC. Three-layer sandwich gel electrophoresis: a method of salt removal and protein concentration in proteome analysis. *J. Proteome Res* 2008;7(10):4256–4265. [PubMed: 18795766]
46. Martin AM, Liu T, Lynn BC, Sinai AP. Elimination of affinity reagent interference for the mass spectrometric detection of low-abundance proteins following immunoprecipitation. *J. Proteome Res* 2007;6(12):4758–4762. [PubMed: 17994686]
47. Huynh MH, Rabenau KE, Harper JM, Beatty WL, Sibley LD, Carruthers VB. Rapid invasion of host cells by *Toxoplasma* requires secretion of the MIC2–M2AP adhesive protein complex. *EMBO J* 2003;22(9):2082–2090. [PubMed: 12727875]

48. Cerede O, Dubremetz JF, Soete M, et al. Synergistic role of micronemal proteins in *Toxoplasma gondii* virulence. *J. Exp. Med* 2005;201(3):453–463. [PubMed: 15684324]
49. Rabenau KE, Sohrabi A, Tripathy A, et al. TgM2AP participates in *Toxoplasma gondii* invasion of host cells and is tightly associated with the adhesive protein TgMIC2. *Mol. Microbiol* 2001;41(3):537–547. [PubMed: 11532123]
50. Alexander DL, Mital J, Ward GE, Bradley P, Boothroyd JC. Identification of the moving junction complex of *Toxoplasma gondii*: a collaboration between distinct secretory organelles. *PLoS Pathog* 2005;1(2):e17. [PubMed: 16244709]
51. Boothroyd JC, Dubremetz JF. Kiss and spit: the dual roles of *Toxoplasma* rhoptries. *Nat. Rev. Microbiol* 2008;6(1):79–88. [PubMed: 18059289]
52. Coppens I, Dunn JD, Romano JD, et al. *Toxoplasma gondii* sequesters lysosomes from mammalian hosts in the vacuolar space. *Cell* 2006;125(2):261–274. [PubMed: 16630815]
53. Bradley PJ, Sibley LD. Rhoptries: an arsenal of secreted virulence factors. *Curr. Opin. Microbiol* 2007;10(6):582–587. [PubMed: 17997128]
54. El Hajj H, Demey E, Poncet J, et al. The ROP2 family of *Toxoplasma gondii* rhoptry proteins: proteomic and genomic characterization and molecular modeling. *Proteomics* 2006;6(21):5773–5784. [PubMed: 17022100]
55. Carruthers VB. Proteolysis and *Toxoplasma* invasion. *Int. J. Parasitol* 2006;36(5):595–600. [PubMed: 16600244]
56. Carruthers VB, Tomley FM. Microneme proteins in apicomplexans. *Subcell. Biochem* 2008;47:33–45. [PubMed: 18512339]
57. Kawase O, Nishikawa Y, Bannai H, et al. Proteomic analysis of calcium-dependent secretion in *Toxoplasma gondii*. *Proteomics* 2007;7(20):3718–3725. [PubMed: 17880006]
- 58••. Hu K, Johnson J, Florens L, et al. Cytoskeletal components of an invasion machine – the apical complex of *Toxoplasma gondii*. *PLoS Pathog* 2006;2(2):e13. [PubMed: 16518471]•• of considerable interestProvides a proteomic analysis of the cytoskeleton in *T. gondii*. Demonstrates that proteomic analysis can generate new hypotheses on protein function and localization
59. Plessmann U, Reiter-Owona I, Lehtreck KF. Posttranslational modifications of α -tubulin of *Toxoplasma gondii*. *Parasitol Res* 2004;94(5):386–389. [PubMed: 15549389]
60. Braun L, Cannella D, Pinheiro AM, et al. The small ubiquitin-like modifier (SUMO)-conjugating system of *Toxoplasma gondii*. *Int. J. Parasitol* 2009;39(1):81–90. [PubMed: 18761012]
61. Novoselov SV, Lobanov AV, Hua D, Kasaikina MV, Hatfield DL, Gladyshev VN. A highly efficient form of the selenocysteine insertion sequence element in protozoan parasites and its use in mammalian cells. *Proc. Natl Acad. Sci. USA* 2007;104(19):7857–7862. [PubMed: 17470795]
62. Dlugonska H, Dytterska K, Reichmann G, Stachelhaus S, Fischer HG. Towards the *Toxoplasma gondii* proteome: position of 13 excretory antigens on a standardized map of two-dimensionally separated tachyzoite proteins. *Parasitol. Res* 2001;87:634–638. [PubMed: 11511000]
63. Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. *Brief Funct. Genomic Proteomic* 2008;7(1):50–62. [PubMed: 18334489]
64. Webb-Robertson BJ, Cannon WR. Current trends in computational inference from mass spectrometry-based proteomics. *Brief Bioinform* 2007;8(5):304–317. [PubMed: 17584764]
65. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 2004;20(16):2878–2879. [PubMed: 15145805]
66. Korf I, Flicek P, Duan D, Brent MR. Integrating genomic homology into gene structure prediction. *Bioinformatics* 2001;17(Suppl 1):S140–S148. [PubMed: 11473003]
67. Radke JR, Behnke MS, Mackey AJ, Radke JB, Roos DS, White MW. The transcriptome of *Toxoplasma gondii*. *BMC Biol* 2005;3:26. [PubMed: 16324218]
68. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25(17):3389–3402. [PubMed: 9254694]
69. Finn RD, Mistry J, Schuster-Bockler B, et al. Pfam: clans, web tools and services. *Nucleic Acids Res* 2006;34:D247–D251. [PubMed: 16381856]

70. Stevens TJ, Arkin IT. Do more complex organisms have a greater proportion of membrane proteins in their genomes? *Proteins* 2000;39(4):417–420. [PubMed: 10813823]
71. Carruthers VB. Armed and dangerous: *Toxoplasma gondii* uses an arsenal of secretory proteins to infect host cells. *Parasitol Int* 1999;48(1):1–10. [PubMed: 11269320]
72. Pierleoni A, Martelli PL, Casadio R. PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* 2008;9:392. [PubMed: 18811934]

Websites

101. ToxodB www.toxodb.org
102. Biodefense Proteomics Research Center. Experimental Proteomics DataBase for *Toxoplasma gondii* and *Cryptosporidium parvum*. www.fiserlab.org/epicdb
103. Biodefense Proteomics Research Center. www.fiserlab.org/biodefense
104. Resource Center for Biodefense Proteomics Research. www.proteomicsresource.org
105. Eukaryotic Pathogens Database Resources. <http://eupathdb.org/eupathdb>
106. Matrix Science. www.matrixscience.com
107. ToxoDB www.toxodb.org/toxo
108. The Global Proteome Machine Organization. Proteomics database and open source software. www.thegpm.org
109. GeneDB eMowse www.genedb.org/genedb/emowseSearch.jsp?organism=pert