Proteomics of *Toxoplasma gondii*

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by Dong Xia

November 2009

Table of Contents

Abstract

The Apicomplexan parasite *Toxoplasma gondii* is an obligate intracellular parasite. Infection by *T .gondii* causes the disease toxoplasmosis, which is one of the most prevalent parasitic diseases of animals and humans. It has been 100 years since the first discovery of the parasite in 1908; research on *T. gondii* has been carried out in many scientific disciplines consistently expanding the understanding of this parasite. In the last ten years, the developments of EST, microarray, genome sequencing and continuing efforts towards genome annotation has centralized the focus of *T. gondii* research on the understanding of gene expression and gene functions on the genome scale. Equipped with the technical advances in mass spectrometry and bioinformatics, proteomics has become established as an integral component in the post-genomics era by providing first-hand data on the functional products of gene expression.

In this study, three complementary proteomic strategies, 1-DE, 2-DE and MudPIT, have been used to characterise the proteome of *T. gondii* tachyzoites. Protein identifications have been acquired for more than two thousand (2252) unique release 4 genes, representing almost one third (29%) of the predicted proteome of all life cycle stages. Functional predictions for each protein were carried out, which provided valuable insights into the composition of the expressed proteome and their potential biological roles. The *T. gondii* proteomic data has been integrated into the publically accessible ToxoDB, where 2477 intron-spanning peptides provided supporting evidence for correct splice site annotation of the release 4 genome annotation. The incompleteness of the release 4 genome annotation has been highlighted using peptide evidence, confirming 421 splice sites that are only predicted by alternative gene models. Analysis has also been carried out on the proteomic data in the light of other genome wide expression data. The comparison of the proteome and transcriptome of *Toxoplasma* and other *Apicomplexa* parasites has revealed important discrepancies between protein and mRNA expression where interesting candidates have been highlighted for further investigation. A preliminary DIGE study has been developed to characterize protein expression changes in *T. gondii* grown in the presence or absence of glucose.

In conclusion, this study has demonstrated the importance of proteomic applications in understanding gene expression profiles and regulation in *T. gondii* and highlighted the importance and potential of proteogenomic approaches in genome annotation process. The importance of temporal and quantitative proteomics as well as the future of systems biology has been discussed.

List of Tables

List of Figures

List of Abbreviations

1-DE	One-dimensional gel electrophoresis
2-DE	Two-dimensional gel electrophoresis
AIST	National Institute of Advanced Industrial Science and Technology
AMA1	Apical membrane antigen 1
Ambic	Ammonium bicarbonate
BAC	Bacterial artificial chromosome
BLAST	Basic Local Alignment Search Tool
BlastP	Protein-protein Blast
BVA	Biological Variation Analysis
CAGE	Cap analysis of gene expression
CAN	Acetonitrile
CDS	Coding sequence
cICAT	Cleavable isotope-coded affinity tags
CID	Collision-induced dissociation
DAS	Distributed annotation system
DHFR-TS	Dihydrofolate reductase-thymidylate synthase
DIA	Differential In-gel Analysis
DIGE	Differential gel electrophoresis
DTT	Dithiothreitol
ECD	Electron-capture dissociation
ER	Endoplasmic reticulum
ESI	Electrospray ionization
EST	Expressed sequence tag
FA	Formic acid

F-actin	Filamentous actin
FDR	False discovery rate
FTICR	Fourier-transform ion cyclotron resonance
G6P	Glucose 6-phosphate
GAP	Gliding-associated protein
Gbrowse	Genome browser
glycerate-2P	Glycerate-2-phosphate
GO	Gene Ontology
GPM	Global Proteome Machine Organization
GPMDB	Global Proteome Machine databases
GRA	Dense granule protein
HMM	Hidden Markov Model
HOST	Host organelle-sequestering tubulo structure
HUPO-PSI	Human Proteome Organization-Proteomics Standards Initiative
IAA	Iodoacetamide
ICPL	Isotope Coded Protein Label
IEF	isoelectric focusing
IF	Intermediate filament
IMC	Inner membrane complex
iTRAQ	Isobaric tags for relative and absolute quantification
KEGG	Kyoto encyclopedia of genes and genomes
LC	Liquid chromatography
Lys-C	Lysine-C
m/z	mass-to-charge ratio
M2AP	MIC2-associated protein

MALDI	Matrix-assisted laser desorption/ionization
MIAPE	Minimum information about a proteomics experiment
MIC	Micronemal protein
MIPS FunCat	Munich Information Centre for Protein Sequences Functional Catalogue
MJ	Moving junction
MNN	Membranous nanotubular network
MPSS	Massively parallel signature sequencing
MRM	Multiple reaction monitoring
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MT	Microtubule
MudPIT	Multidimensional protein identification technology
NCBI	National Centre for Biotechnology Information
NLS	Nuclear localization signal
OAA	Oxaloacetate
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PATS	Prediction of apicoplast targeted sequences
PEP	Phosphoenolpyruvate
PFD	Prefolding dissociation
PFF	Peptide fragment fingerprinting
pI	Isoelectric point
PMF	Peptide mass fingerprinting
PRIDE	Proteomics identifications database

PTM	Post-translational modification
PV	Parasitophorous vacuole
PVM	Parasitophorous vacuole membrane
Q	Quadrupole
QIT	Quadrupole ion trap
RF	Radio frequency
RNA-Seq	RNA sequencing
RON	Rhoptry neck protein
ROP	Rhoptry (bulb) protein
RP	Reversed phase
SAG	Surface antigen
SAGE	Serial analysis of gene expression
SCX	Strong cation exchange
SDS	Sodium dodecyl sulfate
SNP	Single-nucleotide polymorphisms
SPC	Seattle Proteome Center
SRS	SAG related sequence
TCA	Tricarboxylic acid
TIGR	Institute for Genomic Research
TM	Transmembrane
TOF	Time-of-flight
TPP	Trans-Proteomic Pipeline
XIAPE	Xml Information about a proteomics experiment
XML	Extensible markup language
YPD	Yeast Protein Database

Acknowledgements

First of all I would like to express my sincere gratitude to Prof. Jonathan M Wastling for his extraordinary support and guidance, great kindness and constant encouragement throughout my PhD studies. Further thanks are due especially to Dr. Sanya J Sanderson, who has spent a great deal of time giving me support, encouragement, advices on scientific and cultural aspects.

I would like to thank Prof. Rob Beynon, Dr. Andy Jones, Dr. Phillip Whitfield, Dr. Duncan Robertson and Prof. Diana Williams for their valuable support and inspiring discussions; also to Sophie, Beccy, Andy, Sidikki, Nadine, Hos, Emma, Jenna, Sarah, Gianluca and everyone from the Veterinary building for the precious care, assistance and friendship received throughout my PhD studies.

I am very grateful to all the collaboraters, from whom I have gained technical support and valuable discussions: Mark Heiges, Brian Brunk and Bindu Gajria from EuPathDB team; Judith Prieto from the Yates's lab; Arnab Pain, Amandeep Sohal from the Wellcome Trust Sanger Institute and Dhanasekaran Shanmugam from University of Pennsylvania.

I would also like to thank all the good friends I made in Edinburgh and Liverpool, Tijana, Jo, Nicola, Stan, Ken, Dave, Mario, Lee, Sanjay, Peng, Maya, Matt and Laura, Tali and Chris, Lucy and Daniel, Agnes and Martinho... Their love and kindness have beautifully enriched my life in Britain.

Finally, I am forever bonded and indebted to my parents and Anna, for their enormous love, understanding and encouragement.

Author's Declaration

The work presented in this thesis was performed solely by the author except where the assistance of other has been acknowledged.

Dong Xia, November 2009

Chapter 1

Introduction

1.1 *Toxoplasma gondii*

In 1908, Nicolle and Manceaux at the Pasteur Institute in Tunis discovered a parasite in North African rodents, the gundi, *Ctenodactylus gundi* [1]. In the same year, Splendore discovered the same parasite isolated from a rabbit in Brazil [2]. Both groups initially identified this parasite as *Leishmania* before the current name *Toxoplasma gondii* was proposed by Nicolle and Manceaux after extensive microscopic analysis of several tissues and experimental studies in 1909 [3].

In the last 100 years, research on *T. gondii* has been carried out in many scientific disciplines consistently expanding the understanding of this parasite. *T. gondii* is an obligate intracellular protozoan parasite which infects all warm-blooded animals [4]. Humans acquire *T. gondii* through ingestion of undercooked meat, contact with feline faeces, through drinking contaminated water and through transplantation of a contaminated organ [5]. In humans, *T. gondii* is frequently associated with congenital infection and abortion [6]. Infections of *T. gondii* are usually minor and self-limiting but it can cause encephalitis or systemic infections in the immune-compromised, particularly individuals with HIV/AIDS [7].

T. gondii is a member of the phylum *Apicomplexa*, class *Sporozoa* and subclass *Coccidia*. All members of *Apicomplexa* share a common feature of the presence of an apical complex in one or more stages of the life cycle. The *Apicomplexa* include a number of medically and agriculturally significant pathogens such as *Plasmodium, Cryptosporidium, Theileria, Neospora* and *Eimeria*. In common with the other *Apicomplexa, T. gondii* has a complex life-cycle with multiple life-stages.

1.1.1 Life cycle of *T. gondii*

T. gondii is a tissue cyst-forming coccidium with a heteroxenous life cycle. Felidae, for example domestic cats, are the only definitive hosts for *T. gondii* [8] and all other warm-blooded animals are intermediate hosts. There are three infectious stages of *T. gondii*: the tachyzoites, the bradyzoites and the sporozites. They are linked in a complex life cycle which is illustrated in Figure 1.1.

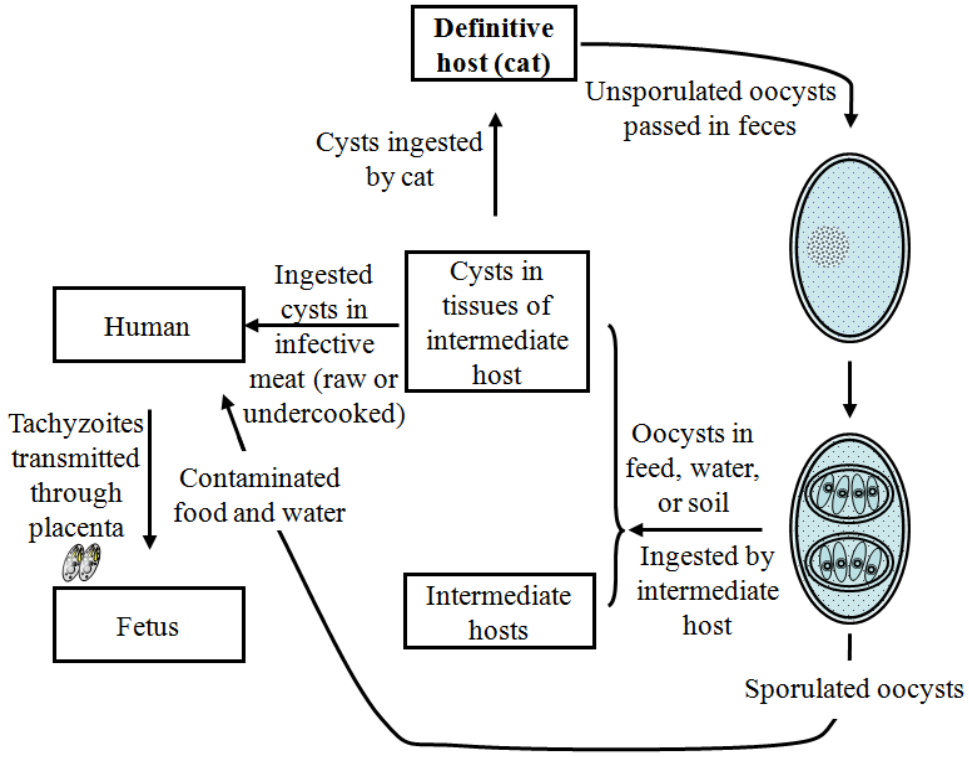


Figure 1.1 Life cycle of *T. gondii* (Adapted from Dubey JP [9])

1.1.1.1 Life cycle in the intermediate hosts

In intermediate hosts, *T. gondii* undergoes two phases of asexual development, the tachyzoite and bradyzoite.

1.1.1.1.1 Tachyzoite

The term “tachyzoite” (Greek: tachos = speed) was proposed by Frenkel [10] to reflect the rapidly multiplied and invasive nature of this life stage. Tachyzoites invade host cells by actively penetrating through host cells (see section 1.1.3). Tachyzoites multiply asexually within the host cell by a repetitive specialized process called endodyogeny, in which two progeny form within the parent parasite [11, 12].

The tachyzoite is crescent-shaped, approximately 2×7 μm with a pointed anterior end (defined by the direction of motility). It consists of a unique cytoskeleton (subpellicular microtubules, conoid, inner membrane complex), secretory organelles (rhoptries, micronemes, dense granules), endosymbiotic derived organelles (mitochondrion, apicoplast), eukaryotic universal organelles (nucleus, endoplasmic reticulum, Golgi apparatus, ribosomes), and specific structures (acidocalcisomes), all enclosed by a complex membranous structure termed the pellicle [13, 14].

In response to the stress of the host cell immune system, the tachyzoites convert to a slow growing form called bradyzoites. Stage differentiation from tachyzoite to bradyzoite is accompanied by a major shift in the antigenic expression profile and alterations to metabolism of the parasites [15-17].

1.1.1.1.2 Bradyzoite and tissue cysts

The term “bradyzoite” (Greek: brady = slow) was also proposed by Frenkel [10] to describe the encysted stage of the parasite in tissues. Bradyzoites multiply slowly by endodyogeny in tissue cysts [15, 18]. Tissue cysts vary in size, young tissue cysts can be as small as 5 μm in diameter and contain only two bradyzoites [19] while

older ones can contain hundreds of bradyzoites and be elongated to 100 μm long [20].

The crescent-shaped bradyzoites are 5-8.5×1-3 μm in size. Structurally, the bradyzoite only differs slightly from the tachyzoite. Bradyzoites are more slender than tachyzoites, and the nucleus is situated more towards the posterior end compared to the centrally located tachyzoite nucleus. Bradyzoites contain several amylopectin granules which stain red with PAS reagent where such material is either in discrete particles or absent in tachyzoites [20].

Tissue cysts are the final life-cycle stage in the intermediate host cells. The parasite may cause immediate toxoplasmosis or remain latent in the host for life. It has been found that a very small percentage (2 of 750, 0.27%) of tissue cysts rupture at any time during chronic infections and result in stage conversion back to the active tachyzoite form [21]. This interconversion mechanism is important to maintain a chronic infection and also causes disease reactivation in people with immune-deficiencies, such as AIDS or malignancies [19, 22].

1.1.1.2 Life cycle in the definitive host

Cats are the definitive host of *T. gondii*. They can be infected by ingesting oocysts, tachyzoites, bradyzoites or transplacentally. The bradyzoite-induced cycle is well-characterized [23, 24]. After the ingestion of tissue cysts by cats, proteolytic enzymes in the stomach and small intestine dissolve the tissue cyst wall and bradyzoites are released. Bradyzoites then penetrate the epithelial cells of the small intestine and initiate the development of numerous generations of *T. gondii*, known as asexual enteroepithelial development [20, 24].

1.1.1.2.1 Asexual development

Five morphologically distinct types of *T. gondii* develop in intestinal epithelia cells in this stage and are designated types A to E [24]. Rather than generations conventionally known as schizonts in other coccidian parasites, "type" is used because there are several generations within each *T. gondii* type (see Figure 1.2).

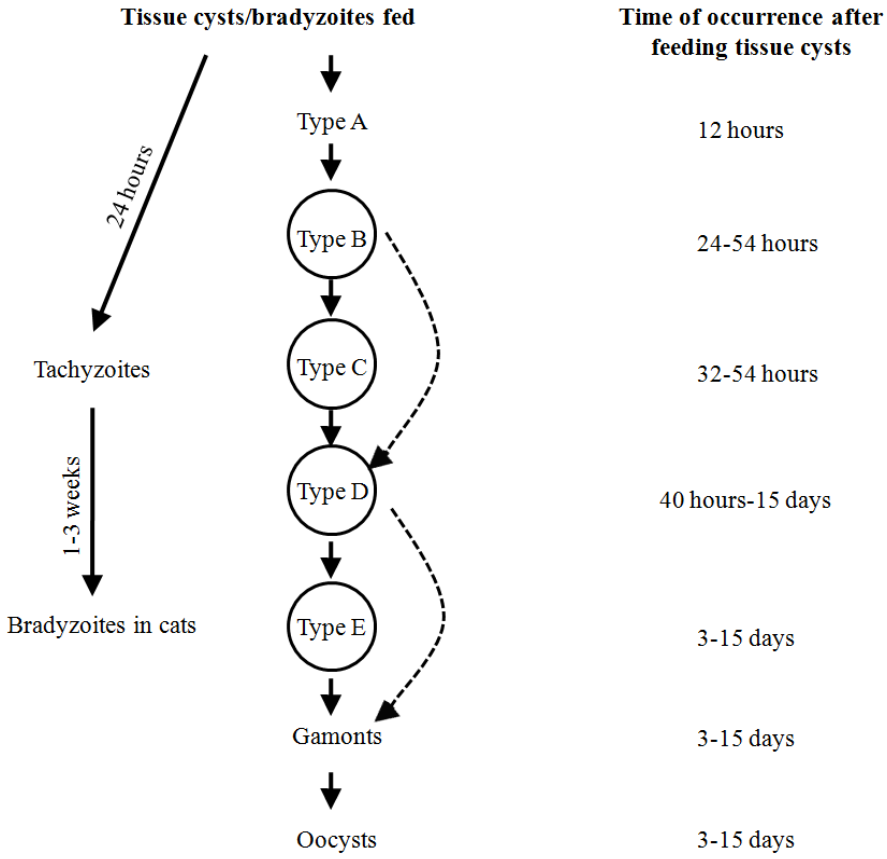


Figure 1.2 Coccidian cycle of *T. gondii* (Adapted from Dubey JP *et al.* [24])

These types were morphologically distinguishable from tachyzoites and bradyzoites that also occur in the cat intestine. Little is known about the structure or biology of type A [19]. Type B schizonts formed merozoites by endodyogeny [25], which has a similar relationship to that described for parasites invading the small intestine of the intermediate host [26, 27]. Type C, D, and E multiply by schizogony, which is also termed as endopolygeny [28]. In schizogony, the nucleus divides two or more times

without cytoplasmic division. Before or simultaneous with the last nuclear division, daughter organism (merozoite) formation is initiated near the centre of the schizont. The merozoites often remain attached to a small amount of residue cytoplasm at the posterior end; some merozoites are released from the host cell into the lumen, where they can reinvade enterocytes [24, 25].

1.1.1.2.2 Sexual development

After the asexual development (types A-E), the sexual cycle starts. Merozoites released from the host cell reinvade new enterocytes and develop into either male (microgametocyte) or female (macrogametocyte) gametocytes [13]. In microgametogony, 15-30 male gametes (microgametes) are produced [29, 30]. Only one female gamete (macrogamete) is formed in macrogametogony [31, 32].

Microgametes use their flagella to swim to and penetrate and fertilize mature macrogametes to form zygotes. That fertilization can occur has been proven from the identification of cross-fertilized parasites [33], however, the necessity and mechanism of fertilization remain uncertain [34]. After fertilization, an oocyst wall is formed around the parasite. Infected epithelial cells rupture and discharge oocysts into the intestinal lumen.

1.1.1.2.3 Oocyst shedding and extracellular sporulation

Cats can shed millions of oocysts in a few days [24, 35]. The oocyst is the only stage of *T. gondii* that can undergo extracellular development. Oocysts are excreted in an unsporulated form and then form two sporocysts by asexual development (sporulation). Each sporocyst contains four sporozoites [13, 36-38]. Sporulated oocysts can survive for long periods under moderate environmental conditions before ingestion by an intermediate host [39, 40].

1.1.2 Population structure of *T. gondii*

Three different invasive forms make *T. gondii* a remarkably successful parasite. The rapid invasion of tachyzoites into virtually all types of animal cells followed by the ability of chronic infection maintained mainly in bradyzoites form and the spread of sporozoites shed in environmental resistant oocysts all contribute to its global distribution, and this suggests a large genetic diversity.

The first genotyping studies on *T. gondii* in North America and Europe was published 15–20 years ago, led to the description of a clonal population structure of *Toxoplasma* with three main lineages, types I, II and III [41-43]. The clonal nature of three *T. gondii* lineages was confirmed by the high-fidelity sequencing of single-copy genetic regions comprising both antigens and introns [44]. Within each lineage, the true rate of divergence were extremely low at less than 1 in 10,000 bp [45] and the genome wide polymorphism rate between three lineages has been estimated to be 0.65% [46]. The absence of diversity within each of the three lineages and low divergence between lineages suggest the three lineages merged as the dominant strains relatively recently. Based on the assumption that mutations arise at a constant rate, it was estimated the three lineages were expanded from a common ancestor 10,000 years ago [44]. The pattern of single-nucleotide polymorphisms (SNP) in the *T. gondii* genome also produced a model that types I and III strains are second and first generation offspring of a cross between a type II strain and one of two ancestral strains [46].

Strains which differ from the three well defined lineages have been described more frequently from more remote geographic areas [47, 48] and several studies on strains from South America revealed that they are genetically distinct from the Eurasian strains studied [49-52]. Combining *T. gondii* strains from Europe, North and South

America in a phylogenetic analysis of polymorphism in introns, a new picture of population structure is given. Eleven distinct haplogroups were recently defined with three major clonal lineages renamed as haplogroups 1, 2 and 3, and notably, all existing haplogroups are predicted to be derived from admixture of four ancestral lineages [53].

Although genetic diversities between the clonal lineages are small, they underlie large phenotypic differences that influence virulence. Studies in human cases (congenital, transplant and AIDS) from North Amedica and Europe indicate a major association with type II strains [42, 48, 54]. Type III strains are only seen in a few human cases with underlying immunodeficient state and type I strains have increased prevalence in some cases of congenital infection and AIDS patients [55, 56]. In mouse models, type I strains have been shown as highly virulent and type II and III strains are considered non-pathogenic [57]. Different degrees of mouse-virulence were shown in different clonal groups in South America [58]. Large molecular-epidemiological studies will be needed to understand the distribution and virulence of different genotypes across the world, which will be important to develop intervention strategies and disease control.

1.1.3 Host cell invasion, intracellular survival of tachyzoites and the role of related secretory organelles

T. gondii is an obligate intracellular parasite; being able to invade a new host cell is crucial for survival and expansion of infection. During infection, most intracellular microbes and parasites take advantage of endocytosis or phagocytosis as a means of entry into cells [59, 60]. In contrast, *T. gondii* does not rely on existing routes, but uses its own active, parasite-driven penetration process [61].

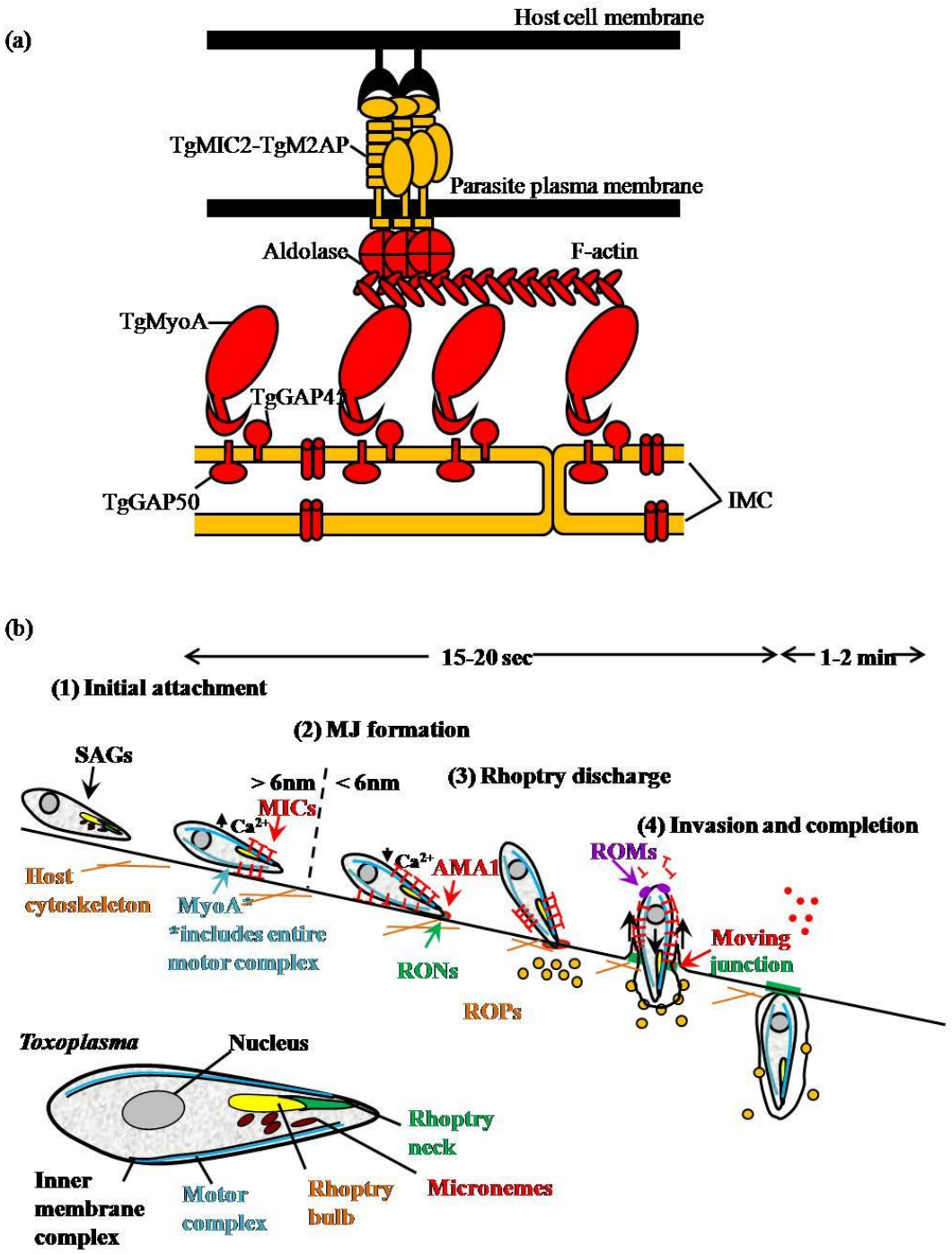


Figure 1.3 Host cell invasion process of tachyzoites (a) Model of the glideosome system of *T. gondii*. As described in greater detail in section 1.1.3.1, the glideosome is the molecular machine that promotes gliding motility. (Adapted from Keeley A *et al*.[62]) (b) An integrated working model of *Toxoplasma* invasion. As decribed further in sections 1.1.3.2-5, the invasion process involves multiple steps including: initial attachment, formation of moving junction, injection of ROP proteins and the completion of invasion. (Adapted from Carruthers V *et al*.[63])

1.1.3.1 Glideosome

T. gondii, as well as other Apicomplexan parasites, utilizes a substrate-dependent locomotion called gliding motility for tissue migration and cell invasion [62, 64] (see Figure 1.3a). Gliding relies on a linear motor system (glideosome) in the pellicle of *T. gondii* between the outer plasma membrane and inner membrane complex (IMC). A class XIV myosin (MyoA), the myosin light chain and two gliding-associated proteins (GAPs) form the motor complex [65]. The motor complex is anchored in the outer membrane of the IMC and connected via filamentous actin (F-actin) and glycolytic enzyme aldolase. Aldolase links the complex to the cytosolic domain of transmembrane adhesive proteins (adhesins) that spans the outer plasma membrane [62]. Tachyzoites use this actin-based motility coupled with regulated protein secretion from apical organelles to actively invade host cells.

1.1.3.2 Initial attachment

An integrated multi-step working model has been shown in Figure 1.3b. The initial attachment of tachyzoites to the host cell surface is mediated by a family of GPI-anchored surface antigens (SAGs) and SAG related sequence (SRS) proteins [66, 67]. A wide distribution of SAG and SRS proteins coated on the parasite surface [68], among which a few are known to dominate the surface of tachyzoites, SAG 1-3 and SRS 1-3 [69]. SAG and SRS proteins provide a low-affinity and lateral interactions between the parasite and host cell surface, which it has been postulated may allow the parasite to survey the host cell surface for an optimal invasion site [63].

The tachyzoite initiates penetration into the target cell exclusively using its apical end. Several lines of evidence suggest micronemes play an important role in the process. Firstly, many micronemal proteins (MICs) have been shown to migrate out to the apical surface of parasite during attachment under a calcium signal dependent

mechanism [70]. Secondly, many MICs possess domains that mediate protein-protein or protein-carbohydrate interactions (for example thrombospondin-like, epidermal growth factor-like, lectin-like and so on) [71]. Finally, the genetic depletion of MIC genes such as MIC1 and MIC3 [72], MIC2 and MIC2-associated protein (M2AP) complex [73-75] substantially attenuates parasites invasion. The study of MIC1 and MIC3 also revealed that individual disruption of either MIC1 or MIC3 expression slightly reduced virulence in the mouse, whereas doubly depleted parasites are non-virulent and fail to produce a lethal infection [72]. Together, this evidence suggests that the parasite expresses a variety of adhesive MIC proteins to target a wide range of host cells and create a robust binding-interface by using multiple receptors.

1.1.3.3 Moving junction

Almost simultaneously as microneme secretion, a calcium-dependent conoid extrusion is observed [63, 76]. The precise role of conoid extrusion is unclear, it could serve to bring the apical tip closer to the host plasma membrane [63] or be an important requisite for microneme secretion [76]. At around the same time of these events, another attachment step is executed which depends on the expression of a micronemal protein apical membrane antigen 1 (AMA1) [77]. During or after secretion, AMA1 forms a stable complex with three rhoptry neck (RON) proteins (RON2, RON4, RON5), which together occupy a structure known as the moving junction (MJ) [78, 79].

The MJ is a ring-like band at the intimate contact of host and parasite plasma membranes. The first role of the MJ is likely to be an anchor, which allows the actin-myosin motor to “pull” the parasites into the nascent vacuole in the host cell. AMA1 is secreted from micronemes onto the plasma membrane and anchored by a

transmembrane domain. While RON4 and RON5 do not have transmembrane domains, RON2 has three predicted transmembrane domains that may act as a bridge by inserting into the host plasma membrane [80]. As invasion proceeds, the MJ migrates from anterior to the posterior end and forms the border or rim of the nascent parasitophorous vacuole (PV). This brings into play the other role of the MJ which is likely to act as a molecular sieve that selectively removes parasite proteins and restricts access of host proteins to the forming vacuole [81, 82]. All of the known MIC proteins other than AMA1 and other transmembrane proteins anchored in the cytoskeleton are selectively excluded suggesting the filtering takes place on the cytoplasmic face of the plasma membrane. Many proteins gain access to the vacuole by partitioning into lipid rafts which raises the possibility that the MJ actively orders lipids within the bilayer and thus influence the protein composition of the vacuolar membrane [80].

1.1.3.4 ROP proteins are injected into host cells

Simultaneously or immediately thereafter the MJ is formed, the proteins located in rhoptry bulb (ROPs) are injected to the host cytoplasm [83]. A detailed proteomic analysis of the contents of the purified rhoptries provided a comprehensive list of RON and ROP proteins [84]. Thirty eight novel proteins were identified and the location of 11 out of 12 novel proteins was verified as rhoptry by the production of antibodies [84]. Furthermore, rhoptry proteins were distinguished according to their sub-organellar location as either bulb (ROP) or neck (RON). This study found that all of the RON proteins identified have homologues in *Plasmodium* which suggests their involvement in processes that are common to the *Apicomplexa* phylum [80, 84]. On the contrary, nearly all of the ROPs are unique to either *Toxoplasma* or *Plasmodium* indicating that ROPs are highly adapted to the intracellular niches that

they occupy [80, 84]. In *Toxoplasma*, ROPs migrate to one of three general locations after injection: the lumen of the nascent PV, the parasitophorous vacuole membrane (PVM) or the host cell nucleus.

ROP1 is released during invasion and accumulates within the lumen of the nascent PV [85]. Interestingly, ROP1 synthesized in one parasite can migrate to the PV of another parasite [86]. The ROP2 family of proteins generally migrates to the host cytosolic side of the PVM [87-89]. The ROP2 family contains a conserved serine/threonine (S/T) kinase domain, although most other members lack key residues predicted to be necessary to phosphorylate proteins [80, 90]. Recently, a highly polymorphic member protein ROP18 was confirmed as having kinase activity and could have potential roles in parasite growth and virulence [91, 92]. ROP2 has also been suggested to be involved in the recruitment of host cell mitochondria by resembling a mitochondrial-import signal [88, 93]. Recent studies pointed out a third destination of ROPs; two rhoptry proteins, PP2C-hn [94] and ROP16 [95] have been observed in the host cell nucleus. PP2C-hn is a protein phosphatase of the 2C class (hn is the abbreviation for host nucleus) and ROP16 is a putative protein kinase; both proteins are directed to the nucleus by a conventional nuclear localization signal (NLS) [89]. ROP16 is likely to modulate host signalling by indirectly inducing the activation of the signal transducer and activator of transcription (STAT 3/6) signalling pathways with consequent effects on host IL-12. Furthermore, ROP16 has shown strain-specific polymorphism where the allele shared by types I and III is effective in mediating sustained phosphorylation of STAT3, while the allele found in type II does not [95]. This interaction may in part explain the much lower levels of IL-12 that are induced by type I or III versus type II strains following infection of

macrophages and subsequently the differences of *Toxoplasma* strain virulence [95, 96].

RONs and ROPs are essential to the host cell invasions and PV establishment, however, the characterization of molecular interactions and biological functions are still at an early stage where great efforts are being carried out using genetic and cellular tools [80, 89].

1.1.3.5 Completion of invasion

It takes only 15-20 seconds before the MJ completes its migration [97]. The invasion arrives to its final stage: pinching off and separation of a complete PVM. The process involves fission of the PVM and host plasma membrane that can take as long as two minutes. The exact mechanism is unclear, for example how are the residual MJ complex or other parasite and host fission proteins involved [63].

1.1.3.6 Intracellular survival

Once inside the host cell, the PV is used as a platform for tachyzoites to modulate several host cell functions that support parasite replication and lead to a long-term chronic infection. Mitochondria and endoplasmic reticulum (ER) from the host cell rapidly surround the PVM and are recruited in the supply of a dedicated nutrient source for the support of parasite division [98]. The host intermediate filaments (IFs) and microtubules (MTs) are also reorganized around the PV providing intracellular localization and structural integrity support [99, 100]. Functionally, host cell modifications induced by the parasites includes the inhibition of inflammatory host responses, interference with regulatory and effector functions of immune cells, and manipulation of host cell apoptosis [101].

Three functionally distinct membranous subcompartments are organized in the PV [102]. The PVM is a thin membranous extension prolonged within the host cell cytosol. The second subcompartment consists of host organelle-sequestering tubulo structures (HOSTs), which are microtubule-based PVM invaginations that channel cholesterol-enriched host endolysosomes into the PV [99]. The third subcompartment is the membranous nanotubular network (MNN), which extends into the vacuolar lumen that link the parasites together and to the PVM [103, 104]. MNN is thought to maintain parasites in an ordered arrangement within PV that allows their synchronous division [102].

Up to now, the *Toxoplasma* proteins identified in the PV originate from either the rhoptries or the dense granules[102]. In addition to ROPs that are injected into the host cell during invasion (section 1.1.3.4), a burst of dense granule proteins (GRAs) are secreted into the PV within the first few minutes of PVM formation [83, 105]. GRA1 is a calcium-binding protein and the only completely soluble GRA protein within the vacuolar space [106]. All other GRAs associate with distinct PV membranous sub-compartments: GRA2, 4, 6, 9 and 12 associate with MNN, while GRA3, 5, 7 and 8 associate predominantly with the PVM and GRA7 with the HOSTs [102, 107]. GRAs share no obvious homology with proteins of known function but are likely to be involved in intracellular parasite development and multiplication and their secretion is likely to continue on a basal level as long as parasite multiplication takes place [102].

1.1.4 *T. gondii* genome

The first output of *T. gondii* genome sequencing project was finished in 2003 [108]. The 10 × shotgun genome sequencing and annotation of the type II strain ME49 was performed by the *Toxoplasma* Genome Consortium, under collaboration between Ian

Paulsen from the Institute for Genomic Research (TIGR) and David Roos from the University of Pennsylvania, and led to a draft version of the 80 Mb genome sequence [108, 109]. In addition to this effort, the Wellcome Trust Sanger Institute determined the 5 × shotgun sequencing of type I strain RH chromosome 1a and 1b as well as a whole genome bacterial artificial chromosome (BAC) library [110]. The type II ME49 strain was the first strain to be sequenced followed by two other strains, GT1 and VEG; all data are available on ToxoDB [108].

A genetic linkage map was used to assemble the gene coordinates from scaffolds to 14 chromosomes of ME49 strain [111]. The chromosome map for ME49 strain was used as a template to create GT1 and VEG strain chromosomes [108]. Gene prediction programs including GLEAN, GlimmerHMM, TigrScan and TwinScan were used for ME49 strain genome annotation in 2003, followed by the release of a major integrated and updated version on ToxoDB (version 4) in 2006 [108]. In the latest release of ToxoDB (version 5), genome sequences of the three strains are between 61-64 Mb in size. A newer version of genome annotation for ME49 strain is also updated together with a brand new genome annotation for GT1 and VEG strains [108]. The numbers of genes for *T. gondii* are currently predicted to be 8102 for the ME49 strain, 8145 for the GT1 strain and 7945 for the VEG strain [108].

All the raw genomic sequences and genome annotations of the latest version of genome sequencing project can be downloaded at ToxoDB [112]. This valuable information provided the foundation of many genome-wide researches of *T. gondii* including the whole proteomic analysis of tachyzoites carried out in this study.

1.2 Proteomics

Proteins are the end products of most genes, the identification of expressed proteins and their functions are central to the understanding of biological meaning of the genome. The term proteome was introduced by Marc Wilkins in 1994 at a conference [113] to describe the entire complement of proteins expressed by a genome, cell, tissue or organism. More specifically, it is the set of expressed proteins at a given time under defined conditions. Papers that began to use the term were published thereafter [114, 115].

The biological importance and interest of studying global protein expression and understanding functional genomics have driven the field of 'proteomics' which has rapidly progressed in recent years. Around 200 research articles and reviews were published each month in the first half of 2009. The rapid growth of proteomics has been supported by technical advances in key areas of proteomics research.

To start with, protein separation techniques were employed such as 2-D electrophoresis and 2-D liquid chromatography (LC) to deal with the complicated protein samples. The next step has benefited from advances in mass spectrometry in which the sensitivity and capability of measuring mass-to-charge ratio (m/z) of ionized peptides and proteins are consistently improving. The final step took advantage of the development of dedicated bioinformatics packages that analyse, visualise and interpret proteomic data to achieve a better understanding of the biological system under study. Simultaneously, the flood of draft and complete genome sequencing projects and new high-throughput sequencing platforms provided essential sequence databases for proteomics data searching.

Proteomics add such a value to biological research that it shifts the paradigm from a 'one-protein-at-a-time' view to a new 'global' view. Hundreds of expressed proteins can be identified and characterized in a high throughput manner that was not possible before. The following sections will discuss the current developments and applications of proteomics in detail.

1.2.1 Proteomic strategies for high-throughput protein identification

Several proteomic strategies have been developed to achieve high-throughput protein identifications. Typically, they can be divided into two groups, 'top-down' and 'bottom-up'. 'Top-down' proteomics is a straight forward approach in which intact proteins are analysed by MS. Due to the instrument limitations, analysing intact protein samples is more difficult to achieve by MS. 'Bottom-up' proteomics is a more commonly used strategy which analyses enzymatically or chemically produced peptides of the protein samples (See Figure 1.4).

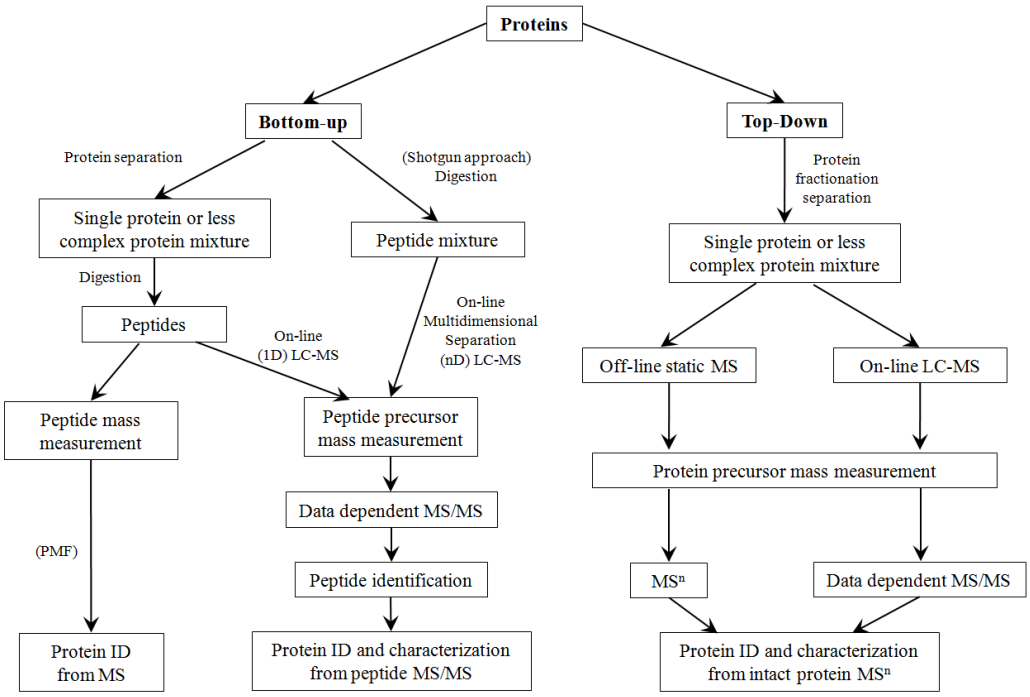


Figure 1.4 Proteomic strategies for protein identification (Adapted from Han et al. [116])

1.2.1.1 Bottom-up proteomics

Bottom-up proteomics uses information acquired from peptide detection to infer the protein identification. According to the method by which the peptide mixture is produced, this approach can be further divided into two strategies.

The first strategy is to separate the complex sample in protein space, by using protein fractionation and separation techniques such as gel electrophoresis. The resulting single protein or appreciably less complex protein mixture is then digested into peptides. The peptides can be analyzed by peptide mass fingerprinting (PMF) or following further peptide separation by LC coupled to a tandem mass spectrometer [116, 117].

The other strategy is to separate the complex sample in peptide space, which is also known as 'shotgun proteomics'. Protein digestion is carried out without or with minimum sample pre-fractionation. The peptides produced are separated by multidimensional chromatography and analyzed by tandem mass spectrometry [117, 118].

1.2.1.2 Top-down proteomics

Top-down proteomics is a relatively new strategy which is mainly restricted to the use of more powerful Fourier-transform ion cyclotron resonance (FTICR) mass spectrometers. In this approach, intact proteins are ionized and fragmented inside the mass spectrometer and high-resolution mass measurement is made [119].

FTICR utilises a new type of fragmentation technique called electron-capture dissociation (ECD). In ECD, the recombination of an electron with the multiply protonated peptide/protein ion makes differences in bond dissociation energies less important and induces protein backbone cleavages [120]. This technique provides

great fragmentation efficiency for small to medium sized proteins [121-123]. For protein ion molecules larger than ~ 50 kDa, protein tertiary structures become more complex with many non-covalent interactions which reduce the fragmentation efficiency and make top-down proteomics inefficient [124]. Recent developments to tackle this problem include prefolding dissociation (PFD) that dissociate ~240 residues from each terminus of protein [125] or using limited digestion to produce larger peptides (>5 kDa) [126].

Top-down proteomic approaches examine the entire protein sequence directly which enables a more complete characterization of protein isoforms and post-translational modifications (PTMs) [127, 128] and lead to exciting *de novo* sequencing [129]. However, the requirement for direct infusion of a single protein or simple protein mixture remains a major challenge for large-scale high throughput proteome characterization. Kelleher's group presented the first large-scale top-down proteomics on a LTQ-FTICR system that uses high-resolution MS/MS data obtained on a chromatographic time scale [130]. The study identified 22 proteins from a single LC-MS/MS run and 38 proteins were unambiguously identified in metabolically labelled proteome experiments.

Despite the rich protein characterization information provided by the top-down proteomic strategy, the sophisticated instrumentation setup, technique restrictions and relatively small protein identification yields under the current development stage mean that bottom-up proteomics is still the most suitable strategy for large-scale protein identification projects. The following sections will discuss the bottom-up proteomic approach in more details.

1.2.2 Sample preparation and separation

Sample preparation and fractionation are crucial in a successful proteomic analysis. The separation and fractionation of a sample benefits downstream applications enormously by improving the resolution of analysis. Proteins in the sample need to be extracted from cells, denatured, disaggregated and solubilised before being analysed by proteomic approaches.

Cell lysis can be achieved by osmotic lysis, detergent lysis, enzymatic lysis of the cell wall, mechanical blending, sonication, freeze/thaw and manual grinding [131]. During or after cell lysis, interfering compounds such as salts, proteolytic enzymes, nucleic acids, polysaccharides, lipids and particulate material must be diluted, inactivated or removed.

As discussed in section 1.2.1, before complex protein samples can be analysed by MS, sample separation needs to be achieved in either protein space or peptide space. Sample separation in protein space is typically achieved by gel electrophoresis followed by enzymatic digestion. Sample separation in peptide space is achieved by direct digestion of complex sample followed by liquid chromatography separations.

1.2.2.1 Sample separation in protein space

One-dimensional and two-dimensional gel electrophoresis (1-DE and 2-DE) are commonly used in proteomic studies. They separate solubilised proteins samples according to their physical properties and provide visualization representations of the proteome studied.

1-DE is a well established widely used technique which separates proteins based on their molecular mass using SDS-polyacrylamide gel electrophoresis (SDS-PAGE) [132]. Proteins are reacted with the anionic detergent SDS to form negatively

charged complexes. Denatured proteins bind to SDS in a relationship proportion to mass and independent of amino acid composition and sequence. Proteins are separated by the polyacrylamide matrix on the basis of molecular mass.

2-DE couples isoelectric focusing (IEF) in the first dimension with SDS-PAGE in the second dimension, and enables protein separation on the basis of isoelectric point (pI) followed by molecular mass [131]. 2-DE can resolve thousands of proteins simultaneously and detect < 1 ng of protein per spot [131].

Separated proteins on the gel need to be visualized by staining methods. One big challenge is the huge differences of protein abundance in the sample, which can range between 7-8 orders of magnitude in a cell [133]. High sensitivity, high linear dynamic range and compatible with downstream protein identification procedures are required for a good staining method. Commonly used methods include organic dyes such as colloidal Coomassie Blue (detection limit ~30 ng) [134], negative stain with metal cations such as zinc chloride (20-50 ng) [135], silver stain such as silver nitrate stains (5-10 ng) [136] and fluorescence stain such as SYPRO Ruby (1-2 ng) [137]. This enables a frozen-in-time view of the proteome which reflects changes in protein isoforms, PTMs and expression levels.

Bands and spots separated by 1-DE and 2-DE are excised followed by in-gel protein digestion. Trypsin is the most widely used enzyme for protein identification based proteomics. It cleaves specifically after arginine or lysine residues, producing peptides with an average size of 800-2000 Da [138].

1.2.2.2 Sample separation in peptide space

Shotgun proteomics separates samples in peptide space. Complex samples are directly digested using one or a combination of proteases such as Lysine-C (Lys-C),

V8 and trypsin. One combination is using Lys-C followed by trypsin. Lys-C fully preserves its activity in the presence of 8 M chaotropic agent urea, the accessibility of this enzyme to peptide bonds is higher than the accessibility of trypsin, which remains active up to 2 M urea.

Sample digestion results in a highly complex of peptide mixture. The peptides are then separated by multidimensional liquid chromatography before analysis by MS/MS. A typical setup using this strategy is called multidimensional protein identification technology (MudPIT) [139, 140]. MudPIT couples a strong cation exchange (SCX) column with a reversed phase (RP) column to separate peptides using 2D liquid chromatography. Peptides are displaced from SCX to the RP column using 12 salt step gradients. The RP column then progressively elutes peptides over a gradient of acetonitrile (ACN) with increasing hydrophobicity. Peptides are then analysed by MS/MS which is reviewed in section 1.2.3.

Although a MudPIT experiment requires less labour with simple automated setup, it results in greatly increased complexity of the generated peptide mixture. Achieving a good result requires highly sensitive and efficient separation. Information is also lost upon the conversion of intact proteins into a mixture of peptides, and it also involves significantly more computer power for data analysis.

1.2.3 Mass spectrometry used in protein identification based proteomics

MS is a technology that measures the mass-to-charge ratio (*m/z*) of molecules. Mass spectrometers consist of three key components. An ion source that converts molecules into gas-phase ions, a mass analyzer that separates charged molecules according to their *m/z* and a detector that records the number of ions at each *m/z*

value. Various ionization sources and analyzers can be combined to facilitate proteomic research.

1.2.3.1 Ionization

Two soft ionization methods are used in proteomics due to their ability to produce intact ions from peptides and proteins, matrix-assisted laser desorption/ionization (MALDI) [141] and electrospray ionization (ESI) [142].

In MALDI, matrix (aromatic acids) is used to protect the analytes from being destroyed by laser light and to assist vaporization and ionization. The analytes are embedded into a crystalline matrix on a metal 'target' plate. The target is then placed in the vacuum of a MALDI source and pulses of laser light (typically a nitrogen laser) are directed at the matrix. The matrix absorbs the laser energy and transfers its charge to the analyte molecules, as the matrix evaporates, analytes are liberated and ionized. The observed ion that contains a neutral molecule [M] is protonated to form a singly charged quasimolecule $[M+H]^+$ [117, 141].

In ESI, the sample is presented in a liquid form and thus can be easily associated with online liquid chromatography. The typical solvents are prepared with water, volatile organic compounds (e.g. methanol, ACN) and acetic acid which increases the conductivity. The solution containing the analytes flows into a capillary that is subject to a high voltage (2-3 kV) which forms the solution into a fine spray of highly charged droplets. The flow of droplets is then directed through a counter-current flow of heated gas, causing the solvent to evaporate and the charge concentration of the surface of the droplets to increase. It then reaches a critical unstable state, known as the Rayleigh limit; the droplets deform into smaller and lower charged particles in a process known as Rayleigh fission [143]. The Rayleigh

fission is repeated until individually charged analyte molecules remain. ESI generally produces a mixture of singly and multiply charged ions $[M+nH]^{n+}$ [117, 142].

1.2.3.2 Mass analysis

Four types of mass analysers are commonly used in proteomic research: quadrupole (Q), ion trap, time-of-flight (TOF), Fourier-transform ion cyclotron resonance (FTICR).

Quadrupole mass analysers separate ions based on the stability of their trajectories in the oscillating electric fields. The quadrupole consist of four parallel metal rods, each opposing rod pair is connected together electrically with a radio frequency (RF) potential applied. A direct current voltage is superimposed on the RF potential to make ions travel along the central axis of the rod. Only ions of a certain *m/z* will reach the detector for a given ratio of potentials while other ions with unstable trajectories will collide with the rods. A range of m/z values can be scanned by continuously varying the voltages [117].

Ion trap mass analysers trap charged molecules using electric or magnetic fields. The Quadrupole ion trap (QIT) is most often used and includes the 3D ion trap (Paul ion trap) and the linear ion trap. In the 3D ion trap, ions are trapped by electric fields produced by a ring-shaped electrode (RF potential) and two end-cap electrodes (dc potential). Ions enter the trap from one of the end-cap electrodes and oscillate at the frequencies that related to their *m/z* values. By changing the voltages applied to electrodes, ions of certain *m/z* become excited and are ejected from the opposite end cap [144, 145]. The linear ion trap is similar to the 3D ion trap except that the electromagnetic signals are designed to trap ions in a rectangular-shaped space. Ions

are confined radially by a set of quadrupole rods with RF potentials and axially by a static electrical potential on end electrodes. Linear ion trap MS provides increased ion storage capacity (10 times compared to the 3D ion trap) and faster scanning speeds [117, 145, 146].

Orbitrap is a new type of ion trap mass analyser invented by Makarov [147]. It provides high mass accuracy and high-resolution capabilities which has the potential to be useful for proteomic research [148, 149]. It consists of an outer barrel-like electrode and a coaxial inner spindle-like electrode. Ions are trapped and orbit around an inner spindle-like electrode, and oscillate harmonically along its axis. The frequency of these harmonic oscillations is independent of the energy and spatial spread of ions and is inversely proportional to the square root of the *m/z.* These oscillations are detected using image current detection and are transformed into mass spectra using Fourier transform similar to FITCR [116, 147, 149].

In a TOF mass analyser, ions that are accelerated in an electrical field, then travel through a field-free vacuum tube towards an ion detector. All ions with the same charge receive the same amount of kinetic energy in the source, while the velocity of the ion depends on their *m/z*. A reflectron with a constant electrostatic field can be used to reverse the path of ions towards the detector. Given the tube length and the measured times of flight, the mass-to-charge ratio of the ion can be calculated. A delayed extraction device can be used to equilibrate the ions which allows the initial velocity of ions to be standardised prior to the entrance of the TOF analyser [117].

FTICR mass analysers provide the greatest capability for mass resolution and mass measurement accuracy. It determines the *m/z* of ions based on the cyclotron frequency of the ions in a fixed magnetic field. FT mass spectrometers consist of a

cubic cell inside a strong magnetic field. Injected ions rotate around the magnetic field with a frequency according to their *m/z*. By varying the electric fields, changes in the ion frequency of rotation can be measured and converted to *m/z* by performing a Fourier transform [117, 150].

1.2.4 Protein identification using MS data

In bottom-up proteomic research, protein identification is made by matching the experimental peptide MS data to a virtual peptide mass database. A virtual database contains known protein sequences acquired from predicted gene models, open reading frames (ORFs) translated from genomic sequences or EST data. Theses sequences are then *in silico* digested using the same cleavage specificity of the protease employed in the experiment to produce theoretically determined peptide mass data.

According to the instrument setup, two types of MS data can be used to determine protein identifications: Peptide mass fingerprinting (PMF) data generated by MALDI-MS and peptide fragment fingerprinting (PFF) data generated by tandem MS, typically ESI-MS/MS.

1.2.4.1 Peptide mass fingerprinting based identification

Using peptide mass fingerprinting to determine protein identification was independently developed in 1993 by several groups [151-155]. This method is based on the use of a list of the molecular masses of digested peptides, served as a fingerprint that uniquely defines a particular protein. Experimental spectra of a candidate protein which contain m/z ratios of digested peptides are compared with theoretical spectra produced by a virtual database and a similarity score is given.

Protein (or proteins, in the situation of homologues or multiple proteins in the sample) in the virtual database with the top-ranking similarity score is considered the protein identification for the spectra. Scoring algorithms for PMF take into account many factors such as dissimilarities in the peptide masses caused by calibration errors, contaminant or missing peaks, chemical modifications and post-translational modifications (PTMs), etc.

Protein identification based on PMF data is fast and efficient; however, several limitations may lead to a poor or false identification. From the database point of view, if a protein sequence or an unknown sequence modification is not present in the virtual sequence database, the similarity score cannot be made or the result represents a false-positive. From the sample point of view, in the case of multiple proteins present in the sample, the increase in complexity of the spectra may result in poor or false-positive identifications; meanwhile, if the size of protein in the sample is too small or the concentration is too low, insufficient peptides MS data can be produced for confident protein identifications.

1.2.4.2 Tandem mass spectrometry based identification

Tandem mass spectrometry (MS/MS) uses two mass analysers in series. The first analyser separates the peptides according to their *m/z* values; selected peptides (precursors) are fragmented and the m/z ratios of the resulting fragments are measured by the second analyser. The most widely used fragment method is collision-induced dissociation (CID). It internally heats precursors by multiple collisions with neutral gas atoms. The C-N bond of the peptide backbone is fragmented into a series of b-fragment (charge remains at the N-terminus) and y-fragment (charge remains at the C-terminus) ions [156].

The principle of peptide fragment fingerprinting (PFF) based identification is very similar to the PMF approach. It correlates the experimental MS/MS spectrum with virtual MS/MS spectra generated from *in silico* digestion of proteins to peptides and fragmentation of these peptides. Protein identification can be made on several independent peptide identifications. The better the overall sequence coverage of the protein, the more confident the protein identification is. Scoring algorithms take into account the mass of the precursor peptide, chemical and post-translational modifications, etc.

PFF based identification has several advantages over PMF. It does not require all the peptides of a given protein to be confirmed to achieve a confident identification. It can work with complex peptide mixtures or to search homologous databases. It can also provide detailed information about peptide sequences and about possible post-translational modifications. However, the PFF approach also has limitations: firstly, same as PMF, a confident identification relies on the presence of protein sequence in the virtual sequence database, although small sequence variations like PTM can be partly compensated for by the identification of other peptides in the protein. Secondly, non-peptide contamination and poor fragmentation can influence the confidence of identification.

1.2.5 Proteomic data interpretation and integration

The output of protein identification based proteomic studies is usually a list of expressed proteins in the sample inferred by MS data. Whilst of considerable value in itself, subsequent characterization of the proteins identified is the next stage, from which important biological information about the sample can be gleaned. A further area in which proteome data is of great value is the integration of proteomic expression data with other genomic resources. The integration centralizes the current

knowledge about a particular organism on a genome expression level and proteomic data can also help confirm and refine existing genome annotations (see section 1.2.6.1).

1.2.5.1 Proteomic data interpretation

Interpretation of proteomic data uses combined evidence from two approaches: first, manual interpretation of information extracted from experimental data in the literature; secondly, information derived by automatically transferring existing knowledge about a homologous sequence to the targeting sequence. This latter approach is the basis of protein prediction programmes designed to predict protein sequence features or motifs based on a set of trained rules.

1.2.5.1.1 Manual interpretation

Manual interpretation process allows input from highly trained and knowledgeable curators. Curators are able to access and extract information from free text in literatures and assess all the information available. Several organism-specific and universal projects have been carried out for this purpose and publicly available online databases are designed to host the information gathered [157-161]. Among these efforts, UniProt is a universal database that commonly used for proteomic data interpretation.

UniProt [157] provides a universal curated protein database that stores and interconnects information from various sources including a non-redundant collection of protein sequences, metagenomic data and functional annotation. The component site UniProtKB/Swiss-Prot contains manual records with the information extracted from literature and curator evaluated computational analysis [162]. Annotated

information includes protein function, catalytic activity, subcellular location, disease, structure and post-translational modifications.

Although manual curation provides the most reliable information for proteomic data interpretation, it is very time consuming. To assist this, text mining methods have been recently developed to automatically extract information from literature [163-165]. However, the current techniques are still under development due to complicated syntactic patterns in natural language and restrictions in the processing non-textual material i.e. figures [165].

1.2.5.1.2 Automatic prediction

The main aim of automatic prediction is to transfer previously characterized protein information to uncharacterized homologous proteins. Homologous proteins are descended from a common ancestral protein that similar sequences or structures are often found and similar functions are often observed. Homologues can be further divided into orthologues and paralogues where orthologues are separated by a speciation event and paralogues are the product of gene duplication [166]. In automatic prediction, both sequence similarity and structure similarity based approaches are commonly used.

While there is no perfect protocol for automatic prediction, the most commonly used tool is Basic Local Alignment Search Tool (BLAST) [167]. A BLAST search compares the query sequence with protein sequences from various databases and reports a similarity score to exactly or partially matched sequences in the databases. However, it does not provide sufficient guidelines on whether the annotations can be safely transferred to the new sequence. Utilizing orthologues group mapping

databases like COG [168] and OrthoMCL[169] can improve confidence of the matching.

A global sequence or structural similarity cannot always be found for a novel protein. In a study of 120 complete genomes, an average of 22.4% of proteins in a genome are reported as singleton (gene families containing only a single member) [170]. In these cases, signature-based resources can be used to infer protein functions when one or more protein signatures can be identified.

Protein signatures are defined by either a regular expression method that shows patterns of conserved amino acid residues [171] or the Hidden Markov Model (HMM) method which provides a statistical profile based on probabilities of finding an amino acid at a given position in the sequence [172]. There are many publicly available signature databases of protein families and domains, including sequence-based PROSITE [171], Pfam [173], PRINTS [174] PANTHER [175] and structure-based SUPERFAMILY [176] and Gene3D [177].

Protein signatures can be used in combination to predict protein functions. For example, proteins with no significant sequence similarity but have similar functions might be expected to share some common features like post-translational modifications, protein-sorting signals and similar subcellular localizations. Universal or organism-specific software packages can be developed to predict certain protein features based on a set of trained rules [178-181].

1.2.5.1.3 Integration of annotations

Manual curation provides the most accurate information about a certain protein, but handling large-scale proteomic data is very labour and time consuming while automatic approaches trade accuracy for a larger coverage and higher speed. As

many annotation methods and databases exist for genomic and proteomic data, using any single method will result in biased or even no predictions. However, trying to use all of them at the same time will not only increase the workload but also lead to confusion in rationalizing the different results obtained. Integration of different sources of data into a single comprehensive source would greatly support proteomic data interpretation.

InterPro [182] is one such database that integrates all the protein signatures from multiple databases. Signatures from member database Gene3D, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY and TIGRFAMs that match the same set of proteins in the same region on the sequence, and that describe the same domain, family, repeat, active site, binding site or post-translational modification, are grouped into single InterPro entries by a curator.

Each InterPro entry contains high-quality manual annotation providing useful information on descriptive abstracts, taxonomy, structural links and cross-reference to external databases such as IntAct [183], MEROPS [184] and Enzyme Commission numbers [185], and Gene Ontology (GO) terms [186] where possible. Recent development has allowed links in InterPro to be made to distributed annotation system (DAS)-related tools such as the SPICE 3D structure viewer [187] and the Dasty Client [188]. DAS comprises a reference server that contains information for other client-servers to communicate and exchange biological annotations with [189].

Integrated information provided by InterPro is used to aid the manual annotation process by curators in UniProtKB/Swiss-Prot as well as being the basis of the automatic annotation system in UniProtKB/TrEMBL [157]. For individual proteomic studies, information provided by InterPro provides valuable addition to

organism-specific manual and automatic data interpretations. Collectively, they provide a preliminary view of biological features contained within the proteomic data.

1.2.5.2 Proteomic data repository and integration

While data interpretation can be carried out at the level of the individual research group, proteomic data integrated with other proteomic resources can benefit the global research community. Using a public repository of proteomic data not only preserves the effort and information made by an individual group, but also provides significant advantages that small scale studies cannot achieve.

Fundamentally, each proteomic study only samples a "snapshot" of the expressed proteins at a given time under a defined condition. By combining efforts from many research groups, a large volume of data with different time points and conditions can be acquired, which can highlight important trends by providing a dynamic view of the biological process. On the other hand, with access to large volumes of data generated from various proteomics platforms, limitations of different platforms can be assessed which will stimulate the development of improved methodology.

1.2.5.2.1 Data format for MS output

With the increasing complexity of evolving and diversifying methods and technologies used in proteomic research, the efficiency of data repository would benefit from a standardized data format. The minimum information about a proteomics experiment (MIAPE) is a project carried out by the Human Proteome Organization-Proteomics Standards Initiative (HUPO-PSI) [190]. This guideline aims to provide sufficient experimental details for publication and allow data exchange and comparison between different datasets.

A data format mzML was developed in 2008 under the auspices of the HUPO-PSI [191]. mzML replaced two competing data formats using vendor-neutral XML (Extensible Markup Language): mzData, developed by HUPO-PSI [192] and mzXML, developed at the Institute for Systems Biology [193]. The format is expected to become the single, unifying format for unprocessed proteomic MS data which is supported by nearly all the instrument and software vendors [191]. The other PSI standard format for peptide and protein identification, analysisXML, is still under development [194].

1.2.5.2.2 Public proteomic data repositories

While the new standardized data formats are being developed and adopted, currently there are four main public repositories actively storing proteomic data for the research communities: the Proteomics identifications database (PRIDE), the Global Proteome Machine databases (GPMDB), PeptideAtlas and Tranche [195].

PRIDE is a database of protein and peptide identifications with their corresponding literature publication [192]. Closely related to HUPO-PSI, PRIDE supports the PSI standard reporting guideline (MIAPE) and data formats (mzML and analysisXML). PRIDE provides public access to the details of experiment by experiment accession number, protein accession number, literature reference and sample parameters including species, tissue, sub-cellular location and disease state [192]. It allows comparison of data generated from different research groups and allows referees to examine and validate the data before publication [196].

GPMDB hosts proteomic data of a number of eukaryotic species and is likely to be the first to categorise the whole human proteome [195]. It is part of the GPM (The Global Proteome Machine Organization), which is an open-source system for

analyzing, storing and validating proteomic data generated from MS/MS spectra [197]. GPM developed peptide identification search engine pipelines including X!Tandem (matching spectra to sequence) [198], X!P3 (proteotypic peptide profiling) [199] and X!Hunter (matching spectra to annotated spectrum libraries) [200]. The GPMDB uses its own data format XIAPE (Xml Information About a Proteomics Experiment) which is a simplification and extension of MIAPE and a new compressed format Common 1.0 to compress and depress MS/MS data files.

PeptideAtlas is a project of Seattle Proteome Center (SPC) aims to achieve a full annotation of eukaryotic genomes through validation of expressed proteins [201]. PeptideAtlas integrates the Trans-Proteomic Pipeline (TPP) developed by SPC which is a collection of tools that uniformly analyse MS/MS data generated from different instruments, and assigned peptides using a variety of different database search programs [202]. All the sequences and spectra in the database are processed through the TPP to achieve a high quality database, along with false discovery rates at the whole atlas level [203].

Tranche software is an open source file sharing tool that enables collections of computers to easily share large amount of data sets [204]. Using the software, Tranche network can be created by anybody or any institution. The ProteomeCommons.org Tranche network is the first instance of a Tranche network in existence. It supports data sharing in proteomics as well as address the problem of data loss through computer hardware failure or changes in staff [195]. Any file type is allowed by the network, including glycomics, metabalomics and 2-D gel data, but MS/MS data is its main focus. Data uploaded to the network are replicated several times to protect against accidental loss, sharing and dissemination of data is secure and all datasets are citable in scientific journals [204].

1.2.6 Data integration and applications

Integration of proteomic data provides biological and technique advantages for proteomic research. Firstly, proteomic data can be used to verify and improve the prediction of genome annotations. Secondly, by comparing proteomic data with transcriptomic data, the importance of post-transcriptional and post-translational control of protein abundance in biological process can be examined.

1.2.6.1 Proteogenomics

Genome annotation normally comprises two components, identification of protein coding sequences (CDS) and the functional annotation of protein products (discussed in the previous sections). Identifying protein-coding genes is the primary process which provides the foundation to all the downstream analyses. While the conventional work flow for bottom-up protein identification based proteomics relies upon the accurate prediction of protein-coding genes within the genome, expression data acquired from proteomic research can also be used to feedback into the gene finding process to improve the quality of the flow from its source.

Despite the significant effort that has been made in genome annotation, the estimate of correct gene structure prediction is only 50% for the human genome [205]. The correction rate is estimated to be higher at two-thirds in eukaryotes with compact genomes such as *Arabidopsis thaliana* [206]. Expressed sequence tags (ESTs) and cDNA alignments studies carried out in Apicomplexan genome annotations revealed slightly better results. However, even in the most studied Apicomplexan genome of *P. falciparum*, approximately 24% of the genes in the current databases are still predicted incorrectly [207]. The correction rate falls in the case of *T. gondii*, where 41% of ToxoDB v4.3 gene models contained at least one inconsistency with full-length cDNAs [208].

By mapping peptides identified against predicted gene models and six-frame translation of the genomic DNA sequences (ORFs), proteomic data can be used to independently validate the prediction of protein-coding genes, confirm the translational expression of these genes and predict new genes. Initial studies tested this idea by querying MS/MS data against translated genomic sequence of bacteria *Haemophilus influenza*, *Mycobacterium tuberculosis*, plant *Arabidopsis thaliana* and human genomes on a small scale [209-212]. The approach was further tested on a small bacterium *Mycoplasma pneumonia* which validated the majority of the predicted genes (ORFs), as well as suggesting 16 new genes [213]. The success of the previous studies lead to the first use of proteomic data as a primary resource for *Mycoplasma mobile* genome annotation [214].

Proteogenomics approaches are increasingly used in various aspects of genome annotation including validating predicted gene models and detecting novel genes [213-216]; confirming the expression of hypothetical genes [217-220] and validating alternative splicing variants [220, 221]. The information gathered in proteogenomic researches can also provide a valuable training set to improve gene identification, predict peptide signals and improve the development of new genome annotation pipelines.

1.2.6.2 The interaction of proteome and transcriptome

Upon the completion of genome annotation, another important post genomic application is transcriptomics. The transcriptome is the complete set of transcripts in a cell at a specific developmental stage or physiological condition.

Two types of technologies are widely used to measure the expression profile of the transcriptome, hybridization based and sequence based. Hybridization based

approaches typically involve incubating fluorescently labelled cDNA with custom-made or commercial oligonucleotide microarrays and measuring the intensity of the fluorescent dye used [222, 223]. Sequence based approaches directly determine the cDNA sequence. Early techniques include the use of relatively low throughput cDNA or EST libraries [224, 225]. Tag-based methods such as serial analysis of gene expression (SAGE) [226], cap analysis of gene expression (CAGE) [227] and massively parallel signature sequencing (MPSS) [228] were developed to provide quantitative expression data in a high throughput manner.

Recent technical development has seen a novel RNA-Seq (RNA sequencing) approach that is based on high-throughput sequencing. In summary, the technique starts with the conversion of a population of RNA (total for small RNAs, such as microRNA or short interfering RNAs, and fractionated for larger RNA, such as poly(A) tail) to a library of cDNA fragments with adaptors attached to one or both ends. Each molecule in the library is then sequenced with a high-throughput sequencing platform to obtain short sequences from one end or both ends [229]. RNA-Seq can be done on a variety of high-throughput sequencing platforms including Illumina IG [230], Applied Biosystems SOLiD [231] and Roche 454 Life Science systems [232]. The resulting sequences can be aligned to a reference genome or reference transcripts (intron eliminated), or assembled *de novo* without the genomic sequence to produce a genome-scale transcription map [229].

One of the main purposes of transcriptomic studies is to test how particular conditions (such as, differentiation, transformation and environment) influence global gene expression. These changes at transcriptional level are then used to assume the same changes happen at the protein expression level under the particular condition and therefore enable the putative functions of proteins to be assigned. This

concept is under the commonly accepted 'guilt-by-association' hypothesis that the function of hypothetical genes may be similar to those of annotated genes that share the same expression profile [233]. However, with the availability of proteomic data, the real changes at the protein expression level can be directly and accurately measured first hand.

In fact, using transcriptomic data to infer protein functions is not only indirect, but potentially very unreliable. The 'guilty-by-association' hypothesis has been challenged by *in silico* approaches in which expression clusters found in microarray data do not in general agree with functional annotation classes [234]. This finding has been further verified and supported by several large scale studies comparing proteomic and transcriptomic expression profiles. Data from combined proteomic and transcriptomic analysis revealed discrepancies in the correlation between mRNA expression levels and protein abundance in plant seeds [235], mouse embryonic stem cells [236], yeast [237] and *Plasmodium* [238, 239].

Although the discrepancies between transcriptome and proteome weaken the application of protein function assignments using transcriptomic data, they enlighten new applications that allow a better understanding of basic biological processes. Despite the possibility that the discrepancies are caused by technical limitations whereby the same level of analytic resolution is not reached between transcriptome and proteome, the other possible explanations have more biological meaning and are worth further investigation. One possible explanation is the selective degradation of proteins. A study found that more than 80% of the cellular proteins are degraded through a proteasome-dependent pathway [240], while another study found the half-lives of 576 human A549 adenocarcinoma cells range from many tens of hours to just 6 min [241], all adding to the complexity of the protein expression profile. On

the other hand, the role of post-transcriptional regulations such as mRNA decay and translational repression may also contribute to the discrepancies between mRNA expression level and protein abundance [242-244]. The integration of transcriptomic and proteomic expression data will provide a dynamic view of gene expression at mRNA and protein level, which will in turn increase the understanding of various biological processes.

1.3 Aims

The knowledge of *T. gondii* biology has been steadily increasing in the last century. The last thirty years have seen a rapid progression in the understanding of gene expression of the parasite along with some major technical developments.

In the 1980s, using immunocytochemical techniques, the identification of specific antibodies led to the cloning of individual genes [106, 245], which allowed the identification of the expression and localization of two surface antigens. In 1998, the development of expression sequence tags (ESTs) [246] rapidly accelerated the rate of gene discovery by generating >7000 ESTs that identified more than 500 new *T. gondii* genes. Moving onto 2002, microarray experiments were developed to provide large scale gene expression profiling of tachyzoite to bradyzoite differentiation [247-249]. In the same year, a pioneering study on the *T. gondii* proteome has been carried out using reproducible 2-DE maps, which led to the identification of around 30 proteins [250]. In 2003, the first output of the *T. gondii* genome sequencing project was completed followed by the continuing efforts towards genome annotation [108, 112]. The focus of *T. gondii* research has gradually moved to the understanding of gene expression and gene functions on the genome scale.

Inspired by the fascinating features of *T. gondii* and technical advantage developed in recent years, this PhD study aims to investigate some important biological and technical issues of the proteomic research of *T. gondii*. This study is achieved through the following steps:

- Multiple proteomic platforms are used to identify *T. gondii* tachyzoite proteome.
- A collection of bioinformatics tools are used to characterize and categorize the expressed tachyzoite proteome.

- The method of integrating the proteomic data with other genomic sources is examined, and the possibilities and applications of proteomic data in the field of proteogenomics are discussed.

- A comparison of proteomes and transcriptomes of several Apicomplexan parasites has been carried out with data acquired from this study and other publicly available expression data, where the meanings of comparison results are discussed.

- Finally, a preliminary case study has been carried out to test the applications of a quantitative proteomic method DIGE in understanding protein expression changes of *T. gondii* tachyzoites grown in the presence or absence of glucose.

Together, this study aims to achieve a comprehensive coverage of the expressed *T. gondii* tachyzoite proteome and at the same time, examine and discuss the applications and potentials of this proteomic data in a broader integrative systems biology view.

Chapter 2

Identifying the *T. gondii* tachyzoite proteome

2.1 Introduction

The sequencing and annotation of the *T. gondii* genome has promoted the rapid development and application of proteomic analysis on the parasite. Studies have been carried out to investigate the mechanisms of host cell invasion, the structure and composition of the apical organelles, the organization of the cytoskeleton and the "entire" proteome of tachyzoites [84, 251-253].

Being the active, infectious stage of *Toxoplasma*, the tachyzoite has been the focus of proteomic studies and so far, no significant data has been published on the other life cycle stages. Type I strain RH tachyzoites have been used for most proteomic studies since they have almost no bradyzoite differentiation *in vitro* under standard culture conditions [254, 255].

The success of a proteomic project is based on the combination of several components, such as good sample isolation and preparation techniques, access to expensive specialist equipment for protein separation and mass spectrometry and finally, comprehensive data analysis capability. Among the pioneering studies on the *T. gondii* proteome, reproducible 2-DE maps were developed in 2002, in which over 1000 individual polypeptide spots could be resolved [250]. Of these, 71 protein spots were analysed by both MALDI-TOF and post-source decay mass spectrometry and due to the limited availability of genomic sequences at that time, this study resulted in around 30 protein identifications [250]. Another early attempt at the *T. gondii* tachyzoite proteome revealed up to 224 protein spots on a 2-DE gel and localized 13 parasite excretory antigens on the gel by Western or T-cell blot [256]. Since then, equipped with a better understanding of parasite biology and number of technical

advances, the research interest has focused on biologically interesting *T. gondii* subproteomes.

2.1.1 *T. gondii* subproteomes

Studies on *T. gondii* subproteomes have been benefited from a greater understanding of *Toxoplasma* biology, which has enabled improved sample preparation methods such as organelle purification based on density [84] and the collection of apical molecules via secretion stimulated by the elevation of Ca^{2+} [257]; the growing availability of sub-cellular markers and the increasingly accurate gene models provided by bioinformatics resources [112].

For example, one study has characterised the proteome of rhoptry organelles [84]. A French press was used to disrupt tachyzoites, followed by the fractionation of a mixture of rhoptries, dense granules, mitochondria and apicoplasts from cell lysates by a Percoll gradient. Sucrose flotation gradient was used to improve the enrichment of rhoptry proteins and the sample was analysed by 1-DE LC-MS/MS [84]. This study identified 38 novel rhoptry protein candidates and distinguished rhoptry proteins according to their sub-organellar location as either bulb (ROP) or neck (RON) [84].

Another subproteomic study prepared excretory/secretory antigens (mainly microneme and dense granule proteins) freely released from *T. gondii* tachyzoite by incubating filter-purified parasites in 1% (v/v) ethanol. The protein contents were characterized by 2-DE followed by N-terminal sequencing or MALDI-MS, and MudPIT [253]. A similar proteomic study used Ca^{2+} ionophore A23187 to stimulate calcium-mediated excretion of proteins from *T. gondii* tachyzoites. The protein

contents were characterized by 2-DE followed by N-terminal sequencing or MALDI-MS [257].

A proteomic analysis of enriched cytoskeletal components partially purified the conoid/apical complex. Parasites were extracted in lysis buffer which left a parasite ghost consisting of the apical complex, attached subpellicular MT and the cortical filament network. The conoid and remnants of the apical complex cytoskeleton were further purified by sonication and differential centrifugation [251]. This study has identified ~200 proteins which represents 70% of the cytoskeletal protein components, and characterized seven novel proteins among which the targeting sequence for recruitment into the cytoskeleton during invasion was determined for five proteins [251].

Together with these studies, many proteomic efforts have been made to understand the key proteins and post-translational modification events involved in host-cell invasion, and intracellular survival processes [258-260]. However, the global characterization of the *T. gondii* tachyzoite proteome has not seen many additions since the early effort in 2002.

2.1.2 Whole proteome profiling of Apicomplexan parasites

While no major global proteomic study has been published on *T. gondii* since 2002, several large-scale studies have investigated the global protein expression profile of other Apicomplexan parasites.

Two whole cell proteome studies have characterized the proteomes of four different life cycle stages of *Plasmodium falciparum* using MudPIT and 1-DE LC-MS/MS in 2002 [261, 262]. Comprehensive proteomic approaches have also been used to analyse the proteome of *Plasmodium berghei, Plasmodium yoelii* and *Plasmodium*

gallinaceum [238, 239, 263]. Together, these studies provided detailed proteomic coverage of several life cycle stages of the *Plasmodium* species making it one of the most studied microbial proteomes.

Whole proteome studies have also been carried out on *Cryptosporidium parvum* sporozoites. One study investigated the non-excysted and excysted forms of sporozoites using LC-MS/MS coupled with iTRAQ isobaric labelling and MudPIT. Together, 303 *C. parvum* proteins were identified among which expression of 26 proteins have been shown to increase significantly during excystation [264]. In addition to this, a comprehensive study to characterise the proteome of excysted sporozoites of *C. parvum* was performed using 2-DE MALDI-TOF or LC-MS/MS, 1-DE LC-MS/MS and MudPIT. In total, 1237 non-redundant proteins were identified in this life cycle stage, which represents approximately 30% of the entire predicted proteome ([265] and Appendix X) and significantly expands upon current knowledge.

2.1.3 Aims

In this chapter, a multi-platform global proteome analysis of *T. gondii* tachyzoites was performed. The strategy was to harness technological advances developed in the field of proteomic research and the recent availability of extensive genomic sequences and genome annotation to enable the rapid identification of proteins at the whole-cell scale. The result had the potential to improve significantly the knowledge of protein expression in this important parasite, complement current transcriptional data and provide a useful dataset from which improvements in gene annotation can be made.

2.2 Materials and methods

2.2.1 *T. gondii* tachyzoite culture

Tachyzoites of *T. gondii* strain RH were maintained twice a week in vero cells (African Green Monkey kidney fibroblast-like cell) (purchased from ECACC, cat. 84113001).

Uninfected vero cells were grown routinely at 37℃ in a 5% CO_2 humidified incubator (BINDER®) in 25 cm^2 bottom (T25) vented cell culture flasks (BD Falcon™) using filter sterilised IMDM medium (Cambrex) supplemented with 5% FCS and 1% Pen/Strep (SIGMA-ALDRICH, 10,000 U/ml). Sub-confluent cultures (70-80%) were detached from the cell culture flask surface by washing the cells twice with 5 ml HEPES buffer and incubating the cells in 5 ml Trypsin-EDTA (Sigma) for 5 min at 37℃. Cells were flushed down and transferred to a centrifuge tube (BD) and centrifuged at 1500 g for 5 min. The cell pellet was resuspended with 5 ml IMDM medium and cells were reseeded at 1×10^5 in 5 ml of IMDM in a T25 flask.

Vero cells were infected with *T. gondii* tachyzoites after 24 hours incubation at a parasite-to-cell ratio of 4:1. Both vero cells and tachyzoites were counted using a neubauer haemacytometer (Assistent). Tachyzoites were incubated for 3 or 4 days prior to sample collection.

T. gondii tachyzoites were scraped off from the cell culture flask using a sterile cell scraper (BD Falcon™) and were separated from the cells by filtration through 3 μm pore-size Nuclepore® polycarbonate filters (Whatman, UK). Tachyzoites were then washed twice with PBS (pH 7.4) with centrifugation at 1500×g for 20 min at 4℃. The pellet was resuspended with 1 ml PBS and transferred to a 1.5ml eppendorf tube;

a final centrifuge was then performed at 16,000×g for 5 min at 4℃. The supernatant was discarded and the pellet was stored at -20℃.

2.2.2 Sample quantification

To estimate protein content in the sample, protein quantifications of tachyzoites were performed as follows. The frozen pellet was first solubilised using 100mM Tris/HCl pH 8.5 and the soluble fraction was quantified with the BCA assay (Bio-Rad Protein Assay). The insoluble fraction was then further solubilised using 2% SDS supplemented with 100mM Dithiothreitol (DTT) and quantified with the 2-D Quant Kit (GE Healthcare).

2.2.2.1 Solubilisation

The *T. gondii* pellet containing approximately 1×10^8 tachyzoites was kept on ice during the whole process. An aliquot of 500 µl of 100mM Tris/HCl pH8.5 was added to the pellet which was incubated on ice for one hour, with vigorous vortexing every 10 min during the incubation. Three cycles of freeze-thaw were performed after incubation. Each cycle consisted of 2 min vigorous vortexing, followed by fast freeze with liquid nitrogen and quick thaw to room temperature. The sample was centrifuged at 16,000×g for 30 min at 4℃. The supernatant was transferred into a new tube and quantified with BCA assay. A further 200 µl of 2% SDS and 100mM DTT buffer was added into the pellet. Three cycles of 5 min heating at 90℃ and 2 min vigorous vortexing were carried out to assist solubilisation. Solubilised protein was then quantified with 2-D Quant Kit.

2.2.2.2 BCA assay

Tachyzoite protein quantifications were performed using the Bio-Rad Protein Assay. Coomassie Plus reagent was diluted 1:4 with water. BSA standards were made by

serial dilutions from 1mg/ml stock. Samples were diluted into 1:10, 1:20, 1:40 and 1:80. Each blank, sample and standard was prepared in triplicate and pipetted into the appropriate well in a 96-well plate. An aliquot of 200 µl of prepared Coomassie reagent was added into every well. The plate was incubated at room temperature between 5 min to 1 hour, and the absorbance was read at 560nm in a microplate photometer (Multiskan Ascent, Thermo). A standard curve of blank-corrected standard absorbance versus protein concentration was plotted and sample protein quantifications were calculated according to their absorbance.

2.2.2.3 2-D Quant Kit

Proteins solubilised with 2% SDS and 100mM DTT were quantified using the 2-D Quant Kit. Working reagent was prepared by mixing 100 parts of colour reagent A with 1 part colour reagent B. BSA standards were set up by adding different volumes of 2 mg/ml BSA standard solution. Sample was added undiluted and diluted 1:10 and prepared in duplicate. Precipitant (500 µl) was added into each tube, vortexed briefly and incubated for 2-3 min at room temperature, then 500 µl of co-precipitant was added into each tube and mixed briefly. Tubes were centrifuged at 16,000×g for 5 min. The supernatant was discarded and 100 µl of copper solution and 400 µl of distilled water were added to each pellet which was then vortexed briefly. Working colour reagent (1 ml) was added into each tube and incubated at room temperature for 15-20 min. Absorbance was read at 480nm with Ultrospec 2100 pro spectrophotometer. A standard curve of blank-corrected standard absorbance versus protein concentration was plotted and protein quantifications were calculated according to their absorbance.

2.2.3 1-D gel electrophoresis (1-DE)

Sample was separated on a 16 cm long and 1.0 mm thick 1-D SDS-PAGE gel using PROTEAN™ II Slab Cell kit (Bio-Rad). Gels were cast following the recipes listed in Table 2.1 for stacking gel and Table 2.2 for separating gel.

Table 2.1 5% Stacking Gel

Reagent	Volume	Comment
dH_2O	13.6 ml	
30% (w/v) Acrylamide	3.4 ml	
0.5 M Tris/HCl (pH 6.8)	5 ml	
10% (w/v) SDS	200 µl	
10% (w/v) APS	200 µl	Add fresh
TEMED	20 µl	Add fresh

Table 2.2 12% Separating Gel

Reagent	Volume	Comment
dH_2O	33.5 ml	
30% (w/v) Acrylamide	40 ml	
1.5 M Tris/HCl (pH 8.8)	25 ml	
10% (w/v) SDS	1 ml	
10% (w/v) APS	500 µl	Add fresh
TEMED	100 µl	Add fresh

Electrode (Running) buffer was prepared as a 10 × stock solution according to Table 2.3 and diluted to 1 × working solution prior to use.

Table 2.3 10 × SDS-PAGE Electrode (Running) Buffer

Reagent	Volume	Comment
Tris base	30.3 g	
Glycine	144 g	
SDS	10 g	
dH_2O	to 1 litre	

40 µl 2× SDS-PAGE loading buffer (see Table 2.4) was added to the pellet which contains 1×10^8 *T. gondii* tachyzoites. Three cycles of 5 min heating at 90 ℃ and 2 min vigorous vortexing were carried out to assist solubilisation followed by a final centrifugation at 16,000 ×g for 3 min.

Table 2.4 2×SDS-PAGE Loading Buffer

Reagent	Concentration	Comment
0.5 M Tris-HCl (pH 6.8)	100 mM	
SDS	4 % (w/v)	
Bromophenol Blue	0.2% (w/v)	
Glycerol	10 % (v/v)	
DTT	200 mM	Add fresh
dH_2O		

The supernatant was loaded into a single sample well and separation was performed by electrophoresis at constant current of 16 mA for stacking and 24mA for separating gels. Run time was between 6-7 hours; after running, the gel was removed from the plate and fixed for 2 hours in 40 % (v/v) ethanol, 10 % (v/v) acetic acid. The gel was then washed twice with dH_2O and stained in colloidal Coomassie blue (1 part methanol, 4 parts colloidal stock (50g Ammonium sulphate, 500 ml dH_2O, 6 ml phosphoric acid and 10 ml 5% (v/v) Coomassie stock)) for 1-7 days.

2.2.4 2-D gel electrophoresis (2-DE)

2.2.4.1 Sample preparation

Two strategies have been used for sample preparation which resulted in similar preparation results determined by the sample quantification assay and the visualization of the stained gels.

a) 120 µl of lysis buffer A (see Table 2.5) was added to the frozen pellet which contains 1×10^8 *T. gondii* tachyzoites. Three cycles of freeze and thaw were performed; each cycle comprised 2 min of vigorous vortexing, followed by fast freeze with liquid nitrogen and quick thaw to room temperature. The sample was centrifuged at 16,000×g for 30 min at 4℃.

Table 2.5 Lysis buffer A

Reagent	Volume	Comment
Urea	8M	
CHAPS	4% (w/v)	
Tris-base	40mM	
DTT	60mM	Add fresh
IPG buffer	0.5% (v/v)	Add fresh
dH_2O		

b) 120 µl of lysis buffer B (see Table 2.6) was added to a 1×10^8 *T. gondii* tachyzoite pellet. The sample was incubated at room temperature for 2-4 hours with a vigorous vortex every half an hour. The sample was then centrifuged at 16,000 ×g for 5 min.

Table 2.6 Lysis buffer B

Reagent	Volume	Comment
Urea	7M	
ThioUrea	2M	
ASB-14	2% (w/v)	
CHAPS	4% (w/v)	
Tris-base	20mM	
Protease Inhibitor Cocktail Tablets (Roche)	1×	
DTT	60mM	Add fresh
IPG buffer	0.5% (v/v)	Add fresh
dH_2O		

2.2.4.2 IPG strip rehydration

From either 1a or 1b, the supernatant was transferred to a new tube and rehydration buffer (supplemented with 40mM DTT and 0.5% IPG buffer) was added to a final volume of 450 µl. The solution was then loaded to one reservoir slot of the immobiline DryStrip reswelling tray (GE healthcare). A 24 cm IPG strip (pH 4-7 or pH 3-11 NL) was placed in the reservoir slot with the gel side down. DryStrip cover fluid (3 ml) was then overlaid onto the strip and the tray covered with the lid. The tray was left at room temperature for a minimum of 10 hours for a complete rehydration.

2.2.4.3 First-dimension Isoelectric Focusing (IEF)

The rehydrated strip was placed on an Ettan™ IPGphor II™ (GE healthcare) following the manufacturer's handbook. IEF was run at 20°C, with 75 μA per strip with the following steps (see Table 2.7).

Table 2.7 IEF Protocol

Step	Voltage Mode	Voltage	Duration (Hour)
1	Step	500 V	2
2	Gradient	1000 V	8
3	Gradient	10,000 V	3
4	Step	10,000 V	4.25

2.2.4.4 Second-dimension SDS-PAGE

Second-dimension SDS-PAGE was performed on a precast DALT Gel 12.5(26 × 20 cm) (GE Healthcare) using the Ettan DALT*six* electrophoresis System (GE Healthcare).

The IPG strip was removed from the Ettan™ IPGphor II™ and two steps of equilibration were carried out. Firstly, 100 μg of DTT was added into 10 ml of equilibration buffer (2% (w/v) SDS, 50 mM Tris-HCl pH 8.8, 6M urea, 30% (v/v) glycerol, and 0.002% (w/v) bromophenol blue). The IPG strip was washed in equilibration buffer I on a shaker for 15 min. Equilibration buffer II was made up by adding 250 μg of iodoacetamide (IAA) into 10 ml of equilibration buffer. The IPG strip was washed in equilibration buffer II on a shaker for 10-15 min.

The equilibrated IPG strip was then assembled onto a DALT precast gel following manufacturer's instruction. The Ettan DALT*six* System was setup according to the product handbook using the electrophoresis buffer kit supplied. The gel was run at 20 ℃ with an initial wattage of 3 W for 0.5 hour and 17 W per gel thereafter.

After electrophoresis, the gel was removed from the plate and fixed for 2 hours in 40 % (v/v) ethanol, 10 % (v/v) acetic acid. The gel was washed twice in dH_2O and stained in colloidal Coomassie blue (1 part methanol, 4 parts colloidal stock (50g Ammonium sulphate, 500 ml dH_2O, 6 ml phosphoric acid and 10 ml 5% Coomassie stock)) for 1-7 days.

2.2.5 Manual tryptic digestion

Slices from 1-DE gels or plugs from 2-DE gels were cut out and placed in separate eppendorf tubes. An aliquot of 15 µl of 50 mM ammonium bicarbonate (Ambic)/ 50% (v/v) acetonitrile was added into each tube and incubated for 10 min at 37 ℃ to destain the plugs/slices. The destain step was repeated 2-3 times until the gel plugs were fully destained and the destain buffer was discarded at each step.

For 1-DE gel slices, two additional reduction steps were carried out before moving onto the next step. DTT (50 µl of 10 mM in 100 mM Ambic) was added to each tube and incubated for 30 min at 37 ℃. The DTT solution was discarded and 50 µl of IAA (55 mM in 100 mM Ambic) was added to each tube, followed by 1 hour incubation at 37 ℃ in the dark. IAA was discarded after incubation.

For both 1-DE slices and 2-DE gel plugs, 15 µl of 100% acetonitrile was added into each tube, followed by 15 min incubation at 37 ℃. This step was repeated until the gel plugs turned completely white indicative of dehydration. The solvent was removed and tubes were incubated at 37 ℃ for 10 min to evaporate the remaining solvent.

25 µg of sequencing grade trypsin (Roche) was first diluted with 250 µl of 50 mM acetic acid to make a stock solution. Stock solution was further diluted into working reagent at a 1:10 ratio with 25 mM Ambic. An aliquot of 10-15 µl of trypsin

(according to the size of the gel plugs) was added to the evaporated plugs and the tubes were incubated at 37 ℃ for 1 hour. An additional 10 µl of 25 mM Ambic was added into each tube, followed by overnight incubation at 37 ℃. The digested solution was run on the LTQ immediately afterwards or stored at -20 ℃.

2.2.6 LTQ (LC-MS/MS)

The LTQ work for the 1-D SDS PAGE gel slices and part of the 2-D SDS PAGE gel analysis was performed by Dr. S.J. Sanderson. Briefly, the LC-MS/MS platform was setup using a LTQ ion-trap mass spectrometer (Thermo-Electron) coupled on-line to a Dionex Ultimate 3000 (Dionex) HPLC system equipped with a nano pepMap100 C18 reversed phase column (75 µm; 3 µm, 100 Angstroms). The column was equilibrated in 98.9% (v/v) water/ 2% (v/v) acetonitrile/ 0.1% (v/v) formic acid (FA) at a flow rate of 300 nl/min. Sample injections of 15 µl of tryptic peptides were loaded onto a C18 TRAP, desalted and washed for 3min at a flow rate of 25 µl/ min prior to being loaded onto a nano pepMap100 C18 column at 300 nl/ min. The peptides were eluted at a flow rate of 300 nl/ min with a linear gradient of 0-50% (v/v) acetonitrile/ 0.1% (v/v) FA over 30 min, followed by 80% (v/v) acetonitrile/ 0.1% (v/v) FA for 5 min. The column was then equilibrated in 98.9% water/ 2% acetonitrile/ 0.1% (v/v) formic acid for 5 min (total run time per sample was 50 min).

Ionised peptides were analysed in the mass spectrometer (0-106 m/z, global and Msx) using the "triple play" mode, consisting initially of a survey (MS) spectrum from which the three most abundant ions were determined (threshold = 200-500 TIC). Collision energy was set at 35% for 30 min. The charge state of each ion was then assigned from the C13 isotope envelope "zoom scan" and finally subjected to a third MS/MS scan. The LTQ was tuned using a 500 fmol/µl solution of glufibrinopeptide (m/z 785.8, $[M+2H]^{2+}$) and calibrated according to the manufacturer's instructions.

The resulting MS/MS spectra (dta files) were merged into an mgf file which was submitted to Mascot searching.

2.2.7 Mascot searching of MS data acquired

Mascot searching was carried out on a local Mascot server. A local *Toxoplasma* database was selected which comprised of the following components retrieved from ToxoDB: ORF>50 aa from clustered EST; ORF>50 aa from whole genome shotgun (10×); Twinscan predictions; TigrScan predictions; GlimmerHMM predictions; Organellar Sequences and Annotated Proteins, release 4 (release 4 gene models). MS/MS Ion Search was used to search the data output from the LTQ. Database search parameters included: fixed carbamidomethyl modification of cysteine residues; variable oxidation of methionine; a peptide tolerance of ± 1.5 Da; MS/MS tolerance ± 0.8 Da; +1, +2, +3 peptide charge state; and a single missed trypsin cleavage. Instrument was set as ESI-TRAP.

2.2.7.1 Manual Validation of Mascot Results

For 1-DE and 2-DE results, additional manual validation was carried out on the proteins identified by Mascot that were based on a single peptide and/ or proteins which returned a Mascot score < 60. The protein identification was accepted if a) a matching peptide possessed an individual ion score above the significant threshold for identity or extensive homology (typically >44), or b) upon manual inspection of individual peptide MS/MS spectra at least 60% percent of the candidate y-ions were at a minimum signal to noise ratio of 10%. Spectra which failed to pass either rule were regarded as false positive identifications, which can result from an accumulation of several peptides with low ion scores.

2.2.8 Multidimensional protein identification technology (MudPIT)

Sample preparation, mass spectrometric analysis and MS data searching were performed by Dr. Judith H. Prieto from John R.Yates's lab, Scripps Research Institute, La Jolla. The procedures are as follows:

2.2.8.1 Sample preparation for MudPIT

A pellet of 1×10^9 tachyzoites resuspended to approximately 800 μg/ml in 500 μl 100 mM Tris/HCl buffer pH 8.5 was lysed by three cycles of freeze/ thaw and the Tris-soluble and insoluble protein fractions separated at 16,000×g for 30 min. **Digestion of soluble fractions:** MS compatible detergent Invitrosol was added to 1% (v/v), the solution heated to 60°C for 5 min, vortexed for 2 min, denatured with 2 M urea, reduced with 5 mM TCEP, carboxyamidomethylated with 10 mM iodoacetamide, followed by addition of 1 mM $CaCl_2$ and trypsin at a ratio of 1:100 (enzyme: protein) and incubated at 37°C overnight. **Digestion of insoluble fractions:** 10% (v/v) Invitrosol was added to the pellet which was heated to 60°C for 5 min, vortexed for 2 min and sonicated for 1 h. The sample was diluted to 1% (v/v) Invitrosol with 8 M urea/ 100 mM Tris/ HCl pH 8.5, reduced and carboxyamidomethylated as before, and digested with endoproteinase Lys-C for 6 h. The solution was diluted to 4 M urea with 100 mM Tris/ HCl pH 8.5 and digested with trypsin as described above.

2.2.8.2 Mass spectrometric analysis by MudPIT

Five soluble replicates and four insoluble samples were each subjected to MudPIT analysis with modifications to the method of Link *et al.*[266], using a quaternary Agilent 1100 series HPLC coupled to a Finnigan LTQ-ion trap mass spectrometer (Thermo, San Jose, CA) with a nano-LC electrospray ionization source. Peptide

mixtures were resolved by strong cation exchange LC upstream of RP-LC. Each sample (100 µg of protein) was loaded onto separate microcolumns and resolved by fully automated 12 step chromatography.

2.2.8.3 SEQUEST searching of MS data acquired

A local *Toxoplasma* database was used (as in section 2.2.7). To identify contaminant host proteins the parasite database was supplemented with a contaminant database (the complete prokaryote and mammalian databases from NCBI). To estimate the amount of false positives a reverse database was added. Poor quality spectra were removed from the data set using an automated spectral quality assessment algorithm. Tandem mass spectra remaining after filtering were searched with the SEQUEST algorithm version 27. All searches were in parallel and were performed on a Beowulf computer cluster consisting of 100 of 1.2 GHz Athlon CPUs. No enzyme specificity was considered for any search. SEQUEST results were assembled and filtered using the DTASelect (version 2.0) program which uses a quadratic discriminate analysis to dynamically set XCorr and DeltaCN thresholds for the entire data set to achieve a user-specified false positive rate (<5% peptides false positive in this analysis). The false positive rates are estimated by the program from the number and quality of spectral matches to the decoy database.

2.3 Results

2.3.1 *T. gondii* tachyzoite proteomic analysis by 1-DE

A large format 1-DE gel was performed to resolve solubilised proteins from 1.1 × 10^8 tachyzoites (220 µg of protein) (see Figure 2.1). In total, 129 gel slices were manually excised from the entire length of the resolving gel; each slice was tryptically digested, submitted to LC-MS/MS and searched against the local *Toxoplasma* database using Mascot (as in section 2.2.7).

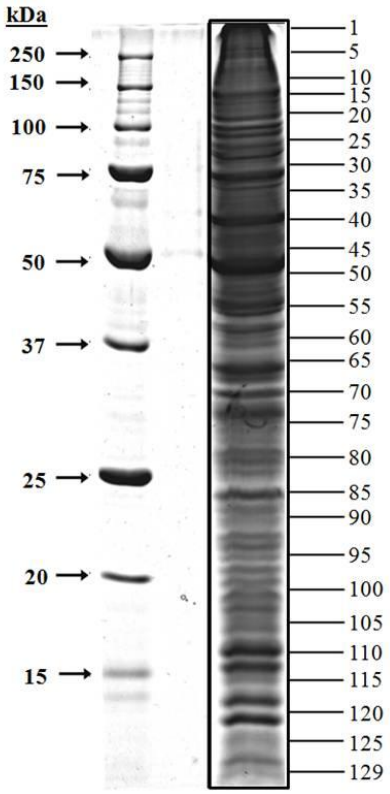
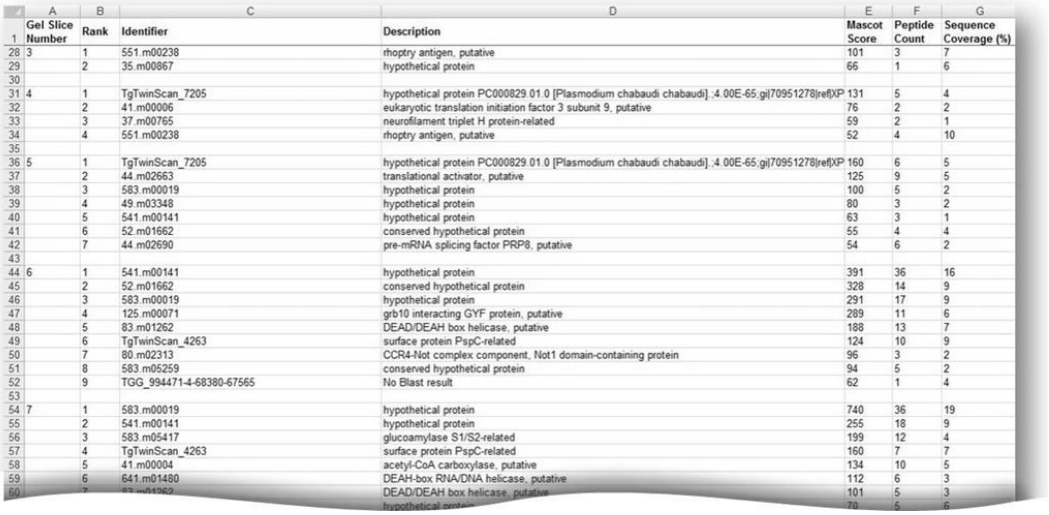


Figure 2.1 Tachyzoite proteins resolved by 1-DE. Proteins from 1.1×10^8 *T. gondii* tachyzoites were resolved on a 16 cm 12% (w/v) SDS-PAGE gel and stained with colloidal Coomassie blue. The gel was cut into 129 contiguous gel slices. Each slice was tryptically digested and analysed by LC-MS/MS. The masses of protein standards and the positions of gel slices are shown.

An average of approximately 20 proteins were identified from each gel slice (ranging from 2 proteins per slice from a region at the top of the gel to 51 proteins per slice in the middle of the gel) and the overall number of individual identification is 2778 (see

Figure 2.2 for a sample of the protein identification table and Appendix I for the list of individual protein identifications).



	A	B	C	D	E	F	G
1	Gel Slice Number	Rank	Identifier	Description	Mascot Score	Peptide Count	Sequence Coverage (%)
28	3	1	551.m00238	rhoptry antigen, putative	101	3	7
29		2	35.m00867	hypothetical protein	66	1	6
30							
31	4	1	TgTwinScan_7205	hypothetical protein PC000829.01.0 [Plasmodium chabaudi chabaudi] ;4.00E-65;gi|70951278|ref|XP	131	5	4
32		2	41.m00006	eukaryotic translation initiation factor 3 subunit 9, putative	76	2	2
33		3	37.m00765	neurofilament triplet H protein-related	59	2	1
34		4	551.m00238	rhoptry antigen, putative	52	4	10
35							
36	5	1	TgTwinScan_7205	hypothetical protein PC000829.01.0 [Plasmodium chabaudi chabaudi] ;4.00E-65;gi|70951278|ref|XP	160	6	5
37		2	44.m02663	translational activator, putative	125	9	5
38		3	583.m00019	hypothetical protein	100	5	2
39		4	49.m03348	hypothetical protein	80	3	2
40		5	541.m00141	hypothetical protein	63	3	1
41		6	52.m01662	conserved hypothetical protein	55	4	4
42		7	44.m02690	pre-mRNA splicing factor PRP8, putative	54	6	2
43							
44	6	1	541.m00141	hypothetical protein	391	36	16
45		2	52.m01662	conserved hypothetical protein	328	14	9
46		3	583.m00019	hypothetical protein	291	17	9
47		4	125.m00071	grb10 interacting GYF protein, putative	289	11	6
48		5	83.m01262	DEAD/DEAH box helicase, putative	188	13	7
49		6	TgTwinScan_4263	surface protein PspC-related	124	10	9
50		7	80.m02313	CCR4-Not complex component, Not1 domain-containing protein	96	3	2
51		8	583.m05259	conserved hypothetical protein	94	5	2
52		9	TGG_994471-4-68380-67565	No Blast result	62	1	4
53							
54	7	1	583.m00019	hypothetical protein	740	36	19
55		2	541.m00141	hypothetical protein	255	18	9
56		3	583.m05417	glucoamylase S1/S2-related	199	12	4
57		4	TgTwinScan_4263	surface protein PspC-related	160	7	7
58		5	41.m00004	acetyl-CoA carboxylase, putative	134	10	5
59		6	641.m01480	DEAH-box RNA/DNA helicase, putative	112	6	3
60		7	[illegible]	DEAD/DEAH box helicase, putative	101	5	3
				hypothetical protein	70	5	6

Figure 2.2 A sample of protein identification table of Mascot results. Listed in the columns (from left to right) are: the gel slice number, ranking of each protein hit returned from the Mascot search for that gel slice, corresponding gene annotations and descriptions, Mascot scores, number of matching peptides to each protein and sequence coverage.

In many instances the same protein was identified in multiple gel slices which could be due to proteolytic processing events, post-translational modifications, isoenzymes or simply that a given protein band spanned more than one gel slice. When redundancy between proteins with the same identification is removed, assuming the variants are the products of a single gene, the expression of 923 individual genes was identified (comprising 857 release 4 genes and 66 alternative gene models and ORFs; discussed in section 2.4.3).

In addition to the first 1-DE gel, another 1-DE experiment was performed with prior sample fractionation of 9.85×10^7 tachyzoites using 100mM Tris/HCl pH 8.5 with the aim of increasing sample resolution and thereby protein identification. Both Tris-soluble (120 µg of protein) and Tris-insoluble (130 µg of protein) fractions were

resolved by 1-DE (see Figure 2.3). The most divergent region (~50-100 kDa) between soluble and insoluble fractions was determined by eye and 25 gel slices were excised correspondingly, followed by tryptic digestion, LC-MS/MS analysis and Mascot searching. When the redundancy was removed, expression of 351 individual genes (335 release 4 genes and 16 alternative gene models and ORFs) was identified from the 50 gel slices on the second gel (see Appendix II for tables listing individual protein identifications).

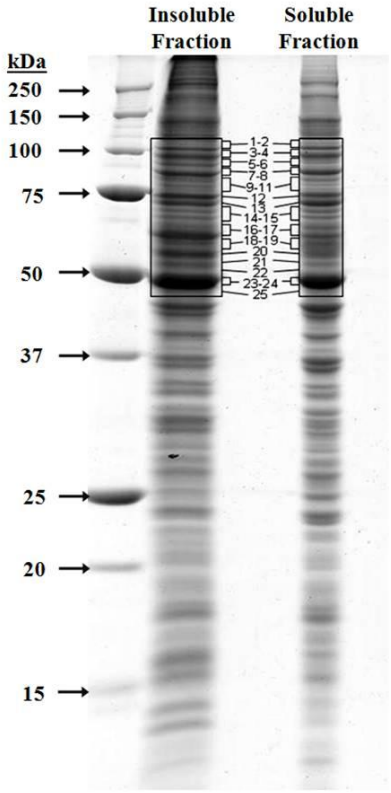


Figure 2.3 Tris-fractionated tachyzoite proteins resolved by 1-DE. Proteins from 9.85×10^7 *T. gondii* tachyzoites previously fractionated into Tris-soluble (120 µg of protein) and Tris-insoluble (130 µg of protein) fractions were resolved on a 16 cm 12% (w/v) SDS-PAGE gel and stained with colloidal Coomassie blue. Contiguous gel slices (25) spanning ~50-100 kDa were excised from both lanes. Each slice was tryptically digested and analysed by LC-MS/MS. The masses of protein standards and the positions of gel slices are shown.

When redundancy between proteins with the same identification from both 1-DE gels was removed, 1012 individual gene products (939 release 4 genes and 73 alternative gene models and ORFs) were identified (see Appendix III for non-redundant list of protein identifications).

2.3.2 *T. gondii* tachyzoite proteomic analysis by 2-DE

Both broad, non-linear (pH 3-10NL) and narrow, linear (pH 4-7) range 2-DE gels were used to resolve tachyzoite proteins (see Figure 2.4). In total, 1217 protein spots were excised from 2-DE gels (783 spots from the pH 3-10NL separation and 434 spots from the pH 4-7 separation). Each protein spot was tryptically digested, analysed by LC-MS/MS and searched for protein identity using Mascot.

Similar to the 1-DE results, many proteins from separate spots shared the same identity and some gel spots contained more than one protein (discussed in section 2.4.2). When redundancy between proteins with the same identification from both 2-DE gels was removed, 616 individual gene products (547 release 4 genes and 69 alternative gene models and ORFs) were identified (see Appendix IV for list of MS evidence obtained from 2-DE proteome maps of *T. gondii* tachyzoite proteins).

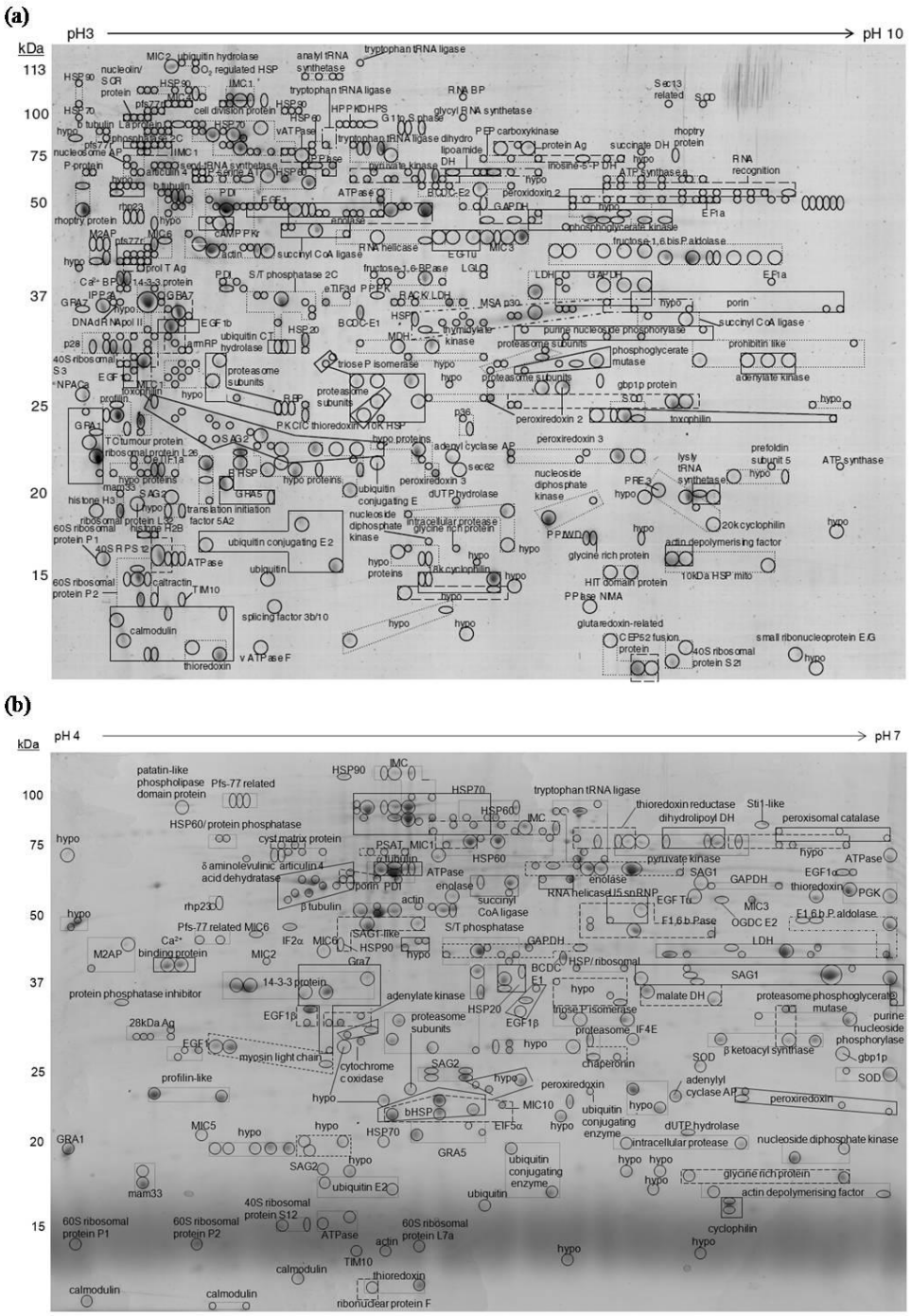


Figure 2.4 Tachyzoite proteins resolved by 2-DE. (a) Soluble proteins from 2.53×10^8 tachyzoites (516 μg of protein) resolved by IEF over a broad, non-linear pH 3-10 range. (b) Soluble proteins from 1×10^8 tachyzoites (200 μg of protein) resolved by IEF over a narrow, linear pH 4-7 range. Both pH3-10NL and pH 4-7 IEF strips were further separated by molecular mass on a 12.5% (w/v) acrylamide gel under denaturing conditions. Protein spots are visualized using colloidal Coomassie. Spots with the same protein identification are boxed. Gel annotation was assisted by Dr. S.J. Sanderson.

2.3.3 MudPIT analysis of *T. gondii* tachyzoites

Whole tachyzoite protein was fractionized into Tris-soluble and Tris-insoluble fractions, and each fraction processed for MudPIT analysis by Dr Judith Prieto in John R. Yates's lab, Scripps Research Institute, La Jolla as detailed in sections 2.2.8.1-3. This resulted in 1300 and 2328 protein identifications, respectively. A total number of 2409 non-redundant proteins were identified, which comprises 2121 release 4 genes and 288 alternative gene models and ORFs. Of the release 4 genes identified, 15.3% were uniquely identified in the Tris-soluble fraction and 48.0% were uniquely identified in the Tris-insoluble fraction (see Appendix V for protein identifications made by MudPIT). The higher number of proteins identified in the Tris-insoluble fractions is likely to reflect incomplete solubilisation in the Tris buffer. This is indicated by the transmembrane domain (TM domain) prediction results where 20.0% of proteins identified in the Tris-insoluble fractions possess at least one TM domain, compared to 8.4% in the Tris-soluble fractions.

2.3.4 Comparison of protein identifications from the three proteomic platforms

Prior to the commencement of this study in 2005, total protein expression evidence was limited to *T. gondii* sub-proteomes and the previous preliminary tachyzoite proteome study from the Wastling group [84, 250, 253]. In total just around 300 release 4 genes were known to be expressed, representing less than 4% of the predicted genome. In this study, when the results from the three proteomic platforms were combined, a total number of 2252 non-redundant release 4 gene identifications were obtained from the tachyzoite stage of the parasite. This represents the expression of approximately 29% of the total number of release 4 predicted genes in

the genome. Notably, 813 of the 2252 (36%) proteins identified are annotated as hypothetical or conserved hypothetical proteins on ToxoDB. With the expression evidence acquired in this study, the status of these proteins can now be changed to "confirmed" proteins. Figure 2.5 illustrates the degree of overlap between the protein identification results derived from each of the three proteomic platforms.

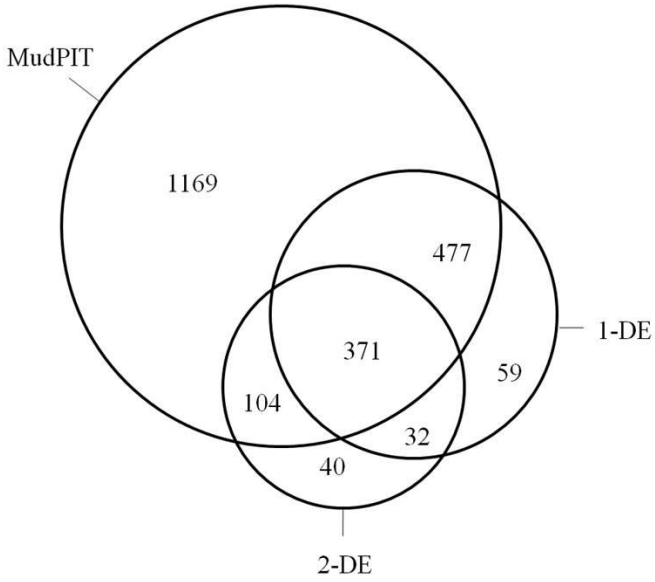


Figure 2.5 *T. gondii* tachyzoite expressed proteome: comparison of proteomic platforms. Venn diagram showing the number of unique and shared non-redundant release 4 gene identifications obtained from each of the three proteomic platforms.

The comparison shows that MudPIT produced the largest number of identifications; however, a number of proteins were uniquely identified from gel-based approaches (59 for 1-DE; 40 for 2-DE). A more detailed comparison of the different proteomic platforms used is discussed in section 2.4.2. In addition to the identified release 4 genes, 394 non-redundant alternative gene models and ORFs were identified from the entire dataset. The implication for genome annotation of these additional identifications out with the release 4 genes is further examined and discussed in chapter 4.

Searching against the contaminant database mostly returns trypsin and keratins with a wide range of identification scores for all three proteomic platforms used. This has been observed by other studies and [267] keratin contamination should be regularly monitored to keep the level to the minimum.

2.4 Discussion

In this study, a significant proportion of the *T. gondii* predicted proteome has been identified. This result benefited from the careful consideration of several important technical challenges such as proteomic platform selection; database design and the searching strategy for MS data acquisition, which are discussed in detail in the following sections.

2.4.1 Coverage of the predicted proteome

Using multiple platforms has proven to be an efficient and powerful approach in protein identification based proteomic research [262, 265, 268, 269]. In this study, three independent proteomic platforms were used to maximise the coverage of the expressed proteome. More than two thousand (2252) unique release 4 genes have been identified, which represents almost one third (29%) of the predicted proteome of all life cycle stages. This work significantly improved the knowledge of the *T. gondii* proteome, increasing the percentage of known expressed genes from ~4% to 29% of the total genome. The coverage was further expanded by the identification of 394 non-redundant alternative gene models and ORFs.

The coverage of the predicted proteome from one life cycle stage in this study is similar to or marginally better than other large scale Apicomplexan proteomic studies. For example, 32% for *C. parvum* excysted sporozoites [265], 20% for sporozoites and 16% for merozoites of *P. falciparum* [261], and 20% in a later *T. gondii* tachyzoite study [270]. Since the exact number of proteins in the predicted proteome of *T. gondii* tachyzoite is unknown, whether the 1/3 coverage of the whole predicted proteome represents a significant proportion of all the expressed proteins from the tachyzoite stage or rather a reflection of the detection limitation of the

proteomic techniques requires further investigation. For example, the detection of low abundance proteins is hampered in the presence of more abundant proteins during MS analysis and considering the continuous synthesis and degradation of proteins with varying turnover rates [271], it is difficult to include all the proteins in a single "snapshot" during sampling. The slightly higher coverage of the predicted proteome achieved by this study and the *C. parvum* study [265] is likely to have benefited from the three complementary proteomic platforms used which enhanced the detection of proteins with different properties. The difference and benefits of the three proteomic platforms are further discussed in section 2.4.2.

In fact, higher proteomic coverage can be expected when more life stages are studied. A proteomic study in *P. falciparum* has revealed that 49% of expressed sporozoite proteins are unique to this stage and an average of 25% of proteins is shared with any other stage. Taking all four stages studied into consideration, only 6% of proteins are common to all stages [261]. A transcriptome study of *T. gondii* found that 50% of expressed tachyzoite genes (day 7) are shared with the genes expressed in mature bradyzoites (day 17) [272]. If a similar overlapping percentage can be applied to the *T. gondii* proteome, a much higher coverage of the predicted proteome can be expected when multiple life cycle stages are studied.

2.4.2 Comparison of the three proteomic platforms used

Complementary proteomic platforms, the gel based techniques of 1-DE and 2-DE followed by LC-MS/MS and the gel free technique, MudPIT were used in this study. While each platform contributed to the overall coverage of the proteome by identifying unique proteins shown in Figure 2.5, they also have their own advantages and limitations in proteomic studies.

The gel based techniques used in this study generally detect more peptides per protein identification and hence deliver higher confidence scores than protein identification obtained from MudPIT. This can be explained by the sample separation in protein space prior to MS analysis, which results in a less complex peptide mixture to be analysed by MS and database searching.

Among the gel based techniques, 1-DE has a wider application in protein separation compared to 2-DE, it has a higher loading capacity and the presence of SDS enables a better solubilisation of hydrophobic proteins (for example, membrane proteins, which are important for parasite invasion and survival in host cells). In fact, 14.2% of proteins identified by 1-DE possess at least one transmembrane domain (TM domain), which is higher than 9.7% for 2-DE. 1-DE has identified more proteins than 2-DE in this study, which has also been observed in previous studies [265, 269]. In addition to that, 1-DE benefits high throughput proteomic analysis with a simpler experiment setup and less labour intensity.

In this study, a similar Tris solubilisation strategy as used for a *Plasmodium* proteome [261] was tested on a second 1-DE experiment. The analysis of 50 slices from the most divergent region led to the identification of 335 release 4 genes, of which 82 genes were not previously identified in the first 1-DE experiment. The 24.4% increase in protein identification may look encouraging initially. However, among 359 release 4 genes identified from the equivalent region on the first 1-DE experiment, 108 genes have not been identified in the second 1-DE experiment. Taking into account that a 23-38% increase of protein identification can be expected simply by analysing the same sample twice by LC-MS/MS [273], and considering the extra labour required to complete the analysis, only the initial 50 slices have been

analysed in the second 1-DE experiment. The same Tris solubilisation strategy was applied for the more automated MudPIT platform.

In comparison with 1-DE and MudPIT, 2-DE has difficulty in resolving very large or small proteins, low abundance proteins, hydrophobic proteins and proteins with extreme pIs. 2-DE experiments are also very labour intensive. However, despite these limitations, the 2-DE approach has uniquely identified 40 release 4 genes in this study. More importantly, 2-DE is able to provide additional information about the expressed proteome that other platforms fail to provide. It delivers a reference map of intact proteins, which reflects changes in protein expression levels, isoforms or post translational modification (PTM).

In this study, clusters of proteins from different gel spots are often found to share the same identification as shown in Figure 2.4. These clusters of proteins are likely to represent isoenzymes or proteins with PTMs. This phenomenon highlights an important application of 2-DE, PTM analysis. For example, protein phosphorylation is a key PTM that is crucial in controlling enzyme activities, protein degradation and particularly, *T. gondii* invasion and host cell interaction [274, 275]. Several phosphoprotein groups identified in this study showed as horizontal strings of spots on the gel since proteins containing negatively charged phosphate groups are separated according to their pI differences. 2-DE provides a rapid and straightforward visualization method of the potential PTM events; peptides from proteins of interest can be enriched by affinity chromatography or chemical approaches [276-278] and analysed further by tandem MS.

MudPIT identified the largest number of proteins in this study. Using a gel free platform, MudPIT can identify proteins with extremes of isoelectric point and

molecular weight [139]. Simplified sample separation in peptide space together with the use of two dimensional chromatography increases the resolution of separation and also provides higher loading capacity. This allows low abundance proteins and hydrophobic proteins to be identified [279]. In this study, 18.0% of proteins identified by MudPIT possess at least one TM domain, which is higher than the results from both 1-DE and 2-DE. In addition to this, MudPIT analysis is readily automated which benefits the high throughput requirement of proteomic research. However, MudPIT generates a highly complex peptide mixture, which requires considerable computing power for MS data searching. The dedicated and costly instrument setup and operational expertise required also prevent easy access for individual labs.

With various protein identification techniques available in this rapidly developing proteomics field, choosing a good combination of techniques is central to providing the largest number of protein identifications whilst utilising limited amounts of sample and labour. In this study, 2-DE gels with two pH gradient ranges were used to enable a broad sampling from pH 3-10, and at the same time achieve increased resolution for the detection of low abundance proteins in pH 4-7 range. By running the second 1-DE experiment, it has been demonstrated that additional protein identifications can be achieved with repeated analysis, but it is unclear whether the extra proteins identified are the results of difference in 1-DE gel separation or the repeated LC-MS/MS analysis. It is hard to justify the extra material and labour required to complete the analysis of the second 1-DE experiment especially when a Tris pre-fractionation method has been applied to a more automated MudPIT platform where five soluble replicates and four insoluble replicates have been analysed.

2.4.3 Database design to maximize protein identification

A successful translation of the raw MS data to protein identification relies on a good sequence database design that can maximize the coverage of potential protein coding sequences (CDS). This will make greatest use of the peptide information acquired in the previous steps. As reviewed in section 1.1.4, various gene prediction algorithms have been used for *T. gondii* genome annotation. These include *ab initio* gene prediction methods TigrScan and GlimmerHMM, and integrative methods such as GLEAN and TwinScan that combine experimental data, comparative genomic alignments and *ab initio* methods [280-282]. A major updated version was made available on ToxoDB version 4 in 2006 (release 4 gene model), which integrated experimental data such as expressed sequence tags (ESTs) [112].

However, despite all the effort that has been made, the release 4 gene model is far from a perfect dataset that can be used as the only sequence database for MS data searching. It was shown that 41% of release 4 gene models are likely to be imperfect containing at least one inconsistency with full-length cDNAs [208]. Comparison of release 4 gene models and genes predicted by other methods revealed a more complicated mixture of gene models. It has been found that any two prediction methods used generally share less than 12% identical predicted genes (head to head comparison) and 68% to 87% of sequences predicted by each prediction methods are unique to other prediction methods [270].

The above finding confirmed that using any single gene model is not sufficient to cover all the possible coding sequences. In this study, a lot of effort has been made to overcome this issue. MS data generated from the 1-DE experiment was initially searched against a sequence database that comprised all applicable coding sequences on ToxoDB version 3. Upon the publication of release 4 gene models in 2006, a new

sequence database was designed to achieve the best coverage of potential *T. gondii* coding sequences. Both release 4 gene models and gene models provided by other prediction methods and experimental evidence such as clustered expressed sequence tags (collectively termed as alternative gene models) were used. These were complemented by the inclusion of open reading frames (ORFs) translated from genomic sequences. MS data generated from the 1-DE experiment was re-searched against the new sequence database and subsequent data analysis was repeated. Searching against the new sequence database enabled us to realize the imperfection of release 4 gene models and led to the development of strategies to improve genome annotation and shorten the time consuming data re-submission process, which is further discussed in Chapter 4.

2.4.4 Search engines and result verification

The successful identification of low-abundance proteins in the sample has been an important issue in proteomic research. For example, the expression of 9 surface antigen related genes in *T. gondii*, which serve important roles in host cell attachment and interfacing with the host immune response during invasion [69], were found to be below the 10 percentile in a microarray experiment [112]. If the expression of these genes remains low at a translational level, the successful identification of these genes in a proteomic study is expected to encounter some technical limitations.

Firstly, the stochastic nature of peptide sampling by the mass spectrometer leads to a bias towards more protein identifications from peptides of higher concentrations. Secondly, different search engines used in MS data analysis also vary the results because of the difference in the search engines themselves as well as different false discovery rates applied [273, 283]. A recent test sample study distributed 20 highly

purified recombinant human proteins to 27 independent laboratories and compared the protein identification results acquired. Although most of the laboratories generated high quality MS data which was sufficient to identify all 20 proteins tested, only seven laboratories correctly reported all the proteins due to the differences in database setup and the search engine used [267].

In this study, manual inspection has been used to preserve the valuable expression evidence of low-abundance proteins acquired by MS data and protein identifications were also verified by searching against decoy and contaminant databases. Protein identification based on a single peptide and proteins that return a Mascot score < 60 are normally regarded as low confidence identifications, but these can also represent real expression of low-abundance proteins. For 1-DE and 2-DE results, manual inspection of individual peptide MS/MS spectra has been carried out on the proteins identified by Mascot (as described in section 2.2.7.1).

For MS data searching of MudPIT results, although manual inspection is not feasible, a decoy database that contains reversed sequences from the target database was used to estimate the false discovery rates (FDRs). The FDRs were calculated at 3.84% for Tris-insoluble fractions and 3.11% for Tris-soluble fractions, which are very respectable figures as < 5% is considered a low FDR for MudPIT [284, 285].

Approaches have been developed to improve the performance of MS data searching by adapting the results from multiple searching engines. This approach can significantly reduce the FDR of the result [284] and improve the confidence of protein identification based on a single peptide [284, 286]. Using multiple search engines for MS data allows on average 35% more peptide identifications to be made at a fixed FDR of 1% compared with using a single search engine [286]. The

development of this multiple search engine platform is nearly completed in the Wastling group in collaboration with Dr Andy Jones, University of Liverpool. It is expected to bring significant benefits to the current study and on-going proteomic studies by re-querying raw MS data using this platform.

2.4.5 Conclusion

Through the years, the rapid improvement of genomic and proteomic techniques has provided us with an ever improving MS detection sensitivity and capability, more accurate genome annotations and more efficient MS data search engines. It is essential to harness the latest technical development in this dynamic proteomic research. In this study, the three complementary high-performance proteomic platforms as well as a carefully designed up-to-date database and searching strategy for MS data acquisition allowed us to achieve a comprehensive coverage of the expressed *T. gondii* tachyzoite proteome. Comparing the results of this study with what had been achieved in 2002, a great improvement has already been seen. It is safe to speculate that a better coverage can be achieved in the near future within this on-going field of proteomic research.

While the identification of several thousand expressed proteins provides a milestone for proteomic research, in combination with the power of bioinformatics interpretation, the protein expression data will further benefit the understanding of *T. gondii* biology, which is investigated and discussed in Chapter 3. To harness the proteomic data in a broader, system biology level, the value of the proteomic data in the process of genome annotation and the comparison between proteomic and transcriptomic data are discussed in Chapters 4 and 5, respectively.

Chapter 3

Bioinformatics interpretation of the proteomic data

3.1 Introduction

Once the expressed proteome has been identified, bioinformatics interpretation plays an important role in understanding the biological functions of the proteins. As reviewed in section 1.2.5, functional assignment of an unknown protein using a bioinformatics approach relies on the assumption that proteins which have similar amino acid sequences or similar structures share similar functions. Data interpretation typically combines evidence from two approaches: databases that provide manually curated data from literature and predictions made by computer programs that automatically transfer existing knowledge about a homologous sequence to the targeting sequence.

For *Toxoplasma gondii*, ToxoDB (http://toxodb.org/toxo/) serves as a functional genomics resource that hosts the largest collection of *T. gondii* genome sequences and annotations [112] and is widely accredited as such by the *T. gondii* research community. On ToxoDB, product description and GO annotation of release 4 genes are provided where applicable [108]. This information provides valuable insights into the function of *T. gondii* genes and was used as the primary resource for data interpretation in this study.

In addition to the information provided on ToxoDB, there are a range of specific and universal prediction programs that can be used to infer the biological functions of the proteins identified. SignalP [178] and TMHMM [179] predict the existence of signal peptides and transmembrane domains in a protein sequence, respectively. The entry of virtually all proteins into the secretory pathway is controlled by signal peptides [287, 288]. In *T. gondii*, signal peptides direct proteins to important localizations central to invasion, such as the extracellular surface during gliding motility [65], the

secretory traffic to apical organelles [83], the dense granule [289] and the apicoplast [290]. Proteins with transmembrane domains also play an important role in the tachyzoite stage such as in gliding motility [62], the moving junction [80] and, the majority of the dense granule proteins (GRAs) are predicted to be transmembrane proteins [291]. The results of these two prediction programs can be used to infer the biological properties of the expressed proteome as well as to examine the quality of proteomic sampling when compared with whole genome predictions.

With the exception of *Cryptosporidium*, most of the clinically important Apicomplexan parasites studied possess two endosymbiotic organelles which carry DNA in addition to the nucleus: the mitochondrion and the apicoplast [292-296]. The growth of *T. gondii* is inhibited by drugs that impair apicoplast autonomy at the level of DNA replication, transcription, RNA processing and translation [297, 298]. Its algal origin also means many proteins and pathways are not shared by the human host. Together, these properties make the apicoplast a very promising drug target. Comparatively little is known about the *Toxoplasma* mitochondrion. This may partly reflect the diversion of research interest to the apicoplast, and the difficulty in defining the mitochondrial genome [299]. However, studies have linked mitochondrial function with *T. gondii* stage conversion whereby the exposure of mitochondrial inhibitors stimulates the transition of tachyzoites to bradyzoites *in vitro* [300, 301]. Additionally, the exact role of the mitochondrion in energy metabolisms still requires more investigation [299, 302]. The importance of these two organelles in Apicomplexan parasites led to the development of two specific prediction programs PATS [180] and PlasMit [303] respectively, which were used in this study.

In addition to the resources provided on ToxoDB and programs that are designed to predict specific properties of a gene, the integrative protein information database InterPro [182] was used in this study to determine protein functions as well as GO annotations. Global sequence analysis tools were also used in the study, BlastP [167] and AmiGO [304] were used to infer functional information of proteins based on sequence similarity, and WoLF PSORT [181] was used to predict subcellular localizations of proteins identified. Together, the results of bioinformatics interpretation will not only represent the predicted distribution of the expressed proteome but also expand the knowledge of the functions of these proteins in the important biological processes of *T. gondii*.

3.2 Materials and Methods

Several bioinformatics prediction programmes were used to assist protein function determination. Briefly, protein sequences were searched using BlastP to characterize protein functions and infer possible subcellular localizations. SignalP was used to predict proteins that contain signal peptides; TMHMM was used to predict transmembrane domains contained within a protein; results returned from PATS, PlasMit and WoLF PSORT together with release 4 gene descriptions and GO cellular component predictions provided by ToxoDB were combined to obtain subcellular localization predictions of proteins.

Functional categorization was constructed using the GO classifications listed on ToxoDB for each release4 gene, which were then assigned to specific MIPS categories within the FunCatDB functional catalogue. Genes without a GO classification were assigned a putative MIPS category using additional information provided by Blast, Pfam domain alignments, InterPro and from independent literature searches.

3.2.1 Protein-protein Blast (BlastP)

BlastP (http://www.ncbi.nlm.nih.gov/BLAST/) is a powerful sequence alignment tool provided by National Centre for Biotechnology Information (NCBI). It compares protein sequences to other previously characterized protein sequences in the database. The alignment and conserved domain results based on sequence similarity are used to conjecture functional and evolutionary information of query sequences. Default settings were applied in this study.

3.2.2 SignalP

SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP/) is an online bioinformatics prediction server which predicts the presence and location of signal peptide cleavage sites in given amino acid sequences provided by The Centre for Biological Sequence Analysis at the Technical University of Denmark [178]. In this study, sequences were uploaded in fasta format, default settings were applied for searching.

3.2.3 TMHMM 2.0

TMHMM 2.0 (http://www.cbs.dtu.dk/services/TMHMM-2.0/) is a bioinformatics prediction server provided by The Centre for Biological Sequence Analysis at the Technical University of Denmark [179]. The server is programmed to predict transmembrane helices in proteins using a hidden Markov model. In this study, sequences were uploaded in fasta format and default settings were applied for searching.

3.2.4 PATS

PATS (prediction of apicoplast targeted sequences) (http://gecco.org.chemie.uni-frankfurt.de/pats/pats-index.php) is an online bioinformatics program provided by the Molecular Design Laboratory, Goethe University, Frankfurt which provides predictions of apicoplast targeted sequences in Apicomplexan parasites. PATS uses a neural network analysis and has been trained with *Plasmodium falciparum* proteins [180]. In this study, default settings were applied.

3.2.5 PlasMit

PlasMit (http://gecco.org.chemie.uni-frankfurt.de/plasmit/index.html) is an online neural network software also provided by the Molecular Design Laboratory, Goethe University, Frankfurt, which predicts mitochondrial transit peptides in *P. falciparum*

[303]. In this study, PlasMit was used to predict possible mitochondrial proteins in conjunction with other prediction software.

3.2.6 WoLF PSORT Prediction

WoLF PSORT (http://wolfpsort.org/) is an online server for protein subcellular localization predictions provided by the Computational Biology Research Centre, National Institute of Advanced Industrial Science and Technology (AIST), Japan [181]. The server has been trained with yeast sequences from SWISS-PROT with the annotation of Yeast Protein Database (YPD). In this study, the organism type was set as animal and default settings were applied. To be considered a valid prediction the possibility percentage of the first localization result must be at least twice that of the second localization percentage.

3.2.7 AmiGO

AmiGO (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi) is the official online database for searching and browsing the Gene Ontology database and is provided by an international consortium [304]. In this study, AmiGO was used to match the protein function information retrieved from literature or other data sources to the controlled vocabulary of terms and identifiers defined by the Gene Ontology project. A Blast tool provided by AmiGO was also used to search GO descriptions of a protein when no GO annotation is available elsewhere.

3.2.8 Munich Information Centre for Protein Sequences Functional Catalogue (MIPS FunCat)

MIPS FunCat (http://mips.gsf.de/projects/funcat) is an annotation scheme for the functional description of proteins from prokaryotes, unicellular eukaryotes, plants and animals [305]. The FunCat consists of 28 main functional categories that cover

general fields like cellular transport, metabolism and cell rescue, defence and virulence. In this study, proteins identified were assigned to the most relevant category according to the functional information provided by GO annotation on ToxoDB, Blast, Pfam domain alignments, InterPro and independent literature searches.

3.2.9 Metabolic Pathway Coverage

The KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway (www.genome.jp/kegg/) [306] was used to examine the coverage of identified proteins on key metabolic pathways. Conversion of *T. gondii* genes to key metabolic pathway components was determined using the Metabolic Pathway Reconstruction tool for *T. gondii* available on ToxoDB (http://roos-compbio2.bio.upenn.edu/~fengchen/pathway/). Glycolysis and gluconeogenesis are important energy production pathways in *Toxoplasma* and were chosen as model pathways in this study. Proteins identified from tachyzoites were mapped onto the reconstructed glycolysis and gluconeogenesis pathways for *T. gondii*.

3.3 Results

3.3.1 SignalP and TMHMM predictions

Proteins identified from *T. gondii* tachyzoites were subjected to SignalP and TMHMM predictions. In this study, 10% of the identified official ToxoDB release 4 genes were predicted to contain a signal peptide and 17.6% of the identified official ToxoDB release 4 genes contain transmembrane domains. Predictions of identified alternative gene models and ORFs returned results of 9% and 21% respectively.

The results are closely similar to *T. gondii* genome predictions for signal peptide and transmembrane containing proteins provided on ToxoDB (11% and 18% respectively) and indicate unbiased sampling of the study. Similar proportions of signal peptide have been reported in the expressed proteome of *Plasmodium falciparum* [261] and similar proportions of transmembrane domain containing proteins are found in the expressed proteome of *Cryptosporidium parvum* and *P. falciparum* [261, 265]. Together they suggest a good sampling standard has been achieved in proteomic studies where signal peptide containing proteins and transmembrane domain containing proteins are not under represented.

3.3.2 Subcellular Localization Prediction

The subcellular localizations of proteins identified were collectively inferred by the predictions made from PATS, PlasMit, WoLF PSORT and release 4 gene descriptions and GO cellular component classification provided by ToxoDB. Figure 3.1 shows the distribution of predicted subcellular localizations of expressed proteins in the tachyzoite stage.

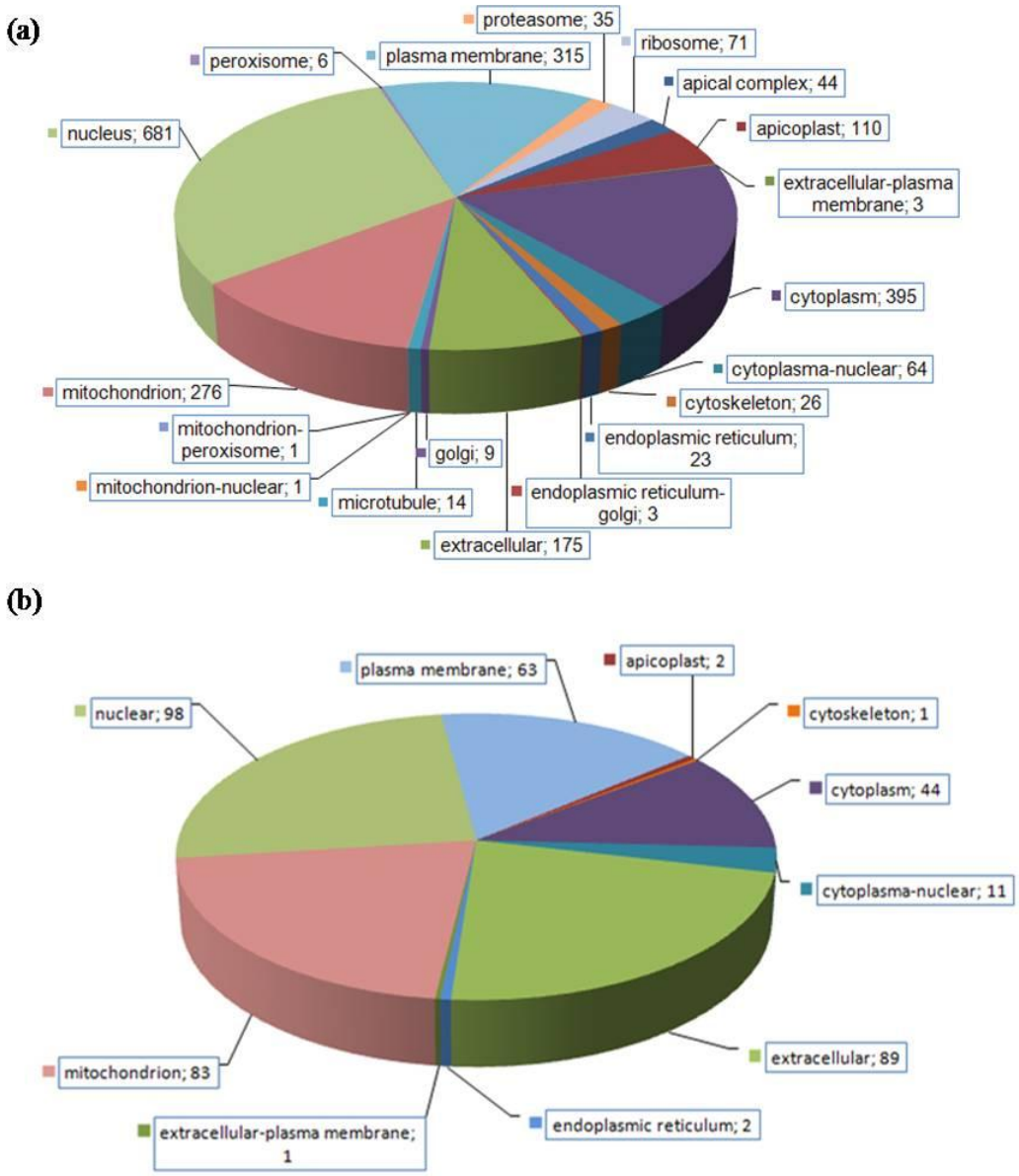


Figure 3.1 Subcellular localization of the expressed tachyzoite proteome. (a) Subcellular localization of identified official ToxoDB release 4 genes. The prediction was first assigned according to gene descriptions and GO annotation provided by ToxoDB. The sequences of proteins with no information provided on ToxoDB were submitted to PATS, PlasMit and WoLF PSORT. The results were combined to obtain predicted subcellular localizations. (b) Subcellular localization of identified alternative gene models and ORFs. The prediction was made by the combined results of PATS, PlasMit and WoLF PSORT.

A wide distribution of subcellular localizations in the expressed proteome was observed. The subcellular localization predictions of identified alternative gene

models and ORFs have shown a different distribution pattern, which was influenced by the resources available. The impact of prediction programs used was discussed in section 3.4.1.3. For release 4 genes identified, nuclear, cytoplasmic, plasma membrane and mitochondrial proteins are among the most represented categories which reflect the rapid gene expression, protein synthesis, extracellular interaction and energy generation events required in the tachyzoite stage which is the rapidly multiplicative, invasive form of the parasite. More than 100 proteins were predicted to be apicoplast proteins. A significant number of apical complex proteins (44) were detected which potentially have important roles in invasion and maintenance in host cells. There are also many proteins that are putatively involved in secretory pathways that are predicted to be located to the endoplasmic reticulum-golgi, plasma membrane and extracellular locations.

3.3.3 Functional Categorization

Functional information for the proteins identified was acquired from gene descriptions and GO annotations listed on ToxoDB, and these were then assigned to specific MIPS categories. Proteins that have no GO classification were assigned a putative MIPS category using additional information provided by BlastP, AmiGO, Pfam domain alignments, InterPro and from independent literature searches. Figure 3.2 shows the functional categorization of proteins identified in this study.

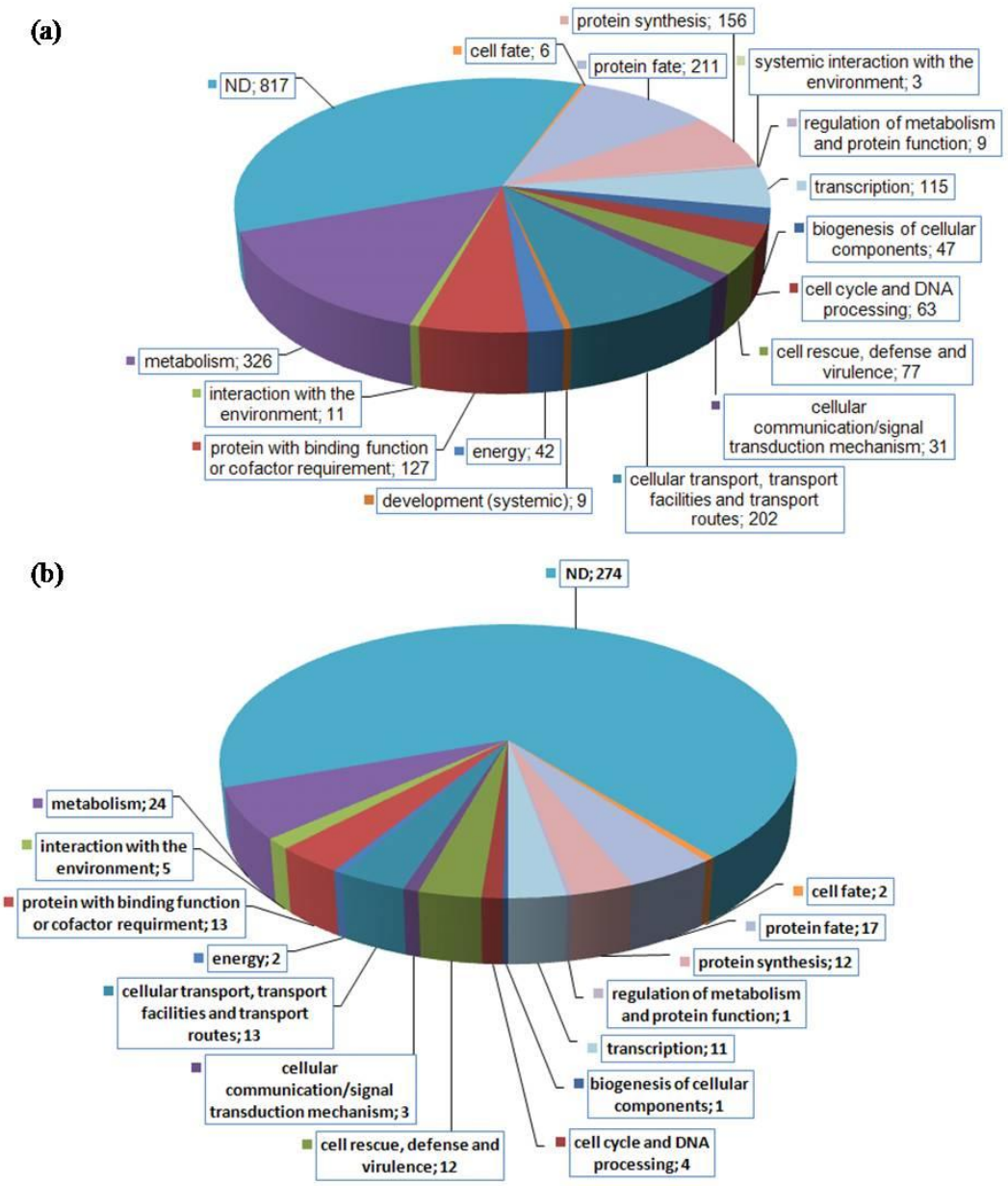


Figure 3.2 Functional categorization of the expressed tachyzoite proteome. Functional assignments of identified official ToxoDB release 4 genes. The prediction was first determined by gene description and GO annotation provided on ToxoDB and then assigned to appropriate MIPS FunCat categories. Putative functional assignment was made to the remainder of identified proteins with information acquired from BlastP, Pfam domain alignments, InterPro and literature searches. (b) Functional assignment of identified alternative gene models and ORFs. The prediction was made by information acquired from BlastP, Pfam domain alignments, AmiGO blast and then assigned to appropriate MIPS FunCat categories.

Similar to subcellular localization predictions, functional categorization of identified alternative gene models and ORFs has shown a different pattern than that of release 4 genes, which is further discussed in section 3.4.1.3. For release 4 genes identified, categories that are highly represented in the results are metabolism, cellular transport, protein synthesis and protein fate (folding, modification, destination) which reflect the highly active metabolism, protein synthesis and cell division functions required in the tachyzoite stage. Importantly, 77 proteins are assigned to the "cell rescue, defence and virulence" category, many proteins of which are required in the invasion of and maintenance in the host cell by the tachyzoite stage.

3.3.4 Metabolic Pathway Coverage

ToxoDB provided a reconstructed version of the KEGG pathway for *T. gondii*. Proteins identified in this study were cross-referenced with the glycolysis and gluconeogenesis pathways. Table 3.1 lists the *Toxoplasma* gene ascribed to each constituent of the glycolysis and gluconeogenesis pathways. The EC numbers are included for reference to Figure 3.3.

Table 3.1 Component enzymes of the glycolysis and gluconeogenesis pathways, EC numbers and corresponding *Toxoplasma* gene identifiers.

EC Numbers	Description	ToxoDB ID
1.1.1.1	Alcohol dehydrogenase	583.m05453
1.1.1.27	L-lactate dehydrogenase	44.m00006
1.1.1.27	L-lactate dehydrogenase	80.m00010
1.2.1.12	Glyceraldehyde 3-phosphate dehydrogenase (phosphorylating)	80.m00003
1.2.1.3	Aldehyde dehydrogenase (NAD(+))	41.m00032
1.2.4.1	Pyruvate dehydrogenase (acetyl-transferring)	50.m03083
1.2.4.1	Pyruvate dehydrogenase (acetyl-transferring)	50.m03618
1.2.4.1	Pyruvate dehydrogenase (acetyl-transferring)	59.m03618
1.8.1.4	Dihydrolipoyl dehydrogenase	20.m03954
2.3.1.12	Dihydrolipoyllysine-residue acetyltransferase	20.m00373
2.3.1.168	Dihydrolipoyllysine-residue (2-methylpropanoyl)transferase	641.m00177

2.7.1.1	Hexokinase	57.m00001
2.7.1.11	6-phosphofructokinase	42.m00123
2.7.1.11	6-phosphofructokinase	49.m03242
2.7.1.40	Pyruvate kinase	129.m00253
2.7.1.40	Pyruvate kinase	55.m00007
2.7.2.-	Phosphotransferases with a carboxy group as acceptor	37.m00745
2.7.2.3	Phosphoglycerate kinase	41.m01331
2.7.2.3	Phosphoglycerate kinase	641.m00193
3.1.3.11	Fructose-bisphosphatase	20.m03907
3.1.3.11	Fructose-bisphosphatase	46.m01668
3.1.3.11	Fructose-bisphosphatase	50.m00005
3.6.1.7	Acylphosphatase	55.m04800
4.1.2.13	Fructose-bisphosphate aldolase	46.m00002
4.1.2.13	Fructose-bisphosphate aldolase	46.m02920
4.2.1.11	Phosphopyruvate hydratase	59.m03410
4.2.1.11	Phosphopyruvate hydratase	59.m03411
4.6.1.-	Phosphorus-oxygen lyases	44.m02781
4.6.1.-	Phosphorus-oxygen lyases	52.m01648
5.3.1.1	Triosephosphate isomerase	42.m00050
5.3.1.1	Triosephosphate isomerase	44.m02801
5.3.1.9	Glucose-6-phosphate isomerase	76.m00001
5.4.2.1	Phosphoglycerate mutase	113.m00016
5.4.2.1	Phosphoglycerate mutase	59.m03656
6.2.1.1	Acetate--CoA ligase	57.m03124

Note: Data acquired from the Metabolic Pathway Reconstruction for *T. gondii* available on the KEGG Pathway site accessed via ToxoDB (http://roos-compbio2.bio.upenn.edu/~fengchen/pathway/).

In the predicted *T. gondii* glycolysis and gluconeogenesis pathways, all but two enzyme components were detected in the expressed proteome, the exceptions being (EC 3.6.1.7-acylphosphatase and EC 1.1.1.1-alcohol dehydrogenase). Figure 3.3 shows the coverage of the expressed tachyzoite proteome on the reconstructed KEGG pathway of glycolysis and gluconeogenesis for *T. gondii*.

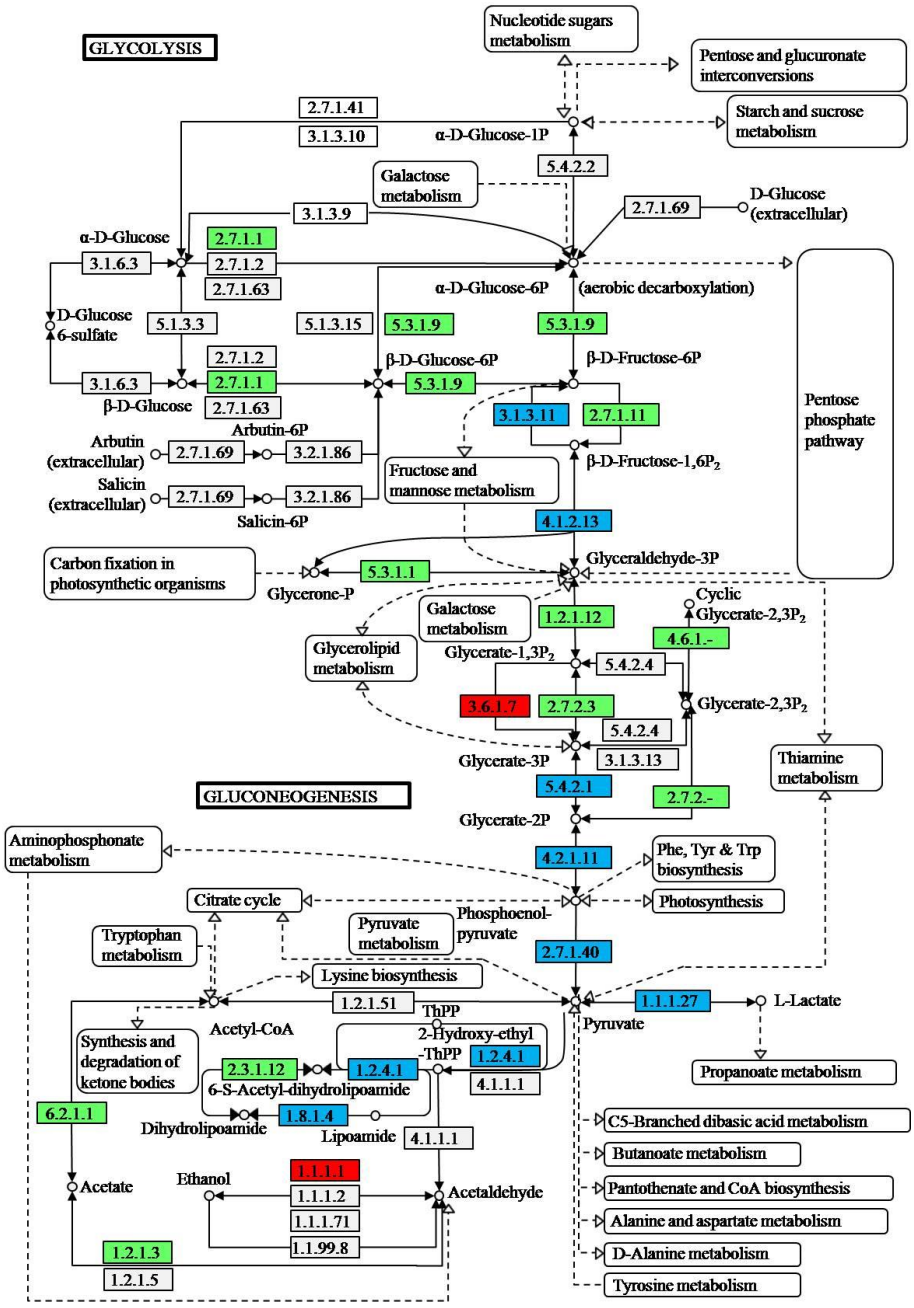


Figure 3.3 Metabolic pathway coverage: Glycolysis and gluconeogenesis. Conversion of *T. gondii* genes to key pathway components was determined using the Metabolic Pathway Reconstruction for *T. gondii* available on the KEGG Pathway site accessed via ToxoDB (http://roos-compbio2.bio.upenn.edu/~fengchen/pathway/). EC number to ToxoDB gene mapping is listed in Table 3.1. Enzymes coloured in green and blue indicate expression evidence which has been confirmed by mass spectrometric data; blue also signifies genes identified by the 2-DE. Enzymes coloured in red are pathway components that have not been identified in this study. All other enzymes (shown in white) are those for which no corresponding *T. gondii* gene has been assigned.

3.4 Discussion

3.4.1 Coverage of the expressed tachyzoite proteome

The results of the bioinformatics interpretation show the expressed proteome of tachyzoites identified in this study covers a wide spectrum of proteins. SignalP and TMHMM results are very similar to the whole genome prediction which probably indicates that the sampling method used in the study was unbiased towards membrane and secreted proteins. It is not known what proportion of the expressed proteome comprises signal peptide or transmembrane domain containing proteins in the individual life cycle stages; however, it is likely that the rapidly dividing, invasive tachyzoite would employ a significant number of secreted proteins. The prediction that nearly one fifth of proteins identified possess at least one transmembrane domain reflected the importance of transmembrane proteins in protein secretion during the host cell invasion and intracellular survival.

A broad distribution of subcellular localizations and functional categories and a near complete coverage of the glycolysis and gluconeogenesis pathways are also evidence that a sensitive and non-biased sampling has been achieved. The profile of subcellular localization and functional categorization reflect well the biological requirements of rapidly dividing, invasive tachyzoites. The expressions of proteins from several important subcellular compartments that serve unique biological functions have been confirmed by proteomic techniques.

Categories with particular biological interest include proteins secreted by the three distinct organelles micronemes, rhoptries and dense granules. The sequential secretion from these organelles plays an important role in host cell invasion and intracellular survival (as discussed in section 1.1.3). Additionally, structural proteins

that support cellular movement during invasion have been identified as well as proteins from the unusual apicoplast organelle.

3.4.1.1 Invasion and survival-apical complex, secretory organelles and others

Despite the lack of organelle specific prediction programs, the expression of several known secretory organelle proteins have been detected in this study according to the ToxoDB gene description. These include 12 microneme proteins, 7 rhoptry proteins and 7 dense granule proteins. The apical complex, which contains micronemes, rhoptries, dense granules, the conoid and cytoskeletal components, is essential in both host cell invasion and survival of Apicomplexan parasites [83, 307, 308]. Forty four proteins identified in this study are annotated to locate to the apical complex. To achieve an accurate result, the assignment was entirely based on ToxoDB either by gene description or annotated GO cellular component. However, there are two other published studies that focused on the sub-proteome of the apical complex which indicate a potentially broader collection of apical complex proteins that exist.

3.4.1.1.1 Comparison with experimental data

In a proteomic study that specifically enriched conoid/apical complex material [251], 179 proteins (release 4 genes) were identified as apical complex proteins after excluding contaminants from other subcellular organelles. Of those, 59 proteins were detected in multiple replicates by multiple peptides which were regarded as good evidence of apical complex proteins. However, as the list of 59 proteins is not released, a direct comparison cannot be made with this study. Among the 179 proteins identified at the first stage, 104 have been identified in this study but only one protein (583.m05259, conserved hypothetical protein) has been assigned as an apical complex protein.

In the other study that characterized the proteome of purified rhoptry organelles [84], 56 proteins (release 4 genes) were identified, of which 50 proteins have been identified in this study. Out of the 50 proteins, seven have been assigned as apical complex proteins on ToxoDB. The comparison with the other organelle specific proteomic studies indicates the sensitivity of this study is high. However, it also highlighted the fact that with limited information confirmed by genome annotation and the lack of apical organelle specific prediction programmes, the results of this study are likely to under-represent apical complex proteins. It also indicates that potential apical complex proteins have been mistakenly assigned to other general categories. For example, among 104 proteins which overlapped with conoid/apical enrichment study, two proteins have been assigned to cytoskeleton and 22 proteins have been assigned to the cytoplasm.

3.4.1.1.2 Comparison with *in silico* predictions

A recent study that employs bioinformatic approaches to predict candidate proteins associated with the apical organelles [309], focused on signature Pfam domains like PAN, TSP-1 and EGF motifs found in microneme proteins, and then searched the domain patterns in all available completed Apicomplexan genome sequences. The results suggested 60 candidate microneme proteins in *T. gondii* [309]. The whole tachyzoite proteome confirmed the expression of 15 of those candidate microneme proteins, 8 of which have not been assigned as "apical" in this chapter as no further localization confirmation can be found elsewhere.

Together, the comparison of the data acquired in the current study with other experimental and computational studies showed that this study has achieved a good coverage of potential apical complex proteins and provided important protein expression evidence for more focused apical complex studies in the future. In

addition to that, based on different results from several studies, no well-defined apical complex proteome has been agreed. The prediction method used here, which is largely based on ToxoDB annotation, is suited to a global proteome survey rather than an in-depth study and the result serves as a good starting point in understanding apical complex proteomes.

3.4.1.1.3 Examples of other important proteins during invasion process

During invasion, the motor complex is formed by a class XIV myosin (MyoA), the myosin light chain and two gliding-associated proteins (GAPs) [65]. There are five known class XIV myosins in *T. gondii*: myosin A, B, C, D and E [310-313]. In this study, 14 proteins identified are annotated as myosin related proteins on ToxoDB. These include the expression of four putative myosins: A, C, D and E. In fact, there is no myosin B annotated on ToxoDB possibly due to the fact that myosin B and myosin C are the products of differential RNA splicing and share the majority of their sequences [310]. Additionally, four myosin light chain proteins have also been identified as well as five actin proteins. Although gliding-associated proteins (GAPs) have not been annotated on ToxoDB from which no protein identification can be made, good proteomic coverage of the glideosome has been achieved in this study.

Another interesting group of proteins are SRS (SAG-related sequences) proteins which are thought to mediate attachment to host cells and activate host immunity thereby regulating the parasite's virulence [69, 314]. There are 46 SRS proteins predicted on ToxoDB version 4, of which 17 proteins were identified in this study. Confirmation of protein expression of these genes in this study will provide higher confidence in further characterization of SRS proteins.

3.4.1.2 Endomembrane system, apicoplast and metabolic pathways

The cell nucleus, cytoplasm and mitochondrion are among the subcellular locations that are well represented by protein constituents. Proteins involved in metabolism, transcription, protein synthesis, protein modification and degradation are required by the parasite for cell division and re-modelling and are highly represented in this study.

The endomembrane system plays an important role in secretory pathways [315]. In this study, 23 proteins identified are predicted to locate to the endoplasmic reticulum (ER). The expression of three COPII-coated vesicle proteins are also detected which mediate the protein transportation between the ER and Golgi [316] as well as 9 proteins that are predicted to locate to the Golgi. There are also a large proportion of proteins that are predicted to be involved in cellular transport and proteins with binding functions, which have potential roles in transporting host nutrients to the parasite and supporting cytoskeleton during invasion. In addition to the ER to Golgi route, proteins encoded by nucleus are also transported through the ER to an important endosymbiotic organelle, the apicoplast [290, 317].

3.4.1.2.1 Protein trafficking to the apicoplast

With a genome size of just 35 kb, the apicoplast only encodes genes required for gene expression (such as ribosomal RNAs and tRNAs) and a few protein coding genes [318]. The majority of genes coding apicoplast proteins have been transferred to the nuclear genome, and their expression products are targeted back to the apicoplast post-translationally [319]. The apicoplast is the location of several anabolic pathways such as the biosynthesis of fatty acids, isoprenoids (i.e. sterols and ubiquinones), and iron-sulfur clusters [290, 320, 321]. Inhibition of the apicoplast metabolic function or interference with its DNA replication is lethal for

the parasite which make it an attractive drug target [298]. Thus, valuable information is provided by a good understanding of the proteins targeted to the apicoplast.

In this study, with the information from both ToxoDB annotation and bioinformatics program PATS, 110 proteins identified are predicted to target to the apicoplast. Comparing this to the 222 apicoplast genes annotated on ToxoDB, expression of 86 (39%) of them have been confirmed in this study. Prediction of the remaining 24 putative apicoplast proteins was made by PATS, a program that can predict apicoplast targeted sequences in Apicomplexan parasites.

These include seven aminoacyl-tRNA synthetases (145.m00322, 145.m00604, 27.m00832, 39.m00356, 50.m00020, 55.m04665 and 80.m00063) and genes involved in anabolic pathways. For example, five genes are annotated to be involved in fatty acid biosynthetic processes (44.m00012, 49.m05646, 42.m03469, 55.m00019 and 76.m01567) and two genes are annotated to be involved in a mevalonate-independent pathway of the isopentenyl diphosphate biosynthetic process (42.m03570 and 55.m04989).

The biosynthetic pathways of the apicoplast require effective mechanisms to provide the organelle with carbon sources, ATP, and reducing power. Carbohydrate metabolism plays a central role in energy production and the synthesis of metabolites. In this study, five proteins identified in the apicoplast are involved in the glycolysis and gluconeogenesis pathways (129.m00253, 20.m00373, 50.m03083, 55.m00007 and 59.m03618) and two proteins are involved in the tricarboxylic acid (TCA) cycle pathway (42.m03524 and 76.m01567). A putative pyruvate dehydrogenase (50.m03083), a central enzyme for the carbohydrate metabolism that provides the link between the glycolytic pathway and the TCA cycle, is also identified in

apicoplast. The localization of pyruvate dehydrogenase (50.m03083) in the apicoplast, rather than its common localization of mitochondrion, has been confirmed by several reports [322-324]. However, a recent study has localized one of the TCA cycle enzymes (42.m03524, aconitate hydratase, putative) to the mitochondrion using a construct with a C-terminal myc epitope fusion [324], which contradicts the PATS prediction used in this study and likely indicates a prediction error. The reliability of PATS prediction is further discussed in section 3.4.2.

When the metabolic pathways were examined with the entire list of the expressed proteins identified in this study, a very promising coverage of glycolysis and gluconeogenesis pathways, as well as other pathways has been shown.

3.4.1.2.2 Metabolic pathway coverage

Glycolysis, gluconeogenesis and the TCA cycle are central pathways of carbohydrate metabolism. They are essential for matching the cellular demand for energy, reducing power and precursors for biosynthesis pathways.

With only two enzymes in the glycolysis and gluconeogenesis pathways not having been identified in this study (EC 3.6.1.7-acylphosphatase and EC 1.1.1.1-alcohol dehydrogenase), it has shown a good coverage of the proteome data on these important energy pathways in the tachyzoite stage. In fact, the protein expression of these two genes have not been detected by other proteomic studies listed on ToxoDB [112], which may reflect a technical limitation of the detection of these genes by proteomic approaches.

A good coverage has also been achieved on other key energy production pathways with 15 out of 18 enzymes identified in the *T. gondii* TCA cycle pathway and 26 out of 34 enzymes identified in oxidative phosphorylation pathway.

3.4.1.3 Coverage of bioinformatics prediction

In functional categorization assignments, despite employing the strategy of using multiple prediction programs, the function of 817 proteins (36% of the proteome identified in this study) cannot be clearly determined based on a sequence similarity approach alone. The same difficulty was seen in a proteomic study of related Apicomplexan parasite *C.parvum* [265] where the function of 39% of the expressed proteome was unclassified. Similarly, in a *P.falciparum* study [261], more than 40% of expressed proteins from various life stages were listed as hypothetical, conserved hypothetical or functional unclassified proteins.

The proportion of sequences with undetermined function is significantly higher for the 394 alternative gene models and ORFs at 70%. This reflects the larger proportion of atypical or truncated sequences that exists in the 394 sequences. At the same time, while subcellular localization prediction programs can make predictions on atypical or truncated sequences, sequence signatures and domain pattern based programs are very likely to misread the target sequences or completely miss an important domain, which can lead to biased prediction results. The reliability and coverage of bioinformatics prediction is further discussed in section 3.4.2.

3.4.2 Choosing the right prediction programs

With many free and commercial bioinformatics tools available for gene localization and function predictions, choosing the right programs is crucial in bioinformatics interpretations of proteome data. In this study, prediction programs were carefully selected. SignalP and TMHMM are commonly used in many studies as reliable prediction tools [265, 325-327]. They are also used in all EuPathDB online databases as standard prediction programs [112, 328-330], which allow the comparison of the proteomic data acquired in this study with *T. gondii* genome prediction on ToxoDB

feasible. For functional and subcellular localization predictions, evidence from universal prediction programs such as BlastP and WoLF PSORT as well as specifically designed Apicomplexan prediction programs such as PATS and PlasMit have been collectively used to achieve a more accurate prediction result.

However, bioinformatics interpretation based largely on sequence and structural similarities to previously characterized proteins has its own drawbacks. The results of the present study illustrate the difficulty in predicting functional information for novel proteins in species where no similarities can be found. For example, 36% of the expressed proteome has no functional annotation. It also has been reported that proteins with highly similar sequences can have different functions *in vivo*, and, conversely, proteins may show similar activities while lacking apparent sequence or structural similarity [331].

Different from functional assignment predictions where little information is available about novel proteins, subcellular localization prediction programs will give a prediction result on any sequences queried. The universal subcellular localization prediction program WoLF-PSORT used in this study gives a likelihood score of localization in percentage. A stringent parsing standard was applied in this study that, to be considered a valid prediction result the possibility percentage of the first localization result must be at least twice that of the second localization percentage. However, as discussed in section 3.4.1, comparing to the results of sub-protoeme studies, confilicts can still be found by using a univeral bioinformatics prediction.

Using more species specific prediction tools with a better defined training dataset provides a more accurate prediction. As mentioned in previous sections, PATS used in this study, which is trained with *P.falciparum* sequences, is the closest program

available that can predict apicoplast sequences for *T. gondii*. However, although both parasites are within the Apicomplexan phylum and share a lot of similarities, using PATS on *T. gondii* sequences can still cause potential problems. The successful prediction of an apicoplast sequence is based on the recognition of two sequence components, a typical endomembrane signal peptide and a plant-like transit peptide [290, 317]. While the first canonical signal peptides are similar, *T. gondii* transit sequences are enriched for serine and threonine [332] while *P.falciparum* transit peptides are enriched for asparagine and lysine residues [319]. Since PATS was primarily trained with *P.falciparum* sequences, the difference of transit peptides compositions may lead to false predictions in *T. gondii* sequences and a *T. gondii* specific prediction tool is needed in this case.

Another useful approach is to simplify the prediction effort by pre-fractionating a complex sample according to different biological properties, and characterizing the resulting sub-proteome. This will provide a useful training dataset for computer learning programs and result in a more focused and accurate bioinformatics interpretation. Several sub-proteome studies of Apicomplexan parasites are available [84, 253, 333, 334] and can be used to facilitate better bioinformatics program design.

3.4.3 Choosing the right categorization system

In addition to choosing the right prediction programs, another issue is to use the right categorization system. While subcellular localizations of proteins can be sorted on a relatively standard system with similar terminologies used, previous publications that study functional categorization of proteomic results often developed their own systems [335-337]. Several functional annotation schemes are available including the Riley scheme [338], MIPS FunCat [305] and GO [339].

The Riley schema was originally proposed for the functional annotation of *E.coli* however it lacks categories that cover parasite specific functions. MIPS FunCat and GO are better candidates since they both have a larger spectrum of functional category coverage.

MIPS FunCat was used in the study to categorise protein functional assignments. Although Gene Ontology provides organizing principles such as biological process and molecular function which make it a valid scheme for functional categorization, the MIPS catalogue has several advantages in this study.

Firstly, despite the considerable effort which has been made to assign GO annotation to *T. gondii* genes, the coverage of GO annotation is still quite low (i.e. around 23% for annotated GO molecular function category). Manually assigning a putative GO biological process and molecular function annotation to a gene would cause potential bias in the following categorization.

Secondly, with both "molecular function" and "biological process" annotation domains available in GO scheme, confusion might be caused whether to categorise protein functions based on their basic molecular functions or towards a broader system level classification of biological process. Although "biological process" was used in favour of "molecular function" in this study to represent a better system biology view, by only accepting one annotation category would make the GO coverage even lower.

Moreover, GO annotation is supported by MIPS FunCatDB where GO numbers can be queried against MIPS FunCat. Proteins with valid GO annotations on ToxoDB have been assigned to correlating MIPS category directly in this study.

In addition to those points, MIPS FunCat is a stable scheme and only four major extensions have been made since 1996 [305] compared to the constantly evolving and changing GO. MIPS FunCat has been used by several large-scale transcriptomic and proteomic studies [234, 340-342] including studies on Apicomplexan parasites *C.parvum* and *P.falciparum* [261, 265].

3.4.4 Conclusion

Bioinformatics interpretation is the first step towards understanding the biological roles of the expressed proteome. Despite a few drawbacks of sequence similarity-based prediction programs, they provide valuable inspections of the expressed proteins and enlighten further detailed studies on proteins of interest.

The following chapters will examine and compare the proteome data with other genomic expression data and make use of the experimental protein expression evidence to improve genome annotation methods.

Chapter 4

Data repository, integration of proteomic data onto ToxoDB and the validation of genome annotation

4.1 Introduction

In previous chapters, the importance of the proteomic data in expanding the current knowledge of protein expression as well as understanding biological functions using bioinformatics interpretation have been investigated and discussed. The standalone proteomic expression data have already added valuable input to the understanding of *T. gondii* biology. In this chapter, an effort has been made to investigate the benefit to the global research community of the integration of the proteomic data with other genomic and proteomic resources.

4.1.1 MS data repository

Several public repositories are actively hosting proteomic data for the research communities, such as the Proteomics identifications database (PRIDE) [192], the Global Proteome Machine databases (GPMDB) [197], PeptideAtlas [201], and Tranche [204]. As reviewed in section 1.2.5.2.2, data storing options offered by the first three platforms involve additional data processing using pipelines developed specifically for that repository. Compatibility issues have emerged due to the lack of a standardized data format for raw MS data. While collaborations between HUPO-PSI and the Institute for Systems Biology have resulted in a new data format, mzML, being proposed [191], more efforts are required for the new data formats to be developed and adopted.

This limitation made Tranche a good option for storing raw MS data. Firstly, any type of data file can be stored by the Tranche network. This allows the raw MS data generated from the various proteomic platforms used in this study to be hosted in one place. The valuable information embedded in raw MS data can be preserved for improved data analysis algorithms in the future. Secondly, by providing permanent

storage and "peer-to-server-to-peer" distributed network, the Tranche network addresses the problem of data loss through computer hardware failure or changes in staff in individual groups as well as providing a rapid data sharing platform for the research communities [195].

4.1.2 Integration of proteomic data onto ToxoDB and genome annotation

While raw MS data can be safely stored in the publically accessible Tranche network, the integration of proteomic data with organism specific genomic and proteomic resources provides further biological and technical advantages to *T. gondii* research. ToxoDB (http://toxodb.org/toxo/) provides such a platform that is ideal for the integration of the proteomic data generated in this study. Various components of ToxoDB such as genome annotation, gene description and GO annotation have been used to facilitate this proteomic study. The addition of detailed whole proteome expression data within these integrated workspaces is able to assist the on-going annotation of the genome of *T. gondii* as discussed in this chapter, as well as enabling the examination of the value of protein expression data in interpreting global gene expression, a topic which is investigated and discussed in Chapter 5.

While a typical bottom-up protein identification based proteomic study relies upon accurate predicted sequence databases, in chapter 2, we have shown the expression evidence for 394 non-redundant alternative gene models and ORFs that cannot be matched to release 4 gene models. This highlighted the incomplete nature of the release 4 genome annotation provided by ToxoDB. The integration of the proteomic data into ToxoDB has provided a good platform to systematically validate the accuracy of release 4 genome annotations such as providing evidence for the existence of genes and confirming exon-intron boundaries. With the ability to visualise proteomic data in a genomic context, discrepancies between proteomic

expression data and release 4 gene models can be examined in more detail leading to the confirmation of alternative gene models or previously unpredicted exons and genes.

In this chapter, the methods of storing and sharing raw MS data and integrating the proteomic data with other genomic sources on ToxoDB are examined. With the facilities provided by ToxoDB, the possibilities and applications of proteomic data in the field of proteogenomics are discussed.

During the production of this manuscript, ToxoDB release 5, the latest version of ToxoDB was released and a new version of genome annotation has been published. The proteomic data generated in this study have been integrated on ToxoDB using the same algorithm described here, followed by the addition of several latest proteomic studies [108]. In order to keep the consistency of nomenclature with analyses carried out in the previous chapters, ToxoDB release 4 was used in the examples shown this chapter. As an on-going effort of genome annotation, the implication of the proteomic data in the latest ToxoDB release 5 is discussed in section 4.4.1.

4.2 Materials and methods

4.2.1 Data depository on Tranche

A java based client was downloaded and installed from the Tranche website (https://proteomecommons.org/tranche/) onto a local computer. Following online instructions, MGF files generated by 1-DE and 2-DE experiments and MS2 files generated by MudPIT experiments were uploaded onto the Tranche network under the project name '*Toxoplasma* Proteome Liverpool'.

4.2.2 Peptide mapping for ToxoDB

4.2.2.1 Collection of peptide expression evidence

Proteomic data acquired in this study and from the *C.parvum* proteomic study performed by Dr. S.J. Sanderson ([265] and Appendix X), were mapped onto the genome scaffold on a peptide level at ToxoDB and CryptoDB, respectively. Identified peptides were collected by computer scripts written by Mark Heiges, University of Georgia (CryptoDB) and Brian Brunk, University of Pennsylvania (ToxoDB). For 1-DE and 2-DE results, the peptides identified were collected from the Mascot html result page of each individual protein identified. Peptides identified from MudPIT experiments of soluble and insoluble fractions were directly collected from Microsoft Excel spreadsheet results reported for the entire run. Each peptide identified was regarded as an individual entry and tagged with the identified protein and the related experimental platform that generated the data.

4.2.2.2 Mapping peptide entries onto the genome scaffold

The collected peptide entries were mapped onto the genome scaffold and aligned with other features (i.e. gene models, ORFs and other genome sources) by Mark Heiges from CryptoDB and Brian Brunk from ToxoDB. The following rules of

mapping were developed based on discussions between the Wastling group and developers from ToxoDB and CryptoDB. For the *C.parvum* proteomic data, only one gene prediction model is available; peptides which identified a gene were directly mapped and the rest of the peptide entries were mapped to ORFs. For the *T. gondii* proteomic data, peptides that hit release 4 gene annotations have been directly mapped onto the ToxoDB genome scaffold. The rest of the peptide entries that hit alternative gene models or ORFs (collectively termed as alternative models) were sent through a six step mapping algorithm, successful mapping at any step would finish for that entry.

Step 1, if all peptides from an alternative model could be mapped to an official release 4 gene, the release 4 gene was adopted and this is termed a 100% match; Step 2, if more than 50% of the peptides from an alternative model could be mapped to a release 4 gene, this was considered a valid mapping and the matching peptides were aligned with the corresponding release 4 gene; Step 3, if a certain set of peptides from an alternative model could be mapped to more than one release 4 gene, the gene that could host most peptides was reported; Step 4, alternative models not conforming to step 2 were mapped to ORFs; Step 5, an alternative model can be mapped to an ORF only if 100% of the peptides can be mapped; Step 6, if any of the peptides from the alternative model cannot be mapped to a release 4 gene or an ORF by step 2-5, the peptides were mapped to the alternative gene models (i.e. TgTwinScan, TgTigrScan and TgGlimmer).

4.2.3 ToxoDB integration and visualization

Following discussions between the Wastling group and collaborators from ToxoDB and CryptoDB, the data integration and visualization process was performed by developers at ToxoDB and CryptoDB. In brief, the following steps were carried out

for *T. gondii* proteomic data generated in this study. Upon the completion of the *T. gondii* proteomic data mapping on the ToxoDB genome scaffold, peptide identifications were integrated into individual gene report pages. A separate tool has been created for querying protein expression data based on mass spectral evidence.

The peptide identifications were also integrated into the ToxoDB genome browser (GBrowse) interface [343], where expressed peptides can be visualized in relation to various gene models and the genomic region from which the sequence is predicted to have been produced.

4.2.4 Validation of release 4 genome annotation by peptide expression data

Peptides identified that align across splice boundaries were reported by developers at ToxoDB during the mapping process. ToxoDB GBrowse was used in this study to visually confirm the correct predictions of gene models via peptide evidence as well as to examine discrepancies between gene predictions and peptide expression data.

4.3 Results

4.3.1 Data repository for raw MS data

All the raw MS data associated with this study can be viewed and downloaded from ProteomeCommons.org Tranche network [204] at https://proteomecommons.org/tranche/, using the following hash:

Ulv/yTYTaaHin5Tv4InpsgoUY1uTJQtdoLRi9HbdtypXqztv+BiVE/wZieBkqu6d3k
U20Vyejo0HYCfswgwiGyPHQPAAAAAAAAOhng==

A brief work flow to retrieve the data is shown in Figure 4.1.

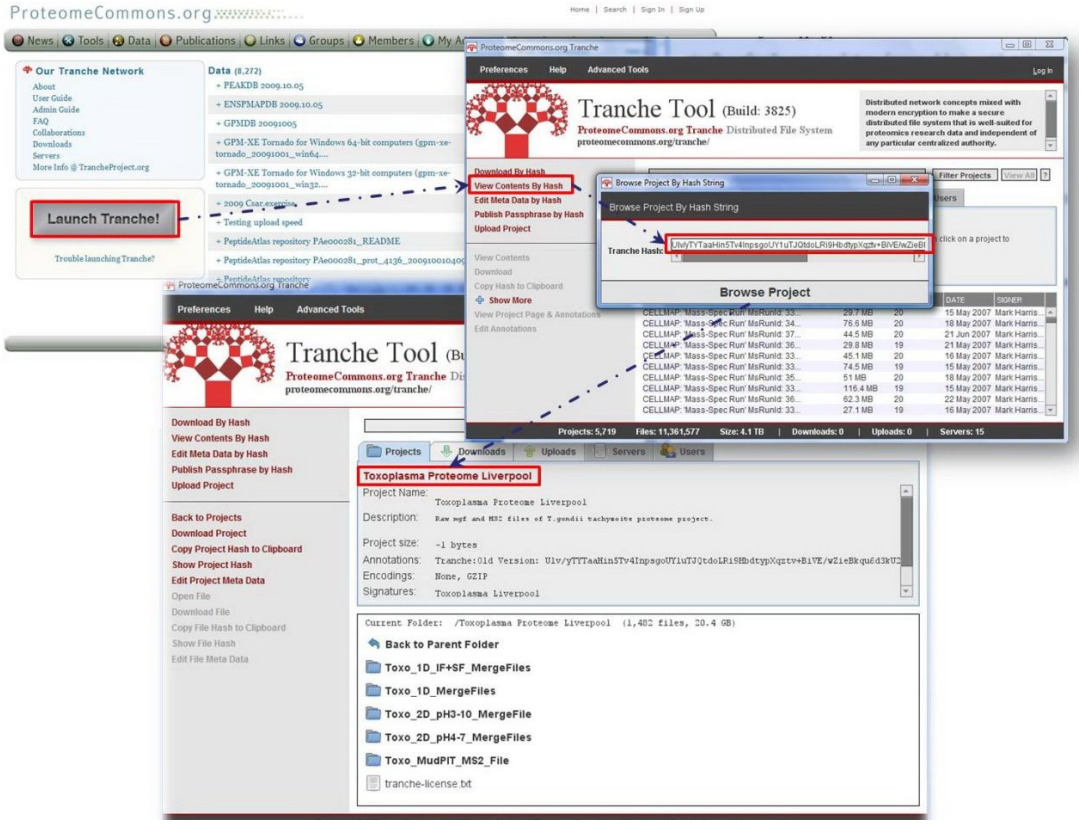


Figure 4.1 Data repository for raw MS data on ProteomeCommons.org Tranche network. All the raw MS data associated with this study have been stored on ProteomeCommons.org Tranche network at https://proteomecommons.org/tranche. The java interface of Tranche can be initiated by clicking 'Launch Tranche!' on the homepage. Inside the java interface, raw data can be viewed by using the following workflow: View Contents By Hash/Input Tranche Hash

(Ulv/yTYTaaHin5Tv4InpsgoUY1uTJQtdoLRi9HbdtypXqztv+BiVE/wZieBkqu6d3kU20Vyejo0HYCfswgwiGyPHQPAAAAAAAAOhng==)/Browse Project. Individual MS files output from each proteomic platforms can be browsed and downloaded under '*Toxoplasma* Proteome Liverpool' project.

4.3.2 Data integration on ToxoDB

The proteomic data was first published on ToxoDB release 4.2 in June 2007. The peptide sequences identified were aligned with release 4 genes and alternative gene models as described in Section 4.2.2. In total, peptide identifications have been mapped to 2252 release 4 genes, and the 394 alternative gene models and ORFs identified have been mapped to 226 ORFs sequences. Peptide identifications can be viewed on individual gene report pages under the "Protein/Protein Features" section. The summary of the dataset can be viewed on ToxoDB through "Queries and Tools/Protein Expression/Mass Spec. Evidence". Proteomic data can be queried based on individual experimental platforms. Parameters can also be set with the minimum number of unique peptide sequences and /or spectra found that match a gene for it to be returned by the query. A related query "Identify ORFs based on Mass Spec. Evidence" is also created to view the peptide identifications that could not be mapped onto release 4 gene models.

The peptide identifications have also been mapped as individual tracks onto the interactive GBrowse on ToxoDB, which hosts approximately 50 GBrowse tracks including predicted gene models, EST alignments, and Affymetrix Probes, etc [112]. The peptide identifications can be viewed by selecting the option "Mass Spec Peptides (Wastling, *et al*.)" for peptides mapped onto release 4 genes and "Mass Spec Peptides (Alternative Models)" for peptides mapped onto alternative gene models and ORFs (see Figure 4.2).

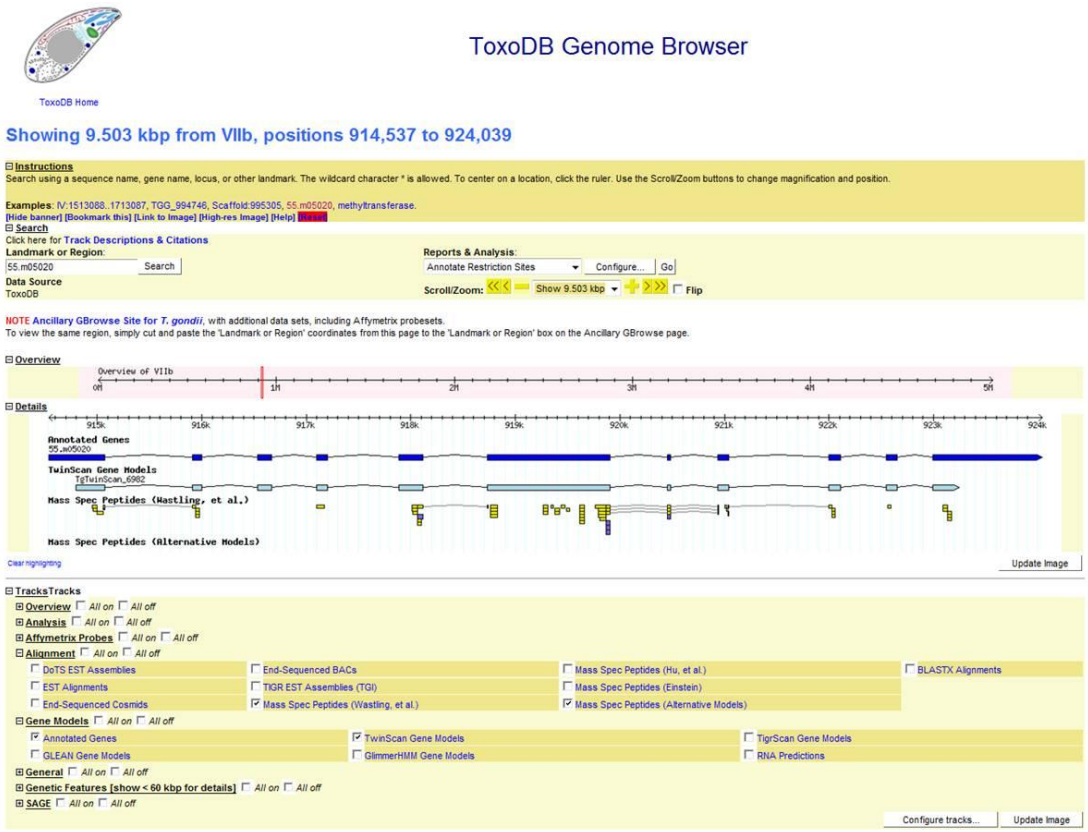


Figure 4.2 Visualization of peptide identifications on the ToxoDB Genome Browser. The peptide identifications are mapped onto ToxoDB GBrowse as individual tracks. Peptides that mapped onto release 4 genes or alternative gene models and ORFs can be viewed by selecting the option "Mass Spec Peptides (Wastling, *et al*.)" and "Mass Spec Peptides (Alternative Models)", respectively.

4.3.3 Examining the accuracy of the release 4 genome annotation

For the majority of the 2252 release 4 gene models identified, the peptide identifications have confirmed the correct ORFs and the positioning of start and stop codons. The current study has also identified 2477 intron spanning peptides in the official release 4 genome annotation, comprising 1110 unique release 4 genes. This has provided important supporting evidence that these exon-intron (splice) sites have been correctly predicted. Examples of peptides spanning splice site can be seen in Figure 4.2.

However, a significant number of peptides do not agree with the splice site predicted by the release 4 gene models. Peptide identifications have confirmed 421 splice sites that are only predicted by alternative gene models, which suggests either incorrect predictions for the release 4 gene models or possibly, alternative splicing events.

By using the GBrowse function, three types of discrepancies between the peptide identification data and release 4 gene models were discovered by this study. Firstly, peptide evidence can be used to indicate the expression of ORFs where release 4 genome annotations failed to predict coding sequence. Secondly, peptide evidence can also support an alternative frame shift or strand orientation to the release 4 gene model predictions. Thirdly, other discrepancies involve the positioning of the exon-intron boundaries as discussed above.

An example of a region of the genome scaffold where peptide evidence supports the presence of an expressed ORF but no release 4 gene is predicted is shown in Figure 4.3. Eleven peptides map to TgGlmHMM_3355 and TgTigrScan_5280, on the largest exon (ORF X-3-4725402-4726856) displayed in the figure. However, no annotated gene has been assigned to this region by the release 4 genome annotation. Additional peptides in this region map to the first exon of the neighbouring gene 46.m02877. Importantly, there are four intron spanning peptides identified that link the extra exon predicted by TgGlmHMM_3355 to the first exon of 46.m02877, indicating that release 4 gene 46.m02877 could have an incorrect start methionine and be missing an amino-terminal exon. Indicated by the peptide expression evidence, the most successful gene prediction model in this region would be TgGlmHMM_3355. Both TgGLEAN_5869 and TgTwinScan_2711 failed to predict the full length of the first exon (ORF X-3-4725402-4726856) that is supported by the

11 peptides. The second exon also reflected some prediction challenges, where neither TgGLEAN_5869 nor TgTwinScan_2711 hosted the full length of the exon supported by the peptide evidence and TgTigrScan_5280 completely failed to predict the existence of the first exon of gene 46.m02877.

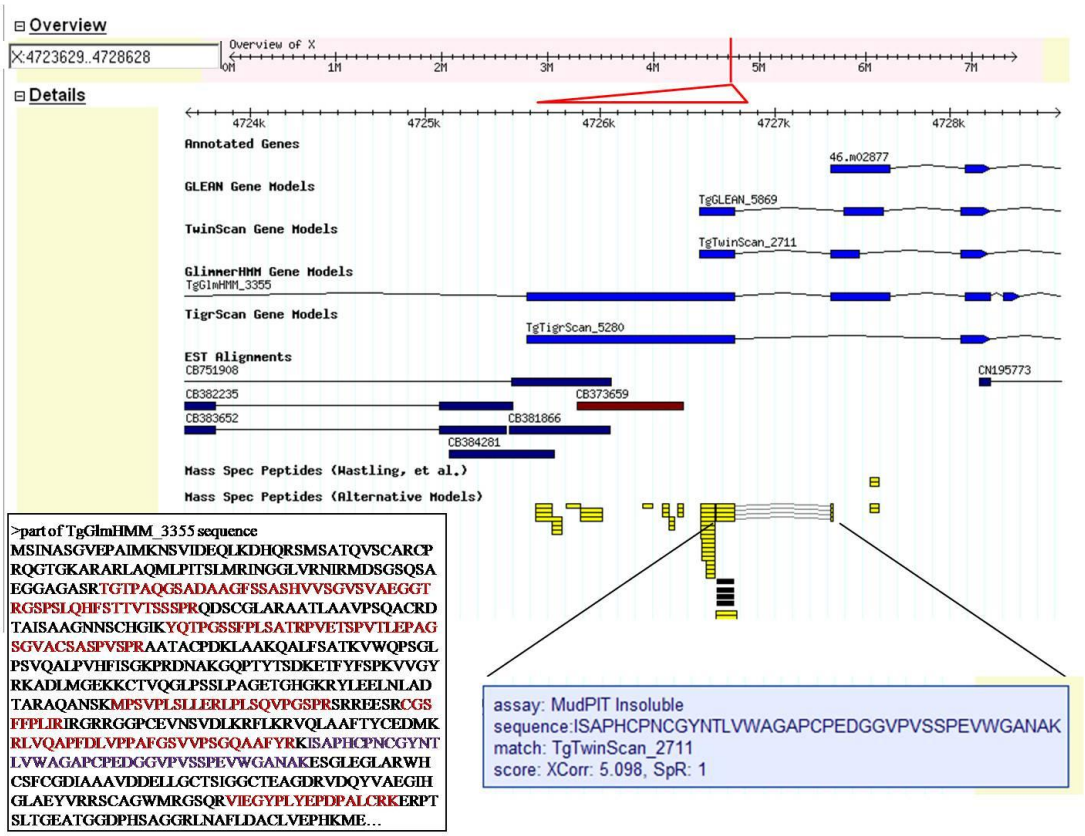


Figure 4.3 Peptide evidence indicating a missing amino-terminal exon predicted by release 4 genome annotation. The position of ORF X-3-4725402-4726856 in the genome scaffold is indicated by a red line on the overview track at the top of the figure and a detailed view is expanded below with the red triangle demarking the ORF length. Different gene annotation models are presented and predicted exons are indicated as blue boxes, linked by zigzag lines to indicate the position of exon/intron boundaries. Part of the predicted sequence for TgGlmHMM_3355 is shown as an insert; sequence for which there is identified peptide evidence is shown in red. The intron spanning peptide is shown in purple. Peptides aligning with this region are shown in yellow and the detailed MS information for one example is shown. EST evidence is shown as dark blue or brown boxes.

The second type of discrepancy is shown in Figure 4.4. Here peptide evidence is able to identify errors in the predicted reading frame or strand orientation. Twelve peptides derived from 35 individual spectra originating from 1-DE and MudPIT approaches provided matching hits to TgGlean_7850, TgTwinScan_4462 and TgGlmHMM_1717, although the various alternative gene models in this region differ in the length and number of exons. The release 4 genome annotation assigned gene 50.m05694 in this region but it is predicted to lie on the opposite strand and TgTigrScan_8273 uses a different reading frame. In this example, peptide expression data have provided supporting evidence for the correct reading frame and the large number of peptide hits to one region only indicates that the gene is likely to comprise a single exon.

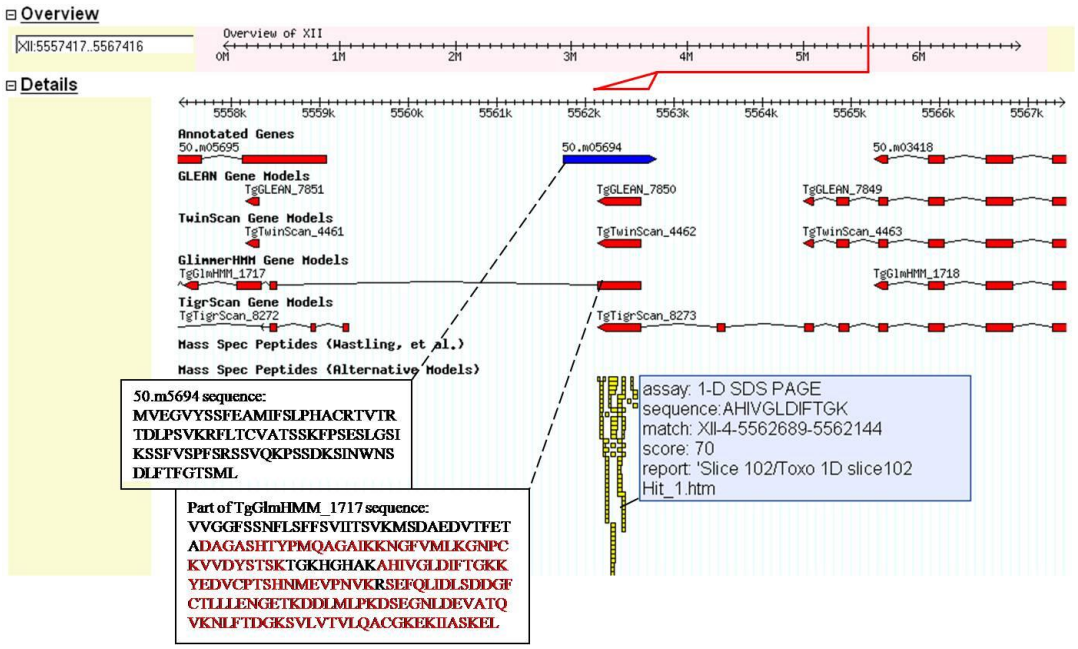


Figure 4.4 Peptide evidence indicate alternative frame shift.

The position of ORF XII-4-5562689-5562144 in the genome scaffold is indicated by a red line on the overview track at the top of the figure and a detailed view is expanded below with the red triangle demarking the ORF length. Different gene annotation models are presented and predicted exons are indicated as blue and red boxes, representing different strand orientation. Exons are linked by zigzag lines to

indicate the position of exon/intron boundaries. Predicted sequences for 50.m05694 and part of TgGlmHMM_1717 are shown as inserts. Sequence for which there is identified peptide evidence is shown in red. Peptides aligning with this region are shown in yellow and the detailed MS information for one example is shown.

An example of the third type of discrepancy is shown in Figure 4.5, where peptide evidence indicates alternative exon-intron boundaries to that predicted by the release 4 genome annotations. In Figure 4.5, 12 peptides identified using the MudPIT approach map to a region of the genome scaffold (X: 3917326-3920484) that is annotated with release 4 gene 28.m00300, comprising two exons. Five of twelve peptides match the second exon of gene 28.m00300. Of the remaining peptides, one maps to the predicted intron region of gene 28.m00300 and although it appears that six peptides match the scaffold in the region of the first exon of 28.m00300, these peptides actually relate to a different frame translation. Alternative gene models also vary considerably in this region in both the number and positioning of the exons, which indicates the difficulty in the prediction of splice sites in this region. All 12 peptides identified in this study only appear in TgGlmHMM_2666, which does not have an intron at this location, providing evidence that this model is most likely to be correct.

Figure 4.5 Peptide evidence indicating alternative exon positioning and splice site. The position of ORF X-1-3917326-3920484 in the genome scaffold is indicated by a red line on the overview track at the top of the figure and a detailed view is expanded below with the red triangle demarking the ORF length. Different gene annotation models are presented and predicted exons are indicated as blue boxes, linked by zigzag lines to indicate the position of exon/intron boundaries. Peptides aligning with this region are shown in yellow. The predicted sequence for ORF X-1-3917326-3920484 is shown as an insert and sequence that is identical to exon 2 of gene 28.m00300 is shown in blue. Sequence for which there is matching peptide evidence is shown in red. Purple lettering indicates the positioning of the 'intron-located' peptide, the detailed MS information for which is shown in the right hand insert.

4.4 Discussion

In this chapter, efforts have been made to present the proteomic data within a broader application. Raw MS data have been made publically accessible online and peptide identification data have been integrated with other genomic resources on ToxoDB. The integration of the proteomic data into ToxoDB revealed an important application of the data to genome annotation, demonstrating the incomplete status of current genome annotation.

By uploading the raw MS data onto the Tranche network, this valuable protein expression information is directly accessible to the research community. Moreover, storing the MS data in its raw format preserves important information about the biological sample and labour involved in the experiments, and enables future analysis to be carried out easily. For example, when the new standardized data format, i.e. mzML, becomes more established, the raw MS data generated from this study can be directly adapted into the new data processing pipelines. The Tranche facility also allows the raw MS data to be searched against the latest gene prediction models using improved search engines, the importance of which is further discussed in section 4.4.2.

4.4.1 Data integration

The integration of proteomic data with other genomic resources on ToxoDB was carried out on a peptide level. Since the genome sequence remains reasonably stable, matching the peptide identifications onto the corresponding genome scaffold avoids continual updating of peptide data mapping each time a new version of genome annotation is released. Peptide expression data can be directly mapped onto the new gene models according to their coordinates. However, as discussed in section 2.4.3,

multiple gene model databases have been used in the study to maximize protein identification. Although the benefit of including alternative gene models has been demonstrated in this chapter where in many cases peptide evidence supports the prediction made by alternative models, several technical issues in handling multiple gene models have been observed through the peptide mapping process.

Firstly, the script used in this study can efficiently collect peptide identifications from various gene models and ORFs expressed. However, due to the requirement of integration with other genomic resources that are already stored on ToxoDB, the peptide mapping algorithm designed was orientated towards release 4 genome annotation. Particular examples are step 2 and step 3 used in the mapping algorithm, where priorities have been given to release 4 gene models during mapping.

Step 2 states that if more than 50% of the peptides from an alternative model could be mapped to a release 4 gene, this was considered a valid mapping and the matching peptides were aligned with the corresponding release 4 gene. The remaining non-matching peptides were separately mounted on the scaffold, aligned with the alternative model. Step 3 states that if a certain set of peptides from an alternative model could be mapped to more than one release 4 gene, the gene that could host most peptides was reported. Again, the remaining peptides were separately mounted on the scaffold, aligned with the alternative model. As shown in this Chapter, peptides identified in the neighbouring region of a release 4 gene model are of great interest to gene expression research as well as genome annotation. They provide strong evidence of alternative splice sites; missing exons and different positioning of start or stop codons, which could alter the function of the gene when expressed. However, due to the peptide mapping algorithm that was orientated towards release 4 genome annotation, these important peptides have been mapped separately to

alternative gene models. Unfortunately the current version of the ToxoDB interface does not allow the user to query thesis alternative gene models. In other words, where once readily available, these peptides are no longer searchable.

Additionally, steps 4, 5, and 6 state that peptides identified from an alternative model can be mapped to an ORF or alternative gene model only if 100% of the peptides can be mapped. The stringent threshold set here was due to the consideration that previous versions of EST and ORF databases used in MS data searching contain small sequencing errors that are not consistent with the release 4 gene models and ORFs. In total, there were 220 TgEST sequences and 184 ORF sequences identified in this study that cannot be mapped onto ToxoDB due to this reason, representing 4.9% of the total number of sequences mapped. This reflected the existence of sequencing errors in previous versions of EST and ORF databases. However, this also resulted in the loss of genuine peptide identifications that mapped to the correct part of the EST or ORF sequence. In addition to this, it was not possible to map peptides identified from 163 alternative gene models in this study (74 of TgGlmHMM, 58 of TgTwinScan and 31 of TgTigrScan) to the release 4 gene models and ORFs. While these peptides are presented on ToxoDB, it is not possible to query MS evidence for older gene models such as TgTwinScan, which means that this subset of peptide data is effectively "lost" to the wider research community. Solutions to this problem are discussed in the next section.

Despite the limitations of the peptide mapping approach due to the multiple gene models and ORF database used, the integration of the proteomic data generated in this study onto ToxoDB has already assisted the genome annotation process by confirming the correct predictions of 2477 intron spanning peptides in the official release 4 genome annotations. More importantly, the discrepancies between the

proteomic data and the release 4 gene models also demonstrate the incompleteness of the release 4 genome annotation. The peptide identification data provided evidence for the expression of 394 alternative models, which were mapped to 226 ORFs, as well as 421 splice sites that have not been predicted by the release 4 genome annotation. In fact, even in the latest release 5.2 genome annotation which was recently published in July 2009, there is strong peptide evidence for the expression of alternative models that were mapped to 203 ORFs. The reduced number of ORFs that have peptide evidence from 226 to 203 reflects the improvement of release 5.2 genome annotation where peptides previously mapped to 23 ORFs are now able to be mapped to release 5.2 genes. However, peptides mapped to those 203 ORFs still possess valuable information for the improvement of release 5.2 genome annotations. Moreover, if the "lost" peptide identifications during the mapping process were to be mapped onto ToxoDB, an even larger discrepancy between expression data and predicted gene models would be evident. This work has highlighted the importance of proteogenomic research which directly incorporates proteomic data into the genome annotation process. It has also highlighted the on-going problem for proteomic analysis of the need to re-submit raw MS data against the latest genome annotations, in order to obtain the highest quality dataset. This is a time-consuming, manual task but which is of significant importance, if one is to avoid the situation of "lost", out-dated and inaccessible annotations.

4.4.2 The application of proteogenomics in *T. gondii* genome annotation

As discussed in section 1.2.6.1, proteomic data can be used in various aspects of the genome annotation process such as validating predicted gene models and detecting novel genes as well as validating alternative splicing variants [213, 215, 218, 220, 221, 344]. In this study, three examples have been shown to demonstrate the

potential usage of proteomic data in indicating missing exons, alternative frame shifts, as well as alternative exon positioning.

The proteomic data acquisition in this study can be used to assist the development of new genome annotation pipelines. Firstly, protein expression data can be used as a valuable training set to improve the prediction of integrative gene prediction programs such as GLEAN [282] and TwinScan [280]. This application is particularly important for microbial genome annotation such as *T. gondii*, where few homologies have been characterized in comparison to the human genome. By analysing the composition and statistical properties of the expressed peptides, programs can be tuned to predict novel genes which homologies have not been previously identified in other organisms.

Secondly, by searching the raw MS data against the latest update of genome annotation in the pipeline, peptide expression data can be directly used to validate the accuracy of the predictions. This information can then be fed back to the automated pipeline and generate an improved version of genome annotation, which can be validated by the raw MS data again. By performing this cycle several times, the accuracy of genome annotation can be rapidly improved. This automated pipeline will also significantly speed up the current proteomic research workflow, where successive upgrades of genome annotation require the raw data to be re-submitted in a slow manual fashion at the moment, as highlighted in the previous section. The pipeline will also resolve the peptide mapping problem on ToxoDB, for example, the raw MS data used to identify those 220 TgEST sequences and 184 ORF sequences that were not able to be mapped on ToxoDB would be preserved and searched against the new annotations. Likewise, the peptide identifications from those 163

alternative gene models that are no longer available for querying on ToxoDB could also be directly entered into the automatic pipeline.

Of course, in order to efficiently initiate the cycle, large scale sampling of peptide identifications from the genome is required. This will enable the maximum number of peptide features to be picked up by gene prediction programs and preserved in the subsequent annotation cycles. Currently, the best approach to achieve the biggest coverage of peptide identifications in a genome is to search the MS data against all the ORFs with a length greater than 50 amino acids from the whole genome sequence database.

Theoretically, the collection of all the ORFs covers the entire potential protein coding sequence (CDS) in the genome. However, the current program setup for ORF marking in the genome only processes the same region of sequence once and identifies a specific ORF as starting from the first start codon it encounters until it comes across a stop codon [112, 345]. This approach is particularly efficient in marking ORFs in organisms with no introns, such as prokaryotes. However, it has several limitations in covering the entire potential coding sequence in a typical eukaryotic gene which contains multiple exons. Firstly, the algorithm marks every ORF from the first start codon to the first stop codon, no matter how many other start codons are within the ORF. This prevents the identification of the second exon starting within the length of the original ORF. Secondly, in a gene that contains multiple exons, a frame shift between different exons means the coding sequence cannot be hosted within a single ORF. This particularly prevents the identification of intron-spanning peptides in the proteomic research which also contains critical information for genome annotation.

In fact, the 163 alternative gene models which could not be mapped onto release 4 gene models and ORFs partly reflected the second limitation of the current ORF marking algorithm, where no single ORF could host 100% of the peptides identified to an alternative gene model. A new algorithm approach to design ORF databases for MS data searching is under development in the Wastling group in collaboration with Dr. Andy Jones, University of Liverpool.

The new ORF database cannot be a simple collection of all the direct translations of genomic sequences from all the start codons which exist, as this would result in a giant sequence database which would require tremendous computing power for MS data searching and which would increase the false discovery rate. One possibility is to harness the latest development in the transcriptome, RNA-Seq. By using the high-throughput sequencing approach offered by RNA-Seq, a genome-scale transcription map can be rapidly achieved [229]. The information can then be used as a reference map for the selection of gene coding ORFs and subsequently reduce the size of the ORF database for MS data searching. An on-going collaborative project between the Wastling group and Dr. Arnab Pain, at the Wellcome Trust Sanger Institute, Cambridge is developing a method for the integration of proteomic and transcriptome data in the genome annotation process for *T. gondii* and *Neospora caninum*.

4.4.3 Conclusion

In this chapter, the proteomic data acquired in this study have been placed in a broader platform. The raw MS data have been stored in the publically accessible Tranche network. The expression data have been integrated on a peptide level with other genomic resources on ToxoDB. Inspired by the issues raised during peptide mapping and the examination of the accuracy of the release 4 genome annotations

using these peptide identifications, the incompleteness of the release 4 genome annotation was highlighted and the potential application of a new genome annotation pipeline was discussed.

In a conventional bottom-up protein identification based proteomic project, peptide identification relies on the predicted gene models. Successive upgrades of genome annotation mean that the proteomic researcher may have to re-submit the data which is a time consuming process. By using the new genome annotation pipeline discussed in this chapter, the manual re-submission can be carried out automatically and proteomic expression data can be directly used to improve genome annotation. Together with the information contained within the transcriptome, a near "perfect" genome annotation can be expected in the near future.

In addition to the application of proteomic data in the field of proteogenomics, the integration of proteomic data into ToxoDB also allows an important comparison to be made, that of the transcriptomic data. This comparison will reveal implications of important biological processes such as protein degradation and post-transcriptional regulation. This interesting subject is investigated and discussed in chapter 5.

Chapter 5

A comparison of the proteome and transcriptome of *Toxoplasma* and other *Apicomplexa* parasites

5.1 Introduction

In the previous chapter, the methods for the integration of proteomic data onto ToxoDB were investigated and the potential for directly using proteomic data in genome annotation was discussed. The integration with another genomic resource in the context of ToxoDB has led to another important application of proteomic data, that of the comparison of the proteome and transcriptome of *Toxoplasma*.

As discussed in section 1.2.6.2, transcriptomic studies were used to infer putative functions of proteins under the commonly accepted 'guilt-by-association' hypothesis [233]. However, with the increasing number of proteomic studies, data from several recent studies have suggested a relatively weak correlation between mRNA expression and protein expression in plant seeds [235], mouse embryonic stem cells [236], yeast [237], and even in Apicomplexan parasite *Plasmodium* [238, 239]. The discrepancies observed not only weaken the application of protein function assignments using transcriptomic data, but also highlight the requirements for more research which would allow a better understanding of basic biological processes. With the first example of a whole cell lysate proteome for *Toxoplasma* acquired in this study, the relationship between proteins and their mRNA can be examined in *T. gondii*.

Several large scale transcriptomic analyses have been carried out on *T. gondii* and the data have been integrated on ToxoDB release 4 under the same genome scaffold and identifiers as used for proteomic data [112]. These large scale transcriptomic studies include transcript expression with microarray evidence, expressed sequence tag (EST) evidence and serial analysis of gene expression (SAGE) tag evidence. The microarray expression profiling of the three archetypal *T. gondii* lineages has been

carried out by Dr. PH Davis and Prof. DS Roos, University of Pennsylvania, USA and the data have been released on ToxoDB pre publication. The EST evidence is mirrored from the dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) site onto ToxoDB which contains various libraries of different strains and life stages. SAGE tag libraries were generated from *T. gondii* parasites at specific stages of development, representing key developmental transitions in primary parasite populations and in three laboratory strains [272]. The large amount of *T. gondii* transcriptomic data and the integration of these data and proteomic data on ToxoDB have made a systematic comparison between proteome and transcriptome of this important parasite feasible.

In addition to the extensive coverage of *T. gondii*, a large amount of proteomic and transcriptomic data have also been acquired for other important Apicomplexan parasites, which enables a cross species comparison that would highlight common features between proteome and transcriptome in *Apicomplexa*. As reviewed in chapter 2, large scale proteomic studies have been carried out on several species of *Plasmodium* [238, 239, 261-263] and *Cryptosporidium parvum* sporozoites [264, 265]. On-going whole cell proteomic studies are also being carried out in *Neospora*, *Theileria* and *Eimeria* by the Wastling group, University of Liverpool, UK. Large scale transcriptomic expression profiling projects have also been carried out. dbEST [346] and EuPathDB [347] host the largest collection of EST data for the *Apicomplexa*. SAGE projects have been carried out for *P. falciparum* [348-350] and microarray expression data are available for *P. falciparum* and *P. berghei* [238, 351-353].

Most of the proteomic and transcriptomic data here have been uploaded to corresponding component sites of EuPathDB, such as CryptoDB, PlasmoDB and ToxoDB [347]. EuPathDB provides a large collection of published data and hosts

valuable data from several unpublished studies that further expand the coverage of expression data for *Apicomplexa*. Moreover, similar to the process discussed in Chapter 4, great efforts have been made to integrate expression data generated from different experimental platforms together under the same identifier system.

In this chapter, equipped with the large amount of expression data and valuable support from EuPathDB, a systematic comparison of the proteome and transcriptome of *Toxoplasma* and other *Apicomplexa* parasites has been carried out. Explanations for observed discrepancies and the potential applications of the interaction of proteomic and transcriptomic data were also discussed.

5.2 Materials and methods

To facilitate a genome wide comparison between proteome and transcriptome, relative expression data were collected primarily from component sites of EuPathDB (formerly known as ApiDB) [347].

5.2.1 Proteomic and transcriptomic data collection for *T. gondii*

The 2252 release 4 genes identified in this study were used as the *T. gondii* tachyzoite proteomic dataset for the comparison. The collection of EST, SAGE and microarray data was carried out using ToxoDB (release 4.2) [112]. To identify the genes that have EST evidence, all tachyzoite EST libraries were selected in the query "ToxoDB Queries and Tools/Transcript Expression/Identify Genes based on EST evidence", with default settings applied. The SAGE expression data were collected using the query "ToxoDB Queries and Tools/Transcript Expression/Identify Genes based on SAGE Tag evidence" with "*T. gondii* 3p SAGE tag frequencies_d6" and "*T. gondii* 3p SAGE tag frequencies_rh" libraries selected and default parameters were applied. The microarray expression percentile for *T. gondii* genes were collected using the query "ToxoDB Queries and Tools/Transcript Expression/ Identify Genes by Microarray Evidence/ Identify Genes based on Expression Percentile (T.g.)". Data comparisons were carried out either on ToxoDB using the "My Query History" function or in Microsoft Office Excel with data downloaded from ToxoDB.

5.2.2 Proteomic and transcriptomic data collection for other Apicomplexan parasites

In order to identify *Apicomplexa* genes which exhibit discrepancies between transcriptomic data and proteomic data, a similar data collection process was carried out on *C. parvum* sporozoites using CryptoDB [330], and *P. falciparum* (all life

stages) using PlasmoDB [329]. All the published gene identifications from major proteome projects listed in CryptoDB (release 3.7) and PlasmoDB (release 5.4) were included in the analysis. Comparative EST libraries were used to collect EST evidence for *C. parvum* and *P. falciparum* and microarray expression data were collected for *P. falciparum*. Data comparisons were carried out either on CryptoDB and PlasmoDB using the "My Query History" function or in Microsoft Office Excel with data downloaded from CryptoDB and PlasmoDB.

For *N. caninum* tachyzoites, preliminary proteome profiling results from a MudPIT experiment generated by the Wastling group was used for the comparison (unpublished). Briefly, *N. caninum* tachyzoites were collected from cell culture by Rebecca Norton. Sample preparation, mass spectrometric analysis and MS data searching of MudPIT analysis were performed by Dr. Judith H. Prieto from John R. Yates's lab, Scripps Research Institute, La Jolla. Proteomic expression data were obtained for 660 of the gene models in the latest set of gene predictions (available via GeneDB at http://www.genedb.org/) at the time of comparison (July, 2008). EST evidence for *N. caninum* was collected from NCBI dbEST (http://www.ncbi.nlm.nih.gov/dbEST/). The alignment of EST evidence to genes that have proteomic evidence was carried out by Amandeep Sohal in Dr. Arnab Pain's group at the Wellcome Trust Sanger Institute, Cambridge. Briefly, EST evidence was mapped onto the genome using the Exonerate software [354] with the following parameters: "exonerate_farm --bestn 1 --showtargetgff yes --showvulgar no --showalignment no --query *Neospora*ESTsequences.fasta --target sequence.dna". Artemis software [355] was then used to highlight the intersections of gene models and EST evidence.

5.2.3 Method for data comparison between three Apicomplexan parasites

The comparison between proteomic and transcriptomic data was carried out at the individual parasite level using their own identifier systems. To determine the number of genes in the intersections between each parasites, OrthoMCL software [169] was used and the analysis was carried out by Amandeep Sohal at the Wellcome Trust Sanger Institute, Cambridge. Briefly, 5460 *P. falciparum* genes, 7793 *T. gondii* genes and 5589 *N. caninum* genes were used in the analysis and 5147 orthologue groups were formed.

5.3 Results

5.3.1 Comparison of gene expression at the transcriptional and translational level in *T. gondii*

A comparison of transcriptomic and proteomic data of *T. gondii* was done in two stages. Firstly, the proteomic data were compared with the entire microarray expression data. Since the microarray data have the largest coverage of the genome in all the current gene expression experiments, this comparison provides a good indication of the level of proteomic sampling. Secondly, the proteomic data were compared with various transcriptomic experiments such as microarray expression data (over 25 expression percentile), SAGE expression data and EST data. The SAGE data provide a more accurate measure of transcripts than microarray data and while providing direct evidence of gene expression, the application of EST data in genome annotation [112] has made it an important transcriptomic dataset. By comparing the proteomic data with all the three large scale transcriptomic datasets available, discrepancies between the transcriptional and translational levels of gene expression can be examined.

5.3.1.1 Comparison with entire microarray expression data

Microarray data provide extensive coverage of the genome; 7764 release 4 genes (99.5% of the genome) were assayed. Four release 4 genes (25.m01905, 8.m00178, 80.m05040 and 83.m00013) which were identified by the proteomic data were not assayed by microarray and for this reason were not included in this comparison. *T. gondii* tachyzoites were assayed and all 7764 genes on the array exhibit some signal. It is difficult to determine the correct signal: noise ratio above which mRNA levels can be considered to be indicative of a gene being switched on since not all genes would be expected to be expressed in the tachyzoite stage.

For the purposes of this comparison, in this study 7764 release 4 genes were divided into quartiles according to the mRNA expression levels determined by microarray. Those genes in the bottom 25% were considered as having zero detectable mRNA above the baseline while those genes in the top 75-100% were the genes that have the highest mRNA expression levels. Figure 5.1illustrates the number of genes in each microarray expression percentile which have been identified by this proteomic study.

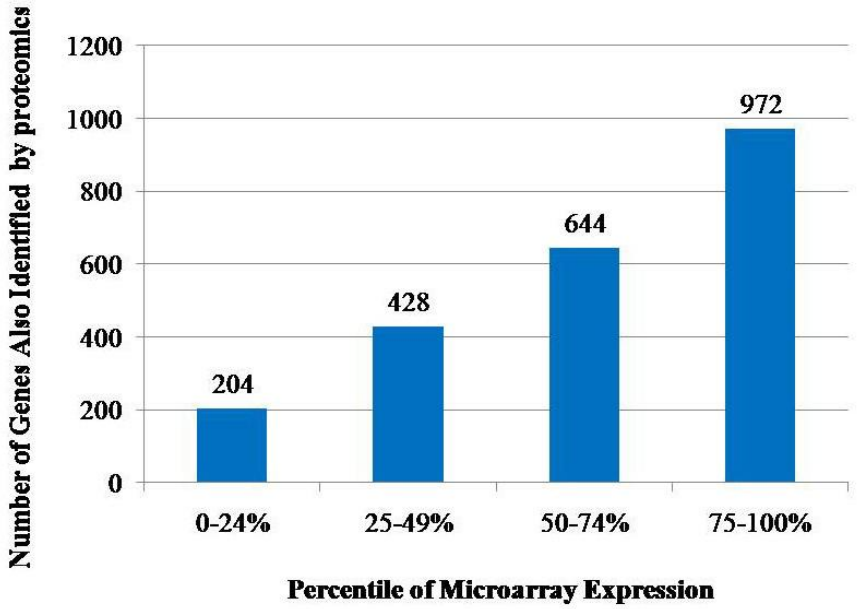


Figure 5.1 Comparison of proteomic data with microarray expression data. Release 4 genes assayed by microarray were divided into quartiles according to mRNA expression levels. The bar chart shows the number of genes also identified by proteomics for each of the four percentile ranges, 0-24%, 25-49%, 50-74% and 75-100%.

While the number of genes contained in each of the quartile ranges is the same (i.e., 1941), there is a general trend for more proteins to have been detected for genes with higher mRNA expression levels (972 genes (50.1%) in the top mRNA expression range 75-100% and 644 genes (33.2%) in the second highest mRNA expression range compared with 22.1% and 10.5% in the lower two percentiles). This may

reflect a correlation between mRNA abundance and protein abundance, and that higher abundance proteins are easier to detect by proteomic approaches.

Proteomic data also provided important expression evidence for 204 genes from the bottom 25% microarray expression range. This highlighted the discrepancies between proteomic expression and microarray expression data where genes in the bottom 25% microarray expression range were described as zero detectable mRNA above baseline. The discrepancies between protein level of expression and transcriptional level of expression were further demonstrated when EST data were included in the comparison.

5.3.2 Comparison of proteomic data with microarray expression data (over 25 expression percentile), SAGE expression data and EST data

The comparison of proteomic data with microarray expression data demonstrated the sensitivity of the proteomic approach and a general correlation between the number of proteins detected and mRNA expression levels. Whilst it is hard to decide a definitive microarray expression percentile for a gene to be considered expressed, 25% was used as a cut-off assuming 75% of the release 4 genes are expressed in the tachyzoite stage. Other important evidences of transcript expression provided on ToxoDB are EST and SAGE. In this study, cDNA evidence from all the tachyzoite EST libraries was collected, which represents a total coverage of 68.4% of all release 4 genes. Expression evidence from laboratory strain RH library and primary VEG strain Day-6 library, which has the closest correlation with RH strain [272], were collected, representing the equivalent sampling of this study. The 1453 genes detected by these two SAGE libraries represent a total coverage of 18.6% of all release 4 genes. The comparisons between 2252 genes identified by proteomic data

and genes identified by various transcriptomic experiments are shown in Figure 5.2 (Detailed gene identifications are listed in Appendix VI).

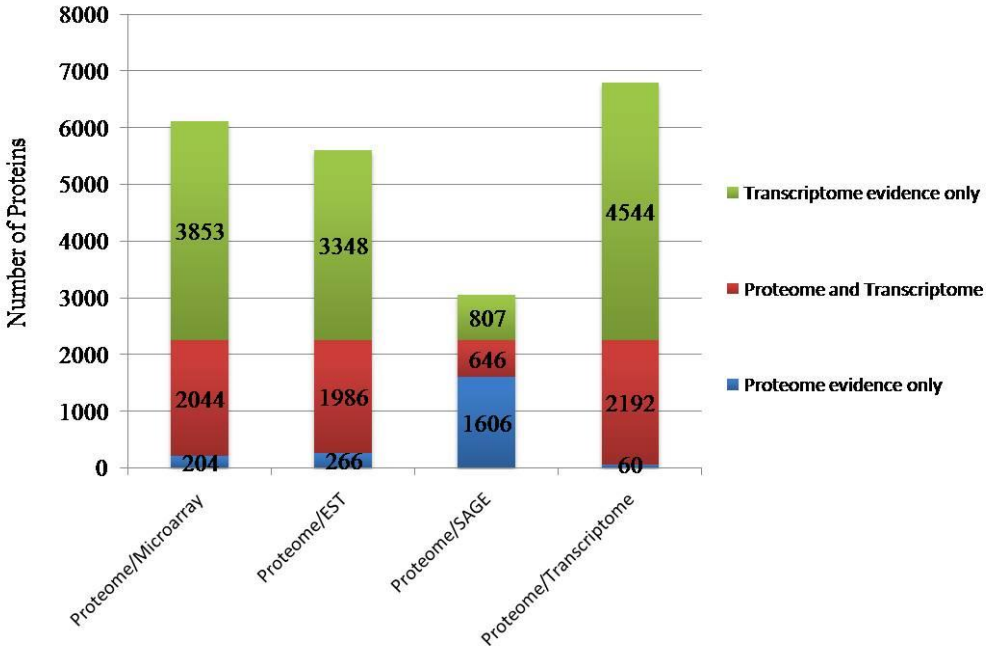


Figure 5.2 Genes with proteome and transcriptome evidence in *T. gondii*.
The relationships between proteomics and various transcriptomic techniques such as Microarray, EST, SAGE and total transcriptomic expression data in *T. gondii* are shown in separated columns. Genes identified by this proteomic study are compared with genes that have transcriptome evidence. The blue portion indicates proteins without transcriptome evidence, the red portion indicates proteins that have both proteome and transcriptome evidence, and the green portion indicates genes without proteome evidence.

When considering all four platforms included in the comparison, 572 genes have expression evidence from all platforms which may reflect their strong expression at both transcriptional and translational levels. Proteomic data has provided protein expression evidence for 204 genes in the bottom 25% microarray expression range, 266 genes that have no EST evidence and 1606 genes that have no SAGE evidence. Furthermore, 60 tachyzoite genes are exclusively identified by proteomic data and have no corresponding transcript expression evidence.

The identification of these 60 genes is of great research interest. This illustrates that some genes with zero or very low mRNA expression can still be identified by proteomic approaches, which may reflect the high sensitivity of the proteomic approach. One clear biological explanation is that these proteins were extant before the tachyzoite differentiated. Or these could be highly stable proteins, with low turnover rates which require low levels of transcription to be maintained in the cell. Another explanation is that substantial quantities of protein can be produced from very low abundance mRNA. In fact, many of those 60 genes were identified by large numbers of peptide identifications which usually indicates high protein abundance. As such these 60 genes represent interesting candidates for understanding the relationship between mRNA and protein abundance levels in *Toxoplasma*. Three examples from this group are 'tubulin beta chain, putative' (28.m00301, 128 peptide hits), 'thioredoxin, putative' (42.m03331, 57 peptide hits) and 'coatomer protein gamma 2-subunit, putative' (59.m00090, 53 peptide hits).

To further understand the properties of genes that only have proteomic data, a proteomic data based comparison was carried out with transcriptomic data across four species of *Apicomplexa*. As shown in Figure 5.2, the SAGE data provide a much smaller overlap with the proteomic data than that of EST and microarray data. In fact, when the SAGE data were excluded from the comparison, 1850 genes were shared among proteomic, EST and microarray experiments which are considerably more than 572 genes when all four platforms are included. The smaller overlap observed between SAGE and proteomics is very interesting, and may represent genuine biological discrepancies or differences in the experimental sampling and detection sensitivities (further discussed in section 5.4.2). However, including the SAGE data to the proteome and transcriptome comparison across four species of

Apicomplexa will add more complexity to the discrepancies while not adding much more value to the whole proteome and transcriptome comparison. Only 77 genes identified by SAGE were not detected by EST and microarray in *T. gondii* and SAGE data are not readily available across other *Apicomplexa* except *Plasmodium.* Due to the above considerations, the SAGE data were not included in the comparison across *Apicomplexa*.

5.3.3 Proteome and transcriptome comparisons across four species of *Apicomplexa*

Given the large amount of good quality transcriptional and translational data across the *Apicomplexa*, a general comparison was carried out to examine the relationship between proteins and their mRNA. The proteomic data based comparison allows us to identify the subsets of proteins for which no transcriptional evidence was acquired. Comparative EST libraries and microarray expression data (no microarray data were available for *Neospora* or *Cryptosporidium*) were compared with their respective proteomic datasets for four species of *Apicomplexa* including *T. gondii* tachyzoites, *C. parvum* sporozoites, *P. falciparum* (all life stages) and *N. caninum* tachyzoites (see Figure 5.3).

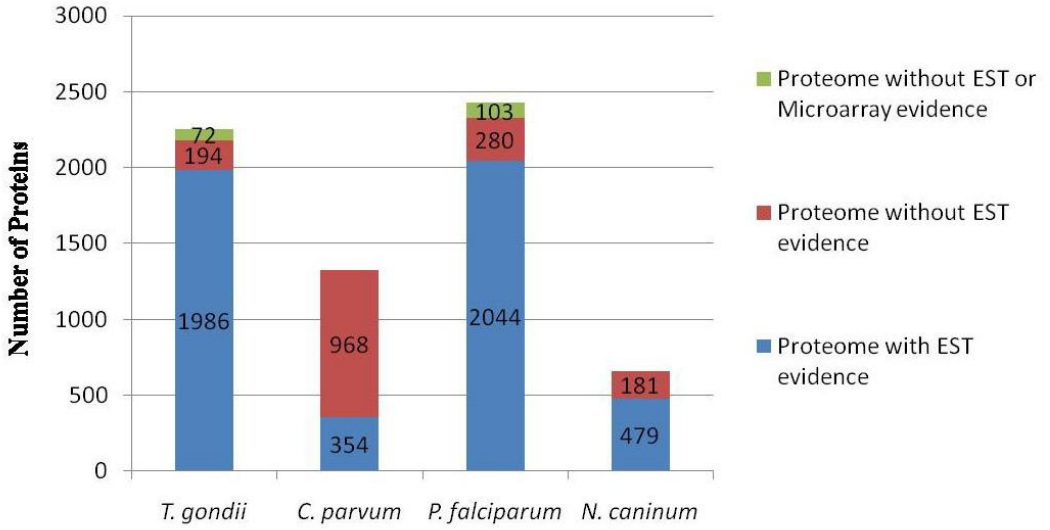


Figure 5.3 Proteome and transcriptome comparisons across four species of *Apicomplexa*. The numbers of proteins identified by peptide evidence in *Toxoplasma gondii* tachyzoites, *Cryptosporidium parvum* sporozoites, *Plasmodium falciparum* (all life-stages) and *Neospora caninum* tachyzoites are shown. The red portion indicates proteins without EST evidence and the green portion indicates genes without EST or microarray evidence (less than 25 expression percentile). No microarray data were available for *Neospora* or *Cryptosporidium* at the time of comparison.

Each column represents the total number of proteins identified by proteomics, with the red portion indicating proteins without any EST evidence and the green portion showing proteins without either EST or microarray data. These data show that except for *C. parvum*, a large number of genes identified by proteomics were also identified at the transcriptional level. Importantly, in addition to those 72 genes identified in *T. gondii*, 103 of *P. falciparum* genes detected by proteomics have neither EST nor microarray evidence over and above the 25% cut-off. There were also proteomic evidence for 968 of *C. parvum* genes and 181 of *N. caninum* genes where no EST evidence existed. The large percentage of *C. parvum* genes (73%) which have proteomic evidence but no EST evidence is likely to represent a relatively poor EST coverage for *C. parvum*. Given this common discrepancy observed between

proteomic data and transcriptomic data across four *Apicomplexa* species, further analyses were carried out to test if these proteins that have no, or relatively low levels of transcription were related at the level of orthology.

5.3.4 *Apicomplexa* genes which exhibit discrepancies between transcriptomic data and proteomic data

Proteomic and transcriptomic data for *P. falciparum*, *T. gondii* and *N. caninum* were collected as described in section 5.2.2. *C. parvum* data were not included because the relatively poor EST coverage is likely to bias the comparison. Genes were sorted into the following three categories (a) transcript present but no protein detected, (b) protein detected but no EST evidence and no transcript detected by microarray at ≥ 25% threshold, and (c) protein detected but no EST evidence. In each category, the numbers of genes which are shared between each species were determined by orthologue mapping using OrthoMCL [169] and the results of the comparison are shown in Figure 5.4 (Detailed gene identifications are listed in Appendix VII).

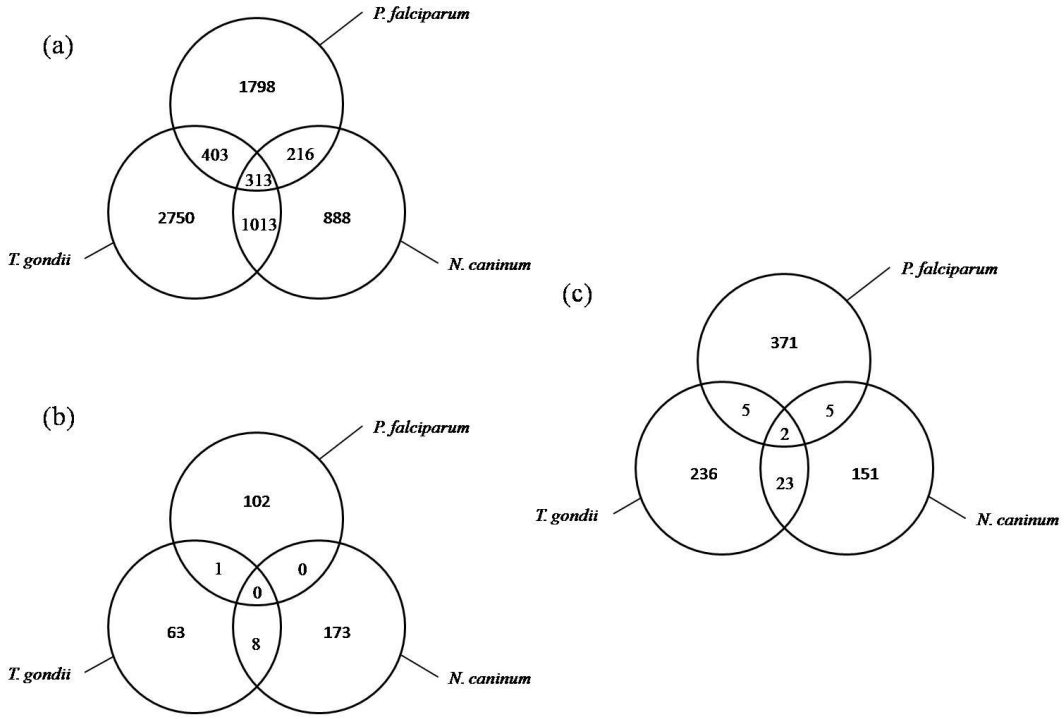


Figure 5.4 Genes from three *Apicomplexa* which exhibit discrepancies between transcriptional data and proteomic data. Each circle represents the number of genes for which a discrepancy was seen between transcriptomic and proteomic data for *Plasmodium falciparum*, *Toxoplasma gondii* and *Neospora caninum* based on: (a) transcript present but no protein detected, (b) protein detected but no EST evidence *and* no transcript detected by microarray ≥ 25% threshold, (c) protein detected but no EST evidence. The numbers of intersections were determined by orthologue mapping using OrthoMCL.

Figure 5.4a shows that of the genes which have transcriptomic evidence but lack proteomic evidence, significant numbers (34% in *Plasmodium*, 39% in *Toxoplasma* and 63% in *Neospora*) had orthologues in other species, and 313 genes were common between all three species although no specific class of genes can be highlighted. This is perhaps a less surprising result, since low levels of protein synthesis or high rates of protein turnover and degradation may contribute to the under-representation of proteomic coverage and the ≥ 25% microarray expression criteria used in this study is rather loose and likely to include transcripts that have not been expressed. Figure 5.4b shows that there are 356 genes across all three

species which have proteomic data but no transcriptomic evidence. However, only several genes are shared between any two species as orthologues. Figure 5.4c shows the comparison only between proteomic and EST data. A larger number of proteins are shared, including two orthologues seen across all three species ('small nuclear ribonucleoprotein, putative', 57.m01848 and 'thioredoxin, putative', 42.m03331; ToxoDB annotation). These data revealed that there is little overlap across the three datasets for orthologous genes for which there is proteomic evidence, but little or no transcriptomic expression evidence (detected by ESTs or microarrays).

5.4 Discussion

Following the effort of data integration discussed in chapter 4, in this chapter, the proteomic data generated from this study have been compared with various transcriptomic expression data using ToxoDB. A broader comparison of proteomic data and transcriptomic data across four species of *Apicomplexa* was also made available with the data and tools provided by EuPathDB. This has further highlighted the importance of data integration in the post-genomic era where large scale genome wide studies can be viewed and analysed together in a way that was not readily accessible before. The study of gene expression has been dominated by transcriptomic approaches where the level of mRNA expression is measured. Now, supplemented with the latest addition of proteomic data, a dynamic view of gene expression pattern can be analysed for the first time on *T. gondii* and other Apicomplexan parasites through the comparison of proteomic data and transcriptomic data.

5.4.1 A weak correlation observed between proteome and transcriptome

In *T. gondii*, the comparison of proteomic data with microarray data has revealed a weak correlation between mRNA abundance and protein abundance, where more genes in the higher microarray expression percentile range have been detected by proteomic approaches than those genes in the lower percentile range. In fact, a quick analysis of the number of EST counts and the number of peptide counts for genes that were identified by this proteomic study has revealed a spearman's rank correlation value of 0.29, which also indicated a weak correlation between mRNA abundance and protein abundance.

Of course, caution needs to be made when using both peptide counts and the number of proteins identified as direct indications of protein abundance. Firstly, variations introduced in the protein separation and digestion steps would affect the distribution of the peptide mixture and, the kinetics of peptides varies during MS analysis. Actually, it has been noted that only a few so-called 'proteotypic' peptides are repeatedly and consistently identified for any given protein present in a mixture in relation to their different physicochemical properties [356]. Secondly, the length of the peptide identifications and the size of the protein are also a source of variations. Larger proteins would generally produce more peptides than proteins of a smaller size and depending upon the fragmentation, two short peptides identified from a protein does not guarantee a two times abundance over a protein identified by a single long peptide. Considering the above limitations, peptide counts are only able to provide a crude indication of the protein abundance. Quantitative approaches can be used for future investigation such as MS-based isotope labelling or lable free methods, as well as gel-based technique differential gel electrophoresis (DIGE), which is further investigated in chapter 6. Nonetheless, proteomic identifications can provide important evidence for the translational level of gene expression. The large number of genes identified by both proteomic and transcriptomic techniques and the general trend of a correlation between them reflected the comparable sensitivity of both techniques.

5.4.2 The significance of discrepancies between proteome and transcriptome

The correlation between proteomic data and transcriptomic data increased the confidence of the feasibility of the comparison between the two techniques and follows the central dogma of Gene-Transcription-Translation. However, perhaps

more significant finding was the discrepancies between proteomic data and transcriptomic data.

The presence of proteomic evidence in the absence of detectable transcriptomic evidence has been noted in several other studies [235-237], as well as in *Plasmodium*, *Cryptosporidium* and *Neospora* as shown in this chapter. The detection of these genes by proteomic studies reflected the high sensitivity of the proteomic approaches. In the case of *T. gondii*, by combining expression evidence from microarray, EST and SAGE, there were 6736 release 4 genes which have transcriptomic evidence, representing 86.4% of the entire genome. It is unlikely that all those 6736 genes are expressed in the tachyzoite stage, especially with the inclusion of 75% of all the genes assayed on microarray; it is likely to include genes expressed in other life stages. Considering the vast coverage of transcriptomic data, the identification of the 60 *T. gondii* genes for which no transcripts were observed is more fascinating.

One possibility of the cause of discrepancies between proteomic data and transcriptomic data is technical limitations, whereby the same level of analytic resolution is hard to reach between the two techniques. Other possibilities involve biological explanations such as selective protein degradation and variations in protein turn-over rates [240, 241] as well as post-translational regulations such as mRNA decay and translational repression [242-244].

Three examples of those 60 genes that have exclusive proteomic evidence indicated in section 5.3.2 were ‘tubulin beta chain, putative’ (28.m00301, 128 peptide hits), ‘thioredoxin, putative’ (42.m03331, 57 peptide hits) and ‘coatomer protein gamma 2-subunit, putative’ (59.m00090, 53 peptide hits). These three genes were identified by large numbers of peptide identifications in the absence of transcript evidence in

tachyzoite stage. Interestingly, although no evidence have been found from Day 6 and RH libraries of SAGE experiment, both the tubulin beta chain, a component of microtubule that is critically important for shape and apical polarity [357, 358] and thioredoxin, a vital component in the antioxidant system of *T. gondii*, which is essential for the adaption and survival of the parasites in macrophages and other immune effecter cells [359] were detected in an earlier library-Day 4 (a VEG primary library representing a mixture of sporozoites and early stage of tachyzoites gene expression). It is possible that sufficient mRNA was produced in the early stage of tachyzoites to retain the required level of protein expression whereby no further mRNA is required in the later stage of the development.

This observation coincides with the finding of the SAGE study that a major shift in gene expression happens from Day 4 to Day 6 libraries [272]. In fact, another 11 genes out of those 60 genes have been detected in the Day 4 SAGE library, including a cell cycle control protein, putative (641.m01576, 7 peptide hits). The major shift in gene expression patterns could also partly explain the smaller size of the SAGE dataset and the reduced overlap with proteomic data compared to EST and microarray data, as shown in Figure 5.2. The SAGE data provided a more accurate measurement of mRNA expression at the specific Day 6 time point (and closely correlated RH strain [272]), while EST data were collected from a larger collection of various time points within the tachyzoite stage, and the arbitrary 25% cut-off used for microarray dataset is likely to include and exclude some genes that are not expressed at this life stage, as discussed earlier.

Interestingly, while the SAGE study indicated the closest correlated mRNA expression to the laboratory strain RH library was the Day 6 library, the proteomic data actually had a larger overlap with an earlier Day 4 SAGE library. In total, 762

genes detected by proteomic data were shared with Day 4 SAGE library while only 428 genes were shared with Day 6 SAGE library. The finding that the expression profile of proteomic data is closer to an earlier time point of transcriptomic data is likely to reflect the rapid changes in gene expression profiles and the temporal differences between mRNA and protein levels.

While the discrepancies between transcriptome and proteome have been observed in *T. gondii* and other Apicomplexan parasites in this study, the analysis that was designed to check for common features of proteins with low levels of transcription across *Apicomplexa* have only identified a few candidates at the level of orthology. The larger overlap of orthologues observed between *T. gondii* and *N. caninum* than the overlaps with *P. falciparum* in all three comparisons shown in Figure 5.4 is likely to reflect a closer phylogenetic distance between *T. gondii* and *N. caninum*, although no specific class of proteins can be highlighted. In all the three Apicomplexan parasites analysed, there was no apparent underlying rule that can explain the discrepancies between proteomes and transcriptomes.

There were some interesting candidates in the comparison, such as a coatomer protein gamma 2-subunit, putative, 59.m00090; ToxoDB annotation. It has been shown in *T. gondii* that although no transcript evidence has been found with microarray, EST and the three libraries of SAGE, convincing peptide evidence (53 peptide hits) were detected. The protein orthologues of this gene have been consistently detected in *P. falciparum*, *N. caninum* and *C. parvum* but no EST evidence has been found in *N. caninum* and *C. parvum*, and only a single corresponding EST has been seen in a *P. falciparum* blood-stage EST library. Coatomer protein gamma 2-subunit resides in Golgi-derived vesicles which mediate both selective and non-selective transport between the ER and Golgi and/or within

the Golgi cisternae [360, 361]. It is possible that this protein has a very long turn-over rate or an extremely high copy number from a single transcript. The precise reason of the consistent discrepancies observed between proteome and transcriptome of this gene across several Apicomplexan parasites and various life stages is very interesting and requires further investigation.

5.4.3 Conclusions and future directions

In this chapter, proteomic data and transcriptomic data have been compared in a systematic way across several Apicomplexan parasites. The general trend of correlations between mRNA abundance and protein abundance identified by both techniques has provided valuable experimental evidence for gene expression following the central dogma of Gene-Transcription-Translation. However, perhaps more importantly is the finding of discrepancies between these two datasets, which could result from either technical limitations or genuine biological phenomena.

The observation that the proteomic data of *T. gondii* better overlaps the Day 4 SAGE library rather than the Day 6 and RH SAGE libraries is particularly interesting. It highlighted the potential latency of the changes to protein expression profile after the mRNA expression has been altered. Combined with many cases where strong proteomic evidence exists in the absence of transcriptomic evidence, the proteomic data have highlighted the limitations of using transcriptomic data in protein function assignments under the largely applied "guilty-by-association" hypothesis [233].

Both proteomics and transcriptomics are still relatively new technologies, representing some of the first generation of genome-wide data to follow the Apicomplexan genome sequencing projects. The attempt to find common features amongst genes that show discrepancies between transcriptome and proteome has

highlighted several interesting observations, but has not reached an overall conclusion due to the limitation of the data coverage and the lack of quantitative data.

As the technology evolves, more sensitive and absolute quantitative proteomic data would provide accurate measurements of protein abundance, which together with more accurate and thorough coverage of transcriptomic expression provided by the latest RNA-Seq technique, a more meaningful comparison between transcriptome and proteome can be made. More importantly, with more temporal data models analysed, a dynamic and comprehensive understanding of the basic biological process involved in gene expression can be achieved in the near future.

Knowing the importance of quantitative data and encouraged by the good proteomic coverage of metabolic pathways shown in chapter 3, a quantitative proteomic study has been carried out in the next chapter to characterize differential protein expression in *T. gondii* under different growing conditions.

Chapter 6

DIGE analysis of *T. gondii* +/- glucose samples

6.1 Introduction

In the previous chapter, the importance of quantitative proteomic data in the integration with other gene expression data was highlighted by the comparison of the proteome and transcriptome of *Toxoplasma* and other *Apicomplexa* parasites. In this chapter, a preliminary case study has been developed to test the applications of a quantitative proteomic method differential gel electrophoresis (DIGE), in understanding the changes of protein expression profiles of *T. gondii* between different growth conditions.

Traditionally, a gel-based method for quantitative protein expression analysis relies on the statistical analysis of a number of replicate sets of one-sample-one-gel for samples of different status. Due to the poor reproducibility of 2-DE, this approach requires several replicate gels to overcome variations in gel running and therefore is very labour intensive and prone to experimental errors [362]. In 1997, an alternative technique which uses optical detection of proteins with a fluorescent tag in conjunction with 2-DE was developed and named DIGE [363]. Figure 6.1 shows a schematic representation of the DIGE workflow.

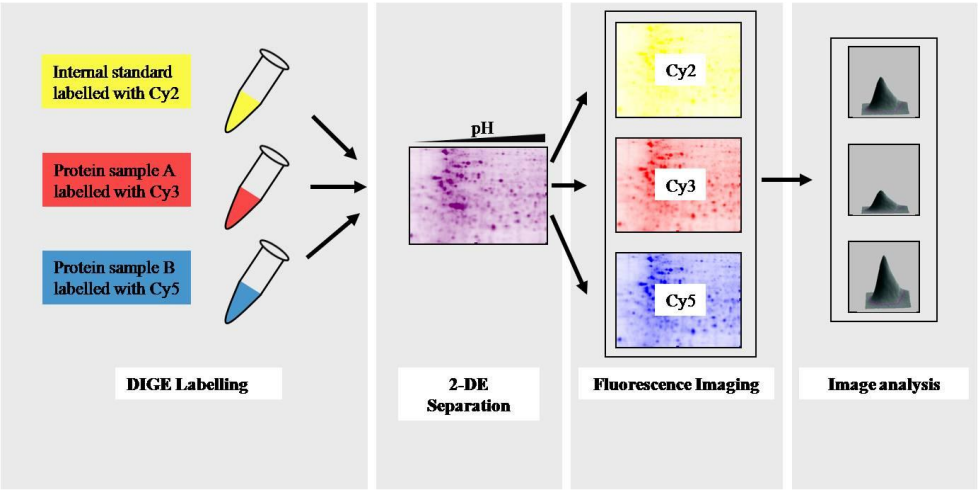


Figure 6.1 Schematic representation of the DIGE workflow.

Equal protein amounts of sample A, sample B and a pooled internal standard are separately labelled by three spectrally distinct fluorescent dyes: Cy2, Cy3 and Cy5. The labelled samples are then mixed and subjected to 2-DE separation followed by fluorescence imaging. Computer-aided image analysis is then performed where the volumes of each spot detected from differentially labelled samples are compared. Four gels that contain four pairs of biological replicates are needed for each experiment to minimize the influences of biological variations in the statistical test. The pooled internal standard is used to calibrate gel-to-gel variations due to heterogeneities during acrylamide polymerisation, gel running and variable precipitation of samples in the first dimension.

The ability of the 2-DE technique in identifying proteins with a broad range of functions and key components of metabolic pathways has been demonstrated in chapter 3. The extended utility of sensitive quantitation of protein expression offered by DIGE has inspired a case study to quantitatively compare changes in the protein expression profile between *T. gondii* grown in the presence and absence of glucose in the cell culture medium.

It has been observed by Dr. Dhanasekaran Shanmugam (University of Pennsylvania, USA) that *T. gondii* tachyzoites are able to grow and replicate for several passages in the total absence of glucose in the growth medium (personal communication). The growth kinetics test of these parasites has revealed that the parasites growing in the absence of glucose are slower than those growing in the presence of glucose by an average of one round of parasite doubling after 48 hrs. Gene expression profiles have been compared between +/-glucose samples and no difference among genes involved in key metabolic pathways for glucose utilization and energy production has been observed (personal communication with Dr. Dhanasekaran Shanmugam).

Metabolomic analysis on +/-glucose samples is also in progress under the collaboration with Dr. Manuel Llinás from Princeton University, USA.

As part of this collaborative project, a preliminary DIGE experiment was carried out and is reported in this chapter. The application and importance of the DIGE technique in the quantitative protein expression profiling and the role of proteomic data in understanding gene expression changes in the context of system is discussed.

6.2 Materials and Methods

DIGE experiments were performed using an Ettan^{TM} DIGE platform, GE Healthcare. The DIGE experiment briefly included the following steps: sample preparation, DIGE labelling, 2-DE, gel imaging, gel analysis using DeCyder^{TM} software and MS analysis.

6.2.1 Sample preparation

Cell culture was carried out by Dr. Dhanasekaran Shanmugam, University of Pennsylvania, USA. Briefly, four pairs of (+) glucose/(-) glucose biological replicates of *T. gondii* tachyzoites were grown in DMEM medium (GIBCO, supplemented with 5.5 mM glucose and 4 mM glutamine for (+) glucose medium, and 4 mM glutamine for (-) glucose medium) on the same batch of confluent HFF host cells plated on T175 tissue culture flasks. For (+) glucose samples, parasites from 1 T175 flask were collected for each replicate while for (-) glucose samples, parasites from 2 T175 flasks were collected for each replicate. All the flasks were inoculated at the same time and 6 hours later the flasks were exchanged with fresh media (either +/-glucose containing as required) to wash out extracellular parasites that had not invaded host cells. Parasites were purified and harvested 48hrs after inoculation using the same method described in section 2.2.1. The frozen samples were then delivered on dry ice.

To each sample (typically containing 2×10^8 parasites), an aliquot of 60 µl of DIGE lysis buffer (8M urea, 4% (w/v) CHAPS, 40 mM Tris-base) was added together with 10µl DNase mix and 5µl of protease inhibitors (Roche). Three cycles of freeze and thaw were performed; each cycle comprised 2 min of vigorous vortexing, followed

by fast freeze with liquid nitrogen and quick thaw to room temperature. The sample was centrifuged at 16,000×g for 30 min at 4 ℃.

The 2-D Clean-Up Kit (GE Healthcare) was used to precipitate protein and remove inferring contaminants such as detergents, salts, lipids, phenolics and nucleic acids. Protein samples were then resuspended in DIGE lysis buffer and the concentrations were adjusted to 5 mg/ml, as determined by 2-D Quant Kit (GE Healthcare). The pH of the sample was adjusted to pH8.5 using 1M NaOH for optimal CyDye^{TM} labelling.

6.2.2 DIGE Labelling

A pooled aliquot of sample A, (+) glucose and sample B, (-) glucose from all batches was prepared as an internal standard. The commercial kit CyDye^{TM} (GE Healthcare) was used for DIGE labelling. The kit contains three different fluorescent dyes Cy2, Cy3 and Cy5. Table 6.1 shows the sample labelling assignments of CyDyes in a typical two sample group comparison setup.

Table 6.1 DIGE labelling Methods

Gel Number	Cy2	Cy3	Cy5
1	Pooled Standard	Sample A 1	Sample B 4
2	Pooled Standard	Sample B 3	Sample A 2
3	Pooled Standard	Sample B 2	Sample A 3
4	Pooled Standard	Sample A 4	Sample B 1

To each CyDye channel, 50 µg of sample was prepared and 1 µl (400 ρmols/µl) of appropriate working dye (Cy2/Cy3/Cy5) was added and vortexed to mix. The sample was then centrifuged briefly and left on ice for 30 min in the dark. To stop the labelling reaction, 1 µl of 10 mM lysine was added to each sample; the sample was centrifuged briefly and left on ice for 10 min. The appropriate Cy2, Cy3 and Cy5 labelled samples were combined into a single tube and vortexed to mix.

6.2.3 2-D gel electrophoresis (2-DE)

First dimension IEF was carried out on 24 cm IPG strips (pH 4-7) following the procedure described in section 2.2.4.3. After the two step equilibration described in section 2.2.4.4, a modified second dimension separation was carried out as DALT precast gels are not compatible with DIGE due to the interference to fluorescent scanning from the plastic back. Low-fluorescence Glass Plates, 27 × 21 cm (GE Healthcare) were used to assemble manually casted gel. All glass plates were thoroughly cleaned with 1% (v/v) decon. The back plates were treated with bind-silane solution (80% (v/v) EtOH, 18% (v/v) dH_2O, 2% (v/v) acetic acid, 0.1% (v/v) γ-methacryloxypropyltrimethoxysilane), which sticks the gel to the glass plate, and left to dry for a minimum one hour at room temperature. Reference markers were applied to the back glass plate and the glass plates were assembled in the gel cast container supplied. SDS-PAGE gels were cast overnight with 30% (w/v) acryalmide, 1.5 M Tris-HCl (pH 8.8), 10% (w/v) SDS, 10% (v/v) TEMED and 10% (w/v) APS.

The equilibrated IPG strips were then assembled onto cast gels and sealed with agarose sealing solution (0.5% (w/v) agarose, trace bromophenol blue). The Ettan DALT*six* System was setup according to the product handbook using the electrophoresis buffer kit supplied. The gel was run at 20 ℃ with an initial wattage of 3 W for 0.5 hour and 17 W per gel thereafter.

A preparative gel was separately prepared with 400 µg of pooled sample from all sample A and sample B replicates. The preparative gel was run together with 4 DIGE gels under same running conditions. The preparative gel was matched to other DIGE gels after the gel analysis and protein spots were picked from the preparative gel for further MS analysis.

6.2.4 Post stain-SYPRO® Ruby

To visualize the proteins present on the preparative gel, SYPRO® Ruby protein gel stain (Invitrogen) was used. The gel was fixed in 40% (v/v) MeOH, 10% (v/v) acetic acid for one hour and stained with SYPRO® Ruby gel stain overnight. The gel was then washed in 10% (v/v) MeOH, 7% (v/v) acetic acid for one hour followed by two 5 min washes in dH_2O prior to gel imaging.

6.2.5 Gel imaging

Both the preparative gel and all four DIGE gels (the latter using three channels of Cy2, Cy3 and Cy5) were scanned using the Ettan™ DIGE Imager (GE Healthcare) following the instructions provided by manufacturer. The gels were scanned at the wavelengths which correspond to the dyes used as shown in Table 6.2.

Table 6.2 The appropriate filter for gel imaging using the Ettan™ DIGE Imager

Dye	Excitation Filter	Emission Filter
Cy2	480/30 nm	530/40 nm
Cy3	540/25 nm	595/25 nm
Cy5	635/30 nm	680/30 nm
SYPRO® Ruby	480/30 nm	595/25 nm

To perform the analytical scan, the pixel size was set to 100 µm required by the image analysis software DeCyder™.

6.2.6 Gel image analysis using the DeCyder™ software

The DeCyder™ image analysis software (GE Healthcare) was used for the analysis of DIGE images. The software consists of five modules. (1) Image Loader: Imports gel image files into the DeCyder database making them accessible for the other modules; (2) DIA (Differential In-gel Analysis): protein spot detection and quantitation on a pair of image channels from the same gel; (3) BVA (Biological Variation Analysis): matches multiple images from different gels to provide

statistical data on differential protein expression levels between multiple groups; (4) Batch Processor: automated image detection and matching of multiple gels without user intervention; and (5) XML Toolbox: generates user specific data reports.

Twelve gel images generated from three channels of four DIGE gels were loaded onto the DeCyder software and the DIA module was used to calculate the number of spots on one gel. All spots in every gel were then detected using the batch processor and the output files were then opened in the BVA module. Gel to gel matching of spots was carried out in the BVA module using the match detection algorithm allowing quantitative comparisons of protein expression across multiple gels. One of the Cy2 images representing internal standard was assigned master gel status and all other images were then matched to it either by automated Batch Processor or manual curation, identifying common protein spots across the gels. The student's T-test was used to determine whether changes in volume of specific spots were significant between samples from different experimental groups. Protein spots which appeared in at least 9 out of 12 images with greater than 1.3 fold changes ($p<0.05$) between (+) glucose and (-) glucose samples were reported. The preparative gel was loaded onto the DeCyder software and assigned as the pick gel in the BVA module. The protein spots of interest were matched onto the pick gel and a pick list was exported by DeCyder software.

6.2.7 Protein identification

The protein spots of interest were picked from the preparative gel by the Ettan^{TM} Spot Picker (GE Healthcare). Manual tryptic digestion was carried out on individual protein spots as described in section 2.2.5. The LTQ (LC-MS/MS) analysis and Mascot searching were performed to acquire protein identifications of each spot using the process described in sections 2.2.6 and 2.2.7.

6.2.8 Protein function annotation

The functional annotation of the proteins identified was carried out using ToxoDB annotation (http://toxodb.org/toxo/), BlastP (http://www.ncbi.nlm.nih.gov/BLAST/), AmiGO (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi) and categorized using MIPS FunCat (http://mips.gsf.de/projects/funcat). The details of each process were described in section 3.2.

6.2.9 Metabolic pathway coverage

Conversion of *T. gondii* genes to key pathway components was determined using the information provided by ToxoDB (http://toxodb.org/toxo/). The enzymes identified in this study were mapped onto KEGG pathways using "Color Objects in KEGG Pathways" tool provided by KEGG (http://www.genome.jp/kegg/tool/color_pathway.html).

6.2.10 Comparison with Microarray data

Microarray expression data of +/-glucose samples were acquired from ToxoDB v4.3 using the query "ToxoDB Queries and Tools/Transcript Expression/ Identify Genes by Microarray Evidence/ Identify Genes based on Differential Expression (T.g.)". The expression library "RH (Type I) in High Glucose vs. RH (Type I) with No Glucose" was selected and the confidence value of ≥0.9 was used. A list of genes showing absolute microarray expression values in the +/-glucose experiment was also downloaded from ToxoDB and compared with DIGE results using Microsoft Office Excel.

6.3 Results

6.3.1 Differential In-gel Analysis

Four biologically replicated pairs of *T. gondii* tachyzoites grown in +/- glucose medium were resolved on four DIGE gels. Around 3800 spots were detected on each gel by the DIA module of DeCyder software. Using separated images generated from different fluorescent dyes on the same gel, the number of differentially expressed proteins on a single gel was determined. Table 6.3 shows the number of gel spots that have greater than a two-fold change between (+) glucose and (-) glucose samples on each of the four DIGE gels.

Table 6.3 Result of individual gel analysis using DIA module of DeCyder

	Up-regulated in (-) glucose sample	Down-regulated in (-) glucose sample	Overall differentially expressed	Spots detected on the gel
Gel 1	55	116	171	3898
Gel 2	49	20	69	3769
Gel 3	29	38	67	3890
Gel 4	28	51	79	3962

6.3.2 Biological Variation Analysis

In order to verify the statistical significance of the differentially expressed proteins determined by the DIA module across all the gels, the BVA module of DeCyder software was used. After matching all the 12 gel images generated from four gels, gel spots that appeared in at least 9 images were considered qualified candidates. Several thresholds were tested to decide the cut-off value to be used for the indication of gene regulation. The higher cut-off value would represent more significant changes between two sample groups but would also decrease the number of gel spots that can be qualified, while choosing a lower cut-off value would allow more gel spots to be analysed but inevitably weaken the significance of changes observed. In this study, if a 2.0 fold change is required, only 3 gel spots would be

qualified to be regulated and only 9 gel spots would be qualified if the threshold is set to a 1.5 fold change. Finally, a threshold of a 1.3 fold change was adopted as it allowed more gel spots to be analysed which would provide a clearer trend of expression profile changes and at the same time, the fold difference was still statistically significant. In total, 27 gel spots showed a greater than 1.3 fold change (p<0.05) between (+) glucose and (-) glucose samples, with 10 gel spots up-regulated in (-) glucose sample and 17 gel spots down-regulated in (-) glucose sample. One example of the gel spot is shown in Figure 6.2.

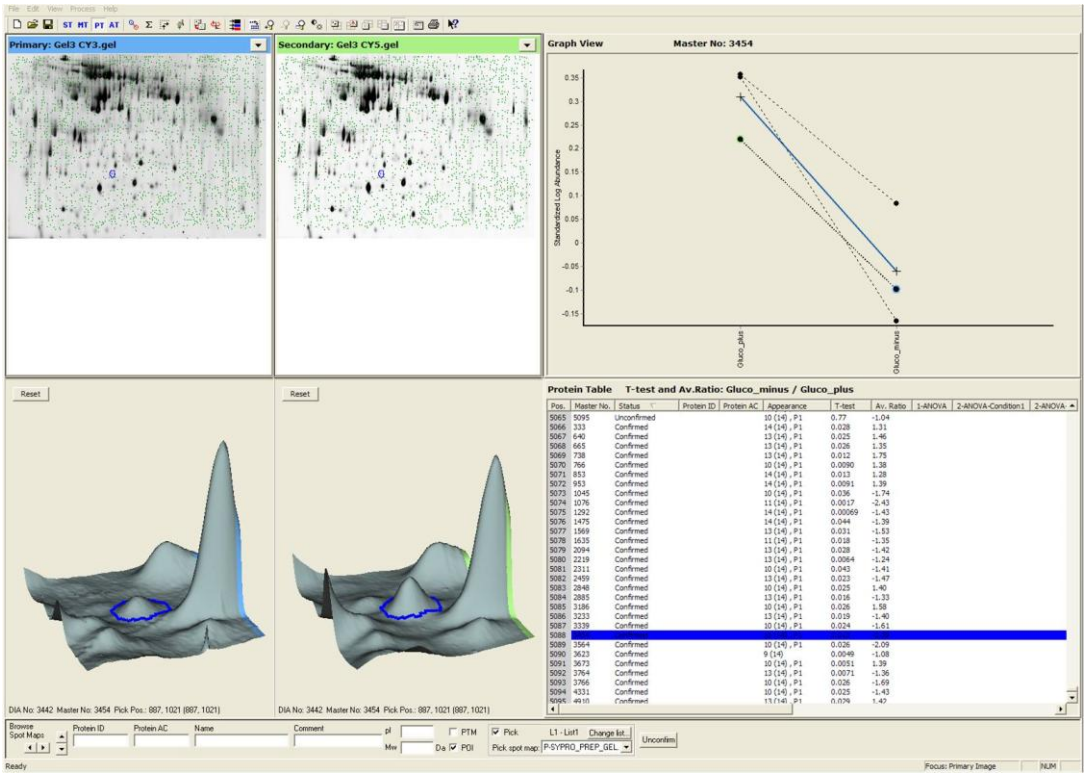


Figure 6.2 Screenshot of the BVA workshop in the Protein Table mode.
The screen is split into 4 sections. The top left section is the gel image view which shows a Cy3 labelled image of (-) glucose sample on the left and the corresponding Cy5 labelled image of (+) glucose sample on the right. The Graph View on the top right section shows the volume comparison of the selected spot (ID: 3454), which is consistently higher in (+) glucose sample than in (-) glucose sample (as measured by standardised log abundance) across all gel images matched. The bottom left section shows the 3D view of the abundance of the matched protein in the two

corresponding gel images selected on the upper left section. The Protein Table on the bottom right section shows spot specific information including the master number of the spot, the T-test value and the average ratio of spot volume.

In the 27 gel spots, 58 release 4 genes and 1 alternative gene model were identified by MS analysis. In many cases, more than one protein has been identified in a single gel spot, an observation seen before in other 2-DE experiments carried out in chapter 2. The influence of this observation to the data interpretation is discussed in section 6.4.1 and the list of proteins identified from each gel spot is provided in Appendix VIII. Of the 59 genes identified, 36 genes were down-regulated and 21 genes were up-regulated in the (-) glucose sample. Interestingly, there were two genes (25.m00211, cytochrome c oxidase, putative and 583.m00712, adenylyl cyclase associated protein) that have been identified from different gel spots that indicated contradictory changes. The possible reasons of this observation are discussed in section 6.4.1.

6.3.3 Functional categorization and metabolic pathway coverage

Functional categorization of the proteins identified was carried out using the same process described in chapter 3 and the result is shown in Figure 6.3. The proteins that have been regulated between +/-glucose were predicted to be involved in a variety of functional categories. Categories that are highly represented in the results are protein fate (folding, modification, destination), such as proteasome and ubiquitin proteins; as well as cell rescue, defence and virulence, including a few heat shock proteins. Notably, both a microneme protein (MIC1) and a dense granule protein (GRA5) were down-regulated over two fold in the (-) glucose sample, which may lead to a decreased invasion rate and disturbed intracellular survival and therefore contribute to the slower growth kinetics of *T. gondii* observed in the glucose depleted medium.

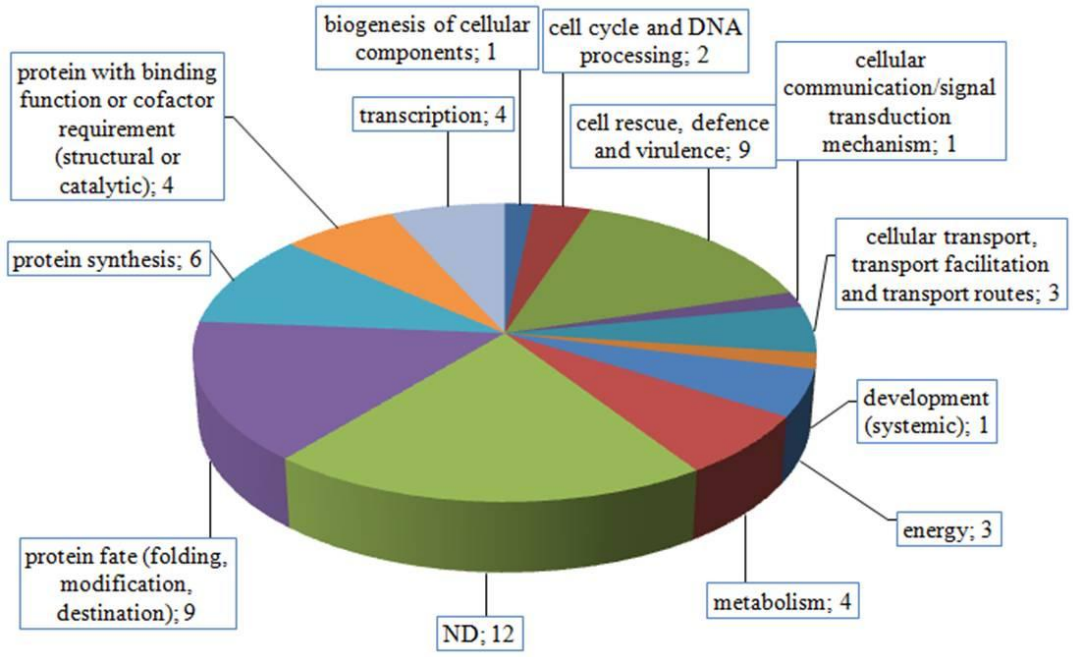


Figure 6.3 Functional categorization of the differentially expressed proteins. Functional assignments of identified official ToxoDB release 4 genes. The prediction was first determined by gene description and GO annotation provided on ToxoDB and then assigned to appropriate MIPS FunCat categories. Putative functional assignment was made to the remainder of identified proteins with information acquired from BlastP, Pfam domain alignments, InterPro and literature searches.

Metabolic pathway coverage of the proteins was determined using the information provided on ToxoDB. Only three proteins that are directly involved in glycolysis and gluconeogenesis were modulated. Figure 6.4 shows the position of these enzymes in relation to other components of the pathway. Enzyme EC 4.2.1.11 (59.m03410, enolase, putative) was down-regulated, and enzyme EC 2.7.1.40 (55.m00007, pyruvate kinase, putative) was up-regulated in the (-) glucose sample. In pyruvate metabolism, enzyme EC 4.1.1.49 (80.m00002, phosphoenolpyruvate carboxykinase (PEP carboxykinase), putative) was up-regulated in the (-) glucose sample.

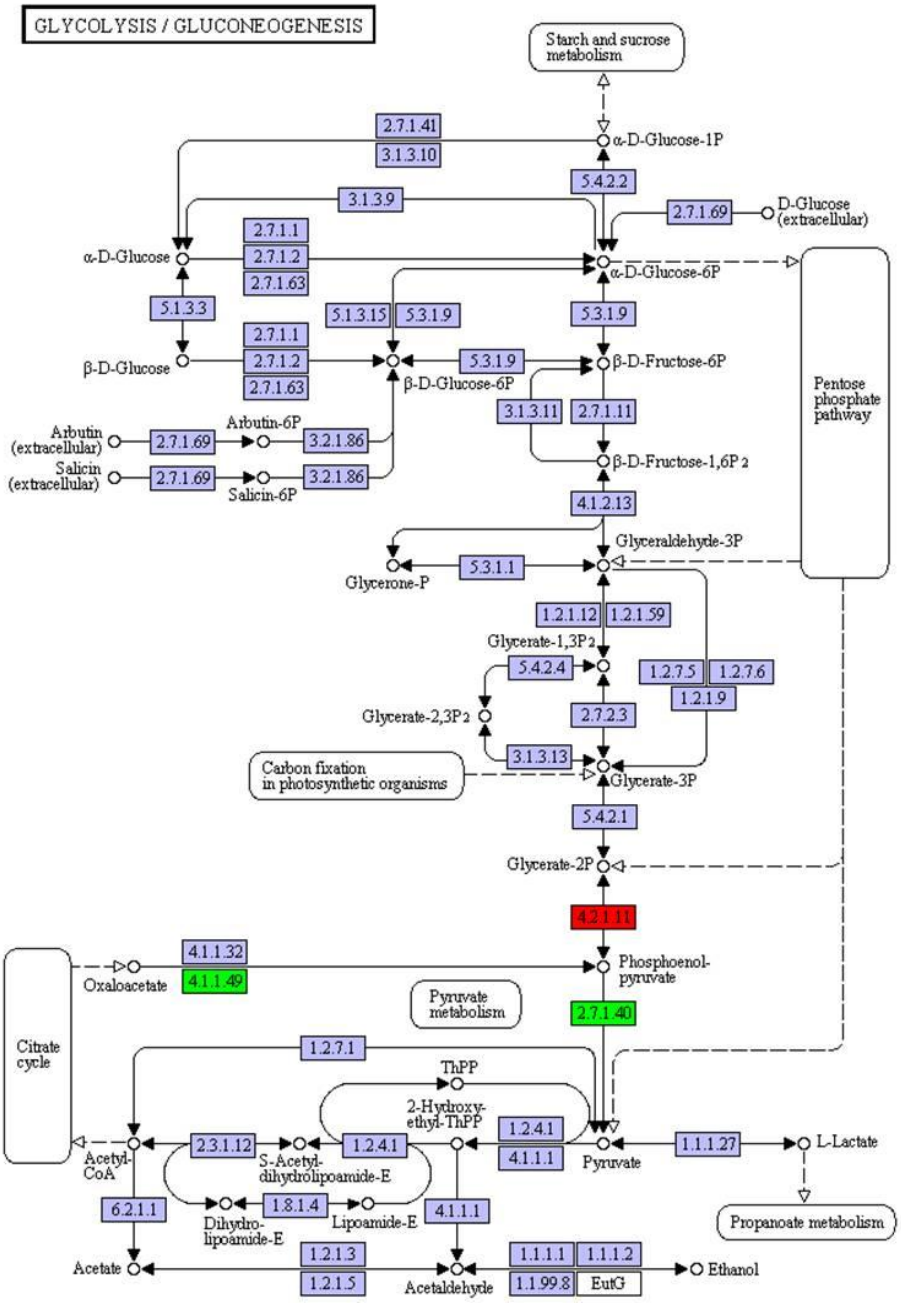


Figure 6.4 Metabolic pathway coverage of the differentially expressed proteins: Glycolysis and gluconeogenesis. Conversion of *T. gondii* genes to key pathway components was determined using the information provided on ToxoDB. The three enzymes identified in this study were mapped onto glycolysis and gluconeogenesis pathways using "Color Objects in KEGG Pathways" tool provided by KEGG (http://www.genome.jp/kegg/tool/color_pathway.html). The enzyme coloured in red is the pathway component that was down-regulated in (-) glucose sample and enzymes coloured in green are pathway components that were up-regulated in (-) glucose sample.

Enolase (EC 4.2.1.11) catalyzes a reversible reaction of the dehydration of glycerate-2-phosphate (glycerate-2P) to form phosphoenolpyruvate (PEP). It has also been found to be involved in invasion [364] as well as play some part in the control of gene regulation during parasite proliferation and differentiation [365, 366]. The precise role of 59.m03410 in the glycolysis and gluconeogenesis requires further investigation. It is possible that the down-regulation of enolase may be caused by the lack of glucose intake and influenced the invasion rate of the parasites.

In gluconeogenesis, pyruvate kinase (EC 2.7.1.40) catalyzes the formation of pyruvate from PEP, and two molecules of ATP are formed for each molecule of glucose. Pyruvate kinase is inactivated by phosphorylation and as it is a control enzyme of flux through glycolysis, phosphorylation will down-regulate glycolysis [367]. In *T. gondii*, the unusual allosteric activation by glucose 6-phosphate (G6P) reported has suggested the involvement of pyruvate kinase to the parasitism [368, 369].

PEP carboxykinase (EC 4. 1.1.49) is involved in the first stages of gluconeogenesis, the synthesis of PEP. After pyruvate converts to oxaloacetate (OAA), PEP carboxykinase then catalyzes it into PEP. This reaction is driven by the hydrolysis of GTP [367]. During starvation, PEP carboxykinase is up-regulated in response to increasing glycogen levels. [367]. In *Toxoplasma*, EC 4.1.1.49 (80.m00002) is the only gene to be predicted in this reaction (there is no evidence for the other enzymes in *Toxoplasma* on ToxoDB). It is possible that the up-regulations of pyruvate kinase and PEP carboxykinase have increased the efficiency of gluconeogenesis by up-regulating the conversion of pyruvate to PEP.

In the biosynthesis pathway of serine, enzyme EC 2.6.1.52 (38.m00011, phosphoserine aminotransferase) was indicated to be down-regulated, which may be an indication that amino acids were used as carbon sources for *T. gondii* in the glucose depleted medium.

6.3.4 Comparison with microarray data

Microarray analysis has also been carried out on the +/- glucose samples. In the 7764 release 4 genes assayed, the expression changes of 4444 release 4 genes between +/- glucose samples were detected by microarray when the confidence value was set at ≥0.9 (90% of them are expected to be correct, according to ToxoDB). In total, 208 genes were up-regulated (> 1.3 fold) and 2405 genes were down-regulated (> 1.3 fold) in the (-) glucose sample. The microarray expression value was downloaded from ToxoDB and compared with the DIGE results. Table 6.4 shows the comparison of DIGE results with the differential expression value indicated by microarray data. For 59 genes identified by the DIGE experiment, only 28 genes were also detected by microarray ≥0.9 confidence data. Eighteen of them showed less than 1.3 fold changes between (+) glucose and (-) glucose samples and are termed "modulated" in the table, while ten of them showed significant changes (greater than 1.3 fold) between the two samples. However, DIGE and microarray have indicated contradictory expression changes for 4 of those 10 genes.

Table 6.4 The comparison of DIGE and microarray results of +/-glucose sample.

ID	Description	Average Ratio (DIGE)	Average Ratio (Microarray)	Correlates
38.m00011	phosphoserine aminotransferase, putative	-2.43	No data	
46.m00027	DEAD/DEAH box helicase, putative	-2.43	No data	
50.m00006	heat shock protein 60	-2.43	No data	
50.m03132	hypothetical protein	-2.43	No data	
50.m03261	hypothetical protein	-2.43	No data	
57.m01695	hypothetical protein	-2.43	-1.44	Yes
80.m00012	microneme protein 1 (MIC1)	-2.43	Modulated	No
44.m02755	small heat shock protein, putative / bradyzoite-specific protein, putative	-2.29	Modulated	No
55.m11049	NAC domain containing protein	-2.29	Modulated	No
44.m04681	hypothetical protein	-2.09	No data	
59.m00008	surface antigen P22	-2.09	No data	
76.m00004	dense granule protein 5 precursor, putative	-2.09	No data	
44.m00031	membrane skeletal protein IMC1, putative	-1.74	7.29	No
59.m03518	asparaginyl-tRNA synthetase, putative	-1.74	Modulated	No
49.m00008	ribosomal protein L23a, putative	-1.61	Modulated	No
583.m00626	hypothetical protein	-1.61	Modulated	No
145.m00347	26S protease regulatory subunit 6a, putative	-1.53	No data	
25.m00007	actin	-1.53	No data	
38.m01067	alanyl-tRNA synthetase, putative	-1.53	No data	
50.m03396	eukaryotic translation initiation factor 3 subunit 3, putative	-1.43	No data	

59.m03410	enolase, putative	-1.43	Modulated	No
645.m00319	RNA recognition motif domain-containing protein	-1.43	No data	
69.m00140	proliferation-associated protein 2G4, putative	-1.43	Modulated	No
76.m00016	elongation factor 1-alpha, putative	-1.43	Modulated	No
541.m01224	hypothetical protein	-1.42	-1.38	Yes
80.m02245	eukaryotic translation initiation factor 3 delta subunit, putative	-1.42	No data	
44.m00051	prohibitin-related	-1.41	No data	
50.m00042	GTP-binding nuclear protein RAN/TC4, putative	-1.41	Modulated	No
583.m00630	purine nucleoside phosphorylase, putative	-1.41	-1.43	Yes
TgTwinScan_5606	No Significant Blast Hit	-1.41	No data	
27.m00003	protein disulfide isomerase, putative	-1.39	No data	
583.m00619	60s ribosomal protein L4, putative	-1.39	No data	
583.m05569	hypothetical protein	-1.36	-1.54	Yes
42.m05838	hypothetical protein	-1.35	No data	
25.m00211	cytochrome c oxidase, putative	-1.33	No data	
55.m08199	hypothetical protein	-1.33	Modulated	No
583.m00712	adenylyl cyclase associated protein	-1.33	No data	
76.m01670	peroxiredoxin family protein/glutaredoxin, putative	-1.33	-1.57	Yes
20.m03912	elongation factor 2, putative	1.31	Modulated	No
38.m01113	heat shock protein, putative	1.31	No data	
55.m04729	heat shock protein, putative	1.31	No data	
55.m05071	26S proteasome non-ATPase regulatory subunit-related	1.31	-1.64	No
50.m00016	bifunctional dihydrofolate reductase-thymidylate synthase (DHFR-TS)	1.38	1.99	Yes
65.m00001	nucleoside-triphosphatase I	1.38	No data	
20.m03930	20k cyclophilin	1.39	No data	

42.m00091	hypothetical protein	1.39	Modulated	No
80.m00002	phosphoenolpyruvate carboxykinase, putative	1.39	No data	
25.m00211	cytochrome c oxidase, putative	1.4	No data	
49.m03152	proteasome subunit alpha type 4, subunit	1.4	No data	
44.m00052	58 kDa phosphoprotein, putative	1.46	Modulated	No
27.m00081	hypothetical protein	1.58	No data	
44.m02723	gi|22035894|emb|CAD43149.1|; putative PDI-like protein [*Toxoplasma gondii*]; 5e-125	1.58	No data	
46.m00017	ubiquitin-conjugating enzyme, putative	1.58	Modulated	No
583.m00712	adenylyl cyclase associated protein	1.58	No data	
59.m03611	heat shock protein HSLV, putative	1.58	-1.89	No
113.m00780	conserved hypothetical protein	1.75	Modulated	No
31.m00869	hypothetical protein	1.75	No data	
42.m00027	26S proteasome non-ATPase regulatory subunit 4, putative	1.75	Modulated	No
55.m00007	pyruvate kinase, putative	1.75	Modulated	No
55.m00139	46 kDa FK506-binding nuclear protein, putative	1.75	No data	
76.m01657	hypothetical protein	1.75	-2.30	No

The first column shows the 59 genes identified by DIGE experiment and the gene descriptions are shown in the second column. The third column shows the average ratio of expression changes between +/-glucose samples indicated by DIGE. The fourth column shows the average ratio of changes indicated by microarray data where changes less than 1.3 fold are termed modulated. The fifth column shows the correlation between the two experiments. Genes coloured in red were down-regulated in (-) glucose sample and genes coloured in green were up-regulated. Genes coloured in blue showed contradicted results from different gel spots on DIGE.

6.4 Discussion

In this chapter, DIGE analysis has been used to characterize the changes in *T. gondii* protein expression profiles in glucose depleted medium. The DIA module has shown that over 3000 protein spots can be detected on each of the four DIGE gels, which highlighted the good resolution of the technique. In the BVA module, the gel matching and statistical analysis helped to overcome the limitations introduced by both biological variations and handling variations. In total, 27 gel spots showed a greater than 1.3 fold change ($p<0.05$) between (+) glucose and (-) glucose samples and 58 release 4 genes and one alternative gene model have been identified. Both the functional categorization and metabolic pathway coverage analysis have highlighted some interesting groups of proteins that deserve further investigations. As the main focus of this chapter was to provide a preliminary test of the application of DIGE in quantitative proteomic studies, considerations are required to the following limitations observed in this chapter.

6.4.1 Multiple proteins identified in a single gel spot

As mentioned before, the directions of protein regulation were not definite due to the multiple proteins identified in a single gel spot and caution is required in the data interpretation. For example, among all the enzymes described in section 6.3.3, only one protein 80.m00002 (PEP carboxykinase) was identified by a single specific gel spot.

In a gel spot from which several proteins can be identified, it is hard to determine which protein has dominated the changes detected at the gel spot level. On the other hand, a genuine protein expression change of a low abundance protein could be masked by a high abundance protein co-migrating that did not have any significant

changes. Multiple proteins identified in a single gel spot have also led to the observation of proteins that have been identified from different gel spots which indicated contradictory expression changes.

The expression changes of a single protein candidate can be confirmed by using a quantitative MS platform that can specifically target signature peptides such as the multiple reaction monitoring (MRM) assay [370]. In this way, DIGE can serve as a preliminary step that provides a screening of protein expression changes in a high throughput fashion, while further MS analysis can be used to unambiguously confirm the expression changes of a specific protein.

6.4.2 Repeatedly identified differentially expressed proteins

A recent study has investigated the occurrence of individual differentially expressed proteins in 2-DE experiments reported in169 articles studying human and mouse or rat samples published in the Proteomics journal between 2004 and 2006 [371]. This meta-analysis study has revealed that among the most frequently identified differentially expressed proteins were enolase, heat shock proteins, proteasome subunits and pyruvate kinase [371] which have also been identified in this chapter.

Since all the commonly identified proteins are highly abundant soluble proteins, it raised the concern whether their frequent identification represents a technical artefact, limitation of the method or universal cellular sensors that respond to multiple different stimuli [371]. Although the full list of 169 articles analysed in this study has not been released, it is likely that many of them were carried out using the conventional one-sample-one-gel setup. In fact, a quick search has revealed that only 56 articles published in the Proteomics journal between 2004 and 2006 have used a DIGE platform (http://www.gopubmed.org/web/gopubmed/). It is possible that these

commonly identified proteins were easy to identify by 2-DE techniques and the changes in gel spot volume caused by gel to gel variations were falsely reported as biological expression differences.

However, a higher confidence can be expected during the interpretation of DIGE results. One of the technical advantages provided by DIGE is the ability to minimize the technical limitations experienced in the conventional 2-DE gel. Firstly, by labelling with three fluorescent CyDyes, different samples can be separated on the same gel under the same running conditions. Secondly, the variations between different biological replicates as well as those introduced during sample handling and first dimension IEF are minimized using pooled internal standard and advanced statistical analysis. As shown in this chapter, around 70 to 170 gels spots on each DIGE gel have shown greater than a two-fold change between (+) glucose and (-) glucose samples in the DIA module. However, after matching the four DIGE gels together, only 27 gels spots have shown greater than a 1.3 fold change reported by the statistical analysis.

Admittedly, the advantages of the DIGE technique over the conventional one-sample-one-gel setup cannot bypass the limitations inherited from 2-DE, such as resolving very large or small proteins, low abundance proteins, hydrophobic proteins and proteins with extreme pIs as discussed in chapter 2. This would inevitably lead to an artificial enrichment of the identifications of high abundance, soluble proteins as observed in the meta-analysis study [371]. However, with the increased confidence provided by DIGE, it is more likely that the identifications of these repeatedly identified differentially expressed proteins in this chapter represent true universal cellular sensors responding to multiple different stimuli.

6.4.3 The comparison of DIGE results with the microarray data

The comparison of DIGE results with the microarray data has shown a surprisingly low number of genes that have been detected by both DIGE and microarray. This may reflect the resolution limitations of the DIGE technique compared to microarray. However, more interestingly are those genes that showed different expression changes indicated by DIGE and microarray.

There were 18 genes identified that have shown significant changes at protein expression level while the mRNA expression level has only been slightly modulated. Since the samples were collected after 48 hours of inoculation, it is possible that the significant changes of mRNA level in the initial response have returned to the basic level while protein expression level remains significantly altered. The similar delay of protein expression has been observed in chapter 5 when proteomic data were compared with SAGE data.

There were also four genes that showed contradictory expression changes between mRNA and protein levels. The biggest discrepancy has been observed for 44.m00031, a membrane skeletal protein (IMC1). The DIGE result indicated the expression of this protein in the (-) glucose sample was down-regulated 1.74 fold, while the mRNA expression was up-regulated 7.29 fold. Membrane skeletons play an important role in the maintenance of cell shape and integrity while IMC also plays a critical role as an anchor for the glideosome [65, 372]. It has been found that IMC1appears to be stable during the early stage of invasion but is degraded rapidly during the late stages of the endodyogeny, while the membrane skeleton of the daughter parasites is assembled at an earlier stage of cell division which does not involve recycling of components of the mother cell network [373]. This may partly contribute to the discrepancies observed between mRNA and protein levels where

the increase of the mRNA expression is required by the daughter parasites and the detected protein expression is affected by the degradation. However, the precise effect of the glucose depleted medium to the expression of IMC1 still requires further investigation.

6.4.4 Conclusion and future directions

In this chapter, the application of the quantitative proteomic approach DIGE in understanding protein expression changes of *T. gondii* tachyzoites in +/- glucose medium were investigated and discussed. Several interesting candidates have been identified such as enzymes that are involved in the glycolysis and gluconeogenesis pathways, as well as proteins involved in host cell invasion and intracellular survival.

The quantitation accuracy of DIGE has been proved to be comparable to the MS-based chemical and metabolic isotope labelling quantitation methods such as ICPL (Isotope Coded Protein Label), iTRAQ (isobaric tags for relative and absolute quantification), and cICAT (cleavable isotope-coded affinity tags) [374-376]. Compared to other MS-based quantitation methods that measure the protein abundances at the peptide level, DIGE measures the protein abundances at the protein level, where information on isoforms-specific expression can be preserved.

As discussed before, DIGE is able to provide a high throughput large scale quantitative screening of protein expression changes. However, considerations are also required to the separation limitations inherited from 2-DE, as well as the frequently observed phenomena of multiple proteins co-migrating to the same gel spot. It should also be noted that, if a spot has been down-regulated to the point of disappearing from the gel, it would not be reported but which nevertheless, could represent a significant event. Data interpretation of DIGE results should be carried

out cautiously and in some cases, further investigations are required to confirm the expression changes observed. One possibility of the discrepancies observed between the DIGE results and microarray data is the temporal differences between mRNA and protein expression. The inclusion of more time-points in the investigation as well as the addition of metabolomic data will provide valuable insights into the dynamic changes of *T. gondii* under different conditions.

Chapter 7

Summary and forward perspectives

In the one hundred years since the discovery of *Toxoplasma gondii*, the understanding of the biology of the parasite has rapidly progressed along with the technical developments. Since the development of expression sequence tags (ESTs) [246], the rate of gene discovery has accelerated. With the first output of the *T. gondii* genome sequencing project completed in 2003, followed by the great and continuing efforts towards genome annotation consecutively published on ToxoDB [108], numerous applications have been made feasible. The focus of *T. gondii* research has gradually moved to the understanding of gene expression and gene functions on the genome scale. Equipped with the combination of protein separation techniques, the latest technical advances of mass spectrometry and bioinformatics, proteomics has rapidly established an important role in post-genomics era applications by providing first-hand data on the functional products of gene expression. In this study, the applications of proteomic technique in the understanding of the proteome of *T. gondii* have been investigated and discussed.

In chapter 2, the use of three complementary proteomic separation platforms followed by powerful MS/MS analyses have harnessed the advantages provided by each technique and ensured the best coverage of the expressed proteome whilst utilising limited amounts of sample and labour. The raw MS data have been searched against a well designed, up-to-date database that provided comprehensive coverage of the predicted *T. gondii* proteome. After the careful verification of the MS data searching results, more than two thousand (2252) unique release 4 genes and 394 non-redundant alternative gene models and ORFs have been identified from the tachyzoite proteome, which represents almost one third (29%) of the entire predicted proteome. Since the proteomic data were only collected from the tachyzoite stage, the actual proteomic coverage of tachyzoite proteome would be expected to be much

higher. Notably, 813 of the 2252 (36%) proteins identified are annotated as hypothetical or conserved hypothetical proteins. With the expression evidence acquired in this study, the status of these proteins can now be changed to "confirmed" proteins. This work has provided the first large scale proteomic profiling of *T. gondii* and greatly increased the knowledge of the *T. gondii* proteome, increasing the percentage of known expressed genes from ~4% in 2002 to 29% of the total genome. The acquisition of protein expression evidence for almost one third of the predicted *T. gondii* proteome has set the foundation for a diverse range of applications in the post-genomic era.

In chapter 3, data interpretation has been carried out using multiple bioinformatics tools, which provided valuable insights into the potential biological roles of the expressed proteins. The SignalP and TMHMM predictions have predicted the composition of signal peptide and transmembrane domains in the expressed proteome and suggested a good sampling standard has been achieved. Both of the subcellular localization predictions and functional categorization have highlighted protein candidates with important biological functions, such as proteins involved in the host cell invasion process and apicoplast proteins. The confirmation of the expression of these proteins will enlighten further detailed studies on chosen proteins of interest. The analysis of metabolic pathway coverage has again demonstrated the sensitivity of this proteomic study and suggested a potential application of proteomic techniques in understanding key metabolism changes, which was further pursued in chapter 7. In addition to the bioinformatics interpretation of the standalone proteomic data, further efforts have been made to integrate proteomic data with other genomic resources in the context of ToxoDB.

In chapter 4, *T. gondii* proteomic data have been stored in the online data repository, Tranche, in the form of raw MS data; it has also been mapped on to ToxoDB in the form of peptide identifications for the first time. Both databases offer publically accessible interfaces for data downloading and examination, which will benefit the global research community. The integration of peptide identifications onto the ToxoDB genome scaffold has highlighted the incompleteness of the release 4 genome annotations. While the majority of the peptide evidence supported the correct ORFs and the positioning of start and stop codons, peptide identifications have confirmed 421 splice sites that are only predicted by alternative gene models. Examined with peptide evidence, three types of annotation errors have been highlighted using GBrowse, such as missing exons, alternative frame shift or strand orientation and the incorrect positioning of the exon-intron boundaries. These important observations have led to the discussion of the limitation of the current peptide mapping algorithms. More importantly was the limitation experienced with the fundamental work flow of conventional bottom-up proteomics which exclusively relied upon the correct prediction of gene coding sequences and suffered labour intensive processing caused by the successive upgrades of genome annotations. Stimulated by the discrepancies observed between the predicted genome annotation and actual, expressed peptide identifications, the importance and potential of a new genome annotation pipeline was discussed. Despite some technical hurdles discussed that require further investigation, this future proteogenomic approach will harness the essence of the raw MS data and combined with the information contained within the transcriptome, should lead to a near "perfect" genome annotation in the future.

In chapter 5, another important application of proteomic data was investigated that has been made available by the integration of the data with other genomic resources,

the comparison of the proteome and transcriptome of *Toxoplasma* and other *Apicomplexa* parasites. In *T. gondii*, a weak correlation between mRNA abundance and protein abundance has been observed, where more proteins have been detected by proteomics for genes with higher mRNA expression levels determined by microarray. More interesting were the discrepancies that were revealed by the comparison of the proteomic data with various transcriptomic data. The 60 tachyzoite genes that were exclusively identified by proteomic data with no corresponding transcript expression evidence illustrated the sensitivity of proteomic approaches. This also provided interesting candidates for understanding the relationship between mRNA and protein abundance levels in *Toxoplasma*. Another interesting observation was that the proteomic data correlated better with a SAGE library from an earlier time point of sampling, which is likely to reflect the rapid changes in gene expression profiles and the temporal differences between mRNA and protein levels. The comparisons across four species of *Apicomplexa* have seen similar discrepancies between the proteome and transcriptome. The attempt to find common features amongst genes that show discrepancies between transcriptome and proteome at the level of orthology has highlighted several interesting observations and candidates that require further study. Although the analysis has not reached an overall conclusion, the process has highlighted the importance of temporal and quantitative data in a better understanding of the correlation between mRNA and protein expression.

In chapter 6, a preliminary case study was carried out to test the applications of the quantitative proteomic approach DIGE in understanding protein expression changes of *T. gondii* tachyzoites grown in the presence or absence of glucose. DIGE has several technical advantages over the conventional gel based one-sample-one-gel

quantitative approaches. The bioinformatics interpretation of the DIGE result has highlighted several interesting candidates such as enzymes that are involved in the glycolysis and gluconeogenesis pathways, as well as proteins involved in host cell invasion and intracellular survival. The comparison with microarray data has shown a surprisingly small number of genes that have been detected by both DIGE and microarray, which may reflect the resolution limitations of the DIGE technique. More interestingly, discrepancies have been observed for 18 genes which have shown significant changes at the protein expression level while the mRNA expression level has only been slightly modulated. In addition to that, four genes have shown contradictory expression changes between mRNA and protein levels. These observations have again highlighted the importance of proteomic data in the understanding of the correlation between mRNA and protein expressions. In order to better understand the DIGE results, considerations are required to the separation limitations inherited from 2-DE and more importantly, the frequently observed phenomena of multiple proteins co-migrating to the sample gel spot. Nonetheless, DIGE is able to provide a preliminary, high throughput, large scale quantitative screening of protein expression changes, while the data verification is readily available using the methods discussed in the chapter such as MRM assay.

In this study, various proteomic techniques have been used to characterize the proteome of the *T. gondii* tachyzoite. The results have expanded the knowledge of diverse disciplines of *T. gondii* biology, such as the composition of the expressed proteome and differential expressed proteins in glucose depleted medium. The proteomic data have shown important applications in the validation of the existing gene models and more importantly, the integration into proteogenomic pipelines to improve future genome annotations. The discrepancies observed between proteomic

and transcriptomic data have highlighted the importance of gene expression regulation following the central dogma of Gene-Transcription-Translation and enlightened further investigations.

Proteomic analysis of *T. gondii* can be expanded in several directions in the future. Firstly, with the current knowledge, little is known about the stage conversion of *T. gondii* from different life cycle forms, the proteomic profiling of more life stages of *T. gondii* would provide valuable insights into the complex life-cycle of the parasite and the stage differentiation.

Secondly, more organelle specific or condition specific sub-proteomic studies would benefit the understanding of *T. gondii* virulence and disease pathogenesis in more detail. For example, the proteomic characterization of proteins involved in host cell invasion, subcellular localizations, proteins targeting to apicoplast, as well as the alteration of the protein expression profile in responds to various environmental changes.

Thirdly, the recent attention of post-translational modifications has already identified some important candidates, such as the glycosylated proteins that are involved in host-cell interactions [259], SUMOylated proteins that play a putative role in host cell invasion and cyst genesis [258]. Further investigations can be carried out in the characterization of important post-translational modifications using techniques such as 2-DE or affinity chromatography.

In addition to that, comparative proteomic studies can be designed in *Apicomplexa* to investigate subjects such as host specificities and the regulation of gene expressions in different life cycle stages. Together, this will lead to a better understanding of *Apicomplexa* biology.

Last but not least, the application of proteogenomics investigated in this study has already shown promising prospects in refining genome annotation, the further developments in this field could result in a closer involvement of proteomic data in genome annotation and fundamentally change the position of proteomic data in the future genome annotation pipelines.

To harness the technical advantages provided by the latest development of proteomics, it is important to examine temporal changes using the quantitative proteomic approaches. This will in turn benefit the integration of proteomic data with other large scale expression data such as transcriptomics and metabolomics. As discussed in chapters 5 and 6, the expression profiles provided by transcriptomics and proteomics have dramatically increased the knowledge of global gene expression and gene function, the ease and pace of discovery and enabled more complex global analysis. Together, they have formed a milestone for a comprehensive understanding of biological processes.

In the past, guided by the hypothesis-driven experimentation, every major technique and conceptual invention has promoted the understanding of biology to a new level [377], such as the invention of the microscope, modern genetics discovered by Mendel and the discovery of the genetic code which is still rapidly powering the developments of modern biology. Various omics techniques such as transcriptomics, proteomics, interactomics and metabolomics have already made a giant step towards the understanding of all the basic biological process. In order to make full use of these expression data, the ultimate aim would be to use an integrative approach that can utilize multiple data sources such as literature, public databases, and high throughput experimental data. These data, incorporated with mathematical modelling and computer simulation tools, will provide a dynamic view of systems biology

within temporal, spatial and physiological contexts, which will in turn extend the future experimentation by generating novel hypotheses for further experiments.

The detailed data standards and modelling algorithms in systems biology are beyond the scope of the current study. However, a well designed algorithm will undoubtedly accelerate the knowledge accumulation process towards the understanding of the entire system. To take an example from ancient Chinese philosophy, a book called "I Ching", written in the 9th century BC, has described a system in which everything can be divided into two complementary parts (yin-yang). Yin and yang are further divided into 64 different hexagrams; each represents a description of a state or process. The hexagrams are connected by an algorithm where they either rule or are ruled (at various levels) by each other. The modern equivalent term to the hexagram would be "variable"; following the algorithm developed in "I Ching", the correct input of a collection of known variables will lead to a prediction of the changes to the unknown variables. The philosophy developed in "I Ching" has influenced the development of various aspects of ancient Chinese science, such as traditional medicine.

The process of knowledge accumulation towards an integrative understanding of every subject system can be summarized in a three-step cycle: macrocosmic-microcosmic-macrocosmic. In the first macrocosmic step, there are too many variables that are unknown, everything is observed in a crude way. That is when "I Ching" was created to organize all the seemingly unrelated observations and guide the exploration. In the second step, when enough evidence and theories have been discovered, they are further divided into diverse disciplines, where everything can be studied into more detail microcosmically. In the third step, when sufficient

knowledge has been acquired, the subject system can be studied at a macrocosmic view again, only to a much higher integrative level.

Through each cycle of the three steps, the knowledge accumulated will drive the development of the new cycle. In the post genomic era, it can be foreseen that the various omics techniques will rapidly fulfil the data requirements for the second microcosmic step in the current cycle. Followed by the macrocosmic step of integrative systems biology, this will soon uncover the mysteries of the majority of biological processes, which will bring the understanding of biology to a new level.

Sharing the excitements of the post genomic era in biology with the latest development of physics, some potential limitations can be predicted for the integrative systems biology, through which perspectives can be speculated that will drive the development of the biology in the new cycle. In his book of "A Brief History of Time", Stephen Hawking has reasoned that: "Even if we do find a complete set of basic laws, there will still be in the years ahead the intellectually challenging task of developing better approximation methods, so that we can make useful predictions of the probable outcomes in complicated and realistic situations" [378]. In physics, the limitations come from the uncertainty principle of quantum mechanics and the difficulty of modelling all theories to equations; whilst in biology, the organic world is much more difficult and complex to predict, with indeterminacies such as genetic mutations and stochastic behaviour.

The proteomic study of *T. gondii* has emerged to be an important component of the rapidly developing integrative systems biology, fascinated by the power and applications of proteomics and other omics techniques in the pursuit of the comprehensive understanding of all basic biological process, I would like to end my

thesis quoting Stephen Hawking's prospect in "A Brief History of Time": "A complete, consistent, unified theory is only the first step: our goal is a complete understanding of the events around us, and of our own existence" [378].

References

1. Nicolle C, Manceaux,L: **Sur une infection à corps de Leishman (ou organismes voisins) du gondi.** *C R Seances Acad Sci* 1908**:**763-766.
2. Splendore A: **Un nuovo protozoa parassita deconigli incontrato nelle lesioni anatomiche d'une malattia che ricorda in molti punti il Kala-azar dell'uoma.** *Rev Soc Sci Sao Paulo* 1908, **3:**109-112.
3. Nicolle C, Manceaux, L: **Sur un protozoaire nouveau du gondi.** *C R Acad Sci* 1909, **148:**369.
4. Tenter AM, Heckeroth AR, Weiss LM: ***Toxoplasma gondii*: from animals to humans.** *Int J Parasitol* 2000, **30:**1217-1258.
5. Hill D, Dubey JP: ***Toxoplasma gondii*: transmission, diagnosis and prevention.** *Clin Microbiol Infect* 2002, **8:**634-640.
6. Benard A, Petersen E, Salamon R, Chene G, Gilbert R, Salmi LR: **Survey of European programmes for the epidemiological surveillance of congenital toxoplasmosis.** *Euro Surveill* 2008, **13**.
7. Porter SB, Sande MA: **Toxoplasmosis of the central nervous system in the acquired immunodeficiency syndrome.** *N Engl J Med* 1992, **327:**1643-1648.
8. Frenkel JK, Dubey JP, Miller NL: ***Toxoplasma gondii* in cats: fecal stages identified as coccidian oocysts.** *Science* 1970, **167:**893-896.
9. Dubey JP: **Toxoplasmosis.** *J Am Vet Med Assoc* 1986, **189:**166-170.
10. Frenkel JK: ***Toxoplasma* in and around us.** *BioScience* 1973, **23:**343-352.
11. Sheffield HG, Melton ML: **The fine structure and reproduction of *Toxoplasma gondii*.** *J Parasitol* 1968, **54:**209-226.
12. Goldman M, Carver RK, Sulzer AJ: **Reproduction of *Toxoplasma gondii* by internal budding.** *J Parasitol* 1958, **44:**161-171.
13. Ferguson DJ, Dubremetz JF: **The ultrastructure of *Toxoplasma gondii*.** In *Toxoplasma gondii: the model* Apicomplexan*: perspectives and methods.* First edition. Edited by Weiss LM, Kim K. London: Academic Press; 2007: 19-48
14. Gustafson PV, Agar HD, Cramer DI: **An electron microscope study of *Toxoplasma*.** *Am J Trop Med Hyg* 1954, **3:**1008-1022.

15. Ferguson DJ, Hutchison WM: **An ultrastructural study of the early development and tissue cyst formation of *Toxoplasma gondii* in the brains of mice.** *Parasitol Res* 1987, **73:**483-491.
16. Denton H, Roberts CW, Alexander J, Thong KW, Coombs GH: **Enzymes of energy metabolism in the bradyzoites and tachyzoites of *Toxoplasma gondii*.** *FEMS Microbiol Lett* 1996, **137:**103-108.
17. Tomavo S: **The differential expression of multiple isoenzyme forms during stage conversion of *Toxoplasma gondii*: an adaptive developmental strategy.** *Int J Parasitol* 2001, **31:**1023-1031.
18. Ferguson DJ, Hutchison WM: **The host-parasite relationship of *Toxoplasma gondii* in the brains of chronically infected mice.** *Virchows Arch A Pathol Anat Histopathol* 1987, **411:**39-43.
19. Dubey JP, Lindsay DS, Speer CA: **Structures of *Toxoplasma gondii* tachyzoites, bradyzoites, and sporozoites and biology and development of tissue cysts.** *Clin Microbiol Rev* 1998, **11:**267-299.
20. Dubey JP: **The life cycle of *Toxoplasma gondii*.** In *Toxoplasma: molecular and cellular biology.* Edited by Ajioka JW, Soldati D. Norfolk: Horizon Bioscience; 2007: 3-16
21. Ferguson DJ, Hutchison WM, Pettersen E: **Tissue cyst rupture in mice chronically infected with *Toxoplasma gondii*. An immunocytochemical and ultrastructural study.** *Parasitol Res* 1989, **75:**599-603.
22. Lyons RE, McLeod R, Roberts CW: ***Toxoplasma gondii* tachyzoite-bradyzoite interconversion.** *Trends Parasitol* 2002, **18:**198-201.
23. Jones JL, Dubey JP: **Waterborne toxoplasmosis - Recent developments.** *Exp Parasitol* 2009.
24. Dubey JP, Frenkel JK: **Cyst-induced toxoplasmosis in cats.** *J Protozool* 1972, **19:**155-177.
25. Speer CA, Dubey JP: **Ultrastructural differentiation of *Toxoplasma gondii* schizonts (types B to E) and gamonts in the intestines of cats fed bradyzoites.** *Int J Parasitol* 2005, **35:**193-206.
26. Dubey JP: **Bradyzoite-induced murine toxoplasmosis: stage conversion, pathogenesis, and tissue cyst formation in mice fed bradyzoites of different strains of *Toxoplasma gondii*.** *J Eukaryot Microbiol* 1997, **44:**592-602.

27. Speer CA, Dubey JP: **Ultrastructure of early stages of infections in mice fed *Toxoplasma gondii* oocysts.** *Parasitology* 1998, **116 (Pt 1):**35-42.
28. Piekarski G, Pelster B, Witte HM: **Endopolygeny in *Toxoplasma gondii*.** *Z Parasitenkd* 1971, **36:**122-130.
29. Ferguson DJ, Hutchison WM, Dunachie JF, Siim JC: **Ultrastructural study of early stages of asexual multiplication and microgametogony of *Toxoplasma gondii* in the small intestine of the cat.** *Acta Pathol Microbiol Scand B Microbiol Immunol* 1974, **82:**167-181.
30. Pelster B, Piekarski G: **Electron microscopical studies on the microgametogeny of *Toxoplasma gondii*.** *Z Parasitenkd* 1971, **37:**267-277.
31. Ferguson DJ, Hutchison WM, Siim JC: **The ultrastructural development of the macrogamete and formation of the oocyst wall of *Toxoplasma gondii*.** *Acta Pathol Microbiol Scand B* 1975, **83:**491-505.
32. Pelster B, Piekarski G: **Ultrastructure of the macrogametes in *Toxoplasma gondii*.** *Z Parasitenkd* 1972, **39:**225-232.
33. Pfefferkorn LC, Pfefferkorn ER: ***Toxoplasma gondii*: genetic recombination between drug resistant mutants.** *Exp Parasitol* 1980, **50:**305-316.
34. Ferguson DJ: ***Toxoplasma gondii* and sex: essential or optional extra?** *Trends Parasitol* 2002, **18:**355-359.
35. Dubey JP: **Oocyst shedding by cats fed isolated bradyzoites and comparison of infectivity of bradyzoites of the VEG strain *Toxoplasma gondii* to cats and mice.** *J Parasitol* 2001, **87:**215-219.
36. Ferguson DJ, Birch-Andersen A, Siim JC, Hutchison WM: **Ultrastructural studies on the sporulation of oocysts of *Toxoplasma gondii*. I. Development of the zygote and formation of the sporoblasts.** *Acta Pathol Microbiol Scand B* 1979, **87B:**171-181.
37. Ferguson DJ, Birch-Andersen A, Siim JC, Hutchison WM: **Ultrastructural studies on the sporulation of oocysts of *Toxoplasma gondii*. II. Formation of the sporocyst and structure of the sporocyst wall.** *Acta Pathol Microbiol Scand B* 1979, **87B:**183-190.
38. Ferguson DJ, Birch-Andersen A, Siim JC, Hutchison WM: **Ultrastructural studies on the sporulation of oocysts of *Toxoplasma gondii*. III.**

Formation of the sporozoites within the sporocysts. *Acta Pathol Microbiol Scand B* 1979, **87:**253-260.

39. Dubey JP: ***Toxoplasma gondii*** **oocyst survival under defined temperatures.** *J Parasitol* 1998, **84:**862-865.

40. Dubey JP: **Toxoplasmosis - a waterborne zoonosis.** *Vet Parasitol* 2004, **126:**57-72.

41. Darde ML, Bouteille B, Pestre-Alexandre M: **Isoenzyme analysis of 35** ***Toxoplasma gondii*** **isolates and the biological and epidemiological implications.** *J Parasitol* 1992, **78:**786-794.

42. Howe DK, Sibley LD: ***Toxoplasma gondii*** **comprises three clonal lineages: correlation of parasite genotype with human disease.** *J Infect Dis* 1995, **172:**1561-1566.

43. Sibley LD, Boothroyd JC: **Virulent strains of** ***Toxoplasma gondii*** **comprise a single clonal lineage.** *Nature* 1992, **359:**82-85.

44. Su C, Evans D, Cole RH, Kissinger JC, Ajioka JW, Sibley LD: **Recent expansion of** ***Toxoplasma*** **through enhanced oral transmission.** *Science* 2003, **299:**414-416.

45. Sibley LD, Ajioka JW: **Population structure of** ***Toxoplasma gondii*****: clonal expansion driven by infrequent recombination and selective sweeps.** *Annu Rev Microbiol* 2008, **62:**329-351.

46. Boyle JP, Rajasekar B, Saeij JP, Ajioka JW, Berriman M, Paulsen I, Roos DS, Sibley LD, White MW, Boothroyd JC: **Just one cross appears capable of dramatically altering the population biology of a eukaryotic pathogen like** ***Toxoplasma gondii*****.** *Proc Natl Acad Sci U S A* 2006, **103:**10514-10519.

47. Ajzenberg D, Banuls AL, Su C, Dumetre A, Demar M, Carme B, Darde ML: **Genetic diversity, clonality and sexuality in** ***Toxoplasma gondii*****.** *Int J Parasitol* 2004, **34:**1185-1196.

48. Ajzenberg D, Cogne N, Paris L, Bessieres MH, Thulliez P, Filisetti D, Pelloux H, Marty P, Darde ML: **Genotype of 86** ***Toxoplasma gondii*** **isolates associated with human congenital toxoplasmosis, and correlation with clinical findings.** *J Infect Dis* 2002, **186:**684-689.

49. Dubey JP, Cortes-Vecino JA, Vargas-Duarte JJ, Sundar N, Velmurugan GV, Bandini LM, Polo LJ, Zambrano L, Mora LE, Kwok OC, et al: **Prevalence of**

Toxoplasma gondii **in dogs from Colombia, South America and genetic characterization of *T. gondii* isolates.** *Vet Parasitol* 2007, **145:**45-50.

50. Dubey JP, Sundar N, Gennari SM, Minervino AH, Farias NA, Ruas JL, dos Santos TR, Cavalcante GT, Kwok OC, Su C: **Biologic and genetic comparison of *Toxoplasma gondii* isolates in free-range chickens from the northern Para state and the southern state Rio Grande do Sul, Brazil revealed highly diverse and distinct parasite populations.** *Vet Parasitol* 2007, **143:**182-188.
51. Ferreira Ade M, Vitor RW, Gazzinelli RT, Melo MN: **Genetic analysis of natural recombinant Brazilian *Toxoplasma gondii* strains by multilocus PCR-RFLP.** *Infect Genet Evol* 2006, **6:**22-31.
52. Lehmann T, Marcet PL, Graham DH, Dahl ER, Dubey JP: **Globalization and the population structure of *Toxoplasma gondii*.** *Proc Natl Acad Sci U S A* 2006, **103:**11423-11428.
53. Khan A, Fux B, Su C, Dubey JP, Darde ML, Ajioka JW, Rosenthal BM, Sibley LD: **Recent transcontinental sweep of *Toxoplasma gondii* driven by a single monomorphic chromosome.** *Proc Natl Acad Sci U S A* 2007, **104:**14872-14877.
54. Howe DK, Honore S, Derouin F, Sibley LD: **Determination of genotypes of *Toxoplasma gondii* strains isolated from patients with toxoplasmosis.** *J Clin Microbiol* 1997, **35:**1411-1414.
55. Fuentes I, Rubio JM, Ramirez C, Alvar J: **Genotypic characterization of *Toxoplasma gondii* strains associated with human toxoplasmosis in Spain: direct analysis from clinical samples.** *J Clin Microbiol* 2001, **39:**1566-1570.
56. Khan A, Su C, German M, Storch GA, Clifford DB, Sibley LD: **Genotyping of *Toxoplasma gondii* strains from immunocompromised patients reveals high prevalence of type I strains.** *J Clin Microbiol* 2005, **43:**5881-5887.
57. Beck HP, Blake D, Darde ML, Felger I, Pedraza-Diaz S, Regidor-Cerrillo J, Gomez-Bautista M, Ortega-Mora LM, Putignani L, Shiels B, et al: **Molecular approaches to diversity of populations of Apicomplexan parasites.** *Int J Parasitol* 2009, **39:**175-189.
58. Pena HF, Gennari SM, Dubey JP, Su C: **Population structure and mouse-virulence of *Toxoplasma gondii* in Brazil.** *Int J Parasitol* 2008, **38:**561-569.

59. Finlay BB, Cossart P: **Exploitation of mammalian host cell functions by bacterial pathogens.** *Science* 1997, **276:**718-725.
60. Isberg RR, Van Nhieu GT: **The mechanism of phagocytic uptake promoted by invasin-integrin interaction.** *Trends Cell Biol* 1995, **5:**120-124.
61. Morisaki JH, Heuser JE, Sibley LD: **Invasion of *Toxoplasma gondii* occurs by active penetration of the host cell.** *J Cell Sci* 1995, **108 (Pt 6):**2457-2464.
62. Keeley A, Soldati D: **The glideosome: a molecular machine powering motility and host-cell invasion by *Apicomplexa*.** *Trends Cell Biol* 2004, **14:**528-532.
63. Carruthers V, Boothroyd JC: **Pulling together: an integrated model of *Toxoplasma* cell invasion.** *Curr Opin Microbiol* 2007, **10:**83-89.
64. Sibley LD: **Intracellular parasite invasion strategies.** *Science* 2004, **304:**248-253.
65. Gaskins E, Gilk S, DeVore N, Mann T, Ward G, Beckers C: **Identification of the membrane receptor of a class XIV myosin in *Toxoplasma gondii*.** *J Cell Biol* 2004, **165:**383-393.
66. Boothroyd JC, Hehl A, Knoll LJ, Manger ID: **The surface of *Toxoplasma*: more and less.** *Int J Parasitol* 1998, **28:**3-9.
67. He XL, Grigg ME, Boothroyd JC, Garcia KC: **Structure of the immunodominant surface antigen from the *Toxoplasma gondii* SRS superfamily.** *Nat Struct Biol* 2002, **9:**606-611.
68. Jung C, Lee CY, Grigg ME: **The SRS superfamily of *Toxoplasma* surface proteins.** *Int J Parasitol* 2004, **34:**285-296.
69. Lekutis C, Ferguson DJ, Grigg ME, Camps M, Boothroyd JC: **Surface antigens of *Toxoplasma gondii*: variations on a theme.** *Int J Parasitol* 2001, **31:**1285-1292.
70. Carruthers VB, Giddings OK, Sibley LD: **Secretion of micronemal proteins is associated with *Toxoplasma* invasion of host cells.** *Cell Microbiol* 1999, **1:**225-235.
71. Soldati D, Dubremetz JF, Lebrun M: **Microneme proteins: structural and functional requirements to promote adhesion and invasion by the**

Apicomplexan parasite *Toxoplasma gondii*. *Int J Parasitol* 2001, **31:**1293-1302.

72. Cerede O, Dubremetz JF, Soete M, Deslee D, Vial H, Bout D, Lebrun M: **Synergistic role of micronemal proteins in *Toxoplasma gondii* virulence.** *J Exp Med* 2005, **201:**453-463.
73. Harper JM, Huynh MH, Coppens I, Parussini F, Moreno S, Carruthers VB: **A cleavable propeptide influences *Toxoplasma* infection by facilitating the trafficking and secretion of the TgMIC2-M2AP invasion complex.** *Mol Biol Cell* 2006, **17:**4551-4563.
74. Huynh MH, Opitz C, Kwok LY, Tomley FM, Carruthers VB, Soldati D: **Trans-genera reconstitution and complementation of an adhesion complex in *Toxoplasma gondii*.** *Cell Microbiol* 2004, **6:**771-782.
75. 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:**2082-2090.
76. Del Carmen MG, Mondragon M, Gonzalez S, Mondragon R: **Induction and regulation of conoid extrusion in *Toxoplasma gondii*.** *Cell Microbiol* 2009, **11:**967-982.
77. Mital J, Meissner M, Soldati D, Ward GE: **Conditional expression of *Toxoplasma gondii* apical membrane antigen-1 (TgAMA1) demonstrates that TgAMA1 plays a critical role in host cell invasion.** *Mol Biol Cell* 2005, **16:**4341-4349.
78. 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:**e17.
79. Lebrun M, Michelin A, El Hajj H, Poncet J, Bradley PJ, Vial H, Dubremetz JF: **The rhoptry neck protein RON4 re-localizes at the moving junction during *Toxoplasma gondii* invasion.** *Cell Microbiol* 2005, **7:**1823-1833.
80. Bradley PJ, Sibley LD: **Rhoptries: an arsenal of secreted virulence factors.** *Curr Opin Microbiol* 2007, **10:**582-587.
81. Charron AJ, Sibley LD: **Molecular partitioning during host cell penetration by *Toxoplasma gondii*.** *Traffic* 2004, **5:**855-867.
82. Mordue DG, Desai N, Dustin M, Sibley LD: **Invasion by *Toxoplasma gondii* establishes a moving junction that selectively excludes host cell**

plasma membrane proteins on the basis of their membrane anchoring. *J Exp Med* 1999, **190:**1783-1792.

83. Carruthers VB, Sibley LD: **Sequential protein secretion from three distinct organelles of *Toxoplasma gondii* accompanies invasion of human fibroblasts.** *Eur J Cell Biol* 1997, **73:**114-123.

84. Bradley PJ, Ward C, Cheng SJ, Alexander DL, Coller S, Coombs GH, Dunn JD, Ferguson DJ, Sanderson SJ, Wastling JM, Boothroyd JC: **Proteomic analysis of rhoptry organelles reveals many novel constituents for host-parasite interactions in *Toxoplasma gondii*.** *J Biol Chem* 2005, **280:**34245-34258.

85. Ossorio PN, Schwartzman JD, Boothroyd JC: **A *Toxoplasma gondii* rhoptry protein associated with host cell penetration has unusual charge asymmetry.** *Mol Biochem Parasitol* 1992, **50:**1-15.

86. Hakansson S, Charron AJ, Sibley LD: ***Toxoplasma* evacuoles: a two-step process of secretion and fusion forms the parasitophorous vacuole.** *EMBO J* 2001, **20:**3132-3144.

87. Beckers CJ, Dubremetz JF, Mercereau-Puijalon O, Joiner KA: **The *Toxoplasma gondii* rhoptry protein ROP 2 is inserted into the parasitophorous vacuole membrane, surrounding the intracellular parasite, and is exposed to the host cell cytoplasm.** *J Cell Biol* 1994, **127:**947-961.

88. Sinai AP, Joiner KA: **The *Toxoplasma gondii* protein ROP2 mediates host organelle association with the parasitophorous vacuole membrane.** *J Cell Biol* 2001, **154:**95-108.

89. Boothroyd JC, Dubremetz JF: **Kiss and spit: the dual roles of *Toxoplasma* rhoptries.** *Nat Rev Microbiol* 2008, **6:**79-88.

90. Carey KL, Jongco AM, Kim K, Ward GE: **The *Toxoplasma gondii* rhoptry protein ROP4 is secreted into the parasitophorous vacuole and becomes phosphorylated in infected cells.** *Eukaryot Cell* 2004, **3:**1320-1330.

91. Taylor S, Barragan A, Su C, Fux B, Fentress SJ, Tang K, Beatty WL, Hajj HE, Jerome M, Behnke MS, et al: **A secreted serine-threonine kinase determines virulence in the eukaryotic pathogen *Toxoplasma gondii*.** *Science* 2006, **314:**1776-1780.

92. El Hajj H, Lebrun M, Arold ST, Vial H, Labesse G, Dubremetz JF: **ROP18 is a rhoptry kinase controlling the intracellular proliferation of *Toxoplasma gondii*.** *PLoS Pathog* 2007, **3:**e14.
93. Sinai AP, Webster P, Joiner KA: **Association of host cell endoplasmic reticulum and mitochondria with the *Toxoplasma gondii* parasitophorous vacuole membrane: a high affinity interaction.** *J Cell Sci* 1997, **110 (Pt 17):**2117-2128.
94. Gilbert LA, Ravindran S, Turetzky JM, Boothroyd JC, Bradley PJ: ***Toxoplasma gondii* targets a protein phosphatase 2C to the nuclei of infected host cells.** *Eukaryot Cell* 2007, **6:**73-83.
95. Saeij JP, Coller S, Boyle JP, Jerome ME, White MW, Boothroyd JC: ***Toxoplasma* co-opts host gene expression by injection of a polymorphic kinase homologue.** *Nature* 2007, **445:**324-327.
96. Robben PM, Mordue DG, Truscott SM, Takeda K, Akira S, Sibley LD: **Production of IL-12 by macrophages infected with *Toxoplasma gondii* depends on the parasite genotype.** *J Immunol* 2004, **172:**3686-3694.
97. Suss-Toby E, Zimmerberg J, Ward GE: ***Toxoplasma* invasion: the parasitophorous vacuole is formed from host cell plasma membrane and pinches off via a fission pore.** *Proc Natl Acad Sci U S A* 1996, **93:**8413-8418.
98. Magno RC, Straker LC, de Souza W, Attias M: **Interrelations between the parasitophorous vacuole of *Toxoplasma gondii* and host cell organelles.** *Microsc Microanal* 2005, **11:**166-174.
99. Coppens I, Dunn JD, Romano JD, Pypaert M, Zhang H, Boothroyd JC, Joiner KA: ***Toxoplasma gondii* sequesters lysosomes from mammalian hosts in the vacuolar space.** *Cell* 2006, **125:**261-274.
100. Halonen SK, Weidner E: **Overcoating of *Toxoplasma* parasitophorous vacuoles with host cell vimentin type intermediate filaments.** *J Eukaryot Microbiol* 1994, **41:**65-71.
101. Luder CG, Stanway RR, Chaussepied M, Langsley G, Heussler VT: **Intracellular survival of Apicomplexan parasites and host cell modification.** *Int J Parasitol* 2009, **39:**163-173.

102. Cesbron-Delauw MF, Gendrin C, Travier L, Ruffiot P, Mercier C: ***Apicomplexa* in mammalian cells: trafficking to the parasitophorous vacuole.** *Traffic* 2008, **9:**657-664.

103. Lemgruber L, De Souza W, Vommaro RC: **Freeze-fracture study of the dynamics of *Toxoplasma gondii* parasitophorous vacuole development.** *Micron* 2008, **39:**177-183.

104. Magno RC, Lemgruber L, Vommaro RC, De Souza W, Attias M: **Intravacuolar network may act as a mechanical support for *Toxoplasma gondii* inside the parasitophorous vacuole.** *Microsc Res Tech* 2005, **67:**45-52.

105. Dubremetz JF, Achbarou A, Bermudes D, Joiner KA: **Kinetics and pattern of organelle exocytosis during *Toxoplasma gondii*/host-cell interaction.** *Parasitol Res* 1993, **79:**402-408.

106. Cesbron-Delauw MF, Guy B, Torpier G, Pierce RJ, Lenzen G, Cesbron JY, Charif H, Lepage P, Darcy F, Lecocq JP, et al.: **Molecular characterization of a 23-kilodalton major antigen secreted by *Toxoplasma gondii*.** *Proc Natl Acad Sci U S A* 1989, **86:**7537-7541.

107. Michelin A, Bittame A, Bordat Y, Travier L, Mercier C, Dubremetz JF, Lebrun M: **GRA12, a *Toxoplasma* dense granule protein associated with the intravacuolar membranous nanotubular network.** *Int J Parasitol* 2009, **39:**299-306.

108. **ToxoDB** [http://toxodb.org]

109. ***Toxoplasma gondii* Genome Project at TIGR** [http://www.tigr.org/tdb/e2k1/tga1/intro.shtml]

110. ***Toxoplasma gondii* Sequencing Project at the Wellcome Trust Sanger Institute** [http://www.sanger.ac.uk/Projects/T_gondii/]

111. Khan A, Taylor S, Su C, Mackey AJ, Boyle J, Cole R, Glover D, Tang K, Paulsen IT, Berriman M, et al: **Composite genome map and recombination parameters derived from three archetypal lineages of *Toxoplasma gondii*.** *Nucleic Acids Res* 2005, **33:**2980-2992.

112. Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, Heiges M, Iodice J, Kissinger JC, Mackey AJ, et al: **ToxoDB: an integrated *Toxoplasma gondii* database resource.** *Nucleic Acids Res* 2008, **36:**D553-556.

113. Wilkins MR: **2D electrophoresis: from protein maps to genomes.** In *First Siena conference*. Siena, Italy; 1994.

114. Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX, Gooley AA, Wilkins MR, Duncan MW, Harris R, Williams KL, Humphery-Smith I: **Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium.** *Electrophoresis* 1995, **16:**1090-1094.

115. Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, Williams KL: **Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it.** *Biotechnol Genet Eng Rev* 1996, **13:**19-50.

116. Han X, Aslanian A, Yates JR, 3rd: **Mass spectrometry for proteomics.** *Curr Opin Chem Biol* 2008, **12:**483-490.

117. Hernandez P, Binz P, Wilkins MR: **Protein Identification in Proteomics.** In *Proteome Research-Concepts, Technology and Application.* 2 edition. Edited by Wilkins MR, Appel RD, Williams KL, Hochstrasser DF. Berlin: Springer; 2007: 41-67

118. McDonald WH, Yates JR, 3rd: **Shotgun proteomics: integrating technologies to answer biological questions.** *Curr Opin Mol Ther* 2003, **5:**302-309.

119. McLafferty FW, Breuker K, Jin M, Han X, Infusini G, Jiang H, Kong X, Begley TP: **Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics.** *FEBS J* 2007, **274:**6256-6268.

120. Zubarev RA, Kelleher NL, McLafferty FW: **Electron capture dissociation of multiply charged protein cations. A nonergodic process.** *J Am Chem Soc* 1998, **120:**3265-3266.

121. Shi SD, Hemling ME, Carr SA, Horn DM, Lindh I, McLafferty FW: **Phosphopeptide/phosphoprotein mapping by electron capture dissociation mass spectrometry.** *Anal Chem* 2001, **73:**19-22.

122. Zubarev R: **Protein primary structure using orthogonal fragmentation techniques in Fourier transform mass spectrometry.** *Expert Rev Proteomics* 2006, **3:**251-261.

123. Zubarev RA, Horn DM, Fridriksson EK, Kelleher NL, Kruger NA, Lewis MA, Carpenter BK, McLafferty FW: **Electron capture dissociation for**

structural characterization of multiply charged protein cations. *Anal Chem* 2000, **72:**563-573.

124. Hicks LM, Mazur MT, Miller LM, Dorrestein PC, Schnarr NA, Khosla C, Kelleher NL: **Investigating nonribosomal peptide and polyketide biosynthesis by direct detection of intermediates on >70 kDa polypeptides by using Fourier-transform mass spectrometry.** *Chembiochem* 2006, **7:**904-907.
125. Han X, Jin M, Breuker K, McLafferty FW: **Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons.** *Science* 2006, **314:**109-112.
126. Garcia BA, Siuti N, Thomas CE, Mizzen CA, Kelleher NL: **Characterization of neurohistone variants and post-translational modifications by electron capture dissociation mass spectrometry.** *Int J Mass Spectrom* 2007, **259:**184-196.
127. Garcia BA, Joshi S, Thomas CE, Chitta RK, Diaz RL, Busby SA, Andrews PC, Ogorzalek Loo RR, Shabanowitz J, Kelleher NL, et al: **Comprehensive phosphoprotein analysis of linker histone H1 from Tetrahymena thermophila.** *Mol Cell Proteomics* 2006, **5:**1593-1609.
128. Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL: **ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry.** *Nucleic Acids Res* 2007, **35:**W701-706.
129. Siuti N, Kelleher NL: **Decoding protein modifications using top-down mass spectrometry.** *Nat Methods* 2007, **4:**817-821.
130. Parks BA, Jiang L, Thomas PM, Wenger CD, Roth MJ, Boyne MT, 2nd, Burke PV, Kwast KE, Kelleher NL: **Top-down proteomics on a chromatographic time scale using linear ion trap fourier transform hybrid mass spectrometers.** *Anal Chem* 2007, **79:**7984-7991.
131. Gorg A, Weiss W, Dunn MJ: **Current two-dimensional electrophoresis technology for proteomics.** *Proteomics* 2004, **4:**3665-3685.
132. Shapiro AL, Vinuela E, Maizel JV, Jr.: **Molecular weight estimation of polypeptide chains by electrophoresis in SDS-polyacrylamide gels.** *Biochem Biophys Res Commun* 1967, **28:**815-820.

133. Anderson NL, Anderson NG: **Proteome and proteomics: new technologies, new concepts, and new words.** *Electrophoresis* 1998, **19:**1853-1861.

134. Neuhoff V, Arold N, Taube D, Ehrhardt W: **Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie Brilliant Blue G-250 and R-250.** *Electrophoresis* 1988, **9:**255-262.

135. Dzandu JK, Johnson JF, Wise GE: **Sodium dodecyl sulfate-gel electrophoresis: staining of polypeptides using heavy metal salts.** *Anal Biochem* 1988, **174:**157-167.

136. Merril CR, Goldman D, Sedman SA, Ebert MH: **Ultrasensitive stain for proteins in polyacrylamide gels shows regional variation in cerebrospinal fluid proteins.** *Science* 1981, **211:**1437-1438.

137. Steinberg TH, Chernokalskaya E, Berggren K, Lopez MF, Diwu Z, Haugland RP, Patton WF: **Ultrasensitive fluorescence protein detection in isoelectric focusing gels using a ruthenium metal chelate stain.** *Electrophoresis* 2000, **21:**486-496.

138. Wilm M, Shevchenko A, Houthaeve T, Breit S, Schweigerer L, Fotsis T, Mann M: **Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry.** *Nature* 1996, **379:**466-469.

139. Washburn MP, Wolters D, Yates JR, 3rd: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology.** *Nat Biotechnol* 2001, **19:**242-247.

140. Wolters DA, Washburn MP, Yates JR, 3rd: **An automated multidimensional protein identification technology for shotgun proteomics.** *Anal Chem* 2001, **73:**5683-5690.

141. Karas M, Hillenkamp F: **Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.** *Anal Chem* 1988, **60:**2299-2301.

142. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM: **Electrospray ionization for mass spectrometry of large biomolecules.** *Science* 1989, **246:**64-71.

143. Li KY, Tu H, Ray AK: **Charge limits on droplets during evaporation.** *Langmuir* 2005, **21:**3786-3794.

144. Jonscher KR, Yates JR, 3rd: **The quadrupole ion trap mass spectrometer--a small solution to a big challenge.** *Anal Biochem* 1997, **244:**1-15.
145. Yates JR, 3rd: **Mass spectral analysis in proteomics.** *Annu Rev Biophys Biomol Struct* 2004, **33:**297-316.
146. Douglas DJ, Frank AJ, Mao D: **Linear ion traps in mass spectrometry.** *Mass Spectrom Rev* 2005, **24:**1-29.
147. Makarov A: **Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis.** *Anal Chem* 2000, **72:**1156-1162.
148. Yates JR, Cociorva D, Liao L, Zabrouskov V: **Performance of a linear ion trap-Orbitrap hybrid for peptide analysis.** *Anal Chem* 2006, **78:**493-500.
149. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R: **The Orbitrap: a new mass spectrometer.** *J Mass Spectrom* 2005, **40:**430-443.
150. Marshall AG, Hendrickson CL, Jackson GS: **Fourier transform ion cyclotron resonance mass spectrometry: a primer.** *Mass Spectrom Rev* 1998, **17:**1-35.
151. Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, Watanabe C: **Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases.** *Proc Natl Acad Sci U S A* 1993, **90:**5011-5015.
152. James P, Quadroni M, Carafoli E, Gonnet G: **Protein identification by mass profile fingerprinting.** *Biochem Biophys Res Commun* 1993, **195:**58-64.
153. Mann M, Hojrup P, Roepstorff P: **Use of mass spectrometric molecular weight information to identify proteins in sequence databases.** *Biol Mass Spectrom* 1993, **22:**338-345.
154. Pappin DJ, Hojrup P, Bleasby AJ: **Rapid identification of proteins by peptide-mass fingerprinting.** *Curr Biol* 1993, **3:**327-332.
155. Yates JR, 3rd, Speicher S, Griffin PR, Hunkapiller T: **Peptide mass maps: a highly informative approach to protein identification.** *Anal Biochem* 1993, **214:**397-408.
156. Shukla AK, Futrell JH: **Tandem mass spectrometry: dissociation of ions by collisional activation.** *J Mass Spectrom* 2000, **35:**1069-1090.
157. **The universal protein resource (UniProt).** *Nucleic Acids Res* 2008, **36:**D190-195.

158. Bieri T, Blasiar D, Ozersky P, Antoshechkin I, Bastiani C, Canaran P, Chan J, Chen N, Chen WJ, Davis P, et al: **WormBase: new content and better access.** *Nucleic Acids Res* 2007, **35:**D506-510.

159. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE: **The mouse genome database (MGD): new features facilitating a model system.** *Nucleic Acids Res* 2007, **35:**D630-637.

160. Grumbling G, Strelets V: **FlyBase: anatomical data, images and queries.** *Nucleic Acids Res* 2006, **34:**D484-488.

161. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2002, **30:**52-55.

162. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-370.

163. Cohen KB, Hunter L: **Getting started in text mining.** *PLoS Comput Biol* 2008, **4:**e20.

164. Zhou D, He Y: **Extracting interactions between proteins from the literature.** *J Biomed Inform* 2008, **41:**393-407.

165. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB: **Frontiers of biomedical text mining: current progress.** *Brief Bioinform* 2007, **8:**358-375.

166. Fitch WM: **Homology a personal view on some of the problems.** *Trends Genet* 2000, **16:**227-231.

167. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.

168. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.

169. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13:**2178-2189.

170. Lee D, Grant A, Marsden RL, Orengo C: **Identification and distribution of protein families in 120 completed genomes using Gene3D.** *Proteins* 2005, **59:**603-615.

171. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P: **PROSITE: a documented database using patterns and profiles as motif descriptors.** *Brief Bioinform* 2002, **3:**265-274.

172. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling.** *J Mol Biol* 1994, **235:**1501-1531.

173. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2008, **36:**D281-288.

174. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, et al: **PRINTS and its automatic supplement, prePRINTS.** *Nucleic Acids Res* 2003, **31:**400-402.

175. Mi H, Guo N, Kejariwal A, Thomas PD: **PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways.** *Nucleic Acids Res* 2007, **35:**D247-252.

176. Wilson D, Madera M, Vogel C, Chothia C, Gough J: **The SUPERFAMILY database in 2007: families and functions.** *Nucleic Acids Res* 2007, **35:**D308-313.

177. Yeats C, Maibaum M, Marsden R, Dibley M, Lee D, Addou S, Orengo CA: **Gene3D: modelling protein structure, function and evolution.** *Nucleic Acids Res* 2006, **34:**D281-284.

178. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol* 2004, **340:**783-795.

179. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305:**567-580.

180. Zuegge J, Ralph S, Schmuker M, McFadden GI, Schneider G: **Deciphering apicoplast targeting signals--feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins.** *Gene* 2001, **280:**19-26.

181. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: **WoLF PSORT: protein localization predictor.** *Nucleic Acids Res* 2007, **35:**W585-587.

182. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al: **InterPro: the integrative protein signature database.** *Nucleic Acids Res* 2009, **37:**D211-215.

183. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, et al: **IntAct--open source resource for molecular interaction data.** *Nucleic Acids Res* 2007, **35:**D561-565.

184. Rawlings ND, Morton FR, Kok CY, Kong J, Barrett AJ: **MEROPS: the peptidase database.** *Nucleic Acids Res* 2008, **36:**D320-325.

185. **Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB), Enzyme Supplement 5 (1999).** *Eur J Biochem* 1999, **264:**610-650.

186. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, **32:**D258-261.

187. Prlic A, Down TA, Hubbard TJ: **Adding some SPICE to DAS.** *Bioinformatics* 2005, **21 Suppl 2:**ii40-41.

188. Jimenez RC, Quinn AF, Garcia A, Labarga A, O'Neill K, Martinez F, Salazar GA, Hermjakob H: **Dasty2, an Ajax protein DAS client.** *Bioinformatics* 2008, **24:**2119-2121.

189. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L: **The distributed annotation system.** *BMC Bioinformatics* 2001, **2:**7.

190. Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK, Jr., Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, et al: **The minimum information about a proteomics experiment (MIAPE).** *Nat Biotechnol* 2007, **25:**887-893.

191. Deutsch E: **mzML: a single, unifying data format for mass spectrometer output.** *Proteomics* 2008, **8:**2776-2777.

192. Jones P, Cote RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H, Apweiler R: **PRIDE: a public repository of protein and peptide**

identifications for the proteomics community. *Nucleic Acids Res* 2006, **34:**D659-663.

193. Pedrioli PG, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, et al: **A common open representation of mass spectrometry data and its application to proteomics research.** *Nat Biotechnol* 2004, **22:**1459-1466.
194. Orchard S, Albar JP, Deutsch EW, Binz PA, Jones AR, Creasy D, Hermjakob H: **Annual spring meeting of the Proteomics Standards Initiative 23-25 April 2008, Toledo, Spain.** *Proteomics* 2008, **8:**4168-4172.
195. Mead JA, Bianco L, Bessant C: **Recent developments in public proteomic MS repositories and pipelines.** *Proteomics* 2009, **9:**861-881.
196. Mead JA, Shadforth IP, Bessant C: **Public proteomic MS repositories and pipelines: available tools and biological applications.** *Proteomics* 2007, **7:**2769-2786.
197. Craig R, Cortens JP, Beavis RC: **Open source system for analyzing, validating, and storing protein identification data.** *J Proteome Res* 2004, **3:**1234-1242.
198. Craig R, Beavis RC: **TANDEM: matching proteins with tandem mass spectra.** *Bioinformatics* 2004, **20:**1466-1467.
199. Craig R, Cortens JP, Beavis RC: **The use of proteotypic peptide libraries for protein identification.** *Rapid Commun Mass Spectrom* 2005, **19:**1844-1850.
200. Craig R, Cortens JC, Fenyo D, Beavis RC: **Using annotated peptide mass spectrum libraries for protein identification.** *J Proteome Res* 2006, **5:**1843-1849.
201. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R: **The PeptideAtlas project.** *Nucleic Acids Res* 2006, **34:**D655-658.
202. Keller A, Eng J, Zhang N, Li XJ, Aebersold R: **A uniform proteomics MS/MS analysis platform utilizing open XML file formats.** *Mol Syst Biol* 2005, **1:**2005 0017.
203. Deutsch EW, Lam H, Aebersold R: **PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows.** *EMBO Rep* 2008, **9:**429-434.

204. **Tranche Project** [https://trancheproject.org/]

205. Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, et al: **EGASP: the human ENCODE Genome Annotation Assessment Project.** *Genome Biol* 2006, **7 Suppl 1:**S2 1-31.

206. Allen JE, Pertea M, Salzberg SL: **Computational gene prediction using multiple sources of evidence.** *Genome Res* 2004, **14:**142-148.

207. Lu F, Jiang H, Ding J, Mu J, Valenzuela JG, Ribeiro JM, Su XZ: **cDNA sequences reveal considerable gene prediction inaccuracy in the *Plasmodium falciparum* genome.** *BMC Genomics* 2007, **8:**255.

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

209. Jungblut PR, Muller EC, Mattow J, Kaufmann SH: **Proteomics reveals open reading frames in Mycobacterium tuberculosis H37Rv not predicted by genomics.** *Infect Immun* 2001, **69:**5905-5907.

210. Choudhary JS, Blackstock WP, Creasy DM, Cottrell JS: **Interrogating the human genome using uninterpreted mass spectrometry data.** *Proteomics* 2001, **1:**651-667.

211. Kuster B, Mortensen P, Andersen JS, Mann M: **Mass spectrometry allows direct identification of proteins in large genomes.** *Proteomics* 2001, **1:**641-650.

212. Link AJ, Hays LG, Carmack EB, Yates JR, 3rd: **Identifying the major proteome components of Haemophilus influenzae type-strain NCTC 8143.** *Electrophoresis* 1997, **18:**1314-1334.

213. Jaffe JD, Berg HC, Church GM: **Proteogenomic mapping as a complementary method to perform genome annotation.** *Proteomics* 2004, **4:**59-77.

214. Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, Butler J, Calvo S, Elkins T, FitzGerald MG, Hafez N, et al: **The complete genome and proteome of Mycoplasma mobile.** *Genome Res* 2004, **14:**1447-1461.

215. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S, et al: **Integration with the**

human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol* 2005, **6:**R9.

216. Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, Ulintz P, Omenn GS, States DJ: **Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics.** *Genome Biol* 2006, **7:**R35.
217. Ansong C, Yoon H, Norbeck AD, Gustin JK, McDermott JE, Mottaz HM, Rue J, Adkins JN, Heffron F, Smith RD: **Proteomics analysis of the causative agent of typhoid fever.** *J Proteome Res* 2008, **7:**546-557.
218. Elias DA, Monroe ME, Marshall MJ, Romine MF, Belieav AS, Fredrickson JK, Anderson GA, Smith RD, Lipton MS: **Global detection and characterization of hypothetical proteins in Shewanella oneidensis MR-1 using LC-MS based proteomics.** *Proteomics* 2005, **5:**3120-3130.
219. Lipton MS, Pasa-Tolic L, Anderson GA, Anderson DJ, Auberry DL, Battista JR, Daly MJ, Fredrickson J, Hixson KK, Kostandarithes H, et al: **Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags.** *Proc Natl Acad Sci U S A* 2002, **99:**11049-11054.
220. Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP, Bafna V: **Improving gene annotation using peptide mass spectrometry.** *Genome Res* 2007, **17:**231-239.
221. Tress ML, Bodenmiller B, Aebersold R, Valencia A: **Proteomics studies confirm the presence of alternative protein isoforms on a large scale.** *Genome Biol* 2008, **9:**R162.
222. Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, Bumgarner RE, Bushel PR, Chaturvedi K, Choi D, et al: **Standardizing global gene expression analysis between laboratories and across platforms.** *Nat Methods* 2005, **2:**351-356.
223. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270:**467-470.
224. Boguski MS, Tolstoshev CM, Bassett DE, Jr.: **Gene discovery in dbEST.** *Science* 1994, **265:**1993-1994.
225. Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, Klein SL, Old S, Rasooly R, Good P, et al: **The status, quality, and**

expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res* 2004, **14:**2121-2127.

226. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270:**484-487.

227. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, et al: **CAGE: cap analysis of gene expression.** *Nat Methods* 2006, **3:**211-222.

228. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, et al: **Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.** *Nat Biotechnol* 2000, **18:**630-634.

229. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10:**57-63.

230. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**621-628.

231. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, et al: **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nat Methods* 2008, **5:**613-619.

232. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.** *Mol Ecol* 2008, **17:**1636-1647.

233. Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klingler T: **Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes.** *Genome Res* 1999, **9:**1198-1203.

234. Clare A, King RD: **How well do we understand the clusters found in microarray data?** *In Silico Biol* 2002, **2:**511-522.

235. Gallardo K, Firnhaber C, Zuber H, Hericher D, Belghazi M, Henry C, Kuster H, Thompson R: **A combined proteome and transcriptome analysis of developing *Medicago truncatula* seeds: evidence for metabolic specialization of maternal and filial tissues.** *Mol Cell Proteomics* 2007, **6:**2165-2179.

236. Williamson AJ, Smith DL, Blinco D, Unwin RD, Pearson S, Wilson C, Miller C, Lancashire L, Lacaud G, Kouskoff V, Whetton AD: **Quantitative proteomics analysis demonstrates post-transcriptional regulation of embryonic stem cell differentiation to hematopoiesis.** *Mol Cell Proteomics* 2008, **7:**459-472.

237. Beyer A, Hollunder J, Nasheuer HP, Wilhelm T: **Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale.** *Mol Cell Proteomics* 2004, **3:**1083-1092.

238. Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, Berriman M, Florens L, Janssen CS, Pain A, Christophides GK, et al: **A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses.** *Science* 2005, **307:**82-86.

239. Tarun AS, Peng X, Dumpit RF, Ogata Y, Silva-Rivera H, Camargo N, Daly TM, Bergman LW, Kappe SH: **A combined transcriptome and proteome survey of malaria parasite liver stages.** *Proc Natl Acad Sci U S A* 2008, **105:**305-310.

240. Yen HC, Xu Q, Chou DM, Zhao Z, Elledge SJ: **Global protein stability profiling in mammalian cells.** *Science* 2008, **322:**918-923.

241. Doherty MK, Hammond DE, Clague MJ, Gaskell SJ, Beynon RJ: **Turnover of the human proteome: determination of protein intracellular stability by dynamic SILAC.** *J Proteome Res* 2009, **8:**104-112.

242. Filipowicz W, Bhattacharyya SN, Sonenberg N: **Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?** *Nat Rev Genet* 2008, **9:**102-114.

243. Shock JL, Fischer KF, DeRisi JL: **Whole-genome analysis of mRNA decay in *Plasmodium falciparum* reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle.** *Genome Biol* 2007, **8:**R134.

244. Hakimi MA, Deitsch KW: **Epigenetics in *Apicomplexa*: control of gene expression during cell cycle progression, differentiation and antigenic variation.** *Curr Opin Microbiol* 2007, **10:**357-362.

245. Burg JL, Perelman D, Kasper LH, Ware PL, Boothroyd JC: **Molecular analysis of the gene encoding the major surface antigen of *Toxoplasma gondii*.** *J Immunol* 1988, **141:**3584-3591.

246. Ajioka JW, Boothroyd JC, Brunk BP, Hehl A, Hillier L, Manger ID, Marra M, Overton GC, Roos DS, Wan KL, et al: **Gene discovery by EST sequencing in *Toxoplasma gondii* reveals sequences restricted to the *Apicomplexa*.** *Genome Res* 1998, **8:**18-28.

247. Cleary MD, Singh U, Blader IJ, Brewer JL, Boothroyd JC: ***Toxoplasma gondii* asexual development: identification of developmentally regulated genes and distinct patterns of gene expression.** *Eukaryot Cell* 2002, **1:**329-340.

248. Matrajt M, Donald RG, Singh U, Roos DS: **Identification and characterization of differentiation mutants in the protozoan parasite *Toxoplasma gondii*.** *Mol Microbiol* 2002, **44:**735-747.

249. Singh U, Brewer JL, Boothroyd JC: **Genetic analysis of tachyzoite to bradyzoite differentiation mutants in *Toxoplasma gondii* reveals a hierarchy of gene induction.** *Mol Microbiol* 2002, **44:**721-733.

250. 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:**39-51.

251. Hu K, Johnson J, Florens L, Fraunholz M, Suravajjala S, DiLullo C, Yates J, Roos DS, Murray JM: **Cytoskeletal components of an invasion machine--the apical complex of *Toxoplasma gondii*.** *PLoS Pathog* 2006, **2:**e13.

252. Xia D, Sanderson SJ, Jones AR, Prieto JH, Yates JR, Bromley E, Tomley FM, Lal K, Sinden RE, Brunk BP, et al: **The proteome of *Toxoplasma gondii*: integration with the genome provides novel insights into gene expression and annotation.** *Genome Biol* 2008, **9:**R116.

253. 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:**34233-34244.

254. Lindsay DS, Toivio-Kinnucan MA, Blagburn BL: **Ultrastructural determination of cystogenesis by various *Toxoplasma gondii* isolates in cell culture.** *J Parasitol* 1993, **79:**289-292.

255. Soete M, Camus D, Dubremetz JF: **Experimental induction of bradyzoite-specific antigen expression and cyst formation by the RH strain of *Toxoplasma gondii* in vitro.** *Exp Parasitol* 1994, **78:**361-370.

256. Dlugonska H, Dytnerska K, Reichmann G, Stachelhaus S, Fischer HG: **Towards the *Toxoplasma gondii* proteome: position of 13 parasite excretory antigens on a standardized map of two-dimensionally separated tachyzoite proteins.** *Parasitol Res* 2001, **87:**634-637.

257. Kawase O, Nishikawa Y, Bannai H, Zhang H, Zhang G, Jin S, Lee EG, Xuan X: **Proteomic analysis of calcium-dependent secretion in *Toxoplasma gondii*.** *Proteomics* 2007, **7:**3718-3725.

258. Braun L, Cannella D, Pinheiro AM, Kieffer S, Belrhali H, Garin J, Hakimi MA: **The small ubiquitin-like modifier (SUMO)-conjugating system of *Toxoplasma gondii*.** *Int J Parasitol* 2009, **39:**81-90.

259. Fauquenoy S, Morelle W, Hovasse A, Bednarczyk A, Slomianny C, Schaeffer C, Van Dorsselaer A, Tomavo S: **Proteomics and glycomics analyses of N-glycosylated structures involved in *Toxoplasma gondii*--host cell interactions.** *Mol Cell Proteomics* 2008, **7:**891-910.

260. 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:**565-576.

261. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, et al: **A proteomic view of the *Plasmodium falciparum* life cycle.** *Nature* 2002, **419:**520-526.

262. Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A, Sauerwein RW, Eling WM, Hall N, Waters AP, Stunnenberg HG, Mann M: **Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry.** *Nature* 2002, **419:**537-542.

263. Khan SM, Franke-Fayard B, Mair GR, Lasonder E, Janse CJ, Mann M, Waters AP: **Proteome analysis of separated male and female gametocytes reveals novel sex-specific *Plasmodium* biology.** *Cell* 2005, **121:**675-687.

264. Snelling WJ, Lin Q, Moore JE, Millar BC, Tosini F, Pozio E, Dooley JS, Lowery CJ: **Proteomics analysis and protein expression during sporozoite excystation of *Cryptosporidium parvum* (Coccidia, *Apicomplexa*).** *Mol Cell Proteomics* 2007, **6:**346-355.

265. Sanderson SJ, Xia D, Prieto H, Yates J, Heiges M, Kissinger JC, Bromley E, Lal K, Sinden RE, Tomley F, Wastling JM: **Determining the protein**

repertoire of *Cryptosporidium parvum* sporozoites. *Proteomics* 2008, **8:**1398-1414.

266. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR, 3rd: **Direct analysis of protein complexes using mass spectrometry.** *Nat Biotechnol* 1999, **17:**676-682.

267. Bell AW, Deutsch EW, Au CE, Kearney RE, Beavis R, Sechi S, Nilsson T, Bergeron JJ, Beardslee TA, Chappell T, et al: **A HUPO test sample study reveals common problems in mass spectrometry-based proteomics.** *Nat Methods* 2009.

268. Slebos RJ, Brock JW, Winters NF, Stuart SR, Martinez MA, Li M, Chambers MC, Zimmerman LJ, Ham AJ, Tabb DL, Liebler DC: **Evaluation of strong cation exchange versus isoelectric focusing of peptides for multidimensional liquid chromatography-tandem mass spectrometry.** *J Proteome Res* 2008, **7:**5286-5294.

269. Breci L, Hattrup E, Keeler M, Letarte J, Johnson R, Haynes PA: **Comprehensive proteomics in yeast using chromatographic fractionation, gas phase fractionation, protein gel electrophoresis, and isoelectric focusing.** *Proteomics* 2005, **5:**2018-2028.

270. Dybas JM, Madrid-Aliste CJ, Che FY, Nieves E, Rykunov D, Angeletti RH, Weiss LM, Kim K, Fiser A: **Computational analysis and experimental validation of gene predictions in *Toxoplasma gondii*.** *PLoS One* 2008, **3:**e3899.

271. Doherty MK, Beynon RJ: **Protein turnover on the scale of the proteome.** *Expert Rev Proteomics* 2006, **3:**97-110.

272. Radke JR, Behnke MS, Mackey AJ, Radke JB, Roos DS, White MW: **The transcriptome of *Toxoplasma gondii*.** *BMC Biol* 2005, **3:**26.

273. Elias JE, Haas W, Faherty BK, Gygi SP: **Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations.** *Nat Methods* 2005, **2:**667-675.

274. Dunn JD, Ravindran S, Kim SK, Boothroyd JC: **The *Toxoplasma gondii* dense granule protein GRA7 is phosphorylated upon invasion and forms an unexpected association with the rhoptry proteins ROP2 and ROP4.** *Infect Immun* 2008, **76:**5853-5861.

275. Gilk SD, Gaskins E, Ward GE, Beckers CJ: **GAP45 phosphorylation controls assembly of the *Toxoplasma* myosin XIV complex.** *Eukaryot Cell* 2009, **8:**190-196.

276. Oda Y, Nagasu T, Chait BT: **Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome.** *Nat Biotechnol* 2001, **19:**379-382.

277. Zhou H, Watts JD, Aebersold R: **A systematic approach to the analysis of protein phosphorylation.** *Nat Biotechnol* 2001, **19:**375-378.

278. Zhou W, Merrick BA, Khaledi MG, Tomer KB: **Detection and sequencing of phosphopeptides affinity bound to immobilized metal ion beads by matrix-assisted laser desorption/ionization mass spectrometry.** *J Am Soc Mass Spectrom* 2000, **11:**273-282.

279. Lu B, McClatchy DB, Kim JY, Yates JR, 3rd: **Strategies for shotgun identification of integral membrane proteins by tandem mass spectrometry.** *Proteomics* 2008, **8:**3947-3955.

280. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction.** *Bioinformatics* 2001, **17 Suppl 1:**S140-148.

281. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.** *Bioinformatics* 2004, **20:**2878-2879.

282. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM: **Creating a honey bee consensus gene set.** *Genome Biol* 2007, **8:**R13.

283. Kapp EA, Schutz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ: **An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis.** *Proteomics* 2005, **5:**3475-3490.

284. Rohrbough JG, Breci L, Merchant N, Miller S, Haynes PA: **Verification of single-peptide protein identifications by the application of complementary database search algorithms.** *J Biomol Tech* 2006, **17:**327-332.

285. Kline KG, Wu CC: **MudPIT analysis: application to human heart tissue.** *Methods Mol Biol* 2009, **528:**281-293.

286. Jones AR, Siepen JA, Hubbard SJ, Paton NW: **Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines.** *Proteomics* 2009, **9:**1220-1229.
287. Gierasch LM: **Signal sequences.** *Biochemistry* 1989, **28:**923-930.
288. von Heijne G: **The signal peptide.** *J Membr Biol* 1990, **115:**195-201.
289. Joiner KA, Roos DS: **Secretory traffic in the eukaryotic parasite *Toxoplasma gondii*: less is more.** *J Cell Biol* 2002, **157:**557-563.
290. Waller RF, Keeling PJ, Donald RG, Striepen B, Handman E, Lang-Unnasch N, Cowman AF, Besra GS, Roos DS, McFadden GI: **Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*.** *Proc Natl Acad Sci U S A* 1998, **95:**12352-12357.
291. Mercier C, Adjogble KD, Daubener W, Delauw MF: **Dense granules: are they key organelles to help understand the parasitophorous vacuole of all *Apicomplexa* parasites?** *Int J Parasitol* 2005, **35:**829-849.
292. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, et al: **The genome of *Cryptosporidium* hominis.** *Nature* 2004, **431:**1107-1112.
293. Zhu G, Marchewka MJ, Keithly JS: ***Cryptosporidium parvum* appears to lack a plastid genome.** *Microbiology* 2000, **146 (Pt 2):**315-321.
294. Fichera ME, Roos DS: **A plastid organelle as a drug target in Apicomplexan parasites.** *Nature* 1997, **390:**407-409.
295. Kohler S, Delwiche CF, Denny PW, Tilney LG, Webster P, Wilson RJ, Palmer JD, Roos DS: **A plastid of probable green algal origin in Apicomplexan parasites.** *Science* 1997, **275:**1485-1489.
296. McFadden GI, Reith ME, Munholland J, Lang-Unnasch N: **Plastid in human parasites.** *Nature* 1996, **381:**482.
297. He CY, Shaw MK, Pletcher CH, Striepen B, Tilney LG, Roos DS: **A plastid segregation defect in the protozoan parasite *Toxoplasma gondii*.** *EMBO J* 2001, **20:**330-339.
298. McFadden GI, Roos DS: **Apicomplexan plastids as drug targets.** *Trends Microbiol* 1999, **7:**328-333.
299. Feagin JE, Parsons M: **The apicoplast and mitochondrion of *Toxoplasma gondii*.** In *Toxoplasma gondii: the model* **Apicomplexan***: perspectives and*

methods. First edition. Edited by Weiss LM, Kim K. London: Academic Press; 2007: 207-244

300. Bohne W, Heesemann J, Gross U: **Reduced replication of *Toxoplasma gondii* is necessary for induction of bradyzoite-specific antigens: a possible role for nitric oxide in triggering stage conversion.** *Infect Immun* 1994, **62:**1761-1767.

301. Tomavo S, Boothroyd JC: **Interconnection between organellar functions, development and drug resistance in the protozoan parasite, *Toxoplasma gondii*.** *Int J Parasitol* 1995, **25:**1293-1299.

302. Mather MW, Vaidya AB: **Mitochondria in malaria and related parasites: ancient, diverse and streamlined.** *J Bioenerg Biomembr* 2008, **40:**425-433.

303. Bender A, van Dooren GG, Ralph SA, McFadden GI, Schneider G: **Properties and prediction of mitochondrial transit peptides from *Plasmodium falciparum*.** *Mol Biochem Parasitol* 2003, **132:**59-66.

304. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S: **AmiGO: online access to ontology and annotation data.** *Bioinformatics* 2009, **25:**288-289.

305. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M, Mewes HW: **The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes.** *Nucleic Acids Res* 2004, **32:**5539-5545.

306. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.

307. Morrissette NS, Sibley LD: **Cytoskeleton of Apicomplexan parasites.** *Microbiol Mol Biol Rev* 2002, **66:**21-38; table of contents.

308. Mondragon R, Frixione E: **Ca(2+)-dependence of conoid extrusion in *Toxoplasma gondii* tachyzoites.** *J Eukaryot Microbiol* 1996, **43:**120-127.

309. Chen Z, Harb OS, Roos DS: **In silico identification of specialized secretory-organelle proteins in Apicomplexan parasites and in vivo validation in *Toxoplasma gondii*.** *PLoS ONE* 2008, **3:**e3611.

310. Heintzelman MB, Schwartzman JD: **A novel class of unconventional myosins from *Toxoplasma gondii*.** *J Mol Biol* 1997, **271:**139-146.

311. Heintzelman MB, Schwartzman JD: **Characterization of myosin-A and myosin-C: two class XIV unconventional myosins from *Toxoplasma gondii*.** *Cell Motil Cytoskeleton* 1999, **44:**58-67.

312. Heintzelman MB, Schwartzman JD: **Myosin diversity in *Apicomplexa*.** *J Parasitol* 2001, **87:**429-432.

313. Hettmann C, Herm A, Geiter A, Frank B, Schwarz E, Soldati T, Soldati D: **A dibasic motif in the tail of a class XIV Apicomplexan myosin is an essential determinant of plasma membrane localization.** *Mol Biol Cell* 2000, **11:**1385-1400.

314. Manger ID, Hehl AB, Boothroyd JC: **The surface of *Toxoplasma* tachyzoites is dominated by a family of glycosylphosphatidylinositol-anchored antigens related to SAG1.** *Infect Immun* 1998, **66:**2237-2244.

315. Sheiner L, Soldati-Favre D: **Protein trafficking inside *Toxoplasma gondii*.** *Traffic* 2008, **9:**636-646.

316. Hoppe HC, Joiner KA: **Cytoplasmic tail motifs mediate endoplasmic reticulum localization and export of transmembrane reporters in the protozoan parasite *Toxoplasma gondii*.** *Cell Microbiol* 2000, **2:**569-578.

317. Waller RF, Reed MB, Cowman AF, McFadden GI: **Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway.** *EMBO J* 2000, **19:**1794-1802.

318. Wilson RJ, Denny PW, Preiser PR, Rangachari K, Roberts K, Roy A, Whyte A, Strath M, Moore DJ, Moore PW, Williamson DH: **Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*.** *J Mol Biol* 1996, **261:**155-172.

319. Foth BJ, Ralph SA, Tonkin CJ, Struck NS, Fraunholz M, Roos DS, Cowman AF, McFadden GI: **Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*.** *Science* 2003, **299:**705-708.

320. Jomaa H, Wiesner J, Sanderbrand S, Altincicek B, Weidemeyer C, Hintz M, Turbachova I, Eberl M, Zeidler J, Lichtenthaler HK, et al: **Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs.** *Science* 1999, **285:**1573-1576.

321. Seeber F: **Biogenesis of iron-sulphur clusters in amitochondriate and Apicomplexan protists.** *Int J Parasitol* 2002, **32:**1207-1217.

322. Foth BJ, Stimmler LM, Handman E, Crabb BS, Hodder AN, McFadden GI: **The malaria parasite *Plasmodium falciparum* has only one pyruvate dehydrogenase complex, which is located in the apicoplast.** *Mol Microbiol* 2005, **55:**39-53.

323. Crawford MJ, Thomsen-Zieger N, Ray M, Schachtner J, Roos DS, Seeber F: ***Toxoplasma gondii* scavenges host-derived lipoic acid despite its de novo synthesis in the apicoplast.** *EMBO J* 2006, **25:**3214-3222.

324. Fleige T, Fischer K, Ferguson DJ, Gross U, Bohne W: **Carbohydrate metabolism in the *Toxoplasma gondii* apicoplast: localization of three glycolytic isoenzymes, the single pyruvate dehydrogenase complex, and a plastid phosphate translocator.** *Eukaryot Cell* 2007, **6:**984-996.

325. Barker AR, Wickstead B, Gluenz E, Gull K: **Bioinformatic insights to the ESAG5 and GRESAG5 gene families in kinetoplastid parasites.** *Mol Biochem Parasitol* 2008, **162:**112-122.

326. Jaouadi B, Ellouz-Chaabouni S, Rhimi M, Bejar S: **Biochemical and molecular characterization of a detergent-stable serine alkaline protease from Bacillus pumilus CBS with high catalytic efficiency.** *Biochimie* 2008, **90:**1291-1305.

327. Verleyen P, Huybrechts J, Schoofs L: **SIFamide illustrates the rapid evolution in Arthropod neuropeptide research.** *Gen Comp Endocrinol* 2009, **162:**27-35.

328. Aurrecoechea C, Brestelli J, Brunk BP, Carlton JM, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, et al: **GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens Giardia lamblia and Trichomonas vaginalis.** *Nucleic Acids Res* 2009, **37:**D526-530.

329. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, et al: **PlasmoDB: a functional genomic database for malaria parasites.** *Nucleic Acids Res* 2009, **37:**D539-543.

330. Heiges M, Wang H, Robinson E, Aurrecoechea C, Gao X, Kaluskar N, Rhodes P, Wang S, He CZ, Su Y, et al: **CryptoDB: a *Cryptosporidium* bioinformatics resource update.** *Nucleic Acids Res* 2006, **34:**D419-422.

331. Gerlt JA, Babbitt PC: **Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies.** *Annu Rev Biochem* 2001, **70:**209-246.

332. DeRocher A, Hagen CB, Froehlich JE, Feagin JE, Parsons M: **Analysis of targeting sequences demonstrates that trafficking to the *Toxoplasma gondii* plastid branches off the secretory system.** *J Cell Sci* 2000, **113 (Pt 22):**3969-3977.

333. Lal K, Prieto JH, Bromley E, Sanderson SJ, Yates JR, 3rd, Wastling JM, Tomley FM, Sinden RE: **Characterisation of *Plasmodium* invasive organelles; an ookinete microneme proteome.** *Proteomics* 2009, **9:**1142-1151.

334. Bousette N, Kislinger T, Fong V, Isserlin R, Hewel J, Emili A, Gramolini A: **Large scale characterization and analysis of the murine cardiac proteome.** *J Proteome Res* 2009.

335. Alfonso P, Nunez A, Madoz-Gurpide J, Lombardia L, Sanchez L, Casal JI: **Proteomic expression analysis of colorectal cancer by two-dimensional differential gel electrophoresis.** *Proteomics* 2005, **5:**2602-2611.

336. Bevan M, Bancroft I, Bent E, Love K, Goodman H, Dean C, Bergkamp R, Dirkse W, Van Staveren M, Stiekema W, et al: **Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of Arabidopsis thaliana.** *Nature* 1998, **391:**485-488.

337. Jiang XS, Tang LY, Dai J, Zhou H, Li SJ, Xia QC, Wu JR, Zeng R: **Quantitative analysis of severe acute respiratory syndrome (SARS)-associated coronavirus-infected cells using proteomic approaches: implications for cellular responses to virus infection.** *Mol Cell Proteomics* 2005, **4:**902-913.

338. Riley M: **Functions of the gene products of Escherichia coli.** *Microbiol Rev* 1993, **57:**862-952.

339. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.

340. Kim HJ, Baek KH, Lee SW, Kim J, Lee BW, Cho HS, Kim WT, Choi D, Hur CG: **Pepper EST database: comprehensive in silico tool for**

analyzing the chili pepper (Capsicum annuum) transcriptome. *BMC Plant Biol* 2008, **8:**101.

341. Balazsi G, Kay KA, Barabasi AL, Oltvai ZN: **Spurious spatial periodicity of co-expression in microarray data due to printing design.** *Nucleic Acids Res* 2003, **31:**4425-4433.

342. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22:**281-285.

343. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12:**1599-1610.

344. Wright JC, Sugden D, Francis-McIntyre S, Riba-Garcia I, Gaskell SJ, Grigoriev IV, Baker SE, Beynon RJ, Hubbard SJ: **Exploiting proteomic data for genome annotation and gene model validation in Aspergillus niger.** *BMC Genomics* 2009, **10:**61.

345. Carver T, Berriman M, Tivey A, Patel C, Bohme U, Barrell BG, Parkhill J, Rajandream MA: **Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database.** *Bioinformatics* 2008, **24:**2672-2676.

346. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST--database for "expressed sequence tags".** *Nat Genet* 1993, **4:**332-333.

347. Aurrecoechea C, Heiges M, Wang H, Wang Z, Fischer S, Rhodes P, Miller J, Kraemer E, Stoeckert CJ, Jr., Roos DS, Kissinger JC: **ApiDB: integrated resources for the Apicomplexan bioinformatics resource center.** *Nucleic Acids Res* 2007, **35:**D427-430.

348. Gunasekera AM, Myrick A, Le Roch K, Winzeler E, Wirth DF: ***Plasmodium falciparum*: genome wide perturbations in transcript profiles among mixed stage cultures after chloroquine treatment.** *Exp Parasitol* 2007, **117:**87-92.

349. Gunasekera AM, Patankar S, Schug J, Eisen G, Kissinger J, Roos D, Wirth DF: **Widespread distribution of antisense transcripts in the *Plasmodium falciparum* genome.** *Mol Biochem Parasitol* 2004, **136:**35-42.

350. Gunasekera AM, Patankar S, Schug J, Eisen G, Wirth DF: **Drug-induced alterations in gene expression of the asexual blood forms of *Plasmodium falciparum*.** *Mol Microbiol* 2003, **50:**1229-1239.

351. Ben Mamoun C, Gluzman IY, Hott C, MacMillan SK, Amarakone AS, Anderson DL, Carlton JM, Dame JB, Chakrabarti D, Martin RK, et al: **Co-ordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite *Plasmodium falciparum* revealed by microarray analysis.** *Mol Microbiol* 2001, **39:**26-36.

352. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*.** *PLoS Biol* 2003, **1:**E5.

353. Kidgell C, Volkman SK, Daily J, Borevitz JO, Plouffe D, Zhou Y, Johnson JR, Le Roch K, Sarr O, Ndir O, et al: **A systematic map of genetic variation in *Plasmodium falciparum*.** *PLoS Pathog* 2006, **2:**e57.

354. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6:**31.

355. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944-945.

356. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, et al: **Computational prediction of proteotypic peptides for quantitative proteomics.** *Nat Biotechnol* 2007, **25:**125-131.

357. Morrissette NS, Roos DS: ***Toxoplasma gondii*: a family of apical antigens associated with the cytoskeleton.** *Exp Parasitol* 1998, **89:**296-303.

358. Stokkermans TJ, Schwartzman JD, Keenan K, Morrissette NS, Tilney LG, Roos DS: **Inhibition of *Toxoplasma gondii* replication by dinitroaniline herbicides.** *Exp Parasitol* 1996, **84:**355-370.

359. Akerman SE, Muller S: **Peroxiredoxin-linked detoxification of hydroperoxides in *Toxoplasma gondii*.** *J Biol Chem* 2005, **280:**564-570.

360. Barlowe C: **Traffic COPs of the early secretory pathway.** *Traffic* 2000, **1:**371-377.

361. Smith SS, Pfluger SL, Hjort E, McArthur AG, Hager KM: **Molecular evolution of the vesicle coat component betaCOP in *Toxoplasma gondii*.** *Mol Phylogenet Evol* 2007, **44:**1284-1294.

362. Righetti PG, Castagna A, Antonucci F, Piubelli C, Cecconi D, Campostrini N, Antonioli P, Astner H, Hamdan M: **Critical survey of quantitative proteomics in two-dimensional electrophoretic approaches.** *J Chromatogr A* 2004, **1051:**3-17.

363. Unlu M, Morgan ME, Minden JS: **Difference gel electrophoresis: a single gel method for detecting changes in protein extracts.** *Electrophoresis* 1997, **18:**2071-2077.

364. Pal-Bhowmick I, Mehta M, Coppens I, Sharma S, Jarori GK: **Protective properties and surface localization of *Plasmodium falciparum* enolase.** *Infect Immun* 2007, **75:**5500-5508.

365. Holmes M, Liwak U, Pricop I, Wang X, Tomavo S, Ananvoranich S: **Silencing of tachyzoite enolase 2 alters nuclear targeting of bradyzoite enolase 1 in *Toxoplasma gondii*.** *Microbes Infect* 2009.

366. Ferguson DJ, Parmley SF, Tomavo S: **Evidence for nuclear localisation of two stage-specific isoenzymes of enolase in *Toxoplasma gondii* correlates with active parasite replication.** *Int J Parasitol* 2002, **32:**1399-1410.

367. McKee T, McKee JR: **Carbohydrate Metabolism.** In *Biochemistry: The molecular basis of life.* Edited by McKee T, McKee JR: McGraw-Hill; 2003: 234-271

368. Denton H, Brown SM, Roberts CW, Alexander J, McDonald V, Thong KW, Coombs GH: **Comparison of the phosphofructokinase and pyruvate kinase activities of *Cryptosporidium parvum,* Eimeria tenella and *Toxoplasma gondii*.** *Mol Biochem Parasitol* 1996, **76:**23-29.

369. Maeda T, Saito T, Oguchi Y, Nakazawa M, Takeuchi T, Asai T: **Expression and characterization of recombinant pyruvate kinase from *Toxoplasma gondii* tachyzoites.** *Parasitol Res* 2003, **89:**259-265.

370. Kitteringham NR, Jenkins RE, Lane CS, Elliott VL, Park BK: **Multiple reaction monitoring for quantitative biomarker analysis in proteomics and metabolomics.** *J Chromatogr B Analyt Technol Biomed Life Sci* 2009, **877:**1229-1239.

371. Petrak J, Ivanek R, Toman O, Cmejla R, Cmejlova J, Vyoral D, Zivny J, Vulpe CD: **Deja vu in proteomics. A hit parade of repeatedly identified differentially expressed proteins.** *Proteomics* 2008, **8:**1744-1749.

372. Dobrowolski JM, Sibley LD: ***Toxoplasma* invasion of mammalian cells is powered by the actin cytoskeleton of the parasite.** *Cell* 1996, **84:**933-939.

373. Mann T, Gaskins E, Beckers C: **Proteolytic processing of TgIMC1 during maturation of the membrane skeleton of *Toxoplasma gondii*.** *J Biol Chem* 2002, **277:**41240-41246.

374. Kolkman A, Dirksen EH, Slijper M, Heck AJ: **Double standards in quantitative proteomics: direct comparative assessment of difference in gel electrophoresis and metabolic stable isotope labeling.** *Mol Cell Proteomics* 2005, **4:**255-266.

375. Turck CW, Falick AM, Kowalak JA, Lane WS, Lilley KS, Phinney BS, Weintraub ST, Witkowska HE, Yates NA: **The Association of Biomolecular Resource Facilities Proteomics Research Group 2006 study: relative protein quantitation.** *Mol Cell Proteomics* 2007, **6:**1291-1298.

376. Wu WW, Wang G, Baek SJ, Shen RF: **Comparative study of three proteomic quantitative methods, DIGE, cICAT, and iTRAQ, using 2D gel- or LC-MALDI TOF/TOF.** *J Proteome Res* 2006, **5:**651-658.

377. Mayr E: *The Growth of Biological Thought: Diversity, Evolution, and Inheritance.* Cambridge, Massachusetts: Belknap Press of Harvard University Press; 1982.

378. Hawking SW: *A brief history of time.* Bantam Dell Publishing Group; 1988.

Appendices

Appendix I MS evidence comprising all (redundant) proteins identified using 1-DE approach

Appendix II MS evidence comprising all (redundant) proteins identified using the Tris fractionated 1-DE approach

Appendix III MS evidence comprising the complete list of non-redundant proteins identified using 1-DE approach

Appendix IV MS evidence obtained from 2-DE proteome maps of *T. gondii* tachyzoite proteins

Appendix V MS evidence comprising all (redundant) proteins identified from the MudPIT fraction

Appendix VI Supplementary Table for Figure 5.2 (Genes with proteome and transcriptome evidence in *T. gondii*)

Appendix VII Supplementary Table for Figure 5.4 (Comparison of transcriptional data and proteomic data between three *Apicomplexa*)

Appendix VIII MS evidence comprising proteins identified in the DIGE experiment of +- Glucose *T. gondii* samples

Appendices I-VIII can be found on the enclosed CD

Appendix IX Xia D, Sanderson SJ, Jones AR, Prieto JH, Yates JR, Bromley E, Tomley FM, Lal K, Sinden RE, Brunk BP, et al: **The proteome of *Toxoplasma gondii*: integration with the genome provides novel insights into gene expression and annotation.** *Genome Biol* 2008, **9:**R116.

Appendix X Sanderson SJ, Xia D, Prieto H, Yates J, Heiges M, Kissinger JC, Bromley E, Lal K, Sinden RE, Tomley F, Wastling JM: **Determining the protein repertoire of *Cryptosporidium parvum* sporozoites.** *Proteomics* 2008, **8:**1398-1414.

Appendix XI 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:**135-143.

Appendices IX-XI are attached here starting from the next page.