International Journal for Parasitology 39 (2009) 135–143

Contents lists available at ScienceDirect

International Journal for Parasitology

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

Invited Review

Proteomes and transcriptomes of the Apicomplexa – Where's the message?

J.M. Wastling [a,*], D. Xia [a], A. Sohal [b], M. Chaussepied [c], A. Pain [b], G. Langsley [c]

[a] Department of Pre-Clinical Veterinary Science, Faculty of Veterinary Science, University of Liverpool, Liverpool L69 7ZJ, UK
[b] Pathogen Genomics, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK
[c] Laboratoire de Biologie Cellulaire Comparative des Apicomplexes, Département de Maladies Infectieuses, Institut Cochin, Université Paris Descartes, CNRS (UMR 8104) and INSERM, U567 Paris, France

ARTICLE INFO

Article history:
Received 8 September 2008
Received in revised form 12 October 2008
Accepted 14 October 2008

Keywords:
Protozoa
Apicomplexa
Proteome
Transcriptome
Host cell

ABSTRACT

The Apicomplexa have some of the most comprehensive and integrated proteome datasets of all pathogenic micro-organisms. Coverage is currently at a level where these data can be used to help predict the potential biological function of proteins in these parasites, without having to defer to measurement of mRNA levels. Transcriptomic data for the Apicomplexa (microarrays, expressed sequence tag (EST) collections, serial analysis of gene expression (SAGE) and massively parallel signature sequencing (MPSS) tags) are also copious, enabling us to investigate the extent to which global mRNA levels correlate with proteomic data. Here, we present a proteomic and transcriptomic perspective of gene expression in key apicomplexan parasites, including *Plasmodium* spp., *Toxoplasma gondii*, *Cryptosporidium parvum*, *Neospora caninum* and *Theileria* spp., and discuss the alternative views of gene expression that they provide. Although proteomic evidence does not exist for every gene, many examples of readily detected proteins whose corresponding genes display little or no detectable transcription, are seen across the Apicomplexa. These examples are not easily explained by the "guilt by association", or "stock and go" hypotheses of gene transcription. With the advent of ultra-high-throughput sequencing technologies there will be a quantum shift in transcriptional analysis which, combined with improving quantitative proteome datasets, will provide a core component of a systems-wide approach to studying the Apicomplexa.

1. Introduction

The last few years has seen proteomics become established as an integral component of the functional genomics repertoire. This growth, which has resulted from fundamental technical advances in mass spectrometry (MS) and bioinformatics, has been accompanied by the emergence of numerous large-scale proteomic experiments with substantial amounts of protein expression data being deposited and displayed into increasingly sophisticated on-line proteome resources. Protozoan parasites have not been left behind in this rush for a proteomic perspective on gene expression; on the contrary, Apicomplexa now have some of the most comprehensive and integrated proteomic datasets of all pathogenic organisms. This continuing appetite for proteomic data follows the recognition that examining the proteome has the potential to reveal far more about putative function than can be accounted for by transcriptional data alone. Furthermore, there has been little slow-down in the pace of technological advances in both MS and the increasing sophistication of the bioinformatic resources that underpinned the emergence of proteomics a little over a decade ago. Importantly, these advances have resulted in a significant increase in the depth and breadth of proteomics coverage that is realistically achievable in an experiment. Whereas a few years ago whole-cell (or so-called "global") proteome surveys could do little more than sample just a small top-slice of the most represented proteins, deep-mining of the proteome is now becoming increasingly feasible and with it the ability to monitor simultaneously the expression of thousands of proteins in a biological system.

Studies on apicomplexan parasites have been especially prominent, promoting a proteomic understanding of gene expression in lower eukaryotes with large-scale proteomic surveys of *Plasmodium falciparum*, (Florens et al., 2002; Lasonder et al., 2002), *Cryptosporidium parvum* (Snelling et al., 2007; Sanderson et al., 2008) and *Toxoplasma gondii* (Xia et al., 2008) being undertaken. Apicomplexan proteomics has also benefited from a range of advances such as improved sub-fractionation of complex protein mixtures prior to analysis (Nirmalan et al., 2007), separation and analysis of apicomplexan sub-proteomes (Bradley et al., 2005; Zhou et al., 2005; Hu et al., 2006) and a strong genome bioinformatic resource populated with increasingly accurate gene models (Bahl et al., 2003; Heiges et al., 2006; Gajria et al., 2008). Proteomic studies have not only provided valuable corroborative evidence for predicted gene models by verifying the existence of thousands of hitherto hypothetical proteins, but have provided sufficient depth of coverage to begin to query the relationship between data acquired

* Corresponding author. Tel.: +44 151 794 4264; fax: +44 151 794 1515.
E-mail address: j.wastling@liv.ac.uk (J.M. Wastling).

0020-7519/$34.00
doi:10.1016/j.ijpara.2008.10.003

from transcriptional surveys, such as those from EST and microarray analysis, and actual protein expression. Such comparative surveys combining datasets from ESTs, microarray expression and proteomics have already raised fascinating questions pertaining to the link between transcription and translation in the Apicomplexa.

Despite significant advances in the accuracy and sensitivity of MS, proteomics still suffers from the disadvantage that, unlike DNA, proteins cannot be amplified to increase the sensitivity of detection. The debate therefore remains on whether current proteomic technologies can provide sufficient depth and breadth of coverage to describe fully global gene expression. However, at a time when technological gaps in proteomics seem to be rapidly closing, questions over the relative biological meaning of proteomic and transcriptomic datasets are timely and especially pertinent to apicomplexan biology. In this paper, we review advances in proteomic and transcriptional studies in the Apicomplexa, which have enabled us to examine the relationship between transcription and translation across this important group of parasites and highlight some fascinating, if not yet fully understood, discrepancies between these types of data.

Although still imperfect, proteomics does after all provide first-hand data on the functional products of gene expression – proteins, and hence their putative function. Some argue that we should even look routinely to proteomics, rather than transcriptional patterns, to give us a more meaningful picture of the biological functions of genes. Certainly, a combination of proteomics and transcriptional analysis provides a better perspective on gene expression, but these technologies are still in their infancy and we still have much to learn about the intimate and complex relationship between the two in the Apicomplexa.

2. A global proteomic perspective of the Apicomplexa

Recent global proteomic studies of apicomplexan parasites have massively increased the amount of protein expression data available for these parasites. In order to maximise the depth of coverage obtained in these analyses, a combination of specialised separation and MS approaches have been adopted. Thus, a typical experiment may involve gel-based analysis of parasite protein (one- or two-dimensional gel electrophoresis; 1-DE, 2-DE) followed by MS of trypsin-digested bands or spots. In addition, the parasite may be analysed by whole shotgun proteome analysis, commonly known as "MudPIT". Whereas gel-based analysis reveals potentially more detailed protein data in the form of semi-quantitation and some post-translational information, shotgun analysis involves the separation of digested peptides in liquid phase, thus avoiding some of the common problems associated with gel separation of hydrophobic proteins or proteins with extreme mass/p*I*. These approaches have enabled up to nearly 50% of the predicted proteome to be resolved on a proteomics platform. A summary of some of the whole-proteome projects in the Apicomplexa is presented in Table 1 and include those for *P. falciparum* (Florens et al., 2002; Lasonder et al., 2002) in which four different life cycle stages were identified using MudPIT and 1-DE LC–MS/MS. Comprehensive proteomic approaches have also been used to analyse the proteome of *Plasmodium berghei* and *Plasmodium yoelii* (Hall et al., 2005; Khan et al., 2005; Tarun et al., 2008). Thus, proteomic analysis of *Plasmodium* has resulted in one of the most comprehensive datasets for any micro-organism, with detailed proteomic coverage of up to five stages of the complex life cycle of *Plasmodium* species. These studies have been aimed at addressing important biological questions such as determining the functional characterisation of previously unknown cellular pathways (e.g. kinase pathways that regulate sex-specific functions in *Plasmodium* described by Khan et al., 2005). Doolan et al. (2003) compared genome and proteome data to identify a large number of sporozoite antigens that are highly expressed in sporozoites and showed high interferon-γ response in the peripheral blood mononuclear cells (PBMCs) of human volunteers, thus providing a list of novel candidates that could be tested as vaccine candidates. In a study which combined the transcriptome and proteome of *P. berghei*, evidence was obtained to demonstrate the developmental stage-specific translational control of mRNA transcripts which gave rise to the "stock and go" hypothesis (Mair et al., 2006). Recently Patra and co-workers (2008) undertook a study on the ookinete/zygote proteome of *Plasmodium gallinaceum*, the results of which represent a detailed proteomic view of *Plasmodium*-mosquito midgut interactions, fundamental to the development of a novel transmission blocking vaccine in malaria.

Large-scale protein expression profiling projects have also been carried out on the tachyzoite stage of *T. gondii* (Cohen et al., 2002; Xia et al., 2008) and similar approaches have been applied in investigating the proteome of *C. parvum* sporozoites (Snelling et al., 2007; Sanderson et al., 2008). These studies have identified between approximately 30–40% of the "total" predicted proteome. Further unpublished proteome data for *Toxoplasma* and *Cryptosporidium* are available via EuPathDB (http://www.eupathdb.org, formerly known as ApiDB and at http://toro.aecom.yu.edu/biodefense/ (unpublished data). More recently, proteome profiling of *Neospora caninum* has also been carried out (Wastling, unpublished data) and peptide evidence has so far been obtained for 660 of the gene models in the current set of gene predictions (available via GeneDB at http://www.genenedb.org). This number is anticipated to increase substantially in the near future.

3. Sub-proteomes of the Apicomplexa

Apicomplexan sub-proteomes have been investigated in some detail, with analysis of the apical invasive organelles leading the field. Bradley and co-workers (2005) pioneered the proteomic investigation of apicomplexan rhoptry organelles, identifying many novel components of the rhoptry and rhoptry neck of *T. gondii*, whilst other key proteins released during host-cell invasion by tachyzoites have also been characterised using 2-DE and MudPIT (Zhou et al., 2004, 2005; Fauquenoy et al., 2008). Rhoptry-enriched fractions have also been investigated in *Plasmodium* merozoites (Sam-Yellowe et al., 2004). The fractionated surface protein of parasite-infected erythrocytes of *P. falciparum* (Florens et al., 2004), the enriched cytoskeleton components of *T. gondii* (Hu et al., 2006) and the cytoskeletal and membrane fractions of both *T. gondii* and *C. parvum* have also been examined (unpublished data, http://toro.aecom.yu.edu/biodefense/).

4. Gene finding and curation in apicomplexans – a proteomic perspective

Except in a very small number of cases where protein sequence is generated by de novo protein sequencing, the quality of proteomic identification is entirely dependent on the accuracy of the gene models against which MS data are searched. Without accurately predicted gene models, proteomic experiments produce only a partial view of the proteome with considerable uncertainty surrounding the nature and number of proteins that may have been identified in any one experiment. Conversely, MS-generated peptide sequence data can be used in reverse logic as a powerful tool not only to provide confirmation, or correction of predicted protein-coding genes, but also to elucidate splicing patterns and as a key input to train gene finding algorithms (Choudhary et al., 2001; Foissac and Schiex, 2005; Fermin et al., 2006; Tanner et al., 2007; Sanderson et al., 2008; Xia et al., 2008). Many of the large-