Published in final edited form as:

J Proteome Res. 2015 November 6; 14(11): 4805–4814. doi:10.1021/acs.jproteome.5b00642.

Top Down Proteomics Reveals Mature Proteoforms Expressed in Subcellular Fractions of the *Echinococcus granulosus* Preadult Stage

Karina R. Lorenzatto[†], Kyunggon Kim^{‡,§}, Ioanna Ntai[‡], Gabriela P. Paludo[†], Jeferson Camargo de Lima[†], Paul M. Thomas[‡], Neil L. Kelleher[‡], and Henrique B. Ferreira^{†,*}

[†]Laboratório de Genômica Estrutural e Funcional, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Avenida Bento Gonçalves, 9500 Porto Alegre, Rio Grande do Sul, Brazil

[‡]Departments of Chemistry and Molecular Biosciences, and the Proteomics Center of Excellence, Northwestern University, 2145 North Sheridan Road, Evanston, Illinois 60208, United States

Abstract

Echinococcus granulosus is the causative agent of cystic hydatid disease, a neglected zoonosis responsible for high morbidity and mortality. Several molecular mechanisms underlying parasite biology remain poorly understood. Here, *E. granulosus* subcellular fractions were analyzed by top down and bottom up proteomics for protein identification and characterization of co-translational and post-translational modifications (CTMs and PTMs, respectively). Nuclear and cytosolic extracts of *E. granulosus* protoscoleces were fractionated by 10% GELFrEE and proteins under 30 kDa were analyzed by LC–MS/MS. By top down analysis, 186 proteins and 207 proteoforms were identified, of which 122 and 52 proteoforms were exclusively detected in nuclear and cytosolic fractions, respectively. CTMs were evident as 71% of the proteoforms had methionine excised and 47% were N-terminal acetylated. In addition, *in silico* internal acetylation prediction coupled with top down MS allowed the characterization of 9 proteins differentially acetylated, including histones. Bottom up analysis increased the overall number of identified proteins in nuclear and

*Corresponding Author. Address: Laboratório de Genômica Estrutural e Funcional, Centro de Biotecnologia, UFRGS, Caixa Postal 15005, 91501-970, Porto Alegre, RS, Brazil. Tel: +55 51 3308 7768. Fax: +55 51 3308 7309. henrique@cbiot.ufrgs.br.

Present Address

Department of Convergence Medicine, University of Ulsan College of Medicine & Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00642.

Figure S-1: Isolation and enrichment of subcellular fractions from *E. granulosus* protoescoleces. Figure S-2: Gene ontology (GO) classification of proteins identified in *E. granulosus* subcellular fractions. Table S-3: Low mass proteins from nuclear and cytosolic fractions reported for the first time in *E. granulosus* protoscoleces using top down approach. Table S-4: Substrates and likely composition of *E. granulosus* NAT complexes. (PDF)

Table S-1: Protein identification, CTMs and PTMs characterization of nuclear and cytosolic fractions using top down MS. (XLSX)

Table S-2: Protein identification of nuclear and cytosolic fractions using bottom up MS. (XLSX)

Table S-5: GO functional classification of proteins identified in nuclear and cytosolic fractions. (XLSX)

The authors declare the following competing financial interest(s): Qualitative protein analyses were performed in ProSightPC, a product commercialized by the Kelleher Research Group.

cytosolic fractions to 154 and 112, respectively. Overall, our results provided the first description of the low mass proteome of *E. granulosus* subcellular fractions and highlighted proteoforms with CTMs and PTMS whose characterization may lead to another level of understanding about molecular mechanisms controlling parasitic flatworm biology.

Graphical abstract

Keywords

top down proteomics; hydatid disease; low mass proteome; subcellular fractions; post-translational modifications

INTRODUCTION

The larval stage of *Echinococcus granulosus* (hydatid cyst or metacestode) is the causative agent of cystic hydatid disease or echinococcosis, which has a cosmopolitan distribution^{1–3} and is one of the 17 neglected tropical diseases prioritized by the World Health Organization (WHO) (http://www.who.int/neglected_diseases/diseases/en/). Molecular mechanisms controlling *E. granulosus* metacestode development, immune evasion and host-parasite interplay are still poorly characterized. The advent of “omics” approaches has greatly impacted the amount of genomic, transcriptomic, and proteomic data for an increasing number of parasites.⁴ The recent descriptions of the *E. granulosus* genome and transcriptome^{5–7} were undoubtedly a major accomplishment and are expected to boost postgenomic studies, and to unveil a new understanding of the biology of this tapeworm.

E. granulosus has two developmental stages in its life cycle: the adult tapeworm, which is the sexual stage, and the cystic larval metacestode, which reproduces asexually.⁸ Within the metacestode, the cyst germinal layer produces the preadult form, called protoscolex. Interestingly, *E. granulosus* protoscoleces present a remarkable developmental plasticity: they can develop into adult worms when ingested by a suitable definitive host (usually dogs), or upon release from a ruptured cyst in an intermediate host (domestic ungulates or humans), they can redifferentiate into secondary hydatid cysts.^{8,9} The molecular mechanisms behind such developmental plasticity are still poorly understood, but it is likely that they involve several levels of regulation. “Omics” approaches have identified potential *E. granulosus* regulatory molecules, including micro-RNAs, found in RNA-Seq studies,^{10,11} and regulatory proteins, such as transcription factors and signaling proteins, identified in proteomic studies.^{12–15}

Protein function regulation in different contexts can be mediated by co-translational and post-translational modifications (CTMs and PTMs, respectively), which can play critical roles in the regulation of a broad range of cellular processes in both prokaryotes and eukaryotes.^{16–19} These modifications should be regulating several biological processes in *E. granulosus* as well, but they remain largely unknown in this parasite. In fact, CTMs and PTMs have been poorly addressed in parasitic flatworms in general, and their characterization may lead to another level of understanding about molecular mechanisms controlling parasitic flatworm biology.

CTM and PTM characterization is not a trivial task, and the availability of tools to achieve complete mapping of all expressed proteins variants (proteoforms) is limited. The term proteoform is currently used to designate all molecular forms in which the protein product of a single gene can be found, including changes due to genetic variation, alternatively spliced RNA transcripts, CTMs, and PTMs.²⁰ Mass spectrometry (MS)-based proteomics has evolved dramatically and is now considered a key technology for identifying and quantifying proteins and PTMs.^{21–24} The bottom up workflow allows high-throughput protein identification, but the assignment of PTM sites is compromised due to the MS dynamic range and the typical low sequence coverage obtained. Peptide enrichment decreases this problem and allows the shotgun approach to be applied to the identification and quantification of PTMs.²³ However, bottom up MS still faces limitations for combinatorial detection of PTMs, unless the modifications are colocated on one peptide. Top down proteomics, on the other hand, is based on the analysis of intact proteins, and is well-suited to the characterization of proteoforms, as it can detect proteins with CTMs and PTMs in a general fashion and can produce a faithful representation of protein variants resulting from maturation within cells.²⁵ Top down MS is not as widely spread as bottom up MS, but recent efforts have enabled its use in a high-throughput fashion.^{26–28}

Here, for the first time, the low mass proteome from subcellular fractions from *E. granulosus* protoscoleces has been addressed. A top down coupled to a bottom up proteomic approach has been used to not only identify, but also characterize, the low mass proteome, highlighting the detection of CTMs in proteoforms expressed in *E. granulosus* larval stage. Our findings revealed new molecular aspects of *E. granulosus*, adding another level of understanding in the biology of this neglected parasite.

MATERIAL AND METHODS

Parasite Material

E. granulosus hydatid cysts were obtained from lungs of cattle slaughtered in a local abattoir (Cooperleo, São Leopoldo, Brazil). Animal slaughtering was conducted according to Brazilian laws and under supervision of the *Serviço de Inspeção Federal* (Brazilian Sanitary Authority) of the Brazilian *Ministério da Agricultura, Pecuária e Abastecimento*. Condemned viscera at post-mortem meat inspection due to the presence of hydatid cysts were collected at the abattoir and dissected in the laboratory, following described protocols,¹³ all approved by the Ethical Committee of the *Universidade Federal do Rio Grande do Sul*. Protoscoleces were collected by aspiration, washed several times in PBS, and kept frozen until genotype confirmation. Parasite genotyping was performed for each

individual cyst, as previously described,²⁹ and all used parasite samples were from the G1 genotype, defined as *E. granulosus sensu stricto* (*E. granulosus* s. s.).³⁰

Protoscolex nuclear and cytosolic proteins were extracted following the protocol described previously,³¹ with some modifications. Briefly, 500 mg of protoscoleces from 9 lung cysts of the G1 genotype was pooled out and homogenized with 3 vol of buffer N (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl₂, 0.5% Nonidet P-40) containing Halt protease and phosphatase inhibitors (Thermo Scientific), on ice, on a Tekmar homogenizer, using 10 bursts of 10 s each and 1 min interval between bursts (at a setting of 60–100). After centrifugation for 5 min at 4000g, the pellet was suspended, homogenized, and centrifuged again under the same conditions. The two supernatants generated in these centrifugation steps were combined, and an aliquot was stained with DAPI and checked on FLoid Cell Imaging Station (Life Technologies), to verify nucleus integrity. Then, this preparation was centrifuged using Type 50.2 Ti rotor (Beckman Coulter) at 100 000g for 1 h at 4 °C, for nuclei and cytosolic proteins isolation. The resulting supernatant enriched in cytosolic proteins (cytosolic fraction) was saved, and the pellet containing nuclei and, therefore, enriched in nuclear proteins (nuclear fraction) was suspended using 1 vol of buffer F (50 mM HEPES, pH 7.5, 5 mM KCl, 5 mM MgCl₂, 0.1 mM EDTA, 0.5 mM DTT) containing Halt protease and phosphatase inhibitors. Subcellular enrichment of fractions was assessed by Western blot. Nuclear and cytosolic fractions were both lyophilized and stored at –20 °C until use.

Sample Preparation

Nuclear and cytosolic lyophilized proteins were reconstituted in 4% SDS, acetone precipitated with 6 vol of cold acetone, and centrifuged at max speed for 10 min. The precipitated pellets were reconstituted in 1% SDS, and protein concentration was determined using BCA assay (ThermoFisher, Rockford, IL). For Gel Elution Liquid-based Fractionation Entrapment Electrophoresis (GELFrEE), 250 µg of protein was fractionated in a GELFrEE 8100 Fractionation system (Expedeon, San Diego, CA) using 10% cartridges. Fractions 1–6 were collected, and a 6 µL aliquot of each 150 µL fraction was used for conventional SDS-PAGE analysis and silver stain visualization.

For intact protein preparation, fractions were precipitated with MeOH/CH₃Cl/H₂O to remove SDS.³² Prior to MS analysis, the samples were resuspended in 20–30 µL of buffer A (95% H₂O, 5% acetonitrile, 0.2% formic acid). For peptide preparation, fractions were precipitated in acetone and pellets were resuspended in 8 M urea in 100 mM NH₄HCO₃, followed by reduction with 20 mM DTT for 30 min at room temperature, and cysteine alkylation with 40 mM iodoacetamide for an additional 30 min at room temperature, in the dark. Protein mixtures were digested with trypsin (Trypsin Gold, Mass Spectrometry grade, Promega), at an enzyme/substrate ratio of 1:50, at 37 °C for 18 h. Peptides were desalted using Zip-Tip C18 (Millipore), dried in a SpeedVac and suspended in 30 µL of buffer A.

LC–MS/MS of Intact Proteins

A 6 µL vol of each GELFrEE fraction was injected in triplicate onto a trap column (150 µm i.d. × 2 cm) and eluted onto an analytical column (75 µm i.d. × 15 cm) and online separated

as described previously.³³ Both trap and analytical columns were packed in-house³⁴ with polymeric reverse phase (PLRP-S, Phenomenex) media (5 μ m, 1000 Å pore size) and connected to 15 μ m nanoelectrospray tips (New Objective, Waltham, MA). Samples were loaded onto the trap using an Ultimate 3000 RPLCnano system (Thermo Scientific Dionex) at a flow rate of 2.5 μ L/min, and then separated on the analytical column and eluted into the mass spectrometer using a flow rate of 300 nL/min and the following gradient: 5% B at 0 min, 10% B at 2 min, 55% B at 52 min, 95% B from 57 to 62 min, 5% B from 65 to 80 min. Buffer A consisted of 95% H₂O, 5% acetonitrile, and 0.2% formic acid, and buffer B consisted of 5% H₂O, 95% acetonitrile, and 0.2% formic acid. Mass spectrometry data were obtained on an Orbitrap Elite (Thermo Scientific). The MS method included the following events: (i) FT scan, four microscans, m/z 500–2,000, and resolution 100 000; and (ii) data-dependent MS/MS on the top two peaks in each spectrum from scan event using higher-energy collisional dissociation (HCD) with normalized collision energy of 25, isolation width 50 m/z , four microscans, and detection of ions with resolving power of 60 000. Dynamic exclusion was enabled with a repeat count of 2, a repeat duration of 120 s, and an exclusion duration of 5000 s. Automatic gain control (AGC) was set to 1E6 ions, and maximum injection time was set to 1 s for both MS1 and MS2. Advanced signal processing was turned on, and data were collected in reduced profile mode. A 15 V offset in the source was used over the entire experiment. The capillary temperature was 320 °C, and the spray voltage was 1.8 kV.

LC–MS/MS of Tryptic Peptides

A 10 μ L volume of each digested GELFrEE fraction was injected in duplicate onto a trap column (150 μ m i.d. $\times$ 2 cm) and eluted onto an analytical column (75 μ m i.d. $\times$ 15 cm). Both trap and analytical columns were packed in-house with 3 μ m Aqua C18 (Phenomenex) and connected to 15 μ m nanoelectrospray tips (New Objective, Waltham, MA). Samples were loaded onto the trap using an Ultimate 3000 RPLCnano system at a flow rate of 3 μ L/min, and then separated on the analytical column and eluted into the mass spectrometer using a flow rate of 300 nL/min and the following gradient: 3% B at 0 min, 8% B at 2 min, 38% B at 50 min, 55% B at 55 min, 90% B from 60 to 64 min, 4% B from 65 to 80 min. Data acquisition on the 7 T LTQ-FT-ICR (Thermo Scientific) consisted of a full scan event, m/z 400–1600 and resolution 50 000 and data-dependent MS/MS on the top eight peaks in each spectrum from scan event 1 using collision-induced dissociation (CID) with normalized collision energy of 35 and isolation width 2 m/z . Dynamic exclusion was enabled with a repeat count of 3, a repeat duration of 45 s, and an exclusion duration of 120 s. The capillary temperature was 200 °C, and the spray voltage was 2.4 kV.

Data Processing

Top down analysis was performed essentially as described previously.²⁷ Raw data files were processed in cRAWler, an in-house software which determines intact and fragment masses and compiles them in ProSight Upload Format (*.puf) files. At first, ProSightPC 3.0 (Thermo Scientific, San Jose, CA) was used to search each *.puf file against *E. granulosus* databases containing deduced amino acid sequences from genome annotation, downloaded from GeneDB⁵ (April 29, 2013) and UniProtKB⁶ (June 2, 2014). An iterative absolute mass search tree was implemented in ProSightPC, starting with a narrow search window of 10 Da

of intact mass tolerance. If the first search failed to find matches, it was followed by a second search using a wide search window of 100 000 Da. The wide search detect hits with PTMs that are not annotated in the database as well as those cases with large differences between theoretical and observed intact mass values. In both cases, a 15 ppm mass tolerance was used for the fragment ions. Low confidence proteoform identifications were excluded by considering only those hits arising from an absolute mass search with an E-value below 1×10^{-4} .

The amino acid sequences of proteins identified in the first search were submitted to an *in silico* prediction of lysine acetylation using the web tool PAIL (<http://bdmpail.biocuckoo.org/prediction.php>). For that, a database including the predicted acetylation events was built and used in a second round of searches on ProSightPC, applying the same parameters set in the first search.

For bottom up analysis, raw data files were processed on COMPASS.³⁵ Peak lists were searched using the MASCOT search engine (Matrix Science, London, U.K.; version 2.3.0) against the *E. granulosus* databases (previously described in this section). Database searches were performed with the following parameters: one missed cleavage; fixed modifications, carbamidomethylation of cysteine; variable modifications, oxidation of methionine; parent ion tolerance, 20 ppm; fragment ion tolerance, 0.6 Da; peptide charges, +2, +3 and +4. To validate MS/MS based peptide and protein identifications, MASCOT *.dat files were loaded on Scaffold (Proteome Software, Portland, OR; version 3.6) and parameters to accept an identification were as follows: peptide probability >95.0%, protein probability >99.0%, and 2 unique peptides.

Functional Classification

Online BLASTP searches were performed using BLAST2GO³⁶ version 3.0 against the NCBI nr database. GO mapping, InterProScan analysis and annotation were performed based on BlastP results (E-value 1.0×10^{-3}). Bar charts were generated using all available GO subcategories of level 3.

RESULTS AND DISCUSSION

Top Down MS Identification of Proteins and Proteoforms

To survey *E. granulosus* s.s. mature proteoforms present in subcellular fractions, nuclei and cytoplasm enriched fractions from protoscoleces were produced. Cells were lysed by a mild treatment, and nucleus integrity was visually inspected (Figure S-1A). Subcellular enrichment of nuclear and cytosolic fractions was assessed by Western blot using antibodies against specific marker proteins (Figure S-1B). Chromodomain helicase was mostly detected in nuclear fraction, whereas the tropomyosin band was less intense in this fraction in comparison with the cytosolic fraction, indicating that the subcellular fractionation allowed the enrichment of nuclear and cytosolic proteins in the corresponding fractions (Figure S-1B). Proteins from nuclear and cytosolic fractions were then resolved by GELFrEE and submitted to top down MS. Representative results of LC separation from a GELFrEE fraction are illustrated in Figure 1A,B. Overall, 186 protein IDs were identified and some of

them appeared in different proteoforms, resulting in the identification of 207 proteoforms (Table S-1A,B). The distribution of exclusive and shared protein IDs and proteoforms between nuclear and cytosolic fractions is shown in Figure 2. There were 122 and 52 proteoforms exclusively detected in nuclear and cytosolic fractions, respectively. Interestingly, the number of proteoforms shared between these fractions is smaller than the number of shared protein IDs. This is explained by the fact that seven protein IDs shared by both subcellular fractions have corresponding proteoforms exclusive of the nuclear or the cytosolic fraction.

Ribosomal proteins (6–29 kDa) and dyneins (9–23 kDa) were detected in several GELFrEE fractions. This is a common result for top down proteomics, due to the high cellular abundance and low molecular weight of these proteins. A perfect separation is not achievable due to overlaps between fractions and also to the occurrence of proteolytic events, both natural or artifactual.²⁵ In addition, the standard mode of data acquisition used here has also led to overlaps in the identification of highly abundant proteins in subsequent LC–MS injections, and the development of new methods is expected to circumvent the overlapping.³⁷ Despite that, less abundant proteins exhibiting confident identifications were also detected, as those shown as examples in Figure 1C.

To compare the sensitivity of the top down approach with that of bottom up proteomics, the same nuclear and cytosolic GELFrEE fractions were also analyzed by a bottom up workflow (Table S-2A–D). In the bottom up analysis, 208 proteins were confidently identified with at least two unique peptides. Of these proteins, 56 were also identified in the top down approach (Figure 3A). There were 152 proteins detected only in the bottom up approach, but most of them have molecular masses larger than 30 kDa (Figure 3B), which makes them less amenable in our top down analysis. There were 130 proteins uniquely identified in the top down approach, and 34 of these proteins have molecular weights below 10 kDa, contrasting with only 4 proteins uniquely identified by bottom up in this same mass range (Figure 3B). Characterization of small proteins represents a challenge for the bottom up workflow due to the inability to generate sufficient tryptic peptides for analysis.³⁸ The ability of top down for low mass protein detection enabled the identification of 25 proteins not previously reported, increasing the repertoire of proteins known to be expressed in *E. granulosus* protoscoleces (Table S-3).

Characterization of Identified Proteoforms

The characterization of CTMs and PTMs is one of the enduring goals of top down proteomics. To accomplish that in a high-throughput workflow, large combinations of PTMs can be annotated and considered within databases.²⁶ For the *E. granulosus* genome, the currently available annotations^{5,6} were solely based on DNA sequencing, and no PTMs have been annotated so far. Even in cases like this, top down can provide the detection of mature proteoforms. ProSightPC can presume a theoretic N-terminal processing, including methionine excision and N-terminal acetylation, enabling the detection of co-translationally modified proteoforms.^{39,40}

The most prevalent types of CTMs identified in nuclear and cytosolic proteoforms from *E. granulosus* protoscoleces in our top down approach were N-terminal methionine excision

(NME) and N-terminal acetylation (NTA). NME was detected in 71% (130 out of 182) of the identified proteoforms (Table S-1C), which is in agreement with the thought that up to 80% of the proteins of any given proteome undergo this modification.⁴¹ In the case of NTA, 47% (85 out of 182) of the proteoforms were found to contain this type of processing (Table S-1D), which is just under the lower limit of the range (50–80%) usually found in other eukaryotes.⁴² This is the first time that NME and NTA are described for *E. granulosus*. However, the detected percentages of proteoforms presenting NMEs and NTAs are likely to be underrepresentations of the actual fraction of proteoforms undergoing these co-translational modifications in *E. granulosus* protoscoleces, as proteoforms with no N-terminal fragment ions were not included in the performed analyses.

NME is an essential process that occurs in the cytoplasm of all organisms and in the two organelles displaying protein synthesis.⁴³ NME involves two types of methionine aminopeptidase (MAP), MAP1 (type-I) and MAP2 (type-II), which have similar three-dimensional structures, despite having only low levels of sequence identity. Higher eukaryotes have at least one MAP1 and one MAP2 in the cytoplasm and one MAP1 in the organelles.⁴⁴ *E. granulosus* genome has at least two genes coding putative MAP1s (UniProt accessions: U6J5U8 and A0A068X0Z5) and one coding a putative MAP2 (UniProt accession: A0A068W7Q0). According to the generally accepted cleavage rule for MAPs (N-end rule), these enzymes remove the N-terminal methionine of all polypeptide chains with a small side chain on their second residues,⁴⁴ although many exceptions are known.⁴⁵ For, *E. granulosus* protoscoleces proteoforms with NME, it was found that most (86%) of them followed the N-end rule, as they presented Ala, Cys, Gly, Pro, Ser, Thr, or Val as second N-terminal residue. The other 14% of potential MAP substrates presented Arg, Asp, Glu, Gln, Ile, Leu, Lys, or Trp as second N-terminal residue. These detected substrates that are not in agreement with the N-end rule may represent exceptions, also found in other organisms.⁴⁵ Future studies will be required for the evaluation of *E. granulosus* MAP activities and substrate specificity.

NTA is another co-translational modification conserved among eukaryotes,⁴⁶ and they are performed by a group of N-terminal acetyltransferases (NATs) complexes, NatA–E. In eukaryotes, these Nat complexes differ both in subunit composition as well as in substrate specificity, and, combined, act on most proteins.⁴² Higher eukaryotes also express NatF, which, due to its distinct substrate preference, is responsible for an overall increase in protein acetylation.⁴⁷ At least 11 genes in the *E. granulosus* genome^{5,6} code for proteins with homology to eukaryotic Nat complex subunits, which is suggestive of the presence of Nat complexes A–F in this parasite.

NTA events are mostly performed by NatA on MAP newly generated N-termini.⁴² This seems to be the case for *E. granulosus* as well, since most (60 out of 85) of the NTA detected events occurred in proteins previously processed by MAP (Figure 4A), and Ser, Ala, and Thr are the top three residues found in the second N-terminal position (Figure 4B). N-termini with larger amino acid residues in the second N-terminal position (such as Asp, Asn, or Gln) are not normally cleaved by MAPs, and potentially acetylated directly on the initiator methionine by NatB, C, and E, depending on the N-terminal sequence.¹⁸ Putative *E. granulosus* NatB, C, and E substrates seem to be at least partially conserved (Figure 4B). *E.*

granulosus NATs possible composition and substrates here detected are shown in Table S-4. The detected N-acetylated sites further support the occurrence and activity of all Nat complexes (A–F) in *E. granulosus*.

While N-terminal processing can be more easily predicted from genomic and proteomic data, the prediction of internal PTMs is undoubtedly a much more difficult task. In the case of organisms with annotated PTMs (human and yeast, for example), the database of possible proteoforms is generated by combinatorial expansion of all potential proteoforms using a modified UniProtKB flat file.^{39,40,48} The size of the database is controlled by setting the maximum features per sequence and 13 is used as default setting, as rarely a protein has more than 13 known modifications.⁴⁰ For *E. granulosus*, however, there are so far no experimental or *in silico* prediction of internal PTMs to allow their more efficient detection by MS. In addition, although we have utilized the same *E. granulosus* s.s. genotype (G1) used for genome sequencing,^{5,6} the known variability among *E. granulosus* s.s. isolates required a nonstringent tolerance to differences between theoretical and experimental masses, hindering the attempts for manual detection of common PTMs based on mass shifts.

In a first attempt to investigate *E. granulosus* internal PTMs, acetylation was *in silico* predicted for all *E. granulosus* nuclear and cytosolic proteoforms identified by MS. The number of predicted modifications extrapolated the default setting of maximum features per sequence. As its increase would make data storage and search time prohibitive, the database of *E. granulosus* predicted proteoforms including acetylation was generated in a modified UniProtKB flat file and keeping the default setting for combinatorial expansion. The impossibility to consider all predicted modifications in a combinatorial fashion has probably limited the detection of modifications, but even with this limitation, the use of a modified database allowed the detection of 9 acetylated proteoforms (Table S-1E), of which 4 are differentially acetylated histones (Figure 5). As the detection of these 9 acetylated proteoforms was database-dependent, we can anticipate that the use, in the future, of a more comprehensive database, including additional *in silico* and experimental data, will significantly improve the identification of *E. granulosus* post-translationally modified proteoforms.

Functional Classification of the Low Mass Proteome of Nuclear and Cytosolic Fractions from *E. granulosus* Protoscoleces

The combination of top down and bottom up approaches resulted in an overall identification of 338 proteins, from which 154 and 112 were exclusively identified in nuclear and cytosolic fractions, respectively (Figure 6A). GO term enrichment analysis was used to classify the functions of the identified proteins, and sequence annotations associated with 280 proteins (83%) were categorized into the three main GO categories: “cellular component” (Figure 6B), and “molecular function” and “biological process” (Figure S-2). The GO terms were summarized into 72, 84, and 73 subcategories for nuclear, cytosolic, and shared proteins, respectively, with GO classifications set at level 3 (Table S-5).

Functional enrichment of the identified proteins revealed that nuclear and cytosolic fractions were enriched in proteins with annotations consistent with nuclear and cytosolic functions, respectively, which points out to the effectivity of the protoscolex subcellular enrichment. In

nuclear fractions, the “structural constituent of ribosome” subcategory from “molecular function” category (Figure S-2A) points out to the enrichment of ribosomal proteins, which are among the major components of the nucleolus. In addition, nuclear proteins assigned to subcategories as “cell part” and “organelle part”, both from the “cellular component” category (Figure 6B), were enriched in nucleosome and spliceosome proteins. In cytosolic fraction, on the other hand, in the “molecular function” category, subcategories as “carbohydrate derivative binding”, “transferase activity” and “protein binding”, with proteins associated with the glycolytic pathway, to oxidative stress and stress response and/or to refolding, respectively, were enriched.

Functional enrichment analysis also pointed out the presence of some less usual subcategories in the nuclear and cytosolic fractions, which are less represented and may be result of subcellular contamination and/or protein shuttling. For instance, the “membrane-bounded organelle” subcategory of the nuclear fraction, besides nuclear envelope proteins, includes also membrane proteins of non-nuclear organelles, such as endoplasmic reticulum (e.g., translocon-associated protein) and mitochondria (e.g., NADH dehydrogenase). The presence of non-nuclear organelle proteins in the nuclear fraction can be explained by the high speed centrifugation step of 100 000g and have been previously described as major sources of contamination in nuclear preparations.⁴⁹ On the other hand, the detection of typical cytosolic proteins (e.g., eukaryotic elongation factor and actin) in the nuclear fraction, as well as the detection of some nuclear proteins (e.g., histone H2B and cyclin) in cytosolic fractions are more likely explained as cases of shuttling between the nucleus and cytosol or vice versa. For these and other proteins, there are previous results showing that they may shuttle between cellular compartments and show moonlighting functions when in ectopic locations and/or in association with other protein partners.^{50–54}

There were 58 proteins unassigned to GO terms and most of them remain annotated as uncharacterized proteins.^{5,6} Interestingly, 24 of these proteins were detected exclusively in the nuclear fractions, and 6 of them (UniProt: A0A068WLY9, A0A068WFZ9, U6JNT9, A0A068WT65, A0A068WEA7, A0A068WUA4) appear to be restricted to the *Echinococcus* genus, since ortholog proteins were found only in *Echinococcus multilocularis*. In the cytosolic fraction, just 1 out of 10 uncharacterized proteins (UniProt: A0A068W6R5) also appears restricted to the *Echinococcus* genus, and, among the proteins shared between nuclear and cytosolic, just 1 out of 13 (UniProt: A0A068WW55) seems to be exclusive from *E. granulosus* s. s. The group of nuclear, cytosolic and shared uncharacterized proteins also includes, respectively, 3 (UniProt: U6JF24, A0A068W906, A0A068X009), 1 (UniProt: A0A068WUI4), and 2 (UniProt: U6JBE2, U6JCZ1) that have orthologs limited to tapeworms; and 2 (UniProt: A0A068X2D5, A0A068WLV4), 1 (UniProt: A0A068WNL1), and 2 (UniProt: A0A068X046, A0A068X0E4) with orthologs limited to helminths. These proteins so far exclusively found in the genus *Echinococcus* or restricted to tapeworms or helminths may carry novel domains likely linked to unrecognized nuclear and cytosolic functions restricted to these organisms, making them intriguing targets to be explored in future studies.

CONCLUSIONS

Top down proteomics allowed a first step toward the characterization of PTMs in *E. granulosus* proteins. The proteoforms detected here support that N-terminal processing is conserved in *E. granulosus* proteins and indirectly suggest the functionality of MAPs and NATs in this parasite. Our modified database also provided the identification of a few internal acetylated sites, allowing the detection of some mature acetylated proteoforms. Moreover, the use of a bottom up approach allowed the identification of IDs not detected using top down MS, adding in the coverage of protein profiles expressed in *E. granulosus* subcellular fractions. Future studies using complementary approaches, exploring not only how *E. granulosus* proteins are decorated, but also addressing their differential occurrence, impact in subcellular location, and specific functions in biological processes, can lead us to unravel further levels of complexity in molecular mechanisms of *E. granulosus* and other related parasites.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Grant 23038.005300/2011-14) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Grant 477649/2012-2) to H.B.F. and by Award Nos. R01GM067193 and P41GM108569 from the National Institute of General Medical Sciences (N.L.K.). K.R.L. was recipient of CNPq and CAPES Ph.D. fellowships (141812/2011-7 and 12676/13-6). G.P.P. was recipient of a CNPq undergraduate fellowship (164139/2012-5). J.C.L. was recipient of a CNPq technical support fellowship (377742/2013-9).

ABBREVIATIONS

CID	collisionally induced dissociation
CTM	co-translational modifications
HCD	higher energy collisional dissociation
MAP	methionine aminopeptidase
MS	mass spectrometry
NAT	N-terminal acetyltransferases
NME	N-terminal methionine excision
NTA	N-terminal acetylation
PTM	posttranslational modification
PUF	ProSight Upload Format
SDS	sodium dodecyl sulfate

REFERENCES

1. Budke CM, Deplazes P, Torgerson PR. Global socioeconomic impact of cystic echinococcosis. *Emerging Infect. Dis.* 2006; 12:296–303. [PubMed: 16494758]
2. Moro P, Schantz PM. Echinococcosis: a review. *Int. J. Infect. Dis.* 2009; 13:125–133. [PubMed: 18938096]
3. Carmena D, Cardona GA. Echinococcosis in wild carnivorous species: epidemiology, genotypic diversity, and implications for veterinary public health. *Vet. Parasitol.* 2014; 202:69–94. [PubMed: 24698659]
4. Thompson RC, Lymbery AJ. Let's not forget the thinkers. *Trends Parasitol.* 2013; 29:581–584. [PubMed: 24211214]
5. Tsai IJ, Zarowiecki M, Holroyd N, Garcarrubio A, Sanchez-Flores A, Brooks KL, Tracey A, Bobes RJ, Fragoso G, Sciutto E, Aslett M, Beasley H, Bennett HM, Cai J, Camicia F, Clark R, Cucher M, De Silva N, Day TA, Deplazes P, Estrada K, Fernández C, Holland PW, Hou J, Hu S, Huckvale T, Hung SS, Kamenetzky L, Keane JA, Kiss F, Koziol U, Lambert O, Liu K, Luo X, Luo Y, Macchiaroli N, Nichol S, Paps J, Parkinson J, Pouchkina-Stantcheva N, Riddiford N, Rosenzvit M, Salinas G, Wasmuth JD, Zamanian M, Zheng Y, Cai X, Soberón X, Olson PD, Lacleite JP, Brehm K, Berriman M, Consortium TsG. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature.* 2013; 496:57–63. [PubMed: 23485966]
6. Zheng H, Zhang W, Zhang L, Zhang Z, Li J, Lu G, Zhu Y, Wang Y, Huang Y, Liu J, Kang H, Chen J, Wang L, Chen A, Yu S, Gao Z, Jin L, Gu W, Wang Z, Zhao L, Shi B, Wen H, Lin R, Jones MK, Brejova B, Vinar T, Zhao G, McManus DP, Chen Z, Zhou Y, Wang S. The genome of the hydatid tapeworm *Echinococcus granulosus*. *Nat. Genet.* 2013; 45:1168–1175. [PubMed: 24013640]
7. Parkinson J, Wasmuth JD, Salinas G, Bizarro CV, Sanford C, Berriman M, Ferreira HB, Zaha A, Blaxter ML, Maizels RM, Fernández C. A transcriptomic analysis of *Echinococcus granulosus* larval stages: implications for parasite biology and host adaptation. *PLoS Neglected Trop. Dis.* 2012; 6:e1897.
8. Smyth JD. Parasites as biological models. *Parasitology.* 1969; 59:73–91. [PubMed: 5769500]
9. Thompson RC, Jenkins DJ. *Echinococcus* as a model system: biology and epidemiology. *Int. J. Parasitol.* 2014; 44:865. [PubMed: 25123067]
10. Cucher M, Prada L, Mourglia-Ettlin G, Dematteis S, Camicia F, Asurmendi S, Rosenzvit M. Identification of *Echinococcus granulosus* microRNAs and their expression in different life cycle stages and parasite genotypes. *Int. J. Parasitol.* 2011; 41:439–448. [PubMed: 21219906]
11. Bai Y, Zhang Z, Jin L, Kang H, Zhu Y, Zhang L, Li X, Ma F, Zhao L, Shi B, Li J, McManus DP, Zhang W, Wang S. Genome-wide sequencing of small RNAs reveals a tissue-specific loss of conserved microRNA families in *Echinococcus granulosus*. *BMC Genomics.* 2014; 15:736. [PubMed: 25168356]
12. Chemale G, van Rossum AJ, Jefferies JR, Barrett J, Brophy PM, Ferreira HB, Zaha A. Proteomic analysis of the larval stage of the parasite *Echinococcus granulosus*: causative agent of cystic hydatid disease. *Proteomics.* 2003; 3:1633–1636. [PubMed: 12923787]
13. Monteiro KM, de Carvalho MO, Zaha A, Ferreira HB. Proteomic analysis of the *Echinococcus granulosus* metacestode during infection of its intermediate host. *Proteomics.* 2010; 10:1985–1999. [PubMed: 20217864]
14. Virginio VG, Monteiro KM, Drumond F, de Carvalho MO, Vargas DM, Zaha A, Ferreira HB. Excretory/secretory products from in vitro-cultured *Echinococcus granulosus* protoscoleces. *Mol. Biochem. Parasitol.* 2012; 183:15–22. [PubMed: 22261090]
15. Cui SJ, Xu LL, Zhang T, Xu M, Yao J, Fang CY, Feng Z, Yang PY, Hu W, Liu F. Proteomic characterization of larval and adult developmental stages in *Echinococcus granulosus* reveals novel insight into host-parasite interactions. *J. Proteomics.* 2013; 84:158–175. [PubMed: 23603110]
16. Cain JA, Solis N, Cordwell SJ. Beyond gene expression: the impact of protein post-translational modifications in bacteria. *J. Proteomics.* 2014; 97:265–286. [PubMed: 23994099]
17. Walsh CT, Garneau-Tsodikova S, Gatto GJ. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew. Chem., Int. Ed.* 2005; 44:7342–7372.

18. Arnesen T, Van Damme P, Polevoda B, Helsens K, Evjenth R, Colaert N, Varhaug JE, Vandekerckhove J, Lillehaug JR, Sherman F, Gevaert K. Proteomics analyses reveal the evolutionary conservation and divergence of N-terminal acetyltransferases from yeast and humans. *Proc. Natl. Acad. Sci. U. S. A.* 2009; 106:8157–8162. [PubMed: 19420222]
19. Mischerikow N, Heck AJ. Targeted large-scale analysis of protein acetylation. *Proteomics.* 2011; 11:571–589. [PubMed: 21246731]
20. Smith LM, Kelleher NL. The Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nat. Methods.* 2013; 10:186–187. [PubMed: 23443629]
21. erný M, Skalák J, Cerna H, Brzobohatý B. Advances in purification and separation of posttranslationally modified proteins. *J. Proteomics.* 2013; 92:2–27. [PubMed: 23777897]
22. Olsen JV, Mann M. Status of large-scale analysis of posttranslational modifications by mass spectrometry. *Mol. Cell. Proteomics.* 2013; 12:3444–3452. [PubMed: 24187339]
23. Silva AM, Vitorino R, Domingues MR, Spickett CM, Domingues P. Post-translational modifications and mass spectrometry detection. *Free Radical Biol. Med.* 2013; 65:925–941. [PubMed: 24002012]
24. Huang J, Wang F, Ye M, Zou H. Enrichment and separation techniques for large-scale proteomics analysis of the protein posttranslational modifications. *J. Chromatogr., A.* 2014; 1372C:1–17. [PubMed: 25465002]
25. Kellie JF, Tran JC, Lee JE, Ahlf DR, Thomas HM, Ntai I, Catherman AD, Durbin KR, Zamdborg L, Vellaichamy A, Thomas PM, Kelleher NL. The emerging process of Top Down mass spectrometry for protein analysis: biomarkers, protein-therapeutics, and achieving high throughput. *Mol. Biosyst.* 2010; 6:1532–1539. [PubMed: 20711533]
26. Ahlf DR, Thomas PM, Kelleher NL. Developing top down proteomics to maximize proteome and sequence coverage from cells and tissues. *Curr. Opin. Chem. Biol.* 2013; 17:787–794. [PubMed: 23988518]
27. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M, Wu C, Sweet SM, Early BP, Siuti N, LeDuc RD, Compton PD, Thomas PM, Kelleher NL. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature.* 2011; 480:254–258. [PubMed: 22037311]
28. Catherman AD, Durbin KR, Ahlf DR, Early BP, Fellers RT, Tran JC, Thomas PM, Kelleher NL. Large-scale top-down proteomics of the human proteome: membrane proteins, mitochondria, and senescence. *Mol. Cell. Proteomics.* 2013; 12:3465–3473. [PubMed: 24023390]
29. Bowles J, Blair D, McManus DP. Genetic variants within the genus *Echinococcus* identified by mitochondrial DNA sequencing. *Mol. Biochem. Parasitol.* 1992; 54:165–173. [PubMed: 1435857]
30. Alvarez Rojas CA, Romig T, Lightowers MW. *Echinococcus granulosus sensu lato* genotypes infecting humans—review of current knowledge. *Int. J. Parasitol.* 2014; 44:9–18. [PubMed: 24269720]
31. Garat B, Esperón P, Picón M, Ehrlich R. *Echinococcus granulosus*: preparation of protein extracts from protoscolex nuclei for mobility-shift assays. *Parasitol. Res.* 1998; 84:598–600. [PubMed: 9694381]
32. Wessel D, Flügge UI. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* 1984; 138:141–143. [PubMed: 6731838]
33. Ntai I, Kim K, Fellers RT, Skinner OS, Smith AD, Early BP, Savaryn JP, LeDuc RD, Thomas PM, Kelleher NL. Applying label-free quantitation to top down proteomics. *Anal. Chem.* 2014; 86:4961–4968. [PubMed: 24807621]
34. Vellaichamy A, Tran JC, Catherman AD, Lee JE, Kellie JF, Sweet SM, Zamdborg L, Thomas PM, Ahlf DR, Durbin KR, Valaskovic GA, Kelleher NL. Size-sorting combined with improved nanocapillary liquid chromatography-mass spectrometry for identification of intact proteins up to 80 kDa. *Anal. Chem.* 2010; 82:1234–1244. [PubMed: 20073486]
35. Wenger CD, Phanstiel DH, Lee MV, Bailey DJ, Coon JJ. COMPASS: a suite of pre- and post-search proteomics software tools for OMSSA. *Proteomics.* 2011; 11:1064–1074. [PubMed: 21298793]

36. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005; 21:3674–3676. [PubMed: 16081474]
37. Durbin KR, Fellers RT, Ntai I, Kelleher NL, Compton PD. Autopilot: an online data acquisition control system for the enhanced high-throughput characterization of intact proteins. *Anal. Chem*. 2014; 86:1485–1492. [PubMed: 24400813]
38. Wu S, Brown RN, Payne SH, Meng D, Zhao R, Toli N, Cao L, Shukla A, Monroe ME, Moore RJ, Lipton MS, Paša-Toli L. Top-Down Characterization of the Post-Translationally Modified Intact Periplasmic Proteome from the Bacterium *Novosphingobium aromaticivorans*. *Int. J. Proteomics*. 2013; 2013:279590. [PubMed: 23555055]
39. Leduc RD, Kelleher NL. Using ProSight PTM and related tools for targeted protein identification and characterization with high mass accuracy tandem MS data. *Curr. Protoc Bioinformatics*. 2007; 19:13.6, 13.6.1.
40. 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–W706. [PubMed: 17586823]
41. Brown JL. The N-terminal region of soluble proteins from procaryotes and eucaryotes. *Biochim. Biophys. Acta, Protein Struct*. 1970; 221:480–488.
42. Arnesen T. Towards a functional understanding of protein N-terminal acetylation. *PLoS Biol*. 2011; 9:e1001074. [PubMed: 21655309]
43. Giglione C, Boularot A, Meinel T. Protein N-terminal methionine excision. *Cell. Mol. Life Sci*. 2004; 61:1455–1474. [PubMed: 15197470]
44. Frottin F, Martinez A, Peynot P, Mitra S, Holz RC, Giglione C, Meinel T. The proteomics of N-terminal methionine cleavage. *Mol. Cell. Proteomics*. 2006; 5:2336–2349. [PubMed: 16963780]
45. Kim HK, Kim RR, Oh JH, Cho H, Varshavsky A, Hwang CS. The N-terminal methionine of cellular proteins as a degradation signal. *Cell*. 2014; 156:158–169. [PubMed: 24361105]
46. Starheim KK, Gevaert K, Arnesen T. Protein N-terminal acetyltransferases: when the start matters. *Trends Biochem. Sci*. 2012; 37:152–161. [PubMed: 22405572]
47. Van Damme P, Hole K, Pimenta-Marques A, Helsens K, Vandekerckhove J, Martinho RG, Gevaert K, Arnesen T. NatF contributes to an evolutionary shift in protein N-terminal acetylation and is important for normal chromosome segregation. *PLoS Genet*. 2011; 7:e1002169. [PubMed: 21750686]
48. LeDuc RD, Taylor GK, Kim YB, Januszyk TE, Bynum LH, Sola JV, Garavelli JS, Kelleher NL. ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic Acids Res*. 2004; 32:W340–W345. [PubMed: 15215407]
49. Oehring SC, Woodcroft BJ, Moes S, Wetzel J, Dietz O, Pulfer A, Dekiwadia C, Maeser P, Flueck C, Witmer K, Brancucci NM, Niederwieser I, Jenoe P, Ralph SA, Voss TS. Organellar proteomics reveals hundreds of novel nuclear proteins in the malaria parasite *Plasmodium falciparum*. *Genome Biol*. 2012; 13:R108. [PubMed: 23181666]
50. Huang HY, Hopper AK. In vivo biochemical analyses reveal distinct roles of β -importins and eEF1A in tRNA subcellular traffic. *Genes Dev*. 2015; 29:772–783. [PubMed: 25838545]
51. Wesolowska N, Lénárt P. Nuclear roles for actin. *Chromosoma*. 2015 DOI: [10.1007/s00412-015-0519-8](https://doi.org/10.1007/s00412-015-0519-8).
52. Lorenzatto KR, Monteiro KM, Paredes R, Paludo GP, da Fonsêca MM, Galanti N, Zaha A, Ferreira HB. Fructose-bisphosphate aldolase and enolase from *Echinococcus granulosus*: genes, expression patterns and protein interactions of two potential moonlighting proteins. *Gene*. 2012; 506:76–84. [PubMed: 22750316]
53. Kobiyama K, Kawashima A, Jounai N, Takeshita F, Ishii KJ, Ito T, Suzuki K. Role of Extrachromosomal Histone H2B on Recognition of DNA Viruses and Cell Damage. *Front. Genet*. 2013; 4:91. [PubMed: 23734163]
54. Keaton MA, Szkotnicki L, Marquitz AR, Harrison J, Zyla TR, Lew DJ. Nucleocytoplasmic trafficking of G2/M regulators in yeast. *Mol. Biol. Cell*. 2008; 19:4006–4018. [PubMed: 18562688]

Figure 1.

Representative data from LC-MS/MS analysis of a single GELFrEE fraction. (A) GELFrEE fractionation of the nuclear sample shown as an example. (B) LC total ion chromatogram from GELFrEE fraction 2. (C) two representative spectra from intact proteins.

Figure 2.

Overlapping between nuclear and cytosolic protein IDs and proteoforms identified by the top down approach. The comparison of proteoforms points out that, although some IDs are common, they exhibit modifications that make them different proteoforms, explaining the decrease in the overlap between nuclear and cytosolic fraction.

Figure 3. Comparison between proteins identified using top down and bottom up approaches. (A) Venn diagram showing a partial overlap between the approaches. (B) molecular mass distribution of proteins identified in both approaches.

Figure 4.

Overview of the CTMs and sequence motifs detected in the identified proteoforms. (A) Venn diagram showing the distribution of NME and NTA events. The overlap indicate proteoforms modified by NME + NTA. (B) Sequence motifs for the first 10 residues of each proteoform identified displaying NME only, NME + NTA (NatA putative substrates), and NTA only (NatC-F putative substrates).

Figure 5.
Detection of two proteoforms from histone H2A. These monoacetylated and diacetylated proteoforms were only detected using our modified database including the acetylation prediction information.

Figure 6.

Overview of the low mass proteome from *E. granulosus* protoscoleces subcellular fractions. (A) Distribution of nuclear and cytosolic proteins identified using both top down and bottom up MS. (B) gene ontology (GO) classification of proteins identified in *E. granulosus* subcellular fractions. The results of BlastP searches against the NCBI nr database were used for GO mapping and annotation. Top 10 most abundant GO terms for cellular component are shown.