The barber’s pole worm CAP protein superfamily - a basis for fundamental discovery and biotechnology advances

Namitha Mohandas\textsuperscript{a}, Neil D. Young\textsuperscript{a}, Abdul Jabbar\textsuperscript{a}, Pasi K. Korhonen\textsuperscript{a}, Anson V. Koehler\textsuperscript{a}, Parisa Amani\textsuperscript{c}, Ross S. Hall\textsuperscript{a}, Paul W. Sternberg\textsuperscript{b}, Aaron R. Jex\textsuperscript{a}, Andreas Hofmann\textsuperscript{c,*}, Robin B. Gasser\textsuperscript{a,*}

\textsuperscript{a} The University of Melbourne, Pathogen Genomics and Genetics Program, Parkville, Victoria, Australia
\textsuperscript{b} California Institute of Technology, Division of Biology, Pasadena, California, USA
\textsuperscript{c} Eskitis Institute, Griffith University, Brisbane, Australia

* Corresponding author:
robinbg@unimelb.edu.au (R. B. Gasser) or a.hofmann@griffith.edu.au (A. Hofmann)
ABSTRACT

Parasitic helminth proteins that belong to the cysteine-rich secretory proteins, antigen 5 and pathogenesis-related 1 (CAP) superfamily are proposed to play key roles in the infection process and the modulation of immune responses in the host. However, there is limited information on these proteins for most socio-economically important nematodes. Here, we review the CAP protein superfamily of Haemonchus contortus (the barber’s pole worm), one of the most important parasitic nematodes (order Strongylida) of small ruminants. To do this, we mined genome and transcriptomic datasets, predicted and curated full-length amino acid sequences (n = 45), undertook systematic phylogenetic analyses of these data and investigated transcription throughout the life cycle of H. contortus. We inferred functions for selected C. elegans orthologs (including vap-1, vap-2, scl-5 and lon-1) based on genetic networking and by integrating data and published information, and were able to infer that a subset of orthologs and their interaction partners play pivotal roles in growth and development via the insulin-like and/or the TGF-beta signaling pathways. The identification of the important and conserved growth regulator LON-1 led us to appraise the three-dimensional structure of this CAP protein by comparative modelling. This model revealed the presence of different topological moieties on the canonical fold of the CAP domain, which coincide with an overall charge separation as indicated by the electrostatic surface potential map. Similar to the molecular mechanism of chemokines, these observations suggest the existence of separate sites for effector binding and receptor interactions and thus support the proposal that these parasite molecules act in similar ways as venoms act as ligands for chemokine receptors or G protein-coupled receptor effectors. In conclusion, this review should guide future molecular studies of these molecules, and could support the development of novel interventions against haemonchosis.

Keywords:
Nematodes
Haemonchus contortus
Genome
Transcriptome
CAP (SCP/TAPS) proteins
Genetic networking
Transcription profiles
Structure models
1. Introduction

*Haemonchus contortus*, also known as the barber’s pole worm, is an economically important parasitic worm (helminth) of livestock worldwide (Veglia, 1915; Sutherland and Scott, 2010). This roundworm (= nematode) infects hundreds of millions of sheep, goats and other ruminants globally, and causes disease, deaths and very substantial production losses. *H. contortus* has a life cycle that takes about three weeks (Veglia, 1915): eggs are excreted in host faeces; in the environment, single first-stage larvae (L1s) develops within eggs to then hatch (in one day) and moult through to the second- (L2s) and third-stage (L3s) larvae; the infective L3s are ingested by the host while grazing, and these develop to fourth-stage larvae (L4s) and then dioecious adults in the abomasum (stomach). The last two stages feed on blood from capillaries and cause haemonchosis (manifested in anaemia, haemorrhagic gastritis, oedema and other complications), often leading to death in severely affected animals (Sutherland and Scott, 2010).

Although a parasite-derived vaccine (Barbervax; http://barbervax.com.au/) has been released, haemonchosis control has relied heavily on the use of anthelmintic drugs. As resistance against some classes of anthelmintics has become very widespread (Wolstenholme et al., 2004; Kaplan and Vidyashankar, 2012; Wolstenholme and Martin, 2014), there is an ongoing need for the development of new compounds (cf. von Samson-Himmelstjerna et al., 2005; Kaminsky et al., 2008; Little et al., 2011). Identifying new intervention targets has been challenging, particularly because of a limited understanding of the molecular biology of *H. contortus* and how it interacts with its host (Nikolaou and Gasser, 2006). Excretory/secretory (ES) molecules from *H. contortus* likely play key roles in the host interplay and the infection process, a statement supported by research revealing abundant expression of transthyretin-like (TTL) protein, peptidase inhibitor, fatty acid binding protein, spermatid-coating protein (SCP)-like protein and lectin genes in the blood-feeding (haematophagous) stages of *H. contortus* (see Laing et al., 2013; Schwarz et al., 2013). In addition, cysteine-rich secretory proteins, antigen 5 and pathogenesis-related 1 (CAP) superfamily members (Gibbs et al., 2008), typified by SCP-like extracellular domains (InterProScan codes: IPR014044 and IPR001283), are also prominent in ES products. Although the two CAP proteins *Hc24* and *Hc40* have been identified amongst ES molecules of *H. contortus* (see Schallig et al., 1997; Rehman and Jasmer, 1998; Yatsuda et al., 2003), there appear to be many more related molecules in this worm, according to recent evidence (Schwarz et al., 2013). This information supports the proposal of a diversified and specific involvement of CAP proteins during *H. contortus* infection and in host interplay. Despite the major socioeconomic impact of haemonchosis globally and some promise for CAP proteins as drug or vaccine targets (Cantacessi et al., 2009, 2012), there is still limited detailed information on this group of proteins for *Haemonchus*.

Although challenging in the past, detailed global molecular explorations have now become feasible because of the accessibility of high throughput nucleic acid sequencing technologies and advanced informatics (Mardis et al., 2008, 2013; Koboldt et al., 2013; Cantacessi et al., 2012, 2015). Using an Illumina sequencing approach, we recently characterized a draft genome and transcriptomes for *H. contortus* from Australia (NCBI BioProject accession no. PRJNA193158; Schwarz et al, 2013), providing an important resource to explore the systems biology of this nematode. In a preliminary assessment, we predicted a sizeable set of CAP protein-encoding genes in this genome (see Schwarz et al., 2013). In the present article, we...
curate full-length open reading frames (ORFs) and predict the full complement of CAP proteins, establish their genetic/phylogenetic relationships, study the transcription profiles of their genes in all key developmental stages and model these proteins in silico. This review should guide future molecular explorations of these CAP proteins and could support the design of novel intervention approaches for haemonchosis as biotechnological outcomes.

2. Methodology

We used the draft genome and transcriptomes of all key developmental stages (egg, L1-L4 and adult) and both sexes (L4 and adult) of *H. contortus* from Australia (NCBI BioProject accession no. PRJNA193158; WormBase (www.wormbase.org) (Schwarz et al., 2013). This genome is 320 Mb and predicted to encode 23,610 proteins (Schwarz et al., 2013). The CAP protein-encoding genes were inferred using an integrative approach of evidence-based gene prediction (assembled transcripts and RNA-seq data) and *ab initio* gene prediction (MAKER and GeneMark). We employed the MAKER2 pipeline (Holt and Yandell, 2011) to predict CAP protein-encoding genes in the *H. contortus* genome assembly. Then, GeneMark-ES (Ter-Hovhannisyan et al., 2008) was used to predict the open reading frames (ORFs) of assembled transcripts. All 167,013 assembled transcripts of *H. contortus* were subjected to an homology-based search against sequences available in the public databases NCBI and InterProScan (Zdobnov et al., 2001; Quevillon et al., 2005) using tBLASTx (Altschul et al., 1990) analysis (e-value cut-off: 10\(^{-5}\)). All *H. contortus* transcripts containing regions of homology (i.e. at least one CAP domain) with known CAP protein-encoding genes were extracted and mapped to *H. contortus* genomic scaffolds using BLAT (Kent, 2002). Gene predictions (Maker2 GFF file), BLAT-transcript alignments (PSL files) and RNA-seq reads mapped to genomic scaffolds (BAM files) were visually inspected and curated using the Integrative Genomics Viewer (IGV; Robinson et al., 2011). Then, we integrated the gene prediction information from multiple sources to obtain a consensus of each gene and corresponding full-length transcript using ORF finder (http://www.ncbi.nlm.nih.gov/projects/gorf/). Then, protein domain searches were conducted using InterProScan (Zdobnov et al., 2001; Hunter et al., 2009), employing default parameters.

Decontaminated and quality filtered paired-end RNA-seq reads were mapped to the set of full-length transcripts using the Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2010). The numbers of the mapped reads per individual transcript were extracted using SAMtools (Li et al., 2009). The resultant read counts per transcript per life cycle stage were used as input data for edgeR (Robinson et al., 2010). Using an edgeR dispersion factor of zero, the genes were considered differentially transcribed if the logarithmic fold-change (FC) compared with the normalised read count data was ≥ 2 and the P-value was ≤ 10\(^{-5}\). Transcription levels were compared in a pairwise manner between different developmental stages of *H. contortus*. Genes were recorded as differentially transcribed using edgeR-calculated common and gene-wise dispersion factors, if the FC compared with the normalised read count data was ≥ 2 and the false discovery rate (FDR) was ≤ 0.05. A heat map (representing mapped reads) was produced using the heatmap.2 function in the gplots package in R (http://www.R-project.org/).

Amino acid (aa) sequences encoded by individual CAP protein genes were translated from full-length transcripts and then aligned using MUSCLE v.3.8 (Edgar, 2004). Because of substantial length differences, the predicted CAP protein sequences
were divided into distinct groups based on their alignments and patterns within CAP domains. For each groups, conserved amino acid residues and PROSITE patterns (Gibbs et al., 2008) were identified. These three sets of data were then individually subjected to phylogenetic analysis using Bayesian inference (BI) employing MrBayes v.3.2.2 (Ronquist and Huelsenbeck, 2003); the optimal model of sequence evolution was assessed utilising a mixed amino acid substitution model and four rate categories approximating a Γ distribution, four chains and 1,000,000 generations, sampling every 100th generation; the first 25% of the generations sampled were removed from the analysis as burn-in.

We used the program GeneMania (Warde-Farley et al., 2010) to display predicted or experimentally validated genetic interactions of *C. elegans* orthologs encoding CAP proteins of *H. contortus*. Those results were compiled and the networks saved in scalable vector graphics (svg) format. Images were labeled and formatted using Inkscape image editor (http://www.inkscape.org/en/).

Annotation and fold appraisal was guided by results from automated model predictions performed with I-TASSER (Yang et al., 2015). Using PSIPRED (Bryson et al., 2005) secondary structure predictions for inferred *H. contortus* CAP proteins, structure-based amino acid sequence alignments were made using SBAL (Wang et al., 2012; Wang and Hofmann, 2015), and the amino acid sequence of SteCRISP was added using the crystal structure deposited in PDB entry 1rc9 (Guo et al., 2005). In preparation for comparative modelling of *Hc*-CAP-15, SteCRISP was identified as the most suitable template (*P*-value: 2x10^-6) by surveying the PDB with pGenThreader (Bryson et al., 2005) for the protein with highest homology. A three-dimensional model was constructed by comparative modelling using the program MODELLER (Sali and Blundell, 1993) in two stages employing the structure of SteCRISP as a template and the structure-based amino acid sequence alignment of *Hc*-CAP-15 and SteCRISP. First, 20 independent models were computed without restraint, and the model with the lowest objective function was selected and visually inspected. The model topology around all free cysteine residues was assessed in relation to potential formation of disulphide bonds. In a second run, geometric restraints were applied to enforce the formation of disulphide bonds (Cys149-Cys159 and Cys151-Cys158). As before, from 20 independently calculated models, the one with the lowest objective function was selected.

3. The full complement of CAP proteins and their relationships

Homology-based searches against sequences in NCBI and InterProScan using tBLASTx analysis identified 1,155 transcripts, which mapped to 261 genomic scaffolds. Manual curation identified 110 genomic scaffolds that are predicted to encode proteins of the CAP superfamily (Gibbs et al., 2008). Subsequent mapping of assembled transcripts to these genomic scaffolds revealed 49 full-length and 61 partial transcripts. Of the 49 amino acid sequences encoding full-length transcripts, 45 of them contained one to three CAP domains and were predicted to represent 33 single-domain, 11 double-domain and one triple-domain CAP proteins.

The features of these 45 predicted proteins (including length, InterProScan domains and signal peptides) are summarised in Table 1. Specifically, the single-domain proteins (184-300 aa) are inferred to be considerably shorter than the double-domain proteins (380-462 aa) and differed in sequence from one another by 0-90% and 1-79%, respectively, upon pairwise comparison; the triple-domain protein identified is 700 amino acids long.
Based on the InterProScan survey, all 45 predicted proteins belong to the cysteine-rich secreted protein-related family (PTHHR10334) with the allergen V5/Tpx-1 family (PR00837) and venom-allergen 5 (PR00838) signatures; 39 of these 45 proteins had the PR-1-like (SSF55797), SCP (SM00198), pathogenesis-related (G3DSA: 3.40.33.10) and CAP (PF00188) signatures (Table 1). In addition, 11 single-domain and seven double-domain proteins had an allergen V5/Tpx-1 related, conserved site CRISP-1 (PS01009). The allergen V5/Tpx-1 related, conserved site CRISP-2 (PS01010) was present in two single-domain but absent from all 11 double-domain proteins (Table 1). Moreover, 27 single-domain proteins and eight double-domain proteins had a signal peptide, which was absent from the other nine predicted proteins (Table 1).

Based on Prosite patterns (Table 1), we initially assigned the 45 proteins to five distinct groups: Group 1 represented all proteins that possessed conserved CAP3 and CAP4 motifs and variable CAP1 and CAP2 motifs; group 2 represented those that did not have a CAP4 motif and were variable in sequence within all the CAP1, CAP2 and CAP3 motifs; group 3 contained proteins that were variable in sequence within the CAP1, CAP2 and CAP4 motifs but had the conserved CAP3 motif. Group 4 represented all 11 double-domain proteins characterised by the conserved CAP3 motif at the C-terminal of the double-domain proteins and the variable CAP1, CAP2 and CAP4 motifs. *Hc*-CAP-44 did not have the CAP4 motif and *Hc*-CAP-41 only consisted of a partial CAP2 motif. Notably, the N-terminal moiety of the double-domain proteins was variable in sequence within all CAP motifs. The only triple-domain protein *Hc*-CAP-45 was classified as group 5.

A phylogenetic analysis of amino acid sequence data displayed the relationships of CAP proteins within individual groups (Fig. 1A). Specifically, within group 1, 12 predicted proteins with close homology to *Na*-ASP-2 and *Hc*24 (Yatsuda et al., 2003; Asojo et al., 2005a) grouped separately (p = 0.59-1.00) to the exclusion of CAP proteins with homology to *C. elegans* LON-1 protein (*Hc*-CAP-15 and -16) and GLIPR-1 (*Hc*-CAP-17) (cf. Shaye and Greenwald, 2011). Within group 2, three predicted proteins (*Hc*-CAP-22, -23 and -24) that shared homology to *Ac*-ASP-7 (Datu et al., 2008; Osman et al., 2012) formed a separate clade to the exclusion of *Hc*-CAP-25. Group 4 double-domain proteins also formed separate clusters according to homology amongst published CAP proteins; *Hc*-CAP-34 and *Hc*-CAP-35, which had homology to *Hc*40 (Rehman and Jasmer, 1998; Yatsuda et al., 2003) formed a separate clade to double-domain proteins which shared sequence homology to ASP-1 (i.e. *Hc*-CAP-36 to *Hc*-CAP-41) and VAP-2 (i.e. *Hc*-CAP-42 and *Hc*-CAP-43). With reference to the phylogenetic trees, we then proceeded to explore transcription profiles for all 45 individual genes in all life stages of *H. contortus*.

4. Transcription profiles

Significant differences in transcription were recorded among many of the 45 genes encoding CAP proteins (i.e. 33 single-, 11 double- and 1 triple-domain proteins) (Fig. 1B), of which 7 single-domain and 7 double-domain proteins were up-regulated in egg and L1-L3 stages, respectively, with respect to other developmental stages (Fig. 1B). Conversely, transcripts encoding 26 single-, 4 double- and one triple domain CAP proteins were significantly up-regulated in haematophagous stages of *H. contortus* (Fig. 1B), initially suggesting that these molecules play key roles in the infection process and/or in maintaining the parasite-host relationship. Although no direct comparison was undertaken here, this up-regulation appears to be distinct from
hookworms *A. caninum*, *A. ceylanicum* and *N. americanus* (see Goud et al., 2005; Datu et al., 2008; Osman 2012; Schwarz et al., 2015), which usually undergo percutaneous, pulmonary and/or somatic migration, and establish in the small intestine rather than the abomasum. This latter difference could explain the apparent expansions of CAP protein genes in *N. americanus* (*n = 128*; Tang et al., 2014) and *A. ceylanicum* (*n = 432*; Schwarz et al., 2015) compared with *H. contortus*. The relatively high proportion (28%) of CAP molecules of *H. contortus* with homologues in *C. elegans* contrasts the situation for *N. americanus* and *A. ceylanicum* (see Tang et al., 2014; Schwarz et al., 2015), in which only six of 137 (4.3%) CAP proteins have orthologues in the free-living nematode. In addition, there are five paralogs encoding related proteins (*Hc-CAP-22, -23, -24, -31 and -32*) in *H. contortus*, which further suggests that they relate to parasite-specific functions.

Gender-enriched transcription was also explored in female and male haematophagous stages of *H. contortus* (Fig. 1B). Significant differences in transcription were recorded for genes encoding 29 single- and seven double-domain CAP proteins, four of which were up-regulated in female haematophagous stages (i.e. L4 and adult) (see Fig. 1B). Conversely, transcripts encoding seven single-, two double-domain and one triple-domain CAP proteins were significantly up-regulated in male haematophagous stages (*p < 0.05*), thus suggesting key roles in reproduction, interaction between the sexes and/or the parasite-host relationship. Most of the gender-enriched molecules encoded by *Hc-cap-7, -10, -22, -23, -29, -31, -32* and -38 were common to L4 and adult stages, but absent from egg and L1-L3 stages (Fig. 1B).

5. Functional annotation based on publicly available information and genetic interactions in *C. elegans*

Of all 45 predicted CAP proteins of *H. contortus*, 13 had *C. elegans* homologs (Supplementary Table 1). Specifically, eight of the 11 double-domain proteins of *H. contortus* had homologs in *C. elegans*, including Ce-VAP-1 (six homologs: *Hc-CAP-36* to *Hc-CAP-41*), Ce-VAP-2 (two homologs: *Hc-CAP-42* and *Hc-CAP-43*) (Supplementary Table 1). Of the 33 single-domain proteins of *H. contortus*, five had homologs in *C. elegans* or human, namely Ce-SCL-5 (*Hc-CAP-14*) and Ce-SCL-19 (*Hc-CAP-13*), Ce-LON-1 (two isoforms; *Hc-CAP-15* and *Hc-CAP-16*) and Ce-F57B7.2 (*Hc-CAP-17*) (Supplementary Table 1).

5.1. Single-domain proteins

Based on information available for *C. elegans* (WormBase), we propose that the LON-1 homologs, *Hc-CAP-15* and -16, regulate body length and polyploidisation in hypodermal cells in adult and larval *H. contortus*. In *C. elegans*, the protein encoded by *lon-1* is a target of TGF-beta signalling and is expressed in hypodermal tissues; the *lon-1* gene is epistatic to *dbl-1* and negatively regulated by Sma/Mab pathway signalling (Maduzia et al., 2002; Morita et al., 2002; Tuck et al., 2014). The disruption of *lon-1* function in *C. elegans* by dsRNAi results in long worms, indicating that it is a negative regulator of body length (Morita et al., 2002). *Ce-lon-1* encodes a 312 amino acid protein with a sequence motif (GHYVQVVW) that is conserved across many metazoan organisms including *H. contortus* (e.g., Kasahara et al., 1989; Lu et al., 1993; Murphy et al., 1995; Kjeldsen et al., 1996; Morita et al., 2002; Delannoy-Normand et al., 2010). Expression studies have also shown that *Ce-LON-1* is repressed by *Ce-DBL-1*, suggesting that *Ce-LON-1* is a downstream component of
the TGF-beta growth regulation pathway (Morita et al., 2002). LON-1 is expressed mainly in adult and larval stages (as is the case in H. contortus) and has dose-dependent effects on body length linked to hypodermal ploidy (cf. Morita et al., 2002).

The lon-1 gene has numerous, mostly proven interactions with genes including Sma-2, -3, -4 (complex), -6, -9, -10, -12, -13, -14, -16, -17, -18 and -19 encoding various Smad proteins (transcription factors), and dbl-1 (see above), daf-4, kin-29 (encoding a kinase involved in regulating the expression of chemosensory receptors and entry into the dauer pathway), rnt-1, crm-1, lgg-1 and che-2 (encodes a protein with G-protein beta-like WD-40 repeats that affects chemotaxis, longevity and dauer formation; expressed in male tail rays, some head neurons and cilia of sensory neurons; e.g., Vowels and Thomas, 1992; Bargmann et al., 1993; Fujiwara et al., 1999; Swoboda et al., 2000; Pujol et al., 2001; Tonkin et al., 2002), most of which are involved (up or down-stream) in TGF-beta signalling. Given its relative conservation between C. elegans and H. contortus, we propose that this complement of genes/proteins regulates body size and specification of sensory structures in the male tail, spicule formation, male tail patterning and morphogenesis, oocyte and germline quality and/or reproductive aging, and are expressed in the pharynx, intestine and/or hypodermis of H. contortus.

Hc-cap-17 is a close ortholog of Ce-F57B7.2 (isoform b) orphan gene of C. elegans (73% identity) and the human gene GLIPR (accession number: AAA82731; 29% identity) (Murphy et al., 1995). The latter gene encodes a member of the Glioma pathogenesis related-1 (GLIPR1) protein subfamily, also referred to as related to testis specific, vespid and pathogenesis related-1 (RTVP1) proteins and being amongst the best characterized CAP subfamily members in mammals (Gibbs et al., 2008). In contrast to the germline-enriched and sex-biased expression profile (cluster A) of Ce-F57B7.2 (isoform b) in C. elegans (see Reinke et al., 2004), Hc-cap-17 had a high and constitutive transcription in all stages of H. contortus studied. Therefore, it seems likely that Hc-cap-17 plays a role in growth and development throughout the life cycle, although this gene or its product could be involved in reproductive processes. Interestingly, evidence from humans shows that GLIPR1L1 is predominantly expressed in the testis, with trace amounts in the bladder (Ren et al., 2006), although the precise cellular localization of GLIPR1L1 is presently unknown. The mouse glipr1l1 gene also has a testis-enriched expression profile; the encoded protein, which undergoes post-translational N-linked glycosylation during spermatogenesis, becomes localized to the connecting piece of elongated spermatids and sperm (Gibbs et al., 2010). After sperm capacitation, mouse GLIPR1L1 has been shown to localize to the anterior area of the sperm head and has a role in the binding of sperm to the zona pellucida surrounding the oocyte (Gibbs et al., 2010). Taken together, this information suggests that, possibly together with other CAP superfamily members and other proteins, GLIPR1L1 is involved in the binding of sperm to the oocyte complex, which strengthens the proposed role of CAP proteins in cellular adhesion (cf. Gibbs et al., 2008).

Ce-scl-5, the C. elegans ortholog of Hc-cap-14, encodes an extracellular CAP protein; the Ce-scl-5 gene is predicted to interact directly with Ce-vap-1 and Ce-vap-2 as well as a transcription factor enhancer encoded by the orphan gene F47B8.2 (Lee et al., 2008). Although its function is presently unknown, abundant transcription of Hc-cap-14 in parasitic stages of H. contortus indicates a role in parasitism or the host-parasite interplay. Another gene, Hc-cap-13, is a Ce-scl-19 ortholog encoding a cysteine-rich secretory protein Limulus clotting factor C (LCCL) domain containing 1
(HGNC: CRISPLD1) protein, with no known interaction with any other gene in *C. elegans*. Based on high levels of expression of CAP orthologs in various embryonic tissues undergoing growth and development (in ear, eye, palate and lips) of mammals (human and mouse) (Robertson et al., 1997; Delrieu et al., 2002; Ahsan et al., 2005; Bhattacharya et al., 2005; Gibbs et al., 2008; Chiquet et al., 2011), we propose that *Hc-CAP-13* is involved in growth and development in *H. contortus*. This statement is supported by high levels of transcription in the egg, L2 and L3 stages and possibly in eggs within the female adult (Fig. 1).

5.2. Double-domain proteins

Six CAP proteins encoded by *Hc-cap-36* to *Hc-cap-41* are homologs of *Ce-VAP-1* (Supplementary Table 1), which is a secreted protein similar to the venom allergen-like (VAL) proteins found in a number of invertebrates, including parasitic nematodes, and a human homolog of cysteine-rich secretory protein 3 isoform 2 precursor (Shaye and Greenwald, 2011). The genes *Hc-cap-36* and *Hc-cap-37* usually exhibited higher transcription in the free-living stages of *H. contortus* (particularly L1 and L2) compared with parasitic (haematophagous) stages in which females displayed higher transcription than males. Although the transcription of *Hc-cap-38* was low in free-living stages, male haematophagous stages had higher transcription than females. For *Hc-cap-41*, highest transcription was seen in the free-living stages, particularly L1 and L2, compared with very low levels in the parasitic stages. The transcription profiles for *Hc-cap-39* and *Hc-cap-40* were similar, with highest levels in L2 and L3 and lowest levels in haematophagous (parasitic) stages. The significant expansion of VAP-1 orthologs in *H. contortus* suggests diversified functional roles of these six molecules in the parasite, which appears to be reflected in considerable differences among their transcription profiles. Based on functional data for *C. elegans*, we hypothesise that at least some of these *Hc-cap* genes are expressed specifically within amphid sheath cells and that they are involved in motility. In *C. elegans*, a Ce-vap-1 reporter fusion is expressed specifically within such cells, and knock-down of Ce-vap-1 results in an Unc phenotype (WBPhenotype:0000643); worms exhibit altered motility (Cronin et al., 2005). In addition, we propose that gender-enriched transcription profiles for *Hc-cap-36* and *Hc-cap-37* (females) or *Hc-cap-38* (males) in haematophagous stages relate also to reproductive and/or developmental processes. A recent molecular investigation (Lozano-Torres et al., 2014) has shown that the silencing of a gene of *Globodera rostochiensis* (potato cyst nematode) with ~35% homology to *Hc-cap-38* suppressed the activation of plant host defences, indicating a role of at least some CAP protein gene homologs in immune modulation.

Using genetic network analysis (cf. Supplementary Fig. 2), Ce-vap-1 is inferred to interact with 17 Ce-scl paralogs (*Ce-scl-1*, -2, -3, -5, -6, -7, -8, -9, -10, -11, -12, -13, -14, -15, -17, -18, -20) (Lee et al., 2008). Various experiments have also shown that *Ce-scl-1* interacts (directly) with *daf-2*, *daf-16* and *age-1* (Larsen, 2003; Libina et al., 2003; Ookuma et al., 2003; Patterson, 2003; Liu et al., 2004; Zhong and Sternberg, 2006; Harada et al., 2007). Although Ce-vap-1 and Ce-vap-2 do not directly interact with each other, all of the 17 Ce-scl paralogs are predicted to interact with Ce-vap-2 (Lee et al., 2008), which itself interacts specifically with egl-9 (Gort et al., 2008). *Ce-scl-20* encodes an ortholog of human cysteine-rich secretory protein (HGNC:GLIPR1) (Shaye and Greenwald, 2011), which is a *p53* target gene with pro-apoptotic activity mediated by an increased production of reactive oxygen species (ROS) (Pinkston-Gosse and Kenyon, 2007; Li et al., 2008). *Ce-scl-18* and -19 encode
orthologs of human cysteine-rich secretory protein LCCL domain containing 1 (HGNC:CRISPLD1), whose function is presently unknown, but which might be involved in the parasite-host interplay and sustaining parasitism.

The two Ce-vap-2 homologs of *H. contortus*, *Hc-cap-42 to Hc-cap-43*, encode an ortholog of human cysteine-rich secretory protein 3 (HGNC:CRISP3), which is involved in receptor-mediated endocytosis (Balklava et al., 2007). These receptors tightly bind extracellular macromolecules (ligands) that they specifically recognize; the plasma-membrane region containing the receptor-ligand complex then undergoes endocytosis and forms a transport vesicle containing the receptor-ligand complex. Although involved in the specific uptake of particular substances (e.g., low density lipoproteins [LDLs] or iron) required by cells, endocytosis has also been linked to the transduction of signals from the periphery of cells to their nuclei (Ceresa and Schmid, 2000). The up-regulation of transcription for *Hc-cap-42 to Hc-cap-43* apparent in the L2 stage might indicate a substantial increase in nutrient uptake in the free-living environment prior to ensheathment, after which the L3 stage (consistent with *C. elegans* dauer) can no longer feed and absorb nutrients from its environment, but rather is reliant on accumulated nutrient resources. The increased transcription of both *Hc-cap-42 to Hc-cap-43* in the female adult stage of *H. contortus* suggests an increased nutrient uptake of progeny within gravid females. This proposal is consistent with evidence for *C. elegans* showing that Ce-vap-2 is involved in the uptake of yolk in oocytes (Grant and Hirsh, 1999; Gort et al., 2008). Experiments in *C. elegans* have demonstrated that knock-down of Ce-vap-2 results in an Rme phenotype (endocytic transport defective or oocyte physiology variant; WBPhenotype:0001425), leading to a lack of yolk uptake in oocytes.

In addition, *Ce-vap-2* interacts genetically with *egl-9* (experimental evidence; Gort et al., 2008), which encodes a proline hydroxylase. Functional genomic studies have shown that *Ce-vap-2* suppresses *egl-9*, leading to egg laying (Egl) defect (http://www.wormbase.org/species/all/phenotype/WBPhenotype:0000006#07153--10; Fraser et al., 2000; Maeda et al., 2001; Kamath and Ahringer, 2003; Simmer et al., 2003; Rual et al., 2004; Sonnichens et al., 2005). The *Ce-EGL-9* protein is known to function in a conserved hypoxia-sensing pathway to negatively regulate *Ce-HIF-1* (hypoxia inducible factor) by hydroxylating prolyl *Ce-HIF-1* residues (Epstein et al., 2001); EGL-9 activity is negatively regulated by its physical association with *Ce-CYSL-1*, a protein with similarity to cysteine synthases, which acts to transduce information about the environmental O₂ levels through hydrogen sulfide (H₂S) signaling (Ma et al., 2012). *Ce-EGL-9* belongs to a superfamily of proteins, such as AlkB and leprecan, which are implicated in oxidative detoxification of alkylated bases and hydroxylation of proteins (Aravind and Koonin, 2001). *Ce-EGL-9* is expressed in musculature, neurons and hypodermis, and is needed for muscle function for egg laying (Trent et al., 1983). Taken together, we propose that *Hc-vap-2* is centrally involved in aspects of reproduction, particularly egg-yolk uptake as well as egg laying, in intimate association with a *Ce-egl-9* ortholog and/or other complementary genes.

The independent interaction of *C. elegans* orthologs *Ce-vap-1* and *Ce-vap-2* with *Ce-scl-1*, which also interacts with parts of the insulin-like signaling pathway, including *Ce-age-1, Ce-daf-2* and *Ce-daf-16* (Supplementary Fig. 2) suggest also a critical and integrated involvement also in the regulation of nematode growth and development, the transition from the free-living L3 (analogous to dauer in *C. elegans*) to the parasitic stage in *H. contortus* and/or the maintenance of a balanced parasite-host relationship.
6. Structure models

In total, 35 (78%) of the 45 predicted CAP protein sequences contained an N-terminal signal peptide (Table 1), whereas the remainder did not (Table 1). These results support the existence of two main “types” of eukaryotic CAP proteins. Both secreted and non-secreted CAP proteins are characterized structurally by the presence of a conserved SCP-domain (Osman et al., 2012). In the present study, the two criteria used to classify CAP proteins were the conservation of CAP motifs and the occurrence of multiple CAP domains (see Fig. 2). Of the 45 protein sequences predicted here, the 33 single-domain proteins were classified into group 1 (n = 16; CAP4 and CAP 3 motifs conserved), group 2 (n = 4; CAP3 motif absent) and group 3 (n = 13; CAP3 motif variable) proteins. Furthermore, 11 double-domain proteins represent group 4 and one triple-domain protein was classified as group 5. Notably, all multi-domain CAP proteins of *H. contortus* have four intact CAP motifs and are thus inferred to adopt a CAP protein fold. This is different from the finding for hookworms, where the C-terminal moiety of some multi-domain proteins does not adopt such a fold (A. Hofmann, unpublished data).

Criteria used previously for the categorization of structural groups (Osman et al., 2012) were not applicable to the predicted CAP proteins of the barber’s pole worm, since all canonical cysteine residues are highly conserved (see Supplementary Fig. 1). For *H. contortus*, the fold harbouring the four CAP motifs comprises an α-β-α sandwich that may be extended at the N- as well as C-terminal ends (see Fig. 3). Generally, variable extensions are linked to the C-terminus of the extracellular domain and include the LCCL (Coch-5b2 and Lgl1), the C-type lectin, the ion channel regulator (Gibbs et al., 2008) and additional extracellular motifs. The abundance of the CAP fold and its extension, by means of a variable set of protein domains with distinct functions, suggest a “vehicle-payload” model for these proteins (Osman et al., 2012), whereby the C-terminally linked domain is delivered by the SCP-domain to sites of molecular action. Such proteins might also be involved in enzymatic activities, recognition or binding (Mason et al., 2014b).

Although conserved domains are known to play key roles in determining protein function, protein-protein interactions, and DNA binding and enzyme activities (Venter et al., 2001), structural analyses of the complete amino acid sequence of proteins should assist investigations of their function, since three-dimensional models obtained by comparative modelling may not accurately reflect the structures of these molecules on an atomic level (cf. Qian et al., 2011; Mason et al., 2014a; Asojo et al., 2011).

The identification of the two homologs, *Hc-CAP-15* and *Hc-CAP-16*, of *Ce-LON-1*, an important growth regulator of many metazoans, led us to construct a homology model for *Hc-CAP-15* (group 1) for further appraisal. *Hc-CAP-15* is identical in amino acid sequence to *Hc-CAP-16*, but their genes differ in sequence by 1.7% and are located on distinct genomic scaffolds. The distinctive mapping of transcripts to these two loci suggests that these genes are transcribed and expressed in different tissues or cells and have a slightly different regulation. Although the transcription profiles of these two genes are very similar across the life cycle of *H. contortus*, *Hc-CAP-15* appears to be transcribed at higher levels in the L3 stage than *Hc-CAP-16* (Fig. 1), suggesting a specific up-regulation in particular tissues. For these two proteins, SteCRISP was identified as the best template for comparative modelling. An appraisal of the predicted three-dimensional structure (see Fig. 3) revealed that the canonical disulphide bridges are indeed conserved. Furthermore, immediately
upstream of the CAP1 motif, two additional disulphide bridges are possible, but do not belong to the canonical set. These four additional cysteine residues are a LON-1 specific feature and are conserved in LON-1 sequences including that of Ce-LON-1. It is noted that the choice of the disulphide bridge formation, as shown in Fig. 3 (Cys149-Cys159, Cys151-Cys158), is arbitrary and another conformation (Cys149-Cys158, Cys151-Cys159) might also be possible.

The electrostatic surface potential map of Hc-CAP-15 shows an obvious charge separation (Fig. 3) that accompanies the separation into (at least) two topological moieties, suggesting that separate receptor docking and effector sites are highly likely on this molecule, thus resembling the structure-function relationships generally observed with chemokines (Saini et al., 2011). Such observations have been made earlier with other CAP proteins of nematodes (Asojo et al., 2005b, Osman et al., 2012), and support the notion that these parasite molecules may act as agonists or ligands for GPCRs such as chemokine receptors, similar to snake and bee toxins that also belong to the CAP superfamily (e.g. Yamazaki et al., 2003).

Interestingly, in the present study, we found no evidence of neutrophil inhibitory factor (NIF) homologs. NIFs can be quite abundant in ES products of some nematodes/helminths (Hewitson et al., 2009). In hookworms, for example, a NIF homolog of Ancylostoma caninum, SCP-1, has been reported and has been shown to bind host integrin CR3 (CD11b/CD18), leading to an inhibition of neutrophil function, including oxidative burst (Moyle et al., 1994; Rieu et al., 1996). The unexpected absence of a CAP protein with homology to NIF from H. contortus might suggest that other molecules assume a role akin to NIF (cf. Anbu and Joshi, 2008) in the parasitic stage of H. contortus, or that NIFs representing CAP proteins are most abundant in parasites that undergo extensive tissue migration or live within tissues.
7. Concluding remarks

Surprisingly, most (82%) of the curated 45 CAP protein sequences inferred here had not been detected previously by proteomic analysis (Yatsuda et al., 2003). The relatively large number of transcripts encoding various CAP proteins in *H. contortus* compares with *Ancylostoma caninum*, *A. ceylanicum* and *N. americanus* (cf. Hawdon et al., 1996, 1999; Datu et al., 2008; Tang et al., 2014; Schwarz et al., 2015), but contrasts with the small numbers of homologous genes inferred from transcriptomic and genomic sequence data sets for some other nematodes studied to date, including *Ascaris suum*, *Toxocara canis* and *Trichuris suis* (see Jex et al., 2011, 2014; Zhu et al., 2015). The reason(s) for this apparent difference is unclear, but could relate to differences in developmental and reproductive biology as well as varying modes of host invasion and immune modulation or evasion among nematode species (cf. Cantacessi et al., 2009). In the future, expanded data sets for other nematodes as well as improved genome assemblies (cf. Tsai et al., 2010) should allow the accurate identification and curation of CAP protein-coding genes, and should enable detailed comparative studies. Such work will provide a basis for in-depth analyses of protein features, thus assisting future functional and structural analyses of CAP protein superfamily members of nematodes.

The availability of the entire genomic and transcriptomic sequence data sets (Schwarz et al., 2013; Laing et al., 2013) and advances in functional genomics (Lok, 2012) should open up opportunities for fundamental investigations of CAP proteins and their genes in different developmental stages of *H. contortus*. The present transcription profiles for genes encoding CAP proteins suggest that these molecules play diverse, but critical roles in development and reproduction at different life cycle stages of this parasite.

In the future, assessing the levels of transcription and expression of CAP protein genes in the tissues of different developmental stages might improve our understanding of the function of these molecules. Moreover, the recent characterization of the gene silencing machinery for *H. contortus* (see Schwarz et al., 2013) could open up interesting opportunities for functional genomic studies of CAP protein genes in the larval stages of this parasite in vitro. For instance, virus-based transduction (Hagen et al., 2014; Hagen et al., 2015a,b) might be useful for the delivery of microRNA-adapted small hairpin RNAs (shRNAmir) into *H. contortus* to achieve effective gene silencing. This area deserves attention and has the potential to overcome the inefficiency of RNAi approaches employed to date on *H. contortus* (reviewed by Knox et al., 2007; Lok, 2012). The development of an effective gene-silencing system for *H. contortus* could lead to a major and rapid advance of our understanding of orphan CAP proteins and their genes, and could help establish their actual involvement in biological and developmental pathways in this important parasite as well as parasite-host interactions at the immuno-molecular level. Such fundamental insights could pave the way to the design of new interventions, such as vaccines, against haemonchosis and other nematodiases of animals and humans. In our opinion, this focus has particular merit, given the recent success in immunizing animals against challenge infection with strongyloid nematodes with selected CAP proteins (Hewitson et al., 2015; Vlamick et al., 2015).
Acknowledgements

Research funding from the National Health and Medical Research Council and Australian Research Council is gratefully acknowledged; support from the Victorian Life Sciences Computation Initiative (VLSCI) is also acknowledged (RBG). We acknowledge the contributions of all staff at WormBase (www.wormbase.org). P.W.S. thanks the Howard Hughes Medical Institute (HHMI) and the National Institutes of Health (NIH). The authors thank Christopher Lee for some assistance with bioinformatics and the collection of some articles.

References


Little PR, Hodge A, Maeder SJ, Wirtherle NC, Nicholas DR, Cox GG, Conder GA. Efficacy of a combined oral formulation of derquantel-abamectin against the adult and larval stages of nematodes in sheep, including anthelmintic-resistant strains. Vet Parasitol 2011;181:180-93.
Lok JB. Nucleic acid transfection and transgenesis in parasitic nematodes. Parasitology 2012;139:574-88.


Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assemblies of short reads to eliminate gaps. Genome Biol 2010;11:R41.


Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amantadites P, Ballew RM, Huson DH, Wortman JR, Zhang Q,


Table 1. Full-length transcripts representing the different CAP protein genes (sd: single domain; dd: double domain; td: triple domain) and the genomic scaffolds to which they map, including the length of inferred amino acid sequences and InterProScan matches (including domains and motifs) for each predicted protein.

Fig. 1. (A) The phylogenetic relationships of CAP proteins predicted from the draft genome and transcriptomes of H. contortus based on Bayesian inference. Groups 1 to 5 are indicated; single-domain [sd], double-domain [dd] and triple-domain [td] proteins. The posterior probability (pp) supporting each clade is indicated. (B) Heatmap displaying transcription profiles for genes Hc-cap-1 to Hc-cap-45 (cf. Table 1). Transcription levels (see colour scale): low (red), medium (orange), high (yellow), very high (white).

Fig. 2. Summary of criteria for the classification of Hc-CAP proteins.

Fig. 3. (A) The homology model of Hc-CAP-15 illustrates the characteristic α-β-α sandwich fold of CAP (= SCP/TAPS) proteins. In case of Hc-CAP-15, this central domain is preceded by a rather extended N-terminal moiety (blue). The location of the four CAP motifs is indicated by pale colours and labels. Disulphide bonds are shown as explicit stick models, and the LON-1 specific disulphide bridges are indicated. (B) The Hc-CAP-15 model is shown as surface representation and colour mapped by the electrostatic surface potential (red: negative, blue: positive). The left image is the same orientation as the cartoon model in (A), the image in (B) shows the view after a 180° turn around the vertical axis. Images were rendered with PyMOL (De Lano, 2002), and the electrostatic potential was calculated using PDB2PQR and APBS (Dolinsky et al., 2007).
Table 1
<table>
<thead>
<tr>
<th>Group</th>
<th>Domain</th>
<th>CAP4</th>
<th>CAP3</th>
<th>CAP1</th>
<th>CAP2</th>
<th>No. of Hc-CAP proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td><strong>Conserved</strong></td>
<td>Conserved</td>
<td>Variable</td>
<td>Variable</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td><strong>Variable</strong></td>
<td><strong>Absent</strong></td>
<td>Variable</td>
<td>Variable</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td><strong>Conserved</strong></td>
<td>Variable</td>
<td>Variable</td>
<td>Variable</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>N-terminus</td>
<td>Conserved</td>
<td>Variable</td>
<td>Variable</td>
<td>Variable</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>C-terminus</td>
<td>Conserved</td>
<td>Variable</td>
<td>Variable</td>
<td>Variable</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>N-terminus</td>
<td>Conserved</td>
<td>Variable</td>
<td>Variable</td>
<td>Variable</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>Conserved</td>
<td><strong>Conserved</strong></td>
<td>Variable</td>
<td>Variable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C-terminus</td>
<td>Conserved</td>
<td>Variable</td>
<td>Variable</td>
<td>Variable</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3

Supplementary Table 1. Pairwise sequence comparisons of 45 predicted CAP protein sequences with *Caenorhabditis elegans* and other parasitic nematodes including *Ancylostoma ceylanicum* (Ace), *Ancylostoma caninum* (Aca), *Ancylostoma duodenale* (Ad), *Necator americanus* (Na), *Ostertagia ostertagi* (Oso), *Onchocerca*
volvulus (Onv), Heligosomoides bakeri/Heligosomoides polygyrus (Hb/Hp), Wuchereria bancrofti (Wb), Dirofilaria immitis (Di) and Brugia malayi (Bm) with percentage identity (%) indicated in brackets.

**Supplementary Fig. 1.** Structure-based amino acid sequence alignments of predicted CAP proteins for individual groups (cf. Fig. 1) reveals relative sequence conservation within groups, and highlights the conservation of secondary structure elements as well as the canonical cysteine residues important for stabilising the tertiary structure. In each alignment, the secondary structure elements are mapped, with α-helices shown in green and β-strands in red; cysteine residues are highlighted in yellow. The first line in each alignment indicates the locations of the different CAP motifs. The alignments were constructed using the program SBAL (Wang et al., 2012) employing results from secondary structure prediction with the program PSIPRED (Bryson et al., 2005).

**Supplementary Fig. 2.** Genetic interactions of *Caenorhabditis elegans* orthologs (yellow) representing CAP protein–encoding genes of *Haemonchus contortus*. Of a total of 45 CAP protein-encoding genes of *H. contortus* (Supplementary Table 1), 13 have *C. elegans* orthologs, of which *Ce*-vap-1 (representing *Hc*-cap-36 to *Hc*-cap-41), *Ce*-vap-2 (representing *Hc*-cap-42 and *Hc*-cap-43), *Ce*-scl-5 (representing *Hc*-cap-14) and *Ce*-lon-1 (representing *Hc*-cap-15 and -16) have predicted (orange) or experimentally proven (green) interactions with other key genes. Specifically, *Ce*-vap-1 interacts with at least 17 *Ce*-scl paralogs (*Ce*-scl-1, -2, -3, -5, -6, -7, -8, -9, -10, -11, -12, -13, -14, -15, -17, -18, -20; grey). Although *Ce*-vap-1 and *Ce*-vap-2 do not directly interact with each other, the 17 *Ce*-scl paralogs are predicted to interact with *Ce*-vap-2 that also interacts specifically with egl-9 (involved in egg laying). The independent interactions of *Ce*-vap-1 and *Ce*-vap-2 with *Ce*-scl-1, which also interacts with components of the insulin-like signaling pathway, including *Ce*-daf-2, *Ce*-daf-16 and *Ce*-age-1 (panel A) (cf. Supplementary Fig. 2) suggest a critical and integrated involvement of these genes in the regulation of nematode growth and development. *Ce*-scl-5 is also predicted to interact directly with a transcription factor enhancer encoded by the orphan gene F47B8.2. The *lon*-1 gene has numerous, mostly proven interactions with genes including *Sma*-2, -3, -4 (complex), -6, -9, -10, -12, -13, -14, -16, -17, -18 and -19 encoding various Smad proteins, and *dbl*-1, *daf*-4, *kin*-29; *rnt*-1, *crm*-1, *lgg*-1 and *che*-2 (panel B), most of which are involved up or down-stream of the TGF-beta signaling pathway. Information extracted from WormBase (www.wormbase.org).

**Supplementary Fig. 3.** Phylogenetic relationships of selected predicted single-domain (A) and double-domain (B) CAP proteins of *Haemonchus contortus* with homologues from *Caenorhabditis elegans*.
### Supplementary Table 1.

<table>
<thead>
<tr>
<th>Gene code</th>
<th>Genomic scaffold</th>
<th>Length of inferred amino acid sequence</th>
<th>Category/homologue/ percentage identity</th>
<th>Other paralogous remnant (percentage identity)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single-domain</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>He-cap-1</td>
<td>S14934</td>
<td>222</td>
<td>-</td>
<td>NHC4 (94.14); AIVASP1 (51.84); AIVASP2 (51.64); RASAASP (46.89)</td>
</tr>
<tr>
<td>He-cap-2</td>
<td>S13677</td>
<td>221</td>
<td>-</td>
<td>NHC4 (89.59); AIVASP1 (95.47); AIVASP2 (90.98); AIVASP3 (75.59)</td>
</tr>
<tr>
<td>He-cap-3</td>
<td>S17432</td>
<td>220</td>
<td>-</td>
<td>NHC4 (88.18); AIVASP1 (90.78); AIVASP2 (89.70); AIVASP3 (85.41)</td>
</tr>
<tr>
<td>He-cap-4</td>
<td>S10126</td>
<td>221</td>
<td>-</td>
<td>NHC4 (69.55); AIVASP1 (48.83); AIVASP2 (48.60); RASAASP (41.45)</td>
</tr>
<tr>
<td>He-cap-5</td>
<td>S19218</td>
<td>221</td>
<td>-</td>
<td>NHC4 (68.18); AIVASP1 (49.30); AIVASP2 (47.66); AIVASP3 (44.98)</td>
</tr>
<tr>
<td>He-cap-6</td>
<td>S19671</td>
<td>184</td>
<td>-</td>
<td>NHC4 (65.91); AIVASP1 (90.23); AIVASP2 (46.60); RASAASP (41.45)</td>
</tr>
<tr>
<td>He-cap-7</td>
<td>C180107</td>
<td>184</td>
<td>-</td>
<td>AIVASP1 (58.33); NHC4 (95.54); AIVASP2 (40.62); AIVASP3 (48.62)</td>
</tr>
<tr>
<td>He-cap-8</td>
<td>C275669</td>
<td>226</td>
<td>-</td>
<td>OanASP1 (59.81); NHC4 (95.54); AIVASP2 (40.62); AIVASP3 (48.62)</td>
</tr>
<tr>
<td>He-cap-9</td>
<td>C469225</td>
<td>213</td>
<td>-</td>
<td>OanASP2 (48.38); AIVASP1 (46.89); AIVASP2 (46.89); RASAASP (46.34)</td>
</tr>
<tr>
<td>He-cap-10</td>
<td>S13971</td>
<td>216</td>
<td>-</td>
<td>OanASP1 (52.58); NASAASP1 (51.46); AIVASP1 (49.29); AIVASP2 (47.17)</td>
</tr>
<tr>
<td>He-cap-11</td>
<td>S3892</td>
<td>225</td>
<td>-</td>
<td>OanASP1 (59.81); AIVASP2 (86.71); AIVASP3 (98); NASAASP (49.51)</td>
</tr>
<tr>
<td>He-cap-12</td>
<td>S6853</td>
<td>233</td>
<td>-</td>
<td>OanASP1 (51.37); NASAASP (47.57); NHC4 (46.66); AIVASP (46.51)</td>
</tr>
<tr>
<td>He-cap-13</td>
<td>S4660</td>
<td>298</td>
<td>CcSOC-L1 (93.54)</td>
<td>NASAASP (36.36); HNLV1L0 (53.49); AIVASP (32.37)</td>
</tr>
<tr>
<td>He-cap-14</td>
<td>S8162</td>
<td>241</td>
<td>CcSOC-L1 (93.54)</td>
<td>AIVASP (29.87); HNLV1L0 (29.28); OanASP1 (26.93)</td>
</tr>
<tr>
<td>He-cap-15</td>
<td>S9916</td>
<td>297</td>
<td>CcLON1a (64.53)</td>
<td>-</td>
</tr>
<tr>
<td>He-cap-16</td>
<td>S18985</td>
<td>297</td>
<td>CcLON1a (64.53)</td>
<td>-</td>
</tr>
<tr>
<td>He-cap-17</td>
<td>S4662</td>
<td>275</td>
<td>FSTB7.2b (52.73)</td>
<td>-</td>
</tr>
<tr>
<td>He-cap-18</td>
<td>S12237</td>
<td>214</td>
<td>-</td>
<td>OanASP1 (52.2); OanASP2 (52.04); OanASP3 (51.71); NHC4 (51.76)</td>
</tr>
<tr>
<td>He-cap-19</td>
<td>C460753</td>
<td>280</td>
<td>-</td>
<td>OanASP1 (54.91); OanASP2 (52.55); CcSOC-L2a (52); OanASP3 (51.94)</td>
</tr>
<tr>
<td>He-cap-20</td>
<td>C460243</td>
<td>214</td>
<td>-</td>
<td>OanASP1 (52.66); OanASP2 (52.52); OanASP3 (52.25); NHC4 (50.73)</td>
</tr>
<tr>
<td>He-cap-21</td>
<td>S11441</td>
<td>240</td>
<td>-</td>
<td>OanASP1 (56.06); HNLV1L1 (55.31); OASAASP (55.51); NASAASP (54.42)</td>
</tr>
<tr>
<td>He-cap-22</td>
<td>S12697</td>
<td>221</td>
<td>-</td>
<td>WbASP1 (27.95); BrHLK1 (26.88); OASAASP (26.34); OASAASP (24.19)</td>
</tr>
<tr>
<td>He-cap-23</td>
<td>S18689</td>
<td>221</td>
<td>-</td>
<td>WbASP1 (28.49); BrHLK1 (27.42); OASAASP (26.88); OASAASP (24.73)</td>
</tr>
<tr>
<td>He-cap-24</td>
<td>S11756</td>
<td>226</td>
<td>-</td>
<td>WbASP1 (29.95); BrHLK1 (27.87); OASAASP (25.83); OASAASP (24.73)</td>
</tr>
<tr>
<td>He-cap-25</td>
<td>S6155</td>
<td>206</td>
<td>-</td>
<td>HNLV1L1 (53.55); OASAASP (24.74); BrHLK1 (24.73); OASAASP (24.73)</td>
</tr>
<tr>
<td>He-cap-26</td>
<td>S3375</td>
<td>236</td>
<td>-</td>
<td>HNLV1L1 (53.48); HNLV1L1 (33.31); OASAASP (32.35); OASAASP (31.96)</td>
</tr>
<tr>
<td>He-cap-27</td>
<td>S2323</td>
<td>241</td>
<td>-</td>
<td>HNLV1L1 (53.81); OASAASP (35.41); OASAASP (34.45); OASAASP (34.13)</td>
</tr>
<tr>
<td>He-cap-28</td>
<td>S10149</td>
<td>216</td>
<td>-</td>
<td>Nasaasp1 (51.96); AIVASP1 (50.24); AIVASP2 (49.13); NHC4 (41.54)</td>
</tr>
<tr>
<td>He-cap-29</td>
<td>S16472</td>
<td>252</td>
<td>-</td>
<td>Nasaasp1 (49.49); AIVASP1 (46.58); AIVASP2 (45.34); OanASP1 (45.07)</td>
</tr>
<tr>
<td>He-cap-30</td>
<td>S2052</td>
<td>100</td>
<td>-</td>
<td>Nasaasp1 (51.07); AIVASP1 (29.69); CcSOC-L2b (29.44); OanASP2 (27.44)</td>
</tr>
<tr>
<td>He-cap-31</td>
<td>S764</td>
<td>298</td>
<td>-</td>
<td>Nasaasp1 (29.81); OanASP1 (29.83); AIVASP1 (24.44); CcSOC-L3 (28.34)</td>
</tr>
<tr>
<td>He-cap-32</td>
<td>C471545</td>
<td>273</td>
<td>-</td>
<td>Nasaasp1 (29.81); OanASP1 (29.83); AIVASP1 (24.44); CcSOC-L3 (28.34)</td>
</tr>
<tr>
<td>He-cap-33</td>
<td>S17928</td>
<td>245</td>
<td>-</td>
<td>Nasaasp1 (34.95); OanASP1 (33.17); AIVASP1 (32.89); CcSOC-L3 (32.49)</td>
</tr>
<tr>
<td><strong>Double-domain</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>He-cap-34</td>
<td>S8316</td>
<td>459</td>
<td>-</td>
<td>HNLV1L1 (92.99); OASAASP1 (34.98); RASAASP1 (32.69); CcVAPI (32.95)</td>
</tr>
<tr>
<td>He-cap-35</td>
<td>S4883</td>
<td>449</td>
<td>-</td>
<td>HNLV1L1 (58.78); CcVAPI (53.86); OASAASP1 (53.66); RASAASP1 (53.66)</td>
</tr>
<tr>
<td>He-cap-36</td>
<td>S10599</td>
<td>431</td>
<td>CcVAPI (47.7)</td>
<td>AIVASP1 (65.99); AIVASP1 (44.34); RASAASP1 (63.98); AIVASP1 (46.51)</td>
</tr>
<tr>
<td>He-cap-37</td>
<td>S17282</td>
<td>431</td>
<td>CcVAPI (46.2)</td>
<td>AIVASP1 (61.72); AIVASP1 (46.07); AIVASP1 (45.98); AIVASP1 (45.51)</td>
</tr>
<tr>
<td>He-cap-38</td>
<td>S7352</td>
<td>429</td>
<td>CcVAPI (44.7)</td>
<td>AIVASP1 (61.38); AIVASP1 (39.2); AIVASP2 (34.28); AIVASP3 (35.02)</td>
</tr>
<tr>
<td>He-cap-39</td>
<td>S5869</td>
<td>453</td>
<td>CcVAPI (43.8)</td>
<td>AIVASP1 (52.83); RASAASP1 (52.72); AIVASP2 (52.59)</td>
</tr>
<tr>
<td>He-cap-40</td>
<td>S18884</td>
<td>459</td>
<td>CcVAPI (44.1)</td>
<td>AIVASP1 (52.83); RASAASP1 (52.72); AIVASP2 (52.59); AIVASP3 (52.36)</td>
</tr>
<tr>
<td>He-cap-41</td>
<td>C460005</td>
<td>462</td>
<td>CcVAPI (42.5)</td>
<td>AIVASP1 (55.56); AIVASP1 (49.76); AIVASP2 (49.32); AIVASP3 (49.16)</td>
</tr>
<tr>
<td>He-cap-42</td>
<td>S14689</td>
<td>380</td>
<td>CcVAPI (44.4)</td>
<td>AIVASP1 (39.25); AIVASP1 (38.67); RASAASP1 (37.29)</td>
</tr>
<tr>
<td>He-cap-43</td>
<td>S19722</td>
<td>394</td>
<td>CcVAPI (44.4)</td>
<td>AIVASP1 (39.25); AIVASP1 (38.67); RASAASP1 (37.29)</td>
</tr>
<tr>
<td>He-cap-44</td>
<td>S17612</td>
<td>429</td>
<td>-</td>
<td>HNLV1L1 (34.18); HNLV1L2 (29.56); AIVASP1 (26.75); CcVAPI (25.12)</td>
</tr>
<tr>
<td>He-cap-45</td>
<td>S18155</td>
<td>700</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### Group 2

**CAP motif**

<table>
<thead>
<tr>
<th>Hc-CAP-23 (S6315)</th>
<th>33333</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hc-CAP-24 (S1756)</td>
<td>46</td>
</tr>
<tr>
<td>Hc-CAP-21 (S2679)</td>
<td>69</td>
</tr>
<tr>
<td>Hc-CAP-23 (S1880)</td>
<td>69</td>
</tr>
</tbody>
</table>

**CAP motif**

<table>
<thead>
<tr>
<th>Hc-CAP-25 (S6315)</th>
<th>113</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hc-CAP-24 (S1756)</td>
<td>122</td>
</tr>
<tr>
<td>Hc-CAP-22 (S2679)</td>
<td>121</td>
</tr>
<tr>
<td>Hc-CAP-23 (S1880)</td>
<td>121</td>
</tr>
</tbody>
</table>

**CAP motif**

<table>
<thead>
<tr>
<th>Hc-CAP-23 (S6315)</th>
<th>179</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hc-CAP-24 (S1756)</td>
<td>185</td>
</tr>
<tr>
<td>Hc-CAP-22 (S2679)</td>
<td>184</td>
</tr>
<tr>
<td>Hc-CAP-23 (S1880)</td>
<td>184</td>
</tr>
</tbody>
</table>

**CAP motif**

<table>
<thead>
<tr>
<th>Hc-CAP-25 (S6315)</th>
<th>204</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hc-CAP-24 (S1756)</td>
<td>226</td>
</tr>
<tr>
<td>Hc-CAP-22 (S2679)</td>
<td>221</td>
</tr>
<tr>
<td>Hc-CAP-23 (S1880)</td>
<td>221</td>
</tr>
</tbody>
</table>
Group 3

CAP motif

HcCAP-8 (C275689)  
HcCAP-26 (S15375)  
HcCAP-21 (S15441)  
HcCAP-27 (S4860)  
HcCAP-20 (C46243)  
HcCAP-18 (S12237)  
HcCAP-30 (S2052)  
HcCAP-14 (S4860)  
HcCAP-31 (S764)  
HcCAP-32 (C471545)  
HcCAP-19 (C460373)  
HcCAP-33 (S17928)  

HcCAP-8 (C275689)  
HcCAP-26 (S15375)  
HcCAP-21 (S15441)  
HcCAP-27 (S4860)  
HcCAP-20 (C46243)  
HcCAP-18 (S12237)  
HcCAP-30 (S2052)  
HcCAP-14 (S4860)  
HcCAP-31 (S764)  
HcCAP-32 (C471545)  
HcCAP-19 (C460373)  
HcCAP-33 (S17928)  

4646

Group 3
Supplementary Fig. 2

A

B
Supplementary Fig. 3

Hc-CAP-17 (S4662)
  └ CeF57B7.2b
    └ Hc-CAP-16 (S19085)
        └ CeLON-1a
            └ Hc-CAP-15 (S9916)
                └ Hc-CAP-14 (S8162)
                    └ CeSCL-5
                        └ Hc-CAP-13 (S4860)
                            └ CeSCL-19
                                └ CeVAP-1b
                                    └ Hc-CAP-36 (S10569)
                                        └ Hc-CAP-37 (S13282)
                                            └ Hc-CAP-38 (S7292)
                                                └ Hc-CAP-39 (S5869)
                                                    └ Hc-CAP-40 (S18884)
                                                        └ Hc-CAP-41 (C464093)
                                                            └ Hc-CAP-42 (S14889)
                                                                └ Hc-CAP-43 (S19722)
                                                                └ CeVAP-2