Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa de Doutorado em Bioinformática

A análise comparativa de proteínas secretadas em helmintos revela diferenças de acordo com diferentes estilos de vida e hospedeiros

Belo Horizonte
2015

Yesid Cuesta Astroz

A análise comparativa de proteínas secretadas em helmintos revela diferenças de acordo com diferentes estilos de vida e hospedeiros

Tese apresentada como requisito parcial para a obtenção do título de Doutor em Bioinformática pelo Programa de Pós-graduação em Bioinformática na Universidade Federal de Minas Gerais.

Orientador: Dr. Guilherme Corrêa de Oliveira

Coorientadora: Dra. Laila Alves Nahum

Belo Horizonte
2015

AGRADECIMENTOS

Ao NIH que através do projeto “Infectious Disease Genomics and Bioinformatics Training in Brazil" (D043-TW007012) me forneceu a bolsa de doutorado durante três anos. Ao Programa de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais por disponibilizar uma bolsa da CAPES para conclui-lo.

Agradeço ao meu orientador Dr. Guilherme Oliveira, pela oportunidade concedida e por contribuir ao meu aperfeiçoamento como pesquisador, fornecendo múltiplas oportunidades para interagir com uma ampla rede de pesquisadores nacionais e internacionais. À minha co-orientadora Dra. Laila Nahum pela orientação, e pelo tempo investido na construção e estruturação de um projeto de pesquisa. Adicionalmente gostaria de agradecê-la pelos conselhos ao longo do meu doutorado. À Dra. Ângela Volpini pela coordenação de múltiplos aspectos logísticos durante a execução do meu projeto.

Ao Programa de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais e ao pessoal administrativo pela ajuda e contribuição à minha formação.

Aos colegas do CEBio e Grupo de Genômica e Biologia Computacional do CPqRR, pela ajuda incalculável durante estes quatro anos de doutorado. Em especial, agradeço aos meus colaboradores durante o projeto: Larissa, Fabiano e Francislon. Agradeço a eles pelas discussões e valiosas contribuições ao meu projeto de pesquisa. Ao Anderson e Fausto pelo suporte técnico.

Ao Dr. Lars Juhl Jensen do Disease Systems Biology Group do NNF Center for Protein Research em Copenhagen, Dinamarca. Pela sua orientação no projeto do interatoma *S. mansoni – H. sapiens*. Ao seu estudante de doutorado Alberto Santos pela amizade, pelas valiosas contribuições no projeto e pelas horas de Skype dedicadas às interações e por ter acreditado nas minhas idéias.

Aos meus amigos, Luis, Fernando, Marinely, Jorge, Vicky, Jose Manuel, obrigado pelo apoio, ajuda, amabilidade e os bons tempos compartilhados em Belo Horizonte.

Aos meus pais e irmã por tudo.

À família Castañeda Gómez pelo apoio brindado.

À Eugenia Castañeda pelo apoio e amor brindado e ao meu filho Martin Cuesta Castañeda

SUMÁRIO

ÍNDICE DE FIGURAS

RESUMO

Análises comparativas dos genes e genomas parciais ou completamente sequenciados em helmintos são importantes para entender a diversidade genômica e evolução dos parasitos e seus hospedeiros em relação às diversas pressões seletivas em seus habitats. Proteínas secretadas pelos parasitos são capazes de modificar o ambiente do hospedeiro e modular seu sistema imune. Pouco se conhece sobre a composição e variabilidade das proteínas secretadas por diferentes espécies de helmintos, além da sua contribuição para o avanço da infecção. Neste projeto, foi feita a predição *in silico* do secretoma em 44 espécies de helmintos (Nematoda: 31 espécies, Platyhelminthes: 13 espécies) visando a compreender a diversidade e evolução dos secretomas. Os resultados indicaram proteínas secretadas associadas com processos de infecção, invasão, adesão e imunoregulação, como inibidores de proteases e citocinas, dentre outras. Analisamos também os domínios proteicos e o conteúdo de arquiteturas nas proteínas secretadas para identificar características específicas do nicho ou hospedeiro. Além disso, foram identificados homólogos da família de cistatinas em três espécies de *Schistosoma* e outros Platyhelminthes. As cistatinas são uma família de inibidores de cisteíno proteases que fazem parte do secretoma. A reconstrução das relações evolutivas destas proteínas permitiu observar sua diversidade no nível molecular e eventos de duplicação gênica moldando a evolução das mesmas ao longo do tempo. Em conjunto, o desenvolvimento deste projeto contribuiu para o conhecimento da biologia de helmintos, incluindo aspectos da interação parasito-hospedeiro. Futuramente, poderemos propor possíveis novos alvos moleculares para o tratamento ou diagnóstico das helmintíases.

Palavras-chave: filogenômica, secretoma, helmintos, biodiversidade, bioinformática, cistatinas.

ABSTRACT

Comparative analyses of partially or completely sequenced genes and genomes in helminths are important to understand the genomic diversity and evolution of parasites and their hosts in terms of different selective pressures in their habitats. Proteins secreted by parasites are able to modify the host's environment and modulate their immune system. Little is known about the composition and variability of secreted proteins in different helminth species besides its contribution in the progress of the infection. In this project we predicted the *in silico* secretome across 44 helminth species (Nematoda: 31 species, Platyhelminthes: 13 species) aiming to understand the diversity and evolution of secretomes. Our results indicated secreted proteins associated with biological processes such as: infection, invasion, adhesion, immunoregulation (protease inhibitors, cytokines), among others. We also analyzed protein domains and domain architectures in secreted proteins to identify specific signatures associated with niches or hosts. Furthermore, cystatin family homologs were identified in three *Schistosoma* species and other Platyhelminthes. Cystatins are a family of cysteine protease inhibitors which are part of the secretome. The reconstruction of evolutionary relationships of these proteins allowed us to observe their diversity at the molecular level and gene duplications events, shaping their evolution over time. In summary, the development of this project contributed to the understanding of of helminth biology, including aspects of the host-parasite interaction. In the future, it will be possible to identify new molecular targets for the treatment or diagnosis of helminthiases.

Keywords: phylogenomics, secretome, helminths, biodiversity, bioinformatics, cistatins.

I – INTRODUÇÃO

1.1 – Biologia dos helmintos

Os helmintos representam um grupo polifilético que inclui os filos Nematoda (nematódeos) e Platyhelminthes (platelmintos). A maioria dos helmintos apresenta ciclo de vida complexo, incluindo estágios de vida livre e parasitária. Em algumas espécies, o estágio assexuado passa por um hospedeiro intermediário, tais como gastrópodes ou insetos. Os helmintos podem ser parasitos de importância socioeconômica e responsáveis por infecções parasitárias em humanos, animais e plantas. Segundo a Organização Mundial de Saúde (World Health Organization - WHO), mais de dois bilhões de pessoas sofrem de helmintíases e muitas outras estão em risco de contrair estas doenças, especialmente em países em desenvolvimento; não obstante, recentemente a esquistossomose causada por *S. haematobium* emergiu na Europa, na ilha francesa da Córsega (Boissier et al. 2016).

Como definido pelo conceito de DALYs ("disease burden as definided by disability-adjusted lifeyears"), a esquistossomose, junto com a ancilostomose e a leishmaniose, estão entre as doenças tropicais negligenciadas com o maior peso epidemiológico quantificado em DALYs (Hotez et al. 2014; Merrifield et al. 2016). A maior parte dos vermes intestinais transmitidos pelo solo (geo-helmintos) são nematódeos incluindo as filárias que causam filariose linfática (elefantíase) e oncocercose. Os platelmintos incluem todas as espécies de Trematoda e Cestoda, causadores da esquistossomose e cisticercose, dentre outras. Infecções crônicas de helmintos estão associadas com anemia, atraso no crescimento, desnutrição, fadiga e comprometimento do desenvolvimento cognitivo (Hotez et al. 2008; Lustigman et al. 2012).

Por outro lado infeções experimentais com helmintos podem limitar a severidade da doença em modelos murinos de artrites (Salinas-Carmona et al. 2009), diabetes tipo 1 (Mishra et al. 2013; Osada et al. 2013), colites (Weinstock 2006), esclerose múltipla (Correale and Farez 2011) e inflamação alérgica das vias respiratórias (Wilson et al. 2005). Tais descobertas aumentaram o interesse do uso desses parasitos e seus produtos de secreção (proteínas), para o tratamento de doenças inflamatórias. Consequentemente, helmintos vivos são atualmente empregados em pelo menos 15 ensaios clínicos em esforços para aliviar doenças alérgicas e autoimunes (Khan and Fallon 2013).

Em plantas, os helmintos causam danos aos cultivos levando a grandes perdas econômicas. Alguns dos nematódeos de plantas mais prejudiciais são os endoparasitos dos gêneros *Meloidogyne spp*, *Heterodera spp.* e *Globodera spp.* (Mehta et al. 2008). Tais organismos invadem as raízes no estágio de larvas juvenis levando a reduções significativas na absorção de água e nutrientes, culminando com a morte das plantas.

O genoma nuclear de diferentes espécies de helmintos de vida livre e parasitária, que afetam humanos e outras espécies de importância agrícola e veterinária, estão sendo sequenciados. Dados do genoma, transcritoma e proteoma predito de helmintos têm fornecido informações importantes para a compreensão dos mecanismos moleculares envolvidos em seu metabolismo, interação parasito-hospedeiro, evasão do sistema imune, evolução molecular, dentre outras abordagens (Brindley et al. 2009; Lustigman et al. 2012; Tsai et al. 2013; Zarowiecki and Berriman 2015). Os genomas nuclear e mitocondrial das três espécies socioeconomicamente mais importantes de *Schistosoma* (*S. haematobium*, *S. japonicum* e *S. mansoni*) foram sequenciados (Berriman et al. 2009; Young et al. 2012; Zhou et al. 2009). A análise de dados genômicos e transcritômicos destas espécies tem aberto novas fronteiras no estudo e controle da esquistossomose (Mourao et al. 2012; Nahum et al. 2012).

1.2 - Genômica comparativa e evolução de helmintos

Abordagens de genômica comparativa levam a importantes descobertas e mostram como as comparações entre gêneros ou espécies podem ajudar a elucidar como os patógenos evoluem em fenótipos específicos que lhes permitam adaptar-se a um novo hospedeiro. Além disso, a identificação de características específicas associadas ao parasitismo requer comparação com os genomas das espécies não parasitas (Jackson et al. 2016). A evolução do parasitismo é um evento recorrente na história da vida e um problema central na biologia evolutiva além de ser uma das áreas onde a genômica comparativa atua.

A maior parte dos métodos de predição funcional depende da análise de similaridade de sequências entre o gene de interesse e os genes com informação funcional disponíveis na literatura (Bork et al. 1998; Nahum et al. 2008; Sjolander 2010). Não obstante, a similaridade de sequências não garante que as mesmas tenham funções idênticas, sendo este método insuficiente para atribuir uma função predita a um gene não caracterizado experimentalmente.

A filogenômica corresponde à interseção entre filogenética e genômica, usando informações evolutivas na predição funcional de genes e produtos gênicos não caracterizados experimentalmente (Eisen et al. 1997; Eisen 1998). Essa abordagem permite melhorar a predição funcional em relação aos métodos baseados estritamente em similaridade, fornecendo uma plataforma comparativa robusta no contexto evolutivo (Eisen 1998; Nahum et al. 2008; Sjolander 2010). Os estudos na área da filogenômica e genômica comparativa oferecem uma visão integrada de um sistema biológico. No caso do estudo de helmintos, esta abordagem é muito importante, pois permite investigar os processos evolutivos no nível molecular, a origem da sua diversidade e evolução da relação parasito-hospedeiro.

Em helmintos encontram-se poucos genes e produtos gênicos caracterizados experimentalmente até o momento. Neste contexto, análises comparativas dos genes e genomas parcial ou completamente sequenciados e reconstrução filogenética dos mesmos permitem a identificação de elementos funcionais conservados entre espécies a partir da identificação das suas sequências e funções homólogas (Mitreva 2012; Tsai et al. 2013). Além disso, tais estudos podem mostrar como a evolução moldou os genomas nas diferentes espécies ao longo do tempo evolutivo, para nos ajudar a compreender sua história e contribuir para a anotação funcional de genomas, genes e seus produtos (Eisen 1998; Tsai et al. 2013; Sjolander 2010).

Os resultados do nosso grupo no CPqRR apontam para excelentes perspectivas quanto à análise filogenômica e genômica comparativa de famílias gênicas específicas e análises em larga escala do proteoma predito de *S. mansoni* e outros helmintos. Na reconstrução do filoma de *S. mansoni,* foi realizada a predição funcional baseada em ortologia de 5,507 proteínas do parasito, sendo 956 previamente desconhecidas (Silva et al. 2012). Esta visão global das relações evolutivas do *S. mansoni* forneceu *insights* sobre o estilo de vida parasitária.

As endopeptidases são outro alvo de estudo do nosso grupo. Foram identificadas famílias de endopeptidases expandidas a partir de eventos de duplicação gênica no proteoma predito de *S. mansoni* com relação a outros metazoários (Silva et al. 2011). Este estudo permitiu identificar adaptações potencialmente relacionadas à vida parasitária. Enzimas modificadoras de histonas de diferentes espécies de *Schistosoma* também têm sido estudadas pelo nosso grupo visando à identificação de alvos terapêuticos contra a esquistossomose (manuscrito em preparação).

1.3 - Proteínas secretadas em helmintos

As proteínas secretadas por uma célula cumprem um papel essencial desde bactérias até mamíferos (Tjalsma et al. 2000). Tais proteínas podem representar entre 8 e 20% do proteoma de um organismo (Greenbaum et al. 2001). As proteínas incluídas no secretoma pertencem a diversas classes funcionais, tais como: citocinas, hormônios, enzimas digestivas, anticorpos, proteases, toxinas, peptídeos antimicrobianos e proteínas associadas ao estresse oxidativo. Algumas delas estão envolvidas em processos biológicos vitais, como adesão celular, migração celular, comunicação célula-célula, diferenciação, proliferação, morfogênese e regulação da resposta imune (Maizels and Yazdanbakhsh 2003). As proteínas secretadas por parasitos são de particular interesse para a compreensão das interações parasito-hospedeiro, uma vez que podem regular a resposta imune do hospedeiro e causar doenças (Maizels and Yazdanbakhsh 2003).

A secreção de proteínas através do retículo endoplasmático está associada a um peptídeo sinal na região N-terminal das mesmas e representa a via clássica de secreção. Algumas proteínas secretadas não contêm peptídeo sinal sendo secretadas pela via não-clássica, por meio de vesículas extracelulares (Figura 1). As proteínas secretadas por esta via têm sido identificadas como enzimas glicolíticas, chaperonas, fatores de tradução, dentre outras, sugerindo que estas proteínas possam ser multifuncionais (*moonlighting proteins*) (Nombela et al. 2006). Trabalhos de microscopia mostraram evidências experimentais da existência de vesículas extracelulares em helmintos, especificamente nos trematódeos *Echinostoma caproni* e *Fasciola hepatica.* Estas vesículas são ativamente liberadas pelos parasitos e podem ser captadas pelas células hospedeiras, além de conter proteínas do secretoma que desempenham um papel importante na interação parasito-hospedeiro (Marcilla et al. 2012).

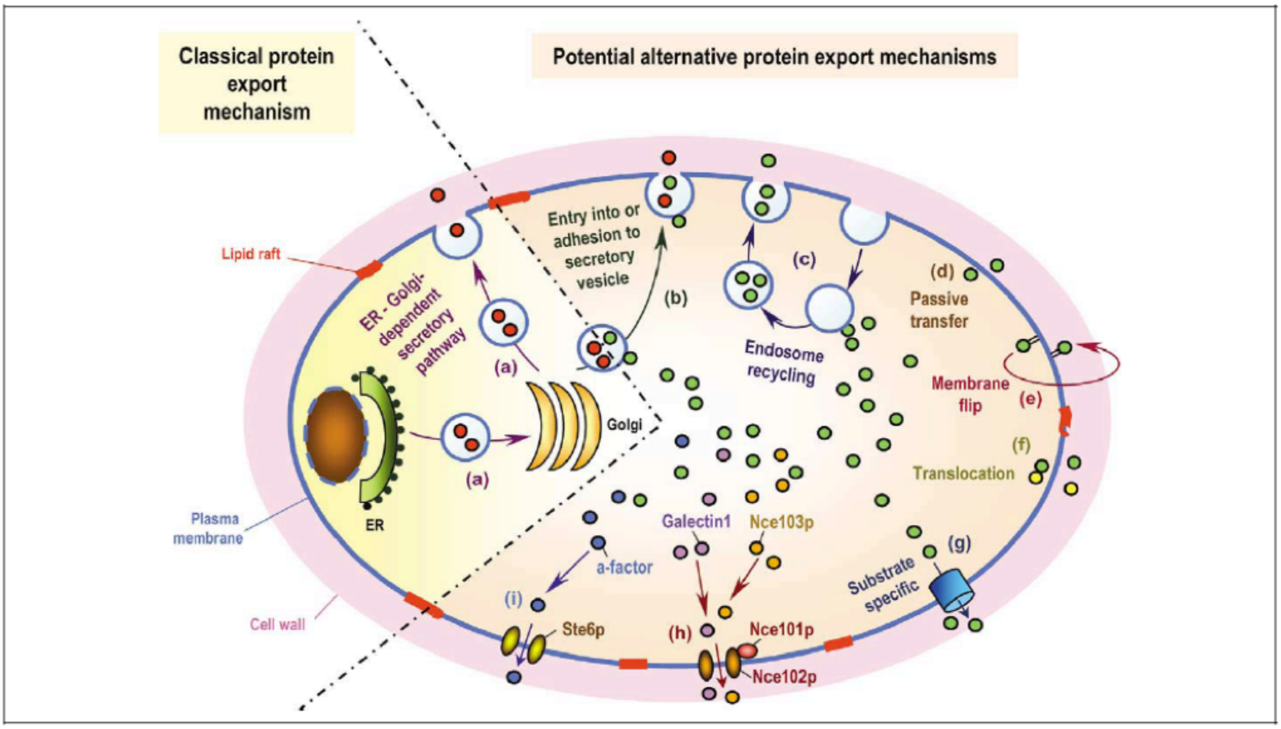

Figura 1. Mecanismos de secreção de proteínas pelas vias clássica e alternativa. Via clássica de secreção envolvendo o retículo endoplasmático e Golgi (a), possíveis vias alternativas de secreção de proteínas através de vesículas (b), subcompartimentos endossomais (c), transferência passiva pela membrana (d), *flipping* de membrana (e), translocação (f) e reconhecimento específico de substrato (g). Baseado em evidências obtidas a partir de leveduras, mamíferos e parasitos. Fonte: Nombela et al. 2006.

As proteínas secretadas podem ser conservadas entre parasitos que compartilham um nicho e podem ser também compartilhadas entre organismos relacionados filogeneticamente. O secretoma é constituído por proteínas relevantes para o estilo de vida e pela capacidade de modular o sistema imune do hospedeiro (Soblik et al. 2011). No secretoma de helmintos parasitos, encontram-se proteases tais como aspartato, cisteíno, metalo e serino proteases, as quais estão envolvidas em processos de coagulação do sangue, fibrinólise, metabolismo de proteínas, reação imune e remodelamento de tecidos (Tort et al. 1999), justificando o estudo do secretoma como potencial alvo de intervenção terapêutica (Soblik et al. 2011).

A modulação do sistema imune do hospedeiro durante a infecção depende da longevidade do parasito no hospedeiro, sendo este processo dependente das proteínas e moléculas secretadas que interagem com o hospedeiro (Hewitson et al. 2009). Atualmente, existe um grande interesse da comunidade científica em compreender melhor as bases moleculares da imunomodulação feita por helmintos. A história de vida dos helmintos, as estratégias de transmissão e os nichos fisiológicos estão relacionados com as atividades imunomodulatórias observadas em três grupos taxonômicos, a saber: Nematoda, Cestoda e Trematoda (Hewitson et al. 2009).

A identificação experimental de proteínas secretadas pode ser um processo demorado e caro. Abordagens bioinformáticas baseadas na análise de genomas sequenciados podem ser usadas para priorizar a análise experimental de novos alvos terapêuticos e de imunodiagnóstico para doenças parasitárias humanas (Gomez et al. 2015) (Figura 2). Para fins de diagnóstico não invasivo, a análise do perfil do secretoma tornou-se um campo emergente na área de descoberta de biomarcadores, isso ocorre porque vários biomarcadores secretados foram identificados como relevantes para o estudo do câncer e outras doenças.

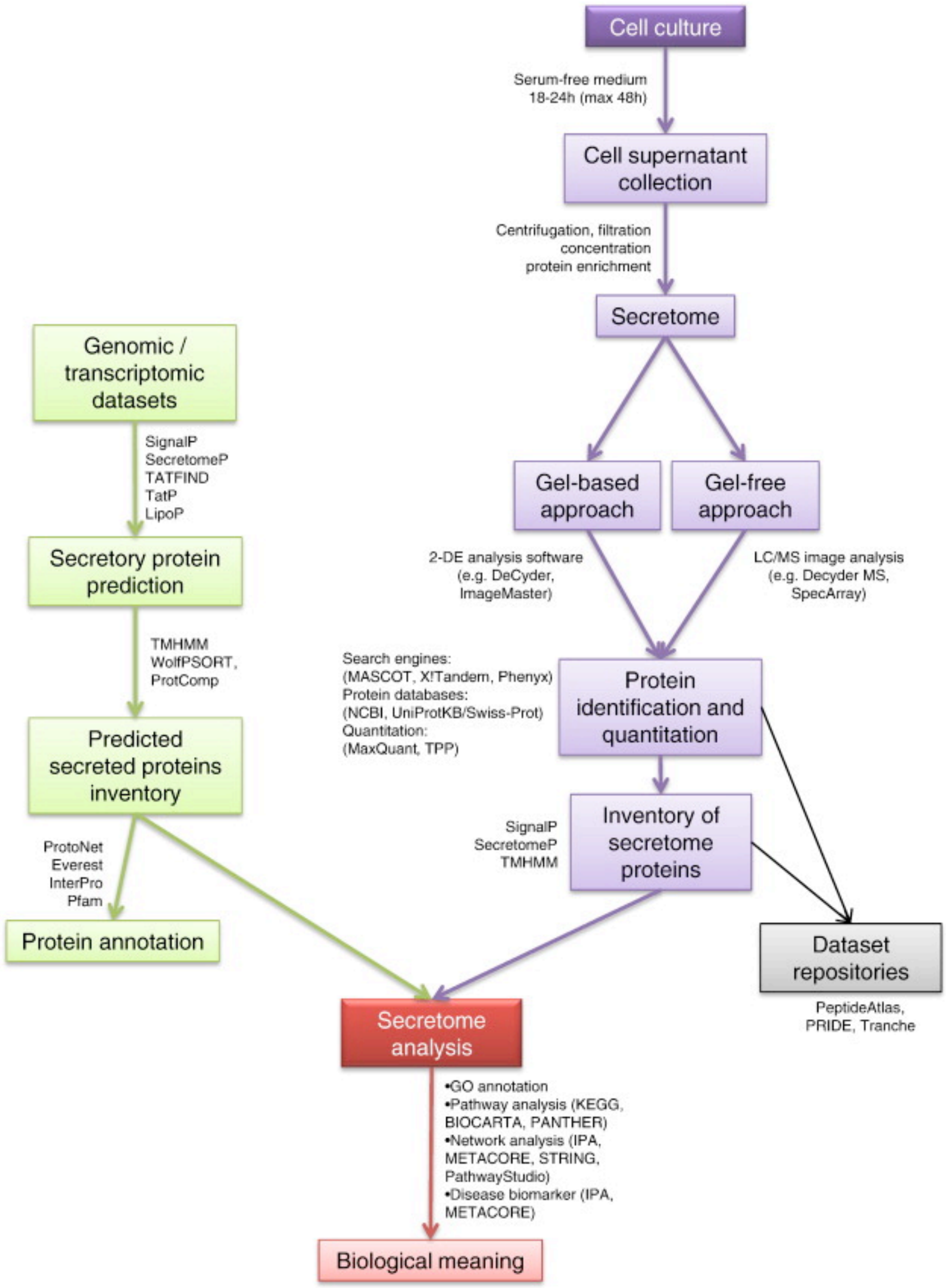

Figura 2. *Workflow* computacional (verde) e experimental (roxo) para a obtenção e análise de secretomas. Análise computacional: na ausência de dados experimentais, o perfil de proteínas secretadas pode ser obtido mediante abordagens que utilizam ferramentas bioinformáticas usando dados genômicos de transcritoma ou proteoma. Análise experimental: para este tipo de análise são necessárias diferentes etapas para a preparação da amostra dependendo da complexidade do estudo. O secretoma pode ser fracionado usando abordagens baseadas em géis (*gel-based*) ou sem utilizar géis (*gel-free*). As proteínas separadas são identificadas e quantificadas por espectrometria de massa. Ambas abordagens produzem uma lista de proteínas secretadas que devem ser posteriormente analisadas no contexto do significado biológico. Esta figura esta baseada em culturas de células mas o *workflow* pode ser aplicado a cultura de parasitos Fonte: Caccia et al. 2013.

1.3.1 - Métodos computacionais para a predição de proteínas secretadas

Na ausência de dados experimentais, o perfil do secretoma de uma célula ou organismo pode ser gerado com abordagens *in silico*. Nesse sentido, diferentes ferramentas bioinformáticas podem ser usadas para a predição de proteínas secretadas em diferentes organismos (Shah et al. 2009). Não obstante, esta metodologia pode apresentar algumas desvantagens. Em primeiro lugar, a sequência do genoma deve estar disponível para o organismo de interesse. Em segundo lugar, a precisão na predição das proteínas secretadas depende do desempenho da ferramenta utilizada e da qualidade das anotações do genoma.

Os métodos computacionais são divididos em três categorias: baseados em matrizes de pesos, alinhamento de sequências e em algoritmos de aprendizado de máquina. O uso de matrizes de peso foi o primeiro método proposto para a predição de peptídeo sinal (von Heijne 1986). Uma matriz de peso é uma medida de probabilidade de encontrar cada resíduo em cada posição na sequência sinal, indicando a similaridade com a coleção de sequências usadas para gerar a matriz. Se a pontuação esta acima de um valor mínimo, isso indica a localização da sequência sinal (Caccia et al. 2013).

Abordagens baseadas em alinhamento de sequências, tais como buscas usando o pacote BLAST, não conseguem identificar peptídeos sinal devido à grande variabilidade de comprimento, baixa similaridade na sequência e os parâmetros do algoritmo que não são otimizados para o alinhamento de sequências curtas (Altschul et al. 1990). Um dos algoritmos desenvolvidos usando esta metodologia é o Signal-BLAST (Frank and Sippl 2008). Este programa compara uma sequência *query* contra dois conjuntos de sequências referência: peptídeo sinal e não peptídeo sinal. Se o melhor alinhamento com a sequência *query* é encontrado no conjunto de sequências com peptídeo sinal, o resultado é secretada e vice-versa.

Ferramentas mais sofisticadas têm sido desenvolvidas com base na abordagem de aprendizado de máquina. Estes métodos "aprendem" a discriminar entre proteínas secretadas e não secretadas numa fase de treino, durante a qual os peptídeos típicos sinal e não sinal são apresentados ao algoritmo e este tenta reproduzir a classificação por meio do ajuste dos seus parâmetros específicos. No final da fase de treinamento, um modelo de classificação é construído e utilizado para categorizar novas sequências de proteínas (Lai et al. 2012). A ferramenta mais usada com estas características é o SignalP que está baseado em ANN (*artificial neural network*) e/ou HMM (*hidden Markov model*) dependendo da versão (Nielsen et al. 1997; Nielsen and Krogh 1998; Bendtsen et al. 2004).

Um estudo comparativo testou 13 ferramentas das categorias mencionadas anteriormente e avaliaram a precisão, especificidade e sensibilidade na discriminação entre peptídeos sinal e não sinal e na identificação da posição de clivagem (Choo et al. 2009). Os resultados mostraram que a maior parte das ferramentas foram mais precisas para eucariotos do que para bactérias. Isto provavelmente foi devido ao fato de existirem mais dados de eucariotos disponíveis para construir os modelos matemáticos. Finalmente, o estudo mostrou que as ferramentas baseadas em abordagens de aprendizado de máquina parecem superar os outros métodos, sendo o SignalP o que têm o melhor desempenho (Caccia et al. 2013).

A predição de proteínas secretadas se torna mais difícil pela elevada semelhança das sequências de peptídeos sinal e as âncoras sinal (e.g. domínios transmembrana), que têm uma composição de aminoácidos semelhantes. Procurar por peptídeos sinal numa escala genômica pode ter muitos falsos positivos devido às regiões transmembrana, mas o SignalP (versões 4.0 e 4.1) tenta superar este desafio através da combinação da identificação do peptídeo sinal com predição da topologia transmembrana (Petersen et al. 2011). Além disso, métodos específicos para a predição de hélices transmembrana têm sido desenvolvidos, incluindo TMHMM, um método baseado em HMM (*hidden Markov model*) que faz uso de matrizes de peso que foram extraídas a partir da análise estatística de TMbase, uma coleção de todas as proteínas transmembrana presentes e anotadas no SwissProt.

As ferramentas referidas acima podem predizer a presença de peptídeo sinal que representa a via clássica de secreção. A via não clássica é predita com ferramentas específicas como

SecretomeP. Nesse caso, os programas estão baseados na ideia de que as proteínas extracelulares compartilham características específicas no nível da sequência independentemente do mecanismo pelo qual elas são secretadas (Bendtsen et al. 2004; Bendtsen et al. 2005).

1.3.2 – Estudos evolutivos nas famílias de proteínas relevantes na interação parasito-hospedeiro: cistatinas, um caso de estudo.

As cistatinas são uma família de inibidores de cisteíno proteases que fazem parte do secretoma e estão presentes em um amplo número de grupos taxonômicos, incluindo Nematoda e Platyhelminthes. A família das cistatinas se divide em três subfamílias. Estefinas (I25A): proteínas intracelulares de 11 kDa que não apresentam pontes dissulfeto. Cistatinas (I25B): proteínas secretadas de 14 kDa com pelo menos uma ponte dissulfeto. Estão envolvidas na regulação da atividade das cisteíno proteases em parasitos e na modulação da resposta imune do hospedeiro (Gregory and Maizels 2008; Khaznadji et al. 2005). Quininogênio (I25C): glicoproteína intracelular com uma massa relativa entre 60 a 120 kDa (Abrahamson 1994). Na literatura também é possível encontrar que esta família de proteínas está dividida em tipo I, II e III, respectivamente. A primeira cistatina descrita em parasitos foi a onchocistatina da subfamília I25B do nematódeo *Onchocerca volvulus* (Lustigman et al. 1992). Inicialmente pensava-se que esta proteína regulava as proteases do parasito durante a muda dos nematódeos. Não obstante, funções adicionais, além daquelas associadas ao processo de muda, são evidenciadas pelo fato de que a cistatina da filária de roedores *Acanthocheilonema vitae* é secretada pelo verme macho e pela microfilária no estágio em sangue (Hartmann et al. 1997).

Cistatinas (subfamília I25B) de filárias possuem um motivo adicional (SND) necessário para inibir diferentes classes de cisteíno proteases, como as legumaínas. A cistatina de *Brugia malayi* tem a capacidade de inibir uma protease tipo legumaína (Manoury et al. 2001). As cistatinas podem ter uma dupla função: inibir as cisteíno proteases de helmintos, assim como as proteases dos hospedeiros (Manoury et al. 2001).

As diferenças entre as cistatinas (I25B) de filárias e *C. elegans* foram observadas testando sua capacidade proliferativa de células T. Enquanto as cistatinas das filárias *O. volvulus* e *A. viteae* interferiram com a proliferação de células T humanas ou murinas, as cistatinas de *C. elegans* não tiveram efeito inibitório (Schierack et al. 2003). Cistatinas (I25B) de nematódeos parasitos e de vida livre diferem em relação a suas propriedades imunomodulatórias (Schierack et al. 2003). Embora as cistatinas apresentem uma baixa similaridade de sequência com as cistatinas dos hospedeiros, suas sequências são idênticas ou conservadas na maioria dos aminoácidos críticos.

Recentemente, o DRG4 (*Disease Reference Group on Helminth Infections*) definiu dez áreas de pesquisa para o desenvolvimento de futuras medidas de controle das helmintíases (http://apps.who.int/iris/bitstream/10665/75922/1/WHO_TRS_972_eng.pdf). Destas, as seguintes áreas estão relacionadas ao nosso projeto: 1) anotação de genomas e transcritomas e o desenvolvimento de novas ferramentas de genômica funcional; 2) melhoramento dos testes de diagnóstico; 3) interações parasito-hospedeiro e como helmintos modulam esta interação.

O presente projeto está centrado no estudo do secretoma em helmintos. Neste estudo, serão preditos os secretomas de diferentes espécies de helmintos, comparando-se as espécies de vida livre, parasitária e diversos hospedeiros, visando a compreender a diversidade e evolução dos secretomas. Para explorar este tema, o presente estudo aborda as seguintes perguntas científicas: 1) A composição do secretoma em termos de proteínas e domínios proteicos é conservada entre diferentes espécies de helmintos? 2) Esta composição do secretoma está relacionada ao estilo de vida de helmintos?

II - OBJETIVOS

2.1 – Objetivo Geral

O objetivo geral deste projeto é analisar a diversidade e evolução do secretoma de helmintos de vida livre e parasitária através da predição *in silico* das proteínas secretadas que potencialmente interagem com seus respectivos hospedeiros.

2.2 – Objetivos Específicos

- Levantamento e filtragem de sequências genômicas de helmintos.
- Análise e identificação do secretoma em 44 espécies de helmintos.
- Comparação de famílias de proteínas que compõem o secretoma de helmintos de diferentes estilos de vida e hospedeiros.
- Identificação de homólogos de cistatinas no proteoma predito de três espécies de *Schistosoma* e outros Platyhelminthes.
- Análise das relações evolutivas da famílias de cistatinas em três espécies de *Schistosoma.*

III – RESULTADOS

3.1 – CAPÍTULO I: Secretoma de Helmintos

Cuesta-Astroz Y, Silva de Oliveira F, Nahum LA, Oliveira G. Helminth secretomes reflect different lifestyles and parasitized hosts. (manuscrito submetido à International Journal for Parasitology).

Proteínas secretadas de parasitos são capazes de modificar o ambiente do hospedeiro e modular seu sistema imunológico. Pouco se sabe sobre a composição e variabilidade das proteínas secretadas em diferentes espécies de helmintos e a sua contribuição na progressão da infecção. O objetivo deste trabalho foi compreender como a diversidade do secretoma é moldada de acordo com diferentes estilos de vida de helmintos e assim identificar características específicas que permitem a sobrevivência dos mesmos em diferentes ambientes. Neste estudo foram preditos os secretomas de 44 espécies de helmintos: 31 nematódeos e 13 platelmintos. Os proteomas preditos foram recuperados do banco de dados WormBase versão 1.0 (http://parasite.wormbase.org). As proteínas secretadas em 44 espécies foram 41,200 proteínas, sendo 31,192 com peptídeo sinal (via clássica de secreção) e 10,008 pela via não clássica. Nossos resultados indicaram proteínas secretadas associadas com infecção, aderência, invasão e processos de regulação imunológica como inibidores de proteases, citocinas, entre outros. Foram identificadas também proteínas específicas relacionadas com os estilos de vida dos helmintos. Proteínas compartilhadas nos secretomas podem revelar mecanismos que parecem ser conservados entre helmintos de plantas, animais e vida livre. Este estudo contribuirá para a compreensão da interação parasito-hospedeiro e possivelmente para a identificação de novos alvos moleculares para o tratamento ou diagnóstico de helmintíases.

Helminth secretomes reflect different lifestyles and parasitized hosts

Yesid Cuesta-Astroz[1,3], Francislon Silva de Oliveira[1,3], Laila Alves Nahum[1,4], Guilherme Oliveira[1,2*]

[1] Centro de Pesquisas René Rachou (CPqRR), Fundação Oswaldo Cruz (FIOCRUZ), Belo Horizonte, MG 30190-002, Brazil

[2] Vale Institute of Technology, Belém, PA 66055-090, Brazil

[3] Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG 31270-901, Brazil

[4] Faculdade Promove de Tecnologia, Belo Horizonte, MG 30130-180, Brazil.

* Corresponding author: Guilherme Oliveira guilherme.oliveira@itv.org

E-mail addresses:

YCA: yesid.cuesta@gmail.com

FSO: francislon@gmail.com

LAN: laila@cpqrr.fiocruz.br

GO: guilherme.oliveira@itv.org

Abstract

Helminths cause a number of medical and agricultural problems and are a major cause of parasitic infections in humans, animals, and plants. Comparative analysis of helminth genes and genomes are important to understand the genomic biodiversity and evolution of parasites and their hosts in terms of different selective pressures in their habitats. The interactions between the infective organisms and their hosts are mediated in large part by secreted proteins, known collectively as the "secretome". Proteins secreted by parasites are able to modify the host's environment and modulate their immune system. The composition and function of this set of proteins varies depending on the ecology, lifestyle, and environment of an organism. The present study aimed at predicting *in silico* the secretome in 44 helminth species including Nematoda (31 species) and Platyhelminthes (13 species) and at understanding the diversity and evolution of secretomes. According to our observations, the secretome and proteome sizes are not related. Secretomes from plant's helminths range from 7.6% to 13.9% of the total proteome with an average of 10.2% and from free-living helminths range from 4.4% to 13% with an average of 9.8% respectively and thus are considerably larger secretomes, in relation to animal helminths secretomes which range from 4.2% to 11.8% of the proteomes, with an average of 7.1%. Across 44 secretomes in different helminth species, we found five conserved domains: 1) PF00014 (Kunitz/Bovine pancreatic trypsin inhibitor domain), 2) PF00046 (Homeobox domain), 3) PF00188 (cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins), 4) PF00085 (Thioredoxin) and 5) PF07679 (Immunoglobulin I-set domain). Our results detected secreted proteins associated with invasion, infection, adhesion, and immunoregulation processes as protease inhibitors, cytokines, among others. In summary, this study will contribute towards the understanding of the host-parasite interaction and possibly identify new molecular targets for the treatment or diagnosis of helmintheases.

Keywords: Secretome, protein evolution, biodiversity, host-parasite interaction, computational biology, helminths.

1. Introduction

Helminths (Nematoda and Platyhelminthes) have great medical, veterinary and agricultural relevance and cause deep socio-economic impacts. According to the World Health Organization (WHO), more than two billion people are infected and many more are at risk of contracting helminthiases, especially in developing countries. Diseases caused by helminths are extremely varied. They include anemia and malnutrition (caused by hookworms such *Ancylostoma ceylanicum*), river blindness (filarial nematode *Onchocerca volvulus*), lymphatic filariasis (filarial nematodes *Brugia malayi, Loa loa*, and *Wuchereria bancrofti*), and impaired cognitive development (Hotez et al. 2008; Lustigman et al. 2012). In some cases, helminths can maintain a chronic infection in the host. Recorded cases include patients with more than 30 years of *Schistosoma mansoni* infection (A. R. C. Harris et al. 1984) and a record of 53 years of *E. granulosus* infection (Spruance 1974).

Helminths are also responsible for considerable losses in agriculture. In plants, helminths cause damage to crops leading to huge economic losses. Some of the most damaging nematodes include plant endoparasites (root-knot nematodes) belonging to the genera *Meloidogyne spp., Heterodera spp.*, and *Globodera spp.* (Mehta et al. 2008). For example, *Meloidogyne spp* impacts both the quantity and quality of harvest, causing an estimated US$80bn in damage annually (Wasmuth et al. 2008). Helminth parasites of livestock, such as cattle and sheep, are the cause for severe economic losses worldwide, billions of dollars are spent annually on treatment and control of nematodes (Wang et al. 2009).

The rationale in the search for anthelmintic targets before the use of genomics was based on the molecular study of individual genes associated with helminth virulence (Brindley et al. 2009). The genome of different free living and parasitic helminth species, which affect humans and other organisms with agricultural and veterinary importance, are being sequenced. Genome, transcriptome, and proteome data from helminths have provided important information for understanding molecular mechanisms involved in metabolism, parasite-host interaction, immune system evasion and molecular evolution, among other approaches (Brindley et al. 2009; Cuesta-Astroz et al. 2014; Nahum, Mourão, and Oliveira 2012; Lustigman et al. 2012; Tsai et al. 2013). Therefore, approaches that consider species biology in a global manner have been increasing deployed (Oliveira and Pierce 2015).

Secreted proteins play essential roles from bacteria to mammals (Tjalsma et al. 2000). Such proteins may represent between eight to 20% of the organism's proteome (Greenbaum et al. 2001). Proteins included in the secretome belong to various functional classes such as cytokines, hormones, digestive enzymes, proteases, toxins, antimicrobial peptides, and proteins associated with oxidative stress. Some of them are involved in vital biological processes such as cell adhesion, cell migration, cell-cell communication, differentiation, proliferation, morphogenesis, and regulation of immune response (Maizels and Yazdanbakhsh 2003). Parasite secreted proteins not only play a role in the organisms that produces it but also have been demonstrated to regulate the host immune response and to be the direct cause of pathology (Maizels and Yazdanbakhsh 2003; Cass et al. 2007; Ferguson et al. 2015; Zhu et al. 2016).

The prediction of secreted proteins depends on the computational identification of specific signals. Protein secretion through the endoplasmic reticulum is associated with a hydrophobic signal peptide at the N-terminus portion, representing the classical secretory pathway. Some secreted proteins do not contain a signal peptide and participate in the non-classical secretion pathways by extracellular vesicles. Secreted proteins by the non-classical pathways have been identified as glycolytic enzymes, chaperones and translation factors, among others, suggesting that these proteins may be multifunctional ("moonlighting" proteins) (Nombela, Gil, and Chaffin 2006). Microscopy studies have shown experimental evidence of extracellular vesicles in helminths, specifically in trematodes *Fasciola hepatica, Echinostoma caproni* (Marcilla et al. 2012), *Schistosoma japonicum* (Zhu et al. 2016) and *Schistosoma mansoni* (Sotillo et al. 2016). These vesicles are actively released by the parasites and are captured by the host cells playing an important role in host-parasite interaction (Zhu et al. 2016).

Little is known about the diversity and evolution of secreted proteins in helminths. Processes shaping secretome diversity according to different niches and lifestyles and specific features allowing parasite survival in different environments remain open to investigation. The present study aimed at performing a comparative analysis of the predicted secretome across 44 species including free-living and parasitic Nematoda and Platyhelminthes in order address the following issues: the conservation of the secretome among different helminth species; the diversity of the secreted repertoire on different secretomes; the correlation between the size and composition of the secretome with the species life style, host and phylogenetic lineage; and the presence of different protein domain features between secreted and non-secreted proteins.

2. Materials and methods

2.1. Organisms and sequence data

Predicted proteomes were retrieved from WormBase (http://parasite.wormbase.org) (T. W. Harris et al. 2014). The original dataset had 92 helminth proteomes distributed in 82 species. The genomes were filtered according to the scheme mentioned below. The final genome dataset was composed of 44 species (31 Nematoda and 13 Platyhelminthes). *C. elegans* data were retrieved from WormBase (https://www.wormbase.org). Data from *O. viverrini* and *G. salaris* were retrieved from the original genome papers (Young et al. 2014; Hahn, Fromm, and Bachmann 2014). Nematoda species were divided according to DNA sequence studies that suggested the existence of five clades (Blaxter et al. 1998). Some closely related species, except for *Schistosoma,* were not included to minimize data duplication (for example, species belonging to the same genus).

2.2. Data filtering

Sequences were scanned using a script to remove possible error sources and to validate them according to the following criteria: 1) starting with a methionine, 2) having no internal stop codons, 3) lacking ambiguous amino acids not represented in the 20 IUPAC amino acid codes, and 4) longer than 100 amino acids. Proteomes with > 65% of sequences retrieved after the filtering process were included. In the case that more than one project was available for the same species, only the project with the best values in the filtering process was included in our analyses.

2.3. Secretome prediction

The *in silico* prediction of secreted proteins was performed using different bioinformatics tools and databases (Figure 1). SignalP 4.1 (Petersen et al. 2011) was used for identifying classical secretory proteins. All proteins identified as not having a signal peptide were analyzed with SecretomeP (Bendtsen et al. 2004) for predicting non-classical secreted proteins. To limit false positive results, only records with NN (neural network) score ≥ 0,9 were considered as secreted proteins. Proteins predicted to be secreted were subsequently scanned for the presence of mitochondrial sequences by TargetP (Emanuelsson et al. 2000) and transmembrane helices by TMHMM (Krogh et al. 2001).

This approach has been used for the prediction of soluble secreted proteins in helminths and arthropods described by other authors (Garg and Ranganathan 2011; Schicht et al. 2013). For consistency, we predicted the secretome datasets using the same methodology. All programs used in this study were linked using Perl and bash shell scripts. A MySQL database was created to store and retrieve information using queries.

2.4. Functional annotation

Putative secreted proteins were mapped to Gene Ontology (GO) terms and annotated using Blast2GO (Conesa et al. 2005) using default parameters (E-Value-Hit-Filter: 1.0E-6; Annotation cut-off: 55; GO weight: 5; Hsp-Hit Coverage cut-off: 0). Additionally, secreted proteins were associated to protein families, domains, and functional sites through InterProScan v.5.0.7 (Jones et al. 2014) in a standalone version. InterProScan combines different protein signature recognition methods into one resource integrating the following databases: Coils, FPrintScan, Gene3D, HMM-Panther, HMM-PIR, HMM-Pfam, HMM-Smart, HMM-Tigr, Phobius, ProfileScan, Prosite, PatternScan, and Superfamily. For the protein domain architecture analysis (presence and order in which domains are arranged within the protein sequence), we used the Pfam database (Finn et al. 2016). Pfam is a collection of manually curated families known as Pfam-A and a set of automatically generated families named Pfam-B. Pfam-A domains were considered in the present study. Additional functional information such as Gene Ontology terms for Pfam domains were retrieved from the InterProScan results. These manual annotations are based on the function of particular domains rather than the function of domain families. These results were deposited in the local MySQL database in order to perform specific queries and retrieve information.

2.5. Orthologs predictions

To detect putative orthologs across predicted secretomes, we performed an OrthoMCL cluster analysis (Fischer et al. 2011) using the default settings (E-value cutoff: 1e-5 and identity: 50%). In this phase, the OrthoMCL maps proteins to groups in OrthoMCL-DB. It performs a BLASTP search against all the proteins in OrthoMCL-DB using a cutoff of 1e-5 and 50% match. Each protein is assigned to the group containing its best hit. If the best matching protein does not have a group, it is assigned to NO_GROUP.

3. Results

We have analyzed 44 predicted proteomes in order to identify putative secreted proteins. To this end, we used a pipeline to identify secreted proteins by classical and non-classical secretion pathways (Figure 1). This pipeline is composed by programs based on protein sequences followed by functional annotations (Gene ontology, protein domains, domain architecture diversity, and cluster of orthologs groups) in order to identify specific signatures in the secretomes according to lifestyle and host. Such analyses generated a list of domains and GOs categories identifying specific and shared features, the annotated GO terms could suggest adaptations to specific niches based on molecular functions and biological process. For the complete list of species and the filtering results, see Table 1.

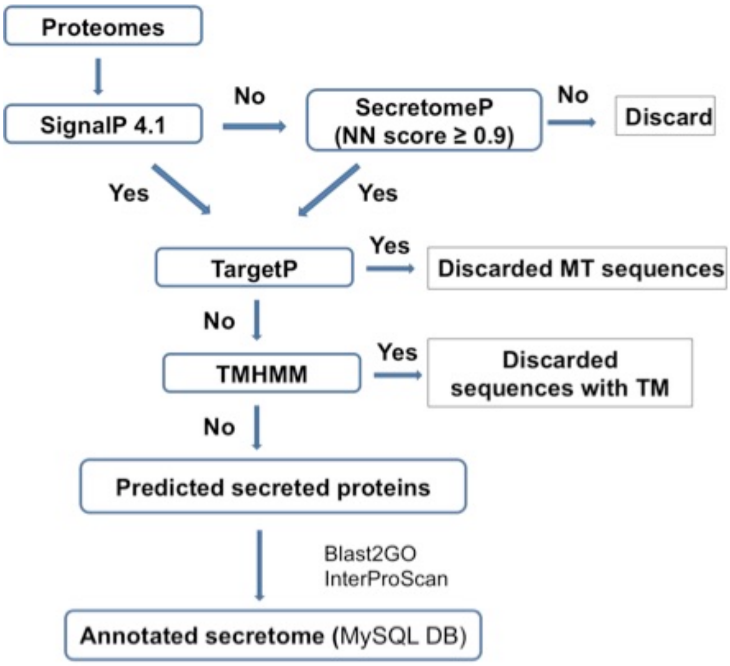
Proteomes
SignalP 4.1
No
SecretomeP (NN score ≥ 0.9)
No
Discard
Yes
Yes
TargetP
Yes
Discarded MT sequences
No
TMHMM
Yes
Discarded sequences with TM
No
Predicted secreted proteins
Blast2GO
InterProScan
Annotated secretome (MySQL DB)

Figure 1. Workflow for secretome prediction and annotation. MT: mitochondrial, TM: transmembrane. See references for accessing databases and tools.

Table 1. Prediction of secreted proteins in 44 helminth species. Nematoda (top) and Platyhelminthes (bottom). Total sequences: number of sequences in the predicted proteome of each species analysed. Valid sequences: sequences after filtering. AP: animal parasite, PP: plant parasite, FL: free living.

Species	Total sequences	Valid sequences (%)	Total Secretome	Lifestyle
Nematoda				
Clade I				
Trichinella spiralis	16,380	13,617 (83.1)	1,146	AP
Trichuris suis	9,831	8,489 (86.3)	727	AP
Clade III				
Acanthocheilonema viteae	10,397	8,337(80.1)	507	AP
Ascaris suum	18,542	14,893 (80.3)	1,158	AP
Brugia malayi	17,750	12,116 (68.2)	800	AP
Dirofilaria immitis	12,857	9,163(71.2)	593	AP
Dracunculus medinensis	9,495	6,559(69.1)	499	AP
Elaeophora elaphi	9,562	7,636(79.8)	525	AP
Enterobius vermicularis	12,063	8,842(73.3)	649	AP
Litomosoides sigmodontis	10,246	8,591(83.8)	542	AP
Loa loa	15,445	12,204 (79.0)	803	AP
Onchocerca volvulus	12,534	9,314 (74.3)	682	AP
Syphacia muris	10,200	7,768(76.1)	584	AP
Thelazia callipaeda	9,999	7,729(77.3)	532	AP
Toxocara canis	16,571	10,668(65)	858	AP
Wuchereria bancrofti	12,625	8,204(65)	517	AP
Clade IV				
Bursaphelenchus xylophilus	17,704	14,948 (84.4)	2,077	PP
Globodera pallida	16,403	13,158 (80.2)	1,218	PP
Meloidogyne hapla	14,420	12,391 (86.0)	943	PP
Parastrongyloides trichosuri	14,957	12,913(86.3)	1,334	AP
Rhabditophanes sp. KR3021	13,493	12,168(90.1)	1,192	AP
Strongyloides ratti	12,430	11,489 (92.4)	973	AP
Clade V				
Ancylostoma ceylanicum	15,892	13,586(85.4)	1,169	AP
Angiostrongylus costaricensis	9,989	6,956(69.6)	509	AP
Dictyocaulus viviparus	13,514	11,510(85.1)	825	AP
Haemonchus contortus	24,747	20,460(82.6)	2,419	AP
Necator americanus	19,153	13,924 (72.7)	1,231	AP
Nippostrongylus brasiliensis	20,234	14,214 (70.2)	1,469	AP

Oesophagostomum dentatum	25,291	16,790 (66.3)	1,531	AP
Pristionchus pacificus	24,217	20,049 (82.7)	2,388	FL
Caenorhabditis elegans	26,018	24,002 (92.2)	3,121	FL
Platyhelminthes				
Cestoda				
Echinococcus multilocularis	10,189	9,145 (89.7)	524	AP
Hydatigera taeniaeformis	10,907	7,614 (69.8)	468	AP
Hymenolepis microstoma	10,077	9,066 (89.9)	454	AP
Mesocestoides corti	9,056	6,520 (72)	406	AP
Taenia solium	12,481	10,406 (83.3)	538	AP
Trematoda				
Clonorchis sinensis	13,634	11,947 (87.6)	656	AP
Fasciola hepática	15,739	12,488 (79.3)	992	AP
Schistosoma haematobium	13,073	8,769 (67.0)	379	AP
Schistosoma japonicum	12,743	9,141 (71.7)	476	AP
Schistosoma mansoni	11,828	10,121 (85.5)	431	AP
Opisthorchis viverrini	16,379	12,117 (73.9)	832	AP
Monogenea				
Gyrodactylus salaris	15,488	11,148 (71.9)	653	AP
Turbellaria				
Schmidtea mediterranea	29,850	19,423 (65.0)	870	FL

3.1. Secretome size

Proteins containing a signal peptide, significant SecretomeP score (NN >= 0.9) and lacking a mitochondrial origin signal and transmembrane domains were considered as belonging to the soluble secretome (41,200 proteins, 31,192 with signal peptide and 10,008 with significant SecretomeP score). Sequences belonging to the secretomes in fasta format per species are available in https://figshare.com/s/6410c0479de5e1a6ece1. The secretome constituted on average 7.6% of the proteome in all the 44 species (8.4% in Nematoda and 5.5% in Platyhelminthes) with extreme values of 13.9% for *Bursaphelenchus xylophilus* and 4.2% for *Schistosoma mansoni* (Figure 2).

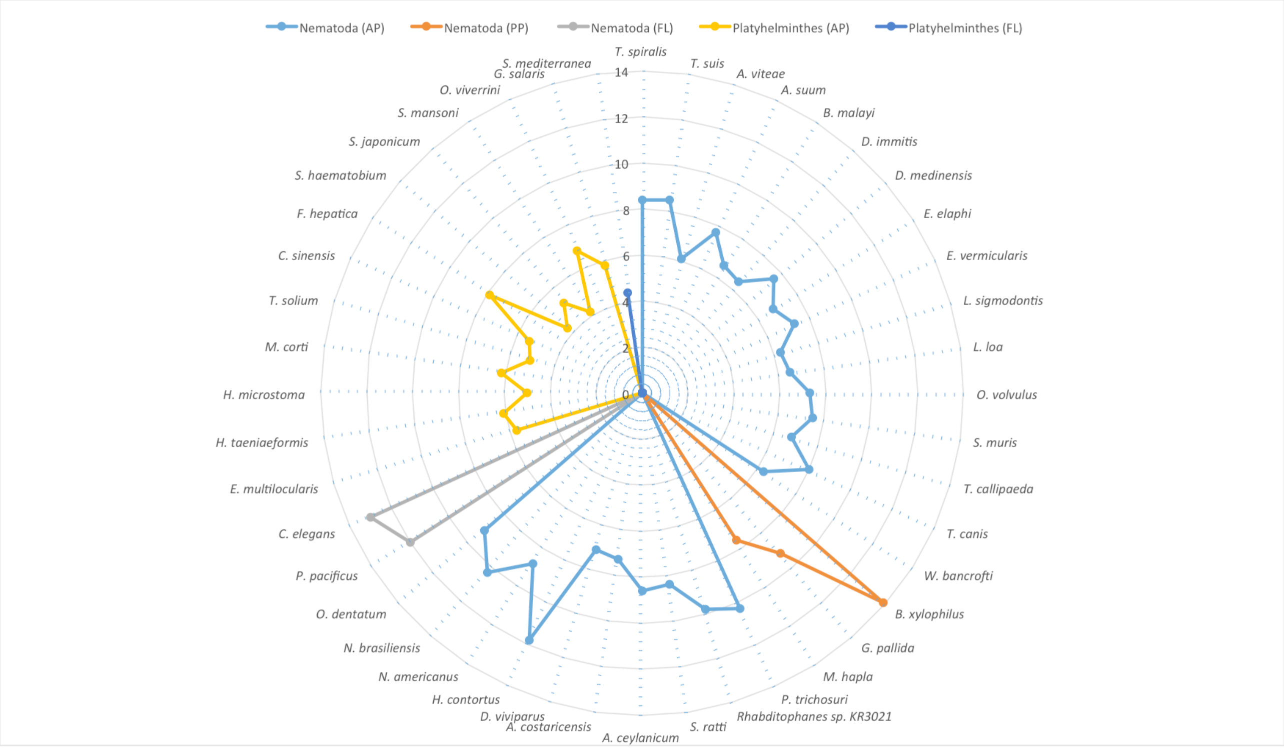

Figure 2. Secretome size (%) distribution across 44 helminth species. Points represent the secretome size in percentage in relation with the whole proteome. AP: animal parasite, PP: plant parasite, FL: free living.

When we analyzed these secretomes in a phylogenetic context, secretomes in Clade I Nematoda had similar sizes, on average 8.4% of the total proteome. Secretomes in Clade III enclosed in average 6.9%, Clade IV 9.8%, and Clade V 9.7% of the proteome. Secretomes from Cestoda contained in average 5.6%, Trematoda to 5.6%, Monogenea to 5.8%, and Turbellaria to 4.4% of the proteome. In Platyhelminthes, the average size was noticeably smaller compared with Nematoda and also between animal and plant infecting helminths. Across Platyhelminthes the relative sizes of the secretomes were similar (Figure 2).

According to the classification used by (Krijger et al. 2014), the 44 species were grouped into three classes. Class 1 contained seven species with secretomes comprising less than 500 proteins (six of them are Platyhelminthes). Class 2 comprised 24 species with secretomes ranging from 500 to 1,100 proteins. Class 3 included the remaining 13 species with more than 1,100 proteins (this secretome size contained exclusively nematodes). According to our observations, the secretome size does not depend on the increase of proteomes (Figure 2) (Table 1). In relation to the hosts, animal helminths secretomes range from 4.2% to 11.8% of the proteomes, with an average of 7.1%. Secretomes from plant helminths range from 7.6% to 13.9% with an average of 10.2% and thus are considerably larger. The secretome size of free-living helminths range from 4.4% to 13% with an average of 9.8% being also considered a larger secretome compared with animal helminths secretomes.

For each secretome, the protein length distribution was analyzed using the number of proteins in defined length intervals (100-300, 301-500, 501-1000 and ≥ 1001 aa) as fractions of the total secretome. The complete secretome of 44 species (41,200 sequences) presented length distribution as shown in supplementary file 1. Figure 3 presents the length distribution for each species. The majority of proteins in the secretomes had less than 300 amino acids and proteins with more than 1,000 amino acids were poorly represented across the secretomes. However species with greater presence of these large proteins were found in the free-living helminths, *C. elegans* and *P. pacificus* (Figure 3). Clade III species (*D. medinensis* and *W. bancrofti*) did not have this type of proteins. In terms of the number of sequences that are part of the secretome, *C. elegans* has the larger secretome containing 3,121 proteins and the smallest was the *S. haematobium* secretome (Figure 3).

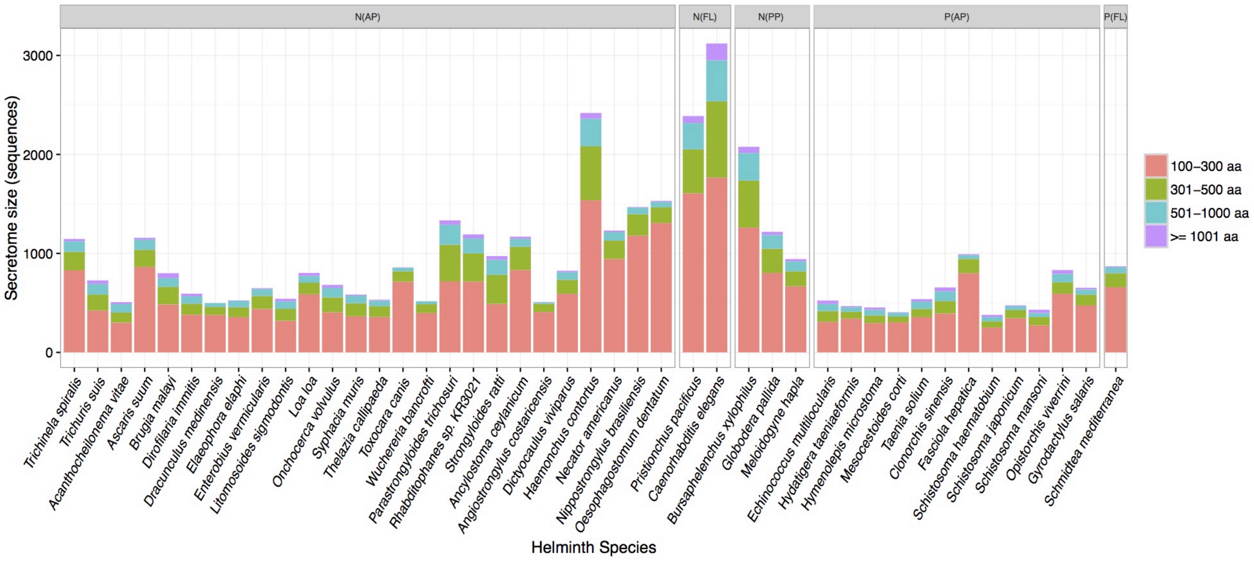

Figure 3. Secretome size distribution (sequences) and sequence size (number of amino acids – aa) across 44 helminth species. N: nematoda, P: platyhelminthes AP: animal parasite, PP: plant parasite, FL: free living.

3.2. Protein domain diversity analysis

We assigned domain information for each protein using InterProScan. We then parsed the Pfam domains results using MySQL queries. Out of 41,200 proteins (across 44 species), 20,607 (50%) proteins had at least one Pfam domain assigned. A total of 2,345 domains were identified across the 44 species and the occurrences per species were counted (supplementary file 2).

In order to have a global view of domain diversity and distribution, we calculated the most represented (top 25) domains across the secretomes (Table 2) and the domain occurrences per species (supplementary file 3). The Shk domain (PF01549) is a potassium channel inhibitor and was the most represented in the secretomes, 705 times (Table 2). The Shk domain is one of the most recurrent domains across protein architectures. Other well represented domains in secreted proteins were peptidase domains involved in hemoglobin degradation and uptake nutrients, redox process, and cell to cell communication (Table 2).

In general, antioxidant molecules, proteases, and cell to cell communication domains were predicted secretomes across helminth species (supplementary file 3). These included peroxiredoxin, thioredoxin, protein-disulfide isomerase, trypsins, lectins, cadherins, laminins, among others. Proteases such as metalloproteases degrade extracellular matrix proteins and may be involved in the degradation of plant and animal tissues. Other domains, however, could mediate the suppression of plant defense by degradation of host proteins involved in pathogen recognition or playing other important roles in defense (Krijger et al. 2014).

Table 2. Top 25 most represented domains found in secreted proteins across 44 helminth species. GO: Gene Ontology.

Pfam ID	Pfam name	Description	Number of proteins	Domain ocorrences	GO terms
PF01549	ShK	ShK is a powerful inhibitor of T lymphocyte voltage-gated potassium channels, in particular Kv1.3. Structural analogues may have use as an immunosuppressants for the treatment of autoimmune diseases	705	1678	-
PF01060	Transthyretn-like family	Apparently nematode-specific protein family.	665	691	GO:0005615 extracellular space
PF01400	Astacin (Peptidase family M12A)	Family of metallopeptidases	527	560	GO:0004222 metalloendopeptidase activity. GO:0006508 proteolysis
PF00188	CAP	CAP protein family (cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins (CAP)) are found in a wide range of organisms, including prokaryotes and non-vertebrate eukaryotes.	518	576	-
PF00059	Lectin_C	A C-type lectin (CLEC) is a type of carbohydrate-binding protein domain.	472	574	GO:0030246 carbohydrate binding
PF00112	Peptidase_C1	Cysteine proteases, are enzymes that degrade proteins.	389	420	GO:0006508 proteolysis GO:0008234 cysteine-type peptidase activity
PF00014	Kunitz_BPTI	Kunitz domains are the active domains of proteins that inhibit the function of protein degrading enzymes (protease inhibitors).	386	1280	GO:0004867 serine-type endopeptidase inhibitor activity

PF00046	Homeobox	Is a protein structural domain that binds DNA or RNA and is thus commonly found in transcription factors.	329	329	GO:0003677 DNA binding
PF00026	Asp	Aspartic proteases are a family of protease enzymes that use an aspartate residue for catalysis of their peptide substrates.	301	316	GO:0004190 aspartic-type endopeptidase GO:0006508 proteolysis
PF02520	DUF148	A domain of unknown function (DUF) is a protein domain that has no characterised function	281	288	-
PF04155	Ground-like	It has been proposed that the domain containing proteins may bind and modulate the activity of Patched-like membrane molecules, reminiscent of the modulating activities of neuropeptides	263	273	-
PF00085	Thioredoxin	Is a class of small redox, it plays a role in many important biological processes, including redox signaling.	242	410	GO:0045454 cell redox homeostasis
PF00135	COesterase	Carboxyl-esterases have been classified into three categories (A, B and C) on the basis of differential patterns of inhibition by organophosphates.	232	261	-
PF01682	DB	This domain has no known function	223	243	-
PF07679	I-set	Are found in several cell adhesion molecules, including vascular (VCAM), intercellular (ICAM), neural (NCAM) and mucosal addressin (MADCAM) cell adhesion molecules, as well as junction adhesion molecules (JAM).	223	973	-
PF00089	Trypsin	Trypsin (EC 3.4.21.4) is a serine protease from the PA clan superfamily	220	258	GO:0004252 serine-type endopeptidase GO:0006508 proteolysis
PF13499	EF-hand_7	helix-loop-helix structural domain or motif found in a family of calcium-binding proteins.	183	257	GO:0005509 calcium ion binding
PF00431	CUB	Is a structural motif of approximately 110 residues found almost exclusively in extracellular and plasma membrane-associated proteins.	180	317	-
PF08246	Inhibitor_I29	Cathepsin propeptide inhibitor domain (I29). protease inhibitors are molecules that inhibit the function of proteases.	173	175	-
PF01764	Lipase_3	Triglyceride lipases are lipases that hydrolyse ester linkages of triglycerides.	167	175	GO:0006629 lipid metabolic process
PF00092	VWA	The von Willebrand factor is a large multimeric glycoprotein found in blood plasma.	162	220	-
PF00328	His_Phos_2	A phosphatase is an enzyme that removes a	158	168	GO:0003993 acid

		phosphate group from its substrate.			phosphatase activity
PF00024	PAN_1	The domain is found in diverse proteins, in some they mediate protein-protein interactions, in others they mediate protein-carbohydrate interactions.	155	266	-
PF00069	Pkinase	The protein kinase domain is a structurally conserved protein domain containing the catalytic function of protein kinases	155	171	GO:0004672 protein kinase activity GO:0005524 ATP binding GO:0006468 protein phosphorylation
PF00090	TSP_1	Thrombospondins (TSP) are secreted proteins with antiangiogenic abilities. Inhibiting the proliferation and migration of endothelial cells	148	447	-

3.2.1. Common Pfam domains

Across 44 secretomes in different helminth species, we found five conserved domains: 1) PF00014 (Kunitz/Bovine pancreatic trypsin inhibitor domain), GO:0004867 - serine-type endopeptidase inhibitor activity. This domain prevents or reduces the activity of serine-type endopeptidases. 2) PF00046 (Homeobox domain), GO:0003677 - DNA binding, is a protein structural domain that binds DNA or RNA and is thus commonly found in transcription factors (Gehring et al., 1992). 3) PF00188 (cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins), no GO identifier available. Members of this domain (PF00188) are most often secreted and have extracellular functions involving processes such as regulation of extracellular matrix, ion channel regulation, and in cell-cell adhesion among others. 4) PF00085 (Thioredoxin), GO:0045454 - cell redox homeostasis, acts as antioxidants by facilitating the reduction of other proteins by cysteine thiol-disulfide exchange. 5) PF07679 (Immunoglobulin I-set domain), no GO identifier available. This domain is found in several cell adhesion molecules. These common domains are involved in universal functionalities across secretomes from diverse species. Ortholog groups retrieved from OrthoMCL were not shared across species.

3.2.2. Species-specific domains

Species-specific domains may represent particular helminth adaptations to specific niches and mechanisms implemented by helminths to establish infection and survival in the host. In terms of exclusive domains, on average 22 domains were found to be species exclusive. Among the species with the highest amount of exclusive domains were *F. hepatica* (71 domains) and *S. mediterranea* (57 domains), *O. volvulus* (seven domains) and *T. callipaeda* (seven domains) had the lowest amount of exclusive domains. The identification of these species exclusive domains may enable the *in silico* selection of potential targets for antihelminthic agents. Supplementary files 4 show the species-specific domains across 44 species.

The ectoparasite *Gyrodactylus salaris* has the domain PF04203 (sortase) as an exclusive domain. According to the Pfam description sortase refers to a group of enzymes that modify surface proteins by recognizing and cleaving a carboxyl-terminal signal. These proteins often play important roles in virulence, infection, and colonization by pathogens. Domains involved in pili assembly in bacteria that appeared to be specific to *G. salaris* are PF03743 (Bacterial conjugation TrbI-like protein), which influences the kinetics of pilus; PF06122 (conjugative relaxosome accessory transposon protein), which is involved in pili formation, and PF06586 (TraK protein), which is known to be essential for pilus assembly, but its exact role in this process is unknown. Another *G. salaris* specific domain is PF02839 (carbohydrate binding domain). This short domain is found in many different glycosyl hydrolases and is structurally similar to the C-terminal chitin-binding domains (ChBD) of chitinase A1 and chitinase B. This domain could be related to degradation of chitin squama observed in the fish host *Salmo salar*.

In the cyst nematode *Globodera pallida* that infect potatoes, the specific domain PF05630 (necrosis inducing protein NPP1) is related to necrosis inducing proteins from oomycetes, fungi, and bacteria (Fellbrich et al. 2002). In *G. pallida*, this protein is involved in early stages of the infection and in callus formation. A domain associated with plant cell-wall hydrolysis such as PF00553 (cellulose binding domain), was another domain involved in host-helminth interaction, more exactly in the plant penetration stage. An unexpected domain related also with pathogenesis is PF07740 (spider toxin), which is a neurotoxin. This domain was found to be specific of *Taenia solium*, but its function remains unclear.

Particular characteristics of the helminth life cycle are the need to penetrate the host and the chronicity of the interaction between it and the host. For this purpose, helminths have developed several strategies. In *Trichuris suis*, an especific domain PF00151 (lipase) hydrolyses ester linkages of host triglycerides reflecting a particularity of its environment. This hydrolytic enzyme could also play a role in nematode penetration of the host by disrupting the tissues in the infected hosts (Bahlool et al. 2013). To maintain a chronic infection, cell protection from oxidative damage by reactive oxygen species (ROS) is an important process, a defense response domain PF00199 (catalase) was identified in the plant nematode *Meloidogyne hapla* as a specific domain.

3.2.3. Nematoda-specific domains and GO terms enrichment

Only three domains were identified as nematode specific. The domain PF01683 (EB module) that has no known function and is found associated with the kunitz domain (PF00014). The PF01682 (DB module) domain which also has no known function and is found associated with Ig (PF00047) and fn3 (PF00041) domains, as well as with some lipases (PF00657). These domains reflect specific functions or biological processes and are accessory domains that work in synergy with other domains. PF01060 (transthyretin-like family) was another nematode specific domain with unknown function.

We found interesting GO terms enriched in nematodes, such as nematode larval development (GO:0002119). This is a nematode specific term and is related to the development progression of the nematode larva over time from its formation to the mature form. Secreted proteins involved in molting process and signaling during this process are included in these GO terms. Metallopeptidase activity (GO:0008237) was also a nematode specific term that is related to proteins associated with tissue migration and haemoglobin degradation. Defense response (GO:0006952), which is related to proteins that act in response to the presence of pathogens and proteins associated to antimicrobial peptide activity are included in this term. Examples of these proteins have the PF15291 domain (Dermcidin, antibiotic peptide) that in *C. elegans* participates in the protection against pathogenic Gram-positive bacteria (Amaral et al. 2012). No OrthoMCL groups were shared across nematode species.

3.2.4. Platyhelminth-specific domains and GO terms enrichment

There were no Platyhelminth specific domains. However, we found enriched GO terms for this taxon. Here, we highlight some of them. Homophilic cell adhesion (GO:0007156), which is related to proteins involved in cell-to-cell comunication such as cadherins, laminins. These proteins are important in the host interaction and recognition (Leontovyč et al. 2016; Rowe et al. 2009). Platelet activation (GO:0030168) that is involved in a series of progressive events that leads to platelets activation. These events could be related to helminth invasion and migration. Leukocyte activation (GO:0045321) participates in the change of morphology and behavior of leukocytes resulting from exposure to a specific antigen, cellular ligand, or soluble factor. This term fits nicely to the proposed effect of secreted proteins and is clearly involved in helminth-host interaction. Negative regulation of cell communication (GO:0010648) is related to any process that decreases the frequency, rate, or extent of cell communication, such as signaling, cell-cell attachment, extracellular matrix interaction, or between a cell and any other aspect of its environment.

Within Platyhelminthes we identified a trematoda specific domain PF08034 (Trematode eggshell synthesis protein). This domain is present in the eggshell protein vitelline protein B1 (vpB1). vpB1 is produced by mature vitelline cells to form the hard protective trematode eggshell and is crucial for eggshell synthesis in Trematoda. There were no Cestoda specific domains. Monogenea and Turbellaria domains are species-specific for *G. salaris* and *S. mediterranea*, respectively. No OrthoMCL groups were shared across Platyhelminthes species, but Trematoda specific orthologs groups were identified such as OG5_222248 and OG5_185048.

3.2.5. Free-living species specific domains and GO terms enrichment

There were no specific domains for free-living helminths, but enriched GO terms for these helminths were identified and point to specific functions and biological processes associated with free-living. Response to gamma radiation (GO:0010332) was identified and is related to any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a gamma radiation stimulus. Another term related to radiation exposition was response to UV (GO:0009411), which is associated to ultraviolet stimulus. Some terms linked to microorganisms defense were identified such as defense response to Gram-negative bacterium (GO:0050829) and defense response to Gram-positive bacterium (GO:0050830). The term GO:0009410 is associated xenobiotic compound stimulus. Proteins involved in redox process and antioxidant are involved in this vital process. Other enriched term was chitin catabolic process (GO:0006032) that results in the breakdown of chitin, which is an abundant protein in the free-living environment.

In free-living nematodes, we found a specific glycoside hydrolases: PF00857 (isochorismatase family), PF05089 (glycoside hydrolase family 89) that includes enzymes with N-acetylglucosaminidase EC 3.2.1.50 activity, and PF12972 (glycoside hydrolase family 89). We identified 22 orthologous groups in free-living nematodes (*C. elegans* and *P. pacificus*).

S. mediterranea, the only representative of free-living Platyhelminthes analyzed in this work, had 57 species-specific Pfam domains, being the helminth with the second largest number of specific domains. An interesting domain PF03173 (putative carbohydrate binding domain) was found as *S. mediterranea* specific, this domain is involved in chitin degradation, which is one of the most abundant polysaccharides on Earth (Tews et al. 1996), its exclusivity could be related to specific aspects of the free living life style of *S. mediterranea*.

3.2.6. Plant-infecting helminth specific domains and GO terms enrichment

Cell-wall-degrading enzymes have no counterpart in most animals. Some examples of these key proteins are cellulases, xylanases, pectate lyases, and other members of the glycosyl hydrolase family (Dieterich and Sommer 2009). We identified Pfam domains related to these important proteins in the plant-helminth interaction.

PF00295 (glycosyl hydrolases family 28) is an *M. hapla* specific domain. This domain appears in plant bacterial pathogens, such as *Erwinia carotovora* or *Ralstonia solanacearum* (*Pseudomonas solanacearum*), and fungal pathogens such as *Aspergillus niger*, and is involved in maceration and soft-rotting of plant tissue. Specific glycosyl hydrolases were identified in *B. xylophilus* such as PF02015 (glycosyl hydrolase family 45) that contains enzymes with only one known activity: endoglucanase (EC 3.2.1.4) and PF00722 (glycosyl hydrolase family 16), which contains enzymes with a number of known activities: lichenase (EC 3.2.1.73), xyloglucan xyloglucosyltransferase (EC 2.4.1.207), agarase (EC 3.2.1.81), kappa-carrageenase (EC 3.2.1.83).

Specific glycosyl hydrolases families were identified in *G. pallida,* such as PF04616 (Glycosyl hydrolase family 43) that includes enzymes with the following activities: beta-xylosidase (EC 3.2.1.37), alpha-L-arabinofuranosidase (EC 3.2.1.55), arabinanase (EC 3.2.1.99), and xylanase (EC 3.2.1.8).

GO terms were enriched for plant helminths reflecting some specific and vital processes such as pectate lyase activity (GO:0030570). PF03211 (pectate lyase) is the domain related to this GO term. The plant-specific ortholog group was OG5_132299, which is also related to pectate lyase activity. This activity is responsible for the maceration and soft rotting of plant tissue and has been implicated in plant disease. Cellulase activity (GO:0008810), another plant-specific enriched term, is related to colonization of the plant by helminths and penetration in plant tissues. Defense response to bacterium (GO:0042742) is related to reactions triggered in response to the presence of bacterium that act to protect the cell or organism such as the effect of antibacterial peptide activity.

3.2.7. Secreted and non-secreted Pfam domains

Comparisons between secreted and nonsecreted protein domains across 44 helminth species allowed profiling the secretome fingerprints. 5,429 domains were identified in non-secreted and 2,345 in secreted proteins. 56 domains were secretome exclusive (supplementary file 5) and 2,289 were shared between secreted and nonsecreted proteins. Secretome specific domains are involved in processes such as: recognition, binding, degradation and uptake of extracellular complex nutrients, signal transduction, and adhesion. The OrthoMCL analysis showed 6,481 shared groups, 938 secretome specific groups, and 15,262 non-specific secretome groups.

3.3. Complex-repetitive secreted proteins and GO terms enrichment

Out of 20,607 secreted proteins with Pfam domain annotations, 98.9% contain ≤ 3 different domains, indicating these proteins have a fairly simple domain organization, the following GO terms were enriched in these proteins: protein disulfide isomerase activity (GO:0003756), cell redox homeostasis (GO:0045454), metalloendopeptidase activity (GO:0004222), serine-type endopeptidase inhibitor activity (GO:0004867), cysteine-type endopeptidase activity (GO:0004197), and serine-type carboxypeptidase activity (GO:0004185).

Our results also showed that 1.1% of the secreted proteins contained ≥ 4 different domains, which according to (Suh and Hutter 2012) are named “complex secreted proteins”. In these proteins, we found enriched GO terms such as: cell adhesion mediated by integrin (GO:0033627), basement membrane organization (GO:0071711), positive regulation of endopeptidase activity (GO:0010950), positive regulation of locomotion (GO:0040017), response to misfolded protein (GO:0051788), and regulation of cell proliferation (GO:0042127). Among these terms, there are proteins involved in protein-protein interaction including laminin, integrins, and proteins highly enriched in EGF domains and trombospondin repeats. Table 3 indicates the number of complex proteins per species according to the classification related to the number of different domains present in a protein. *S. mansoni* had the highest percentage (3.4%) of proteins with four or more different domains. *T. canis* did not have complex secreted proteins in the secretome.

The majority of potentially secreted proteins contained a small number of domains. Out of 4,451 architectures, 2,276 (51.1%) had a single domain (unidomain), which is represented in 14,444 proteins (70,1%). 168 (3.7%) architectures contain more than 10 domains, which are represented in 237 (1.1%) proteins. Only 62 architectures (1.4%) contain more than 20 domains, which are represented in 103 (0.5%) proteins. Comparing with non-secreted proteins, out of 25,708 architectures, 5,439 (21.1%) were unidomain represented in 202,862 (43.2%). 719 (2.8%) architectures contain more than 10 domains that are equivalent to 1,159 (0.24%) proteins and 103 (0.40%) contain more than 20 domains equivalent to 354 (0.07%) proteins (supplementary file 6). In general, secreted proteins are simpler in terms of the domain architecture than non-secreted proteins. However, large proteins containing more than 10 and 20 domains were overrepresented in the secreted proteins in comparison with non-secreted. This is explained by the presence in the secretome of proteins that have highly repetitive domains and are involved in cell to cell communication, protein binding, adhesion, among others.

Repetitive secreted proteins contain multiple copies of one or two different domains. The top 50 highly repetitive secreted proteins contain 83 proteins. Mainly nematode proteins were part of this classification (supplementary file 7). *C. elegans* was the species with more proteins (19 in total). Among the top 50 architectures of proteins highly repetitive most had a signal peptide, which means that they are proteins secreted via the classical pathway. Only two proteins had no peptide signal. The top 50 highly repetitive proteins contained 21 different domains (supplementary file 7). Domains involved in protein-protein interactions, present in cystein-rich proteins (particularly characteristic of secreted proteins) and endopeptidase inhibitor activity among others were characteristic of the repetitive proteins. Domains of unknown function were also present. Many of them were found exclusively in nematodes.

Table 3. Complex secreted proteins. Number of proteins low (≤ 3 domains) or high (≥ 4 domains) complexity.

Species	Proteins with ≤ 3 different domains (%)	Proteins with ≥ 4 different domains (%)
Nematoda		
Clade I		
Trichinella spiralis	406 (98.5)	6 (1.5)
Trichuris suis	403 (98.3)	7 (1.7)
Clade III		
Acanthocheilonema viteae	307 (97.5)	8 (2.5)
Ascaris suum	527 (99.2)	4 (0.8%)
Brugia malayi	456 (97.2)	13 (2.8)
Dirofilaria immitis	310 (97.1)	9 (2.9)
Dracunculus medinensis	279 (99.6)	1 (0.4)
Elaeophora elaphi	306 (99.6)	1 (0.4)
Enterobius vermicularis	347 (99.1)	3 (0.9)
Litomosoides sigmodontis	302 (97.4)	8 (2.6)
Loa loa	364 (97.8)	8 (2,2)
Onchocerca volvulus	352 (98.0)	7 (2)
Syphacia muris	320 (99.6)	1 (0.4)
Thelazia callipaeda	268 (98.9)	3 (1.1)
Toxocara canis	405 (100)	0
Wuchereria bancrofti	266 (99.6)	1 (0.4)
Clade IV		
Bursaphelenchus xylophilus	999 (98.9)	11 (1.1)
Globodera pallida	488 (99.1)	4 (0.9)
Meloidogyne hapla	359 (98.6)	5 (1.4)
Parastrongyloides trichosuri	759 (99.8)	1 (0.2)
Rhabditophanes sp. KR3021	585 (98.8)	7 (1.2)

Strongyloides ratti	537 (99.2)	4 (0.8)
Clade V		
Ancylostoma ceylanicum	732 (99.7)	2 (0.3)
Angiostrongylus costaricensis	254 (99.6)	1 (0.4)
Dictyocaulus viviparus	433 (99.0)	4 (1)
Haemonchus contortus	1,227 (98.9)	13 (1)
Necator americanus	601 (99.5)	3 (0.5)
Nippostrongylus brasiliensis	594 (99.6)	2 (0.4)
Oesophagostomum dentatum	763 (99.6)	3 (0.4)
Pristionchus pacificus	926 (99.3)	6 (0.7)
Caenorhabditis elegans	1,798 (98)	35 (2)
Platyhelminthes		
Cestoda		
Echinococcus multilocularis	253 (98.4)	4 (1.6)
Hydatigera taeniaeformis	229 (99.5)	1 (0.5)
Hymenolepis microstoma	212 (97.2)	6 (2.8)
Mesocestoides corti	175 (99.4)	1 (0.6)
Taenia solium	246 (99.6)	1 (0.4)
Trematoda		
Clonorchis sinensis	319 (98.7)	4 (1.3)
Fasciola hepática	449 (99.3)	3 (0.7)
Schistosoma haematobium	200 (98.5)	3 (1.5)
Schistosoma japonicum	246 (99.6)	1 (0.4)
Schistosoma mansoni	227 (96.6)	8 (3.4)
Opisthorchis viverrini	296 (98)	6 (2)
Monogenea		
Gyrodactylus salaris	265 (98.8)	3 (1.2)
Turbellaria		
Schmidtea mediterranea	592 (99.5)	3 (0.5)

3.4. Orthology classification of secreted proteins

OrthoMCL was used to arrange the secreted proteins into clusters and to identify groups of the most conserved proteins among secretomes. A total of 26,870 proteins (65%) out of 41,200 secreted proteins were identified across 44 species and were sorted in 7,419 orthologous clusters. 2,066 proteins matched other sequences in the OrthoMCL online database, but had "NO_ GROUP" (designation by OrthoMCL DB) (Table 4). Additionally 12,265 proteins had no OrthoMCL database hits and were considered unique. Certain clusters contained a large number of proteins. The most abundant ortholog group was OG5_186610 with 169 sequences. The proteins included in this ortholog group belong to the protein family CAP (PF00188, cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins) (supplementary file 8).

Table 4 shows OrthoMCL distributions of secretome proteins by species. Although the *S. mansoni* and *S. haematobium* secretomes were smaller, a higher proportion of the *S. mansoni* secretome (425/431 proteins) and *S. haematobium* secretome (364/379) had OrthoMCL group assignments. These proportions were comparable with *C. elegans* (3,099/3,121). We propose a ratio called OG_ diversity, which is result of number of orthologous groups/Total secretome. This value reflects the diversity of orthologs groups (OGs) per species. *S. haematobium* had the largest number (0.75) of OGs according to total secretome. The secretome with the least diversity of OGs was *T. spiralis* (0.29). We did not find an orthologous group shared by all 44 helminths species. However, we found specific orthologous groups shared between Nematoda or Platyhelminthes (supplementary file 9). We also compared orthologous groups between secreted *versus* non-secreted proteins; free-living *versus* non-free-living; plant helminths versus non-plant helminths (supplementary file 9).

Table 4. OrthoMCL results. Number of proteins observed in each category. Hits (%): OrthoMCL database hits. OGs: orthologs groups. OG_diversity: OGs/Secretome.

Species	Secretome (seqs)	Hits (%)	OGs	No_Group	No hits	OG_diversity
Nematoda						
Clade I						
Trichinella spiralis	1,146	464 (40.4)	342	8	682	0.29
Trichuris suis	727	474 (65.2)	365	14	253	0.50
Clade III						
Acanthocheilonema viteae	507	451 (88.9)	362	57	56	0.71
Ascaris suum	1,158	772 (66.6)	618	52	386	0.53
Brugia malayi	800	729 (91.1)	473	142	71	0.59

Dirofilaria immitis	593	500 (84.3)	378	98	93	0.63
Dracunculus medinensis	499	356 (71.3)	274	29	143	0.54
Elaeophora elaphi	525	470 (89.5)	361	92	55	0.68
Enterobius vermicularis	649	447 (68.8)	374	28	202	0.57
Litomosoides sigmodontis	542	488 (90)	377	91	54	0.69
Loa loa	803	609 (75.8)	441	114	194	0.54
Onchocerca volvulus	682	566 (82.9)	408	96	116	0.59
Syphacia muris	584	431 (73.8)	359	26	153	0.61
Thelazia callipaeda	532	416 (78.2)	336	59	116	0.63
Toxocara canis	858	558 (65)	454	52	300	0.52
Wuchereria bancrofti	517	461 (89.1)	348	94	56	0.67
Clade IV						
Bursaphelenchus xylophilus	2,077	1,197 (57.6)	759	35	880	0.36
Globodera pallida	1,218	580 (47.6)	430	22	638	0.35
Meloidogyne hapla	943	450 (47.7)	378	13	493	0.40
Parastrongyloides trichosuri	1,334	889 (66.6)	546	28	445	0.40
Rhabditophanes sp. KR3021	1,192	739 (61.9)	545	33	453	0.45
Strongyloides ratti	973	680 (69.8)	480	27	293	0.49
Clade V						
Ancylostoma ceylanicum	1,169	951 (81.3)	672	24	218	0.57
Angiostrongylus costaricensis	509	362 (71.1)	318	10	147	0.62
Dictyocaulus viviparus	825	639 (77.4)	570	16	186	0.69
Haemonchus contortus	2,419	1,756 (72.6)	934	39	663	0.38
Necator americanus	1,231	905 (73.5)	735	26	326	0.59
Nippostrongylus brasiliensis	1,469	928 (63.1)	721	27	541	0.49
Oesophagostomum dentatum	1,531	1,115 (72.8)	799	20	416	0.52
Pristionchus pacificus	2,388	1,274 (53.3)	849	58	1,114	0.35
Caenorhabditis elegans	3,121	3,099 (99.3)	1,958	315	22	0.62
Platyhelminthes						
Cestoda						
Echinococcus multilocularis	524	319 (60.8)	268	16	205	0.51
Hydatigera taeniaeformis	468	276 (58.9)	241	13	192	0.51
Hymenolepis microstoma	454	267 (58.8)	205	13	187	0.45
Mesocestoides corti	406	223 (54.9)	191	5	183	0.47
Taenia solium	538	303 (56.3)	256	10	235	0.47
Trematoda						
Clonorchis sinensis	656	427 (65)	322	34	229	0.49
Fasciola hepática	992	756 (76.2)	647	28	236	0.65
Schistosoma haematobium	379	364 (96)	286	44	15	0.75
Schistosoma japonicum	476	408 (85.7)	351	35	68	0.73
Schistosoma mansoni	431	425 (98.6)	270	72	6	0.62

Opisthorchis viverrini	832	388 (46.6)	286	23	444	0.34
Monogenea						
Gyrodactylus salaris	653	307 (47)	263	10	346	0.40
Turbellaria						
Schmidtea mediterranea	870	716 (82.3)	461	18	154	0.52

3.5. Dynamics of domain architecture

The emergence of proteins with new and species-specific domains or domain combinations could be one of the main mechanisms of secretome evolution and diversity. Therefore, distinct domain architectures can give rise to new functions and new molecular interaction alternatives. The protein architectures found in the secretomes are available in the supplementary file 6.

On average, each secreted protein had 1,73 domains, a proportion that did not vary significantly among taxonomic groups. *B. malayi* (2,89) had the highest amount of domains per protein and *F. hepatica* (1,27) the lowest one. The number of domains in non-secreted proteins was 1.65 per protein on average. Architecture diversity in a species is the number of architectures observed divided by the number of proteins with Pfam domains. If the ratio is close to 1 there is a greater diversity of architectures, because it means that every protein represents a specific architecture. According to our results, we can see that Platyhelminthes have the highest architecture diversity in secreted proteins (Figure 4).

Overall, 4,451 unique domain architectures were identified across 44 secretomes, with 2,830 of them (63.5%) appearing exclusively in a single secretome and only two ‘core’ architectures (0,04%) that were present in all secretomes. Non-secreted proteins displayed a richer domain architecture with 25,709 unique architectures identified, 13,934 (54.1%) of them appear exclusively in a single secretome and 158 ‘core’ architectures (0.61%). The ‘core’ architectures of the secretomes were composed by only one domain, corresponding to proteins that were related to essential and conserved functions such as Homeobox domain (PF00046) secreted by non-classical pathway and cysteine-rich secretory family (PF00188, SP) a classical pathway secreted protein.

The average number of exclusive architectures per species was 64, ranging from 24 to 157. *F. hepatica* had the highest number of exclusive domain architectures (41,8%). The lowest number of exclusive architectures was found in *T. callipaeda* (11%). Free-living helminths had values above 26% with highest value in *S. mediterranea* (32%). Plant parasitic helminths and the ectoparasite *G. salaris* also had high values of exclusive architectures. The highest numbers of architectures were found in *C. elegans* (594) *and M. corti* (143) with the lowest values.

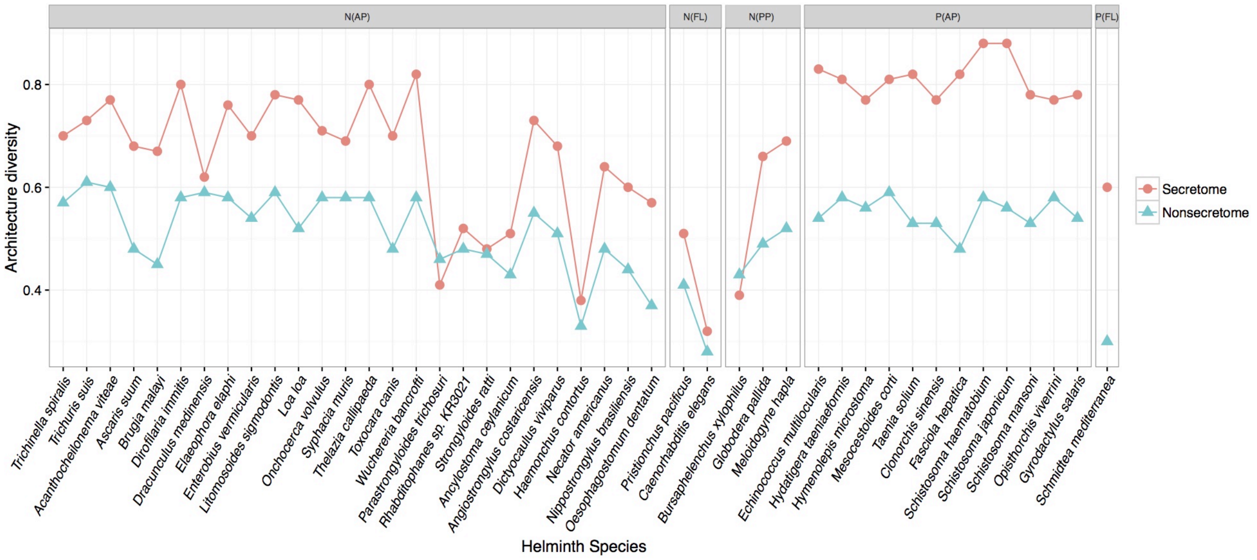

Figure 4. Protein architecture diversity (number of architectures per proteins with Pfam domains) distribution across secreted and non-secreted proteins in 44 helminth species. N: nematoda, P: platyhelminthes AP: animal parasite, PP: plant parasite, FL: free living.

4. Discussion

The identification of secreted proteins may provide a catalog of potential new immunomodulators, the development of new diagnostic tests, potential new drug targets and treatments (Geary et al. 2012). The need for new methods for diagnosis and control is well recognized (Liang et al. 2003; Melman et al. 2009). For this reason, the study of helminth secretomes may provide new venues for the development of control measures. The secretome of helminths also provides the repertoire of proteins that shows the imprint of adaptation to several habitats (Krijger et al. 2014).

The helminth expressed secretome is dynamic and adjusts to the developmental stage of the worm, the milieu in which each life cycle stage is exposed to and the state of the host immune system, among others. For example, it has been shown for *B. malayi* microfilariae that the composition of the secreted proteins depends on the local environment of the microfilarie in the human host (Moreno and Geary 2008). On the other hand, it has been demonstrated that not only secreted proteins are involved, but also carbohydrates (Jenkins et al. 2005; Thomas et al. 2003) and lipid mediators play important roles in the modulation of the host immune system by worms (Brattig et al. 2009; Van der Kleij et al. 2002).

The computationally set of secreted proteins we obtained still needs experimental validation, but it already provides strong clues as to which proteins to expect and the adaptations that evolved in organisms with distinct life styles.

We found that platyhelminthes have smaller secretomes, which is in agreement with previous studies (Gomez et al. 2015; Garg and Ranganathan 2012; Tsai et al. 2013). Small secretomes could be related to an environment or niche with compounds or nutrients that are easier to obtain and metabolize (Krijger et al. 2014). The *C. elegans* genome encodes a large proportion of secreted proteins as compared to other invertebrate and vertebrate organisms (Suh and Hutter 2012). According to our results, 3,121 proteins were predicted as secreted. This represents 13% of the filtered proteome, being the largest secretome in this study. *C. elegans* possesses an elaborated set of secreted proteins illustrating that genetic complexity does not necessarily correlate with anatomical complexity (Suh and Hutter 2012).

The higher number of secreted proteins in the free living nematodes *C. elegans* and *P. pacificus* permits the use of a wider variety of substrates present in soil or plant debris, which are likely more difficult to degrade than those available in animal hosts (Krijger et al. 2014). The plant helminth *B. xylophilus* also had a larger secretome than other animal helminths. This finding may be explained assuming that animal hosts represent a nutritionally simpler environment than plant hosts. Competition with other microorganisms in free-living and other plant helminths are possibly more severe than in animal hosts, which may affect the number of secreted proteins involved in counteracting competitors (Krijger et al. 2014). Therefore, the secretome size could be related to lifestyle and parasite environment rather than with proteome size. *S. haematobium* has the smallest secretome across 44 helminth species. The small size of animal helminth secretomes may reflect an adaptation or a selective advantage in order to evade recognition by the immune system (Krijger et al. 2014).

However, according to our results *H. contortus*, a pathogenic nematode of ruminants, had a large secretome, almost equal to free living nematodes and plant helminthes. The *H. contortus* secretome is particularly rich in peptidases linked to key roles in host invasion, locomotion, migration into stomach tissue (during the histotropic phase), degradation of blood, and other proteins as evasive mechanisms (Schwarz et al. 2013).

The study of protein domains has provided solutions to some human diseases. One example is the case of Kunitz domains (protease inhibitor domain) that are stable as standalone peptides, able to recognize specific protein structures. These properties have led to attempts at developing biopharmaceutical drugs targeting this domain (Lehmann 2008). The first of these drugs to be marketed was a kallikrein inhibitor called ecallantide, which was used for the treatment of angioedema (Lehmann 2008).

Protein domains are an independent, compact and stable protein structural units that folds independently of other units, thus with potentially different biological functions (Barrera et al. 2014) and may catalyze different reactions (Barrera et al. 2014). The study of domain diversity in secretomes is a fast and effective way to characterize protein diversity and may provide clues to the different lifestyles and environments in which these organisms live. The architecture of secreted proteins are therefore relevant for the understanding of the interaction of the helminth and the environment. In addition, the prediction of domain architectures enables the determination of the overall protein function and diversity and has been used to transfer genomic annotations in newly sequenced genomes (Barrera et al. 2014).

Few domains were found in common across the secretomes, possibly because helminths have evolved separately as parasitic worms, leading to specific adaptations for their particular niche (Zarowiecki and Berriman 2015). Only five domains were common across 44 secretomes, these domains were universal and are involved in biological functions such as peptidases, inhibitors, antioxidants, and cell adhesion. All these are important processes in the interaction with the host. Of the five domains, the Kunitz domain (PF00014) was the most frequent and is present in proteins containing 117 distinct architectures suggesting that this domain is relevant in very diverse activities, one of them being the inhibition of proteases.

Species-specific domains are involved in particular functions related to lifestyle or ecological niche in helminths. In *G. salaris*, a salmon ectoparasite, it was possible to identify domains related to pilus formation in bacteria. These domains could be participating in the attachment of the ectoparasite to the host and have important roles in virulence, infection, and colonization by pathogens. However, the presence of these bacterial domains in the secretome of *G. salaris* could be also related to the ectoparasite lifestyle, because *Gyrodactilus spp* feeds on host mucus and epithelial cells (Cable and Harris 2002). On the other hand, Bird and colleagues (2009) postulated that pilus formation domains have been acquired from bacteria by ancestral nematodes via horizontal gene transfer, and such events would be relevant for the establishment of the parasitic life style (Bird et al. 2009).

Exclusive plant domains such as the ones found in cellulases, pectate lyases, and cell wall degrading enzymes are key adaptations towards plant parasitism most probably achieved by horizontal gene transfer from a rhizobial bacterium (Dieterich and Sommer 2009). In the potato parasite helminth *G. pallida,* we found the exclusive domain necrosis inducing protein (NPP1)

(PF05630) with the architecture SP (signal peptide) and PF05630, that is involved in early stages of the infection and is present in the potato fungus *Phytopthora infestans*, suggesting that plant parasitism has evolved from fungal associations (Dieterich and Sommer 2009).

Specific glycoside hydrolases (GH) found in *B. xylophilus* suggested that they play important roles in fungal cell wall degradation. GH 16 is involved in fungal cell wall degradation with endo-beta-1,3-glucanase activity and beta 1,3 glucan is one of the main components of the cell wall (Adams 2004). This protein was only present in *B. xylophilus*, reflecting the difference in their food sources. Some secreted GH in *B. xylophilus* were acquired from other organisms by horizontal gene transfer as supported by phylogenetic evidence, which showed that they were gained from ascomycete fungi and bacteria (Shinya et al. 2013; Kikuchi et al. 2011). Another interesting GH protein identified was GH 18 (chitinase) that is involved in insect cuticle degradation and have antifungal roles (Staats et al. 2014). This protein was present in some helminths that have insect vectors.

The most represented domains across helminth secretomes were ranked in a top 25 classification. The ShK (PF01549) domain was the most common domain. It is suggested that this domain is important in parasitic interactions (Heizer et al. 2013). ShK proteins can inhibit calcium-dependent lymphocyte activation (Tudor et al., 1996). This suggests a direct immunomodulatory role for ShK homologs in helminths and its potential biopharmaceutical applications. Kunitz_BPTI (PF00014) and Inhibitor_I29 (PF08246) are protease inhibitors. It has been suggested that they are involved in protecting helminths from host molecules, in particular, those derived from the gastrointestinal tract, such as a broad diversity of peptidases (Heizer et al. 2013). In this way gastrointestinal helminths can safely navigate and survive within host digestive tract.

CAP (PF00188) domains were also among the most prevalent domains across the helminth secretomes analyzed here. They play roles in larval migration and evasion of the host's immune response (Sotillo et al. 2014). CAP domains have been also found associated with proteins with immunodulatory activity (Cantacessi et al. 2009) and have been studied in some parasitic nematode species such as the hookworm *A. caninum* (Hawdon, Narasimhan, and Hotez 1999), and the murine strongyloid nematode, *Heligmosomoides polygyrus* (Moreno et al. 2011).

The trypsin (PF00089) domain is involved in the breakdown of proteins. Trypsin domains were up-regulated in the parasitic stages of the nematodes *Cooperia oncophora* and *Ostertagia ostertagi* (Heizer et al. 2013). Proteins associated with this domain play a role in the feeding process (Goyal et al. 2005). These secreted proteases may also participate in countering the host immune responses by hydrolyzing antibodies or in parasite establishment in the host (Heizer et al. 2013).

The lectin_C (PF00059) domain is related to extracellular metazoan proteins with diverse functions. In general, it is involved in calcium dependent carbohydrate binding. This domain has been linked to proteins involved in the host-parasite interface, which may assist in evading the host immune response (Loukas and Maizels 2000). Some nematode C-type lectins have been observed to the parasite surface such as the epicuticle.

Hosts use oxidative stress as a means of combating parasites (Schirmer et al. 1987). It was hypothesized that parasites would have a very well developed redox system to defend themselves against reactive oxygen species attacks (Zarowiecki and Berriman 2015). Domains involved in detoxification process were part of the most represented domains found in the present work such as transthyretin-like family (PF01060) and thioredoxin (PF00085). Other interesting process in host-helminth interaction was lipid catabolism due to the lipid-rich environment in which helminths reside. Lipase_3 (PF01764) and coesterase (PF00135) are involved in these processes and are placed among the most represented domains across the helminth secretomes. Domains involved in cell adhesion or protein-protein interaction were placed in this top 25 ranking such as: I-set (PF07679), CUB (PF00431), PAN_1 (PF00024), and TSP_1 (PF00090).

Among the most represented predicted protein families across secretomes, we found proteases, which are involved in blood coagulation, protein metabolism, immune reactions, and tissue remodeling (Dzik 2006). The role of proteases in extracellular matrix degradation seems to be exclusive of parasites. Thus, free-living helminths do not secrete extracellular matrix degradation proteases (Lackey et al. 1989). The specific action and release of peptidases after host infection is an important function in the transition from a free-living organism to parasite (Hawdon et al. 1995; Gamble and Mansfield 1996). Metallo and cisteine proteases (PF01400 and PF00112, respectively) are involved in the tissue/cell invasion process and nutrient uptake. Astacin (PF01400) is most represented domain in nematodes. This could support the idea that these proteins had an expansion in nematodes as reported by (Park et al. 2010). Furthermore, studies suggest that cysteine proteases appeared early in the evolution degrading intra and extracellular proteins (Sajid and McKerrow 2002). On the other hand, aspartic proteases correspond to a major enzymatic class from parasitic helminths and play a key role in the ability to degrade haemoglobin (Brinkworth et al. 2001). Such are the targets of protective antibodies against the human parasite *Necator americanus* (Pearson et al. 2009).

Domain repetitions in multidomain proteins are important for the overall domain function (Messih et al. 2012). Domain repetition is a predominant mechanism for protein diversity and evolution. For example, in the glutamate receptor interacting protein (GRIP) contains seven PDZ domains, two of them interact with a α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor, but only in the presence of the adjacent copies (Messih et al. 2012).

Our results showed that domains such as PAN, TSP-1, VWA, I-set, and EGF are usually present as tandem copies, sometimes in combination with other adhesive domains. The presence of these repeats suggests that they play a common functional role during the invasion process (host interaction). Adhesive proteins play important roles in ligand binding, cell-cell and cell-extracellular matrix interactions (Bork and Rohde 1991).

In the present study, we provide access to the full list of core domains, exclusive domains, and domain architectures of the secreted proteins across 44 helminths. This information can be a useful resource for researchers interested in comparative studies of secretomes across different helminths in order to know more about the protein evolution in these crucial proteins and their interaction with the host and to understand protein function and evolution. The number and order of the domains will determine the function, as is the case of multimodular enzymes. The rational is that proteins containing the same domain composition could be similarly annotated (Barrera et al. 2014).

Secreted proteins have higher architecture diversity comparing with non-secreted proteins. The only exceptions were the secretomes of *B. xylophilus* and *P. trichosuri*. This metrics was calculated in order shed light on the diversity of secreted proteins in terms of protein domains and architectures, once domain and architecture diversity point toward the existence of different mechanisms to attach or interact with host components and the environment. On average, Platyhelminthes had the smaller secretomes across 44 helminth species, albeit more diverse architectures.

5. Conclusions

This is the first proteome-wide comparative study of predicted secretomes in helminths using species with different lifestyles and environments to provide relevant information towards the understanding of their diversity, function, and adaptation. The comparison of secretome profiles across 44 helminth species revealed their differences and similarities, reflecting the organisms' life style and generating a list of proteins that should be considered for development of new helminth control strategies. This is important once helminths have evolved various strategies to invade host tissues and to evade or even manipulate the immune system. Only five domains were conserved across 44 secretomes and reveal mechanisms that appear to be conserved among plant, animal, and free-living helminths. Our findings also indicate that the secretome composition is not conserved across species and the differences suggest possible unique adaptations to specific niche.

According to our observations, the secretome size does not depend on the increase of proteome size. The majority of proteins in the secretomes had less than 300 amino acids, which means that these proteins have a fairly simple domain organization (unidomain proteins). Proteins with more than 1,000 amino acids were poorly represented across the secretomes. In platyhelminthes, the secretome average size was noticeably smaller compared with nematodes and also among animal and plant helminths.

Comparisons between secreted and non-secreted proteins allowed profiling the helminth secretomes. Secretome-specific domains are involved in biological processes such as: recognition, binding, degradation and uptake of complex extracellular nutrients, signal transduction and adhesion. In general, secreted proteins are simpler in terms of the domain number than non-secreted proteins. However, large proteins containing more than 10 and 20 domains were over represented in secreted in comparison to non-secreted proteins. This is explained by the presence in the secretome of proteins involved in cell to cell communication, protein binding, adhesion, among others that contain repetitive domains such as PAN, TSP-1, VWA, I-set, and EGF. Proteins with these repeats may play a common functional role during the invasion process (host interaction). Regarding the architecture diversity, secreted proteins have the highest diversity comparing with non-secreted proteins, which could be related to selective diversifying pressures.

Our approach is based on data available from recently sequenced genomes. All the predicted proteins in this study are preliminary predictions, which need to be experimentally validated assays or refined in future improved versions of genome assemblies.

Acknowledgments

The authors acknowledge the use of computing resources of the Bioinformatics Platform at Fiocruz-Minas. This work was supported by the National Institutes of Health-NIH/Fogarty International Center (TW007012 and 1P50AI098507-01) to GO, CAPES (REDE 21/2015 and 070/13) to GO, FAPEMIG (RED-00014-14 and PPM-00189-13) to GO and CNPq (304138/2014-2 and) to GO. GO is a CNPq fellow (309312/2012-4). FSO a FAPEMIG fellow and YCA a CAPES fellow.

References

Adams, David J. 2004. “Fungal Cell Wall Chitinases and Glucanases.” *Microbiology*. doi:10.1099/mic.0.26980-0.

Amaral, André C., Osmar N. Silva, Nathália C C R Mundim, Maria J A De Carvalho, Ludovico Migliolo, Jose R S A Leite, Maura V. Prates, Anamélia L. Bocca, Octávio L. Franco, and Maria S S Felipe. 2012. “Predicting Antimicrobial Peptides from Eukaryotic Genomes: In Silico Strategies to Develop Antibiotics.” *Peptides* 37 (2): 301–8. doi:10.1016/j.peptides.2012.07.021.

Bahlool, Qusay Z M, Alf Skovgaard, Per W Kania, and Kurt Buchmann. 2013. “Effects of Excretory/secretory Products from Anisakis Simplex (Nematoda) on Immune Gene Expression in Rainbow Trout (Oncorhynchus Mykiss).” *Fish & Shellfish Immunology* 35 (3): 734–39. doi:10.1016/j.fsi.2013.06.007.

Barrera, Alejandro, Ana Alastruey-Izquierdo, María J. Martín, Isabel Cuesta, and Juan Antonio Vizcaíno. 2014. “Analysis of the Protein Domain and Domain Architecture Content in Fungi and Its Application in the Search of New Antifungal Targets.” *PLoS Computational Biology* 10 (7). doi:10.1371/journal.pcbi.1003733.

Bendtsen, Jannick Dyrløv, Lars Juhl Jensen, Nikolaj Blom, Gunnar Von Heijne, and Søren Brunak. 2004. “Feature-Based Prediction of Non-Classical and Leaderless Protein Secretion.” *Protein Engineering, Design and Selection* 17 (4): 349–56. doi:10.1093/protein/gzh037.

Bird, David McK, Valerie M Williamson, Pierre Abad, James McCarter, Etienne G J Danchin, Philippe Castagnone-Sereno, and Charles H Opperman. 2009. “The Genomes of Root-Knot Nematodes.” *Annual Review of Phytopathology* 47: 333–51. doi:10.1146/annurev-phyto-080508-081839.

Blaxter, M L, P De Ley, J R Garey, L X Liu, P Scheldeman, A Vierstraete, J R Vanfleteren, et al. 1998. “A Molecular Evolutionary Framework for the Phylum Nematoda.” *Nature* 392 (6671): 71–75. doi:10.1038/32160.

Bork, P, and K Rohde. 1991. “More von Willebrand Factor Type A Domains? Sequence Similarities with Malaria Thrombospondin-Related Anonymous Protein, Dihydropyridine-Sensitive Calcium Channel and Inter-Alpha-Trypsin Inhibitor.” *The Biochemical Journal*, November, 908–10. http://www.ncbi.nlm.nih.gov/pubmed/1659389.

Brattig, Norbert W., Arline Schwohl, Achim Hoerauf, and Dietrich W. Büttner. 2009. “Identification of the Lipid Mediator Prostaglandin E2 in Tissue Immune Cells of Humans Infected with the Filaria Onchocerca Volvulus.” *Acta Tropica* 112 (2): 231–35. doi:10.1016/j.actatropica.2009.07.018.

Brindley, Paul J., Makedonka Mitreva, Elodie Ghedin, and Sara Lustigman. 2009. “Helminth Genomics: The Implications for Human Health.” *PLoS Neglected Tropical Diseases*. doi:10.1371/journal.pntd.0000538.

Brinkworth, Ross I., Paul Prociv, Alex Loukas, and Paul J. Brindley. 2001. “Hemoglobin-Degrading, Aspartic Proteases of Blood-Feeding Parasites. Substrate Specificity Revealed by Homology Models.” *Journal of Biological Chemistry* 276 (42): 38844–51. doi:10.1074/jbc.M101934200.

Cable, J., and P. D. Harris. 2002. “Gyrodactylid Developmental Biology: Historical Review, Current Status and Future Trends.” In *International Journal for Parasitology*, 32:255–80.

doi:10.1016/S0020-7519(01)00330-7.

Cantacessi, C., B. E. Campbell, A. Visser, P. Geldhof, M. J. Nolan, A. J. Nisbet, J. B. Matthews, et al. 2009. "A Portrait of the 'SCP/TAPS' proteins of Eukaryotes - Developing a Framework for Fundamental Research and Biotechnological Outcomes." *Biotechnology Advances*. doi:10.1016/j.biotechadv.2009.02.005.

Cass, Cynthia L, Jeffrey R Johnson, Lindsay L Califf, Tao Xu, Hector J Hernandez, Miguel J Stadecker, John R Yates, and David L Williams. 2007. "Proteomic Analysis of Schistosoma Mansoni Egg Secretions." *Molecular and Biochemical Parasitology* 155 (2): 84–93. doi:10.1016/j.molbiopara.2007.06.002.

Conesa, Ana, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. 2005. "Blast2GO: A Universal Tool for Annotation, Visualization and Analysis in Functional Genomics Research." *Bioinformatics* 21 (18): 3674–76. doi:10.1093/bioinformatics/bti610.

Cuesta-Astroz, Yesid, Larissa L S Scholte, Fabiano Sviatopolk Mirsky Pais, Guilherme Oliveira, and Laila A. Nahum. 2014. "Evolutionary Analysis of the Cystatin Family in Three Schistosoma Species." *Frontiers in Genetics* 5 (JUL). doi:10.3389/fgene.2014.00206.

Dieterich, Christoph, and Ralf J. Sommer. 2009. "How to Become a Parasite - Lessons from the Genomes of Nematodes." *Trends in Genetics* 25 (5): 203–9. doi:10.1016/j.tig.2009.03.006.

Dzik, Jolanta M. 2006. "Molecules Released by Helminth Parasites Involved in Host Colonization." *Acta Biochimica Polonica* 53 (1): 33–64. doi:20061201 [pii].

Emanuelsson, O, H Nielsen, S Brunak, and G von Heijne. 2000. "Predicting Subcellular Localization of Proteins Based on Their N-Terminal Amino Acid Sequence." *Journal of Molecular Biology* 300 (4): 1005–16. doi:10.1006/jmbi.2000.3903.

Fellbrich, Guido, Annette Romanski, Anne Varet, Beatrix Blume, Frédéric Brunner, Stefan Engelhardt, Georg Felix, Birgit Kemmerling, Magdalena Krzymowska, and Thorsten Nürnberger. 2002. "NPP1, a Phytophthora-Associated Trigger of Plant Defense in Parsley and Arabidopsis." *Plant Journal* 32 (3): 375–90. doi:10.1046/j.1365-313X.2002.01454.x.

Ferguson, Brian J, Stephen A Newland, Sarah E Gibbs, Panagiotis Tourlomousis, Paula Fernandes dos Santos, Meghana N Patel, Samuel W Hall, et al. 2015. "The Schistosoma Mansoni T2 Ribonuclease Omega-1 Modulates Inflammasome-Dependent IL-1β Secretion in Macrophages." *International Journal for Parasitology* 45 (13): 809–13. doi:10.1016/j.ijpara.2015.08.005.

Finn, Robert D, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Jaina Mistry, Alex L Mitchell, Simon C Potter, et al. 2016. "The Pfam Protein Families Database: Towards a More Sustainable Future." *Nucleic Acids Research* 44 (D1): D279-85. doi:10.1093/nar/gkv1344.

Fischer, Steve, Brian P Brunk, Feng Chen, Xin Gao, Omar S Harb, John B Iodice, Dhanasekaran Shanmugam, David S Roos, and Christian J Stoeckert. 2011. "Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes into New Ortholog Groups." *Current Protocols in Bioinformatics* Chapter 6 (September): Unit 6.12.1-19. doi:10.1002/0471250953.bi0612s35.

Gamble, H. Ray, and Linda S. Mansfield. 1996. "Characterization of Excretory-Secretory Products from Larval Stages of Haemonchus Contortus Cultured in Vitro." *Veterinary Parasitology* 62 (3–4): 291–305. doi:10.1016/0304-4017(95)00871-3.

Garg, Gagan, and Shoba Ranganathan. 2011. “In Silico Secretome Analysis Approach for next Generation Sequencing Transcriptomic Data.” *BMC Genomics* 12 Suppl 3 (Suppl 3): S14. doi:10.1186/1471-2164-12-S3-S14.

———. 2012. “Helminth Secretome Database (HSD): A Collection of Helminth Excretory/secretory Proteins Predicted from Expressed Sequence Tags (ESTs).” *BMC Genomics* 13 Suppl 7: S8. doi:10.1186/1471-2164-13-S7-S8.

Geary, James, Mohamed Satti, Yovany Moreno, Nicole Madrill, Doug Whitten, Selwyn a Headley, Dalen Agnew, Timothy Geary, and Charles Mackenzie. 2012. “First Analysis of the Secretome of the Canine Heartworm, Dirofilaria Immitis.” *Parasites & Vectors* 5 (1): 140. doi:10.1186/1756-3305-5-140.

Gomez, Sandra, Laura Adalid-Peralta, Hector Palafox-Fonseca, Vito Adrian Cantu-Robles, Xavier Soberón, Edda Sciutto, Gladis Fragoso, et al. 2015. “Genome Analysis of Excretory/Secretory Proteins in Taenia Solium Reveals Their Abundance of Antigenic Regions (AAR).” *Scientific Reports* 5 (May): 9683. doi:10.1038/srep09683.

Goyal, Kshamata, Claudia Pinelli, Sarah L. Maslen, Rakesh K. Rastogi, Elaine Stephens, and Alan Tunnacliffe. 2005. “Dehydration-Regulated Processing of Late Embryogenesis Abundant Protein in a Desiccation-Tolerant Nematode.” *FEBS Letters* 579 (19): 4093–98. doi:10.1016/j.febslet.2005.06.036.

Greenbaum, D, N M Luscombe, R Jansen, J Qian, and M Gerstein. 2001. “Interrelating Different Types of Genomic Data, from Proteome to Secretome: ’Oming in on Function.” *Genome Research* 11 (9): 1463–68. doi:10.1101/gr.207401.

Hahn, Christoph, Bastian Fromm, and Lutz Bachmann. 2014. “Comparative Genomics of Flatworms (Platyhelminthes) Reveals Shared Genomic Features of Ecto- and Endoparastic Neodermata.” *Genome Biology and Evolution* 6 (5): 1105–17. doi:10.1093/gbe/evu078.

Harris, Arthur R C, Robert J Russell, Alan D Charters, Royal Perth Hospital, and X G P O Box. 1984. “A Review of Schistosomiasis in Immigrants in Western Australia, Demonstrating the Unusual Longevity of Schistosoma Mansoni.” *Trans R Soc Trop Med Hyg* 78 (3): 385–88.

Harris, Todd W., Joachim Baran, Tamberlyn Bieri, Abigail Cabunoc, Juancarlos Chan, Wen J. Chen, Paul Davis, et al. 2014. “WormBase 2014: New Views of Curated Biology.” *Nucleic Acids Research* 42 (D1). doi:10.1093/nar/gkt1063.

Hawdon, J M, B F Jones, M A Perregaux, and P J Hotez. 1995. “Ancylostoma Caninum: Metalloprotease Release Coincides with Activation of Infective Larvae in Vitro.” *Experimental Parasitology* 80 (2): 205–11. doi:10.1006/expr.1995.1025.

Hawdon, J M, S Narasimhan, and P J Hotez. 1999. “Ancylostoma Secreted Protein 2: Cloning and Characterization of a Second Member of a Family of Nematode Secreted Proteins from Ancylostoma Caninum.” *Molecular and Biochemical Parasitology* 99 (2): 149–65. http://www.ncbi.nlm.nih.gov/pubmed/10340481.

Heizer, Esley, Dante S Zarlenga, Bruce Rosa, Xin Gao, Robin B Gasser, Jessie De Graef, Peter Geldhof, and Makedonka Mitreva. 2013. “Transcriptome Analyses Reveal Protein and Domain Families That Delineate Stage-Related Development in the Economically Important Parasitic Nematodes, Ostertagia Ostertagi and Cooperia Oncophora.” *BMC Genomics* 14 (1): 118. doi:10.1186/1471-2164-14-118.

Hotez, P.J., P.J. Brindley, J.M. Bethony, C.H. King, E.J. Pearce, and Julie Jacobson. 2008. “Helminth Infections: The Great Neglected Tropical Diseases.” *The Journal of Clinical*

Investigation 118 (4): 1311–21. doi:10.1172/JCI34261.tion.

Jenkins, Stephen John, James Philip Hewitson, Stephanie Ferret-Bernard, and Adrian Paul Mountford. 2005. “Schistosome Larvae Stimulate Macrophage Cytokine Production through TLR4-Dependent and -Independent Pathways.” *International Immunology* 17 (11): 1409–18. doi:10.1093/intimm/dxh319.

Jones, Philip, David Binns, Hsin Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. “InterProScan 5: Genome-Scale Protein Function Classification.” *Bioinformatics* 30 (9): 1236–40. doi:10.1093/bioinformatics/btu031.

Kikuchi, Taisei, James A Cotton, Jonathan J Dalzell, Koichi Hasegawa, Natsumi Kanzaki, Paul McVeigh, Takuma Takanashi, et al. 2011. “Genomic Insights into the Origin of Parasitism in the Emerging Plant Pathogen <italic>Bursaphelenchus Xylophilus</italic>.” *PLoS Pathog* 7 (9): e1002219. doi:10.1371/journal.ppat.1002219.

Krijger, Jorrit-Jan, Michael R Thon, Holger B Deising, and Stefan G R Wirsel. 2014. “Compositions of Fungal Secretomes Indicate a Greater Impact of Phylogenetic History than Lifestyle Adaptation.” *BMC Genomics* 15: 722. doi:10.1186/1471-2164-15-722.

Krogh, a, B Larsson, G von Heijne, and E L Sonnhammer. 2001. “Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes.” *Journal of Molecular Biology* 305 (3): 567–80. doi:10.1006/jmbi.2000.4315.

Lackey, Angela, Eric R. James, Judy A. Sakanari, Steven D. Resnick, Margaret Brown, Albert E. Bianco, and James H. McKerrow. 1989. “Extracellular Proteases of Onchocerca.” *Experimental Parasitology* 68 (2): 176–85. doi:10.1016/0014-4894(89)90095-7.

Lehmann, Andreas. 2008. “Ecallantide (DX-88), a Plasma Kallikrein Inhibitor for the Treatment of Hereditary Angioedema and the Prevention of Blood Loss in on-Pump Cardiothoracic Surgery.” *Expert Opinion on Biological Therapy* 8 (8): 1187–99. doi:10.1517/14712598.8.8.1187.

Leontovyč, Roman, Neil D Young, Pasi K Korhonen, Ross S Hall, Patrick Tan, Libor Mikeš, Martin Kašný, Petr Horák, and Robin B Gasser. 2016. “Comparative Transcriptomic Exploration Reveals Unique Molecular Adaptations of Neuropathogenic Trichobilharzia to Invade and Parasitize Its Avian Definitive Host.” *PLoS Neglected Tropical Diseases* 10 (2): e0004406. doi:10.1371/journal.pntd.0004406.

Liang, You Sheng, Jian Rong Dai, Yin Chang Zhu, Gerald C. Coles, and Michael J. Doenhoff. 2003. “Genetic Analysis of Praziquantel Resistance in Schistosoma Mansoni.” *Southeast Asian Journal of Tropical Medicine and Public Health* 34 (2): 274–80.

Loukas, Alex, and Rick M. Maizels. 2000. “Helminth C-Type Lectins and Host-Parasite Interactions.” *Parasitology Today*. doi:10.1016/S0169-4758(00)01704-X.

Lustigman, Sara, Roger K. Prichard, Andrea Gazzinelli, Warwick N. Grant, Boakye A. Boatin, James S. McCarthy, and María Gloria Basáñez. 2012. “A Research Agenda for Helminth Diseases of Humans: The Problem of Helminthiases.” *PLoS Neglected Tropical Diseases*. doi:10.1371/journal.pntd.0001582.

Maizels, Rick M, and Maria Yazdanbakhsh. 2003. “Immune Regulation by Helminth Parasites: Cellular and Molecular Mechanisms.” *Nature Reviews. Immunology* 3 (9): 733–44. doi:10.1038/nri1183.

Marcilla, Antonio, María Trelis, Alba Cortés, Javier Sotillo, Fernando Cantalapiedra, María Teresa

Minguez, María Luz Valero, et al. 2012. “Extracellular Vesicles from Parasitic Helminths Contain Specific Excretory/secretory Proteins and Are Internalized in Intestinal Host Cells.” *PloS One* 7 (9): e45974. doi:10.1371/journal.pone.0045974.

Mehta, Angela, Ana C M Brasileiro, Djair S L Souza, Eduardo Romano, Magnólia A. Campos, Maria F. Grossi-De-Sá, Marília S. Silva, et al. 2008. “Plant-Pathogen Interactions: What Is Proteomics Telling Us?” *FEBS Journal*. doi:10.1111/j.1742-4658.2008.06528.x.

Melman, Sandra D., Michelle L. Steinauer, Charles Cunningham, Laura S. Kubatko, Ibrahim N. Mwangi, Nirvana Barker Wynn, Martin W. Mutuku, et al. 2009. “Reduced Susceptibility to Praziquantel among Naturally Occurring Kenyan Isolates of Schistosoma Mansoni.” *PLoS Neglected Tropical Diseases* 3 (8). doi:10.1371/journal.pntd.0000504.

Messih, Mario Abdel, Meghana Chitale, Vladimir B. Bajic, Daisuke Kihara, and Xin Gao. 2012. “Protein Domain Recurrence and Order Can Enhance Prediction of Protein Functions.” *Bioinformatics* 28 (18). doi:10.1093/bioinformatics/bts398.

Moreno, Yovany, and Timothy G Geary. 2008. “Stage- and Gender-Specific Proteomic Analysis of Brugia Malayi Excretory-Secretory Products.” *PLoS Neglected Tropical Diseases* 2 (10): e326. doi:10.1371/journal.pntd.0000326.

Moreno, Yovany, Pierre Paul Gros, Mifong Tam, Mariela Segura, Rajesh Valanparambil, Timothy G. Geary, and Mary M. Stevenson. 2011. “Proteomic Analysis of Excretory-Secretory Products of Heligmosomoides Polygyrus Assessed with Next-Generation Sequencing Transcriptomic Information.” *PLoS Neglected Tropical Diseases* 5 (10). doi:10.1371/journal.pntd.0001370.

Nahum, Laila A., Marina M. Mourão, and Guilherme Oliveira. 2012. “New Frontiers in Schistosoma Genomics and Transcriptomics.” *Journal of Parasitology Research*. doi:10.1155/2012/849132.

Nombela, César, Concha Gil, and W. LaJean Chaffin. 2006. “Non-Conventional Protein Secretion in Yeast.” *Trends in Microbiology* 14 (1): 15–21. doi:10.1016/j.tim.2005.11.009.

Oliveira, Guilherme, and Raymond J Pierce. 2015. “How Has the Genomics Era Impacted Schistosomiasis Drug Discovery?” *Future Medicinal Chemistry* 7 (6): 685–87. doi:10.4155/fmc.15.30.

Park, Ja-On, Jie Pan, Frank Möhrlen, Marcus-Oliver Schupp, Robert Johnsen, David L Baillie, Richard Zapf, Donald G Moerman, and Harald Hutter. 2010. “Characterization of the Astacin Family of Metalloproteases in C. Elegans.” *BMC Developmental Biology* 10 (1): 14. doi:10.1186/1471-213X-10-14.

Pearson, Mark S, Jeffrey M Bethony, Darren A Pickering, Luciana M de Oliveira, Amar Jariwala, Helton Santiago, Aaron P Miles, et al. 2009. “An Enzymatically Inactivated Hemoglobinase from Necator Americanus Induces Neutralizing Antibodies against Multiple Hookworm Species and Protects Dogs against Heterologous Hookworm Infection.” *FASEB Journal* 23 (9): 3007–19. doi:10.1096/fj.09-131433.

Petersen, Thomas Nordahl, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2011. “SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions.” *Nature Methods* 8 (10): 785–86. doi:10.1038/nmeth.1701.

Rowe, J Alexandra, Antoine Claessens, Ruth A Corrigan, and Mònica Arman. 2009. “Adhesion of Plasmodium Falciparum-Infected Erythrocytes to Human Cells: Molecular Mechanisms and Therapeutic Implications.” *Expert Reviews in Molecular Medicine* 11 (May): e16.

doi:10.1017/S1462399409001082.

Sajid, M, and J H McKerrow. 2002. “Cysteine Proteases of Parasitic Organisms.” *Molecular and Biochemical Parasitology* 120 (1): 1–21. doi:10.1016/S0166-6851(02)00043-9.

Schicht, Sabine, Weihong Qi, Lucy Poveda, and Christina Strube. 2013. “The Predicted Secretome and Transmembranome of the Poultry Red Mite Dermanyssus Gallinae.” *Parasites & Vectors* 6 (1): 259. doi:10.1186/1756-3305-6-259.

Schirmer, R H, T Schöllhammer, G Eisenbrand, and R L Krauth-Siegel. 1987. “Oxidative Stress as a Defense Mechanism against Parasitic Infections.” *Free Radical Research Communications* 3 (1–5): 3–12.

Schwarz, Erich M, Pasi K Korhonen, Bronwyn E Campbell, Neil D Young, Aaron R Jex, Abdul Jabbar, Ross S Hall, et al. 2013. “The Genome and Developmental Transcriptome of the Strongylid Nematode Haemonchus Contortus.” *Genome Biology* 14 (8): R89. doi:10.1186/gb-2013-14-8-r89.

Shinya, Ryoji, Hironobu Morisaka, Taisei Kikuchi, Yuko Takeuchi, Mitsuyoshi Ueda, and Kazuyoshi Futai. 2013. “Secretome Analysis of the Pine Wood Nematode Bursaphelenchus Xylophilus Reveals the Tangled Roots of Parasitism and Its Potential for Molecular Mimicry.” *PloS One* 8 (6): e67377. doi:10.1371/journal.pone.0067377.

Sotillo, Javier, Mark Pearson, Jeremy Potriquet, Luke Becker, Darren Pickering, Jason Mulvenna, and Alex Loukas. 2016. “Extracellular Vesicles Secreted by Schistosoma Mansoni Contain Protein Vaccine Candidates.” *International Journal for Parasitology* 46 (1): 1–5. doi:10.1016/j.ijpara.2015.09.002.

Sotillo, Javier, Alejandro Sanchez-Flores, Cinzia Cantacessi, Yvonne Harcus, Darren Pickering, Tiffany Bouchery, Mali Camberis, et al. 2014. “Secreted Proteomes of Different Developmental Stages of the Gastrointestinal Nematode Nippostrongylus Brasiliensis.” *Molecular & Cellular Proteomics : MCP* 13 (10): 2736–51. doi:10.1074/mcp.M114.038950.

Spruance, S. L. 1974. “Latent Period of 53 Years in a Case of Hydatid Cyst Disease.” *Arch Intern Med* 134 (4): 741–42.

Staats, Charley Christian, Angela Junges, Rafael Lucas Muniz Guedes, Claudia Elizabeth Thompson, Guilherme Loss de Morais, Juliano Tomazzoni Boldo, Luiz Gonzaga Paula de Almeida, et al. 2014. “Comparative Genome Analysis of Entomopathogenic Fungi Reveals a Complex Set of Secreted Proteins.” *BMC Genomics* 15 (1): 822. doi:10.1186/1471-2164-15-822.

Suh, Jinkyo, and Harald Hutter. 2012. “A Survey of Putative Secreted and Transmembrane Proteins Encoded in the C. Elegans Genome.” *BMC Genomics* 13 (1): 333. doi:10.1186/1471-2164-13-333.

Tews, I, a Perrakis, a Oppenheim, Z Dauter, K S Wilson, and C E Vorgias. 1996. “Bacterial Chitobiase Structure Provides Insight into Catalytic Mechanism and the Basis of Tay-Sachs Disease.” *Nature Structural Biology*. doi:10.1038/nsb0196-95.

Thomas, Paul G, Michele R Carter, Olga Atochina, Akram A Da’Dara, Danuta Piskorska, Edward McGuire, and Donald A Harn. 2003. “Maturation of Dendritic Cell 2 Phenotype by a Helminth Glycan Uses a Toll-like Receptor 4-Dependent Mechanism.” *Journal of Immunology (Baltimore, Md. : 1950)* 171 (11): 5837–41. doi:10.4049/jimmunol.171.11.5837.

Tjalsma, H, A Bolhuis, J D Jongbloed, S Bron, and J M van Dijl. 2000. “Signal Peptide-Dependent

Protein Transport in Bacillus Subtilis: A Genome-Based Survey of the Secretome." *Microbiology and Molecular Biology Reviews : MMBR* 64 (3): 515–47. doi:10.1128/MMBR.64.3.515-547.2000.

Tsai, Isheng J, Magdalena Zarowiecki, Nancy Holroyd, Alejandro Garciarrubio, Alejandro Sanchez-Flores, Karen L Brooks, Alan Tracey, et al. 2013. "The Genomes of Four Tapeworm Species Reveal Adaptations to Parasitism." *Nature* 496 (7443): 57–63. doi:10.1038/nature12031.

Tudor JE, Pallaghy PK, Pennington MW, Norton RS. 1996. "Solution Structure of ShK Toxin, a Novel Potassium Channel Inhibitor from a Sea Anemone." *Nat Struct Biol* 3 (4): 317–20.

Van der Kleij, Desiree, Eicke Latz, Jos F H M Brouwers, Yvonne C M Kruize, Marion Schmitz, Evelyn A. Kurt-Jones, Terje Espevik, et al. 2002. "A Novel Host-Parasite Lipid Cross-Talk. Schistosomal Lyso-Phosphatidylserine Activates Toll-like Receptor 2 and Affects Immune Polarization." *Journal of Biological Chemistry* 277 (50): 48122–29. doi:10.1074/jbc.M206941200.

Wang, Zhengyuan, John Martin, Sahar Abubucker, Yong Yin, Robin B Gasser, and Makedonka Mitreva. 2009. "Systematic Analysis of Insertions and Deletions Specific to Nematode Proteins and Their Proposed Functional and Evolutionary Relevance." *BMC Evolutionary Biology* 9 (1): 23. doi:10.1186/1471-2148-9-23.

Wasmuth, James, Ralf Schmid, Ann Hedley, and Mark Blaxter. 2008. "On the Extent and Origins of Genic Novelty in the Phylum Nematoda." *PLoS Neglected Tropical Diseases* 2 (7): e258. doi:10.1371/journal.pntd.0000258.

Young, Neil D, Niranjan Nagarajan, Suling Joyce Lin, Pasi K Korhonen, Aaron R Jex, Ross S Hall, Helena Safavi-Hemami, et al. 2014. "The Opisthorchis Viverrini Genome Provides Insights into Life in the Bile Duct." *Nature Communications* 5: 4378. doi:10.1038/ncomms5378.

Zarowiecki, Magdalena, and Matt Berriman. 2015. "What Helminth Genomes Have Taught Us about Parasite Evolution." *Parasitology* 142 Suppl (S1): S85-97. doi:10.1017/S0031182014001449.

Zhu, Lihui, Juntao Liu, Jinwei Dao, Ke Lu, Hao Li, Huiming Gu, Jinming Liu, Xingang Feng, and Guofeng Cheng. 2016. "Molecular Characterization of S. Japonicum Exosome-like Vesicles Reveals Their Regulatory Roles in Parasite-Host Interactions." *Scientific Reports* 6 (May): 25885. doi:10.1038/srep25885.

3.2 – CAPÍTULO II: Análise evolutiva das cistatinas de *Schistosoma*

Cuesta-Astroz Y, Scholte LL, Pais FS, Oliveira G, Nahum LA. Evolutionary analysis of the cystatin family in three *Schistosoma* species. Front. Genet. 5:206. doi: 10.3389/fgene.2014.00206. PMID: 25071834.

Neste estudo foram identificados homólogos de cistatinas no proteoma predito de três espécies de *Schistosoma* e outros Platyhelminthes. Cistatinas são uma família de inibidores de cisteino proteases distribuídos em três subfamílias (I25A-C). Membros desta família sem atividade de cistatina são considerados "não classificados". Pouco se sabe sobre a evolução das cistatinas em *Schistosoma* bem como suas funções fisiológicas e padrões de expressão no ciclo de vida dos parasitos. Foi analisada a diversidade de sequências de aminoácidos focando na identificação de assinaturas proteicas para se estabelecer as relações evolutivas entre cistatinas de *Schistosoma* e humanas validadas experimentalmente. Padrões de expressão de genes foram obtidos a partir de diferentes estádios de desenvolvimento no *S. mansoni* utilizando dados de *microarray*. Em *Schistosoma,* foram identificadas I25A e I25B refletindo pouca diversificação funcional. I25C e os membros da subfamília de "não classificados" não foram identificados nas espécies de platelmintos analisadas. A filogenia obtida coloca as cistatinas em diferentes clados, refletindo a sua diversidade molecular. Nossos resultados sugerem que as cistatinas de *Schistosoma* são muito divergentes de seus homólogos humanos, especialmente em relação a subfamília I25B. Cistatinas de *Schistosoma* também são diferentes de seus homólogos em outros platelmintos. Finalmente, os dados transcritômicos indicaram que os genes de I25A e I25B são constitutivamente expressos e, portanto, poderiam ser essenciais para a progressão do ciclo de vida de *Schistosoma.* Em resumo, este estudo fornece *insights* sobre a evolução, classificação e diversificação funcional de cistatinas em *Schistosoma* e outros platelmintos, melhorando a compreensão da biologia do parasita e abrindo novas fronteiras na identificação de novos alvos terapêuticos contra helmintíases.

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 09 July 2014
doi: 10.3389/fgene.2014.00206

Evolutionary analysis of the cystatin family in three *Schistosoma* species

Yesid Cuesta-Astroz ^{1,2†}***, Larissa L. S. Scholte*** ^{1,2†}***, Fabiano Sviatopolk-Mirsky Pais*** ^{1,3}***,***
Guilherme Oliveira ^{1} ***and Laila A. Nahum*** ^{1,3}*******

[1] Grupo de Genômica e Biologia Computacional, Centro de Excelência em Bioinformática, Instituto Nacional de Ciência e Tecnologia em Doenças Tropicais, Centro de Pesquisas René Rachou (CPqRR), Fundação Oswaldo Cruz (Fiocruz), Belo Horizonte, Brazil
[2] Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[3] Faculdade Infórium de Tecnologia, Belo Horizonte, Brazil

Edited by:
Arnon Dias Jurberg, Fundação Oswaldo Cruz (Fiocruz), Brazil

Reviewed by:
Tim Littlewood, Natural History Museum, UK
Gustavo Coutinho Cerqueira, Broad Institute, USA

****Correspondence:***
Laila A. Nahum, Centro de Pesquisas René Rachou, Avenida Augusto de Lima, 1715, Barro Preto., Belo Horizonte, MG 30190-002, Brazil
e-mail: laila@nahum.com.br

†These authors have contributed equally to this work.

The cystatin family comprises cysteine protease inhibitors distributed in 3 subfamilies (I25A–C). Family members lacking cystatin activity are currently unclassified. Little is known about the evolution of *Schistosoma* cystatins, their physiological roles, and expression patterns in the parasite life cycle. The present study aimed to identify cystatin homologs in the predicted proteome of three *Schistosoma* species and other Platyhelminthes. We analyzed the amino acid sequence diversity focused in the identification of protein signatures and to establish evolutionary relationships among *Schistosoma* and experimentally validated human cystatins. Gene expression patterns were obtained from different developmental stages in *Schistosoma mansoni* using microarray data. In *Schistosoma*, only I25A and I25B proteins were identified, reflecting little functional diversification. I25C and unclassified subfamily members were not identified in platyhelminth species here analyzed. The resulting phylogeny placed cystatins in different clades, reflecting their molecular diversity. Our findings suggest that *Schistosoma* cystatins are very divergent from their human homologs, especially regarding the I25B subfamily. *Schistosoma* cystatins also differ significantly from other platyhelminth homologs. Finally, transcriptome data publicly available indicated that I25A and I25B genes are constitutively expressed thus could be essential for schistosome life cycle progression. In summary, this study provides insights into the evolution, classification, and functional diversification of cystatins in *Schistosoma* and other Platyhelminthes, improving our understanding of parasite biology and opening new frontiers in the identification of novel therapeutic targets against helminthiases.

Keywords: schistosomiasis, proteinase inhibitor, phylogenomics, bayesian inference, function prediction

INTRODUCTION

Five species of the genus *Schistosoma* (Trematoda) are involved in the human infection, being the main etiologic agents of human schistosomiasis: *Schistosoma mansoni* and *Schistosoma japonicum* causing intestinal schistosomiasis, and *Schistosoma haematobium* causing urinary schistosomiasis. According to the World Health Organization, schistosomiasis is endemic in 77 countries, affects more than 200 million people worldwide, and other 779 million live in areas at risk of infection (WHO, 2012). Schistosomiasis control relies mainly on praziquantel treatment but its efficacy is limited. Furthermore, evidence of praziquantel resistant parasites was obtained in the laboratory and in endemic regions (Liang et al., 2003; Melman et al., 2009; Coeli et al., 2013). Hence schistosomiasis is still one of the most prevalent infectious and parasitic diseases worldwide being a major source of morbidity and mortality in developing countries.

The urgent need to develop novel drugs or a vaccine for *Schistosoma* species has encouraged an interest in the function prediction of relevant proteins for parasitism. The search for new drug targets based on evolutionary analyses using *S. mansoni* genomic/proteomic data has been performed (Silva et al., 2011, 2012). Such studies have improved the *S. mansoni* functional annotation, allowed for a deeper understanding of the genomic complexity and lineage-specific adaptations potentially related to the parasitic lifestyle, and pointed out several proteins as potential drug targets, including proteases.

Cysteine proteases, one of the four major classes of proteolytic enzymes, have been found in a wide range of taxonomic groups, from viruses to vertebrates. These peptidases are involved in many biological processes, such as catabolism, antigen processing, inflammation, dystrophy, and metastasis (Henskens et al., 1996). Protease inhibitors, such as cystatins, inhibit the enzymatic activity of cysteine proteases. Cystatins comprise a family of cysteine protease inhibitors identified in diverse taxonomic groups, including Platyhelminthes and Nematoda (Kordis and Turk, 2009). In humans, cystatins have evolved widely not only to regulate enzymes in pathways but also as a defense mechanism against proteases of invading pathogens (Toh et al., 2010).

In parasites, cystatins participate in normal physiological processes, but are also important pathogenicity factors, being directly involved in host-parasite interactions (Hartmann et al., 1997; Manoury et al., 2001; Schierack et al., 2003; Harnett, 2014).

Based on sequence similarity, the presence or lack of disulfide bonds, and physiological localization, cystatins were first classified in three families: family 1 (e.g., stefins), family 2 (e.g., cystatins), and family 3 (e.g., kininogens) (Barrett, 1986). Afterwards, in terms of number of cystatin domains and the presence of sequence features these proteins were classified into type 1, 2, and 3 (Rawlings and Barrett, 1990). In the present study we adopted the classification proposed by MEROPS database a resource for peptidases and protein inhibitors (Rawlings et al., 2014). The database uses a hierarchical structure-based classification in which each peptidase and inhibitor amino acid sequences are grouped into families based on statistically significant similarities. MEROPS classifies cystatin proteins as members of the I25 family, further subdivided into four subfamilies: I25A, I25B, I25C, and unclassified (**Figure 1**). This classification system is based on similarities between protein sequences and three dimensional structures. According to MEROPS classification, proteins containing a single inhibitor unit are termed simple inhibitor, and those containing multiple inhibitor units are termed as a compound inhibitor (Rawlings et al., 2014). However, several proteins containing cystatin domains cannot be easily included in a classification scheme, resulting in a number of cystatin family members that remain without classification in the subfamily level (Cornwall et al., 2003; Kordis and Turk, 2009; Siricoon et al., 2012).

One of the first cystatin proteins described in parasitic organisms was the onchocystatin (I25B), a highly antigenic protein encoded by the nematode *Onchocerca volvulus* (Lustigman et al., 1991, 1992). Onchocystatin was initially proposed to be involved in parasite protease regulation during the molting process in Nematoda. Afterwards, it was shown that this protein is also involved in modulation of host immune responses (Hartmann et al., 1997). The molecular interactions of parasite cystatins and host molecules have not yet been clearly determined, but it is believed that the mechanisms are similar to those demonstrated for other species (Klotz et al., 2011). Some examples of known host parasite interactions were previously described in nematodes in which I25B secreted cystatins inhibit host cathepsins such as B and H by *Haemonchus contortus* (Newlands et al., 2001), L and S by *Acanthocheilonema viteae* (Vray et al., 2002), and B and L by *Nippostrongylus brasiliensis* (Dainichi et al., 2001).

Although cystatin family members have been the subject of many studies in different organisms, little is known regarding functional diversification and evolution in *Schistosoma*. In this context, the information available for human homologs can be used in comparative studies, at sequence and structure level, in order to understand the interactions of *Schistosoma* cystatins and host cysteine proteases. The present study aimed to identify cystatin homologs on predicted proteomes of three *Schistosoma* species and other Platyhelminthes in order to have a landscape

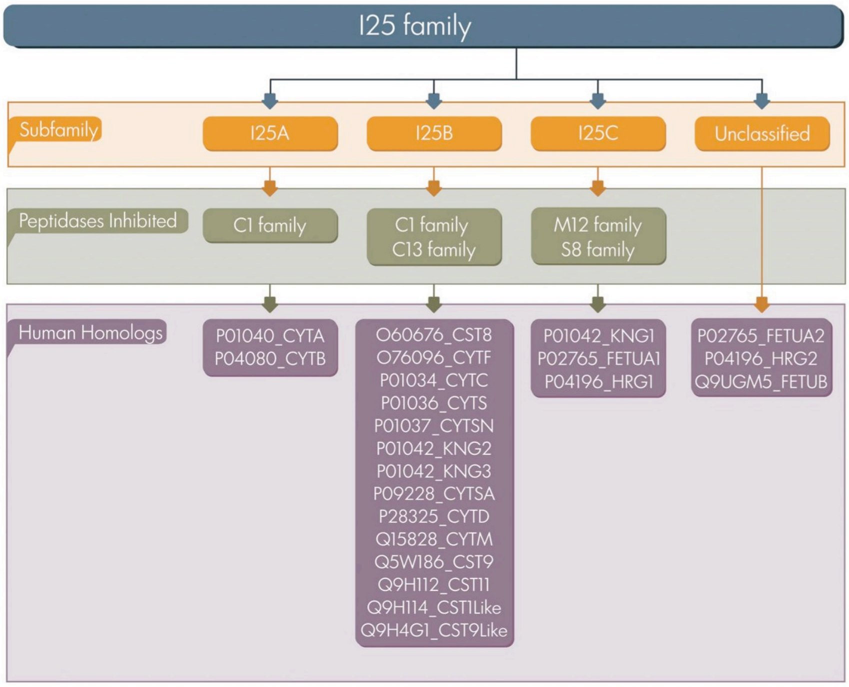

FIGURE 1 | Cystatin classification. The I25 family is classified in three subfamilies (I25A–C) according to MEROPS. Several family members remain unclassified. UniProt accession numbers of human homologs are listed.

view of the functional diversification in this phylum. In addition, evolutionary analyses were reconstructed for *Schistosoma* and human homologs based on the information at the sequence level, signatures, and phylogenetic relationships. Additionally, we evaluated cystatins' expression in different stages of the parasite life cycle in order to answer the following questions: How many cystatin homologs are present in *Schistosoma* species and in other Platyhelminthes? Do potential homologs have characteristic sequence features? What are the evolutionary relationships of the cystatin family members in *Schistosoma* species and their human homologs? Is the transcription of cystatin members during the *S. mansoni* life cycle stage-specific or is it conserved through the stages assessed?

In summary, we used predicted proteome data currently available for three *Schistosoma* species (Berriman et al., 2009; Zhou et al., 2009; Young et al., 2012), three Cestoda (Tsai et al., 2013), and the free living *Schmidtea mediterranea* (unpublished data) to identify potential cystatin homologs encoded by Platyhelminthes. Using combined computational approaches, we identified proteins belonging to the I25 family and reported members classified in two subfamilies: I25A and I25B. We also assessed microarray public datasets to investigate gene expression in different stages of the *Schistosoma mansoni* life cycle. This study provides insights into the evolution and potential functional diversification of Platyhelminthes cystatins improving our understanding of parasite biology and opening new frontiers in the identification of novel therapeutic targets against helminthiases.

MATERIALS AND METHODS

ORGANISMS AND SEQUENCE DATA

The dataset of selected species comprises three *Schistosoma* species: *S. haematobium* (NCBI taxid: 6185), *S. japonicum* (6182), and *S. mansoni* (6183); four other Platyhelminthes: *Echinococcus granulosus* (6210), *Hymenolepis microstoma* (85433), *Schmidtea mediterranea* (412041), and *Taenia solium* (6204); and *Homo sapiens* (9606). *Schistosoma* predicted proteomes were downloaded from SchistoDB 3.0 (beta.schistodb.net) (Zerlotini et al., 2013). Cestoda proteome data was obtained from the Sanger Institute FTP site (ftp.sanger.ac.uk/pub/pathogens). *S. mediterranea* proteome data was kindly provided by Dr. Eric Ross from Stowers Institute for Medical Research (USA). Predicted proteomes from each genome project were used in order to obtain evidence of protein gain or loss and a more accurate identification of cystatin homologs. *H. sapiens* I25 family members were retrieved from the Human Protein Reference Database (www.hprd.org) (Keshava Prasad et al., 2009). Functional information regarding the cystatin family is available on the MEROPS peptidase database (Rawlings et al., 2014) via the I25 inhibitor family identifier.

HOMOLOGS IDENTIFICATION

Potential cysteine protease inhibitors encoded by platyhelminth genomes were identified by using the hmmscan software included in the HMMER 3.0 package (Eddy, 2011). Each proteome was compared against Pfam-A HMM profiles, which were retrieved from the Pfam database (Finn et al., 2014). Such analyses were performed in order to identify the presence and architecture of proteins comprising the cystatin domain (Pfam: PF00031). The significance of the Pfam-A match is based on the resulting score. A match is considered significant when the score is greater than or equal to the gathering threshold for the Pfam domain. To date, the current threshold for the cystatin domain (Pfam: PF00031) is 20.9. Proteins containing significant or insignificant matches with the target domain were selected. Insignificant matches although less informative than significant ones, can be used for identifying functionally conserved regions when no significant matches are found. For this reason, insignificant matches were also initially selected in this work. In addition, information on accessory domains and protein signatures (Q-x-V-x-G motif, PW motif, LP motif, SND/SNS/TND motifs, and disulfide bonds) were considered to define potential I25 homologs. The presence of signal peptide in potential cysteine protease inhibitors was predicted by SignalP 4.1 using the neural network method with default D-cutoff values and using eukaryotes as "organism group" (Petersen et al., 2011). The illustrations of protein domain architectures were generated using DOG 2.0 (Ren et al., 2009).

PHYLOGENETIC ANALYSIS

Aiming at establishing evolutionary relationships among *Schistosoma* and experimentally validated human cystatins, I25A and I25B amino acid domain sequences from *S. haematobium*, *S. japonicum*, *S. mansoni*, and *H. sapiens* were selected for phylogenetic reconstruction. The evolutionary relationships between *Schistosoma* and human cystatins may provide cues about functions performed by parasites' orthologs. Human PF00031 domains classified into I25C subfamily or inhibitor units not assigned to a subfamily were not included in this analysis once they have no cysteine protease inhibitor activity. To optimize the dataset for phylogenetic analysis we removed redundancy and sequences too distantly related using the Decrease Redundancy tool, available as a resource at ExPaSy (www.expasy.org). The Decrease Redundancy parameters were set as 98 for "% max similarity" and 30 for "% min similarity." The filtered set of amino acid sequences, corresponding to the conserved domain (PF00031) were aligned using MAFFT 7 with iterative refinement by the G-INS-i strategy (Katoh et al., 2009). The multiple sequence alignment comprising 22 sequences and 96 sites was manually refined using Jalview (Waterhouse et al., 2009) and further used in phylogenetic analysis. To reconstruct the phylogenetic tree we used MrBayes 3.2.1, which performs Bayesian inference using a variant of the Markov Chain Monte Carlo (Ronquist and Huelsenbeck, 2003). MCMC analyses were run as four chains, one cold and three heated chains, for 10,000,000 generations and sampled every 100 generations. Twenty-five percentage of the initial samples were discarded as "burn-in." Mixed models were applied as a parameter to estimate the best-fit evolutionary model. Support values were estimated as Bayesian posterior probabilities. The evolutionary history of *Schistosoma* and human cystatins was also reconstructed based on the maximum likelihood method (ML), as implemented in PhyML (Guindon et al., 2010). For the phylogenetic reconstruction we tested 12 different evolutionary models (JTT, LG, DCMut, MtREV, MtMam, MtArt, Dayhoff, WAG, RtREV, CpREV, Blosum62, and VT) using the ProtTest 2.4 software (Abascal et al., 2005). The evolutionary model best fitting the data (best fit model) was determined by comparing the likelihood of the tested models according to the Akaike Information Criterion. Trees were visualized and edited using the FigTree software (tree.bio.ed.ac.uk/software/figtree).

TRANSCRIPTIONAL PROFILES

Data from 35,437 oligonucleotide microarray probes from *S. mansoni* transcriptomic analyses (Fitzpatrick et al., 2009) were interrogated in order to identify the transcription patterns of two cystatin family members: Smp_006390 (I25A) and Smp_034420.2 (I25B). Thirteen development stages were covered and the complete set of raw and normalized data were downloaded from ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) under the experiment accession number E-MEXP-2094. For differential expression analysis, mean fluorescence normalized values were linear model fitted using three replicates per stage and a total of 19 evolutionary pairwise comparisons were made (see Fitzpatrick et al., 2009 for details). Additionally, recently published RNAseq transcription data (Protasio et al., 2012) was also interrogated for gene expression pattern and gene model evaluation. In this case, four developmental stages of *S. mansoni* were covered. Raw sequence datasets (three from cercariae stage, two from 3 h post-infection mechanically transformed schistosomula, two from 24 h post-infection schistosomulas, and one from adult worms), were downloaded from ArrayExpress under the accession number E-MTAB-451. The RNAseq reads were stored in a local server and aligned to the most recent version of the *S. mansoni* genome (v.5). Reads were mapped with Tophat-v.2.0.8 (Trapnell et al., 2012) and transcripts were assembled with Cufflinks-v.2.0.2 (Trapnell et al., 2012). Cuffdiff, a program from the Cufflinks suite, was used to estimate expression of transcripts across samples. CummeRbund, an R package, and Integrative Genomics Viewer -IGV (Thorvaldsdóttir et al., 2013) were used to visualize results.

RESULTS

In this study we have mined platyhelminth proteomes in order to identify proteins belonging to the I25 family and its respective subfamilies. To this end, we used intrinsic methods at sequence level followed by multiple sequence alignment and phylogenetic analysis. Such analyses generated an evolutionary view of potential cystatin proteins in three *Schistosoma* species. We also analyzed the amino acid sequence diversity focused on the identification of protein signatures. Finally, we verified the transcriptional profiles of cystatins. Overall, a framework for functional analysis of parasite cystatins is provided. In summary, our findings contribute to a better understanding of host-parasite interactions and pathogenesis, once analysis and cystatins appear as relevant molecules in these processes.

IDENTIFICATION OF CYSTATIN FAMILY MEMBERS

Cystatin family (I25) members were identified using an intrinsic method. Platyhelminth proteomes were scanned by hmmscan (Eddy, 2011) and potential homologs were retrieved based on the presence of significant or insignificant matches with the conserved cystatin domain (PF00031) (**Table 1**). In cases where insignificant PF00031 matches were recovered, we also searched for critical residues that mediate protease inhibition to define the query protein as a potential cysteine protease inhibitor (**Table 2**). Based on “start” and “end” alignment positions of potential homologs identified overlapping the PF00031 HMM profile, truncated regions were assigned. It is important to emphasize that the Pfam database (Finn et al., 2014) is built from the most recent UniProt (UniProt Consortium, 2014) release and that no single protein database covers all diversity existing in nature. More specifically, the total of platyhelminth cystatins available at UniProt is underrepresented when compared, for instance, to mammals. Thus, it is possible that the presence of divergent regions reflects their degree of divergence to other proteins available at the database. On the other hand, it is also important to consider that the difference between the PF00031 HMM profile and the query sequences can be related to the presence of pseudogenes or errors in the gene models.

Considering alternative splicing products (Smp_034420.1, Smp_034420.2, and Smp_034420.3; Sha_109477 and Sha_109478), we identified in *Schistosoma* species ten proteins that contain the conserved domain (**Table 1**). Three proteins were retrieved in *S. haematobium*, three in *S. japonicum*, and four in *S. mansoni*. These single domain proteins vary in length and in domain size. In order to classify the identified homologs in subfamilies (I25A–C or unclassified), we searched for the presence of signal peptide and other evolutionarily conserved residues (**Table 2**), which are involved in the formation of a wedge-like structure that is complementary to the active site of target proteases. Features as Q-x-V-x-G, LP, and PW motifs which are considered essential for binding and inhibiting cysteine proteases activity were identified. To remove potentially redundant sequences as well as too distantly related proteins we filtered alternative splicing products and run the Decrease Redundancy program using the previously mentioned parameters. In total, four sequences were filtered out: Sha_109478, Sjc_0094540, Smp_034420.1, and Smp_034420.3.

Concerning other Platyhelminthes species, we identified 14 cystatin proteins encoded by three Cestoda (*E. granulosus*, *H. microstoma*, and *T. solium*) and a free living Turbellaria (*S. mediterranea*) (**Table 1**). Contrary to *Schistosoma* species, hmmscan searches retrieved additional domains in some of those homologs. Such domains were retrieved as insignificant matches and a few showed overlapping regions with the conserved domain (PF00031) (**Table 1**). The information of additional domains may suggest potential lineage-specific innovations that happened in cystatin family members over evolutionary time. On the other hand, it can reflect the caveat of data quality in organisms for which we have only draft genomes.

We also analyzed cystatin diversity at the sequence level in terms of critical motifs, amino acid conservation, or variants that could lead to differences in the inhibitory capability of the cystatins (**Table 2** and **Figure 3**). The alignment of identified I25A and I25B cystatin sequences point out four conserved regions: a Glycine residue within the N-terminal region, a Q-x-V-x-G motif in one hairpin loop, and a PW or LP motifs in the second loop (**Figure 3**). Those regions can dock with the substrate-binding site of family C1 of cysteine proteases (Dickinson, 2002). One disulfide bridge exclusively present in I25B proteins was also identified.

Table 1 | Cystatin predictions across *Schistosoma* and other Platyhelminthes.

Taxon	TaxID	Accession	Length	Domain	Start	End	*E*-value	Score	Significant
Schistosoma haematobium	6185	Sha_109477	160	PF00031	83	142	2.2×10^{-1}	11.6	No
		Sha_109478	145	PF00031	38	127	4.5×10^{-2}	13.8	No
		Sha_300402	101	PF00031	35	91	1×10^{-6}	28.7	Yes
Schistosoma japonicum	6182	Sjc_0005780	145	PF00031	37	126	1.2×10^{-8}	35	Yes
		Sjc_0066340	101	PF00031	40	88	2.7×10^{-7}	30.6	Yes
		Sjc_0094540	123	PF00031	40	85	1.8×10^{-4}	21.5	Yes
Schistosoma mansoni	6183	Smp_006390.1	101	PF00031	35	91	1×10^{-5}	25.5	Yes
		Smp_034420.1	117	PF00031	38	98	8.3×10^{-7}	29	Yes
		Smp_034420.2	148	PF00031	38	129	7.9×10^{-8}	32.3	Yes
		Smp_034420.3	145	PF00031	38	126	1.5×10^{-2}	15.3	No
Echinococcus granulosus	6210	EgrG_000159200.1	98	PF00031	16	77	4.5×10^{-5}	23.5	Yes
		EgrG_000159200.1	98	PF03672	29	51	2.2×10^{-3}	17.5	No
		EgrG_000543900.1	111	PF00031	34	79	5.1×10^{-3}	16.9	No
		EgrG_000849600.1	274	PF00031	45	98	2.3×10^{-10}	40.4	Yes
		EgrG_000849600.1	274	PF00031	167	224	4.8×10^{-2}	13.7	No
		EgrG_000849600.1	274	PF13549	37	104	1.1×10^{-1}	11.7	No
Hymenolepis microstoma	85433	HmN_000582300.1	180	PF00031	41	83	2.4×10^{-2}	14.7	No
		HmN_000582400.1	107	PF00031	26	78	1.5×10^{-2}	15.4	No
		HmN_000582400.1	107	PF06050	18	51	4.4×10^{-2}	12.5	No
		HmN_000842000.1	295	PF00031	63	115	3.2×10^{-4}	20.7	No
Taenia solium	6204	TsM_000671000	274	PF00031	45	98	5.2×10^{-10}	39.3	Yes
		TsM_000671000	274	PF00031	175	224	1.7×10^{-2}	15.2	No
		TsM_000687900	98	PF00031	9	76	2.6×10^{-4}	21	Yes
		TsM_000687900	98	PF03672	29	51	2.2×10^{-2}	14.3	No
		TsM_000687900	98	PF13805	8	73	7.1×10^{-2}	12.2	No
		TsM_001154200	115	PF00031	37	84	4×10^{-3}	17.2	No
		TsM_001154200	115	PF14073	37	85	1.4×10^{-1}	11.9	No
		TsM_001154200	115	PF15606	39	87	3×10^{-2}	14.2	No
		TsM_001154300	137	PF00031	33	95	1.2×10^{-4}	22.1	Yes
Schmidtea mediterranea	412041	mk4.000249.00.01	93	PF00031	4	51	8.6×10^{-2}	12.9	No
		mk4.000249.04.01	93	PF00031	4	51	9.6×10^{-2}	12.8	No
		mk4.004385.02.01	119	PF00031	33	104	1.3×10^{-13}	50.8	Yes
		mk4.027397.00.01	176	PF00031	33	121	6.1×10^{-14}	51.9	Yes

TaxID: identifier at NCBI taxonomy database. Accession: accession number in the source genome project database. Length: number of amino acids. Domain: domain prediction based on HMM models from the Pfam database using the HMMscan tool. Start and End: alignment positions of the retrieved domain in the query sequence. Significant (Yes/No): statistically significant or insignificant score values according to the gathering threshold for each Pfam domain.

I25A subfamily members are predominantly intracellular single-domain proteins of about 11 kDa and ~100 amino acid residues, which do not contain disulfide bridges and the PW motif. I25A inhibitors have three evolutionarily highly conserved regions: a glycine residue within the N-terminal region, a central Q-x-V-x-G motif, and a C-terminal LP pair (Klotz et al., 2011).

Figure 3 shows five *Schistosoma* and human homologs that have these conserved features, being therefore classified into the I25A subfamily. Following the same pattern of conserved features, when analyzing the proteome data of others Platyhelminthes (**Table 2**), we identified potential I25A subfamily member in *E. granulosus* (EgrG_000159200.1) and in *T. solium* (TsM_000687900). In *H. microstoma* three potential cystatin proteins without signal peptide were identified, something uncommon in other platyhelminth predictions. Therefore, this result should be further evaluated carefully, before being considered an evolutionary innovation. In *S. mediterranea* we identified two proteins belonging to the I25A subfamily. However the identified cystatins mk4.000249.04.01 and mk4.000249.00.01 are identical. For this reason we took into account the probability of redundancy and considered one of them (**Table 2**) as a cystatin homolog.

I25B inhibitors are secreted single-domain proteins around 14 kDa, ~120 residues long with at least one disulfide bridge and

Table 2 | Sequence features of I25 family members in selected taxa.

TaxID	Accession	SP	SND/SNS/TND	Q-x-V-x-G	S-S	LP	PW
6185	Sha_109477	Yes	N/A	Yes	Yes	No	Yes
	Sha_109478	Yes	N/A	No	Yes	No	Yes
	Sha_300402	No	N/A	Yes	No	Yes	No
6182	Sjc_0005780	Yes	N/A	Yes	Yes	No	Yes
	Sjc_0066340	No	N/A	Yes	No	Yes	No
	Sjc_0094540	No	N/A	Yes	Yes	No	No
6183	Smp_006390.1	No	N/A	Yes	No	Yes	No
	Smp_034420.1	Yes	N/A	Yes	Yes	No	No
	Smp_034420.2	Yes	N/A	Yes	Yes	No	Yes
	Smp_034420.3	Yes	N/A	No	Yes	No	Yes
6210	EgrG_000159200.1	No	N/A	Yes	Yes	Yes	No
	EgrG_000543900.1	Yes	N/A	Yes	No	No	No
	EgrG_000849600.1_x	Yes	N/A	Yes	No	No	Yes
	EgrG_000849600.1_y	No	N/A	No	No	No	No
85433	HmN_000582300.1	No	N/A	Yes	No	No	No
	HmN_000582400.1	No	N/A	Yes	No	No	No
	HmN_000842000.1	No	N/A	Yes	No	No	No
6204	TsM_000671000_x	Yes	N/A	Yes	No	No	Yes
	TsM_000671000_y	No	N/A	No	No	No	No
	TsM_000687900	No	N/A	Yes	No	Yes	No
	TsM_001154200	Yes	N/A	No	No	No	No
	TsM_001154300	Yes	N/A	Yes	No	No	No
412041	mk4.000249.00.01	No	N/A	Yes	No	Yes	No
	mk4.000249.04.01	No	N/A	Yes	No	Yes	No
	mk4.004385.02.01	Yes	N/A	Yes	Yes	No	No
	mk4.027397.00.01	Yes	N/A	Yes	Yes	No	Yes
9606	O60676_CST8	Yes	No	No	Yes	No	Yes
	O76096_CYTF	Yes	Yes	Yes	Yes	No	Yes
	P01034_CYTC	Yes	Yes	Yes	Yes	No	Yes
	P01036_CYTS	Yes	No	No	Yes	No	Yes
	P01037_CYTSN	Yes	No	Yes	Yes	No	Yes
	P01040_CYTA	No	No	Yes	No	Yes	No
	P01042_KNG1	Yes	No	No	Yes	No	No
	P01042_KNG2	N/A	No	Yes	Yes	No	No
	P01042_KNG3	N/A	No	Yes	Yes	No	Yes
	P04080_CYTB	No	No	Yes	No	Yes	No
	P09228_CYTSA	Yes	No	Yes	Yes	No	Yes
	P28325_CYTD	Yes	No	Yes	Yes	No	Yes
	Q15828_CYTM	Yes	Yes	Yes	Yes	No	Yes
	Q5W186_CST9	Yes	No	No	Yes	No	No
	Q9H112_CST11	Yes	No	No	Yes	No	Yes
	Q9H114_CST1Like	Yes	Yes	No	Yes	No	Yes
	Q9H4G1_CST9Like	Yes	No	No	Yes	No	Yes

TaxID: identifier at NCBI taxonomy database. 9606: Homo sapiens. Accession: accession number in the source genome project database or in UniProt (human proteins). Sequence features include a signal peptide (SP), disulfide bridge (S-S), and distinct motifs (SND/SNS/TND, Q-x-V-x-G, LP, and PW), in which amino acids are indicated by the one-letter code. Yes or No: presence or absence of conserved features. N/A: Not applicable.

a signal peptide. I25B inhibitors have two of the three conserved regions previously mentioned: the N-terminal Gly residue and a central Q-x-V-x-G motif. Instead of a C-terminal LP pair, I25B inhibitors have a PW motif at the C-terminal segment. Besides, some I25B members also possess a distinct conserved SND, SNS or TND motifs between the first conserved glycine and the central Q-x-V-x-G motif (**Table 2**). The presence of these additional motifs allow cystatin proteins to inhibit either legumain or asparaginyl endopeptidases (Alvarez-Fernandez et al., 1999; Zavasnik-Bergant, 2008; Klotz et al., 2011; Schwarz et al., 2012). In parasites, I25B subfamily members were demonstrated to be involved in modulation of host immune responses (Khaznadji et al., 2005; Gregory and Maizels, 2008). **Figure 3** shows I25B *Schistosoma* and human sequences identified according the sequence features previously mentioned. In three Platyhelminthes species (*E. granulosus*, *H. microstoma*, *T. solium*) it was not possible to identify I25B homologs (**Table 2**). In *S. mediterranea* we identified two similar sequences. However, it seems like that mk4.004385.02.01 is a fragment of mk4.027397.00.01. In this case we have chosen the mk4.027397.00.01 protein as potentially true I25B homolog due to the presence of expected sequence features (**Table 2**).

I25C subfamily members act mostly on serine proteases classified into the family S8 (Cornwall et al., 2003) and metalloproteases from the family M12 (Valente et al., 2001). Fetuins and histidine rich proteins are also multi-domain secreted proteins, but lack cystatin activity and are called as unclassified (Rawlings et al., 2014). Kininogens and fetuins are much younger than I25A and I25B proteins and their occurrence are restricted to vertebrates (Kordis and Turk, 2009). According to our findings, the I25C and unclassified subfamily members are not encoded by the genomes of platyhelminth species here analyzed.

Human cystatin protein subunits previously characterized as protease inhibitors were classified as belonging to one of the three cystatin subfamilies 125A–C (**Table 2**). In total, 15 proteins were retrieved from the Human Protein Reference Database (Keshava Prasad et al., 2009) and evolutionarily conserved residues were identified (**Table 2** and **Figure 3**). As already described in the literature, one additional motif was found (SND/SNS/TND), which is related to legumain inhibition. A multidomain protein P01042 was retrieved for each separate subunit denoted as KNG1, KNG2, and KNG3 (**Table 2**).

EVOLUTIONARY RELATIONSHIPS AMONG *SCHISTOSOMA* AND *H. SAPIENS* CYSTATINS

The evolutionary relationships in cystatins were reconstructed from an alignment containing 22 sequences corresponding to the conserved domain (PF00031) and 96 sites (**Figure 3**) using maximum likelihood and Bayesian inference. Both methods retrieved the same tree topology. Statistical support was also calculated for each node by both phylogenetic inference methods (Bayesian inference/maximum likelihood). Protein sequences are represented on the phylogenetic tree by UniProt (UniProt Consortium, 2014) and SchistoDB (Zerlotini et al., 2013) identifiers. Based on the phylogeny we were able to identify two well supported monophyletic subfamilies (100/100): I25A and I25B (**Figure 4**). I25C and unclassified homologs were not identified in *Schistosoma* species. Two domains of the kininogen protein P01042 (KNG2 and KNG3) were placed in I25B clade due to its cysteine protease inhibitory activity. The KNG1 domain from the same protein belongs to the I25C subfamily and was not included in this analysis because it has lost its inhibitory activity due to mutations in structurally important regions (Kordis and Turk, 2009) and may act as a calcium transporter (Higashiyama et al., 1987).

According to the phylogeny (**Figure 4**), three *Schistosoma* proteins (Sha_300402, Sjp_0066340, and Smp_006390.1) and two human cystatin proteins (P04080_CYTB, P01040_CYTA) were grouped into the I25A subfamily clade. These proteins share the evolutionarily conserved residues: a N-terminal Gly, a central Q-x-V-x-G, and a C-terminal LP (**Table 2** and **Figure 3**). Based on the information available on the literature and in protein databases, both human homologs (P04080_CYTB, P01040_CYTA) were experimentally and structurally characterized. On the other hand, only two I25A *Schistosoma* proteins (Smp_006390.1 and Sjp_0066340) were experimentally characterized at the protein level (Morales et al., 2004; He et al., 2011). Those proteins are involved in a intracellular modulator role. For instance, Smp_006390.1 was able to inhibit the formation of hemozoin by live schistosomula, suggesting a possible role in the gut of the schistosomula (Morales et al., 2004).

In I25B subfamily clade three *Schistosoma* homologs were grouped with 13 human proteins (14 cystatin domains). The conserved features (signal peptide, Gly residue, Q-x-V-x-G motif, disulfide bridge, and PW motif) were not detected in all protein domain sequences (**Table 2** and **Figure 3**). Contrary to *Schistosoma* I25A subfamily members, I25B cystatins were not experimentally characterized. Human sequences placed on the I25B subfamily clade reveal a significant expansion of such subfamily in *H. sapiens*. The phylogenetic analysis shows that these homologs originated from successive post speciation gene duplication events. Most human I25B members present the typical protein signatures (**Table 2** and **Figure 3**).

In summary, *Schistosoma* cystatin clades of I25A and I25B subfamily members were well supported by both branch support methods showing 100/100 (aLRT and posterior probability). *S. haematobium* and *S. mansoni* cystatins are closest related to each other when compare to the *S. japonicum* homolog. The phylogenetic analysis showed that human and *Schistosoma* cystatins are placed in different clades, reflecting its diversity at molecular level.

CYSTATIN EXPRESSION IN *SCHISTOSOMA MANSONI*

According to the microarray data analysis (Fitzpatrick et al., 2009), both I25A and I25B members, Smp_006390.1 and Smp_034420.2 respectively, were constitutively expressed in several stages of the *S. mansoni* life cycle (**Figure 5**). These findings corroborate the pairwise comparisons of key developmental stages performed by Fitzpatrick et al. (2009) that did not indicate any differential expression of both cystatin mRNAs. Comparisons were made with an adjusted p value (Adj $p < 0.05$), corrected using the Benjamini and Hochberg method for multiple testing, for which 3 replicates per life cycle were assessed. Both cystatins expression were also confirmed in the RNA sequencing work of Protasio et al. (2012). Transcripts were expressed constitutively in all four stages evaluated (data not shown). No differential expression assessment was considered in this case, mostly because those experiments did not include adult worm sample replicates. Therefore, without replicates, it is impossible to estimate sample variability.

As RNAseq data can be used to improve gene model annotations, current gene models for both I25A and I25B members were investigated. We analyzed the read coverage by mapping the reads against the reference genome. It was performed by Tophat, a tool that allows read alignments containing gaps in regions spanning introns. Therefore, predicted gene models for both cystatin family members were confirmed by visual inspection using IGV (Thorvaldsdóttir et al., 2013). All three exons of I25A and four exons of I25B, as well as both 5′ and 3′ UTR regions, located in the *S. mansoni* SuperContig 0138 and chromosome 2 respectively, had several reads correctly mapped within the exon limits (data not shown).

DISCUSSION

CONSERVED SEQUENCE FEATURES IN CYSTATINS

The I25A cystatin subfamily is a predominantly intracellular protein and does not present disulfide bridges. The inhibitor domain Q-x-V-x-G is present in the first hairpin loop and contains the glycine residue in the first position of the cystatin domain. This amino acid is also conserved in other cystatins of the I25B subfamily (**Figure 3**). This glycine residue allows the N-terminal region to interact with the sub-sites S4, S3, and S2 of the cysteine proteases. In *Schistosoma* species, besides its role in degradation of hemoglobin, I25A can also act intracellularly as a general regulator of protease activity (Morales et al., 2004). The constant and ubiquitous expression in *S. mansoni*, as shown by transcriptomic analysis, supports this idea. In the I25A subfamily we identified the highly conserved LP motif in positions 73–74 (**Figure 3**). The LP motif is essential for high affinity binding to papain (Pol and Bjork, 1999). The conserved motif differences between I25A and I25B subfamilies could reflect differences in the inhibitory spectrum of these proteins during evolution of function (Dickinson, 2002).

Kordis and Turk (2009) postulated that the progenitor of this family was most probably intracellular, lacked a signal peptide and disulfide bridges. The hypothesis is that throughout cystatins evolution, gene duplications combined with deletions and insertions of genetic material resulted in single and multidomain proteins with or without disulpfide bonds, glycosylated or not. Accordingly members of I25B subfamily likely evolved from I25A ancestors, which lack cysteine residues, acquiring disulfide bridges and signal peptide during evolutionary processes (Brown and Dziegielewska, 1997; Gregory and Maizels, 2008). In *Schistosoma* we identified a single I25B cystatin in each species, all of them containing critical motifs (**Table 2**). On the other hand, the *E. granulosus* protein EgrG_000849600.1 has signal peptide and two cystatin domains. One of them has an insignificant match with the PF00031 HMM profile (**Table 2** and **Figure 2**). This multidomain protein did not show disulfide bridges, which is not typical for secreted proteins. The *T. solium* protein TsM_000671000 displayed architecture very similar to *E. granulosus* (**Table 2** and **Figure 2**). Perhaps this indicates a protein architecture that is lineage-specific of Cestoda, although both of these species fall within the same cyclophyllidean family, and may not be representative of all tapeworms. The *H. microstoma* cystatins displayed an interesting protein architecture. They do not contain a signal peptide and also lack the LP motifs, unlike other I25A members (**Table 2**).

The legumain inhibitory motifs (SND/SNS/TND) are distinct from the papain binding motif (Q-x-V-x-G) (Alvarez-Fernandez et al., 1999). These sites are located on the opposite side of the papain binding site (Gregory and Maizels, 2008) and were present in four human sequences P01034_CYTC (SND), Q15828_CYTM (SNS), O76096_CYTF (TND), Q9H114_CST1Like (SND) in position 28 of the cystatin domain (**Table 2** and **Figure 3**). In *Schistosoma* and other Platyhelminthes, these motifs belonging to bifunctional cystatins were not present. However, in the nematode *Brugia malayi* the SND motif was identified in the secreted cystatin Bm-CPI-2 and was able to block the activity of mammalian legumain (Manoury et al., 2001). This inhibition profile is shared with other nematodes implying a dual function for nematode cystatins. In terms of adaptation to parasitism, Hartmann and Lucius (2003) compared I25B cystatins from filariae to those of *C. elegans*, and observed a distinct pattern of enzyme inhibitory activity and immunological properties.

Our results point to the diversity in terms of the presence and absence of sequence features used to classify cystatins (**Table 2**). An accurate cystatin classification based just upon these features is challenging. We used strict classification criteria, but we must take into account that many of the genomes here analyzed are still in their early versions and may contain inaccurate gene models, suggesting it is necessary to undertake manual curation for unambiguous annotation of cystatins and other proteins.

PHYLOGENETIC RELATIONSHIPS IN THE CYSTATIN FAMILY

We identified two cystatin subfamilies, I25A and I25B, in the *Schistosoma*-Human phylogeny (**Figure 4**). In this work, a comparative analysis of *Schistosoma* cystatins and other 15 experimentally validated human cystatins belonging to diverse subgroups provided insights into the abundance, diversity, and evolution of *Schistosoma* cystatin family members. According to our results, the I25 cystatin family has a few members in *Schistosoma*, including I25A and I25B subfamilies' representatives in each species. Human cystatins have diversified significantly during the course of evolution, both at the sequence and functional levels, indicating that the cystatin domain is a protein-protein interaction module that can interact with novel targets (Alvarez-Fernandez et al., 1999; Dickinson, 2002; Abrahamson et al., 2003; Cornwall and Hsia, 2003; Kordis and Turk, 2009).

A subgroup in the I25B subfamily, named Cres (cystatin-related epididymal spermatogenic)/Testatin, was identified in our study with statistical support 80/86. This clade comprises I25B *Schistosoma* sequences. In human, glycoproteins of the Cres/Testatin subgroup are expressed in reproductive tissues and their function may be related to reproduction (Frygelius et al., 2010). The topology of this subgroup showed in **Figure 4** is consistent with the phylogeny reported by Frygelius et al. (2010). Interestingly this subgroup lacks the consensus Q-x-V-x-G motif (**Table 2** and **Figure 3**). **Figure 4** shows that these proteins have a

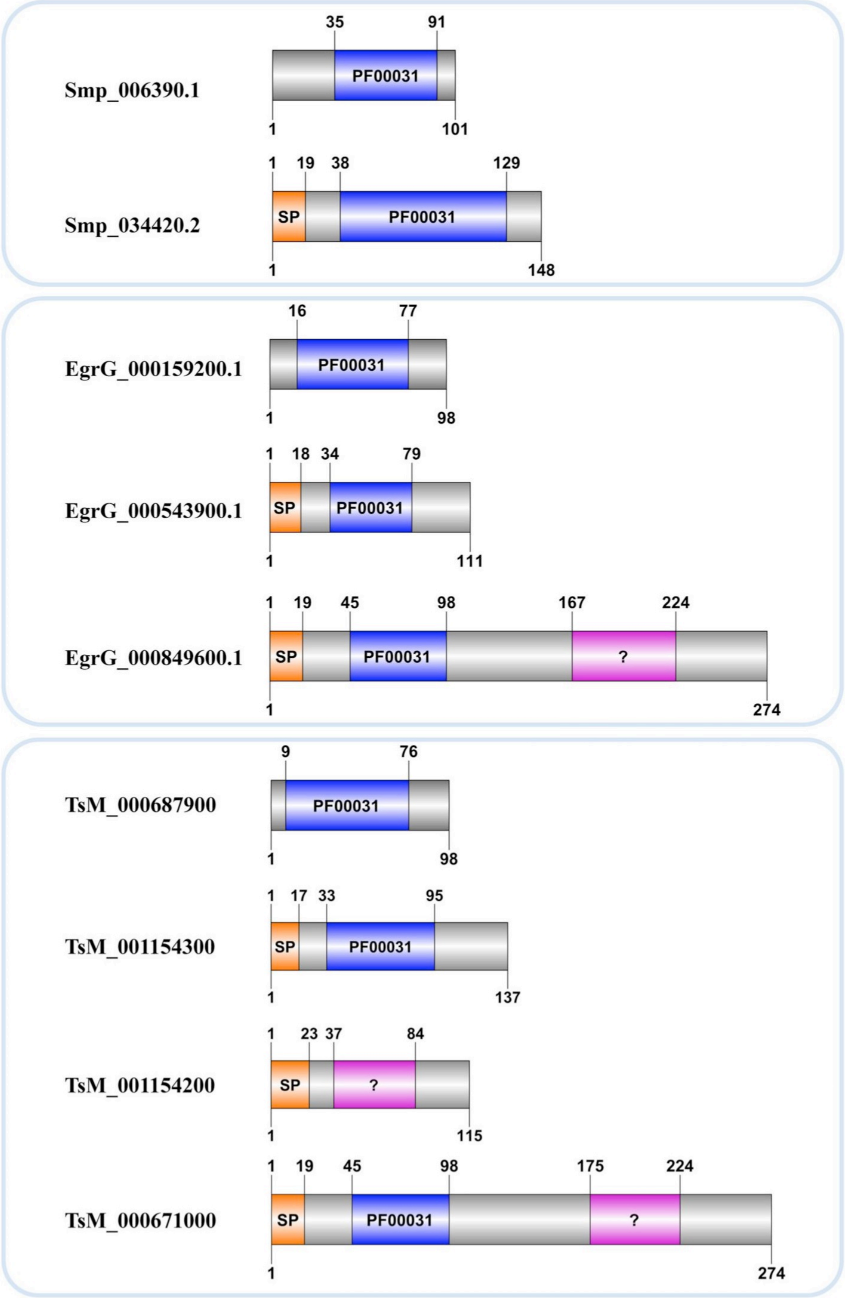

FIGURE 2 | Cystatin architecture in three Platyhelminthes. Accession numbers correspond to those assigned in each genome project. Domain limits (above) and sequence length (below) are provided for each protein. A signal peptide (SP) and the conserved domain (PF00031) with significant (blue) or insignificant (?) matches are indicated.

common origin and may represent a new subgroup within I25 family. Phylogenetic and comparative analyses show that genes involved in reproduction as Cres/Testatin and host pathogen interaction are under strong positive selection (Frygelius et al., 2010).

The phyletic distribution of the multidomain cystatins is limited and phylogenomic analyses suggest that multidomain cystatins are not monophyletic. Evidence suggests, they originated independently several times during evolution of eukaryotes (Kordis and Turk, 2009). Kininogen proteins (e.g., P01042) are

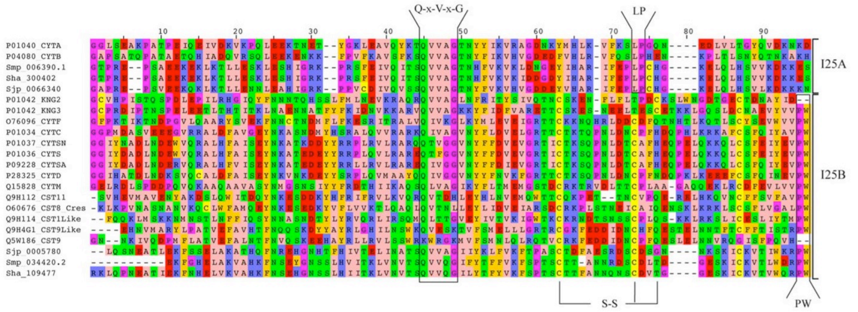

FIGURE 3 | Alignment of *Schistosoma* and human homologs. Multiple sequence alignment of the conserved domain (PF00031) of I25A and I25B proteins of *S. haematobium* (Sha), *S. japonicum* (Sjp), *S. mansoni* (Smp), and *Homo sapiens* (UniProt accession numbers). Amino acid sequences were aligned using MAFFT with iterative refinement by the G-INS-i strategy. Conserved motifs (Q-x-V-x-G and PW) and disulfide bridges (S-S) are indicated.

multidomain and divergent cystatins containing three domains with different inhibitory properties. In our phylogenetic analysis we discarded the first domain as it lacks inhibitory activity (Rawlings et al., 2014). The two remaining kininogen domains (P01042_KNG2; P01042_KNG3) were placed in the I25B subfamily clade. The P01042_KNG2 and P01042_KNG3 domains contain the Q-x-V-x-G residues critical for inhibitory activity (**Table 2**). Both domains are grouped together with other human cystatins Q15828_CYTM and O76096_CYTF that inhibit both cysteine and asparaginyl protease due to the presence of the SNS and TND motifs, respectively (**Table 2**).

Our results suggest that *Schistosoma* species contain only two cystatin subfamilies, reflecting little functional diversification. Due to the presence of highly divergent sequences in I25B clade, the recognition of orthologous sequences is a difficult task. The intracellular cystatins belonging to I25A subfamily are more conserved than the divergent extracellular cystatins (I25B) (**Figure 3**), as reported for other proteins with similar features (Julenius and Pedersen, 2006).

Khaznadji et al. (2005) reported the first I25A multidomain protein in invertebrates, a multidomain I25A in the platyhelminth *Fasciola hepatica* containing six cystatin like domains, two of which are well conserved (Khaznadji et al., 2005). The intracellular and multidomain I25A inhibits parasite cathepsin L1 activity. The methods used by Khaznadji et al. (2005) to determine the domain architecture of this cystatin differs from those applied by MEROPS (Rawlings et al., 2014), which indicates the presence of a single domain in this protein. In addition, the presence of multidomain proteins are not the only novelty in the cystatin family, several I25A cystatins from unicellular eukaryotic organisms have gained the signal peptide, which is absent in the majority of metazoan and eukaryotic I25A cystatins. The presence of signal peptide as observed in some unicellular eukaryotic I25A cystatins (Kordis and Turk, 2009) and in *Fasciola gigantica* (Siricoon et al., 2012) may lead to the gain of novel defense functions.

In synthesis, the major obstacle to the identification and classification of cystatins using amino acid sequences is the fact that many of the proteins contain multiple homologous inhibitor domains in a single protein (Rawlings et al., 2014). Furthermore, phylogenetic analysis of cystatin family members is hampered by short protein length often added to the sequence divergence. In addition, different branches appear to have evolved at different rates (Dickinson, 2002).

CYSTATIN EXPRESSION

In the present work, we interrogated publicly available gene expression datasets in order to investigate mRNA expression of both cystatin members I25A (Smp_006390.1) and I25B (Smp_034420.2) in *S. mansoni*. Microarray data by Fitzpatrick et al. (2009) assessed three ecological niches of *S. mansoni* life cycle (freshwater, molluscan, and definitive vertebrate host) and indicated similar expression levels of both cystatins among the parasite life cycle (**Figure 5**). The constitutive expression may be essential for schistosome life cycle progression. Published RNAseq data (Protasio et al., 2012) also point to the expression of cystatins in cercariae, schistosomula, and adults stages. Additional reports (Morales et al., 2004) suggest that I25A is expressed equally by adult males, females, and schistosomula stages. Therefore, both I25A and I25B are expressed throughout the *S. mansoni* life cycle.

In *S. japonicum*, He et al. (2011) observed not just the expression levels of the stefin Sjp_0066340, a I25A subfamily member, in egg, schistosomula, and adult stages by RT-PCR. He et al. (2011) also performed immunohistochemistry studies, which revealed that the *S. japonicum* stefin is mainly localized at the epithelial cells lining the gut as well as the tegument on the surface of adult worms. Additionally, the stefin of *Clonorchis sinensis* was also found mainly localized in the epithelial cells lining the intestine of the parasite (Kang et al., 2014). The stefin of *F. gigantica* was also localized in the intestinal epithelium and the tegumental type cell bodies together with the tegumental syncytium (Tarasuk et al., 2009). Altogether, the expression of parasite I25A proteins in the host-parasite interface point to a possible role in molecular interactions with host proteins, which are mostly inhibitors of host cysteine proteases such as cathepsins (Tarasuk et al., 2009; He et al., 2011; Kang et al., 2014).

Recently, an unusual secreted form of I25A member was characterized in *F. gigantica* (Siricoon et al., 2012). Although this awkward cystatin does present a signal peptide, typical of I25B proteins, sequence analysis does correlate it to the I25A subfamily. Nevertheless, Siricoon et al. (2012) provided evidence of a

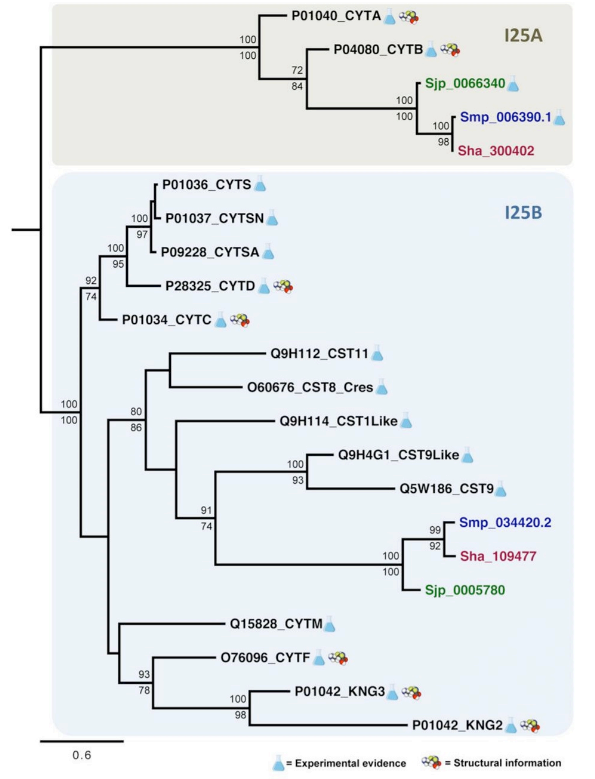

FIGURE 4 | Evolutionary relationships of I25A and I25B cysteine protease inhibitors. A total of 22 amino acid sequences and 96 sites comprising the conserved domain (PF00031) of homologs encoded by *S. haematobium* (brown), *S. japonicum* (green), *S. mansoni* (blue), and *Homo sapiens* (black) were analyzed. The phylogeny was reconstructed by two methods using WAG was as the best fit model. In the Bayesian inference, support values for each node were estimated as posterior probability (above). In the maximum likelihood analysis, they were estimated using the Akaike Likelihood Ratio Test (aLRT) (below). Only support values higher than 70% are shown.

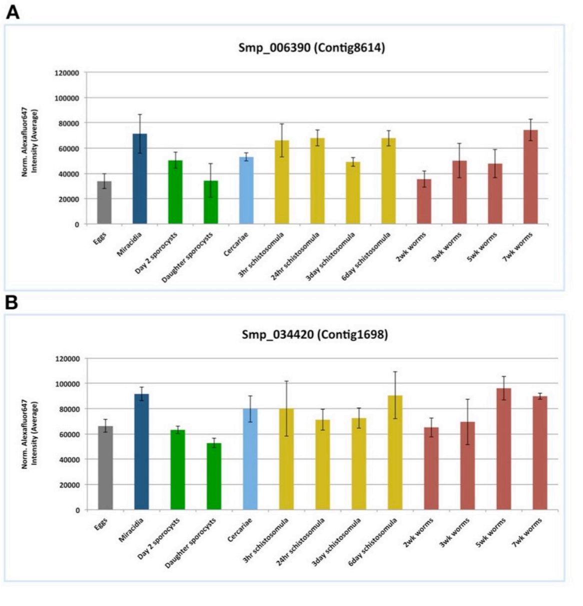

FIGURE 5 | Cystatins mRNA expression patterns in the *S. mansoni* life cycle. Public microarray data available at ArrayExpress (E-MEXP-2094) was downloaded to a local server in order to identify cystatins expression patterns of the I25A subfamily member Smp_006390 **(A)** and the I25B subfamily member Smp_034420 **(B)** in *S. mansoni*. Bars correspond to the mean normalized values for each oligonucleotide probe named Smp_006390 and Smp_034420 in 13 different life stages.

secreted form of a cystatin protein in Platyhelminthes that was again observed in the intestinal epithelium in all developmental stages. Moreover, the secreted 125A protein was also found expressed in the prostate gland in the adult stage of *F. gigantica*, which suggests a regulative role of cysteine protease activity in reproductive system. Similarly, the expression of human cystatin I25B subgroup proteins, also called Cres/Testatin, was localized at the reproductive tissues and their function may be related to reproduction (Frygelius et al., 2010). Interestingly, human Cres/Testatin subgroup was placed in the same clade with the *Schistosoma* cystatins I25B (**Figure 4**).

Based on the evidence of expression in related organisms and given the constitutive expression of cystatins I25A and I25B in *S. mansoni*, it is possible that cystatin functions can be involved in key processes in *Schistosoma*. Such proteins may be required to keep its proteolytic activity balanced as well as to protect the parasite against degradation by host or endogenous proteins. Nevertheless, cystatin tissue-specific expression, such as those identified at the reproductive system in human and *F. gigantica*, could evidence a more specialized role against specific cysteine proteases.

CONCLUSIONS

In summary, our evolutionary analysis using genomic, transcriptomic, and proteomic data for three *Schistosoma* species and other Platyhelminthes has provided the first insights into the evolution, classification, and functional diversification of platyhelminth cystatins. These findings improve our understanding concerning the diversity, at the molecular level, of cystatins encoded by such species. Only two subfamilies (I25A and I25B) were clearly identified in *Schistosoma* and Platyhelminthes reflecting the low diversification of this family when compared to human. Regarding Cestoda, it is necessary to implement an exhaustive study in order to better understand the domain composition revealed in our work.

We expect that this study will encourage experimental and structural characterization of cystatins in *Schistosoma* and other closely related parasites. Altogether, studies involving parasite cystatins will help to elucidate the functions performed by those proteins as well their correlation with parasite biology and host-parasite interaction. The importance of new insights revealed by functional genomics as RNAi experiments and comparative expression patterns across different life cycle stages in *Schistosoma* and other Platyhelminthes will provide a functional landscape of the cystatin role in the parasite life cycle.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: Yesid Cuesta-Astroz, Larissa L. S. Scholte, Fabiano Sviatopolk-Mirsky Pais, Guilherme Oliveira, and Laila A. Nahum. Carried out homologs and protein signatures identification: Yesid Cuesta-Astroz and Larissa L. S. Scholte. Performed expression analysis: Fabiano Sviatopolk-Mirsky Pais and Yesid Cuesta-Astroz. Performed the phylogenetic studies: Yesid Cuesta-Astroz and Larissa L. S. Scholte. Wrote the manuscript: Yesid Cuesta-Astroz, Larissa L. S. Scholte, Fabiano Sviatopolk-Mirsky Pais, Guilherme Oliveira and Laila A. Nahum. Reviewed and revised the manuscript: Laila A. Nahum and Guilherme Oliveira. Coordinated this study: Laila A. Nahum and Guilherme Oliveira. All authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health-NIH/Fogarty International Center (TW007012 to Guilherme Oliveira), the National Council for Research and Development, CNPq (CNPq Research Fellowship 309312/2012-4 to Guilherme Oliveira, INCT-DT 573839/2008-5 to Guilherme Oliveira, and CNPq-Universal 476036/2010-0 to Laila A. Nahum), and the Research Foundation of the State of Minas Gerais, FAPEMIG (CBB-1181/08 and PPM-00439-10 to Guilherme Oliveira). The authors are grateful to Dr. Eric Ross (Stowers Institute for Medical Research, USA) for providing the *Schmidtea mediterranea* proteome data. The authors acknowledge the use of the computing resources of the Center for Excellence in Bioinformatics (CEBio) and thank Mariana Aparecida de Oliveira Dias (CEBio) for the graphic design of **Figure 1**. Authors are very grateful to the two anonymous reviewers for their comments and suggestions to improve this work.

REFERENCES

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21, 2104–2105. doi: 10.1093/bioinformatics/bti263

Abrahamson, M., Alvarez-Fernandez, M., and Nathanson, C. M. (2003). Cystatins. *Biochem. Soc. Symp.* 70, 179–199.

Alvarez-Fernandez, M., Barrett, A. J., Gerhartz, B., Dando, P. M., Ni, J., and Abrahamson, M. (1999). Inhibition of mammalian legumain by some cystatins is due to a novel second reactive site. *J. Biol. Chem.* 274, 19195–19203. doi: 10.1074/jbc.274.27.19195

Barrett, A. J. (1986). The cystatins: a diverse superfamily of cysteine peptidase inhibitors. *Biomed. Biochim. Acta* 45, 1363–1374.

Berriman, M., Haas, B. J., Loverde, P. T., Wilson, R. A., Dillon, G. P., Cerqueira, G. C., et al. (2009). The genome of the blood fluke *Schistosoma mansoni*. *Nature* 460, 352–358. doi: 10.1038/nature08160

Brown, W. M., and Dziegielewska, K. M. (1997). Friends and relations of the cystatin superfamily–new members and their evolution. *Protein Sci.* 6, 5–12. doi: 10.1002/pro.5560060102

Coeli, R., Baba, E. H., Araujo, N., Coelho, P. M., and Oliveira, G. (2013). Praziquantel treatment decreases *Schistosoma mansoni* genetic diversity in experimental infections. *PLoS Negl. Trop. Dis.* 7:e2596. doi: 10.1371/journal.pntd.0002596

Cornwall, G. A., Cameron, A., Lindberg, I., Hardy, D. M., Cormier, N., and Hsia, N. (2003). The cystatin-related epididymal spermatogenic protein inhibits the serine protease prohormone convertase 2. *Endocrinology* 144, 901–908. doi: 10.1210/en.2002-220997

Cornwall, G. A., and Hsia, N. (2003). A new subgroup of the family 2 cystatins. *Mol. Cell. Endocrinol.* 200, 1–8. doi: 10.1016/S0303-7207(02)00408-2

Dainichi, T., Maekawa, Y., Ishii, K., Zhang, T., Nashed, B. F., Sakai, T., et al. (2001). Nyppocystatin, a cysteine protease inhibitor from *Nippostrongylus brasiliensis,* inhibits antigen processing and modulates antigen-specific immune response. *Infect. Immun.* 69, 7380–7386. doi: 10.1128/IAI.69.12.7380-7386.2001

Dickinson, D. P. (2002). Salivary (SD-type) cystatins: over one billion years in the making–but to what purpose? *Crit. Rev. Oral Biol. Med.* 13, 485–508. doi: 10.1177/154411130201300606

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Fitzpatrick, J. M., Peak, E., Perally, S., Chalmers, I. W., Barrett, J., Yoshino, T. P., et al. (2009). Anti-schistosomal intervention targets identified by lifecycle transcriptomic analyses. *PLoS Negl. Trop. Dis.* 3:e543. doi: 10.1371/journal.pntd.0000543

Frygelius, J., Arvestad, L., Wedell, A., and Tohonen, V. (2010). Evolution and human tissue expression of the Cres/Testatin subgroup genes, a reproductive tissue specific subgroup of the type 2 cystatins. *Evol. Dev.* 12, 329–342. doi: 10.1111/j.1525-142X.2010.00418.x

Gregory, W. F., and Maizels, R. M. (2008). Cystatins from filarial parasites: evolution, adaptation and function in the host-parasite relationship. *Int. J. Biochem. Cell Biol.* 40, 1389–1398. doi: 10.1016/j.biocel.2007.11.012

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Harnett, W. (2014). Secretory products of helminth parasites as immunomodulators. *Mol. Biochem. Parasitol.* doi: 10.1016/j.molbiopara.2014.03.007. [Epub ahead of print].

Hartmann, S., Adam, R., Marti, T., Kirsten, C., Seidinger, S., and Lucius, R. (1997). A 41-kDa antigen of the rodent filaria *Acanthocheilonema viteae* with homologies to tropomyosin induces host-protective immune responses. *Parasitol. Res.* 83, 390–393. doi: 10.1007/s004360050269

Hartmann, S., and Lucius, R. (2003). Modulation of host immune responses by nematode cystatins. *Int. J. Parasitol.* 33, 1291–1302. doi: 10.1016/S0020-7519(03)00163-2

He, B., Cai, G., Ni, Y., Li, Y., Zong, H., and He, L. (2011). Characterization and expression of a novel cystatin gene from *Schistosoma japonicum*. *Mol. Cell. Probes* 25, 186–193. doi: 10.1016/j.mcp.2011.05.001

Henskens, Y. M., Veerman, E. C., and Nieuw Amerongen, A. V. (1996). Cystatins in health and disease. *Biol. Chem. Hoppe Seyler* 377, 71–86.

Higashiyama, S., Ohkubo, I., Ishiguro, H., Sasaki, M., Matsuda, T., and Nakamura, R. (1987). Heavy chain of human high molecular weight and low molecular weight kininogens binds calcium ion. *Biochemistry* 26, 7450–7458. doi: 10.1021/bi00397a038

Julenius, K., and Pedersen, A. G. (2006). Protein evolution is faster outside the cell. *Mol. Biol. Evol.* 23, 2039–2048. doi: 10.1093/molbev/msl081

Kang, J. M., Ju, H. L., Lee, K. H., Kim, T. S., Pak, J. H., Sohn, W. M., et al. (2014). Identification and characterization of the second cysteine protease inhibitor of *Clonorchis sinensis* (CsStefin-2). *Parasitol. Res.* 113, 47–58. doi: 10.1007/s00436-013-3624-8

Katoh, K., Asimenos, G., and Toh, H. (2009). Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.* 537, 39–64. doi: 10.1007/978-1-59745-251-9_3

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Res.* 37, D767–D772. doi: 10.1093/nar/gkn892

Khaznadji, E., Collins, P., Dalton, J. P., Bigot, Y., and Moire, N. (2005). A new multidomain member of the cystatin superfamily expressed by *Fasciola hepatica*. *Int. J. Parasitol.* 35, 1115–1125. doi: 10.1016/j.ijpara.2005.05.001

Klotz, C., Ziegler, T., Danilowicz-Luebert, E., and Hartmann, S. (2011). Cystatins of parasitic organisms. *Adv. Exp. Med. Biol.* 712, 208–221. doi: 10.1007/978-1-4419-8414-2_13

Kordis, D., and Turk, V. (2009). Phylogenomic analysis of the cystatin superfamily in eukaryotes and prokaryotes. *BMC Evol. Biol.* 9:266. doi: 10.1186/1471-2148-9-266

Liang, Y. S., Dai, J. R., Zhu, Y. C., Coles, G. C., and Doenhoff, M. J. (2003). Genetic analysis of praziquantel resistance in *Schistosoma mansoni*. *Southeast Asian J. Trop. Med. Public Health* 34, 274–280.

Lustigman, S., Brotman, B., Huima, T., and Prince, A. M. (1991). Characterization of an *Onchocerca volvulus* cDNA clone encoding a genus specific antigen present in infective larvae and adult worms. *Mol. Biochem. Parasitol.* 45, 65–75. doi: 10.1016/0166-6851(91)90028-5

Lustigman, S., Brotman, B., Huima, T., Prince, A. M., and McKerrow, J. H. (1992). Molecular cloning and characterization of onchocystatin, a cysteine proteinase inhibitor of *Onchocerca volvulus*. *J. Biol. Chem.* 267, 17339–17346.

Manoury, B., Gregory, W. F., Maizels, R. M., and Watts, C. (2001). Bm-CPI-2, a cystatin homolog secreted by the filarial parasite *Brugia malayi*, inhibits class II MHC-restricted antigen processing. *Curr. Biol.* 11, 447–451. doi: 10.1016/S0960-9822(01)00118-X

Melman, S. D., Steinauer, M. L., Cunningham, C., Kubatko, L. S., Mwangi, I. N., Wynn, N. B., et al. (2009). Reduced susceptibility to praziquantel among naturally occurring Kenyan isolates of *Schistosoma mansoni*. *PLoS Negl. Trop. Dis.* 3:e504. doi: 10.1371/journal.pntd.0000504

Morales, F. C., Furtado, D. R., and Rumjanek, F. D. (2004). The N-terminus moiety of the cystatin SmCys from *Schistosoma mansoni* regulates its inhibitory activity *in vitro* and *in vivo*. *Mol. Biochem. Parasitol.* 134, 65–73. doi: 10.1016/j.molbiopara.2003.10.016

Newlands, G. F., Skuce, P. J., Knox, D. P., and Smith, W. D. (2001). Cloning and expression of cystatin, a potent cysteine protease inhibitor from the gut of *Haemonchus contortus*. *Parasitology* 122, 371–378. doi: 10.1017/S0031182001007302

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Pol, E., and Bjork, I. (1999). Importance of the second binding loop and the C-terminal end of cystatin B (stefin B) for inhibition of cysteine proteinases. *Biochemistry* 38, 10519–10526. doi: 10.1021/bi990488k

Protasio, A. V., Tsai, I. J., Babbage, A., Nichol, S., Hunt, M., Aslett, M. A., et al. (2012). A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS Negl. Trop. Dis.* 6:e1455. doi: 10.1371/journal.pntd.0001455

Rawlings, N. D., and Barrett, A. J. (1990). Evolution of proteins of the cystatin superfamily. *J. Mol. Evol.* 30, 60–71. doi: 10.1007/BF02102453

Rawlings, N. D., Waller, M., Barrett, A. J., and Bateman, A. (2014). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* 42, D503–D509. doi: 10.1093/nar/gkt953

Ren, J., Wen, L., Gao, X., Jin, C., Xue, Y., and Yao, X. (2009). DOG 1.0: illustrator of protein domain structures. *Cell Res.* 19, 271–273. doi: 10.1038/cr.2009.6

Ronquist, F., and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. doi: 10.1093/bioinformatics/btg180

Schierack, P., Lucius, R., Sonnenburg, B., Schilling, K., and Hartmann, S. (2003). Parasite-specific immunomodulatory functions of filarial cystatin. *Infect. Immun.* 71, 2422–2429. doi: 10.1128/IAI.71.5.2422-2429.2003

Schwarz, A., Valdes, J. J., and Kotsyfakis, M. (2012). The role of cystatins in tick physiology and blood feeding. *Ticks Tick Borne Dis.* 3, 117–127. doi: 10.1016/j.ttbdis.2012.03.004

Silva, L. L., Marcet-Houben, M., Nahum, L. A., Zerlotini, A., Gabaldon, T., and Oliveira, G. (2012). The *Schistosoma mansoni* phylome: using evolutionary genomics to gain insight into a parasite's biology. *BMC Genomics* 13:617. doi: 10.1186/1471-2164-13-617

Silva, L. L., Marcet-Houben, M., Zerlotini, A., Gabaldon, T., Oliveira, G., and Nahum, L. A. (2011). Evolutionary histories of expanded peptidase families in *Schistosoma mansoni*. *Mem. Inst. Oswaldo Cruz* 106, 864–877. doi: 10.1590/S0074-02762011000700013

Siricoon, S., Grams, S. V., and Grams, R. (2012). Efficient inhibition of cathepsin B by a secreted type 1 cystatin of *Fasciola gigantica*. *Mol. Biochem. Parasitol.* 186, 126–133. doi: 10.1016/j.molbiopara.2012.10.003

Tarasuk, M., Vichasri Grams, S., Viyanant, V., and Grams, R. (2009). Type I cystatin (stefin) is a major component of *Fasciola gigantica* excretion/secretion product. *Mol. Biochem. Parasitol.* 167, 60–71. doi: 10.1016/j.molbiopara.2009.04.010

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics* 14, 178–192. doi: 10.1093/bib/bbs017

Toh, E. C., Huq, N. L., Dashper, S. G., and Reynolds, E. C. (2010). Cysteine protease inhibitors: from evolutionary relationships to modern chemotherapeutic design for the treatment of infectious diseases. *Curr. Protein Pept. Sci.* 11, 725–743. doi: 10.2174/138920310794557646

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Tsai, I. J., Zarowiecki, M., Holroyd, N., Garciarrubio, A., Sanchez-Flores, A., Brooks, K. L., et al. (2013). The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* 496, 57–63. doi: 10.1038/nature12031

UniProt Consortium. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 42, D191–D198. doi: 10.1093/nar/gkt1140

Valente, R. H., Dragulev, B., Perales, J., Fox, J. W., and Domont, G. B. (2001). BJ46a, a snake venom metalloproteinase inhibitor. Isolation, characterization, cloning and insights into its mechanism of action. *Eur. J. Biochem.* 268, 3042–3052. doi: 10.1046/j.1432-1327.2001.02199.x

Vray, B., Hartmann, S., and Hoebeke, J. (2002). Immunomodulatory properties of cystatins. *Cell. Mol. Life Sci.* 59, 3042–3052. doi: 10.1007/s00018-002-8525-4

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033

WHO. (2012). Schistosomiasis: population requiring preventive chemotherapy and number of people treated in 2010. *Wkly. Epidemiol. Rec.* 87, 37–44.

Young, N. D., Jex, A. R., Li, B., Liu, S., Yang, L., Xiong, Z., et al. (2012). Whole-genome sequence of *Schistosoma haematobium*. *Nat. Genet.* 44, 221–225. doi: 10.1038/ng.1065

Zavasnik-Bergant, T. (2008). Cystatin protease inhibitors and immune functions. *Front. Biosci.* 13:4625–4637. doi: 10.2741/3028

Zerlotini, A., Aguiar, E. R., Yu, F., Xu, H., Li, Y., Young, N. D., et al. (2013). SchistoDB: an updated genome resource for the three key schistosomes of humans. *Nucleic Acids Res.* 41, D728–D731. doi: 10.1093/nar/gks1087

Zhou, Y., Zheng, H., Chen, Y., Zhang, L., Wang, K., Guo, J., et al. (2009). The *Schistosoma japonicum* genome reveals features of host-parasite interplay. *Nature* 460, 345–351. doi: 10.1038/nature08140

Conflict of Interest Statement: The Guest Associate Editor Arnon Dias Jurberg declares that, despite being affiliated to the same institution as authors Cuesta-Astroz, Scholte, Pais, Oliveira and Nahum, the review process was handled objectively and no conflict of interest exists. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 26 April 2014; accepted: 18 June 2014; published online: 09 July 2014.
Citation: Cuesta-Astroz Y, Scholte LLS, Pais FS-M, Oliveira G and Nahum LA (2014) Evolutionary analysis of the cystatin family in three Schistosoma species. Front. Genet. 5:206. doi: 10.3389/fgene.2014.00206
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics.

IV – DISCUSSÃO

Apesar da quantidade extraordinária de dados gerados pelos projetos de sequenciamento de DNA (genomas ou genes específicos), existe uma grande demanda por análises que possam auxiliar a interpretação destes dados integrando a informação genotípica à diversidade fenotípica dos organismos e sua adaptação a ambientes diversos (Aguilar-Diaz et al. 2006; Berriman et al. 2009; Saarma et al. 2009). Além disso, existe uma grande demanda pela identificação de novos alvos terapêuticos e desenvolvimento de novos fármacos para tratamento das helmintíases, especialmente pelas limitações apresentadas pelas drogas disponíveis.

As proteínas secretadas são capazes de modificar ou manipular o ambiente do hospedeiro, além de modular a sua imunidade. Pouco se conhece acerca da função e evolução das proteínas secretadas pelos helmintos e sua interação com seus hospedeiros. A identificação de proteínas secretadas pode proporcionar um catálogo de potenciais novos imunomoduladores e o desenvolvimento de novos testes diagnósticos ou potenciais novos tratamentos, sendo o secretoma uma interessante e promissória rota de intervenção e terapia, além da identificação de novos alvos terapêuticos (Liang et al. 2003; Melman et al. 2009; Coeli et al. 2013).

A disponibilidade de dados do proteoma predito de alguns helmintos de vida livre e parasitária e sua análise comparativa pode nos dar uma ideia da diversidade dos secretomas destes organismos e contribuir para o entendimento da evolução do parasitismo incluindo as interações parasito-hospedeiro. Métodos experimentais são normalmente trabalhosos e extremamente caros. Assim predições computacionais são uma abordagem atraente, particularmente no contexto da avaliação em grande escala devido a sua relativa facilidade e rapidez. No entanto, como todas as ferramentas de predição, há uma taxa de erro em função de falsos positivos e negativos como deveria ser esperado (Torto et al. 2003).

A função de uma proteína está altamente relacionada com o contexto. Neste caso, as proteínas secretadas estão relacionadas por exemplo com o estágio do ciclo de vida do parasito entre outros fatores. Além disso, as proteínas secretadas podem ter uma regulação pós-traducional ou ter uma retenção do peptídeo sinal que impede a sua secreção. A integração dos resultados computacionais com dados transcritômicos e proteômicos pode melhorar a identificação de proteínas que estão sendo secretadas (Caccia et al. 2013) além de melhorar a confiabilidade da interpretação funcional dos nossos resultados.

As proteínas estão sendo cada vez mais identificadas experimentalmente em locais inesperados, apresentando um desafio importante nos estudos bioinformáticos do secretoma, assim como na análise geral dos dados biológicos (Caccia et al. 2013). Algumas dessas proteínas podem ter uma localização incorreta e representarem artefatos, enquanto outras podem ser mal interpretadas. Também têm sido observado que proteínas específicas podem executar funções diferentes dependendo da sua localização subcelular. O fenômeno pelo qual diferentes formas da mesma proteína desempenham diferentes funções em locais diferentes é denominado *moonlighting* e pode desafiar a interpretação dos dados de proteômica e a compreensão do significado biológico da localização das proteínas (Copley 2012; Butler and Overall 2009).

Devido às limitações das ferramentas computacionais e nosso incompleto conhecimento da biologia das proteínas secretadas e exportadas e na ausência da prova experimental direta da secreção, é improvável que um secretoma predito *in silico* represente exatamente o verdadeiro secretoma. Uma nova geração de ferramentas bioinformáticas terão que incluir análises integradas realizadas em diversas bases de dados "ômicos", mineração de dados e recursos de mineração de texto para a extração de conhecimento e interoperabilidade com ontologias para capturar propriedades funcionais das proteínas. Além disso, análises experimentais avançadas devem ser desenvolvidas para a verificação dos secretomas.

Motivados pelas vantagens dos métodos computacionais e com a finalidade de fazer um estudo comparativo abrangente em helmintos, identificamos computacionalmente as proteínas secretadas preditas em 44 espécies de helmintos com diferentes estilos de vida, linhagens filogenéticas e hospedeiros. O conjunto de proteínas secretadas evidencia relações evolutivas e adaptação a vários habitats (Capitulo I). Os secretomas obtidos foram representados em 41,200 proteínas, com 31,192 preditas pela via clássica (presença de peptídeo sinal) e 10,008 identificadas pela via não clássica.

Diferentes tamanhos de secretomas foram identificados, com os platelmintos apresentando secretomas com menor tamanho. Secretomas pequenos podem estar relacionados com um ambiente ou nicho com compostos ou nutrientes que são fáceis de metabolizar (Krijger et al. 2014). Os secretomas maiores foram encontrados em helmintos nematoides de vida livre e plantas, com exceção de *H. contortus* que apresentou um secretoma comparável com os anteriormente citados. O alto número de proteínas secretadas nos nematoides de vida livre *C. elegans* e *P. pacificus* permite usar uma ampla variedade de substratos presentes no solo ou restos de plantas, que são provavelmente mais difíceis de degradar do que os substratos disponíveis no hospedeiro animal (Krijger et al. 2014).

O estudo da diversidade de domínios nos secretomas é uma maneira rápida e eficaz para caracterizar a diversidade de proteínas e pode fornecer pistas para os estilos de vida e ambientes em que esses organismos vivem. Foram encontrados poucos pontos em comum nos secretomas em termos de domínios proteicos (Pfam), possivelmente devido ao fato dos helmintos parasitas terem evoluído independentemente, sofrendo adaptações específicas para o seu nicho especial (Zarowiecki and Berriman 2015). Apenas cinco domínios foram comuns aos 44 secretomas (PF00014, PF07679, PF00085, PF00046, PF00188). Esses domínios estão envolvidos em processos biológicos tais como peptidases e inibidores, antioxidantes e adesão celular, sendo todos esses processos importantes na interação com o hospedeiro.

Por outro lado foram encontrados domínios espécie-específicos. Estes domínios estão envolvidos em funções específicas relacionadas ao estilo de vida ou nicho em helmintos. Dentre os domínios específicos encontrados destacam-se aqueles associados a helmintos de plantas, como celulases, pectato liase e enzimas que degradam a parede celular as quais são adaptações chave para o parasitismo de plantas. Provavelmente estes foram adquiridos por transferência horizontal de genes a partir de rizobios e, portanto, funções bacterianas têm sido propostas como uma importante força motriz na evolução do parasitismo (Dieterich and Sommer 2009). No helminto parasita de batata *G. pallida* foi encontrado o domínio exclusivo NPP1 (*necrosis inducing protein*, PF05630) que está envolvido na fase inicial da infecção e está presente no fungo da batata *Phytophtora infestans,* o que sugere que o parasitismo de plantas evoluiu de associações fúngicas (Dieterich and Sommer 2009).

Em adição aos domínios específicos de helmintos de plantas, os específicos de helmintos de vida livre foram identificados. Este grupo de helmintos são expostos a um maior número de substratos mais complexos. Em geral, é provável encontrar mais estresse oxidativo e xenobióticos do que em endoparasitas obrigatórios, por isto organismos de vida livre têm sistemas redoxes mais complexos e variados (Zarowiecki and Berriman 2015). De acordo com nossos resultados encontramos termos GO (*Gene Ontology*) enriquecidos relacionados com o ambiente de helmintos de vida livre, tais como resposta a UV (GO:0009411) e resposta a estímulo xenobiótico (GO:0009410).

Um outro grupo de domínios de proteínas analisado foram os domínios mais representados nos secretomas como as proteases e inibidores, quinases, domínios envolvidos em adesão celular e interação proteína-proteína. Um dos domínios com maior representação nas proteínas secretadas foi o Kunitz_BPTI (PF00014) que é um domínio inibidor de proteases e tem sido sugerido como envolvido em proteger os helmintos das moléculas do hospedeiro, em especial os derivados do trato gastrointestinal, tais como uma ampla diversidade de peptidases (Heizer et al. 2013). O estudo do domínio Kunitz tem fornecido soluções para doenças humanas, levando ao desenvolvimento de produtos farmacêuticos a partir deste domínio, tais como o inibidor de calicreina (serino protease) e foi chamado *ecallantide* usado para o tratamento de angioedema (Lehmann 2008). Este é um exemplo de que os estudos da diversidade de domínios poderia ser um campo promissor na busca de potenciais fármacos antihelmínticos.

Além da análise da presença de domínios e a associação com o estilo de vida de helmintos, a importância das repetições de domínios em proteínas secretadas tem sido estudada. A repetição de domínios é um mecanismo predominante para a diversidade de proteínas e sua evolução. Em alguns casos, a repetição de domínios em uma proteína multidomínio é essencial para a função da proteína (Messih et al. 2012). Nossos resultados mostraram que domínios PAN, TSP-1, VWA, I-set e EGF estão normalmente presentes como cópias em série (*tandem*), as vezes em combinação com outros domínios de adesão, sendo as proteínas que contêm esta arquitetura importantes no processo de invasão.

Por outro lado, estudos evolutivos e de expressão em famílias proteicas que fazem parte do secretoma são relevantes na compreensão da interação parasito-hospedeiro. A família das cistatinas foi nosso estudo de caso. Tais proteínas são inibidores de cisteíno proteases e estão envolvidas na regulação da atividade proteolítica e na modulação do sistema imune do hospedeiro (Capitulo II). Com o intuito de identificar, classificar e anotar funcionalmente as cistatinas no proteoma predito de três espécies de *Schistosoma* e outros platelmintos, adotamos uma abordagem computacional baseada em modelos ocultos de Markov (*hidden Markov models*), abordagem filogenética para anotação funcional de homólogos e análise de expressão em diferentes estágios do ciclo de vida de *S. mansoni.* Como resultado, em diferentes espécies de *Schistosoma* somente foram identificadas cistatinas I25A e I25B refletindo pouca diversificação funcional. I25C e a subfamília de "não classificados" não foram identificadas nas espécies de platelmintos analisadas.

A análise filogenética indicou que as cistatinas pertencem a diferentes clados, refletindo a sua diversidade molecular. Nossos resultados sugerem que as cistatinas de *Schistosoma* são muito divergentes dos homólogos humanos, especialmente a subfamília I25B. A análise dos dados transcritômicos indicou que os genes de I25A e I25B são constitutivamente expressos e, portanto, podem ser essenciais para a progressão do ciclo de vida de *Schistosoma.*

O trabalho apresentado nesta tese de doutorado é pioneiro no estudo comparativo de secretomas preditos em escala genômica em helmintos, analisando espécies com diferentes estilos de vida e hospedeiros para contribuir na compreensão da diversidade, função e adaptação aos seus nichos específicos. Além da interação parasito-hospedeiro mediada pelas proteínas secretadas e sua diversidade que reflete especificidades dos helmintos a diferentes ambientes, este estudo pode auxiliar na identificação de novos alvos terapêuticos ou moléculas diagnóstico.

V – CONCLUSÕES

O avanço das tecnologias de sequenciamento associadas ao aumento do poder computacional levou um crescimento no volume de dados genômicos, transcritômicos e proteômicos em diferentes espécies de parasitos o que abriu novas fronteiras na compreensão da biologia de parasitos e na identificação de novos alvos terapêuticos. Neste contexto, o estudo das proteínas secretadas disponibiliza uma fonte de informações sobre novos alvos de drogas e marcadores diagnóstico. Por este motivo, o estudo dos secretomas de helmintos terá um grande impacto para controlar as helmintíases futuramente. Como um primeiro passo para elucidar as bases moleculares da colonização do hospedeiro torna-se necessário determinar o repertório de proteínas secretadas pelo parasito. Além disso, o secretoma de helmintos irá mostrar o *fingerprinting* da adaptação a vários habitats. Esta abordagem tem impacto também no estudo da interação parasito-hospedeiro que é importante para a compreensão os mecanismos de infecção pelos quais o helminto entra no hospedeiro e persiste.

VI – REFERENCIAS

Abrahamson, M. 1994. “Cystatins.” *Methods Enzymol* 244: 685–700. http://www.ncbi.nlm.nih.gov/pubmed/7845245.

Aguilar-Diaz, H, R J Bobes, J C Carrero, R Camacho-Carranza, C Cervantes, M A Cevallos, G Davila, et al. 2006. “The Genome Project of Taenia Solium.” *Parasitol Int* 55 Suppl: S127-30. doi:10.1016/j.parint.2005.11.020.

Altschul, S F, W Gish, W Miller, E W Myers, and D J Lipman. 1990. “Basic Local Alignment Search Tool.” *J Mol Biol* 215 (3): 403–10. doi:10.1016/S0022-2836(05)80360-2.

Bendtsen, Jannick Dyrløv, Lars Juhl Jensen, Nikolaj Blom, Gunnar Von Heijne, and Søren Brunak. 2004. “Feature-Based Prediction of Non-Classical and Leaderless Protein Secretion.” *Protein Engineering, Design and Selection* 17 (4): 349–56. doi:10.1093/protein/gzh037.

Bendtsen, Jannick Dyrløv, Henrik Nielsen, David Widdick, Tracy Palmer, and Søren Brunak. 2005. “Prediction of Twin-Arginine Signal Peptides.” *BMC Bioinformatics* 6 (July): 167. doi:10.1186/1471-2105-6-167.

Berriman, M, B J Haas, P T LoVerde, R A Wilson, G P Dillon, G C Cerqueira, S T Mashiyama, et al. 2009. “The Genome of the Blood Fluke Schistosoma Mansoni.” *Nature* 460 (7253): 352–58. doi:10.1038/nature08160.

Boissier, Jérôme, Sébastien Grech-Angelini, Bonnie L Webster, Jean-François Allienne, Tine Huyse, Santiago Mas-Coma, Eve Toulza, et al. 2016. “Outbreak of Urogenital Schistosomiasis in Corsica (France): An Epidemiological Case Study.” *The Lancet Infectious Diseases* 16 (8): 971–79. doi:10.1016/S1473-3099(16)00175-4.

Bork, P, T Dandekar, Y Diaz-Lazcoz, F Eisenhaber, M Huynen, and Y Yuan. 1998. “Predicting Function: From Genes to Genomes and Back.” *J Mol Biol* 283 (4): 707–25. doi:10.1006/jmbi.1998.2144.

Brindley, Paul J., Makedonka Mitreva, Elodie Ghedin, and Sara Lustigman. 2009. “Helminth Genomics: The Implications for Human Health.” *PLoS Neglected Tropical Diseases*. doi:10.1371/journal.pntd.0000538.

Butler, Georgina S., and Christopher M. Overall. 2009. “Proteomic Identification of Multitasking Proteins in Unexpected Locations Complicates Drug Targeting.” *Nature Reviews Drug Discovery* 8 (12): 935–48. doi:10.1038/nrd2945.

Caccia, Dario, Matteo Dugo, Maurizio Callari, and Italia Bongarzone. 2013. “Bioinformatics Tools for Secretome Analysis.” *Biochimica et Biophysica Acta* 1834 (11): 2442–53. doi:10.1016/j.bbapap.2013.01.039.

Choo, Khar, Tin Tan, and Shoba Ranganathan. 2009. “A Comprehensive Assessment of N-Terminal Signal Peptides Prediction Methods.” *BMC Bioinformatics* 10 (Suppl 15): S2. doi:10.1186/1471-2105-10-S15-S2.

Coeli, R, E H Baba, N Araujo, P M Coelho, and G Oliveira. 2013. “Praziquantel Treatment Decreases Schistosoma Mansoni Genetic Diversity in Experimental Infections.” *PLoS Negl Trop Dis* 7 (12): e2596. doi:10.1371/journal.pntd.0002596.

Copley, Shelley D. 2012. “Moonlighting Is Mainstream: Paradigm Adjustment Required.” *BioEssays* 34 (7): 578–88. doi:10.1002/bies.201100191.

Correale, J., and M. F. Farez. 2011. “The Impact of Environmental Infections (Parasites) on MS Activity.” *Multiple Sclerosis Journal* 17 (10): 1162–69. doi:10.1177/1352458511418027.

Dieterich, Christoph, and Ralf J. Sommer. 2009. “How to Become a Parasite - Lessons from the Genomes of Nematodes.” *Trends in Genetics* 25 (5): 203–9. doi:10.1016/j.tig.2009.03.006.

Eisen, J A. 1998. “Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis.” *Genome Res* 8 (3): 163–67. http://www.ncbi.nlm.nih.gov/pubmed/9521918.

Eisen, J A, D Kaiser, and R M Myers. 1997. “Gastrogenomic Delights: A Movable Feast.” *Nature Medicine* 3 (10): 1076–78. http://www.ncbi.nlm.nih.gov/pubmed/9334711.

Frank, K., and M. J. Sippl. 2008. “High-Performance Signal Peptide Prediction Based on Sequence

Alignment Techniques." *Bioinformatics* 24 (19): 2172–76. doi:10.1093/bioinformatics/btn422.

Gomez, Sandra, Laura Adalid-Peralta, Hector Palafox-Fonseca, Vito Adrian Cantu-Robles, Xavier Soberón, Edda Sciutto, Gladis Fragoso, et al. 2015. "Genome Analysis of Excretory/Secretory Proteins in Taenia Solium Reveals Their Abundance of Antigenic Regions (AAR)." *Scientific Reports* 5 (May): 9683. doi:10.1038/srep09683.

Greenbaum, D, N M Luscombe, R Jansen, J Qian, and M Gerstein. 2001. "Interrelating Different Types of Genomic Data, from Proteome to Secretome: 'Oming in on Function." *Genome Research* 11 (9): 1463–68. doi:10.1101/gr.207401.

Gregory, W F, and R M Maizels. 2008. "Cystatins from Filarial Parasites: Evolution, Adaptation and Function in the Host-Parasite Relationship." *Int J Biochem Cell Biol* 40 (6–7): 1389–98. doi:10.1016/j.biocel.2007.11.012.

Hartmann, S, B Kyewski, B Sonnenburg, and R Lucius. 1997. "A Filarial Cysteine Protease Inhibitor down-Regulates T Cell Proliferation and Enhances Interleukin-10 Production." *Eur J Immunol* 27 (9): 2253–60. doi:10.1002/eji.1830270920.

Heizer, Esley, Dante S Zarlenga, Bruce Rosa, Xin Gao, Robin B Gasser, Jessie De Graef, Peter Geldhof, and Makedonka Mitreva. 2013. "Transcriptome Analyses Reveal Protein and Domain Families That Delineate Stage-Related Development in the Economically Important Parasitic Nematodes, Ostertagia Ostertagi and Cooperia Oncophora." *BMC Genomics* 14 (1): 118. doi:10.1186/1471-2164-14-118.

Hewitson, J P, J R Grainger, and R M Maizels. 2009. "Helminth Immunoregulation: The Role of Parasite Secreted Proteins in Modulating Host Immunity." *Mol Biochem Parasitol* 167 (1): 1–11. doi:10.1016/j.molbiopara.2009.04.008.

Hotez, P.J., P.J. Brindley, J.M. Bethony, C.H. King, E.J. Pearce, and Julie Jacobson. 2008. "Helminth Infections: The Great Neglected Tropical Diseases." *The Journal of Clinical Investigation* 118 (4): 1311–21. doi:10.1172/JCI34261.tion.

Hotez, Peter J., Miriam Alvarado, María-Gloria Basáñez, Ian Bolliger, Rupert Bourne, Michel Boussinesq, Simon J. Brooker, et al. 2014. "The Global Burden of Disease Study 2010: Interpretation and Implications for the Neglected Tropical Diseases." Edited by Nilanthi de Silva. *PLoS Neglected Tropical Diseases* 8 (7): e2865. doi:10.1371/journal.pntd.0002865.

Jackson, Andrew P., Thomas D. Otto, Martin Aslett, Stuart D. Armstrong, Frederic Bringaud, Alexander Schlacht, Catherine Hartley, et al. 2016. "Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism." *Current Biology* 26 (2): 161–72. doi:10.1016/j.cub.2015.11.055.

Khan, Adnan R., and Padraic G. Fallon. 2013. "Helminth Therapies: Translating the Unknown Unknowns to Known Knowns." *International Journal for Parasitology* 43 (3–4): 293–99. doi:10.1016/j.ijpara.2012.12.002.

Khaznadji, E, P Collins, J P Dalton, Y Bigot, and N Moire. 2005. "A New Multi-Domain Member of the Cystatin Superfamily Expressed by Fasciola Hepatica." *Int J Parasitol* 35 (10): 1115–25. doi:10.1016/j.ijpara.2005.05.001.

Krijger, Jorrit-Jan, Michael R Thon, Holger B Deising, and Stefan G R Wirsel. 2014. "Compositions of Fungal Secretomes Indicate a Greater Impact of Phylogenetic History than Lifestyle Adaptation." *BMC Genomics* 15: 722. doi:10.1186/1471-2164-15-722.

Lai, Jhih-Siang, Cheng-Wei Cheng, Ting-Yi Sung, and Wen-Lian Hsu. 2012. "Computational Comparative Study of Tuberculosis Proteomes Using a Model Learned from Signal Peptide Structures." Edited by Bin Xue. *PLoS ONE* 7 (4): e35018. doi:10.1371/journal.pone.0035018.

Lehmann, Andreas. 2008. "Ecallantide (DX-88), a Plasma Kallikrein Inhibitor for the Treatment of Hereditary Angioedema and the Prevention of Blood Loss in on-Pump Cardiothoracic Surgery." *Expert Opinion on Biological Therapy* 8 (8): 1187–99. doi:10.1517/14712598.8.8.1187.

Liang, You Sheng, Jian Rong Dai, Yin Chang Zhu, Gerald C. Coles, and Michael J. Doenhoff. 2003. "Genetic Analysis of Praziquantel Resistance in Schistosoma Mansoni." *Southeast Asian Journal of Tropical Medicine and Public Health* 34 (2): 274–80.

Lustigman, S, B Brotman, T Huima, A M Prince, and J H McKerrow. 1992. “Molecular Cloning and Characterization of Onchocystatin, a Cysteine Proteinase Inhibitor of Onchocerca Volvulus.” *J Biol Chem* 267 (24): 17339–46. http://www.ncbi.nlm.nih.gov/pubmed/1512269.

Lustigman, Sara, Roger K. Prichard, Andrea Gazzinelli, Warwick N. Grant, Boakye A. Boatin, James S. McCarthy, and María Gloria Basáñez. 2012. “A Research Agenda for Helminth Diseases of Humans: The Problem of Helminthiases.” *PLoS Neglected Tropical Diseases*. doi:10.1371/journal.pntd.0001582.

Maizels, Rick M, and Maria Yazdanbakhsh. 2003. “Immune Regulation by Helminth Parasites: Cellular and Molecular Mechanisms.” *Nature Reviews. Immunology* 3 (9): 733–44. doi:10.1038/nri1183.

Manoury, B, W F Gregory, R M Maizels, and C Watts. 2001. “Bm-CPI-2, a Cystatin Homolog Secreted by the Filarial Parasite Brugia Malayi, Inhibits Class II MHC-Restricted Antigen Processing.” *Curr Biol* 11 (6): 447–51. http://www.ncbi.nlm.nih.gov/pubmed/11301256.

Marcilla, Antonio, María Trelis, Alba Cortés, Javier Sotillo, Fernando Cantalapiedra, María Teresa Minguez, María Luz Valero, et al. 2012. “Extracellular Vesicles from Parasitic Helminths Contain Specific Excretory/secretory Proteins and Are Internalized in Intestinal Host Cells.” *PloS One* 7 (9): e45974. doi:10.1371/journal.pone.0045974.

Mehta, Angela, Ana C M Brasileiro, Djair S L Souza, Eduardo Romano, Magnólia A. Campos, Maria F. Grossi-De-Sá, Marília S. Silva, et al. 2008. “Plant-Pathogen Interactions: What Is Proteomics Telling Us?” *FEBS Journal*. doi:10.1111/j.1742-4658.2008.06528.x.

Melman, Sandra D., Michelle L. Steinauer, Charles Cunningham, Laura S. Kubatko, Ibrahim N. Mwangi, Nirvana Barker Wynn, Martin W. Mutuku, et al. 2009. “Reduced Susceptibility to Praziquantel among Naturally Occurring Kenyan Isolates of Schistosoma Mansoni.” *PLoS Neglected Tropical Diseases* 3 (8). doi:10.1371/journal.pntd.0000504.

Merrifield, Maureen, Peter J. Hotez, Coreen M. Beaumier, Portia Gillespie, Ulrich Strych, Tara Hayward, and Maria Elena Bottazzi. 2016. “Advancing a Vaccine to Prevent Human Schistosomiasis.” *Vaccine* 34 (26): 2988–91. doi:10.1016/j.vaccine.2016.03.079.

Messih, Mario Abdel, Meghana Chitale, Vladimir B. Bajic, Daisuke Kihara, and Xin Gao. 2012. “Protein Domain Recurrence and Order Can Enhance Prediction of Protein Functions.” *Bioinformatics* 28 (18). doi:10.1093/bioinformatics/bts398.

Mishra, P K, N Patel, W Wu, D Bleich, and W C Gause. 2013. “Prevention of Type 1 Diabetes through Infection with an Intestinal Nematode Parasite Requires IL-10 in the Absence of a Th2-Type Response.” *Mucosal Immunology* 6 (2): 297–308. doi:10.1038/mi.2012.71.

Mitreva, M. 2012. “The Genome of a Blood Fluke Associated with Human Cancer.” *Nat Genet* 44 (2): 116–18. doi:10.1038/ng.1082.

Mourao, M M, C Grunau, P T LoVerde, M K Jones, and G Oliveira. 2012. “Recent Advances in Schistosoma Genomics.” *Parasite Immunol* 34 (2–3): 151–62. doi:10.1111/j.1365-3024.2011.01349.x.

Nahum, Laila A., Marina M. Mourão, and Guilherme Oliveira. 2012. “New Frontiers in Schistosoma Genomics and Transcriptomics.” *Journal of Parasitology Research*. doi:10.1155/2012/849132.

Nahum Laila, Pereira Sergio. 2008. “Phylogenomics, Protein Family Evolution, and the Tree of Life: An Integrated Approach between Molecular Evolution and Computational Intelligence.” In *Studies in Computational Intelligence*, 259–279.

Nielsen, H, J Engelbrecht, S Brunak, and G von Heijne. 1997. “Identification of Prokaryotic and Eukaryotic Signal Peptides and Prediction of Their Cleavage Sites.” *Protein Engineering* 10 (1): 1–6. http://www.ncbi.nlm.nih.gov/pubmed/9051728.

Nielsen, H, and A Krogh. 1998. “Prediction of Signal Peptides and Signal Anchors by a Hidden Markov Model.” *Proceedings. International Conference on Intelligent Systems for Molecular Biology* 6: 122–30. http://www.ncbi.nlm.nih.gov/pubmed/9783217.

Nombela, César, Concha Gil, and W. LaJean Chaffin. 2006. “Non-Conventional Protein Secretion in Yeast.” *Trends in Microbiology* 14 (1): 15–21. doi:10.1016/j.tim.2005.11.009.

Osada, Yoshio, Sohsuke Yamada, Atsunori Nabeshima, Yasunobu Yamagishi, Kenji Ishiwata, Susumu Nakae, Katsuko Sudo, and Tamotsu Kanazawa. 2013. "Heligmosomoides Polygyrus Infection Reduces Severity of Type 1 Diabetes Induced by Multiple Low-Dose Streptozotocin in Mice via STAT6- and IL-10-Independent Mechanisms." *Experimental Parasitology* 135 (2): 388–96. doi:10.1016/j.exppara.2013.08.003.

Petersen, Thomas Nordahl, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2011. "SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions." *Nature Methods* 8 (10): 785–86. doi:10.1038/nmeth.1701.

Saarma, U, I Jogisalu, E Moks, A Varcasia, A Lavikainen, A Oksanen, S Simsek, et al. 2009. "A Novel Phylogeny for the Genus Echinococcus, Based on Nuclear Data, Challenges Relationships Based on Mitochondrial Evidence." *Parasitology* 136 (3): 317–28. doi:10.1017/S0031182008005453.

Salinas-Carmona, Mario C., Guadalupe de la Cruz-Galicia, Isabel Pérez-Rivera, Juan M. Solís-Soto, Juan C. Segoviano-Ramirez, Anna Velia Vázquez, and Mario A. Garza. 2009. "Spontaneous Arthritis in MRL/ *Lpr* Mice Is Aggravated by *Staphylococcus Aureus* and Ameliorated by *Nippostrongylus Brasiliensis* Infections." *Autoimmunity* 42 (1): 25–32. doi:10.1080/08916930802228290.

Schierack, P, R Lucius, B Sonnenburg, K Schilling, and S Hartmann. 2003. "Parasite-Specific Immunomodulatory Functions of Filarial Cystatin." *Infect Immun* 71 (5): 2422–29. http://www.ncbi.nlm.nih.gov/pubmed/12704112.

Shah, R., Y. Lu, C. C. Hinkle, F. C. McGillicuddy, R. Kim, S. Hannenhalli, T. P. Cappola, et al. 2009. "Gene Profiling of Human Adipose Tissue During Evoked Inflammation In Vivo." *Diabetes* 58 (10): 2211–19. doi:10.2337/db09-0256.

Silva, L L, M Marcet-Houben, L A Nahum, A Zerlotini, T Gabaldon, and G Oliveira. 2012. "The Schistosoma Mansoni Phylome: Using Evolutionary Genomics to Gain Insight into a Parasite's Biology." *BMC Genomics* 13: 617. doi:10.1186/1471-2164-13-617.

Silva, L L, M Marcet-Houben, A Zerlotini, T Gabaldon, G Oliveira, and L A Nahum. 2011. "Evolutionary Histories of Expanded Peptidase Families in Schistosoma Mansoni." *Mem Inst Oswaldo Cruz* 106 (7): 864–77. http://www.ncbi.nlm.nih.gov/pubmed/22124560.

Sjolander, K. 2010. "Getting Started in Structural Phylogenomics." *PLoS Comput Biol* 6 (1): e1000621. doi:10.1371/journal.pcbi.1000621.

Soblik, H, A E Younis, M Mitreva, B Y Renard, M Kirchner, F Geisinger, H Steen, and N W Brattig. 2011. "Life Cycle Stage-Resolved Proteomic Analysis of the Excretome/secretome from Strongyloides Ratti--Identification of Stage-Specific Proteases." *Mol Cell Proteomics* 10 (12): M111 010157. doi:10.1074/mcp.M111.010157.

Tjalsma, H, A Bolhuis, J D Jongbloed, S Bron, and J M van Dijl. 2000. "Signal Peptide-Dependent Protein Transport in Bacillus Subtilis: A Genome-Based Survey of the Secretome." *Microbiology and Molecular Biology Reviews : MMBR* 64 (3): 515–47. doi:10.1128/MMBR.64.3.515-547.2000.

Tort, J, P J Brindley, D Knox, K H Wolfe, and J P Dalton. 1999. "Proteinases and Associated Genes of Parasitic Helminths." *Adv Parasitol* 43: 161–266. http://www.ncbi.nlm.nih.gov/pubmed/10214692.

Torto, T A, S Li, A Styer, E Huitema, A Testa, N A Gow, P van West, and S Kamoun. 2003. "EST Mining and Functional Expression Assays Identify Extracellular Effector Proteins from the Plant Pathogen Phytophthora." *Genome Res* 13 (7): 1675–85. doi:10.1101/gr.910003.

Tsai, Isheng J, Magdalena Zarowiecki, Nancy Holroyd, Alejandro Garciarrubio, Alejandro Sanchez-Flores, Karen L Brooks, Alan Tracey, et al. 2013. "The Genomes of Four Tapeworm Species Reveal Adaptations to Parasitism." *Nature* 496 (7443): 57–63. doi:10.1038/nature12031.

von Heijne, G. 1986. "A New Method for Predicting Signal Sequence Cleavage Sites." *Nucleic Acids Research* 14 (11): 4683–90. http://www.ncbi.nlm.nih.gov/pubmed/3714490.

WEINSTOCK, J. V. 2006. "Helminths and Mucosal Immune Modulation." *Annals of the New York*

Academy of Sciences 1072 (1): 356–64. doi:10.1196/annals.1326.033.

Wilson, Mark S., Matthew D. Taylor, Adam Balic, Constance A.M. Finney, Jonathan R. Lamb, and Rick M. Maizels. 2005. "Suppression of Allergic Airway Inflammation by Helminth-Induced Regulatory T Cells." *The Journal of Experimental Medicine* 202 (9): 1199–1212. doi:10.1084/jem.20042572.

Young, N D, A R Jex, B Li, S Liu, L Yang, Z Xiong, Y Li, et al. 2012. "Whole-Genome Sequence of Schistosoma Haematobium." *Nat Genet* 44 (2): 221–25. doi:10.1038/ng.1065.

Zarowiecki, Magdalena, and Matt Berriman. 2015. "What Helminth Genomes Have Taught Us about Parasite Evolution." *Parasitology* 142 Suppl (S1): S85-97. doi:10.1017/S0031182014001449.

Zhou, Yan, Huajun Zheng, Yangyi Chen, Lei Zhang, Kai Wang, Jing Guo, Zhen Huang, et al. 2009. "The Schistosoma Japonicum Genome Reveals Features of Host–parasite Interplay." *Nature* 460 (7253): 345–51. doi:10.1038/nature08140.

VII – ANEXOS

Além dos trabalhos já mencionados, atuo em outros estudos. Uma síntese dos projetos em andamento encontra-se a seguir.

7.1 - An integrative approach to unravel the human-*Schistosoma mansoni* interactome: who, when and where.

Cuesta-Astroz Y[1, 2], Santos A[3], Juhl Jensen L[3], Oliveira G[1].

[1] Grupo de Genômica e Biologia Computacional, Centro de Pesquisas René Rachou (CpqRR), Fundaçao Oswaldo Cruz (FIOCRUZ), Belo Horizonte, Brazil.
[2] Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil.
[3]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhaguen, 2200 Copenhagen, Denmark.

Keywords: systems biology, host-pathogen interaction, *Schistosoma mansoni,* protein-protein interaction, drug targets.

The study of molecular host–parasite interactions is essential to understand parasite infection and local adaptation within the host. Recent efforts use several strategies to identify inter-species protein–protein interactions (PPIs) between the host and parasites, viruses and bacterias. Here, we investigate the inferred PPI network between human and *S. mansoni*, one of the parasites causing Schistosomiasis, a neglected tropical disease. To this end, we propose an integrative approach that gives context to the interactions according to the parasite's life cycle and subcellular localization of the proteins. We use a homology-based method to predict interactions by looking at intra-species interactions among all organisms within the closest ancestral group common to both, human and *S. mansoni* and uses conservation of interactions as a measure of confidence. Besides, we used publicly available datasets of domain-domain interactions to identify possible PPIs based on common domains. To contextualize the interactions, we limit the interactions to human membrane and extracellular proteins expressed in tissues that support the parasite's tropism (skin, blood, lung, liver and intestine). In total our approach predicted 34,586 PPIs and 1,392 after filtering, which showed crosstalk between parasite and host proteins enriched in biological process and tissue-specific secretory pathways essential in the life cycle of the parasite. An initial manual curation of some of the interactions revealed tissue-specific interactions that are also stage associated according to expression data available for *S. mansoni*. We believe that applying this systems biology approach will certainly help uncover targetable mechanisms for the therapy of Schistosomiasis, and also opens the possibility for the analyses of any host-parasite pair.

7.2 - Whole genome analysis of *Biomphalaria glabrata* (Lophotrochozoa), a snail intermediate host for transmission of schistosomiasis.

Biomphalaria glabrata Genome Consortium

Freshwater snails of the genus *Biomphalaria* are important snail hosts for the widespread transmission of schistosomiasis in humans. Schistosomiasis is a debilitating parasitic disease with high morbidity and high loss of Disability-Adjusted Life Year (DALYs), ranking second among neglected tropical diseases only to malaria in its negative impact on global human health. Here we characterize the genome of *B. glabrata* and describe a variety of biological properties that enable their persistence in complex aquatic environments and define this snail as suitable hosts for *S. mansoni*, including aspects of immunity and gene regulation. These efforts may improve snail management and perhaps foster developments to successfully interrupt snail-mediated parasite transmission in support of schistosomiasis eradication.

My contribution to this project was the *in silico* prediction of secreted proteins. Secreted proteins were also clustered based on RNAseq expression patterns in 12 snail tissues. We identified 583 secreted proteins corresponding to 4.1% of the *B. glabrata* proteome. The secretomes included different functional classes related with different process such as digestive enzymes, protease inhibitors, lectins, kinases and antioxidant enzymes. Secreted proteins were mapped onto KEGG pathways revealing a great deal of functional diversity. Most secreted proteins belonged to categories as carbohydrate metabolism, glycan biosynthesis, folding sorting and degradation, signal transduction and transport and catabolism. In order to have a dynamic view of the secretomes, all the RNAseq data for the 12 tissues were clustered according protein expression. We found a similar expression pattern of the secretomes in foot and salivary glands and a different pattern in ovotestis. Secreted proteins expression in different tissues could point to those that may be directly involved in snail-parasite interaction. Further studies of the secretomes could help to unravel relevant aspects of snail-parasite interactions that are essential for combating infection.