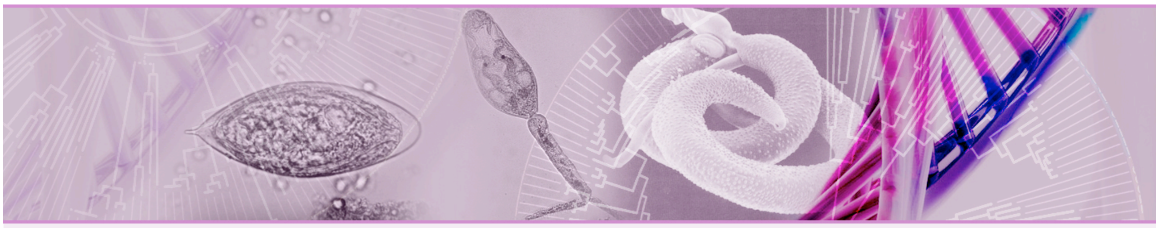
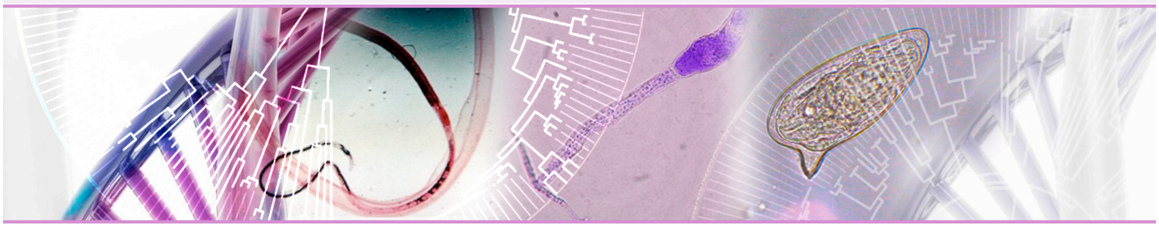
UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA GERAL
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

Área de Concentração: Biotecnologia, Genômica e Bioinformática.

TESE DE DOUTORADO



GENÔMICA COMPARATIVA DE *Schistosoma mansoni*: BIODIVERSIDADE MOLECULAR À LUZ DA EVOLUÇÃO



Orientada: Larissa Lopes Silva
Orientador: Dr. Guilherme Corrêa de Oliveira
Coorientadora: Dra. Laila Alves Nahum

Belo Horizonte
2013

Larissa Lopes Silva

GENÔMICA COMPARATIVA DE *Schistosoma mansoni*: BIODIVERSIDADE MOLECULAR À LUZ DA EVOLUÇÃO

Tese de doutorado apresentada ao Programa de Pós-Graduação em Genética do Departamento de Biologia Geral do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Genética.

Orientador: Dr. Guilherme Corrêa de Oliveira
Coorientadora: Dra. Laila Alves Nahum

Belo Horizonte
2013

Silva, Larissa Lopes.
Genômica comparativa de *Schistosoma mansoni*: biodiversidade molecular à luz da evolução. [manuscrito] / Larissa Lopes Silva. – 2013.
114 f. : il. ; 29,5 cm.

Orientador: Guilherme Corrêa de Oliveira. Coorientadora: Laila Alves Nahum.

Tese (doutorado) – Universidade Federal de Minas Gerais, Departamento de Biologia Geral.

1. Schistosoma mansoni - Teses. 2. Esquistossomose - Teses. 3. Parasito – Teses. 4. Evolução molecular – Teses. 5. Genética – Teses. 6. Filogenômica. 7. Genômica comparativa. I. Oliveira, Guilherme Corrêa de. II. Nahum, Laila Alves. III. Universidade Federal de Minas Gerais. Departamento de Biologia Geral. IV. Título.

CDU: 575.852

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

GENÔMICA COMPARATIVA DE *Schistosoma mansoni*: BIODIVERSIDADE MOLECULAR À LUZ DA EVOLUÇÃO

por

Larissa Lopes Silva

Este trabalho foi realizado no grupo de Genômica e Biologia Computacional do Centro de Pesquisas René Rachou[1] e no Centro de Excelência em Bioinformática[2] sob orientação do Dr. Guilherme Corrêa de Oliveira e coorientação da Dra. Laila Alves Nahum.

1 - Grupo de Genômica e Biologia Computacional, Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz – FIOCRUZ, Belo Horizonte, MG, Brasil.
2 - Centro de Excelência em Bioinformática, FIOCRUZ, Belo Horizonte, MG, Brasil.

Belo Horizonte

2013

AGRADECIMENTOS

Agradeço ao meu orientador Dr. Guilherme Oliveira, por inúmeras oportunidades a mim concedidas. Cada uma delas, de forma muito especial, é parte estruturante da minha formação. Agradeço também pela confiança, paciência, pelo convívio amistoso com esta pessoa gentil, bem humorada, extremamente educada e tão otimista que nos faz acreditar que o impossível pode ser realizado.

À minha co-orientadora Dra. Laila Nahum, pela orientação, conselhos e conhecimento compartilhado.

Ao Programa de Pós-Graduação em Genética pela oportunidade que tanto contribuiu para minha formação.

Aos amigos do CEBIO e Grupo de Genômica e Biologia Computacional do CPqRR, Yesid, Juliana, Izinara, Laura, Fabiano, Sara, Fernanda, Éric, Luiz, Francislon, Márcia, Fernanda Ludolf, Marina, Rosiane, Flávio, Anna, Elisângela e Aristóteles pela ajuda incalculável e por tornarem esta jornada súper agradável. Em especial, agradeço aos queridos Ângela, Eliane, Anderson, Mari Oliveira, Mari Magalhães. Vocês são meus anjos da guarda no CEBIO e como todo bom anjo da guarda, são sinônimo de competência e disponibilidade em ajudar.

Ao Omar Carvalho e à Roberta Caldeira do Laboratório de Helmintologia e Malacologia Médica (LHMM) do Centro de Pesquisas René Rachou (CPqRR), responsáveis pelo meu ingresso na ciência.

Aos amigos queridos do LHMM, Cristiano Massara e Cristiane Lafetá.

Aos amigos do laboratório do Prof. Walter Salzburger do Instituto de Biologia Evolutiva da Universidade de Basel, em especial ao Hugo, Emília, Laura, Nico, Brita, Bernd, Yuri e Zuzana.

Aos amigos de Basel, Paola, Sara, Isa e Aurélio.

Às famílias Lopes e Silva, por tudo. Não consigo descrever o tamanho do meu amor e gratidão.

Ao Rô pelo amor, paciência, companheirismo, conselhos e por tornar minha vida ainda mais feliz. Amor, esse ano é nosso!

Às famílias Scholte e Carvalho por terem me acolhido com tanto carinho.

Aos amigos Kelly e Márcio pelo carinho e apoio de sempre.

Às minhas mães Cléa e Zélia, ao meu pai Márcio, às minhas irmãs Érika, Natália e Giselle, ao meu irmão Marcinho e ao Etê, pelo incentivo constante, por me fazerem acreditar na vida, nas pessoas, no amor e por serem parte de mim e permitirem que eu seja parte de vocês.

A todos vocês meu sincero e carinhoso MUITO OBRIGADA!

“Around here, however, we don't
look backwards for very long.
We keep moving forward, opening up
new doors and doing new things,
because we're curious...
and curiosity keeps leading
us down new paths.”

Walt Disney

Dedico este trabalho ao Rô que me apresentou
o amor verdadeiro, a confiança, a plenitude;
que aceita minha alma em eterna infância
e considera natural minha alegria de viver;
e que como um bom fã da Disney me
ensinou que "*If you can dream it,
you can do it*. Walt Disney".

SUMÁRIO

ÍNDICE DE FIGURAS

RESUMO

Schistosoma mansoni é um dos três principais agentes etiológicos da esquistossomose humana, uma doença que possui altos índices de prevalência e morbidade, causando um vasto impacto socioeconômico. A esquistossomose é endêmica em 78 países onde aproximadamente 243 milhões de pessoas requerem tratamento preventivo e outras 779 milhões vivem em áreas de risco de infecção. O genoma de *S. mansoni* foi completamente sequenciado e o proteoma predito contém mais de 11.000 proteínas. Entretanto, mais de 45% permanecem sem função predita ou caracterização experimental. A disponibilidade de informações genômicas permite a aceleração do conhecimento da Biologia de Sistemas de *S. mansoni* e das relações estabelecidas com os hospedeiros, resultando em avanço no entendimento da esquistossomose e, dentro das melhores expectativas, novos métodos para controle da doença. A presente tese está centrada na aplicação de análises comparativas para abordar a complexidade das funções biológicas no nível sistêmico através da reconstrução da história evolutiva das macromoléculas codificadas no genoma de *S. mansoni*. Além disso, concentramos esforços no aprimoramento da anotação funcional do genoma deste parasito, sendo a predição baseada em homólogos com função conhecida em outros organismos. Em síntese, os quatro capítulos que estruturam este trabalho contribuem para a compreensão de alguns desafios científicos que visam responder às seguintes perguntas: Quais são os genes ganhos, perdidos ou duplicados no genoma de *S. mansoni*? Quais são os processos evolutivos que deram origem à atual biodiversidade molecular deste parasito? Como a coevolução parasito-hospedeiro resultou na possibilidade deste helminto sobreviver por anos em um ambiente potencialmente hostil da circulação sanguínea humana, protegido da ação do sistema imunológico e/ou intervindo ativamente, tornando a resposta do hospedeiro ineficaz? Quais são novos potenciais alvos terapêuticos para o desenvolvimento de drogas alternativas e vacina contra a esquistossomose? Utilizando uma abordagem multidisciplinar e integrativa, promovemos uma melhora significativa na predição funcional do proteoma predito de *S. mansoni*, contribuímos para o conhecimento das relações evolutivas do parasito em relação a outros eucariotos, identificamos famílias proteicas cuja expansão está relacionada à biologia parasitária e identificamos potenciais alvos terapêuticos que podem auxiliar o combate à esquistossomose. Dando continuidade a este trabalho, análises comparativas envolvendo dados genômicos, transcritômicos e proteômicos de outros helmintos e vetores, assim como a caracterização experimental de alvos promissores, aumentarão nosso conhecimento sobre a biologia parasitária rumo à prevenção do progresso da doença.

ABSTRACT

Schistosoma mansoni is one of the three main causative agents of human schistosomiasis, a disease with high prevalence and morbidity, causing a vast socio-economic impact. Schistosomiasis is endemic in 78 countries where 243 million people require preventive chemotherapy and other 779 million live in areas of risk of infection. The *S. mansoni* genome was completely sequenced and the predicted proteome contains over 11.000 proteins. However, more than 45% remain without a predicted function or experimental characterization. The availability of genomic information allows the acceleration of *S. mansoni* Systems Biology knowledge as well as the evolutionary relationships with hosts, resulting in improved understanding about schistosomiasis and hopefully the creation of new disease control methods. This thesis is focused on the application of the comparative analyses to address the complexity of biological functions at a systems level through the reconstruction of the evolutionary history of macromolecules encoded by *S. mansoni* genome. Furthermore, additional efforts were made in order to improve the functional annotation of *S. mansoni* proteome based on the identification of homologues with known function in other organisms. In summary, the four chapters that structure this thesis contribute to solving persisting scientific challenges that aim at understanding of: Which genes were gained, lost or duplicated in the *S. mansoni* genome? Which were the evolutionary processes that shaped this parasite's molecular biodiversity? How host-parasite co-evolution allowed this helminth to survive for years in human bloodstream, a potentially hostile environment, protected against host immune system or actively reacting making host immune response ineffective? What are potential new therapeutic targets to alternative drugs and vaccine development? Using a multidisciplinary and integrative approach, we improved the functional annotation of *S. mansoni* predicted proteome, contributed to the knowledge of evolutionary relationships of this parasite in comparison to other eukaryotes, identified expanded protein families that are related to parasite biology and highlighted potential therapeutic targets that can be used against schistosomiasis. Continuing this work, comparative analyses involving genomic, transcriptomic, and proteomic data from other helminth parasites and vectors as well as experimental characterization of promising targets will supply more information regarding parasite biology towards prevention of disease progression.

I - INTRODUÇÃO

1.1 *Schistosoma mansoni* e Esquistossomose

A esquistossomose, uma doença causada por trematódeos digenéticos que pertencem à família Schistosomatidae, é uma das doenças infecciosas e parasitárias mais prevalentes no mundo, sendo, portanto, uma das doenças tropicais mais importantes em termos de saúde pública, com número de casos inferior apenas à malária (Gryseels et al., 2006; Waknine-Grinberg et al., 2012). Cinco espécies do gênero *Schistosoma* estão envolvidas na infecção humana, sendo três os agentes principais: *Schistosoma mansoni* e *S. japonicum*, causadores da esquistossomose intestinal, e *S. haematobium*, agente etiológico da esquistossomose urinária.

Segundo a Organização Mundial de Saúde (*World Health Organization* - WHO), a esquistossomose é endêmica em 78 países onde cerca de 243 milhões de pessoas requerem quimioterapia, sendo dessas mais de 100 milhões crianças em idade escolar (WHO, 2013) (Figura 1). Outras 700 milhões de pessoas vivem em

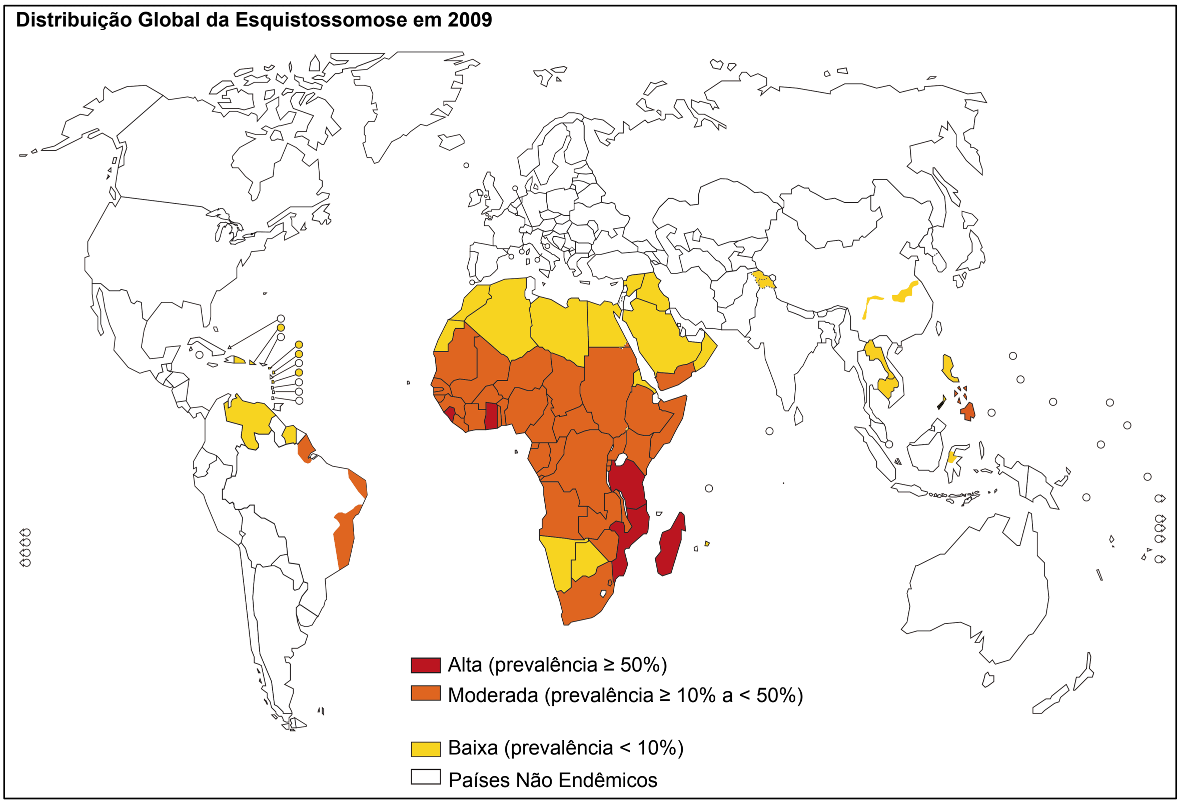


Figura 1: Distribuição global da esquistossomose (Adaptado de WHO, 2011).

áreas endêmicas sob o risco de infecção (Steinmann et al., 2006).

S. mansoni ocorre em grande parte da África subsaariana, regiões do Oriente Médio, Venezuela, Caribe e é a espécie responsável pela ocorrência da esquistossomose no Brasil (Gryseels et al., 2006). Os vermes adultos possuem coloração branca ou acinzentada, sendo que os machos medem aproximadamente 1 cm e as fêmeas 1,5 cm. Ao contrário de outros trematódeos, estes organismos são dióicos. O macho aloja e fecunda a fêmea em um canal ginecóforo constituído por dobras no sentido longitudinal. O ciclo biológico de *S. mansoni* apresenta uma alternância de gerações entre os hospedeiros intermediários, moluscos do gênero *Biomphalaria*, e os hospedeiros definitivos vertebrados, como roedores, primatas e o homem (Figura 2).

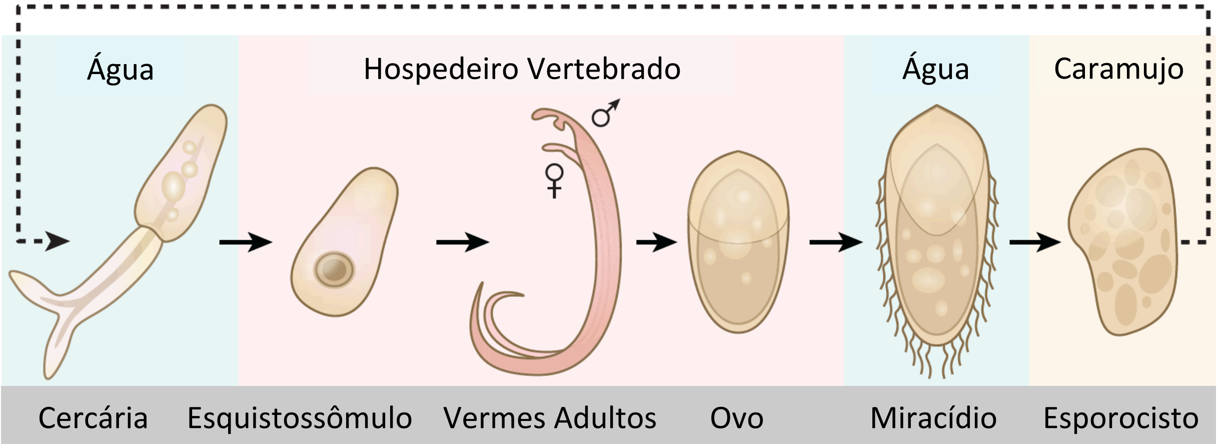


Figura 2: Estágios de desenvolvimento do ciclo de vida de parasitos do gênero *Schistosoma*. Adaptado de Han et al., 2009.

No Brasil estima-se que cerca de 25 milhões de pessoas residam em áreas endêmicas e que 4 a 6 milhões encontram-se infectadas (Lambertuci, 2010). Em 2011, um total de 26.777 pessoas receberam tratamento em 22 estados brasileiros, representando uma diminuição de 35% no número de pessoas tratadas em comparação ao ano anterior (WHO, 2013). Atualmente a esquistossomose tem registros de ocorrência em todas as regiões do brasileiras (Figura 3), sendo os estados das regiões Nordeste e Sudeste os mais afetados. A distribuição da doença está diretamente relacionada à presença de hospedeiros intermediários e hospedeiros definitivos infectados.

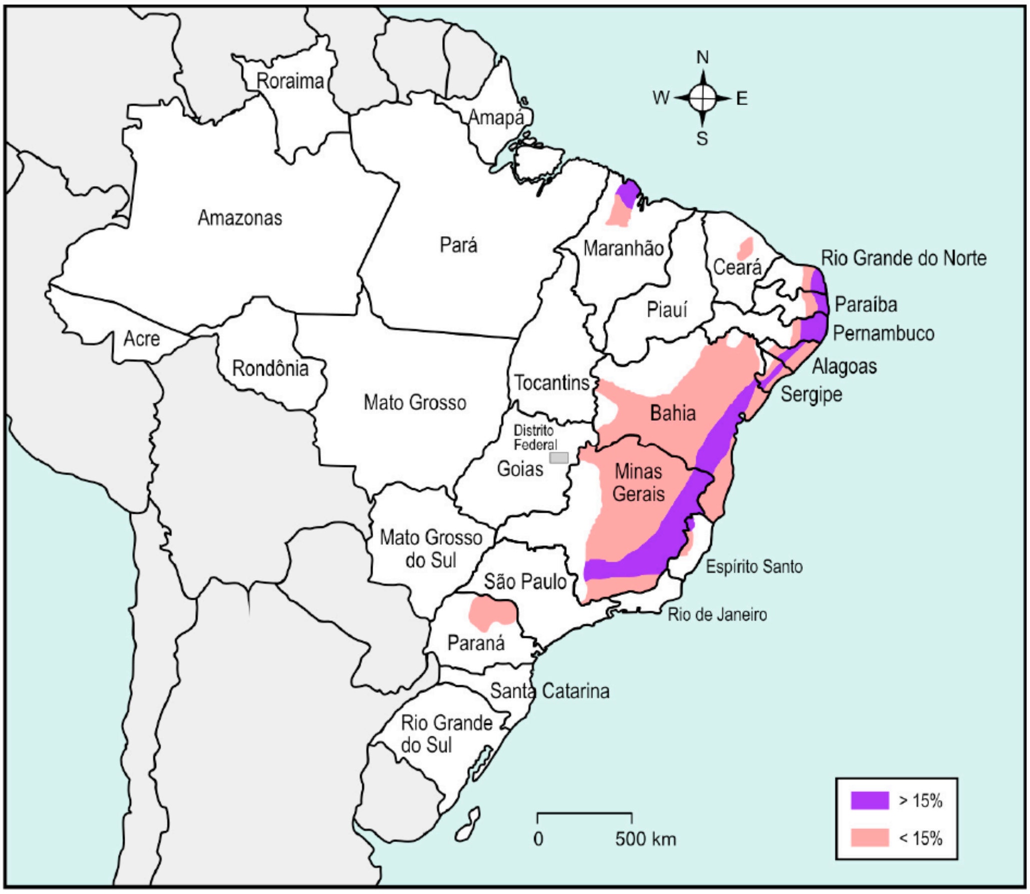


Figura 3: Distribuição da esquistossomose no Brasil. Fonte: Rokni et al., 2012.

Atualmente o controle da esquistossomose depende principalmente do tratamento de pacientes infectados com praziquantel® (Figura 4), a única droga disponível para tratamento (Greenberg, 2013). Entretanto, como as demais drogas, o praziquantel não previne contra a reinfecção e a eficácia é variável dependendo do sexo do parasito, estágios de desenvolvimento e período de infecção, além de evidências de parasitos resistentes a este medicamento terem sido obtidas em laboratório e no campo (Liang et al., 2003; Pica-Mattoccia & Cioli, 2004; Botros & Bennet, 2007; Melman et al., 2009). Em virtude do exposto, a esquistossomose continua sendo uma das doenças infecciosas e parasitárias mais prevalentes no mundo, responsável por altos índices de morbidade e mortalidade, e gerando perdas econômicas significativas, especialmente em países em desenvolvimento.

A fim de identificar novos alvos para o desenvolvimento de drogas, vacinas e diagnóstico, a comunidade científica uniu esforços para sequenciar o genoma de *S. mansoni* como parte de uma iniciativa da Organização Mundial de Saúde de apoio ao sequenciamento do genoma de parasitos (LoVerde et al., 2004).

Os dados do genoma de *S. mansoni* foram publicados em meados de 2009 e uma nova versão foi recentemente disponibilizada totalizando 364,5 megabases (Mb) distribuídas em 885 *scaffolds* (supercontigs), metade dos quais possuem tamanho maior que 2 Kilobases (Kb) (Berriman et al., 2009; Protasio et al., 2012). Aproximadamente 81% do tamanho estimado do genoma foi ordenado em cromossomos constituindo o mapa genético de *S. mansoni* (Criscione et al., 2009; Protasio et al., 2012). Um total de 10.852 genes foram identificados, codificando mais de 11.000 proteínas, 45% das quais permanecem sem anotação funcional (Protasio et al., 2012; Zerlotini et al., 2009).

Embora o conhecimento da função gênica seja de extrema importância para a compreensão dos processos biológicos, cerca de 30-35% dos genes na maioria dos organismos permanece sem caracterização experimental ou predita. Além disso, grande parte das demais sequências codificantes possuem apenas uma anotação funcional muito genérica (Galperin & Koonin, 2000). Por este motivo, pesquisas têm sido desenvolvidas com o objetivo de anotar automaticamente as funções proteicas em resposta à alta taxa de geração de sequências em comparação com a taxa de caracterização experimental (Galperin & Koonin, 2000; Gabaldón & Huynen, 2004; Sjölander, 2004; Hawkins & Kihara, 2007; Nahum & Pereira, 2008; Gabaldón, 2008; Jiang, 2008; Andrade et al., 2011).

Na tentativa de aumentar a acurácia da predição funcional em larga escala e identificar inovações evolutivas relacionadas à adaptação de organismos aos mais diversos ambientes, cientistas têm associado métodos filogenéticos aos estudos genômicos, tirando proveito do poder que uma perspectiva evolutiva fornece às análises biológicas comparativas. Este novo campo de atuação tem provido informações importantes no que diz respeito à predição funcional dos genes, bem como sobre a história evolutiva dos genomas e organismos, tendo se mostrado uma alternativa eficaz para otimizar o processo de predição funcional e evitar a propagação de erros associados à predição baseada apenas em similaridade de sequências (Eisen, 1998; Eisen & Wu 2002; Nahum et al. 2009; Frickey & Lupas, 2004; Brown & Sjolander, 2006; Nahum & Pereira, 2008; Huerta-Cepas *et al.*, 2010).

A ampla quantidade de dados genômicos, transcritômicos e proteômicos, associados ao desenvolvimento de poder computacional, devem contribuir para um melhor entendimento acerca da evolução e desenvolvimento de parasitos do gênero *Schistosoma*, bem como da sua interação com os hospedeiros definitivo e intermediário, evasão do sistema imune e patogênese da esquistossomose.

1.2 Genômica Comparativa

A genômica comparativa permite estudar a relação entre sequência, estrutura e função dos genomas de diferentes espécimes, espécies e/ou linhagens biológicas. Esta abordagem tem como objetivo identificar as assinaturas da seleção natural para compreender os processos evolutivos que atuam sobre a biodiversidade molecular no nível genômico (Touchman, 2010). Segundo Hardison (2003), a informação genética por si só não explica a diversidade fenotípica e uma das principais contribuições da genômica comparativa seria encontrar as partes funcionais dos dados genômicos para promover a saúde dos indivíduos e da sociedade.

Análises de genômica comparativa podem ser utilizadas para responder diferentes perguntas biológicas como: Quais características ou informações genômicas estão relacionadas à biologia dos organismos (e.g. ecologia, estilo de vida)? Qual é o conjunto de proteínas compartilhadas pelos metazoários? Quais processos modelam a evolução dos genomas? Quais genes/proteínas estão sob seleção purificadora, diversificadora, positiva, negativa ou neutra? Como resultado, estudos comparativos podem fornecer informações sobre os processos que modelaram a arquitetura genômica em parasitos e organismos de vida livre, permitindo a identificação de potenciais alvos para o desenvolvimento de novas drogas e vacina, assim como genes/proteínas diretamente relacionados à biologia parasitária (Swain et al., 2012; Tsai et al., 2013).

A comparação de dados genômicos pode ser dividida em três principais tipos de análises que envolvem: i) estrutura genômica - incluindo repetições, rearranjos, sintenia, pontos de quebra e conteúdo GC; ii) regiões codificantes, conteúdo proteico e relações de homologia (e.g. ortologia e paralogia); iii) regiões não codificantes - incluindo o estudo de elementos regulatórios (Wei et al., 2002). Portanto, além de contribuir para a identificação de regiões funcionais, estudos de genômica comparativa também contribuem para a identificação de outras classes importantes como RNAs não codificantes e elementos regulatórios (Touchman, 2010). Dentre as ferramentas computacionais utilizadas para tais análises destacam-se os algoritmos para alinhamento de sequências moleculares, reconstrução de filogenia em larga escala, identificação de genes e predição de função baseada ou não em homologia (e.g. Brown & Sjolander, 2006; Huerta-Cepas et al., 2007; Huerta-Cepas et al., 2010).

É claro e bem estabelecido que vários aspectos da biologia comparativa podem se beneficiar da associação com estudos evolutivos (e.g. Felsenstein, 1985; Eisen & Wu 2002). A filogenômica, estudo das relações evolutivas baseado em análises genômicas comparativas, tem o potencial de analisar novos dados utilizando uma abordagem filogenética para responder questões fundamentais como a origem e evolução do parasitismo.

O genoma mitocondrial e nuclear de algumas espécies do gênero *Schistosoma* foram completamente sequenciados abrindo novas fronteiras para análises comparativas (Nahum et al., 2012; Swain et al., 2012). A disponibilidade desses dados tem impacto ainda maior na compreensão da história evolutiva de agentes etiológicos da esquistossomose devido à ausência de registros fósseis.

Análises comparativas de *Schistosoma* vêm contribuindo não só para o entendimento sobre divergência, especiação e biologia destes vermes, mas também para o conhecimento das rotas migratórias destes parasitas. Um exemplo das contribuições oriundas deste tipo de estudo pode ser obtida no trabalho de Lawton e colaboradores (2012). Segundo Lawton, análises genômicas comparativas sugerem que parasitos do gênero *Schistosoma* se originaram na Ásia com posterior migração para a África, em conjunto com a migração e dispersão de hospedeiros intermediários e definitivos. Este movimento foi subsequentemente seguido por uma radiação de espécies *Schistosoma*, dando origem aos grupos de *S. mansoni* e *S. haematobium*, bem como ao grupo de *S. indicum* que regressou à Ásia mais tarde. Cada um desses grandes eventos evolutivos foram marcados por mudanças distintas na estrutura genômica desses organismos.

Na presente tese, utilizamos distintas abordagens computacionais para realizar análises genômicas comparativas. Para investigar as regiões codificantes de *S. mansoni*, em comparação com outros metazoários, utilizamos a abordagem filogenômica com o intuito de contribuir para o conhecimento da evolução do parasitismo e identificar eventos potencialmente relacionados à biologia parasitária.

1.3 Filogenômica

O termo filogenômica, interseção entre filogenética e genômica, foi criado no final da década de noventa pelo pesquisador Jonathan Eisen com o objetivo de otimizar a predição funcional gênica em escala genômica (Eisen et al., 1997). Nos últimos anos, o crescente número de genomas sequenciados aliados ao desenvolvimento do poder computacional e novas ferramentas de Bioinformática, abriram caminho para as análises comparativas envolvendo diversos grupos taxonômicos (Medina, 2005; Huerta-Cepas et al., 2007). Estas análises permitiram a integração da informação genotípica (sequência, estrutura e função de genes e seus respectivos produtos) à diversidade fenotípica dos organismos contribuindo para a compreensão dos processos que moldam a evolução dos genes, genomas, organismos e comunidades dos mais diversos ambientes (Nahum & Pereira, 2008; Huerta-Cepas et al., 2007).

Neste contexto, a filogenômica surgiu como uma plataforma integrativa que permite reconstruir a história evolutiva de processos biológicos complexos à partir de dados genômicos, além de possibilitar a predição funcional acurada, fornecendo elementos para a priorização de certos alvos no delineamento experimental (Eisen, 1998; Brown & Sjolander, 2006; Nahum & Pereira, 2008). A disponibilização de informações resultantes da reconstrução da história evolutiva à partir de dados genômicos, transcritômicos e proteômicos fornece à comunidade científica uma ampla visão da evolução da arquitetura dos genomas sob a perspectiva de todos os componentes do repertório gênico ou proteico de certa espécie, além de propiciar acesso ao conhecimento em diferentes níveis de divergência da árvore da vida (e.g. Medina, 2005; Huerta-Cepas et al., 2007).

Estas análises, que integram todos os elementos de um sistema biológico, em contraste com estudos isolados, deram origem à abordagem chamada Biologia de Sistemas (Hood, 2003). Esta abordagem possui especial relevância no contexto de genomas recém sequenciados, uma vez que através da genômica ou proteômica comparativa é possível realizar a predição funcional baseada em relações de ortologia (e.g. Huerta-Cepas et al., 2010). Assim, a correlação entre Biologia Evolutiva, Biologia de Sistemas e Genômica surge para solucionar questões fundamentais em Biologia: a função dos sistemas biológicos complexos e a compreensão da evolução da biodiversidade no nível molecular.

O presente projeto está centrado na aplicação da filogenômica para abordar a complexidade das funções biológicas no nível sistêmico. Para atingir nossos objetivos, adotamos uma abordagem filogenômica para reconstruir a história evolutiva do proteoma predito de *S. mansoni* em comparação com outros eucariotos, parasitos e de vida livre, a fim de contribuir para um melhor entendimento dos processos que modelaram as famílias proteicas deste parasito ao longo do tempo evolutivo, bem como seu impacto na biologia parasitária.

Utilizando esta abordagem multidisciplinar e integrativa, aprimoramos a predição funcional do proteoma predito de *S. mansoni*, contribuímos para o conhecimento das relações evolutivas deste parasito em relação a outros eucariotos, identificamos famílias proteicas cuja expansão está potencialmente relacionada à biologia parasitária e identificamos potenciais alvos terapêuticos que possam auxiliar o combate à esquistossomose.

Os resultados obtidos estão estruturados em quatro capítulos que abordam relações evolutivas do genoma e proteoma predito e estudos de famílias de proteínas, incluindo potenciais alvos terapêuticos no tratamento da esquistossomose e outras helmintoses.

- Capítulo I: Filoma de *Schistosoma mansoni*. Este artigo aborda a história evolutiva de cada proteína codificada no genoma do *S. mansoni*. Para reconstruir o filoma a partir do proteoma predito do parasito utilizamos um pipeline automático. Este estudo permitiu identificar famílias de proteínas expandidas em *S. mansoni*, além de contribuir significativamente para a melhoria da anotação funcional das proteínas do parasito sem prévia caracterização.

- Capítulo II: Genômica Evolutiva de Famílias de Endopeptidases. Neste estudo, realizamos a análise de três famílias de peptidases expandidas no proteoma de *S. mansoni* em relação a outros eucariotos de distintos grupos taxonômicos. Este trabalho forneceu uma ampla visão das relações evolutivas de membros dessas famílias, da complexidade genômica linhagem-específica e adaptações potencialmente relacionadas à vida parasitária.

- Capítulo III: Filogenômica e Anotação Funcional de Proteína Quinases. Com o objetivo de identificar todas as proteína quinases eucarióticas de *S. mansoni* (*ePKinome*), realizamos uma extensa análise computacional. Este estudo permitiu a identificação e classificação do conjunto de proteínas quinase eucarióticas deste parasito, contribuiu para a anotação funcional do proteoma predito e identificou potenciais alvos para o desenvolvimento de drogas alternativas para o tratamento da esquistossomose.

- Capítulo IV: Investigação do Papel Biológico de MAPKs. Dando continuidade ao trabalho sobre proteína quinases eucarióticas, caracterizamos experimentalmente MAPKs através da interferência por RNA (RNAi), um processo de silenciamento da expressão gênica desencadeado por pequenos RNAs. Os resultados obtidos mostram que duas MAPKs (SmERK e SmJNK) são alvos de interesse para o desenvolvimento de drogas visto que um inibidor destas proteínas provavelmente irá interromper o progresso da doença.

Além dos trabalhos já mencionados, desde o ingresso no Doutorado atuo também como colaboradora em outros projetos e uma pequena síntese desses estudos é apresentada na seção anexos.

II - OBJETIVOS

2.1 - Objetivo Geral:

Analisar o proteoma predito de *S. mansoni* em relação a outros organismos utilizando métodos filogenéticos, bem como análises experimentais, visando contribuir para a compreensão da Biologia de Sistemas deste parasito.

2.2 - Objetivos Específicos:

- ✓ Analisar as relações evolutivas de *S. mansoni* em relação a outros metazoários e identificar os processos evolutivos que modelaram o proteoma deste parasito ao longo do tempo evolutivo.
- ✓ Contribuir para a anotação funcional de genes e produtos gênicos codificados pelo genoma de *S. mansoni.*
- ✓ Identificar famílias proteicas cuja expansão está relacionada à evasão do sistema imune e/ou adaptação ao hospedeiro definitivo.
- ✓ Identificar potenciais alvos terapêuticos que possam auxiliar o combate à esquistossomose.

III - CAPÍTULOS

3.1 - CAPÍTULO I: Filoma de *S. mansoni*

Silva LL, Marcet-Houben M, Nahum LA, Zerlotini A, Gabaldón T, Oliveira G. The *Schistosoma mansoni* phylome: using evolutionary genomics to gain insight into a parasite's biology. BMC Genomics. 2012 Nov 13;13(1):617. PMID: 23148687.

Neste estudo em colaboração com o grupo de Genômica Comparativa liderado pelo Dr. Toni Gabaldón (Centre for Genomic Regulation, Espanha), utilizando uma abordagem filogenômica, analisamos o proteoma predito de *S. mansoni* em comparação com outros 12 eucariotos: um coanoflagelado, dez invertebrados, e um vertebrado. Os táxons selecionados cobrem importantes inovações evolutivas, tornando-os especialmente adequados para estudar as inovações em *Schistosoma* no contexto da evolução dos metazoários. Este estudo teve como principais objetivos i) reconstruir a história evolutiva de cada proteína codificada no genoma deste parasito - filoma; ii) contribuir para o enriquecimento da anotação funcional do proteoma predito de *S. mansoni*, uma vez que mais de 45% das proteínas careciam de anotação funcional. O filoma resultante possui 7.964 árvores filogenéticas obtidas a partir da análise comparativa de 11.763 proteínas de *S. mansoni* com os potenciais homólogos nos outros 12 organismos. Utilizando a abordagem filogenética para a predição de ortologia foi possível transferir anotação funcional (termos GO) para 3.451 proteínas de *S. mansoni*, das quais 790 eram anotadas previamente como “proteínas hipotéticas” que correspondem a proteínas cuja função é completamente desconhecida. Além de promover uma melhora significativa na predição funcional do proteoma de *S. mansoni*, os resultados desta análise fornecem informações importantes sobre a evolução do genoma deste parasito, como a identificação de duplicações gênicas que podem estar relacionadas a especificidades morfológicas ou fisiológicas deste organismo.

Silva *et al. BMC Genomics* 2012, **13**:617
http://www.biomedcentral.com/1471-2164/13/617

BMC Genomics

RESEARCH ARTICLE **Open Access**

The *Schistosoma mansoni* phylome: using evolutionary genomics to gain insight into a parasite's biology

Larissa Lopes Silva[1,2,3], Marina Marcet-Houben[4,5], Laila Alves Nahum[1,2,6], Adhemar Zerlotini[2,7], Toni Gabaldón[4,5] and Guilherme Oliveira[1,2*]

Abstract

Background: *Schistosoma mansoni* is one of the causative agents of schistosomiasis, a neglected tropical disease that affects about 237 million people worldwide. Despite recent efforts, we still lack a general understanding of the relevant host-parasite interactions, and the possible treatments are limited by the emergence of resistant strains and the absence of a vaccine. The *S. mansoni* genome was completely sequenced and still under continuous annotation. Nevertheless, more than 45% of the encoded proteins remain without experimental characterization or even functional prediction. To improve our knowledge regarding the biology of this parasite, we conducted a proteome-wide evolutionary analysis to provide a broad view of the *S. mansoni*'s proteome evolution and to improve its functional annotation.

Results: Using a phylogenomic approach, we reconstructed the *S. mansoni* phylome, which comprises the evolutionary histories of all parasite proteins and their homologs across 12 other organisms. The analysis of a total of 7,964 phylogenies allowed a deeper understanding of genomic complexity and evolutionary adaptations to a parasitic lifestyle. In particular, the identification of lineage-specific gene duplications pointed to the diversification of several protein families that are relevant for host-parasite interaction, including proteases, tetraspanins, fucosyltransferases, venom allergen-like proteins, and tegumental-allergen-like proteins. In addition to the evolutionary knowledge, the phylome data enabled us to automatically re-annotate 3,451 proteins through a phylogenetic-based approach rather than solely sequence similarity searches. To allow further exploitation of this valuable data, all information has been made available at PhylomeDB (http://www.phylomedb.org).

(Continued on next page)

---

* Correspondence: oliveira@cpqrr.fiocruz.br
[1]Grupo de Genômica e Biologia Computacional, Centro de Pesquisas René Rachou. Instituto Nacional de Ciência e Tecnologia em Doenças Tropicais. Fundação Oswaldo Cruz - FIOCRUZ, Belo Horizonte, MG 30190-002, Brazil
[2]Centro de Excelência em Bioinformática, Fundação Oswaldo Cruz – FIOCRUZ, Belo Horizonte, MG, Brazil
Full list of author information is available at the end of the article

BioMed Central

© 2012 Silva et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

(Continued from previous page)

Conclusions: In this study, we used an evolutionary approach to assess *S. mansoni* parasite biology, improve genome/proteome functional annotation, and provide insights into host-parasite interactions. Taking advantage of a proteome-wide perspective rather than focusing on individual proteins, we identified that this parasite has experienced specific gene duplication events, particularly affecting genes that are potentially related to the parasitic lifestyle. These innovations may be related to the mechanisms that protect *S. mansoni* against host immune responses being important adaptations for the parasite survival in a potentially hostile environment. Continuing this work, a comparative analysis involving genomic, transcriptomic, and proteomic data from other helminth parasites, other parasites, and vectors will supply more information regarding parasite's biology as well as host-parasite interactions.

Keywords: Phylogenomics, Maximum likelihood analysis, Homology prediction, Functional annotation, Paralogous families, Parasite genomics, Schistosomiasis

Background

Schistosoma mansoni, S. haematobium, and S. japonicum (Platyhelminthes: Trematoda) are the main causative agents of human schistosomiasis, a neglected tropical disease that is endemic in 77 countries where more than 237 million people require preventive chemotherapy and other 779 million live in areas of risk of infection [1-4]. The genomes of these parasites have been recently published providing insights into parasite's development, infection, and host-parasite interactions [5-7]. However, even with the progress made over the last years, schistosomiasis control depends primarily on the treatment of infected patients with Praziquantel®, the only drug available for mass treatment (e.g. [5,8,9]). Drawbacks of this drug are that it does not prevent against reinfection and its effectiveness varies depending on several factors such as the parasite's gender, developmental stage, and the time of infection. Furthermore, Praziquantel®-resistant parasites have been found both in the laboratory and in the field, thus increasing the urgent need for new effective drugs and vaccines [10-13].

Schistosoma mansoni infects 7.1 million people in America, 95% of which in Brazil, and 54 million people in Sub-Saharan Africa causing intestinal and hepatosplenic schistosomiasis [14,15]. The *S. mansoni* genome sequencing data was published in 2009 and a new version was recently released [5,16]. The improved genome has 364.5 megabases (Mb) assembled in 885 scaffolds, half of which are represented in scaffolds greater than 2 kilobases [16]. A total of 10,852 genes were identified, encoding over 11,000 proteins, 45% of which remain without known or predicted function [5,16,17]. 81% of the genome was assembled onto the parasite's chromosomes, providing a partial genetic map [16,18]. The availability of genomic data offers new opportunities for innovation in the control of schistosomiasis, by providing information that allows for the identification of novel drug targets and vaccine candidates through a system-wide perspective [5,19,20].

Making accurate functional predictions for genes or proteins is a key step in every genome sequencing project. However, on average, 30 to 50% of the predicted proteome remains uncharacterized while for the remaining set only general predictions are made. To deal with the gap between the rapid progress in genome sequencing and experimental characterization of genes and gene products, computational methods have been developed [21-23]. Two main approaches are generally used for functional prediction of genes and their products: one based on sequence similarity searches and another on phylogenetic analysis.

Owing to the computational cost and complexity of large scale phylogenetic analysis, the accurate identification of orthology relationships remains a challenge in comparative genomics and most of the orthology prediction methods rely on similarity-based search (e.g. BLAST [24], OrthoMCL [25], InParanoid [26]). In these cases, functional prediction is obtained based on the transfer of information from the most similar sequences in the database to the gene or protein of interest (e.g. [24]). However, several limitations are associated with this method, mainly the lack of a straightforward relationship between sequence similarity and protein function [21,27-29]. Since this approach is fast, simple, and can be automated to analyze thousands of genes, it has been used frequently to predict functional products encoded by newly sequenced genomes. Over the last years this practice has generated systematic errors, the extent of which is not completely known [22,27-32].

In an attempt to improve the accuracy of functional prediction at a large scale, phylogenetic methods may be applied [33,34]. The advantage of such methods is that they focus on the evolutionary history of genes rather than merely on their sequence similarity [30,35,36]. Ideally, functional transfer in the genomic context or for specific genes/proteins should be performed only when there is any experimental evidence for those used as source of information. However, in databases as UniProt, only 3% of proteins have experimental support for their annotations [28]. To deal with the absence of experimental support for most part of the available proteomes, transfer of functional annotation aiming to provide hints regarding the gene/protein function needs to follow strict requirements to avoid, as much as possible, misclassifications. In the last decade, the publication of a large number of genomic and proteomic data and the development of faster and powerful computers, new software, and automated pipelines have allowed for the reconstruction of phylogenetic trees of the complete set of proteins encoded in a genome – the so called phylome [37].

The phylome data may give a broad view of the evolution of an organism, since it comprises the phylogenies of all proteins encoded in its genome [37]. Most notably, a phylome can be used to detect specific evolutionary scenarios, to quantify the fraction of individual phylogenies whose topologies are consistent with a given hypothesis, and to improve functional annotation of proteins and biological systems [38,39]. Furthermore, comparing genomes or proteomes through an evolutionary perspective may provide insights to the understanding of the metabolism, physiology, pathogenicity, and the adaptation to a particular life style of organisms. In this context, the availability of *S. mansoni* genomic data provides the opportunity to study this parasite from a genome-wide perspective rather than from individual gene or protein analyses.

Taking advantage of the benefits provided by a genome-wide approach combined with an evolutionary perspective, we reconstructed the *S. mansoni* phylome with the goals of i) gaining insight into lineage-specific evolutionary events potentially related to the parasitic lifestyle, and ii) improving the functional annotation of the genome/proteome.

Phylogenetic techniques used in the present work included multiple sequence alignment [40-43] alignment trimming [44], neighbor-joining tree building [45], evolutionary model testing, and maximum likelihood analysis [46]. The resulting phylome data contains 7,964 protein phylogenetic trees, covering the analysis of 11,763 *S. mansoni* proteins and their homologs in 12 other organisms, out of which we identified evolutionary events and homology relationships. The results provided useful information about the parasite's genome evolution such as the identification of gene duplication events and expanded protein families such as proteases, tetraspanins, fucosyltransferases, venom allergen-like proteins (also called as SmVAL or SCP-like), tegumental-allergen-like proteins (SmTAL), among others. Altogether, the results obtained are likely to pave the way for a better understanding of the parasite's biology including host-parasite interactions. This, in turn will accelerate the search for new drugs and vaccine directed toward the control and eradication of schistosomiasis.

Results and discussion

Reconstruction of the *S. mansoni* phylome

The *S. mansoni* phylome reconstructed in this work was derived from the comparative analysis of all proteins encoded in the parasite genome (predicted proteome) and their homologs in 12 other eukaryotic proteomes whose genomes were completely sequenced (Table 1). The set of selected species is particularly rich in metazoans (11 species), including ten invertebrates, one tunicate, and one vertebrate. One choanoflagellate, *Monosiga brevicollis*, was included as outgroup of the phylogenetic reconstruction. The metazoan species selected represent important evolutionary innovations, e.g. the origin of the third germ layer, the development of organs, systems, complex patterns of communication, and the emergence of the adaptive immune system, making this dataset set especially suitable for addressing the evolutionary innovations in *S. mansoni* in the context of metazoan evolution.

To perform the phylogenetic analyses, we applied an automated pipeline similar to the one used for the human phylome project [39]. This pipeline is illustrated here (Figure 1). The resulting alignments, phylogenies, and orthology predictions can be accessed at PhylomeDB [47] (http://phylomedb.org).

Using this phylogenomic approach, we analyzed 11,763 *S. mansoni* proteins and obtained 7,964 phylogenetic trees covering 70% of the parasite's proteome. This coverage is remarkably similar to that of other phylome data of newly sequenced genomes such as that of the pea aphid *Acyrthosiphon pisum* (67%) [38].

The absence of trees for the remaining 3,490 proteins is either due to a possible high degree of divergence between the *S. mansoni* proteins and their homologs in the other selected species, an indication of the uniqueness of the parasite's proteome, or it reflects the presence of errors in gene models. Out of the 7,964 phylogenetic trees, 3.038 (38%) correspond to trees with "seed" proteins with a completely unknown function and without any GO [48] assignment in SchistoDB [17].

Phylogeny-based orthology prediction

In order to create a complete list of orthology and paralogy relationships among *S. mansoni* proteins and those encoded in the other eukaryotic proteomes included in this work, we analyzed the parasite's phylome using a *species-overlap* algorithm as previously described [39]. The comprehensive catalogue of phylogeny-based orthology and paralogy relationships among *S. mansoni* and other species was made publicly available at PhylomeDB [47].

Table 1 Proteomes selected for the *S. mansoni* phylome reconstruction

Scientific Name	UniProt Species Code[1]	TaxID[2]	Proteins[3]	Source[4]	Download
Monosiga brevicollis	MONBE	81824	9,170	JGI	2011-06-01
Ciona Intestinalis	CIOIN	7719	14,048	UniProt Reference Proteomes	2011-07-09
Nematostella vectensis	NEMVE	45351	24,424	UniProt Reference Proteomes	2011-07-09
Schistosoma haematobium	SCHHA	6185	12,767	SchistoDB	2012-03-09
Schistosoma mansoni	SCHMA	6183	11,103	SchistoDB	2012-03-09
Schistosoma japonicum	SCHJA	6182	12,636	SchistoDB	2012-03-09
Caenorhabditis elegans	CAEEL	6239	19,758	UniProt Reference Proteomes	2011-07-09
Ascaris suum	ASCSU	6253	18,430	WormBase	2012-03-09
Brugia malayi	BRUMA	6279	19,916	WormBase	2012-03-09
Trichinella spiralis	TRISP	6334	15,878	WormBase	2012-03-09
Drosphila melanogaster	DROME	7227	11,794	FlyBase	2011-09-13
Tribolium castaneum	TRICA	7070	16,533	BeetleBASE - HGSC	2011-12-16
Homo sapiens	HUMAN	9606	20,965	UniProt Reference Proteomes	2011-07-09

1 - Code assigned to each species in the *S. mansoni* phylome. 2 - Taxonomic identifier at NCBI (TaxID). 3 - Number of proteins analyzed per species. 4 - Database from which the protein data were retrieved.

Owing to the increasing rate at which new fully sequenced genomes are released, the accumulation of genomic and proteomic data has been much higher than the rates at which genes or proteins are experimentally characterized. Aiming at producing a high confidence set of functional predictions for *S. mansoni* proteins, we used the evolutionary relationships as inferred from phylogenetic trees to obtain subsets of one-to-one (single homolog in *S. mansoni* and in other species) homology relationships among *S. mansoni* proteins and the homologs from other species included in the present study (Figure 2).

By using such phylogeny-based approach, we transferred 10,175 functional annotations (GO terms [48]) to 3,451 *S. mansoni* proteins, from which 790 (7% of the parasite's proteome) were previously annotated as "hypothetical protein", corresponding to proteins whose function had not been predicted or experimentally tested before (Additional file 1 Table S1). The transfer was performed from each ortholog with known function in the selected taxa to the *S. mansoni* "seed" protein. For the other proteins that already had any functional prediction, the annotation was confirmed or improved. Consequently, a "seed" protein could receive more than one functional description. In these cases, all functional annotations were maintained allowing the user to choose the closest related transferred functional annotation, those that came from model organisms, or even to create a consensus based on all of them.

To validate the applied methodology, we retrieved reviewed *S. mansoni* proteins from UniProt [49], including experimentally confirmed ones, to evaluate the annotation transferred by the phylogenomic approach. The functional annotations performed by PhylomeDB correspond to known functions in the aforementioned database (Additional file 1 Table S2). Even though the BLAST search may detect distant homologs with additional domains, our subsequent phylogenetic reconstruction and our selection of orthologs will select those orthologs that are likely to have similar domain architecture. This is an additional reason why an orthology-based annotation is preferred over sequence similarity searches, since orthologs as compared to paralogs have a higher tendency to share a similar domain architecture [50].

Although less reliable than those based on one-to-one orthology relationships, annotation transfer based on more complex subsets (one-to-many, many-to-one, or many-to-many) may provide important hints to predict the biological function of *S. mansoni* proteins. However, in these cases, one or more genes are co-orthologous to a set of genes in another genome due to lineage-specific duplication(s) that can be associated with functional shifts, affecting the reliability of the functional transfer [38,51]. An example of a one-to-one transfer from a *Drosophila melanogaster* protein to a *S. mansoni* protein comes from the phylogenetic reconstruction of the Phy000V14T_SCHMA (Smp_170950) protein, potentially related to the glycine cleavage system, and its homologs in the selected species (Figure 3). The analysis of this tree resulted in six transfers of functional annotation from homologous proteins to the *S. mansoni* "seed" protein. The GO terms in all six functional annotations are related to aminomethyltransferase activity and glycine catabolic process providing further support for the annotation transfer. In this example, to illustrate a case of a one-to-one transfer, we chose the functional annotation transferred from *Drosophila melanogaster* once, according to the information available in UniProt [49], it is one of the orthologs with known function and experimental validation. Tags for homologous sequences with experimental validation are not available in

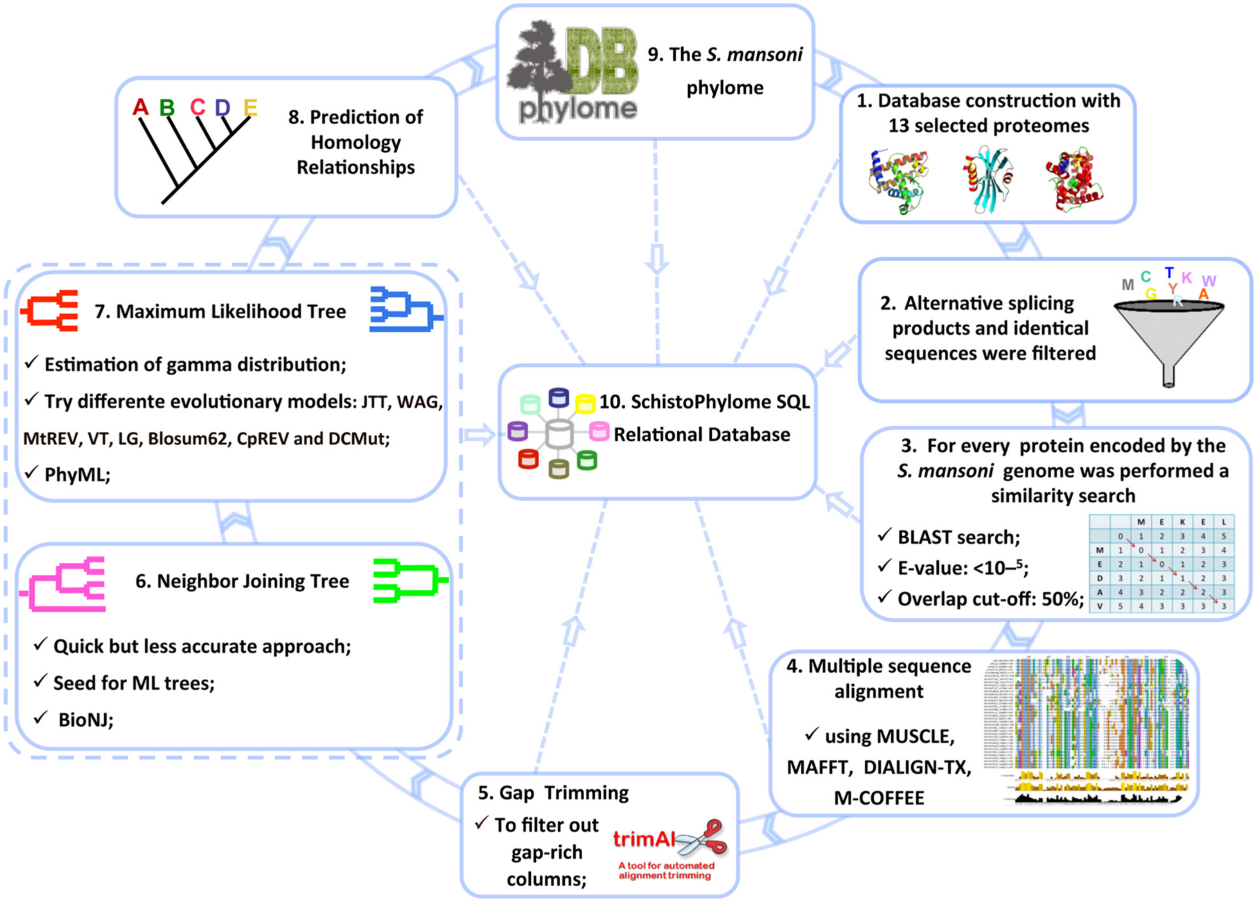


Figure 1 Pipeline used to reconstruct and analyze the *S. mansoni* phylome. Each protein sequence encoded in the parasite genome was compared against a database of proteins from other 12 fully sequenced eukaryotic proteomes (Table 1) to select putative homologous proteins. Groups of potential homologs were aligned and subsequently trimmed to remove gap-rich regions. The refined alignment was used to build a NJ tree, which was then used as a "seed" tree to perform a ML likelihood analysis as implemented in PhyML. In the ML analysis, up to five different evolutionary models were tested and the model best fitting to the data was determined by the Akaike Information Criterion (AIC). Different algorithms were used to identify homology relationships and lineage-specific duplications. To extract and interpret the large data set obtained a Structured Query Language (SQL) relational database was built. This database was the main resource for data mining in this work. Adapted from [39].

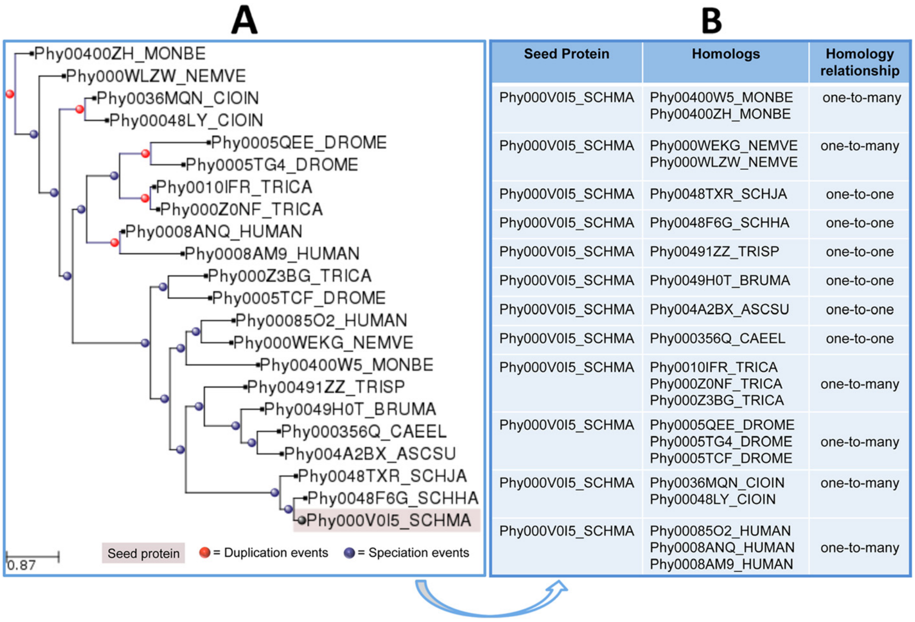
PhylomeDB [47]. However, links to UniProt [49] and other databases are provided.

To explore the benefits offered by comparative genomics in order to improve functional annotation of genes and gene products, it is also necessary to consider the limitations involved in this approach. Although it is generally accepted that functional annotation through orthology, rather than just homology relationship, constitutes one of the most promising annotation approaches, these surveys are designed to provide predictions regarding the likely protein function, but it does not substitute experimental confirmation [36,52]. Functional diversity is often associated with significant divergence at the sequence level, but high levels of identity do not ensure that two or more proteins perform the same function, since subtle changes in active sites are able to completely change the protein function [53].

As we previously mentioned, evolutionary analysis involving fully sequenced genomes/proteomes remains a challenge. Although the tools here applied were not originally designed for large scale phylogenetic analysis, we adapted them to work on a large scale, since we strongly believe that a system-wide perspective on evolutionary processes can greatly improve the understanding on how genomes came to be and what evolutionary process took them there. Functional prediction as described in the present work could be used as a starting point for future projects, prioritizing the selection of certain genes or proteins for new experimental studies.

Detection of gene duplications in *S. mansoni*

An additional advantage of the phylogeny-based approach is that it readily provides a collection of gene evolutionary histories that can be mined for particular



Seed Protein	Homologs	Homology relationship
Phy000V0I5_SCHMA	Phy00400W5_MONBE Phy00400ZH_MONBE	one-to-many
Phy000V0I5_SCHMA	Phy000WEKG_NEMVE Phy000WLZW_NEMVE	one-to-many
Phy000V0I5_SCHMA	Phy0048TXR_SCHJA	one-to-one
Phy000V0I5_SCHMA	Phy0048F6G_SCHHA	one-to-one
Phy000V0I5_SCHMA	Phy00491ZZ_TRISP	one-to-one
Phy000V0I5_SCHMA	Phy0049H0T_BRUMA	one-to-one
Phy000V0I5_SCHMA	Phy004A2BX_ASCSU	one-to-one
Phy000V0I5_SCHMA	Phy000356Q_CAEEL	one-to-one
Phy000V0I5_SCHMA	Phy0010IFR_TRICA Phy000Z0NF_TRICA Phy000Z3BG_TRICA	one-to-many
Phy000V0I5_SCHMA	Phy0005QEE_DROME Phy0005TG4_DROME Phy0005TCF_DROME	one-to-many
Phy000V0I5_SCHMA	Phy0036MQN_CIOIN Phy00048LY_CIOIN	one-to-many
Phy000V0I5_SCHMA	Phy00085O2_HUMAN Phy0008ANQ_HUMAN Phy0008AM9_HUMAN	one-to-many

Figure 2 Homology relationships and evolutionary events inferred from the analysis of a *S. mansoni* protein. A) Phylogenetic tree reconstructed for the parasite “seed” protein Phy000V0I5_SCHMA (Smp_175750). **B**) Homology relationships identified between the “seed” protein and its homologs in the other species.

events. Since gene duplication is considered one of the main mechanisms for functional innovation and diversification [54], we explored the *S. mansoni* phylome to identify protein families that have been specifically expanded in this lineage, since its diversification from the other sequenced metazoans. We used the above-mentioned *species-overlap* algorithm that identifies duplication nodes and also provides clues of the relative dating of the duplication event [39,55].

Such analysis revealed that in 3,051 reconstructed phylogenetic trees there is at least one paralog connected to the “seed” protein through a duplication node (Additional file 1 Table S3). Among these, 211 phylogenies show lineage-specific duplications in the three *Schistosoma* species in comparison with the other taxa. These expansions are small-to-moderate in size, resulting in a total of two to ten paralogs, and include some of the most significant expansions as discussed below.

The inclusion of *S. haematobium* and *S. japonicum* proteomes gave us a high resolution within *Schistosoma* genus and allowed us to make comparisons across this taxon. In general, the expansions observed in *S. mansoni* can also be observed in the other two *Schistosoma* species, although with variable number of paralogs in each species. As previously observed by evolutionary relationships, cytogenetic data, and syntenic analyses, the present study shows that *S. mansoni* is more closely related to *S. haematobium* than to the *S. japonicum* [56-59]. Moreover, 170 evolutionary trees have only *S. mansoni* and *S. haematobium* proteins, while only six phylogenies have solely *S. mansoni* and *S. japonicum* proteins. Meanwhile, most of the homologous proteins shared by *S. mansoni* and *S. haematobium* are annotated as “hypothetical protein” and do not have any predicted function or significant hits with known proteins in public databases as UniProt [49], Pfam [60], or non-redundant (nr) NCBI database (ftp://ftp.ncbi.nih.gov/blast/db).

A small number of phylogenetic trees (1,45%) had only sequences of *S. mansoni*. These could be the result of very recent duplication events of proteins that are specific to this species. However, many of these genes were not found in the genetic map of *S. mansoni* [16,18] and they do not contain protein domains traceable at Pfam [60]. BLAST searches against the non-redundant (nr) NCBI database detected a few non-*Schistosoma* proteins as significant hits that were annotated as hypothetical in all cases. For these reasons we rather believe that these sequences correspond to spurious predictions. Further analyses will be conducted in the future in order to confirm or refute this hypothesis.

Among the most significant protein expansions in *S. mansoni* we identified tetraspanins, fucosyltransferases, venom allergen-like proteins (SmVAL), tegumental-allergen-like proteins (SmTAL), leishmanolysins, and elastases, which were previously proposed as drug targets, once they can be related to morphological or physiological specificities of this parasite [5,20,61-65]. In these cases, the protein family membership ranged from 6 to 23 paralogs encoded in the parasite’s genome.

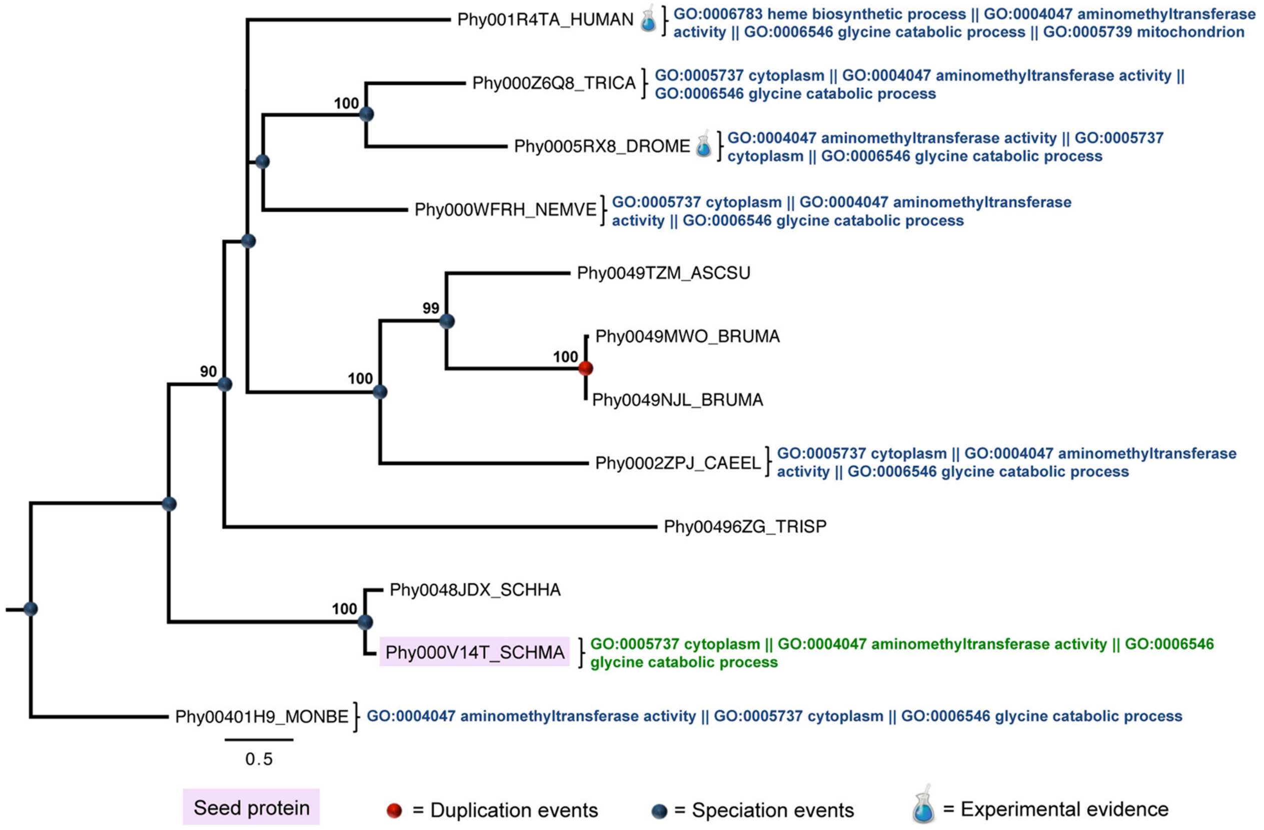


Figure 3 Example of functional prediction based on phylogenetic analysis. The protein sequences are represented by the internal identifier in PhylomeDB. Relationships among the parasite Phy000V14T_SCHMA "seed" protein (Smp_170950) and its homologs in other species (Table 1) as inferred by maximum likelihood method implemented in PhyML. Support values were computed by approximate likelihood ratio test (aLTR). Curly brackets hold Gene Ontology (GO) terms for proteins in this dataset.

Tetraspanins are small proteins with four transmembrane domains involved in the coordination of intra and intercellular processes, such as signal transduction, cell proliferation, adhesion, and migration, cell fusion and host-parasite interactions [66,67]. The function of schistosome tetraspanins are not completely understood, but cell-cell interactions and maintenance of cell membrane integrity might be performed by these proteins as well as they can be receptors for host ligands, acting on immune evasion [61]. The suppression of two tetraspanin genes (*Sm-tsp-1* and *Sm-tsp-2*) by RNA interference in mice also suggests that these proteins play important structural roles in the parasite's tegument, being a good target for anti-schistosomal vaccine [68]. Figure 4 illustrates an example of tetraspanin lineage-specific duplications. In this case, the number of homologs in the three Schistosoma species varies from six to eight. Tree topology shows distinct well-supported clades suggesting that structural and/or functional variants might be present. Three proteins in this dataset have experimental evidence: Phy0048JNS_SCHHA (Q26499), Phy0048WJL_SCHMA (P19331), and Phy0005UU9_DROME (O46101) [49,69,70].

Venom allergen-like proteins (SmVAL), also called sperm-coating protein-like (SCP-like), are structurally related proteins members of the SCP/TAPS family. In Platyhelminthes, these proteins have been linked as potential modulators of immune function and components of sexual development [71]. Although the specific function of each SmVAL family member is unknown, there is evidence suggesting potential roles in larval penetration, host immune response modulation, and adult worm development [63,71]. Furthermore, analyses of SmVAL transcripts demonstrated that the corresponding genes are upregulated in infective stages of the parasite, highlighting SmVAL proteins as candidates for novel vaccine strategies [71,72].

Fucosyltransferases are enzymes that catalyses the fucose transfer from the donor guanosine-diphosphate fucose to different acceptor molecules such as oligosaccharides, glycoproteins, and glycolipids [73]. In schistosomes, fucosyltransferases are involved in producing immunomodulatory epitopes during infection, granuloma formation, egg/endothelium interactions, and were previously highlighted as anti-schistosomal candidates [63,74].

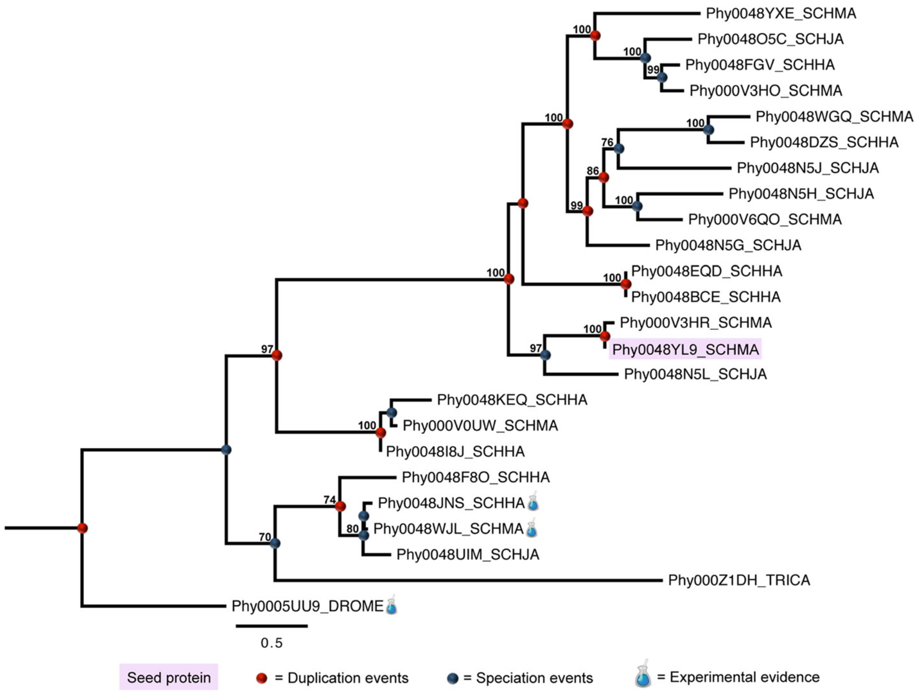


Figure 4 Phylogenetic relationships of schistosome lineage-specific duplicated tetraspanins. Analysis was performed with trimmed sequence alignment by using the maximum likelihood method as implemented in PhyML. Best fit model (WAG) and support values for each node were estimated by the Akaike Likelihood Ratio Test (aLRT). Sequence labels follow the PhylomeDB internal identifier. For details, see supplementary data (Additional file 1 Table S3).

Tegumental-Allergen-Like proteins (SmTALs) are members of a protein family present in parasitic Platyhelminthes [64,75]. These proteins are located inside the tegument and have different life-cycle expression patterns [64]. The tegumental protein Sm22.6 is considered the main target for human IgE in *S. mansoni* and human IgE response against this protein is associated with the development of age-dependent partial immunity to *S. mansoni* infections in endemic areas [64,76].

Leishmanolysin, (also called invadolysin and SmPepM8), is a major surface protease member of the metallopeptidase M8 family. This protein can perform activities in schistosomes similar to those performed in *Leishmania* where these proteins are involved in different types of processes like degradation of the extracellular matrix and inhibition or perturbations of host cell interactions [63,72]. In turn, elastases are serine proteases that in schistosomes play a pivotal role in the penetration by cercariae of host skin to initiate infection. Recent studies have also revealed that these proteases can be employed by schistosomes to overcome or evade the host immune response [77,78]. Members of *S. mansoni* peptidase families such as leishmanolysins, cercarial elastases, and cathepsin D proteins were subjected to a detailed study in respect to their domain architectures, functional properties, and evolutionary relationships as described elsewhere [65].

Another specific feature of schistosomes is related to their tegument. Distinct from nematodes, which have a cuticle covering and protecting the organism body, schistosomes are covered by a living syncytium bounded by a complex multilaminate surface, which undergoes several adaptations soon after infection is initiated [79-81]. The external double membrane plays a crucial role in host-parasite interactions, being responsible for diverse mechanisms of survival [19,82,83]. The development of a tegument, highly specialized and resistant to immune damage, was accompanied by evolutionary adaptations, for example, the expansions of other protein families encoding annexins, cadherins, and innexins.

Annexins are widely distributed in eukaryotes performing a broad range of important biological processes related to tegument membrane [84-86]. In schistosomes, annexins appear to be involved in parasite's stability protecting against immune attack by the host as well as against structural breakdown [85,86]. Cadherins are adhesion molecules that mediate Ca^{2+}-dependent cell-cell adhesion and whose duplication events happened probably in parallel to the advent of a third germ layer in flatworms [5,87]. Innexins are components of gap-junction proteins, the intercellular channels that allow for the exchange of ions and other small signal molecules [88,89]. In *C. elegans*, innexins have been implicated in different processes like electrical coupling between pharyngeal muscles, calcium propagation in the gut, gap junction-mediated oocyte, and sensory neuron identity [89].

In summary, we identified that approximately 45% of the *S. mansoni* predicted proteins that were covered by this phylogenomic analysis have, at least, one paralog encoded in the parasite genome that might have arisen by gene duplication events that occurred after its divergence from other selected taxa (Additional file 1 Table S3). In other eukaryotic genomes this value ranges from 30 and 65% [90], whereas in *C. elegans* this value is equal to 49% [91].

Altogether, the present results indicate that besides the exploitation of host endocrine and immune signals, the parasite genome exhibit multiple events of gene duplication which may be, at least partially, an adaptive response related to the parasitic lifestyle. These expansions probably reflect the intriguing complexity of evolutionary events that happened over time, resulting in important characteristics in schistosome's biology with consequences to the disease it causes. Taking into account the host environment and the selective forces that it imposes to a parasite, the phylogeny of host(s) and parasite(s) are probably closely related, once this coevolution will be responsible for the continuity or elimination of such an interaction. Nonetheless, previous empirical experiments involving schistosomes and the intermediate host provide further support to suggest the potential for host-schistosome coevolution [92].

In this context, it is important to analyze the evolutionary history of protein families during screening for potential targets for drug and vaccine development. Incorporating the evolutionary perspective in drug development studies can improve our understanding regarding drug resistance and effectiveness, as well as to guide new strategies of drug discovery. Gene duplication events as well as adaptive evolution should be considered during this process, since an anti-parasitic drug could bind a single protein or in all proteins encoded by a multi-gene family [93]. As a consequence, therapies which target a subset of genes that arose by duplication may not be effective at low doses. To solve this problem, the drug's effectiveness can be increased when a single-copy gene is targeted and its function is inactivated causing complete perturbation of a vital pathway [93,94].

Conclusions

Through a systemic approach, we may accelerate the advance towards the understanding of schistosomiasis, its etiologic agents, and host-parasite interactions, optimizing the discovery of therapeutic targets to the development of new drugs and vaccines. Besides promoting a significant improvement in the functional annotation of the *S. mansoni* predicted proteome, our approach provided relevant information about the parasite's genome evolution such as the identification of gene duplication events and expanded protein families, supplying important information regarding the mechanisms involved in *Schistosoma*'s genome evolution. Among the parasite paralog groups, we identified proteases, tetraspanins, fucosyltransferases, venom allergen-like proteins (also called as SmVAL or SCP-like), and tegumental-allergen-like proteins (SmTAL) that may be related to morphological or physiological specificities of this parasite. In addition, we strongly believe that the *S. mansoni* phylome data will pave the way for other, more detailed analysis, such as those that have been already performed on expanded peptidases families [65].

One of the remaining challenges is to understand which evasion strategies enable this parasite to survive for years in a potentially hostile environment, protected from the host immune system action and/or actively making the host response ineffective. Different mechanisms may be involved in these processes, including the generation of variant proteins by expression of micro-exon genes (MEG), which have been pointed as a potential strategy [94], and small non-coding RNAs which perform many essential regulatory functions [95].

Insights obtained through this phylogenomic approach will help us to guide forward genetic approaches to better understand the host-pathogen relationships toward to the elucidation of novel drug targets and vaccine candidates urgently needed to reduce the morbidity and mortality caused by schistosomiasis worldwide. Continuing this work, a comparative analysis involving genomic, transcriptomic, and proteomic data from other helminth species as *Taenia solium*, *Echinococcus multiloculares*, *Echinococcus granulosus*, *Fasciola hepatica*, other parasites, and vectors will provide valuable information from a system-wide perspective of a broad range of organisms, improving our understanding regarding the parasitic lifestyle.

Methods

Organisms and sequence data

Predicted proteomes from 13 fully sequenced eukaryotic genomes were downloaded from JGI Genome Projects, SchistoDB, Quest For Orthologs, WormBase, BeetleBASE, and FlyBase (Table 1). The taxon sampling was selected according to the availability of the predicted proteomes and based on the phylogenetic position of each species. The comprehensive taxa selected cover important evolutionary innovations making this dataset set especially suitable for addressing the evolutionary innovations in schistosomes in the context of metazoan evolution. Model organisms were also included to provide functional annotations that could be potentially transferred to *S. mansoni* homologous proteins.

Phylome reconstruction

To reconstruct the complete collection of phylogenetic trees for all *S. mansoni* proteins and their homologs in other 12 fully sequenced organisms (Table 1), we used a similar automated pipeline to that described earlier for the human proteome [39] (Figure 1). A local database was created containing data from the *S. mansoni* proteome and those of 12 other completely sequenced genomes/proteomes. Alternative splicing products and identical sequences from the *S. mansoni* proteome were filtered out. For each protein encoded in the *S. mansoni* genome ("seed"), a Smith-Waterman search [96] (E-value $\leq 10^{-5}$) was performed against the above mentioned database to retrieve proteins with significant sequence similarity. Sequences that aligned with a continuous region longer than 50% of the query sequence were selected and aligned using MUSCLE 3.6 [40], MAFFT [41], DIALIGN-TX [42], and M-Coffee [43] with default parameters. Positions in the alignment containing a high number of gaps were eliminated using trimAl [44], with a consistency cutoff of 0.1667 and a gap score cutoff of 0.1. Neighbor-joining trees were derived from the trimmed alignments using *scoredist* distances as implemented in BioNJ [45] and maximum likelihood trees were obtained as implemented in PhyML using the NJ tree as a starting point [46]. For each "seed" protein phylogenetic reconstruction, we tested four different evolutionary models (JTT, WAG, BLOSUM62, VT, LG, CpREV, and DCMut). In all cases a discrete gamma-distribution model with four rate categories plus invariant positions was assumed, the gamma parameter and the fraction of invariant positions were estimated from the data. Tree support values were computed by approximate likelihood ratio test (aLTR) as implemented in PhyML [46,97]. The evolutionary model best fitting the data was determined by comparing the likelihood of the used models according to the Akaike Information Criterion (AIC) [98].

Prediction of homology relationships

To derive orthology and paralogy relationships among *S. mansoni* proteins and those encoded in the other genomes included in this study we used a *species-overlap* algorithm as described in [39] and as implemented in ETE (Environment for Tree Exploration) [99]. This algorithm uses the level of species overlap between the two daughter partitions of a given node to define it as duplication or a speciation event. The analysis starts at the protein used to generate the tree ("seed" protein) and runs through the internal nodes of the tree until it reaches the root. All the trees were rooted at the midpoint. If the two partitions share any species (if there is species overlap), the node is defined as a duplication node and the proteins are considered paralogous ones. Otherwise (if there is no overlap) the node is defined as a speciation node leading to orthologous proteins. Once all the nodes have been classified, the algorithm establishes the orthology and paralogy relationships between the "seed" protein and other proteins included in the tree according to the original definition of these terms [39,100]. A previous study has shown that the *species-overlap* algorithm produces reliable orthology predictions with higher sensibility than a strict reconciliation method [101].

Orthology-based functional annotation

Based on the list of orthology and paralogy relationship we performed the transfer of functional annotation from each ortholog with known function to the *S. mansoni* "seed" proteins. To produce a confident set of functional predictions for *S. mansoni* proteins, we classified the list of orthologs in different subsets of orthology relationships (one-to-one, one-to-many, many-to-one, and many-to-many) between the *S. mansoni* proteins and the other proteins included in this phylome data. If no duplication has occurred since the speciation, the two genes form a one-to-one relationship. If subsequent duplications have occurred, other types of orthology relationships (one-to-many or many-to-many) were assigned [51]. One example of this classification is provided (Figure 2).

To further analyze such large data set, we built the SchistoPhylomeSQL a Structured Query Language relational database using MySQL as a database management system. This local database integrates information from PhylomeDB (http://phylomedb.org) and SchistoDB (http://www.schistodb.net). Access to the database was obtained using DbVisualizer version 7.0.5 (http://www.dbvis.com), a graphical user interface that allows developing and accessing database management system (DBMSs) in different operating systems. The SchistoPhylomeSQL database was the main resource for data mining in this work. Perl scripts and SQL queries were implemented to parse the text files and load them to the database.

Detection of *S. mansoni* gene expansions

Using ETE [99], we analyzed the *S. mansoni* phylome data to identify protein families that were specifically expanded in the *S. mansoni* lineage since its diversification from the other metazoans (Additional file 1 Table S3). The duplication events defined by the *species-overlap* algorithm that only comprised paralogs from *S. mansoni* were considered lineage-specific duplications. In cases where the information extracted from more than one phylogenetic tree contained the same paralogous proteins, changing only the "seed" protein position, the data was filtered to obtain a non-redundant list of in-paralogs.

Additional file

Additional file 1: Table S1: Functional prediction for the*S. mansoni* "seed" proteins. **Table S2**: Comparison between reviewed proteins with functional annotation at UniProt and GO terms transferred by phylogenomic approach. **Table S3**: Homologous proteins to the *S. mansoni* "seed" proteins.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

LLS: carried out the phylogenetic and functional annotation studies, and drafted the manuscript. MM: performed the phylome reconstruction and functional annotation transfer. LAN: participated in the coordination of this study, and co-wrote this manuscript. AZ: wrote the Perl scripts for data manipulation and provided computational support for this study. TG: participated in the coordination of this study, supervised the phylome reconstruction, and co-wrote this manuscript. GO: participated in the design and coordination of this study, and co-wrote this manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors acknowledge the use of the computing resources of the Centre de Regulació Genòmica (CRG, Spain) and the Center for Excellence in Bioinformatics (CEBio/Fiocruz, Brazil). The authors are grateful to Jaime Huerta-Cepas (CRG), Eric Aguiar, and Francislon Silva (CEBio) for bioinformatics technical support and to Mariana de Oliveira (CEBio) for help with illustrations in this work. This work was funded by the National Institutes of Health - NIH/Fogarty International Center (TW007012 to GO), the National Council for Research and Development - CNPq (CNPq Research Fellowship 306879/2009-3 to GO, INCT-DT 573839/2008-5 to GO, and CNPq-Universal 476036/2010-0 to LAN), the Spanish Ministry of Science and Innovation (BFU2009-09168 to TG), and the Research Foundation of the State of Minas Gerais - FAPEMIG (CBB-1181/08 and PPM-00439-10 to GO).

Author details

[1]Grupo de Genômica e Biologia Computacional, Centro de Pesquisas René Rachou. Instituto Nacional de Ciência e Tecnologia em Doenças Tropicais. Fundação Oswaldo Cruz - FIOCRUZ, Belo Horizonte, MG 30190-002, Brazil. [2]Centro de Excelência em Bioinformática, Fundação Oswaldo Cruz – FIOCRUZ, Belo Horizonte, MG, Brazil. [3]Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais – UFMG, Belo Horizonte, MG, Brazil. [4]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Dr. Aiguader, 88, 08003, Barcelona, Spain. [5]Universitat Pompeu Fabra (UPF), 08003, Barcelona, Spain. [6]Faculdade Infórium de Tecnologia, Belo Horizonte, MG 30130-180, Brazil. [7]Laboratório Multiusuário de Bioinformática, Embrapa Informática Agropecuária, Campinas, São Paulo, Brazil.

Received: 30 September 2012 Accepted: 22 October 2012
Published: 13 November 2012

References

1. Engels D, Chitsulo L, Montresor A, Savioli L: **The global epidemiological situation of schistosomiasis and new approaches to control and research.** *Acta Trop* 2002, **82**(2):139–146.
2. Steinmann P, Keiser J, Bos R, Tanner M, Utzinger J: **Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk.** *Lancet Infect Dis* 2006, **6**(7):411–425.
3. Gryseels B: **Schistosomiasis.** *Infect Dis Clin North Am* 2012, **26**(2):383–397.
4. Organization WH: **Schistosomiasis: population requiring preventive chemotherapy and number of people treated in 2010.** *Wkly Epidemiol Rec* 2012, **87**(4):37–44.
5. Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyama ST, Al-Lazikani B, Andrade LF, Ashton PD, Aslett MA, Bartholomeu DC, Blandin G, Caffrey CR, Coghlan A, Coulson R, Day TA, Delcher A, DeMarco R, Djikeng A, Eyre T, Gamble JA, Ghedin E, Gu Y, Hertz-Fowler C, Hirai H, Hirai Y, Houston R, Ivens A, Johnston DA, Lacerda D, Macedo CD, McVeigh P, Ning Z, Oliveira G, Overington JP, Parkhill J, Pertea M, Pierce RJ, Protasio AV, Quail MA, Rajandream MA, Rogers J, Sajid M, Salzberg SL, Stanke M, Tivey AR, White O, Williams DL, Wortman J, Wu W, Zamanian M, Zerlotini A, Fraser-Liggett CM, Barrell BG, El-Sayed NM: **The genome of the blood fluke Schistosoma mansoni.** *Nature* 2009, **460**(7253):352–358.
6. SjGSaFA C: **The Schistosoma japonicum genome reveals features of host-parasite interplay.** *Nature* 2009, **460**(7253):345–351.
7. Young ND, Jex AR, Li B, Liu S, Yang L, Xiong Z, Li Y, Cantacessi C, Hall RS, Xu X, Chen F, Wu X, Zerlotini A, Oliveira G, Hofmann A, Zhang G, Fang X, Kang Y, Campbell BE, Loukas A, Ranganathan S, Rollinson D, Rinaldi G, Brindley PJ, Yang H, Wang J, Gasser RB: **Whole-genome sequence of Schistosoma haematobium.** *Nat Genet* 2012, **44**(2):221–225.
8. TDR: *Tropical Disease Research: progress 2005–2006*. Geneva: World Health Organization, Tropical Disease Research; 2007:112.
9. Bruun B, Aagaard-Hansen J: *The social context of schistosomiasis and its control: an introduction and annotated bibliography*. Geneva: World Health Organization; 2008.
10. Liang YS, Dai JR, Zhu YC, Coles GC, Doenhoff MJ: **Genetic analysis of praziquantel resistance in Schistosoma mansoni.** *Southeast Asian J Trop Med Public Health* 2003, **34**(2):274–280.
11. Pica-Mattoccia L, Cioli D: **Sex- and stage-related sensitivity of Schistosoma mansoni to in vivo and in vitro praziquantel treatment.** *Int J Parasitol* 2004, **34**(4):527–533.
12. Botros SS, Bennett JL: **Praziquantel resistance.** *Expert Opin Drug Discov* 2007, **2**:S35–S40.
13. Melman SD, Steinauer ML, Cunningham C, Kubatko LS, Mwangi IN, Wynn NB, Mutuku MW, Karanja DM, Colley DG, Black CL, Secor WE, Mkoji GM, Loker ES: **Reduced susceptibility to praziquantel among naturally occurring Kenyan isolates of Schistosoma mansoni.** *PLoS Negl Trop Dis* 2009, **3**(8):e504.
14. Rokni MB: *Schistosomiasis*. InTech; 2002.
15. van der Werf MJ, de Vlas SJ, Brooker S, Looman CW, Nagelkerke NJ, Habbema JD, Engels D: **Quantification of clinical morbidity associated with schistosome infection in sub-Saharan Africa.** *Acta Trop* 2003, **86**(2–3):125–139.
16. Protasio AV, Tsai IJ, Babbage A, Nichol S, Hunt M, Aslett MA, De Silva N, Velarde GS, Anderson TJ, Clark RC, Davidson C, Dillon GP, Holroyd NE, LoVerde PT, Lloyd C, McQuillan J, Oliveira G, Otto TD, Parker-Manuel SJ, Quail MA, Wilson RA, Zerlotini A, Dunne DW, Berriman M: **A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni.** *PLoS Negl Trop Dis* 2012, **6**(1):e1455.
17. Zerlotini A, Heiges M, Wang H, Moraes RL, Dominitini AJ, Ruiz JC, Kissinger JC, Oliveira G: **SchistoDB: a Schistosoma mansoni genome resource.** *Nucleic Acids Res* 2009, **37**(Database issue):579–582.
18. Criscione CD, Valentim CL, Hirai H, LoVerde PT, Anderson TJ: **Genomic linkage map of the human blood fluke Schistosoma mansoni.** *Genome Biol* 2009, **10**(6):R71.
19. Han ZG, Brindley PJ, Wang SY, Chen Z: **Schistosoma genomics: new perspectives on schistosome biology and host-parasite interaction.** *Annu Rev Genomics Hum Genet* 2009, **10**:211–240.
20. DeMarco R, Verjovski-Almeida S: **Schistosomes–proteomics studies for potential novel vaccines and drug targets.** *Drug Discov Today* 2009, **14**(9–10):472–478.
21. Hawkins T, Kihara D: **Function prediction of uncharacterized proteins.** *J Bioinform Comput Biol* 2007, **5**(1):1–30.
22. Nahum LA, Pereira SL: **Phylogenomics, Protein Family Evolution, and the Tree of Life: An Integrated Approach between Molecular Evolution and Computational Intelligence**. In *Studies in Computational Intelligence (SCI)*. vol. 122nd edition. Edited by Smolinski TG, Milanova MG, Hassanien AE. Berlin Heidelberg: Springer-Verlag; 2008:259–279.
23. Jiang Z: **Protein function predictions based on the phylogenetic profile method.** *Crit Rev Biotechnol* 2008, **28**(4):233–238.
24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403–410.
25. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**(9):2178–2189.
26. Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer EL: **InParanoid 7: new algorithms and tools for eukaryotic orthology analysis.** *Nucleic Acids Res* 2010, **38**(Database issue):196–203.
27. Brenner SE: **Errors in genome annotation.** *Trends Genet* 1999, **15**(4):132–133.
28. Brown D, Sjölander K: **Functional classification using phylogenomic inference.** *PLoS Comput Biol* 2006, **2**(6):e77.

29. Gabaldón T: **Comparative genomics-based prediction of protein function.** *Methods Mol Biol* 2008, **439**:387–401.
30. Eisen JA, Wu M: **Phylogenetic analysis and gene functional predictions: phylogenomics in action.** *Theor Popul Biol* 2002, **61**(4):481–487.
31. Galperin MY, Koonin EV: **Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption.** *In Silico Biol* 1998, **1**(1):55–67.
32. Koski LB, Golding GB: **The closest BLAST hit is often not the nearest neighbor.** *J Mol Evol* 2001, **52**(6):540–542.
33. Andrade LF, Nahum LA, Avelar LG, Silva LL, Zerlotini A, Ruiz JC, Oliveira G: **Eukaryotic Protein Kinases (ePKs) of the Helminth Parasite Schistosoma mansoni.** *BMC Genomics* 2011, **12**:215.
34. Nahum LA, Goswami S, Serres MH: **Protein families reflect the metabolic diversity of organisms and provide support for functional prediction.** *Physiol Genomics* 2009, **38**(3):250–260.
35. Eisen JA: **Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis.** *Genome Res* 1998, **8**(3):163–167.
36. Gabaldón T, Dessimoz C, Huxley-Jones J, Vilella AJ, Sonnhammer EL, Lewis S: **Joining forces in the quest for orthologs.** *Genome Biol* 2009, **10**(9):403.
37. Sicheritz-Pontén T, Andersson SG: **A phylogenomic approach to microbial evolution.** *Nucleic Acids Res* 2001, **29**(2):545–552.
38. Huerta-Cepas J, Marcet-Houben M, Pignatelli M, Moya A, Gabaldón T: **The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for Acyrthosiphon pisum genes.** *Insect Mol Biol* 2010, **19**(Suppl 2):13–21.
39. Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T: **The human phylome.** *Genome Biol* 2007, **8**(6):R109.
40. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinforma* 2004, **5**:113.
41. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program.** *Brief Bioinform* 2008, **9**(4):286–298.
42. Subramanian AR, Kaufmann M, Morgenstern B: **DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment.** *Algorithms Mol Biol* 2008, **3**:6.
43. Wallace IM, O'Sullivan O, Higgins DG, Notredame C: **M-Coffee: combining multiple sequence alignment methods with T-Coffee.** *Nucleic Acids Res* 2006, **34**(6):1692–1699.
44. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T: **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.** *Bioinformatics* 2009, **25**(15):1972–1973.
45. Gascuel O: **BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data.** *Mol Biol Evol* 1997, **14**(7):685–695.
46. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52**(5):696–704.
47. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Denisov I, Kormes D, Marcet-Houben M, Gabaldón T: **PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions.** *Nucleic Acids Res* 2011, **39** (Database issue):556–560.
48. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**(1):25–29.
49. Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Antunes R, Barrell D, Bely B, *et al*: **Ongoing and future developments at the Universal Protein Resource.** *Nucleic Acids Res* 2011, **39**(Database issue):214–219.
50. Forslund K, Pekkari I, Sonnhammer EL: **Domain architecture conservation in orthologs.** *BMC Bioinforma* 2011, **12**:326.
51. Koonin EV: **Orthologs, paralogs, and evolutionary genomics.** *Annu Rev Genet* 2005, **39**:309–338.
52. Kuzniar A, van Ham RC, Pongor S, Leunissen JA: **The quest for orthologs: finding the corresponding gene across genomes.** *Trends Genet* 2008, **24**(11):539–551.
53. Gerlt JA, Babbitt PC: **Can sequence determine function?** *Genome Biol* 2000, **1**(5):reviews0005.0001–reviews0005.0010.
54. Ohno S: *Evolution by gene duplication*. 1st edition. New York Heidelberg Berlin: Springer-Verlag; 1970.
55. Huerta-Cepas J, Gabaldón T: **Assigning duplication events to relative temporal scales in genome-wide studies.** *Bioinformatics* 2011, **27**(1):38–45.
56. Lawton SP, Hirai H, Ironside JE, Johnston DA, Rollinson D: **Genomes and geography: genomic insights into the evolution and phylogeography of the genus Schistosoma.** *Parasit Vectors* 2011, **4**:131.
57. Webster BL, Southgate VR, Littlewood DT: **A revision of the interrelationships of Schistosoma including the recently described Schistosoma guineensis.** *Int J Parasitol* 2006, **36**(8):947–955.
58. Littlewood DT, Lockyer AE, Webster BL, Johnston DA, Le TH: **The complete mitochondrial genomes of Schistosoma haematobium and Schistosoma spindale and the evolutionary history of mitochondrial genome changes among parasitic flatworms.** *Molecular phylogenetics and evolution* 2006, **39**(2):452–467.
59. Swain MT, Larkin DM, Caffrey CR, Davies SJ, Loukas A, Skelly PJ, Hoffmann KF: **Schistosoma comparative genomics: integrating genome structure, parasite biology and anthelmintic discovery.** *Trends Parasitol* 2011, **27**(12):555–564.
60. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38**(Database issue):211–222.
61. Tran MH, Pearson MS, Bethony JM, Smyth DJ, Jones MK, Duke M, Don TA, McManus DP, Correa-Oliveira R, Loukas A: **Tetraspanins on the surface of Schistosoma mansoni are protective antigens against schistosomiasis.** *Nat Med* 2006, **12**(7):835–840.
62. McManus DP, Loukas A: **Current status of vaccines for schistosomiasis.** *Clin Microbiol Rev* 2008, **21**(1):225–242.
63. Fitzpatrick JM, Peak E, Perally S, Chalmers IW, Barrett J, Yoshino TP, Ivens AC, Hoffmann KF: **Anti-schistosomal intervention targets identified by lifecycle transcriptomic analyses.** *PLoS Negl Trop Dis* 2009, **3**(11):e543.
64. Fitzsimmons CM, Jones FM, Stearn A, Chalmers IW, Hoffmann KF, Wawrzyniak J, Wilson S, Kabatereine NB, Dunne DW: **The Schistosoma mansoni tegumental-allergen-like (TAL) protein family: influence of developmental expression on human IgE responses.** *PLoS Negl Trop Dis* 2012, **6**(4):e1593.
65. Silva LL, Marcet-Houben M, Zerlotini A, Gabaldón T, Oliveira G, Nahum LA: *Evolutionary Histories of Expanded Peptidase Families in Schistosoma mansoni.*: in press. Mem Inst Oswaldo Cruz; 2011.
66. Levy S, Shoham T: **Protein-protein interactions in the tetraspanin web.** *Physiology (Bethesda)* 2005, **20**:218–224.
67. Yáñez-Mó M, Barreiro O, Gordon-Alonso M, Sala-Valdés M, Sánchez-Madrid F: **Tetraspanin-enriched microdomains: a functional unit in cell plasma membranes.** *Trends Cell Biol* 2009, **19**(9):434–446.
68. Tran MH, Freitas TC, Cooper L, Gaze S, Gatton ML, Jones MK, Lovas E, Pearce EJ, Loukas A: **Suppression of mRNAs encoding tegument tetraspanins from Schistosoma mansoni results in impaired tegument turnover.** *PLoS Pathog* 2010, **6**(4):e1000840.
69. Lee KW, Shalaby KA, Medhat AM, Shi H, Yang Q, Karim AM, LoVerde PT: **Schistosoma mansoni: characterization of the gene encoding Sm23, an integral membrane protein.** *Exp Parasitol* 1995, **80**(1):155–158.
70. Inal J, Bickle Q: **Sequence and immunogenicity of the 23-kDa transmembrane antigen of Schistosoma haematobium.** *Mol Biochem Parasitol* 1995, **74**(2):217–221.
71. Chalmers IW, McArdle AJ, Coulson RM, Wagner MA, Schmid R, Hirai H, Hoffmann KF: **Developmentally regulated expression, alternative splicing and distinct sub-groupings in members of the Schistosoma mansoni venom allergen-like (SmVAL) gene family.** *BMC Genomics* 2008, **9**:89.
72. Curwen RS, Ashton PD, Sundaralingam S, Wilson RA: **Identification of novel proteases and immunomodulators in the secretions of schistosome cercariae that facilitate host entry.** *Mol Cell Proteomics* 2006, **5**(5):835–844.
73. Ma B, Simala-Grant JL, Taylor DE: **Fucosylation in prokaryotes and eukaryotes.** *Glycobiology* 2006, **16**(12):158R–184R.
74. Marques ET Jr, Ichikawa Y, Strand M, August JT, Hart GW, Schnaar RL: **Fucosyltransferases in Schistosoma mansoni development.** *Glycobiology* 2001, **11**(3):249–259.
75. Vichasri-Grams S, Subpipattana P, Sobhon P, Viyanant V, Grams R: **An analysis of the calcium-binding protein 1 of Fasciola gigantica with a comparison to its homologs in the phylum Platyhelminthes.** *Mol Biochem Parasitol* 2006, **146**(1):10–23.
76. Dunne DW, Butterworth AE, Fulford AJ, Kariuki HC, Langley JG, Ouma JH, Capron A, Pierce RJ, Sturrock RF: **Immunity after treatment of human**

schistosomiasis: association between IgE antibodies to adult worm antigens and resistance to reinfection. *Eur J Immunol* 1992, **22**(6):1483–1494.
77. Salter JP, Choe Y, Albrecht H, Franklin C, Lim KC, Craik CS, McKerrow JH: Cercarial elastase is encoded by a functionally conserved gene family across multiple species of schistosomes. *J Biol Chem* 2002, **277**(27):24618–24624.
78. Aslam A, Quinn P, McIntosh RS, Shi J, Ghumra A, McKerrow JH, Bunting KA, Dunne DW, Doenhoff MJ, Morrison SL, Zhang K, Pleass RJ: Proteases from Schistosoma mansoni cercariae cleave IgE at solvent exposed interdomain regions. *Mol Immunol* 2008, **45**(2):567–574.
79. Hockley DJ, McLaren DJ: Schistosoma mansoni: changes in the outer membrane of the tegument during development from cercaria to adult worm. *Int J Parasitol* 1973, **3**(1):13–25.
80. Pearce EJ, Sher A: Mechanisms of immune evasion in schistosomiasis. *Contrib Microbiol Immunol* 1987, **8**:219–232.
81. Braschi S, Borges WC, Wilson RA: Proteomic analysis of the schistosome tegument and its surface membranes. *Mem Inst Oswaldo Cruz* 2006, **101**(Suppl 1):205–212.
82. Mclaren DJ, Hockley DJ: Blood flukes have a double outer membrane. *Nature* 1977, **269**(5624):147–149.
83. Escobedo G, Roberts CW, Carrero JC, Morales-Montor J: Parasite regulation by host hormones: an old mechanism of host exploitation? *Trends Parasitol* 2005, **21**(12):588–593.
84. Rescher U, Gerke V: Annexins–unique membrane binding proteins with diverse functions. *J Cell Sci* 2004, **117**(Pt 13):2631–2639.
85. Tararam CA, Farias LP, Wilson RA, Leite LC: Schistosoma mansoni Annexin 2: molecular characterization and immunolocalization. *Exp Parasitol* 2010, **126**(2):146–155.
86. Hofmann A, Osman A, Leow CY, Driguez P, McManus DP, Jones MK: Parasite annexins–new molecules with potential for drug and vaccine development. *BioEssays* 2010, **32**(11):967–976.
87. Takeichi M: Cadherins: a molecular family important in selective cell-cell adhesion. *Annu Rev Biochem* 1990, **59**:237–252.
88. Phelan P, Bacon JP, Davies JA, Stebbings LA, Todman MG, Avery L, Baines RA, Barnes TM, Ford C, Hekimi S, Lee R, Shaw JE, Starich TA, Curtin KD, Sun YA, Wyman RJ: Innexins: a family of invertebrate gap-junction proteins. *Trends Genet* 1998, **14**(9):348–349.
89. Yeh E, Kawano T, Ng S, Fetter R, Hung W, Wang Y, Zhen M: Caenorhabditis elegans innexins regulate active zone differentiation. *J Neurosci* 2009, **29**(16):5207–5217.
90. Zhang J: Evolution by gene duplication: an update. *Trends Ecol Evol* 2003, **18**(6):292–298.
91. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski MP, Misra S, Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P, Celniker SE, Chervitz SA, Coates D, Cravchik A, Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LS, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJ, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vosshall LB, Zhang J, Zhao Q, Zheng XH, Lewis S: Comparative genomics of the eukaryotes. *Science* 2000, 287(5461):2204–2215.
92. Webster JP, Shrivastava J, Johnson PJ, Blair L: Is host-schistosome coevolution going anywhere? *BMC Evol Biol* 2007, **7**:91.
93. Emes RD, Yang Z: Duplicated paralogous genes subject to positive selection in the genome of Trypanosoma brucei. *PLoS One* 2008, 3(5):e2295.
94. DeMarco R, Mathieson W, Manuel SJ, Dillon GP, Curwen RS, Ashton PD, Ivens AC, Berriman M, Verjovski-Almeida S, Wilson RA: Protein variation in blood-dwelling schistosome worms generated by differential splicing of micro-exon gene transcripts. *Genome Res* 2010, **20**(8):1112–1121.
95. Oliveira KC, Carvalho ML, Maracaja-Coutinho V, Kitajima JP, Verjovski-Almeida S: Non-coding RNAs in schistosomes: an unexplored world. *An Acad Bras Cienc* 2011, **83**(2):673–694.
96. Smith TF, Waterman MS: Identification of common molecular subsequences. *J Mol Biol* 1981, 147(1):195–197.
97. Anisimova M, Gascuel O: Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol* 2006, **55**(4):539–552.
98. Akaike H: Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory 1973*. Edited by Petrov BN, Csàki F. Budapest: Akademiai kiado; 1973:267–281.
99. Huerta-Cepas J, Dopazo J, Gabaldón T: ETE: a python Environment for Tree Exploration. *BMC Bioinforma* 2010, **11**:24.
100. Fitch WM: Distinguishing homologous from analogous proteins. *Syst Zool* 1970, **19**(2):99–113.
101. Marcet-Houben M, Gabaldón T: The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. *PLoS One* 2009, **4**(2):e4357.

doi:10.1186/1471-2164-13-617
Cite this article as: Silva *et al.*: The *Schistosoma mansoni* phylome: using evolutionary genomics to gain insight into a parasite's biology. *BMC Genomics* 2012 **13**:617.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

3.2 - CAPÍTULO II: Genômica Evolutiva de Famílias de Endopeptidases

Silva LL, Marcet-Houben M, Zerlotini A, Gabaldón T, Oliveira G, Nahum LA. Evolutionary histories of expanded peptidase families in *Schistosoma mansoni*. Mem Inst Oswaldo Cruz. 2011 Nov;106(7):864-77. PMID: 22124560.

Neste estudo em colaboração com o grupo de Genômica Comparativa liderado pelo Dr. Toni Gabaldón (Centre for Genomic Regulation, Espanha). Analisamos o proteoma predito de *S. mansoni* em comparação com outros 16 eucariotos cujos genomas foram sequenciados e o proteoma predito encontra-se disponível. Estes incluem uma planta (*A. thaliana*), três fungos (*U. maydis, S. cerevisiae, N. crassa*), sete invertebrados (*N. vectensis, C. elegans, C. briggsae, D. melanogaster, A. gambiae, B. mori, S. purpuratus*), um tunicado (*C. intestinalis*), um cefalocordado (*B. floridae*) e três vertebrados (*D. rerio, M. musculus, H. sapiens*). Utilizando uma abordagem filogenômica, identificamos e analisamos famílias de proteases/peptidases expandidas no proteoma predito de *S. mansoni*. Em função da importância destas enzimas no desenvolvimento e patologia da esquistossomose, focamos no estudo das mesmas com o objetivo de contribuir para um melhor conhecimento da origem e evolução destes potenciais alvos terapêuticos. Nossos resultados revelam que a complexidade biológica de *S. mansoni* está relacionada a inovações evolutivas que ocorreram antes e após sua diversificação dos demais eucariotos analisados neste estudo e que, muito provavelmente, as diferentes proteases se originaram a partir de sucessivos eventos de duplicação gênica neste parasito. Dentre os membros de famílias de proteases expandidas identificamos: catepsina D (aspartil proteases da família A01), elastase cercarial (serino proteases da família S01) e SmPepM8 (também chamada de leishmanolisina, metaloproteases da família M8). Estas proteínas estão envolvidas na penetração da cercária no hospedeiro definitivo, digestão de hemoglobina, evasão do sistema imune do hospedeiro e iniciação de processos inflamatórios, motivo pelo qual foram previamente apontadas como alvos promissores para o desenvolvimento de drogas e vacina contra a esquistossomose. Este trabalho fornece uma visão evolutiva destas três famílias de endopeptidases de *S. mansoni*, permitindo um melhor entendimento da complexidade genômica linhagem-específica e adaptações potencialmente relacionadas à vida parasitária.

864 *Mem Inst Oswaldo Cruz*, Rio de Janeiro, Vol. *106*(7): 864-877, November 2011

Evolutionary histories of expanded peptidase families in *Schistosoma mansoni*

Larissa Lopes Silva[1,2,3], Marina Marcet-Houben[4], Adhemar Zerlotini[1,2], Toni Gabaldón[4], Guilherme Oliveira[1,2], Laila Alves Nahum[1,2]/+

[1]Grupo de Genômica e Biologia Computacional, Instituto de Pesquisas René Rachou,
Instituto Nacional de Ciência e Tecnologia em Doenças Tropicais [2]Centro de Excelência em Bioinformática-Fiocruz,
Belo Horizonte, MG, Brasil [3]Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[4]Bioinformatics and Genomics Programme, Centre de Regulació Genòmica, Universitat Pompeu Fabra, Barcelona, Spain

Schistosoma mansoni *is one of the three main causative agents of human schistosomiasis, a major health problem with a vast socio-economic impact. Recent advances in the proteomic analysis of schistosomes have revealed that peptidases are the main virulence factors involved in the pathogenesis of this disease. In this context, evolutionary studies can be applied to identify peptidase families that have been expanded in genomes over time in response to different selection pressures. Using a phylogenomic approach, we searched for expanded endopeptidase families in the* S. mansoni *predicted proteome with the aim of contributing to the knowledge of such enzymes as potential therapeutic targets. We found three endopeptidase families that comprise leishmanolysins (metallopeptidase M8 family), cercarial elastases (serine peptidase S1 family) and cathepsin D proteins (aspartic peptidase A1 family). Our results suggest that the* Schistosoma *members of these families originated from successive gene duplication events in the parasite lineage after its diversification from other metazoans. Overall, critical residues are conserved among the duplicated genes/proteins. Furthermore, each protein family displays a distinct evolutionary history. Altogether, this work provides an evolutionary view of three* S. mansoni *peptidase families, which allows for a deeper understanding of the genomic complexity and lineage-specific adaptations potentially related to the parasitic lifestyle.*

Key words: phylogenomics - maximum likelihood analysis - homology prediction - functional annotation - proteases - paralogous families - parasite genomics

Schistosomiasis, which is caused by different species from the *Schistosoma* genus, remains one of the most prevalent tropical neglected diseases, affects 210 million people worldwide, and is responsible for at least 280,000 deaths every year (van der Werf et al. 2003, Steinmann et al. 2006, Han et al. 2009). *Schistosoma mansoni* is one of the three major species that infect humans and is the causative agent of intestinal and hepatic schistosomiasis mainly in Africa and South America (Han et al. 2009). Measures to control schistosomiasis rely almost entirely on praziquantel®, which is the only drug available for mass chemotherapy. Despite the effectiveness of this treatment, re-infection is common and drug-resistant parasites have been found in the laboratory and in the field, which demonstrate the urgent need to develop additional chemotherapeutic agents and effective vaccines (Liang et al. 2003, Pica-Mattoccia & Cioli 2004, Botros & Bennett 2007, Melman et al. 2009).

Over the past several years, advances in the molecular analysis of major parasites have identified some key factors involved in parasitic diseases and peptidases as one of the major factors of pathogenicity (McKerrow et al. 2006, Kasný et al. 2009). These enzymes have been implicated in processes that are crucial to the development and survival of helminth parasites, including digestion, invasion from host tissues, activation of inflammation and evasion of the host immune system (McKerrow et al. 2006, Kasný et al. 2009).

Peptidases (also termed proteases, proteinases or proteolytic enzymes) are hydrolytic enzymes that cleave peptide bonds in proteins. Endopeptidases cleave internal peptide bonds, whereas exopeptidases hydrolyse the amino terminus (aminopeptidases) or carboxy terminus (carboxypeptidases) of different proteins. Enzymatic specificity is determined based on the chemical groups responsible for catalysis in the peptide's active site. Thus, peptidases are classified into one of the following classes: asparagine, aspartic, cysteine, glutamic, metallo, serine, threonine and unknown peptidases (Rawlings & Barrett 1993, Rawlings et al. 2010).

Asparagine peptidases are enzymes that have active sites composed of an aspartic acid and an asparagine, the latter being the P1 residue, the amino acid or molecule, which can be found at a specific location in the cleavage site (Rawlings et al. 2010). In turn, aspartic peptidases have their catalytic centres formed by two aspartate residues that activate a water molecule that mediates the nucleophilic attack on the peptide bond (James 2004, Rawlings

---

Financial support: NIH/FIC (TW007012 to GO), CNPq (CNPq Research Fellowship 306879/2009-3 and INCT-DT 573839/2008-5 to GO, CNPq-Universal 476036/2010-0 to LAN), MICINN (BFU2009-09168 to TG), FAPEMIG (CBB-1181/08 and PPM-00439-10 to GO)
+ Corresponding author: laila@nahum.com.br
Received 20 April 2011
Accepted 9 August 2011

et al. 2010). In general, cysteine peptidases have cysteine and histidine residues forming their "catalytic dyad". Meanwhile, other active site residues have been found. Glutamic peptidases have glutamic acid residues as their primary catalytic residues, which are probably the nucleophilic attack mediators involved in the catalysis (Fujinaga et al. 2004, Rawlings et al. 2010). In metallopeptidases, the catalytic mechanism usually involves a single catalytic zinc ion tetrahedrally coordinated by one glutamate and two histidine residues (Rawlings et al. 2010). Serine peptidases have serine residues at their active sites, which together with two other variable amino acids constitute the "catalytic triad" (Hedstrom 2002, Rawlings et al. 2010). Threonine peptidases have threonine residues as their nucleophiles during catalysis. For unknown peptidases, the active site residues have not yet been determined.

Evolutionary analyses have been applied to a broad range of studies, which include the identification of gene/protein families that have expanded in a specific lineage over evolutionary time and possibly indicate the existence of selective pressure (Irving et al. 2003, Sargeant et al. 2006, Nahum & Pereira 2008, Robinson et al. 2008, Wu et al. 2009, Huzurbazar et al. 2010). The availability of faster and more powerful computers combined with the development of automated pipelines has enabled the investigation of such evolutionary processes through the reconstruction of phylogenetic trees for the complete set of proteins encoded in a genome (known as phylome). The results obtained by this analysis provide a broad view of the evolution of an organism's genome and proteome, which allows for a deeper understanding of genomic complexity and lineage-specific adaptations (Huerta-Cepas et al. 2007, 2010b).

In a previous study, we described the reconstruction of the *S. mansoni* phylome to improve gene/protein functional annotation and provide insights into parasite's biology (phylomedb.org). By applying an automated pipeline, we also identified lineage-specific gene duplications, which may have led to a potential diversification of several protein families that are relevant for host-parasite interactions, such as tetraspanins, fucosyltransferases and sperm-coating protein-like proteins. Here, we explore the *S. mansoni* phylome data to analyse three endopeptidase families that expanded in this lineage since its diversification from 15 other metazoan species with the aim of contributing to the available knowledge of parasite biology and host-parasite interactions from an evolutionary perspective. The members of these families include leishmanolysins (metallopeptidase M8 family), cercarial elastases (serine peptidase S1 family) and cathepsin D proteins (aspartic peptidase A1 family).

The present paper is centred on two main research questions: (i) Did any peptidase families expand in the *S. mansoni* genome/proteome and if so, which ones? (ii) What are the evolutionary histories of these peptidase families? To address these questions, we used a so-called species-overlap algorithm (Huerta-Cepas et al. 2007) to detect lineage-specific duplications that occurred during the evolution of the parasite's genome. We also integrated information on sequence alignments, phylogenetic trees, protein architecture and the conservation of critical residues to characterise these proteins. Our results indicate that each peptidase family has a unique evolutionary history within/across the analysed species. Furthermore, our data support the hypothesis that gene duplication events followed by divergence is the main mechanism shaping the evolution of *S. mansoni*-specific paralogous groups.

The analysis of the evolutionary histories of these three *S. mansoni* families is relevant to functional genomics, evolutionary biology, medicine and biotechnology, especially taking into account the importance of *S. mansoni* peptidases in the development of schistosomiasis and that they have been described as promising vaccine and drug targets (McKerrow et al. 2006, Abdulla et al. 2007, Kasný et al. 2009).

MATERIALS AND METHODS

Organisms and sequence data - The dataset of species selected for analysis includes eight invertebrates (*Nematostella vectensis*, *Caenorhabditis elegans*, *Caenorhabditis briggsae*, *S. mansoni*, *Drosophila melanogaster*, *Anopheles gambiae*, *Bombyx mori* and *Strongylocentrotus purpuratus*), one tunicate (*Ciona intestinalis*), one cephalochordate (*Branchiostoma floridae*), three vertebrates (*Danio rerio*, *Mus musculus* and *Homo sapiens*), three fungi (*Neurospora crassa*, *Saccharomyces cerevisiae* and *Ustilago maydis*) and one plant (*Arabidopsis thaliana*). Information on the selected taxa is provided as Supplementary data.

This dataset is particularly rich in metazoans (76% of the selected species) that cover important evolutionary innovations, for example, the origin of bilateral symmetry, the third germ layer, the development of organs, systems, complex patterns of communication and the emergence of the adaptive immune system, which makes it especially suitable for addressing the evolutionary innovations in *S. mansoni* in comparison with other metazoan species (phylomedb.org).

The *S. mansoni* predicted proteome dataset was downloaded from SchistoDB version 2.0 (schistodb.net) (Zerlotini et al. 2009). Proteomes derived from the 16 fully sequenced genomes were downloaded from the Broad Institute *Ustilago maydis* Database, Ensembl, Intergr8, JGI Genome Projects, National Center for Biotechnology Information Genome Database and SilkDB, which can be collectively accessed through the Genomes OnLine Database (genomesonline.org).

Endopeptidase protein families - Peptidases are hydrolases that act on peptide bonds [Enzyme Commission (EC) 3.4]. Three endopeptidase families were selected and analysed in detail in the present work. They include the metallopeptidase M8 family (EC 3.4.24.-), serine peptidase S1 family (EC 3.4.21.-) and aspartic peptidase A1 family (EC 3.4.23.-) members and belong to three peptidase clans (MA, PA and AA, respectively), as described in the MEROPS database (Rawlings et al. 2010).

Information on enzymes was collected from the literature and database references and included in the Supplementary data. The EC numbers were collected from the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology database, which is available online (chem.qmul.ac.uk/iubmb/enzyme/).

Alignments and phylogenetic trees - Sequence alignments and phylogenetic trees of the endopeptidase families selected for analysis were retrieved from the *S. mansoni* phylome data, which were reconstructed through a comparative analysis among all proteins encoded by the parasite genome and their potential homologs in 16 other eukaryotic species (phylomedb.org) (Huerta-Cepas et al. 2011).

Briefly, the *S. mansoni* phylome was reconstructed using each protein encoded in the *S. mansoni* genome ("seed" proteins) and the potential homologs identified through similarity-based searches (Smith & Waterman 1981) against the dataset of selected proteome data described above. The groups of homologous sequences were aligned using MUSCLE v3.6 (Edgar 2004) and gap-rich columns were filtered using trimAl (Capella-Gutiérrez et al. 2009). Phylogenetic analyses were performed using the neighbour-joining and maximum likelihood (ML) methods, as implemented in PhyML (Guindon & Gascuel 2003).

For the phylogenetic reconstruction of each "seed protein", we tested four different evolutionary models (JTT, WAG, BLOSUM62 and VT). In all cases, a discrete gamma-distribution model with four rate categories plus invariant positions was assumed with the gamma parameter and the fraction of invariant positions estimated from the data. Tree support values were computed using the approximate likelihood ratio test as implemented in PhyML (Guindon & Gascuel 2003, Anisimova & Gascuel 2006). The evolutionary model best fitting the data was determined by comparing the likelihood of the used models according to the Akaike Information Criterion (Akaike 1973). The resulting alignments, phylogenies and homology prediction can be accessed at PhylomeDB (phylomedb.org) (Huerta-Cepas et al. 2011) through protein sequence identifiers (e.g., UniProt: C4PZH6; SchistoDB: Smp_127030; PhylomeDB: Phy000V7EC_SCHMA).

To integrate information from SchistoDB (Zerlotini et al. 2009) and PhylomeDB (Huerta-Cepas et al. 2011), we built a local relational database, named SchistoPhylomeSQL, which allowed us to extract and interpret the large amount of data in this work (Fig. 1). Access to this local database was implemented using DbVisualizer version 7.0.5 (dbvis.com). The SchistoPhylomeSQL database was the main resource for data mining in this work. In-house Perl scripts and Structured Query Language queries were used to parse data files during the database building and searching processes.

Paralogy and orthology relationships - To derive a complete catalogue of the paralogy and orthology relationships between *S. mansoni* proteins and those from other eukaryotic proteomes, we applied a "species-overlap" algorithm, as previously described (Huerta-Cepas et al. 2007). This algorithm uses the level of species overlap between the two daughter partitions of a given node to define it as a duplication or speciation event, which give rise to paralogs and orthologs, respectively. Once all the nodes have been classified, the algorithm establishes the paralogy and orthology relationships between the "seed protein" and other proteins included in the phylogenetic tree, according to the original definition of these terms (Fitch 1970, Gabaldón 2008).

Lineage-specific duplications - Using a python Environment for Tree Exploration (Huerta-Cepas et al. 2010a), we analysed the *S. mansoni* phylome data (phylomedb.org) to identify protein families that were specifically expanded in the *S. mansoni* lineage since its diversification from the other selected taxa (Supplementary data). The duplication events defined by the "species-overlap" algorithm that only comprised paralogs from *S. mansoni* were considered lineage-specific duplications. In cases where more than one phylogenetic tree contained the same paralogous proteins, by changing only the "seed" protein position, the data were filtered to obtain a non-redundant list of in-paralogs.

Protein architecture and critical residues - In this study, we used the Pfam database (Finn et al. 2010) to identify the presence and organisation of protein sequence domains as well as critical residues present in the three *S. mansoni* endopeptidase families. Pfam is a large and widely used database of protein domains families. This database contains multiple sequence alignments and profile hidden Markov models (profile HMMs) for each protein family. Pfam-A entries are derived from the underlying sequence database, which is termed Pfamseq. This database is built from the most recent release of UniProtKB at a given time point (Finn et al. 2010, Apweiler et al. 2011). To predict active sites in new sequences, Pfam uses the information available in UniProtKB for homologous proteins, whose catalytic residues have been experimentally characterized (Mistry et al. 2007). Based on Pfam information, the illustrations of the *S. mansoni* protein domain architectures were generated using DOG 2.0 (Ren et al. 2009).

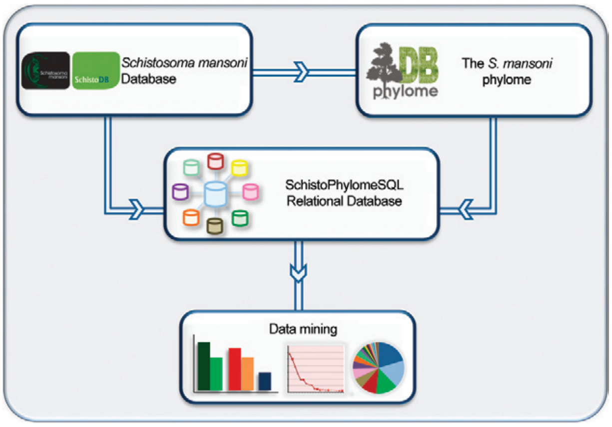


Fig. 1: flowchart of the applied methodology. The *Schistosoma mansoni* proteome data was retrieved from SchistoDB and each protein was used as "seed" to reconstruct the *S. mansoni* phylome. The resulting alignments, phylogenies, and homology predictions are available at PhylomeDB. To integrate information from SchistoDB and PhylomeDB, we built the SchistoPhylomeSQL, a local relational database as the main resource for data mining in this work.

RESULTS

Comparative genomics has revealed a great deal of sequence and/or functional diversity within and across organisms with respect to gene/protein family size, composition and the relatedness of their members (Huerta-Cepas et al. 2007, 2010b, Nahum et al. 2009, Andrade et al. 2011, Avelar et al. 2011). The rationale underlying the present work is that lineage-specific duplications may reflect molecular biodiversity and that the adaptation of organisms to different environments may ultimately help to identify potential therapeutic targets against parasitic diseases.

Our previous work identified lineage-specific gene duplications that led to the diversification of several families in *S. mansoni* (phylomedb.org). Furthermore, recent advances in proteomic analyses of schistosomes have revealed that peptidases are one of the main virulence factors involved in the pathogenesis of schistosomiasis (McKerrow et al. 2006, Kasný et al. 2009). In this work, we performed a phylogenomic analysis to address the two main questions of (i) whether peptidase families are expanded in the *S. mansoni* proteome and (ii) whether they share similar evolutionary histories.

Endopeptidase family members are duplicated in S. mansoni - To investigate which peptidase families are expanded in the *S. mansoni* genome, we explored the parasite phylome data available at PhylomeDB (Huerta-Cepas et al. 2011). Phylogenetic analyses were performed using an automated pipeline and a complete list of the paralogy relationships among the *S. mansoni* proteins was retrieved using a "species-overlap" algorithm that identifies family members originated by lineage-specific duplication events (Huerta-Cepas et al. 2007).

Based on the functional annotation available from the SchistoDB (Zerlotini et al. 2009) and UniProt (Apweiler et al. 2011) databases, the results revealed that the most significant peptidase expansions in the *S. mansoni* proteome corresponded to endopeptidases such as leishmanolysins, cercarial elastases and cathepsin D proteins. These enzymes belong to three distinct endopeptidase families, metallopeptidase M8 family (EC 3.4.24.-), serine peptidase S1 family (EC 3.4.21.-) and aspartic peptidase A1 family (EC 3.4.23.-), as described in the MEROPS database (Rawlings et al. 2010) and represent promising targets for vaccine and drug development.

In total, we identified 12 leishmanolysins, 13 cercarial elastases (Supplementary data) and 11 cathepsin D proteins (Supplementary data) in the predicted *S. mansoni* proteome. These proteins vary in length and sequence composition, but they are highly conserved with respect to the presence of a conserved sequence domain, which is distinct for each protein family as defined by the Pfam database (see details below). Currently, no crystal structure has been obtained for the *S. mansoni* peptidases described here.

Leishmanolysin (also called invadolysin) is a major surface peptidase member of the metallopeptidase M8 family. Leishmanolysins are believed to share the same mechanism used by the other zinc metalloproteinases, such as thermolysin. The conserved glutamate residue in the catalytic site acts in conjunction with a zinc ion to deprotonate and activate a water molecule. In turn, the activated water molecule acts as a nucleophile to attack the carbonyl of the peptide bond of a variety of substrates (Macdonald et al. 1995, Schlagenhauf et al. 1998). In *Leishmania*, these proteins are involved in different types of processes, such as the inhibition or perturbations of host cell interactions and the degradation of the extracellular matrix (Fitzpatrick et al. 2009). These proteins may have similar activities in schistosomes. Indeed, the *S. mansoni* protein, SmPepM8 (Smp_090100), is the second most abundant constituent in cercarial secretions, which provides insight on how it may contribute to tissue invasion by schistosomes and suggests this protein as a potential anti-parasitic target (Curwen et al. 2006, Fitzpatrick et al. 2009).

The catalytic triad of serine, histidine and aspartate residues is conserved in members of the serine protease family (Wilmouth et al. 2001, Hajjar et al. 2010). In elastases, this triad and an essential water molecule are involved in the catalysis. The peptide to be cleaved is bound noncovalently in the enzyme near the catalytic triad. In the first reaction step, the hydroxyl of the serine residue performs a nucleophilic attack on the substrate amide bond to form an ester. The amino terminus of the substrate is then covalently bound to the enzyme. The histidine residue abstracts a proton from a water molecule, which then attaches to the ester carbon to give rise to an oxyanion intermediate. Cercarial elastases play a key role in the penetration by the cercariae of mammalian skin to initiate infection and recent studies have revealed that these peptidases are also employed by the schistosomes to overcome or evade the host immune response (Salter et al. 2002, Aslam et al. 2008).

Cathepsin D is a member of the aspartic protease family. The active site of cathepsin D contains two aspartate residues, which perform an acid-base catalysis. This enzymatic mechanism involves the deprotonation of water by an ionised aspartate residue. This water molecule attacks the peptide carbonyl and there is a simultaneous protonation of the carbonyl oxygen by the other aspartate residue (e.g., Northrop 2001). Schistosome cathepsin D is involved in haemoglobin digestion, a process that provides the parasite with its main source of amino acid nutrients and that is essential for its development, growth and reproduction (Brindley et al. 2001, Caffrey et al. 2004, Delcroix et al. 2006). Given the essential function of cathepsin D in parasite nutrition and the ability of recombinant forms to cleave human immunoglobulin G, this protein is considered a potential target for novel anti-parasitic interventions (Verity et al. 2001, Morales et al. 2008).

The phylogenetic relationships of each endopeptidase family (Figs 2-4) are shown with protein sequences represented by identifiers in PhylomeDB (phylomedb.org) (Huerta-Cepas et al. 2011), UniProt (uniprot.org) (Apweiler et al. 2011) and/or SchistoDB (schistodb.net) (Zerlotini et al. 2009). In each phylogenetic tree, the *S. mansoni* endopeptidases form a well-supported clade of closely related proteins.

Together, the analysis of the *S. mansoni* proteome through an evolutionary approach identified endopeptidase family members that arose by gene duplication after the divergence of this parasite from the other eukaryotic species studied in this work. These lineage-specific duplications are related to the parasite's biology and evolution.

Leishmanolysins (metallopeptidase M8 family) - Our pipeline identified 12 *S. mansoni* leishmanolysins (Supplementary data). Proteins Smp_171330 and Smp_171340 are located in the same genomic region of Smp_090100 and Smp_090110, respectively, and could not be retrieved from the UniProt (Apweiler et al. 2011) and GeneDB (genedb.org) databases, which suggests that these genes were incorrectly annotated and probably deleted from these databases. Similar findings were obtained in two previous studies (Berriman et al. 2009, Bos et al. 2009).

To reconstruct the evolutionary history of *S. mansoni* leishmanolysins and their homologs in selected taxa, we performed a sequence alignment of 32 protein sequences identified as potential homologs by our pipeline. The trimmed alignment contained 1,822 sites, which cover most of the conserved protein domain identified in these proteins.

By analysing the phylogenetic tree (Fig. 2), it is possible to demonstrate that *S. mansoni* leishmanolysins have homologs in most species analysed in the present work, with the exception of *C. intestinalis* (tunicata) and fungi. However, this result does not completely discard the presence of homologous proteins in other organisms because they may be very divergent from the others in the database and therefore be missed by the pipeline search. The same is true for the other protein families mentioned in this paper.

Based on the information available in the literature and curated databases, three leishmanolysin homologs have been experimentally confirmed in *D. melanogaster*, *M. musculus* and *H. sapiens* and their function is related to the coordination of mitotic progression and cell migration (for details see Supplementary data). Although predicted functions or experimental evidence are not yet available, the metallopeptidase M8 family is also expanded in the sea anemone (*N. vectensis*) and sea urchin (*S. purpuratus*). The metallopeptidase M8 family also has more paralogs in the schistosomes (12 proteins)

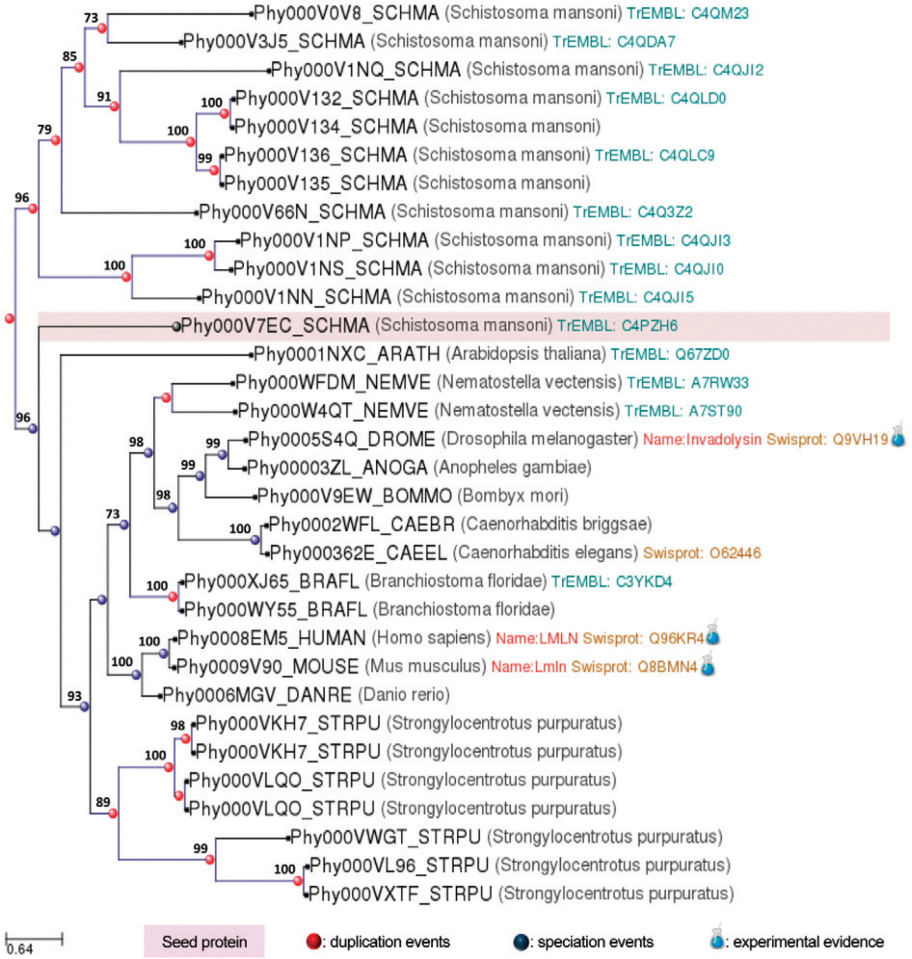


Fig. 2: phylogenetic relationships of the *S. mansoni* leishmanolysins and their homologs in selected taxa. Analysis was performed with trimmed sequence alignment by using the maximum likelihood method as implemented in PhyML. Best fit model (WAG) and support values for each node were estimated by the Akaike Likelihood Ratio Test (aLRT). Sequence labels follow the PhylomeDB internal identifier. For details, see Supplementary data.

compared to *H. sapiens* (4 proteins), which is in contrast with what is normally observed for the human peptidase families (Berriman et al. 2009).

A conserved protein domain (Pfam: PF01457), which characterises members of the metallopeptidase M8 family, was identified in all *S. mansoni* proteins analysed here (Fig. 5). Length variation and conservation of active sites were also observed. According to the Pfam profile HMMs, truncated domains were identified in all proteins, which possibly reflects the presence of different protein isoforms, as has been described elsewhere (Floris et al. 2008). The truncated domains could also indicate that parts of the sequences are missing at the N-terminal, C-terminal regions, or both due to annotation issues.

The data also reveals that the protein domain is duplicated in Smp_167090, Smp_167120 and Smp_135530. Seven *S. mansoni* proteins (Smp_090100, Smp_090110, Smp_127030, Smp_135530, Smp_153930, Smp_167090 and Smp_173070) were identified as active due to the presence of expected active site residues and metal ligand sites in the correct positions based on alignments with reference sequences, as previously described (Berriman et al. 2009).

Cercarial elastases (serine peptidase S1 family) - Our analysis identified a total of 13 cercarial elastases encoded in the *S. mansoni* genome (Supplementary data). Similar results were obtained by Berriman et al. (2009). However, with TreeFam (Ruan et al. 2008), Berriman et al. (2009) also

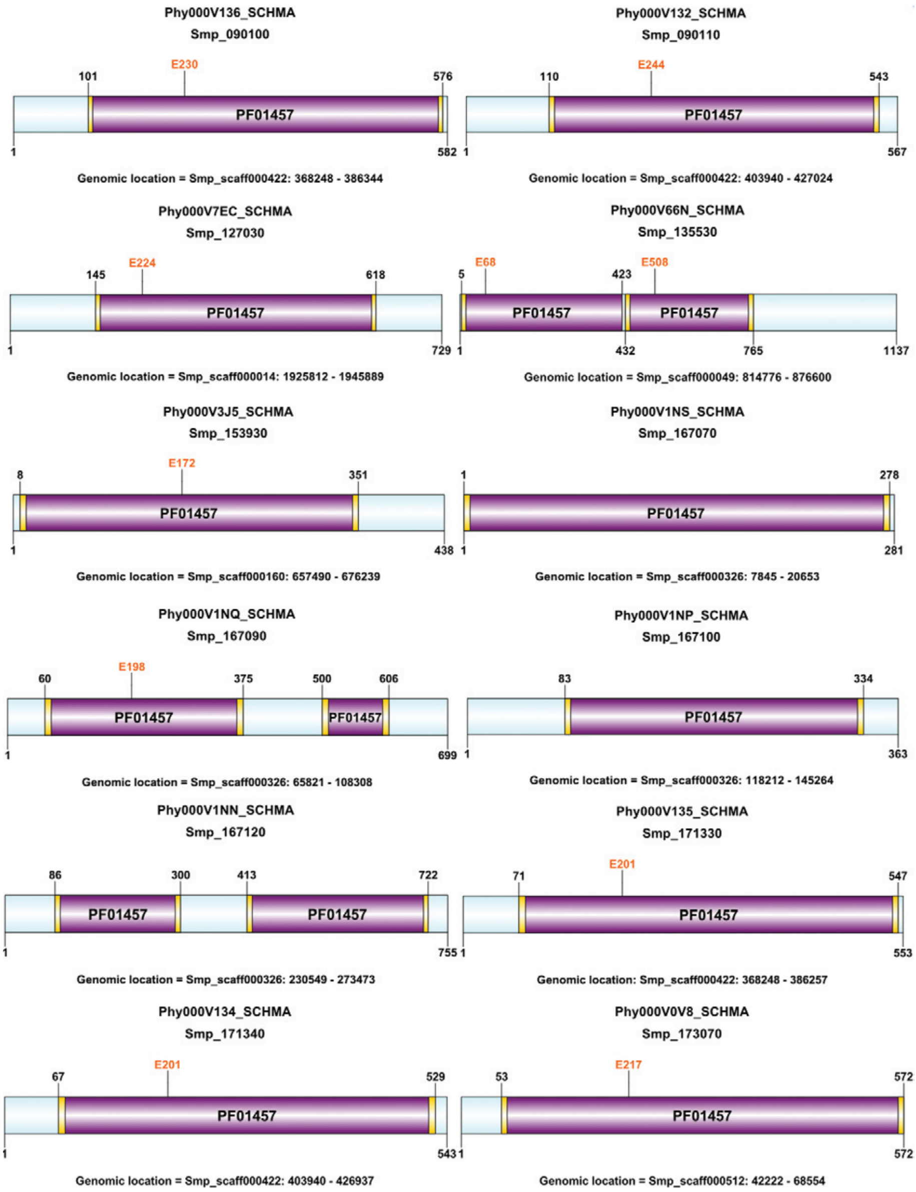


Fig. 3: conserved protein domain architecture of the *Schistosoma mansoni* leishmanolysins. Protein identifiers were assigned in SchistoDB. The conserved protein domain according to Pfam (PF01457) is present in all proteins. Truncated regions (yellow block) are indicated. Sequence length, domain limits, and active sites are also shown.

identified the Smp_192850 protein, which is annotated as a hypothetical protein and only contains 69 amino acids.

Two proteins, Smp_152560.2 and Smp_056680.2, are encoded in the same genomic location and could not be recovered in UniProt (Apweiler et al. 2011). Searches for the former protein in GeneDB (genedb.org) retrieved only the latter (Smp_056680), which indicated that the Smp_152560.2 gene was improperly annotated and thus was eliminated from both databases. In the original version of the *S. mansoni* genome, some sequences were interpreted as isoforms and different gene models were constructed. However, further studies indicated that these were actually mistakes in the genome assembly/annotation due to low sequence coverage. In the new version of the parasite genome, which is to be released by the Wellcome Trust Sanger Institute (sanger.ac.uk), many of these sequences have been collapsed.

Whole amino acid sequences from 35 proteins were aligned and filtered to remove gap-rich columns as previously described. The trimmed alignment contains 583 sites, which cover the conserved protein domain.

The phylogenetic analysis of the *S. mansoni* elastases and their homologs in the other species included in this work was performed as already described. The parasite elastases form a well-supported monophyletic clade, which suggests that these proteins originated from a common ancestor by gene duplication events followed by divergence in the *Schistosoma* lineage.

In observing the resulting phylogeny (Fig. 3), it is possible to demonstrate that *S. mansoni* elastases have homologs in six of the 16 other species considered in this analysis (*N. vectensis*, *D. melanogaster*, *An. gambiae*, *B. floridae*, *M. musculus* and *H. sapiens*). The serine peptidase S1 family is also expanded in all of these species except for one, *D. melanogaster*. According to the information available in UniProt (Apweiler et al. 2011), seven homologs have been experimentally confirmed in *D. melanogaster*, *M. musculus* and *H. sapiens,* and their function is related to a digestive process and immune response (Supplementary data). It is believed that similar activities are performed by elastases in schistosomes (Salter et al. 2002, Aslam et al. 2008).

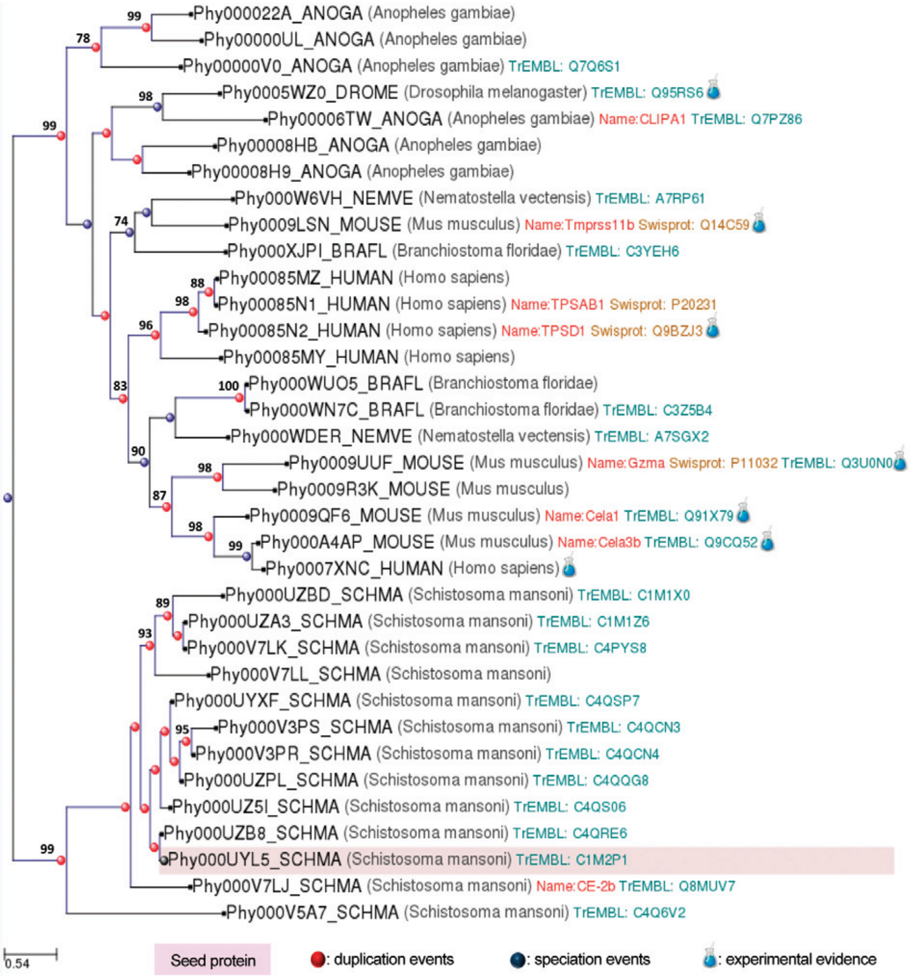


Fig. 4: phylogenetic relationships of the *Schistosoma mansoni* elastases and their homologs in selected taxa. Analysis was performed with trimmed sequence alignment by using the maximum likelihood method as implemented in PhyML. Best fit model (WAG) and support values for each node were estimated by the Akaike Likelihood Ratio Test (aLRT). Sequence labels follow the PhylomeDB internal identifier. For details, see Supplementary data.

A conserved protein sequence domain (Pfam: PF00089), which is found in all characterised members of the serine peptidase S1 family, was identified in the *S. mansoni* elastases and ranges in length from 141-265 amino acids (Fig. 6). The catalytic triad of histidine, aspartate and serine residues is present in most of these proteins. Based on profile HMMs available in Pfam, truncated regions were assigned to all 12 of these elastases, perhaps reflecting their degree of divergence in relation to other proteins in the database. Meanwhile, it is important to emphasise that protein databases do not cover all of the existing diversity in nature.

Together, these results indicate that the correct number of cercarial elastases encoded in the *S. mansoni* genome is 12 and not 13 as described before. However, only Smp_006510, Smp_006520 and Smp_141450 were previously predicted as active proteins (Berriman et al. 2009). Smp_194800 has a much shorter domain compared to others. This difference could reflect either the presence of an elastase pseudogene in the parasite genome or that the sequence was incorrectly annotated due to an error in the gene model. Considering that the first-pass annotation of the *S. mansoni* genome was produced by a combination of gene-finding algorithms (Augustus,

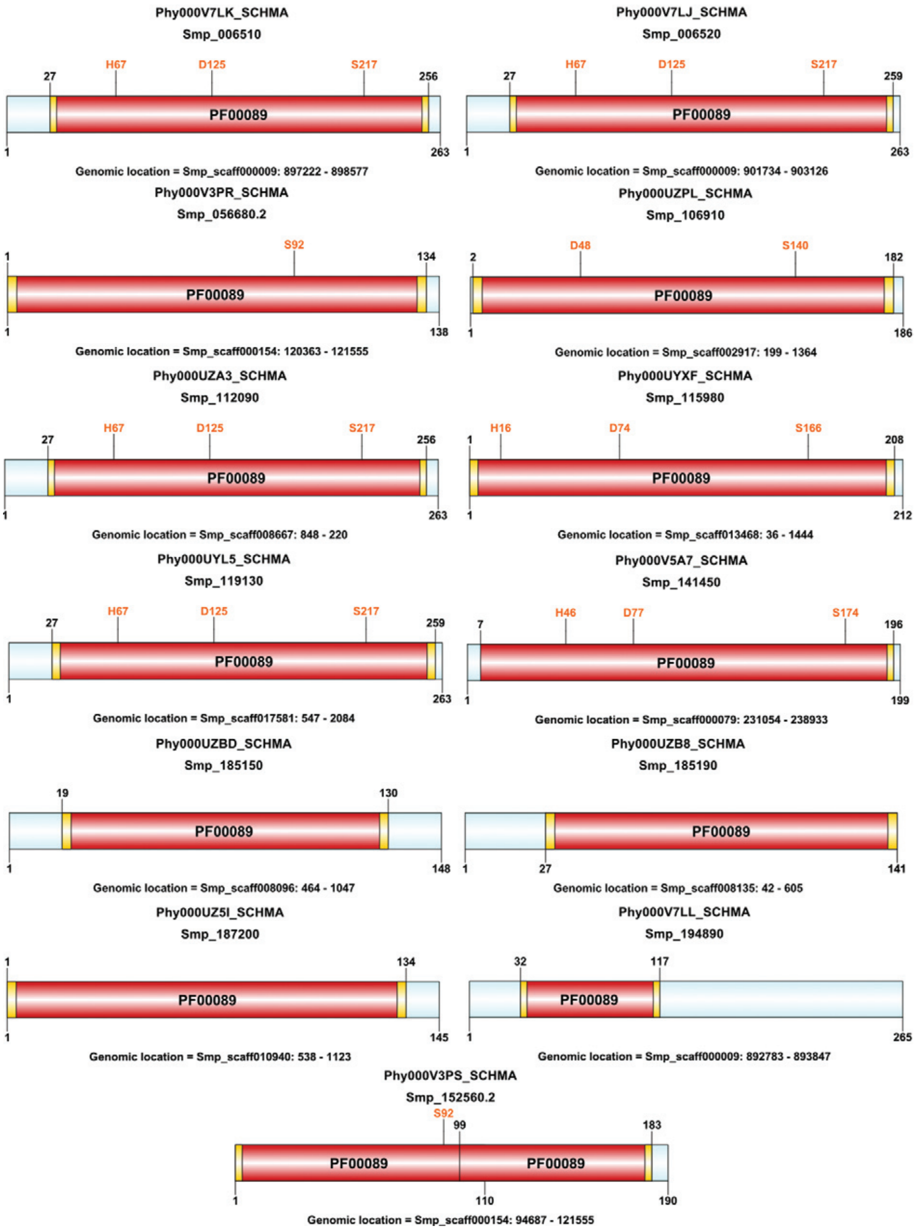


Fig. 5: conserved protein domain architecture of the *Schistosoma mansoni* elastases. Protein identifiers were assigned in SchistoDB. The conserved protein domain according to Pfam (PF00089) is present in all proteins. Truncated regions (yellow block) are indicated. Sequence length, domain limits, and the catalytic triad of histidine (H), aspartate (D), and serine (S) are also shown.

Fig. 6: phylogenetic relationships of the *Schistosoma mansoni* cathepsins D and their homologs in selected taxa. Analysis was performed with trimmed sequence alignment by using the maximum likelihood method as implemented in PhyML. Best fit model (WAG) and support values for each node were estimated by the Akaike Likelihood Ratio Test (aLRT). Sequence labels follow the PhylomeDB internal identifier. For details, see Supplementary data.

Twinscan and GlimmerHMM) (Berriman et al. 2009), this genome has not received extensive manual curation and therefore, many gene models will be refined in the future. Furthermore, EVidenceModeler (Haas et al. 2008) has also been used to incorporate expressed sequence tag (EST) evidence into the data.

Cathepsin D proteins (aspartic peptidase A1 family) - Our pipeline identified 11 *S. mansoni* cathepsin D proteins (Supplementary data) that were duplicated after the divergence of *S. mansoni* from the other metazoans analysed here. The evolutionary history of cathepsin D proteins was reconstructed from the sequence alignment of 111 protein sequences from *S. mansoni* and the selected taxa. The final trimmed alignment contained 1,676 sites, which covered most of the conserved protein domain (Pfam: PF00026). Two *S. mansoni* proteins corresponded to alternative splicing products (Smp_136830.2 and Smp_013040.2). Similar results were found by Berriman et al. (2009).

The phylogenetic tree indicates that the *S. mansoni* cathepsin D proteins have homologs in all but one species (*S. purpuratus*) analysed in this work (Fig. 4). The aspartic peptidase A1 family has also been expanded in 12 of the 15 species in which homologous proteins were identified (*A. thaliana*, *U. maydis*, *S. cerevisiae*, *N. crassa*, *C. elegans*, *C. briggsae*, *D. melanogaster*, *C. intestinalis*, *B. floridae*, *D. rerio*, *M. musculus* and *H. sapiens*). The number of paralogous proteins ranges from two-17 and includes different aspartic peptidases, such as pepsins, renins, gastricsin and cathepsin D proteins. Based on the information available in the literature and curated databases, these homologous proteins are involved in digestion and protein degradation (Supplementary data). In schistosomes, cathepsin D proteins play an integral role in haemoglobin proteolysis (Brindley et al. 2001, Caffrey et al. 2004, Delcroix et al. 2006).

To predict the protein domain architecture of *S. mansoni* cathepsin D proteins, we applied the same methodology as previously described. The conserved domain (Pfam: PF00026), which has been found in all characterised aspartic peptidase A1 family members, was also identified in the *S. mansoni* proteins with sequence lengths ranging from 94-430 amino acids (Fig. 7). Active sites are also indicated. Based on the profile HMMs available in Pfam, truncated regions were observed in the N-terminal, C-terminal or both regions. The data also indicate that an additional short sequence domain (Pfam: PF07966), which is known as the A1 propeptide domain, is present at the N-terminal region of two *S. mansoni* proteins, Smp_013040.1 and Smp_013040.2. Smp_136840 has a much shorter domain compared to other proteins in the same family.

In a previous study, four *S. mansoni* cathepsin D proteins (Smp_013040.1, Smp_013040.2, Smp_136730 and Smp_136830.2) were identified as active proteins (Berriman et al. 2009), but the variation in the domain architecture and its implications in functional complexity were not investigated. One interesting study would be to analyse the functional properties of Smp_013040.1 and Smp_013040.2, which contain the A1 propeptide domain (PF07966).

DISCUSSION

We found that three endopeptidase families are expanded in the helminth parasite *S. mansoni*, which include members of the metallopeptidases (M8 family), serine peptidases (S1 family) and aspartic peptidases (A1 family). In this work, a comparative analysis of these three protein families in *S. mansoni* and 16 other eukaryotic proteomes revealed their distinct evolutionary histories and provided further information with respect to the sequence and functional features of the parasite family members.

Based on the *S. mansoni* genomic data, 335 peptidases were identified, which comprise 2.5% of the predicted proteome (Berriman et al. 2009). They include members of five major classes of peptidases (aspartic, cysteine, metallo, serine and threonine). Of the 61 peptidase families, 44 are expanded in this parasite and the number of paralogous proteins range from two-26.

Using a computational approach, Bos et al. (2009) analysed all putative peptidases encoded in the parasite's genome in addition to using EST data, which is similar to work by Berriman et al. (2009). After removing redundant sequences, inactive homologs, likely pseudogenes and sequences smaller than 100 amino acids from the dataset, they identified a total of 255 peptidase sequences from the five catalytic classes.

Our results are not fully comparable to those obtained by Bos et al. (2009) with respect to elastases and cathepsin D proteins. However, it is worth noting that the phylogenetic analysis of the serine peptidase S1 family performed by these authors also indicated a well-supported clade of four *S. mansoni* elastases, which are corroborated by our findings. The other homologs with high similarities to the cercarial elastases were likely pseudogenes and, for this reason, they were excluded from the analysis by Bos et al. (2009).

Our results suggest that *Schistosoma* members of these endopeptidase families originated from successive gene duplication events in the parasite lineage after its diversification from the other metazoans analysed here. These results were corroborated by previous proteomic and phylogenetic analyses on *Fasciola hepatica* peptidases, which showed that the repertoire of virulence-associated cathepsin L proteins was established by a series of gene duplication events (Irving et al. 2003, Robinson et al. 2008). These studies also indicate that the gene duplications were followed by active site residue refinements, which interfere with the substrate specificity of the *F. hepatica* cathepsin L proteins. Whether the *S. mansoni* proteins share a similar refinement remains to be established.

Gene duplication followed by divergence is known to be the most predominant mechanism of molecular evolution and represents the main source of raw material for the generation of new genes and proteins through the processes of neo and sub-functionalisation (Ohno 1970, Conant & Wolfe 2008, Nahum & Pereira 2008, Hamilton et al. 2009). Although in some cases sequences have diverged to the extent that it is impossible to recognise homologous relationships, different proteins that arose by gene duplication may be related at distinct levels, such as sequence, structure, function or a combination of these features and can be grouped into families and superfamilies (Nahum & Pereira 2008).

Gene fusion, gene fission and domain shuffling were not observed as mechanisms shaping the evolution of the *S. mansoni* endopeptidase families analysed in this work. Whether gene fusion/fission also plays a role in the evolution of the *S. mansoni* genome will be a subject of a future work. Our previous study indicated that domain shuffling is one of the main evolutionary forces driving the sequence and functional diversification of the protein kinases of this parasite (Andrade et al. 2011, Avelar et al. 2011).

Peptidases have been implicated in various processes that are crucial to the development and survival of parasites, including host invasion, degradation of haemoglobin in blood feeding, immune evasion and activation of inflammation (McKerrow et al. 2006, Kasný et al. 2009).

Experimental work suggests that the SmPepM8 metallopeptidase (leishmanolysin) may contribute to tissue invasion by schistosome cercariae. This peptidase was the second most abundant protein released during the transformation of *S. mansoni* cercariae into schistosomula (Curwen et al. 2006). Leishmanolysins are a major surface peptidase member of the metallopeptidase M8 family, which in leishmaniasis are involved in different types of processes, such as the inhibition or perturbation of host cell interactions and the degradation of the extracellular matrix (Fitzpatrick et al. 2009). It is speculated that these proteins could perform similar activities in schistosomes during host-parasite interactions (Curwen et al. 2006, Fitzpatrick et al. 2009).

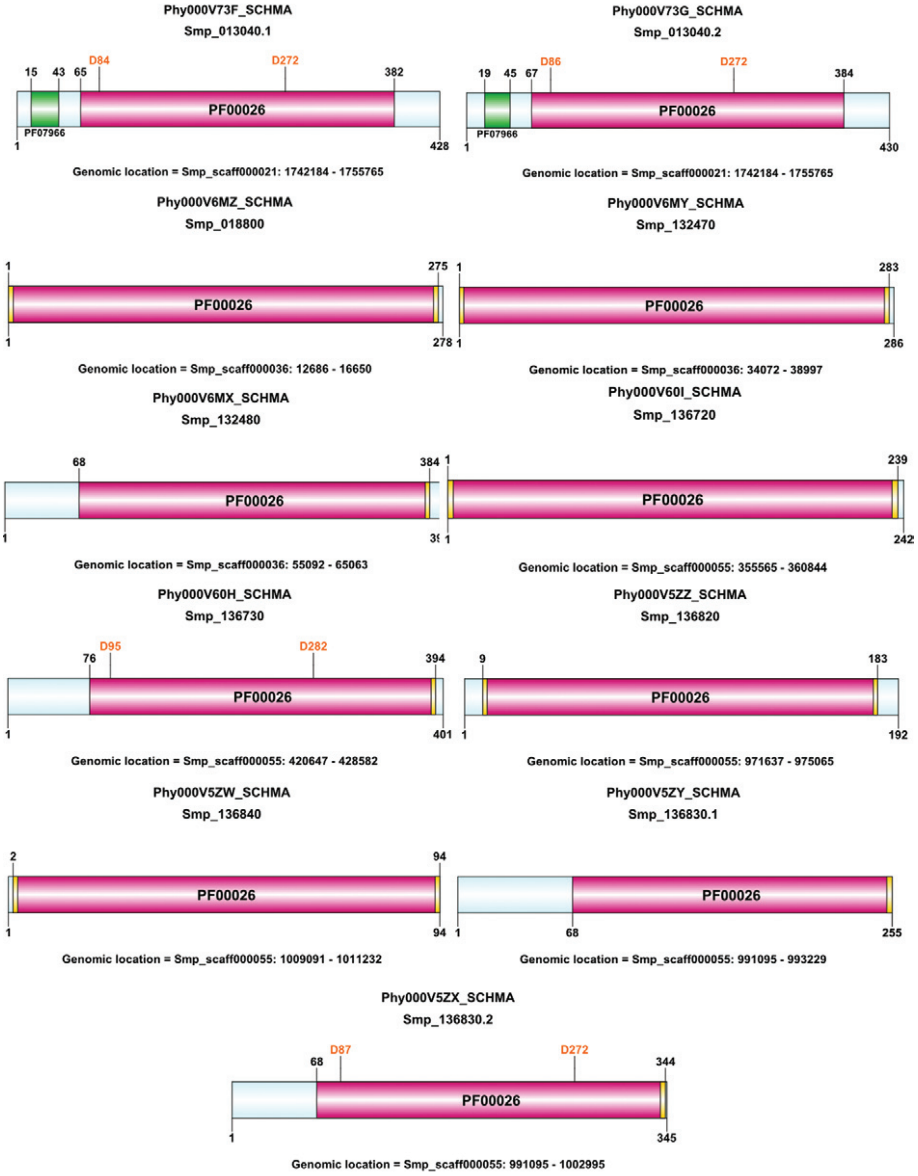


Fig. 7: conserved protein domain architecture of the *S. mansoni* cathepsin Ds. Protein identifiers were assigned in SchistoDB. The conserved protein domain according to Pfam (PF00026) is present in all proteins. The PF07966 additional N-terminal domain was identified in two proteins. Truncated regions (yellow block) are indicated. Sequence length, domain limits, and active sites are also shown.

Invasion of host skin is the initial event in establishing an infection in mammalian hosts. Considering the complexity of host skin barriers that the cercariae must go through during invasion, it has been suggested that multiple enzyme activities are required for this process (Salter et al. 2002). However, only one peptidase (cercarial elastase) has been identified as a major secretory product released during skin penetration (Knudsen et al. 2005, Hansell et al. 2008). These proteins may also be involved in eliminating the outer layer of the cercariae during transformation. Although cercarial elastases were named based on their ability to degrade insoluble elastin, numerous substrates for these enzymes have been identified, which include collagen, keratin and extracellular matrix proteins (Salter et al. 2002, McKerrow 2003, Knudsen et al. 2005).

Orthologous genes encoding elastase proteins were found in *Schistosoma haematobium*, *Schistosoma japonicum* and *Schistosoma douthitti* (Salter et al. 2002, Zhou et al. 2009). The expression of *S. japonicum* cercarial elastases was confirmed in both the sporocyst and cercarial stages and evidence that this peptidase is released by the parasite during the invasion of mammalian skin was obtained by anti-recombinant SjCE antibodies in infected mouse skin (Zhou et al. 2009). However, orthologous peptidases to *S. mansoni* cercarial elastases were not detected in the acetabular secretions of *S. japonicum* (Dvorák et al. 2008). Furthermore, the faster penetration by *S. japonicum* into the host skin may reflect the differential use of proteolytic enzymes in addition to those characterised in *S. mansoni* or even involve new peptidases not yet characterised (Chlichlia et al. 2005, He et al. 2005). Recent studies have also demonstrated that *S. mansoni* elastases are capable of cleaving IgE molecules from human, mouse and rat, indicating that the parasite may be able to overcome or evade the IgE response (Aslam et al. 2008). However, this subject remains controversial.

The biological complexity of *S. mansoni* is related to evolutionary innovations that took place before and after its diversification from other metazoans. Because duplicated genes are important substrates for improving an organism's adaptation to its environment, understanding how members of protein families evolved may link evolutionary studies to parasite biology. In turn, this knowledge will provide insights into host-parasite relationships and accelerate the identification of novel vaccine and drug targets aimed at the treatment and eradication of schistosomiasis.

In conclusion, this paper provides an evolutionary view of three *S. mansoni* peptidase families, thus allowing for a deeper understanding of the genomic complexity and lineage-specific adaptations potentially related to the parasitic lifestyle. In the future, our results obtained using a systemic approach (proteome-wide analyses) may accelerate the understanding of schistosomiasis, its etiologic agents and host-parasite interactions and optimise the discovery of therapeutic targets for the development of new drugs and vaccines.

ACKNOWLEDGEMENTS

To the use of the computing resources of CRG (Spain) and CEBio-Fiocruz (Brazil), to Jaime Huerta-Cepas (CRG), Eric Aguiar and Francislon Silva (CEBio), for bioinformatics technical support, to Mariana de Oliveira (CEBio), for help with illustrations, and to the two anonymous reviewers, for the valuable suggestions that improved this paper.

REFERENCES

Abdulla MH, Lim KC, Sajid M, McKerrow JH, Caffrey CR 2007. Schistosomiasis *mansoni*: novel chemotherapy using a cysteine protease inhibitor. *PLoS Med 4*: e14.

Akaike H 1973. Information theory and an extension of the maximum likelihood principle. In BN Petrov, F Csaki (eds.), *Second International Symposium on Information Theory*, Academiai Kiado, Budapest, p. 267-281.

Andrade LF, Nahum LA, Avelar LG, Silva LL, Zerlotini A, Ruiz JC, Oliveira G 2011. Eukaryotic protein kinases (ePKs) of the helminth parasite *Schistosoma mansoni*. *BMC Genomics 12*: 215.

Anisimova M, Gascuel O 2006. Approximate likelihood-ratio test for branches: a fast, accurate and powerful alternative. *Syst Biol 55*: 539-552.

Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Antunes R, Barrell D, Bely B, Bingley M, Binns D, Bower L, Browne P, Chan WM, Dimmer E, Eberhardt R, Fazzini F, Fedotov A, Foulger R, Garavelli J, Castro LG, Huntley R, Jacobsen J, Kleen M, Laiho K, Legge D, Lin Q, Liu W, Luo J, Orchard S, Patient S, Pichler K, Poggioli D, Pontikos N, Pruess M, Rosanoff S, Sawford T, Sehra H, Turner E, Corbett M, Donnelly M, van Rensburg P, Xenarios I, Bougueleret L, Auchincloss A, Argoud-Puy G, Axelsen K, Bairoch A, Baratin D, Blatter MC, Boeckmann B, Bolleman J, Bollondi L, Boutet E, Quintaje SB, Breuza L, Bridge A, deCastro E, Coudert E, Cusin I, Doche M, Dornevil D, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gehant S, Ferro S, Gasteiger E, Gateau A, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, Hulo N, James J, Jimenez S, Jungo F, Kappler T, Keller G, Lara V, Lemercier P, Lieberherr D, Martin X, Masson P, Moinat M, Morgat A, Paesano S, Pedruzzi I, Pilbout S, Poux S, Pozzato M, Redaschi N, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stanley E, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Wu CH, Arighi CN, Arminski L, Barker WC, Chen C, Chen Y, Dubey P, Huang H, Mazumder R, McGarvey P, Natale DA, Natarajan TG, Nchoutmboube J, Roberts NV, Suzek BE, Ugochukwu U, Vinayaka CR, Wang Q, Wang Y, Yeh LS, Zhang J 2011. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res 39*: D214-219.

Aslam A, Quinn P, McIntosh RS, Shi J, Ghumra A, McKerrow JH, Bunting KA, Dunne DW, Doenhoff MJ, Morrison SL, Zhang K, Pleass RJ 2008. Proteases from *Schistosoma mansoni* cercariae cleave IgE at solvent exposed interdomain regions. *Mol Immunol 45*: 567-574.

Avelar LG, Nahum LA, Andrade LF, Oliveira G 2011. Functional diversity of the *Schistosoma mansoni* tyrosine kinases. *J Signal Transduct*: 603290.

Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyama ST, Al-Lazikani B, Andrade LF, Ashton PD, Aslett MA, Bartholomeu DC, Blandin G, Caffrey CR, Coghlan A, Coulson R, Day TA, Delcher A, DeMarco R, Djikeng A, Eyre T, Gamble JA, Ghedin E, Gu Y, Hertz-Fowler C, Hirai H, Hirai Y, Houston R, Ivens A, Johnston DA, Lacerda D, Macedo CD, McVeigh P, Ning Z, Oliveira G, Overington JP, Parkhill J, Pertea M, Pierce RJ, Protasio AV, Quail MA, Rajandream MA, Rogers J, Sajid M, Salzberg SL, Stanke M, Tivey AR, White O, Williams DL, Wortman J, Wu W, Zamanian M, Zerlotini A, Fraser-Liggett CM, Barrell BG, El-Sayed NM 2009. The genome of the blood fluke *Schistosoma mansoni*. *Nature 460*: 352-358.

Bos DH, Mayfield C, Minchella DJ 2009. Analysis of regulatory protease sequences identified through bioinformatic data mining of the *Schistosoma mansoni* genome. *BMC Genomics 10*: 488.

Botros SS, Bennett JL 2007. Praziquantel resistance. *Expert Opin Drug Discov 2* (Suppl. 1): 35-40.

Brindley PJ, Kalinna BH, Wong JY, Bogitsh BJ, King LT, Smyth DJ, Verity CK, Abbenante G, Brinkworth RI, Fairlie DP, Smythe ML, Milburn PJ, Bielefeldt-Ohmann H, Zheng Y, McManus DP 2001. Proteolysis of human hemoglobin by schistosome cathepsin D. *Mol Biochem Parasitol 112*: 103-112.

Caffrey CR, McKerrow JH, Salter JP, Sajid M 2004. Blood 'n' guts: an update on schistosome digestive peptidases. *Trends Parasitol 20*: 241-248.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics 25*: 1972-1973.

Chlichlia K, Schauwienold B, Kirsten C, Doenhoff MJ, Fishelson Z, Ruppel A 2005. *Schistosoma japonicum* reveals distinct reactivity with antisera directed to proteases mediating host infection and invasion by cercariae of *S. mansoni* or *S. haematobium*. *Parasite Immunol 27*: 97-102.

Conant GC, Wolfe KH 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet 9*: 938-950.

Curwen RS, Ashton PD, Sundaralingam S, Wilson RA 2006. Identification of novel proteases and immunomodulators in the secretions of schistosome cercariae that facilitate host entry. *Mol Cell Proteomics 5*: 835-844.

Delcroix M, Sajid M, Caffrey CR, Lim KC, Dvorák J, Hsieh I, Bahgat M, Dissous C, McKerrow JH 2006. A multienzyme network functions in intestinal protein digestion by a platyhelminth parasite. *J Biol Chem 281*: 39316-39329.

Dvorák J, Mashiyama ST, Braschi S, Sajid M, Knudsen GM, Hansell E, Lim KC, Hsieh I, Bahgat M, Mackenzie B, Medzihradszky KF, Babbitt PC, Caffrey CR, McKerrow JH 2008. Differential use of protease families for invasion by schistosome cercariae. *Biochimie 90*: 345-358.

Edgar RC 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res 32*: 1792-1797.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A 2010. The Pfam protein families database. *Nucleic Acids Res 38*: D211-222.

Fitch WM 1970. Distinguishing homologous from analogous proteins. *Syst Zool 19*: 99-113.

Fitzpatrick JM, Peak E, Perally S, Chalmers IW, Barrett J, Yoshino TP, Ivens AC, Hoffmann KF 2009. Anti-schistosomal intervention targets identified by lifecycle transcriptomic analyses. *PLoS Negl Trop Dis 3*: e543.

Floris M, Orsini M, Thanaraj TA 2008. Splice-mediated variants of proteins (SpliVaP) - data and characterization of changes in signatures among protein isoforms due to alternative splicing. *BMC Genomics 9*: 453.

Fujinaga M, Cherney MM, Oyama H, Oda K, James MN 2004. The molecular structure and catalytic mechanism of a novel carboxyl peptidase from *Scytalidium lignicolum*. *Proc Natl Acad Sci USA 101*: 3364-3369.

Gabaldón T 2008. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol 9*: 235.

Guindon S, Gascuel O 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol 52*: 696-704.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol 9*: R7.

Hajjar E, Broemstrup T, Kantari C, Witko-Sarsat V, Reuter N 2010. Structures of human proteinase 3 and neutrophil elastase - so similar yet so different. *FEBS J 277*: 2238-2254.

Hamilton PB, Adams ER, Njiokou F, Gibson WC, Cuny G, Herder S 2009. Phylogenetic analysis reveals the presence of the *Trypanosoma cruzi* clade in African terrestrial mammals. *Infect Genet Evol 9*: 81-86.

Han ZG, Brindley PJ, Wang SY, Chen Z 2009. Schistosoma genomics: new perspectives on schistosome biology and host-parasite interaction. *Annu Rev Genomics Hum Genet 10*: 211-240.

Hansell E, Braschi S, Medzihradszky KF, Sajid M, Debnath M, Ingram J, Lim KC, McKerrow JH 2008. Proteomic analysis of skin invasion by blood fluke larvae. *PLoS Negl Trop Dis 2*: e262.

He YX, Salafsky B, Ramaswamy K 2005. Comparison of skin invasion among three major species of *Schistosoma*. *Trends Parasitol 21*: 201-203.

Hedstrom L 2002. Serine protease mechanism and specificity. *Chem Rev 102*: 4501-4524.

Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Denisov I, Kormes D, Marcet-Houben M, Gabaldón T 2011. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res 39*: D556-560.

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T 2007. The human phylome. *Genome Biol 8*: R109.

Huerta-Cepas J, Dopazo J, Gabaldón T 2010a. ETE: a python environment for tree exploration. *BMC Bioinformatics 11*: 24.

Huerta-Cepas J, Marcet-Houben M, Pignatelli M, Moya A, Gabaldón T 2010b. The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for *Acyrthosiphon pisum* genes. *Insect Mol Biol 19* (Suppl. 2): 13-21.

Huzurbazar S, Kolesov G, Massey SE, Harris KC, Churbanov A, Liberles DA 2010. Lineage-specific differences in the amino acid substitution process. *J Mol Biol 396*: 1410-1421.

Irving JA, Spithill TW, Pike RN, Whisstock JC, Smooker PM 2003. The evolution of enzyme specificity in *Fasciola* spp. *J Mol Evol 57*: 1-15.

James MNG 2004. Catalytic pathway of aspartic peptidases. In AJ Barrett, ND Rawlings, JF Woessner (eds.), *Handbook of proteolytic enzymes*, Elsevier Science, London, p. 12-19.

Kasný M, Mikes L, Hampl V, Dvorák J, Caffrey CR, Dalton JP, Horák P 2009. Chapter 4. Peptidases of trematodes. *Adv Parasitol 69*: 205-297.

Knudsen GM, Medzihradszky KF, Lim KC, Hansell E, McKerrow JH 2005. Proteomic analysis of *Schistosoma mansoni* cercarial secretions. *Mol Cell Proteomics 4*: 1862-1875.

Liang YS, Dai JR, Zhu YC, Coles GC, Doenhoff MJ 2003. Genetic analysis of praziquantel resistance in *Schistosoma mansoni*. *Southeast Asian J Trop Med Public Health 34*: 274-280.

Macdonald MH, Morrison CJ, McMaster WR 1995. Analysis of the active site and activation mechanism of the *Leishmania* surface metalloproteinase GP63. *Biochim Biophys Acta 1253*: 199-207.

McKerrow JH, Caffrey C, Kelly B, Loke P, Sajid M 2006. Proteases in parasitic diseases. *Annu Rev Pathol 1*: 497-536.

McKerrow JJ 2003. Invasion of skin by schistosome cercariae: some neglected facts - Response. *Trends Parasitol 19*: 63-66.

Melman SD, Steinauer ML, Cunningham C, Kubatko LS, Mwangi IN, Wynn NB, Mutuku MW, Karanja DM, Colley DG, Black CL, Secor WE, Mkoji GM, Loker ES 2009. Reduced susceptibility to praziquantel among naturally occurring Kenyan isolates of *Schistosoma mansoni*. *PLoS Negl Trop Dis 3*: e504.

Mistry J, Bateman A, Finn RD 2007. Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics 8*: 298.

Morales ME, Rinaldi G, Gobert GN, Kines KJ, Tort JF, Brindley PJ 2008. RNA interference of *Schistosoma mansoni* cathepsin D, the apical enzyme of the hemoglobin proteolysis cascade. *Mol Biochem Parasitol 157*: 160-168.

Nahum LA, Goswami S, Serres MH 2009. Protein families reflect the metabolic diversity of organisms and provide support for functional prediction. *Physiol Genomics 38*: 250-260.

Nahum LA, Pereira SL 2008. Phylogenomics, protein family evolution and the tree of life: an integrated approach between molecular evolution and computational intelligence. In TG Smolinski, MG Milanova, AE Hassanien (eds.), *Studies in computational intelligence (SCI)*, Springer-Verlag, Berlin, pp. 259-279.

Northrop DB 2001. Follow the protons: a low-barrier hydrogen bond unifies the mechanisms of the aspartic proteases. *Acc Chem Res 34*: 790-797.

Ohno S 1970. *Evolution by gene duplication*, 1st ed., Springer-Verlag, Heidelberg, 160 pp.

Pica-Mattoccia L, Cioli D 2004. Sex- and stage-related sensitivity of *Schistosoma mansoni* to *in vivo* and *in vitro* praziquantel treatment. *Int J Parasitol 34*: 527-533.

Rawlings ND, Barrett AJ 1993. Evolutionary families of peptidases. *Biochem J 290*: 205-218.

Rawlings ND, Barrett AJ, Bateman A 2010. MEROPS: the peptidase database. *Nucleic Acids Res 38*: D227-233.

Ren J, Wen L, Gao X, Jin C, Xue Y, Yao X 2009. DOG 1.0: illustrator of protein domain structures. *Cell Res 19*: 271-273.

Robinson MW, Tort JF, Lowther J, Donnelly SM, Wong E, Xu W, Stack CM, Padula M, Herbert B, Dalton JP 2008. Proteomics and phylogenetic analysis of the cathepsin L protease family of the helminth pathogen *Fasciola hepatica*: expansion of a repertoire of virulence-associated factors. *Mol Cell Proteomics 7*: 1111-1123.

Ruan J, Li H, Chen Z, Coghlan A, Coin LJ, Guo Y, Hériché JK, Hu Y, Kristiansen K, Li R, Liu T, Moses A, Qin J, Vang S, Vilella AJ, Ureta-Vidal A, Bolund L, Wang J, Durbin R 2008. TreeFam: 2008 Update. *Nucleic Acids Res 36*: D735-740.

Salter JP, Choe Y, Albrecht H, Franklin C, Lim KC, Craik CS, McKerrow JH 2002. Cercarial elastase is encoded by a functionally conserved gene family across multiple species of schistosomes. *J Biol Chem 277*: 24618-24624.

Sargeant TJ, Marti M, Caler E, Carlton JM, Simpson K, Speed TP, Cowman AF 2006. Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biol 7*: R12.

Schlagenhauf E, Etges R, Metcalf P 1998. The crystal structure of the *Leishmania major* surface proteinase leishmanolysin (gp63). *Structure 6*: 1035-1046.

Smith TF, Waterman MS 1981. Identification of common molecular subsequences. *J Mol Biol 147*: 195-197.

Steinmann P, Keiser J, Bos R, Tanner M, Utzinger J 2006. Schistosomiasis and water resources development: systematic review, meta-analysis and estimates of people at risk. *Lancet Infect Dis 6*: 411-425.

van der Werf MJ, de Vlas SJ, Brooker S, Looman CW, Nagelkerke NJ, Habbema JD, Engels D 2003. Quantification of clinical morbidity associated with schistosome infection in sub-Saharan Africa. *Acta Trop 86*: 125-139.

Verity CK, McManus DP, Brindley PJ 2001. Vaccine efficacy of recombinant cathepsin D aspartic protease from *Schistosoma japonicum*. *Parasite Immunol 23*: 153-162.

Wilmouth RC, Edman K, Neutze R, Wright PA, Clifton IJ, Schneider TR, Schofield CJ, Hajdu J 2001. X-ray snapshots of serine protease catalysis reveal a tetrahedral intermediate. *Nat Struct Biol 8*: 689-694.

Wu DD, Wang GD, Irwin DM, Zhang YP 2009. A profound role for the expansion of trypsin-like serine protease family in the evolution of hematophagy in mosquito. *Mol Biol Evol 26*: 2333-2341.

Zerlotini A, Heiges M, Wang H, Moraes RL, Dominitini AJ, Ruiz JC, Kissinger JC, Oliveira G 2009. SchistoDB: a *Schistosoma mansoni* genome resource. *Nucleic Acids Res 37*: D579-582.

Zhou Y, Zheng H, Chen Y, Zhang L, Wang K 2009. The *Schistosoma japonicum* genome reveals features of host-parasite interplay. *Nature 460*: 345-351.

3.3 - CAPÍTULO III: Filogenômica e Anotação Funcional de Proteína Quinases

Andrade LF, Nahum LA, Avelar LG, Silva LL, Zerlotini A, Ruiz JC, Oliveira G. Eukaryotic protein kinases (ePKs) of the helminth parasite *Schistosoma mansoni*. BMC Genomics. 2011 May 6;12:215. PMID: 21548963.

Neste estudo em colaboração com o Dr. Jerônimo Ruiz (CPqRR/FIOCRUZ) e Adhemar Zerlotini (EMBRAPA), analisamos as proteína quinases eucarióticas de *S. mansoni* (ePKs). Estas proteínas desempenham um papel central na mediação da transdução de sinal em redes complexas e são consideradas alvos para o desenvolvimento de fármacos. A identificação e classificação das ePKs de *S. mansoni* foi realizada utilizando-se diferentes abordagens computacionais, como a construção de modelos ocultos de Markov (HMMs) e reconstrução filogenética. HMMs representando a sequência do domínio catalítico destas enzimas foram utilizados nas buscas por similaridade para recuperar potenciais homólogos. Desta forma, um total de 252 ePKs de *S. mansoni* foram identificadas. O total de ePKs corresponde a cerca de 2.0% do proteoma predito de *S. mansoni*, indicando que a fosforilação de proteínas é um mecanismo importante para a regulação do complexo ciclo de vida deste parasito. Utilizando-se uma abordagem filogenômica, analisamos as ePKs de *S. mansoni* em comparação com seis outros eucariotos: *Saccharomyces cerevisiae, Caenorhabditis elegans*, *Brugia malayi, Drosophila melanogaster*, *Mus musculus* e *Homo sapiens*. Através da reconstrução da história evolutiva destas proteínas, foi possível classificá-las em grupos, famílias e, em alguns casos, em subfamílias proteicas conforme a hierarquia adotada pelo KinBase (http://kinase.com/kinbase). Além de identificar e classificar as ePKs de *S. mansoni,* este estudo promoveu a melhoria da anotação funcional de mais de 40% das ePKs codificadas no genoma deste parasito e apresentou um cenário das relações evolutivas dessas enzimas em relação aos potenciais homólogos nos demais organismos selecionados, além de identificar potenciais alvos para o desenvolvimento de novas drogas contra a esquistossomose.

Andrade *et al. BMC Genomics* 2011, **12**:215
http://www.biomedcentral.com/1471-2164/12/215

BMC Genomics

RESEARCH ARTICLE **Open Access**

Eukaryotic Protein Kinases (ePKs) of the Helminth Parasite *Schistosoma mansoni*

Luiza F Andrade[1†], Laila A Nahum[1,2†], Lívia GA Avelar[1,3], Larissa L Silva[1,2,3], Adhemar Zerlotini[2], Jerônimo C Ruiz[1] and Guilherme Oliveira[1,2*]

Abstract

Background: Schistosomiasis remains an important parasitic disease and a major economic problem in many countries. The *Schistosoma mansoni* genome and predicted proteome sequences were recently published providing the opportunity to identify new drug candidates. Eukaryotic protein kinases (ePKs) play a central role in mediating signal transduction through complex networks and are considered druggable targets from the medical and chemical viewpoints. Our work aimed at analyzing the *S. mansoni* predicted proteome in order to identify and classify all ePKs of this parasite through combined computational approaches. Functional annotation was performed mainly to yield insights into the parasite signaling processes relevant to its complex lifestyle and to select some ePKs as potential drug targets.

Results: We have identified 252 ePKs, which corresponds to 1.9% of the *S. mansoni* predicted proteome, through sequence similarity searches using HMMs (Hidden Markov Models). Amino acid sequences corresponding to the conserved catalytic domain of ePKs were aligned by MAFFT and further used in distance-based phylogenetic analysis as implemented in PHYLIP. Our analysis also included the ePK homologs from six other eukaryotes. The results show that *S. mansoni* has proteins in all ePK groups. Most of them are clearly clustered with known ePKs in other eukaryotes according to the phylogenetic analysis. None of the ePKs are exclusively found in *S. mansoni* or belong to an expanded family in this parasite. Only 16 *S. mansoni* ePKs were experimentally studied, 12 proteins are predicted to be catalytically inactive and approximately 2% of the parasite ePKs remain unclassified. Some proteins were mentioned as good target for drug development since they have a predicted essential function for the parasite.

Conclusions: Our approach has improved the functional annotation of 40% of *S. mansoni* ePKs through combined similarity and phylogenetic-based approaches. As we continue this work, we will highlight the biochemical and physiological adaptations of *S. mansoni* in response to diverse environments during the parasite development, vector interaction, and host infection.

Background

Human schistosomiasis caused by blood fluke parasites of *Schistosoma* genus, remains an important parasitic disease and a major health economic problem in many tropical and subtropical countries. Schistosomes have a complex life cycle that includes six different stages (cercariae, schistosomula, adult worms - male and female, egg, miracidia and sporocyst) in different environments: water, definitive host (mammals) and intermediate host (snail). During parasite development, signals from the environment are sensed and stimulate physiological, morphological and, biochemical adaptations. Oils are shown to stimulate cercarial penetration; hormones and exposure to the snail haemolymph trigger specific physiological adaptations [1-3]. The free living parasite forms display light and geotropism and female development is dependent on signals from the male adult worm through mechanisms not completely understood [4,5]. It has been demonstrated that worm pairing induces changes in gene expression in the female vitelline gland [4] and the accumulation of glutathione and lipids in the male [5]. Furthermore, microarray analysis revealed distinct differential gene expression

---

* Correspondence: oliveira@cpqrr.fiocruz.br
† Contributed equally
[1] Genomics and Computational Biology Group, Instituto Nacional de Ciência e Tecnologia em Doenças Tropicais, Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz - FIOCRUZ, Belo Horizonte, MG- 30190-002, Brazil
Full list of author information is available at the end of the article

BioMed Central
© 2011 Andrade et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

profiles between males and females [6-8]. Therefore, the success of the parasite infection depends on the assessment at the cellular and molecular levels of the environment and the transmission of signals to physiological regulatory networks that will collectively stimulate adaptations.

The maintenance of homeostasis and complex cellular adaptations in *Schistosoma mansoni* require specific extracellular signals that must be integrated to generate an appropriate response from the sensory receptor via intracellular proteins [3]. Signal transduction involves non-linearly integrated networks that interact mostly by switching activity status via phosphorylation (protein kinases) and dephosphorylation (protein phosphatases) of amino acid residues, or the incorporation of GTP. Other cellular non-protein messengers include cyclic AMP, Ca^{2+} and diacylglycerol.

Protein kinases (PKs) play a central role in mediating intracellular signals by adding a phosphate group from ATP or GTP to an amino acid residue leading to a conformational change in the target protein that will switch its activation status [9]. Most PKs have a catalytic domain, which binds and phosphorylates target proteins, and a regulatory region. Many PKs are autophosphorylated or may be phosphorylated by other PKs, an interaction regulated by the accessory protein domains [10].

PKs are classified into two superfamilies containing the eukaryotic or conventional protein kinases (ePK) that share a conserved catalytic domain, and the atypical protein kinases (aPKs) (Table 1). The catalytic domain of ePKs is composed of 250-300 amino acids and is divided into 12 subdomains with highly conserved individual amino acids and motifs [11]. aPKs are reported to have biochemical kinase activity, but lack sequence similarity to the ePK catalytic domain [12]. According to their substrate recognition sites, ePKs are divided broadly into two major classes, serine/threonine kinases (STKs) and tyrosine kinases (TKs). Dual specificity kinases (Hybrid), which phosphorylate serine, threonine, and tyrosine, are also found. ePKs have been further classified into eight groups based on sequence similarity of their catalytic domains, the presence of accessory domains, and their modes of regulation [9,12-14]. According to KinBase [15], a database that holds information of PKs encoded in the human genome and their homologs in other eukaryotes, the eight ePK groups are: AGC (cAMP-dependent protein kinase/protein kinase G/protein kinase C extended), CAMK (Calcium/Calmodulin regulated kinases), CK1 (Cell Kinase I), CMGC (Cyclin-dependent Kinases and other close relatives), RGC (Receptor Guanylate Cyclases), STE (MAP Kinase cascade kinases), TK (Protein Tyrosine Kinase) and TKL (Tyrosine Kinase Like). A ninth group, called ''Other'', consists of a mixed collection of kinases that cannot be classified easily into the previous families [14] (Table 1).

Table 1 Protein kinase classification

Superfamily	Class	Group
Eukaryotic Protein Kinases (ePKs)	Serine/Threonine Kinases	AGC
		CaMK
		CMGC
		STE
		CK1
	Tyrosine Kinases	TK
		RGC
	Hybrid	TK like
		Other

Protein kinases are broadly classified into two major superfamilies (ePKs and aPKs). The ePKs groups, families, and subfamilies adopted in the present work followed the proposed hierarchy described elsewhere [15,16,27].

PKs are considered druggable targets from the medical and chemical viewpoints as a growing number of PKs inhibitors have been developed and approved for treatment of different human disease [16]. An example of a successful PK inhibitor is Gleevac®, that induces a conformational change in PTK and mimics substrate binding and therefore prevents activation by upstream kinases [17]. Beyond this, PKs have gained interest as targets treatment strategies to fight many parasites, including *S. mansoni* [18-21].

The current schistosomiasis treatment frequently does not cure 100% of those treated in high-risk communities and the emergence of *Schistosoma* resistant strains is a real possibility [22-25]. Thus, the identification of potential drug targets should be further emphasized. The recent sequencing of *S. mansoni* genome and large-scale transcriptome projects have yielded crucial information to the identification of new candidate drugs [26-29]. Understanding protein structure and function in many model organisms can help elucidate the function of their parasite homologs and further enable the application of such information in drug design and development. The study of the kinase complement (kinome) is therefore of major importance for the understanding of the physiology of the organism and also provides insights into how to disrupt the fine adaptative mechanisms. The present work aimed at analyzing the *S. mansoni* predicted proteome data in order to identify all ePKs encoded in the genome of this parasite. For this purpose, we combined computational approaches such as sequence similarity searches using Hidden Markov Models (HMMs) and distance-based phylogenetic analyses. The functional annotation was performed mainly to yield insights into the signaling process related to the complex lifestyle of *S. mansoni.*

Results and discussion

The *Schistosoma mansoni* ePKinome

The ePK complement of *S. mansoni*, defined as the ePKinome, was identified by searching the parasite predict proteome with a HMM profile of the ePK catalytic domain of five selected organisms. This analysis revealed 252 ePKs in the *S. mansoni* predicted proteome, representing 1.9% of the total proteins encoded in the parasite genome. Although the total number of protein kinases found across the analyzed species varies greatly (from 82 to 503), the percentage values in respect to the genomes of protozoan and helminth parasites as well as other eukaryotes from KinBase range only between 1.5 to 2% (Figure 1).

Amino acid sequences corresponding to the conserved catalytic domain of ePKs were aligned by MAFFT [30] and further used in phylogenetic analysis based on a distance method as implemented in PHYLIP [31]. The dataset for each ePK group also included the ePK homologs from six other eukaryotes: *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Brugia malayi*. This approach allowed us to classify the *S. mansoni* ePKinome at the group, family, and/or subfamily levels based on the hierarchy proposed elsewhere [12,13,32,32], and sometimes provided insights into kinase function and evolution. Detailed information is available in the Additional file 1 that contains, among other things, all *S. mansoni* ePKs with the corresponding identifier from the genome project linked to SchistoDB database [29]. SchistoDB http://www.schistodb.net allows the community to access to all sequences, annotations and other data types integrated into the genomic information. It also provides several tools to analyze retrieve and display the data. In the SchistoDB it is possible to encounter, for each ePK, the development expression stages by EST evidence, information about orthologs, Gene Onthology (GO) function, metabolic pathways, structural information, PDB structures, and links to external databases such as the TDR database [33]. The TDR database contains additional information for *S. mansoni* genes like antigenicity, essentiality, phenotypes and associated compounds (druggability).

As shown in Figure 2, *S. mansoni* proteins have representatives in the main ePK groups. ePKs that do not fall into these groups are categorized as “Other” in which multiple families have been defined. The *S. mansoni* largest ePK group is CMGC, a feature unique to this parasite, and the smallest group is RGC, a common feature shared with many of the analyzed organisms (Figure 2).

Of the 252 ePKs identified in *S. mansoni* proteome, only 16 were experimentally studied as highlighted in the supplementary material (Additional file 1) and the others 236 ePKs were previously annotated only by automatic methods based on sequence similarity searches [26,29].

S. mansoni ePKs were examined for the presence of the 12 smaller subdomains present in the catalytic domain and also for the presence of a lysine in subdomain II and aspartic acids residues in subdomain VIb and VII, which are known to play essential roles in the kinase function [9,12,34]. According to our analysis, 12 proteins are predicted to be catalytically inactive ePKs, as they lack one or more of the three essential amino acid residues in the catalytic domain (Additional file 1), including all members of *S. mansoni* RGC group (see below).

Approximately 2% of the *S. mansoni* ePK remain unclassified once they do not have similarity to any

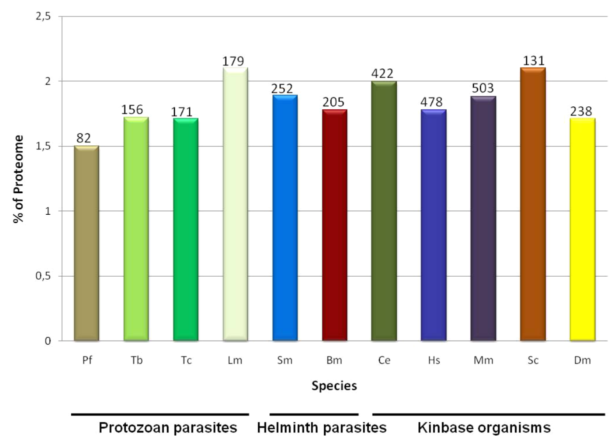


Figure 1 ePKinome in the predicted proteomes of diverse taxa. A total of 252 PKs were identified in the predicted proteome of *S. mansoni*. For comparison, the percentage (%) of the total predicted proteome that codes for kinases and the total number of ePKs (shown on top of each bar) is shown for four protozoan parasites: Pf - *Plasmodium falciparum* [107]; Tc - *Trypanosoma cruzi*, Tb - *Trypanosoma brucei*, Lm - *Leishmania major* [89]; two helminth parasites: Bm - *Brugia malayi* [108] and *S. mansoni* [26]; and five model organisms of KinBase Ce - *Caernorhabditis elegans* [51], Hm - *Homo sapiens* [13], Mm - *Mus musculus* [109], Dm - *Drosophila melanogaster* [110] and Sc - *Saccharomyces cerevisiae* [10].

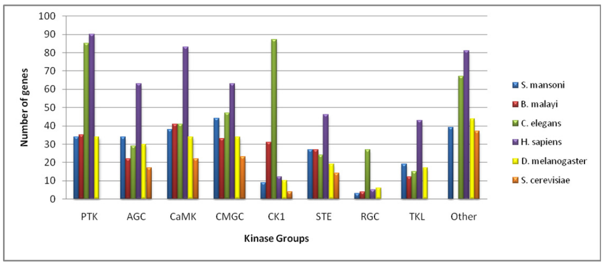


Figure 2 Distribution of ePKs groups in *S. mansoni* and model organisms. *S. mansoni* proteins were classified according to KinBase [15] by combining sequence searches (HMMs) and phylogenetic analysis. For comparison, occurrence of the ePKs in *B. malayi, C. elegans*, *H. sapiens*, *D. melanogaster*, *S. cerevisiae*, and is shown. The ePK groups include: PTK (Protein Tyrosine Kinase), AGC (cAMP-dependent protein kinase/protein kinase G/protein kinase C extended), CaMK (Calcium/Calmodulin regulated kinases), CMGC (Cyclin-dependent Kinases and other close relatives), CK1 (Cell Kinase I), STE (MAP Kinase cascade kinases), RGC (Receptor Guanylate Cyclases), TKL (Tyrosine Kinase Like), and Other.

known PK family. All these proteins have a truncated catalytic domain probably because of an incorrect protein prediction. The unclassified ePKs from *C. elegans, D. melanogaster, H. sapiens* and *S. cerevisiae* range from 19% to about 38% their kinomes.

Serine/Threonine kinases

AGC group

Around 13 families have been classified as part of the AGC group in eukaryotic organisms [11]. In *S. mansoni*, most AGC proteins belong to PKA (Protein Kinase A) (5 proteins), DMPK (Myotonic Dystrophy Protein Kinase) (4 proteins and 1 product of alternative splicing), PKC (Protein Kinase C) (4 proteins) and PKG (4 proteins) families. Other *S. mansoni* proteins have only one representative in the remaining AGC families (Additional file 1). According to our phylogenetic analysis, *S. mansoni* has no homolog of the YANK (Yet Another Novel Kinase) family (Additional file 2).

The most significant difference between PKA and PKG family members is that in PKA, the regulatory and catalytic activities are performed by separate gene products known as PKA-R and PKA-C, respectively, whereas in PKG the cNMP-binding (cyclic nucleotide-binding domain) and catalytic domains are usually present in the same polypeptide [35]. The inactive conformation of PKA is a heterotetramer of two PKA-R and two PKA-C subunits, while PKG exists as a homodimer [35]. *S. mansoni* processes five homologs of the PKA-C subunit (Additional file 1), and six predicted of PKC-R subunit (Smp_131050, Smp_147320, Smp_079010, Smp_030400, Smp_019280, Smp_022100) allowing for a variety of different holoenzymes to be formed in this parasite. Some studies demonstrated that PKG proteins of *Toxoplasma* [36] and *Eimeria* [37] and PKG and PKA proteins of *Plasmodium* [38,39] are essential as the inhibitors causes an anti-parasite effect in these organisms. Recently it was shown that inhibition of the SmPKA-C subunit (Smp_152330), expressed in adult worms of *S. mansoni*, resulted in the death of the parasites [40]. This result and the range of holoenzymes that can be formed, indicate that genes in this family are critical for the development of *S. mansoni* and may represent good targets for drug development.

PKC belongs to a large protein family that is classified into four important subfamilies: PKC Alpha subfamily, that contain the conventional PKCs (γ, βI, βII, and α) and are sensitive to diacylglycerol (DAG) and Ca^{2+}; PKC Eta and Delta subfamilies containing the *novel* PKCs (ε, δ, η, and θ) which are regulated by DAG alone; and PKC Iota subfamily, that contain the atypical PKCs (ζ and ι), and are insensitive to both compounds [41,42]. PKC is considered to be a mechanistic regulator of development in vertebrates, playing a key role in cell growth and differentiation [43-45]. *S. mansoni* has representatives in the three main PKC subfamilies mentioned above (Iota, Eta and Alpha) but lacks homologs in the Delta subfamily, present in *C. elegans*, *D. melanogaster, M. musculus*, and *H. sapiens*. The two PKC Alpha proteins found in *S. mansoni* (Smp_128480 and Smp_176360), belong to the PKCβI isoform and were recently characterized [46-48]. Both are associated with the neural mass, excretory vesicle, ridge cyton, tegument and germinal cells in schistosomula and miracidium, suggesting a possible role in larval transformation [46-48].

One protein in AGC group, Smp_157370, remains unclassified. In the phylogenetic tree, this protein appears more closely related to the GRK (G-protein coupled Receptor Kinase) family (Additional file 2), despite the good conservation of the catalytic domain, this protein lacks the accessory domain that is characteristic of the GRK proteins (Additiona file 2). Furthermore, Smp_157370 does not form a clade with the GRK family members according to our phylogenetic tree, which corroborates its divergence in relation to GRK homologs in other eukaryotes (Additional file 2).

Interestingly, according to SchistoDB [29] EST evidences, the two most highly transcribed ePKs (Smp_151140 and Smp_158560.1) in *S. mansoni*, belong to the DMPK family of the AGC group, mainly in cercariae, schistosomula, eggs and adult worms. This finding is interesting as these are the four life cycle stages of the parasite which are in contact with the definitive host. In *C. elegans p*roteins of DMPK family are expressed in hypodermal cells and are involved in embryonic elongation [49].

CaMK group

The divalent cation calcium (Ca^{2+}) is one of the ions most widely used as a second messenger in cellular signaling. A significant portion of calcium-mediated signaling is controlled by calmodulin-binding kinases. Some members of the CaMK group are dependent on the binding of Ca^{2+}/CaM [50]. In the *S. mansoni* ePKinome, 32 proteins were classified as CaMK with the vast majority (18 proteins) belonging to the CaMKL (Calcium/Calmodulin Regulated Kinase) - like family. A similar number was found in other organisms analyzed here (Additional file 3). *S. mansoni* also contain members of DAPK (death associated protein kinase), MAPKAPK (MAPK associated protein kinase), MLCK (myosin light chain kinase), and PHK (phosphorylase kinase) families in the CaMK group (Additional file 4).

MLCK is a Ca^{2+}/calmodulin-dependent protein kinase whose only known substrate is myosin II regulatory light chain [51]. The primary function of MLCK is to stimulate muscle contraction through the phosphorylation of the myosin II regulatory light chain (RLC), a eukaryotic motor protein that interacts with filamentous actin. Although MLCK has only one known substrate (RLC), this protein is linked to a variety of cellular processes due to the diverse biological function of myosin II [50]. Two distinct smooth muscle MLCK genes were identified in *S. mansoni* (Smp_121780, Smp_126240), although no homologs were identified for the non-smooth muscle vertebrate MLCK through our phylogenetic analysis. This likely reflects the absence of a striated muscle in this parasite.

DCAMKL (Doublecortin and CaMK-like) is a protein that regulates the microtubule cytoskeleton and in the chick is specifically expressed in the developing brain [52,53]. CASK is a protein that participates in cell adhesion [54]. According to our phylogenetic analysis, a single homolog of the DCAMKL (Smp_053560) and CASK (Smp_131690) families were found in *S. mansoni* (Additional file 4).

While the CaMK2 (CaMK family 2) family is encoded by four genes in humans, only a single CaMK2 gene, with two predicted alternative spliced transcripts, was identified in the *S. mansoni* genome (Additional file 1). *S. mansoni* CaMK2 was recently identified as putative target for drug development after comparative chemogenomics approach using the *S. mansoni* proteome and the proteome of two model organisms, *C. elegans* and *D. melanogaster*. [55]. The function of this protein in *S. mansoni* is still unknown. In sea urchin, CaMK2 is required for nuclear envelope breakdown following fertilization [56].

CMGC group

CMGC kinases are relatively abundant in *S. mansoni*, a feature that can be explained by the requirement to control cell proliferation and to ensure correct replication and segregation of organelles, which together are essential mechanisms for parasites with a complex life cycle. In the CMGC group, all of the main families are conserved between *S. cerevisiae*, *C. elegans*, *M. musculus*, *H. sapiens*, and *S. mansoni*, including CDK (Cyclin Dependent Kinase), MAPK (Mitogen Activated Protein Kinase), GSK (Gycogen Synthase 3 Kinase), CLK (CDC-Like Kinase), SRPK (SR Protein Kinase), CK2 (Cell Kinase 2), and DYRK (Dual-specificity Tyrosine Regulated Kinase) (Additional file 5) and RCK.

S. mansoni has 14 CDKs, the same number was found in *C. elegans* (compared with only seven in *S. cerevisiae*), including homologs of all subfamilies (CDK7, CDK4, CDK8, CRK7, CDK9, PITSLRE, CDK10, PCTAIRE, PFTAIRE, VDK5 and CDC2) (Additional file 5). On the other hand, only one RCK family protein (Smp_132890) was identified in the parasite. The RCK proteins are similar to mammalian MAK (male germ cell-associated kinase), which have been implicated in spermatogenic meiosis and in signal transduction pathways for sight and smell [51].

GSK family is represented by 3 proteins in *S. mansoni*. One of those (Smp_008260.1) was selected as putative target for drug development after comparative chemogenomics approach [55]. GSK proteins are involved in development and cell proliferation, are overexpressed in colon carcinomas and positively regulates the Wnt signaling pathway during embryonic development and oocyte-to-embryo transition in *C. elegans* [57].

The MAPK signaling pathways are some of the best characterized signaling systems. *S. mansoni* contains nine MAPKs, compared to seven in *D. melanogaster* and 14 in *C. elegans*. As shown in Figure 3, mammals have, at least five MAPK cascades described; these include the extracellular signal-regulated kinase (ERK) cascade, which regulates cell growth and differentiation, the c-Jun N-terminal kinase (JNK)/stress-activated protein kinase (SAPK), and the p38 MAPK cascades, which function mainly in stress responses such as inflammation and apoptosis [58]. In *D. melanogaster* and *C. elegans*, the MAPK pathways are involved in critical cellular and developmental processes [59,60]. *S. cerevisiae* has four distinct MAPK signaling pathways that are likely mediators of responses to pheromone, nutritional starvation, and cellular or osmotic stress [51]. The MAPK signaling pathways are well conserved in *S. mansoni* (Figure 3), including representatives of the subfamilies ERK, p38, JNK, and, NLK but lacks members of ERK5 that are part of a signaling pathways found mainly in mammals (Additional file 5). Each subfamily is activated by different stimuli that generate different biological responses [61-63]. In *S. mansoni* only one protein was identified in JNK (Smp_172240) subfamily. JNK proteins play key roles in human cell function [62] and in the development of *C. elegans* worms [64]. JNK may have an important role in schistosome survival and represent a good target for experimental approaches.

STE group

In *S. mansoni*, the STE group includes seven STE7 (MEK or MAPKK), two STE11 (MEKK or MAPKKK), and 13 STE20 (MEKKK) kinases (Additional file 6). The large number of STE family members in *S. mansoni* could translate into an enormous potential for downstream signal specificity and diversity. SmSLK (Smp_150260) is a Ste20 family protein, recently characterized in *S. mansoni*, which is able to activate protein MAPK/JNK in human embryonic kidney (HEK) cells as well as in *Xenopus* oocytes. In addition, imunofluorescence showed that SmSLK was abundant in the tegument of adult schistosomes [65]. These findings indicate that signals sensed in the environment by many different proteins may activate the MAPK cascade that will generate an adaptive physiological response. Futhermore, molecules that activate the MAPK pathways, as some hormone and cytokine signals, are not found in the *S. mansoni* predicted proteome (Figure 3). It has been demonstrated that the parasite takes advantage of

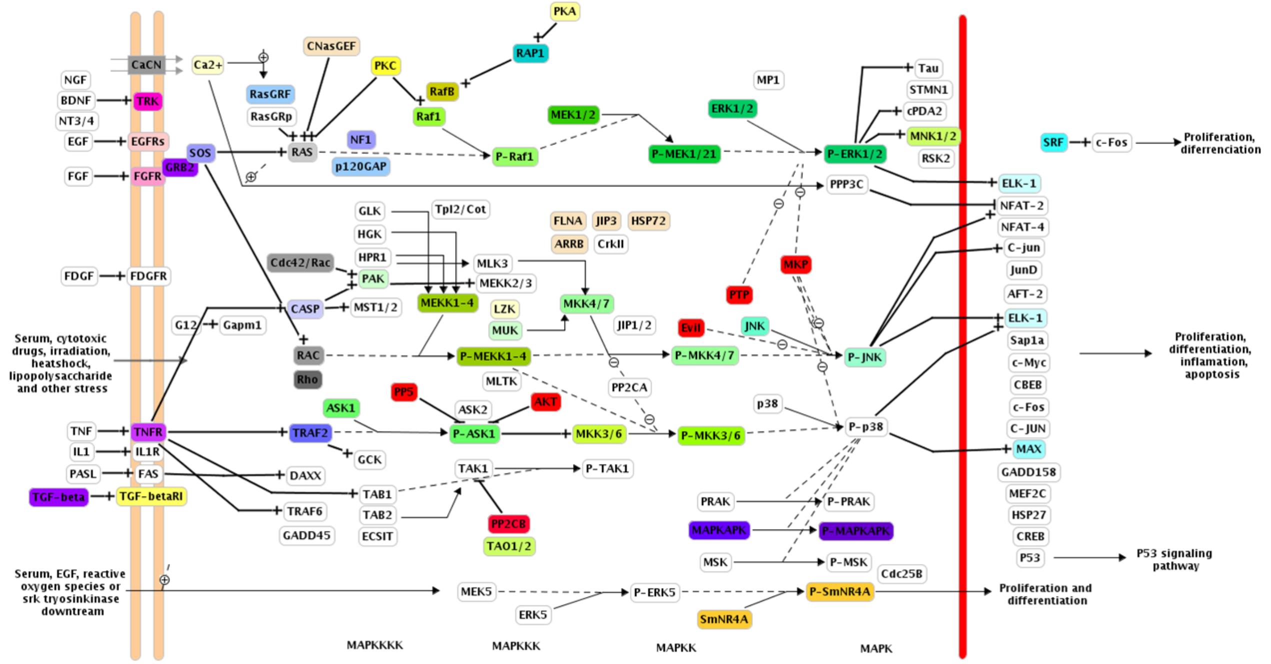


Figure 3 MAPK signaling pathway. MAPKs are expressed in all eukaryotic cells and are activated by diverse *stimuli* ranging from cytokines, growth factors, neurotransmitters, hormones, cellular stress, and cell adherence. The basic assembly of the MAPK pathway is a three-component module that includes three kinases that establish a sequential activation pathway comprising a MAPK kinase kinase (MKKK, MEKK, STE11, or STE13), MAPK kinase (MKK, MEK or STE7), and MAPK (green). The activated MAPK may translocate to the nucleus and bind to transcription factors (light blue). The mammalian MAPK can be subdivided into five families: ERK, p38, JNK, nml, and ERK5. Each MAPK family has distinct biological functions. Colored blocks correspond to proteins indentified in the *S. mansoni* predicted proteome and white blocks are mammals' proteins with no homologs in *S. mansoni* predicted proteome. The + signal represents protein activation and - signal protein inhibition by protein phosphatases (red colored proteins). *S. mansoni* has all representatives of ERK, p38, and JNK pathways including proteins of STE (MEK1/2, MKK), AGC (PKC and PKA), CaMK (MAPKAPK), TK (EGFR, FGFR), and TKL (Raf and TGFbeta-RI) groups.

host proteins for its growth and development [66,67]. Other ePKs such as members of the PKA, PKC, Raf and receptor protein tyrosine kinases (RTKs) families, also participate in MAPK signaling pathway. RTKs are anchored to the membrane and have an important role in transmitting the signal from the extracellular to cytoplasm (Figure 3) [64].

In *C. elegans* genome studies such as classical forward genetic and RNA interference screens and systematic targeted gene knockout revealed genes that are essential to the organism [68]. Although the off-target and non-specific effect of RNAi [69,70], in *S. mansoni* this is one of the best approaches to explore the functional property of the genes since the knockout experiments are not yet available for schistosomes [71]. By analyzing the phylogenetic trees of the present work, it was possible to identify the proteins of *S. mansoni* that have homologs in *C. elegans* and display lethality and sterile phenotypes by RNAi (Table 2). Interestingly, most essential proteins in table 2 belong to the CMGC and STE groups, suggesting the involvement of these proteins in signaling pathways that culminate in essential cellular processes.

CK1 Group

The two smallest groups found in the *S. mansoni* ePKinome were CK1 and RGC (Figure 2). In contrast, in *C. elegans* CK1 is the largest group and RGC is dramatically expanded. However, these expansions are a unique feature of *C. elegans*, as compared to other eukaryotes selected for this analysis (Figure 4). The CK1 group consists of three main ePK families: CK1, VRK (Vaccinia Related Kinase), and TTBK (Tau Tubulin Kinase) that formed three individual clusters in the phylogenic tree (Figure 4). *S. mansoni* has representatives in each of these families also found in *C. elegans*, *D. melanogaster*, *M. musculus*, *H. sapiens*, *S. cerevisiae* and *B. malayi* kinomes. The nematodes, *C. elegans* and *B. malayi*, still have two other families that seem to be specific to this taxonomic group, TTBKL and Worm6. The Worm8 family was identified only in *Caenorhabditis* so far. The diversification of the CK1 group in *C. elegans* may be an adaptation allowing for enhanced DNA repair in response to excessive exposure to environmental mutagens [51]. One CK1 encoding gene (*spe-6*) functions in spermatogenesis, and at least half of the proteins in this group are selectively expressed in *C. elegans* sperm as

Table 2 Orthology relationships among ePKs of *S. mansoni*, *B. malayi*, and *C. elegans* and RNAi phenotype for *C. elegans* proteins

Group	*S. mansoni*	*C. elegans*	*Brugi malayi*	*C. elegans RNAi phenotype*
CaMK				
	Smp_053560	C44C8.6_Ce e K08F8.1	-	Embryonic Lethal
AGC				
	Smp_123640	F19C6.1_Ce	-	Embryonic lethal, maternal sterile, organism morphology variant.
	Smp_136750	Y47D3A.16_Ce	Bm1_44635	Larval lethal, fertility reduced
	Smp_158560	K08B12.5_Ce	Bm1_02320	Body morphology defect, slow growth
	Smp_096310	pkc-3_Ce_A	Bm1_03335	Embryonic lethal, sterile, body morphology defect
TK				
	Smp_139480	vab-1_Ce_T	Bm1_03410	Embryonic lethal, body morphology defect
STE				
	Smp_026510, Smp_151670	mek-2_Ce	-	Embryonic lethal, sterile, reduced brood size, exploted through vulva, slow growth
	Smp_096640	gck_1_Ce	-	cortical dynamics defective early embryonic, maternal sterile
	Smp_163420	mig_15_Ce	Bm1_32540	Abnormal cell migration, protruding vulva, developmental delay
	Smp_068060	kin_18_Ce	Bm1_55590	Embryonic lethal, body morphology defect, slow growth, sterile, larval lethal, dumpy
	Smp_131800	Y59A8B.23_CE	Bm1_11845	More depolarized oocytes
	Smp_150260	C04A11.3_Ce	Bm1_26435	Slow growth, larval lethal late L3/L4
TKL				
	Smp_079760	pat4_Ce_TK	Bm1_20815	Embryonic lethal, body morphology defect, larval lethal, sterile, paralyzed, uncoordinated movement, paralyzed
	Smp_176990	lin45_Ce_T	Bm1_40290	Sterile, osmotic integrity problems
CMGC				
	Smp_068960	F22D6.5_Ce	Bm1_12230	Embryonic lethal, slow growth, sterile progeny, uncoordinateed movement
	Smp_041770	spk_1_Ce	-	Embryonic lethal, larval lethal, maternal sterile
	Smp_172700	cdk_4_Ce	-	Locomotion variant, larval lethal, sterile progeny
	Smp_133020	pmk-1_Ce, pmk-2_Ce	-	Apoptosis increased, sterile progeny, reduced brood size, embryonic lethal
	Smp_134260, Smp_133490, Smp_133500	C05D10.2	-	Reduced brood size, embryonic lethal
	Smp_080730	cdk_1_Ce	-	embryonic lethal, sterile progeny
	Smp_155720, Smp_125310, Smp_008260.1	Y18D10A.5_Ce, C44H4.6_Ce	-	Larval lethal, slow growth, embryonic lethal
	Smp_156990	F39H11.3_Ce	Bm1_33825	Sterile
	Smp_176620	B0495.2_Ce, ZC504.3_Ce	Bm1_17240	Embryonic lethal
	Smp_003000	H25P06.2_Ce	-	Embryonic lethal, slow growth, uncoordinated movement
	Smp_150040	Cdk_7_Ce	Bm1_14135	Embryonic lethal
	Smp_140700, Smp_074080	lit_1_Ce	Bm1_130	Embryonic lethal, larval lethal
	Smp_141230	mbk_2_Ce	Bm1_38345	Embryonic lethal, sterile progeny

Orthology relationships were inferred based on phylogenetic trees. RNAi phenotype for *C. elegans* were visualized by recent work [68] that relied on identification and analysis of essential genes.

shown by microarray analysis [72,73]. The role of these proteins in the parasite *S. mansoni* is unclear.

Tyrosine kinases

TK group

PTKs can be classified, based on the presence or absence of transmembrane domains, into receptor tyrosine kinase (RTK) that relay intracellular signals [74], and cytoplasmatic tyrosine kinase (CTK). *S. mansoni* kinome contains 15 RTKs and 19 CTKs. The 15 RTK include two InsRs (Insulin Receptors), four EGFRs (Epidermal Growth Factor Receptor), two VKRs (Venus Flytrap Kinase Receptors), a representative for Ephs (Ephrin receptors), Ror, CCK4 (Colon Carcinoma Kinase 4), and MUSK (Muscle_Specific kinase) families, besides three unknown receptors.

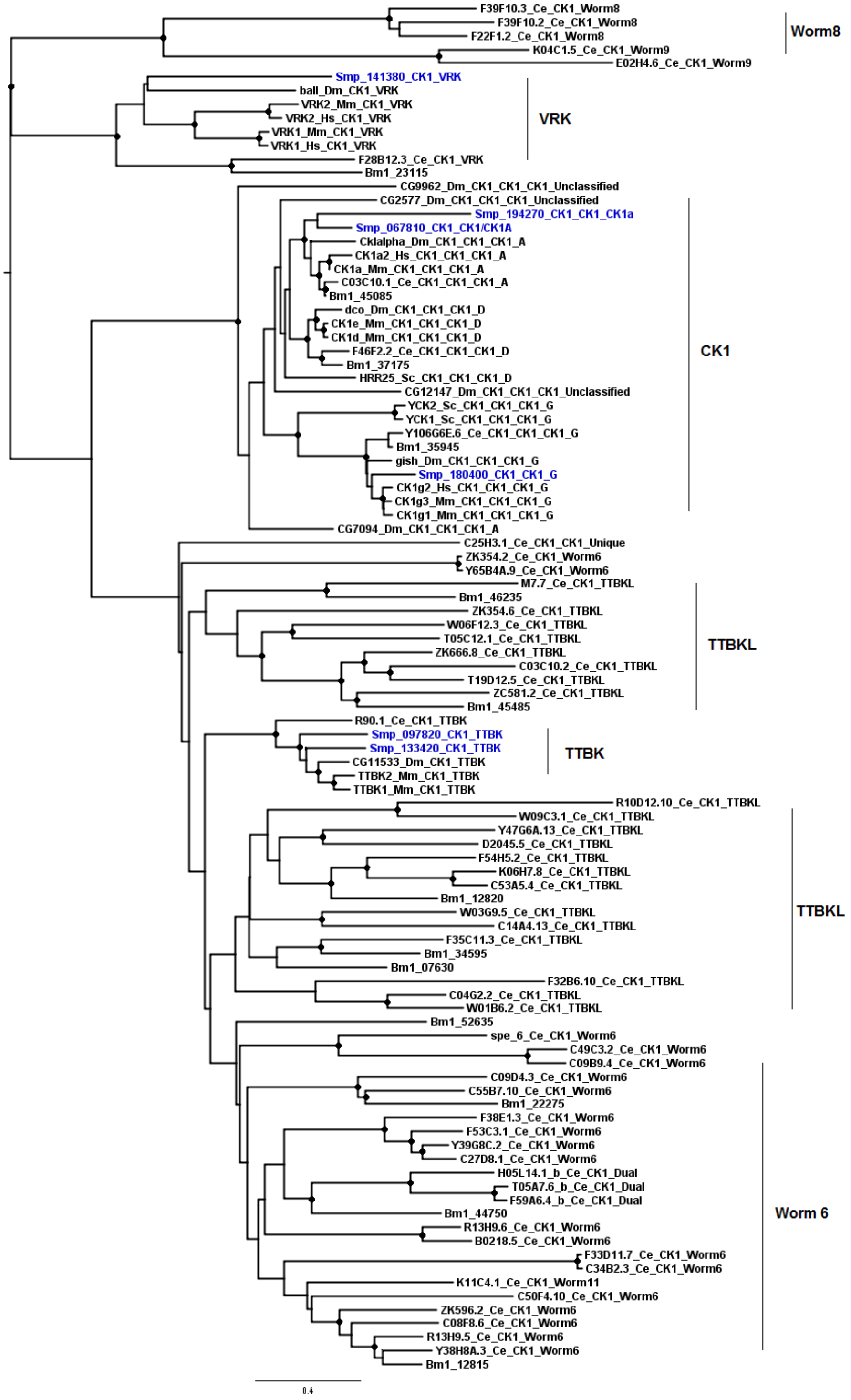


Figure 4 Phylogenetic analysis of the CK1 group. Amino acid sequences of the catalytic domain of the CK1 proteins of *S. mansoni* (blue labels), *C. elegans*, *D. melanogaster*, *S. cerevisiae*, *H. sapiens*, and *B. malayi* were used to construct a distance-based phylogenetic tree using PHYLIP programs. Bootstrap values (1000 replicates) equal or higher than 80% are indicated (●) as cutoff values to support the family/subfamily classification. Functional classification is indicated individually in the protein labels. Protein families are highlighted by the vertical bars and include: VRK (Vaccinia Related Kinase), CK1 (Casein Kinase 1), TTBK (Tau Tubulin Kinase), Worm 8 (a specific family of *Caenorhabditis*), TTBKL (TTBK-Like kinase), and worm 6, two specific families of *C. elegans* and *B. malayi*.

Two InsRs in *S. mansoni*, SmIR-1 (Smp_009990) and SmIR-2 (Smp_074030) present distinct functions during parasite development. These two receptors are well clustered within the InsR families but showed to be more divergent than the mammalian and *D. melanogaster* proteins (Additional file 7). SmIR-1 was localized in the muscles, intestinal epithelium, and basal membrane of adult male and female worms and at the periphery of schistosomula, mainly in the tegument [75]. SmIR-1 co-localized in schistosome tegument with glucose transporters suggesting a role in the regulation of glucose uptake which is an essential nutrient for the intra-mammalian stages of *S. mansoni*. SmIR-2, in contrast, was distributed in the parenchyma of adult males and females indicating a possible involvement of the receptor in parasite growth. *S. mansoni* is the first invertebrate with two insulin receptors characterized that seem to have distinct functions, as in vertebrates [18,75,76]. Mammals have two InsR members; insulin-like growth factor receptor (IGFR), which has a role in controlling growth, and (InR) which has specialized in metabolic regulation [77].

In *C. elegans* EGFR signaling induces behavioral quiescence [78]. One *S. mansoni* EGFR homolog (Smp_173590) was localized in the parasite muscle and perhaps related to muscle development or function [79]. Vertebrate EGF activates *S. mansoni* EGFR and the downstream classical ERK pathway (Figure 3), indicating the conservation of EGFR function in *S. mansoni* [79]. Moreover, human EGF was shown to increase protein and DNA synthesis as well as protein phosphorylation in parasites, supporting the hypothesis that host EGF could regulate schistosome development [80]. The similarity of schistosome proteins to sex hormone receptors of mammalian hosts provides a good example of host parasite relationship, where the adult worm depends on the host hormone synthesis for their maturation and reproduction [3].

Five *S. mansoni* proteins are not clustered with the main RTK families as shown in our phylogenetic analyses (Additional file 7). Three of them have a truncated catalytic domain (Smp_175590, Smp_093500 and Smp_157300) and two are specific RTK with a venus flytrap domain (VKR family). VKR is a family of receptors found in invertebrates, especially in insects. One *S. mansoni* VKR protein, Smp_153500 (SmVKR), was recently studied [81]. We identified another protein (Smp_019790) clustering with SmVKR (Additional file 7) with a high similarity. Despite the similarity of the catalytic domain of VKR protein with the IRs, these two proteins are not clustered with InsR family. In this respect, the most interesting finding is that VKR family members are not found in mammals and could represent good targets for drug development as a specific inhibitor for this family will probably not affect any protein of the host [81].

The CTKs in *S. mansoni* are represented by 11 different families (Additional file 7). SmTK3 (Smp_054500) and SmTK5 (Smp_136300) - src family members, and SmTK4 (Smp_149460) - syk family, are present in reproductive organs and possibly involved in the development of gonads and multiplication of germinal and vitelline cells [82-84]. Abl proteins of *S. mansoni* (Smp_128790 and Smp_169230) were recently studied using a Abl specific inhibitor (Imatinib, Gleevac$^{®}$). The results showed an important morphological alteration in adult worms of *S. mansoni* that led to the death of the parasites [21]. *C. elegans* contains 42 members of the Fer family, while only a single member, SmFes, was found in *S. mansoni*. The Fer gene of *S. mansoni* (SmFes, Smp_164810) exhibits the characteristic features of Fes/Fps/Fer (fes, feline sarcoma; fps, Fujinami poultry sarcoma; fer, fes related) PTKs. By immunolocalization assays it was shown that SmFes is particularly expressed at the terebratorium of miracidia and tegument of cercaria and schistosomula skin-stage. These findings suggest that SmFes may play a role in signal transduction pathways involved in larval transformation after penetration into intermediate and definitive hosts [85,86].

RGC group

Proteins in this group share sequence similarity to the catalytic domain (Pfam: PF07714) found in proteins of the TK group [87]. The RGC group is underrepresented in most species, except in *C. elegans* that has a large expansion of these proteins and *S. cerevisiae* that has no protein with similarity to the TK catalytic domain (Figure 2). Only three RGC members were identified in the *S. mansoni* ePKinome. All of them are more closely related to the mammalian and insect families than the worm family. *C. elegans* and *B. malayi* RGC proteins form at least two different families noticeably more divergent from *S. mansoni*, *D. melanogaster*, *M. musculus*, and *H. sapiens* families as suggested by our phylogenetic analysis (Additional file 8). Most RGC proteins remain functionally uncharacterized. In *C. elegans*, several RGC proteins are highly expressed in restricted sets of neurons and are implicated in chemosensation. One RGC is involved in dauer stage formation [88]. Other parasites such as *L. major, T. brucei, T. cruzi* and *P. falciparum* also lack homologs in the RGC group [20,89]. The three *S. mansoni* RGC proteins have an amino acid substitution in the aspartic acid in subdomain VIb of the catalytic domain, rendering them catalytically inactive. Although the catalytic center of an enzyme is usually highly conserved, there have been reports of proteins, like those of the RGC group of ePKs, with substitutions at essential catalytic positions, which convert the enzyme into a catalytically inactive form. A recent study showed that inactive enzymes are found in a large variety of families conserved among metazoan species and they have lost their catalytic activity, have adopted new functions, and are involved in regulatory processes [34,90].

Hybrid protein kinase

TKL Group

TKL consists of a divergent group that is phylogenetically close to the tyrosine kinases (Figure 3). However, TKL proteins have an unusual catalytic domain that is a hybrid between the serine/threonine and tyrosine kinases [32]. The catalytic domain may display greater similarity to the tyrosine catalytic domain (Pfam: PF07714) or to the serine/threonine catalytic domains (Pfam: PF00069) [10,11]. In *S. mansoni*, the TKL group includes MLK (Miked Lineage Kinases), LISK (Family containing closely related LIMK and TESK sub-families), Raf, RIPK (Receptor Interacting Protein Kinase), STKR (Serine/threonine kinase Receptors for activin and TGFβ ligands), and LRRK (Leucine Rich Repeat Kinase) (type 1 and type 2) families. Of the 19 TKL proteins found in *S. mansoni*, 15 display greater similarity to the serine/threonine catalytic domain and four (of Raf and MKL/ILK families) to the tyrosine catalytic domain. *S. mansoni* has no homologous proteins of the IRAK (interleukin-1 (IL-1) receptor-associated kinase) family that is present in *C. elegans*, *B. malayi, D. melanogaster, Homo sapiens*, and *M. musculus* (Additional file 9). Although *S. cerevisiae* does not have any TKL protein homologue, other fungal species do contain such proteins [14]. Raf (also known as MAPKKK) is a TKL family that plays an important role in the activation of STE proteins in the signaling cascade that culminates in the activation of ERK1/2 (Figure 4) [63]. A recent study showed that blocking the expression of the homolog of the *S. mansoni* Raf protein (Smp_176990) in *C. elegans* by RNAi, generate a sterile phenotype, which supports the hypothesis of the involvement of Raf protein in the germline development, somatic gonad development, oogenesis, spermatogenesis, ovulation or fertilization (Table 2). Raf protein may represents a good target for drug development in *S. mansoni*.

A STKR member that binds to TGFβ (Transforming growth factor) is a membrane receptor that can be divided into two subclasses (Type I and Type II). The type II receptor binds TGFβ and then recruits the type I receptor. The TGFβ type I receptor was cloned in *S. mansoni* (SmTβRI - Smp_173400.2) and it was found to be localized in the parasite surface [91,92]. Other type I STRK (Smp_049760) was identified in the *S. mansoni* predicted proteome and was not experimentally characterized so far (Additional file 9). Three type II STKRs (TGFβ type II receptors) are proteins identified in the same contig which were predicted to be a product of alternative splicing. A recent study revealed the presence of two transcripts that are translated into two different isoforms of type II receptor [93]. These transcripts are produced from the same gene by alternative splicing of the last two exons. The authors indicated that these different type II receptors might signal in different cells or development stages. Furthermore, that study showed that in the presence of human TGFβ, SmTβRII (Smp_165310) activated SmTβRI. The results also provide evidence for the role for the TGF-β signaling pathway in male-induced female reproductive development [94,95].

Other Group

The Other group consists of a mixed collection of kinases with representatives in higher eukaryotes, including SCY1, NEK (Mitotic Kinase family, also known as NRK), PEK, Haspin, WEE, NAK (Numb-Associated Kinase), ULK (Unc-51 Like Kinase), IRE (Inositol Requiring), PLK (Polo Like Kinases), AUR (Aurora Kinase), and CDC7 (Cell Division Control 7) families (Additional file 1). Our analysis showed that 15% of the *S. mansoni* ePKinome do not fall into any of the eight major groups, but include 20 smaller and conserved families.

Accessory Domains

The structure of the catalytic domain of many ePKs is highly conserved across distinct organisms because of the fact that all ePKs recognize and bind ATP at common sites. However, only the catalytic domain (architecture/sequence) is sufficiently divergent to enable the discrimination of groups, families, and subfamilies (Figure 5).

Most ePKs also have a second domain that is involved in protein-protein interaction and allosteric regulation of the catalytic domain [17]. In this work, only the catalytic domain sequence was used in the phylogenetic analyses. Interestingly, when the information on the ePK accessory domains was integrated into the phylogenies, we observed a correlation between diversity of protein architecture and the phylogenetic patterning. We also believe that the diversification of the ePKs happened a long time ago.

The analysis of the sequence domain data from Pfam [96] showed that approximately 30% of *S. mansoni* ePKs are multi-domain proteins containing various regulatory and signaling domains tethered to catalytic kinase domains (Figures 6 and 7). It is known that the distinct protein architectures reflect functional differences among proteins [32]. Hence, understanding the mechanisms that generate such diverse repertoire of protein architectures is essential to the comprehension of the biological function of the ePKs. Furthermore, we observed in ePKs of *S. mansoni* some unusual architecture that probably occurs by domain fusion and recruitment (see some examples

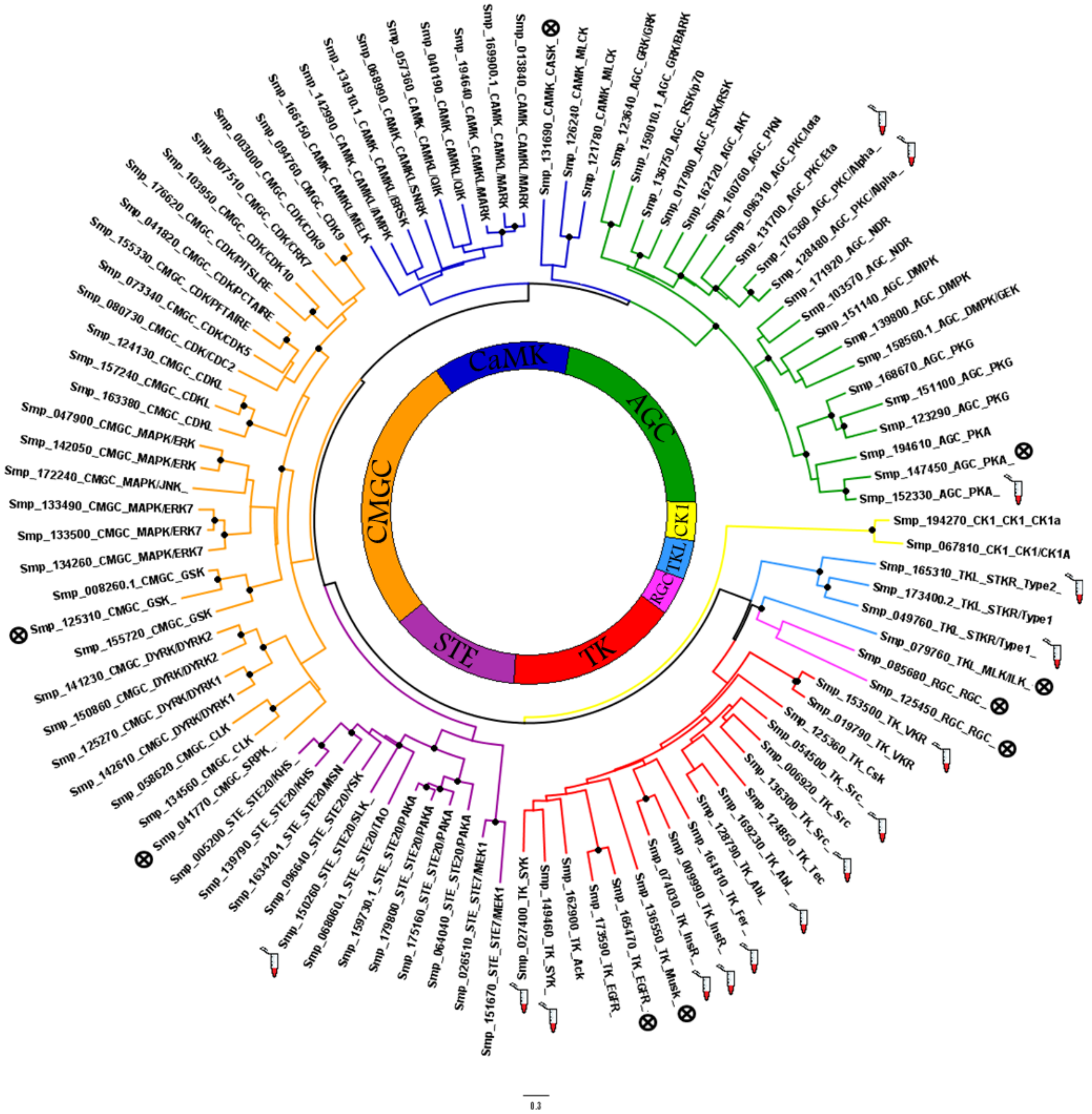


Figure 5 Phylogenetic analysis of the paralogous ePKs groups of *S. Mansoni*. The catalytic domain of *S. mansoni* ePKs was used to construct a distance-based phylogenetic tree using PHYLIP programs. Some ePK were excluded from the tree after filtered the alignments to keep proteins with 30% to 98% pairwise sequence identity. The major ePK groups are color coded and include: CaMK (dark blue), CMGC (orange), TK (red), AGC (green), STE (pink), TKL (light blue), CK1 (yellow), and RGC (light pink). Functional classification is indicated individually in the protein. Proteins with experimental evidence (🧪) and those predicted to be inactive (⊗) due one or more substitutions in important residues in the catalytic domain are indicated. Bootstrap values (100 replicates) equal or higher than 80% are indicated (●).

below), generating specificity towards cognate substrates and regulators in this parasite.

The most common Pfam accessory domains found in *S. mansoni* kinases are Pkinase_C (Pfam: PF00433) all found in the AGC group; C1_1 (Pfam: PF00130) found in the AGC and TKL groups; SH2 (PF00017) all found in the TK group; and SH3 (Pfam: PF00018) found in TK and TKL groups. These domains are commonly found in protein kinase families as we observed in other species from KinBase [15].

More than 40% of *S. mansoni* AGC group have the PKinase_C domain associated with the catalytic domain. The

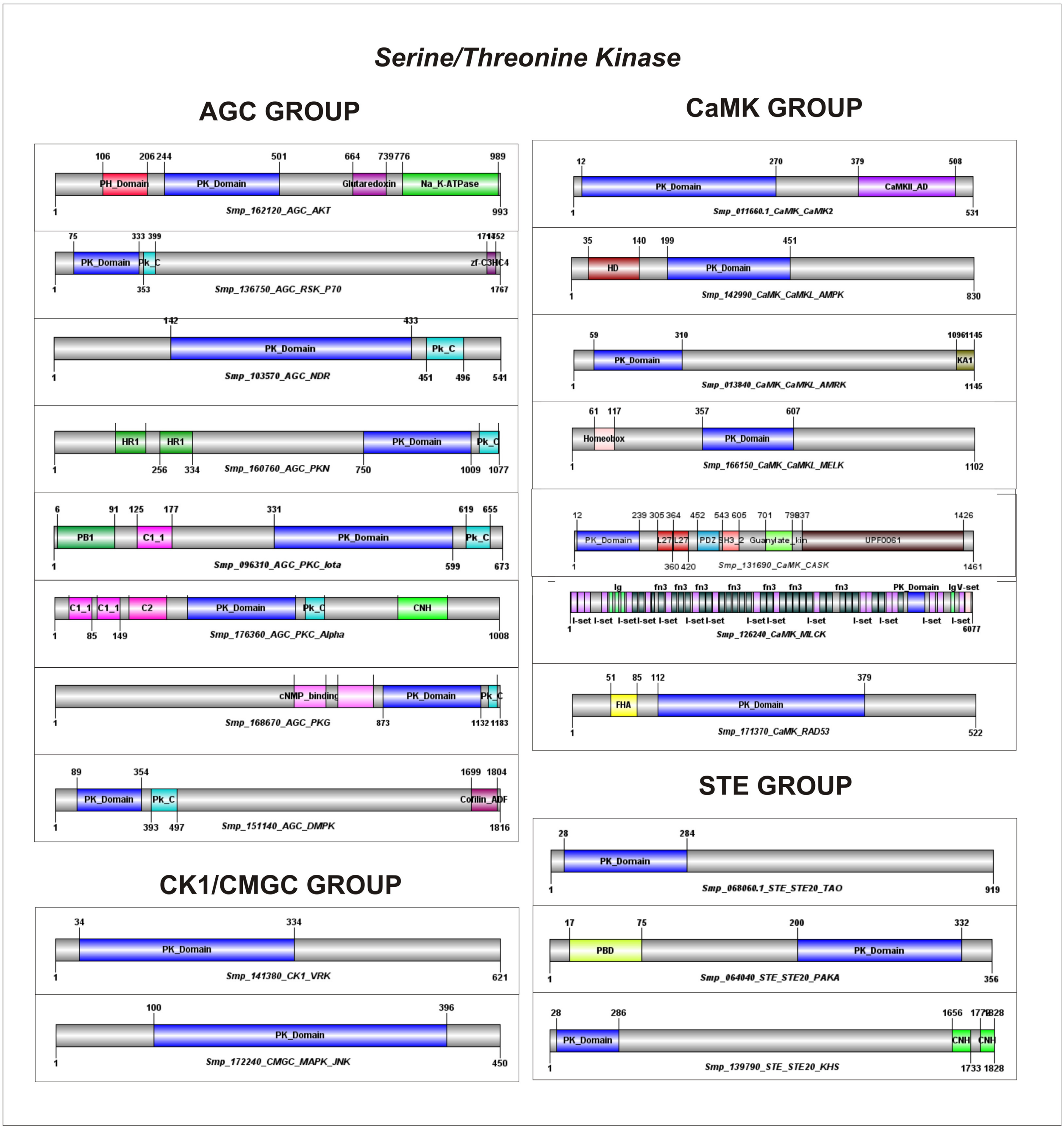


Figure 6 ***S. mansoni*** **ePKs domain architectures**. Representative domain organizations of some *S. mansoni* ePKs belonging to the AGC, CaMK, CK1, CMGC, and STE groups are shown. Each protein ID and classification is shown below each image. Abbreviations followed are: PK_Domain (Protein kinase domain), PH_Domain (Pleckstrin Homology domain), Na_K_ATPase (Sodium/potassium ATPase beta chain), PK_C- (Protein Kinase C terminal domain), Zf-C3HC4 (zinc-finger, C3HC4 type RING finger), HR1 (Hr1 repeat), PB1 (Phox and Bem1p domain), C1_1 (Phorbol esters/diacylglycerol binding domain), C2 (Ca^{2+}-dependent domain), CNH (citron homology domain), cNMP_binding (cyclic nucleotide-binding domain), Cofilin_ADF (Cofilin/tropomyosin-type actin-binding protein), CaMKII_AD (Calcium/calmodulin dependent protein kinase II Association), HD (HD homeobox domain), L27 (L27 domain), PDZ (PDZ domain, also known as DHR or GLGF), SH3_2 (Variant SH3 domain), Guanylate_kin (Guanylate kinase), UPF0061 (Uncharacterized ACR, YdiU/UPF0061 family), Ig (Immunoglobulin domain), fn3 (Fibronectin type III domain), V-set (Immunoglobulin V-set domain), FHA (Forkhead-associated domain), and PBD (P21-Rho-binding domain). The protein domain architectures were generated using DOG 1.0 [111] based on the Pfam domain limits [96].

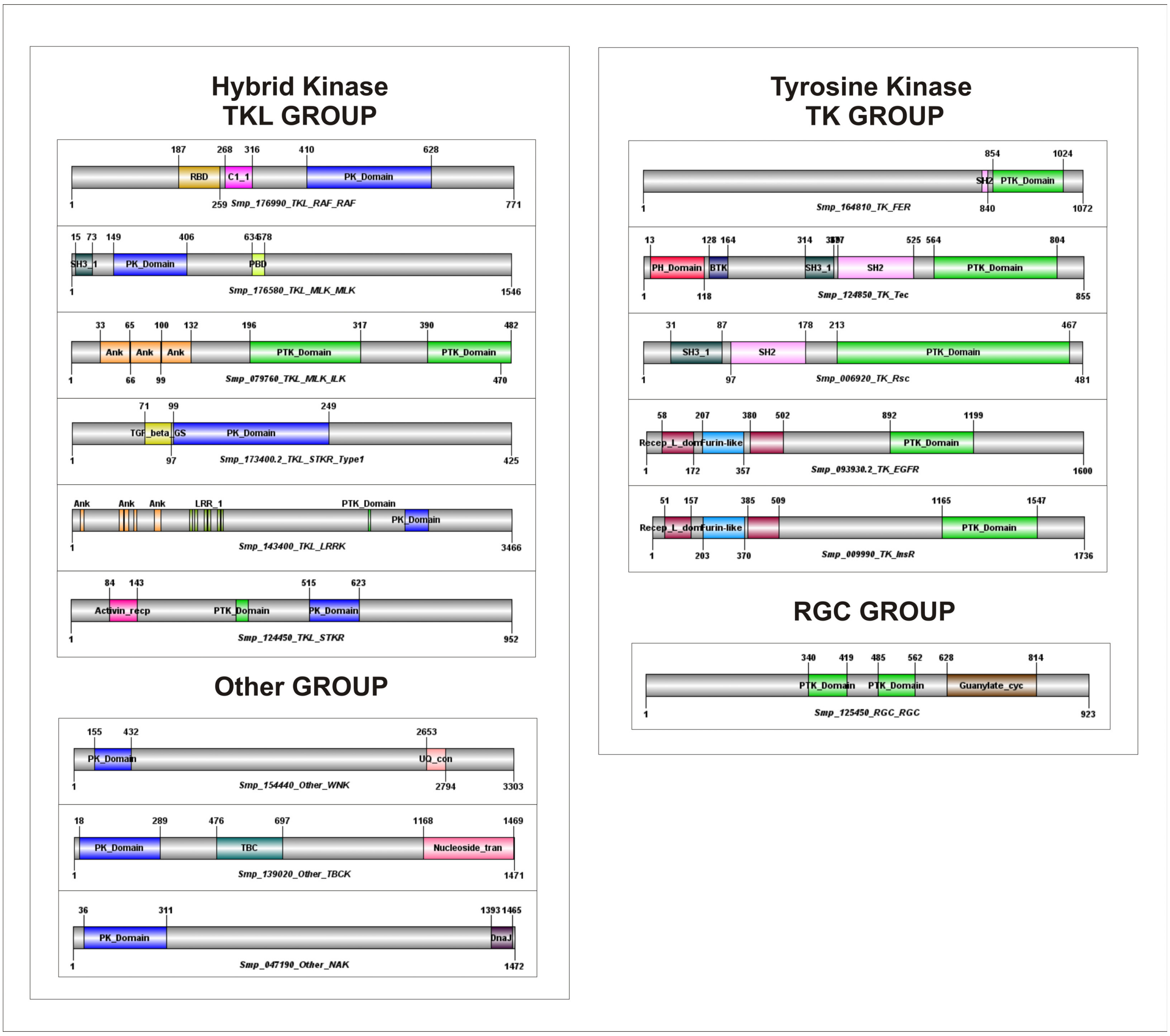


Figure 7 Accessory Domains. Representative domain organizations of some *S. mansoni* ePKs belonging to the TKL, Other, TK and, RGC groups are shown. Each protein ID and classification is shown below each image. Abbreviations followed are: PK_domain (Protein kinase domain), RBD (Raf-like Ras-binding domain), C1_1 (Phorbol esters/diacylglycerol binding domain), SH3_1 (Src homology 3 domain), PBD (P21-Rho-binding domain), Ank (Ankyrin repeat), PTK_domain (Protein tyrosine kinase domain), TGF_beta_GS (Transforming growth factor beta type I GS-motif), LRR_1 (Leucine Rich Repeat), Activin_recep (Activin types I and II receptor domain) UQ_con (Ubiquitin-conjugating enzyme) TBC (TBC domain), Nucleoside_tran (Nucleoside transporter) DnaJ (DnaJ domains), SH2 (Src homology 2 domain) PH_domain (Pleckstrin Homology domain) BTK (Bruton's tyrosine kinase motif) Recep_L_domain (Receptor L domain) furin-like (Furin-like cysteine rich region) Guanylate_cyc (Receptor family ligand binding region). The protein domain architectures were generated using DOG 1.0 [111] based on the Pfam domain limits [96].

C1_1 domain is conserved in N-terminal regions of all PKC proteins of *S. mansoni* (Figure 6) and has been shown to bind PE (phorbol esters) and DAG (diacylglycerol). DAG is an important second messenger and Phorbol esters are analogues of DAG [32,96]. The C1_1 domain is present in one or two copies depending on the isozyme of PKC (Figure 6). cNMP_binding is a N-terminal domain of PKG proteins that bind cyclic nucleotides (cAMP or cGMP) to relieve the inhibition of the catalytic domain [15,35]. The AKT protein of *S. mansoni* (Smp_162120) has an unusual domain combination (Figure 6) as the two C-terminal domains (Glutaredoxin and Na_K-ATPase) are not found in *D. melanogaster, C. elegans, M. musculus* and *H. sapiens.*

CASK is a member of the CaMK group and plays a key role in establishing inter-cellular contacts and plasticity at cellular junctions [54]. The accessory domains found in *S. mansoni* CASK protein (L27 and PDZ domains, which serve as protein interaction modules, SH3, and a C-terminal guanylate kinase domains) are conserved in higher eukaryotes. However, the UPF0061 (Pfam: PF02696) is uncharacterized [96] and possesses an unusual domain found in the C-terminal region of *S. mansoni* CASK protein (Figure 6). The long protein kinase MLCK (Figure 5a) possesses a large number of Ig repeats (I-set, V-set and Ig) that, in other species, are involved in a variety of functions, including cell-cell recognition, cell-surface receptors, muscle structure and the immune system [96], and fn3 repeats, that is an approximately 100 amino acid domain commonly found in a variety of organisms.

The CMGC and CK1 groups have none or a few accessory domains in *S. mansoni*. However, it is known that small regions in these proteins play an important role in recognizing and binding to the substrate [97,98]. For example, the CD domain (common docking domain) is a C-terminal region of MAPK proteins composed of a set of negatively charged amino acids that is used to anchor protein activators (such as STE proteins), substrates (such as MAPKAPK) and inactivating proteins (such as MAPK phosphatases) [99]. Thus, this region governs a series of signal transduction in the cascade of reactions of MAPKs. Other regions, including the ED (ERK docking) site, working with the CD domain and ensuring specificity and interaction strength [99].

PBD (p21-Rho-binding) and C-terminal CNH domain are usually found in the STE20 families (Figure 6). PBD binds to cdc42-GTPases activating the signaling cascade which act upstream in the MAPK cascade. The CNH domain interacts with the small GTPase and regulating the actin cytoskeleton [96].

The SH3 and SH2 (Pfam: PF00018 and PF00017, respectively) domains are common found in CTK proteins. SH2 function as regulatory modules of intracellular signaling cascades and it was found in eight out of 19 *S. mansoni* CTKs. Fer PTK is usually composed of three domains, FHC domain, SH2, and C-terminal kinase domain as it occurs in Fer proteins of *H. sapiens, M. musculus*, and *D. melanogaster*. However, the *S. mansoni* Fer protein (SmFes - Smp_164810) [86] and the 42 Fer proteins of *C. elegans* seems to have lost the N-terminal FHC domain (Figure 7). RTKs are characterized by an extracellular domains, a membrane spanning segment and an intracellular kinase domain [32]. The extracellular ligand binding domain of EGFR and InsR proteins are composed of two receptor_L sandwiching a Furin_like domain (Figure 7). SmVKR is composed of an unusual extracellular Venus flytrap module (VFT) linked through a single transmembrane domain to an intracellular tyrosine catalytic domain similar to that of the insulin receptor and a putative function in reproduction and development was observed [81]. Other extracellular domains found in *S. mansoni* are Ephrin_lbd (Pfam: PF01404) in the Ephrin recptors (Eph) and Ig domains (Pfam: PF00047) in CCK4 proteins (Additional file 1).

In conclusion, the protein architecture, including the accessory domains, may indicate potential protein partners. Signaling roles of schistosome specificities or unusual architectures are of special biological interest.

Conclusions

This study allowed us to identify and classify 252 ePKs encoded in the predicted proteome of *S. mansoni*. Together, these proteins represent 1.9% of the proteome and indicate that protein phosphorylation is an important mechanism for regulating the complex life cycle of the parasite. We improve the functional annotation of 40% of *S. mansoni* ePKs (Table 3) by applying a phylogenetic framework. Moreover, it was possible to gain insights into kinase function once 94% of the *S. mansoni* ePKinome had previously an unknown function. *S. mansoni* has proteins in each ePKs group. Most of them are clearly clustered with known kinases from other eukaryotes with no family being exclusively found or expanded in *S. mansoni*. Some proteins are not clustered with the main ePK family as the catalytic domain is truncate, indicating that the current gene/protein predictions require further refinement. Proteins were mentioned as potential targets for drug design and development as they may play an essential function in the parasite. Furthermore new and effective drugs bind PKs close but not in the ATP site and occlude ATP access to the kinase to retard enzyme activity [17]. So, proteins of *S. mansoni* with a sequence highly similar to host proteins can be used as protein

Table 3 *S. mansoni* ePKinome annotation improved by phylogenomics

	After phylogenetic analysis	
Modifications	**Number of changes**	**% of *S. mansoni* ePKinome**
Change group classification	14	5,55%
Change family classification	18	7,14%
Add subfamily	61	24,20%
Classification of unknown proteins	11	3,96%
Total of changes	**104**	**40,85%**

Through phylogenetic analysis, we have improved the functional annotation of 40% of *S. mansoni* ePKs by changing protein classification at the group and family levels, assigning subfamilies, and classifying proteins without any previous classification.

targets since the inhibitor binds in non-conserved residues outside the ATP site. Also, the unusual domains found in *S. mansoni* can be used for constructing more specific *S. mansoni* inhibitors. Moreover, as we continue this work, we will highlight the biochemical and physiological adaptations of *S. mansoni* in response to diverse environments during parasite development, vector interaction, and host infection.

Methods

Organisms and Sequences

S. mansoni [NCBI taxid: 6183] and six other organisms were selected for this work including *Homo sapiens* [taxid: 9606], *Mus musculus* [taxid: 10090], *Drosophila melanogaster* [taxid: 7227], *Caenorhabditis elegans* [taxid: 6239], *Brugia malayi* [taxid: 6279], and *Saccharomyces cerevisiae* [taxid: 4932]. The *S. mansoni* predicted proteome data was downloaded from *SchistoDB*, version [2.0] [29], which contains the original gene and genomic information provided by the Wellcome Trust Institute and described elsewhere [26]. Datasets of protein kinases from the other organisms were downloaded from the kinase database at Sugen/Salk - KinBase [15], except for *Brugia malayi*, which was retrieved from KEGG [100].

Functional Classification

Functional classification of protein kinases into groups, families, and subfamilies followed the proposed hierarchy described elsewhere [12,13,32]. Potential protein kinases of *S. mansoni* were identified and characterized by combined approaches based on sequence similarity and phylogenetic relationships. These proteins were first identified by similarity to Hidden Markov Models (HMMs) as described below. Also based on sequence similarity, each predicted protein kinase was manually annotated by integrating data from InterProScan [101] and reverse PSI-BLAST (rpsblast) [102] output searches into Artemis [103]. Further analysis was performed by HMMs searching for non-catalytic (accessory) domains associated to the conserved catalytic domain of protein kinases based on data available at the Protein families database - Pfam [96]. Functional classification was also devised based on the literature and on the assumption of a broad conservation of the molecular functions. Phylogenetic analyses of the ePK kinases groups performed in the present work corroborated this classification as well as supported new functional assignments for previously uncharacterized proteins (see below).

Hidden Markov Models

In order to identify potential homologs in *S. mansoni*, amino acid sequences of known protein kinases of five model organisms (*H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, and *S. cerevisiae*) were selected. A total of 68 diverse amino acid sequences corresponding to the kinase catalytic domain and sharing less than 50% sequence identity were aligned in MAFFT [30] and manually-edited for further analysis. Local and global HMMs were built with the HMMer package http://hmmer.janelia.org from multiple sequence alignments and used for sensitive searches against the *S. mansoni* proteome [26].

Phylogenetic Analyses

Amino acid sequences corresponding to the conserved catalytic domain (Pfam: PF0069 or PF07714) of each group of protein kinases were separately aligned using the default parameters of MAFFT [30]. Multiple sequence alignments were filtered to keep proteins sharing 50% to 90% pairwise sequence identity using the decreased redundancy tool [104] and manually edited to remove ambiguous regions using BioEdit [105]. Final alignments were used in phylogenetic reconstructions through multiple programs available in the Phylogeny Inference Package - PHYLIP, version 3.69 [31]. Initially, 1000 random datasets (replicates) were created for each alignment using *seqboot* with default parameters. For each dataset, it was calculated a distance matrix under the JTT model with gamma-distributed sites by *protdist*. Next, phylogenies were estimated from distance matrix data adopting the *Fitch-Margoliash* criterion as implemented in *fitch*. Finally, the results from the random datasets were summarized by *consense*, which computes consensus trees by the majority-rule consensus tree method. Phylogenetic trees were visualized and edited using the Tree Figure Drawing Tool - *FigTree*, version 1.3.1 [106]. Nodes with at least 80% bootstrap values were considered to support functional prediction.

Additional material

Additional file 1: Eukaryotic protein kinases of *Schistosoma mansoni*.

Additional file 2: Phylogenetic analysis of the AGC group. Amino acid sequences of the catalytic domain of the AGC proteins of *S. mansoni* (blue labels), *C. elegans*, *D. melanogaster*, *S. cerevisiae*, *H. sapiens*, and *B. malayi* were used to construct a distance-based phylogenetic tree using PHYLIP programs. Bootstrap values (1000 replicates) equal or higher than 80% are indicated (●) as cutoff values to support the family/subfamily classification. Functional classification is indicated individually in the protein labels. Protein families are highlighted by the vertical bars and include: RSK (Ribosomal S6 Kinase), PKA (Protein Kinase A), PKG (Protein Kinase G), PDK1(phosphoinositide-dependent protein kinase 1), NDR (nuclear Dbf2-related kinases), DMPK (Myotonic Dystrophy Protein Kinase), MAST (Microtubule Associated Serine/Threonine Kinase), RSKL (RSK-like); PKC (Protein Kinase C), PKN (Protein Kinase N), Akt (also known as PKB, Protein Kinase B), SGK (Serum and Glucocorticoid Resposive Kinase), GRK (G-protein coupled Receptor Kinase), YANK (Yet Another Novel Kinase).

Additional file 3: Distribution of some ePKs Families in *S. mansoni* and model organisms. *S. mansoni* proteins were classified according to KinBase [15] by combining sequence similarity searches (HMMs) and phylogenetic analysis (this work). For comparison, occurrence of the ePKs families in *C. elegans*, *D. melanogaster*, *S. cerevisiae*, and *H. sapiens* is shown.

Additional file 4: Phylogenetic analysis of the CaMK group. Amino acid sequences of the catalytic domain of the CaMK proteins of *S. mansoni* (blue labels), *C. elegans*, *D. melanogaster*, *S. cerevisiae*, *H. sapiens*, and *B. malayi* were used to construct a distance-based phylogenetic tree using PHYLIP programs. Bootstrap values (1000 replicates) equal or higher than 80% are indicated (●) as cutoff values to support the family/subfamily classification. Functional classification is indicated individually in the protein labels. Protein families are highlighted by the vertical bars and include: MAPKAPK (MAP Kinase Associated Protein Kinase), CAMK1 (CAMK family 1), PHK (Phosphorylase Kinase), Trio, DAPK (Death Associated Protein Kinase), MLCK (Myosin Light Chain Kinases), CASK, CAMK2 (CAMK family 2), DCAMKL (Doublecortin and CaMK-Like), PKD (Protein Kinase D), RAD53, Trbl, CAMKL (CAMKL-like).

Additional file 5: Phylogenetic analysis of the CMGC group. Amino acid sequences of the catalytic domain of the CMGC proteins of *S. mansoni* (blue labels), *C. elegans*, *D. melanogaster*, *S. cerevisiae*, *H. sapiens*, and *B. malayi* were used to construct a distance-based phylogenetic tree using PHYLIP programs. Bootstrap values (1000 replicates) equal or higher than 80% are indicated (●) as cutoff values to support the family/subfamily classification. Functional classification is indicated individually in the protein labels. Protein families are highlighted by the vertical bars and include: GSK (Gycogen Synthase 3 Kinase), MAPK (Mitogen Activated Protein Kinase), CK2 (Cell Kinase 2), CDKL (Cyclin Dependent Kinase Like), CDK (Cyclin Dependent Kinase), SRPK (SR Protein Kinase; phosphorylates SR splicing factors), DYRK (Dual-specificity Y (tyrosine) Regulated Kinase), CLK (CDC-Like Kinase).

Additional file 6: Phylogenetic analysis of the STE group. Amino acid sequences of the catalytic domain of the STE proteins of *S. mansoni* (blue labels), *C. elegans*, *D. melanogaster*, *S. cerevisiae*, *H. sapiens*, and *B. malayi* were used to construct a distance-based phylogenetic tree using PHYLIP programs. Bootstrap values (1000 replicates) equal or higher than 80% are indicated (●) as cutoff values to support the family/subfamily classification. Functional classification is indicated individually in the protein labels. Protein families are highlighted by the vertical bars and include: STE11 (MAP3K (MAP kinase kinase kinase) genes), STE20 (MAP4K (MAP kinase kinase kinase kinase) genes), STE7 (MAP2K (MAP kinase kinase) genes).

Additional file 7: Phylogenetic analysis of the TK group. Amino acid sequences of the catalytic domain of the TK proteins of *S. mansoni* (blue labels), *C. elegans*, *D. melanogaster*, *S. cerevisiae*, *H. sapiens*, and *B. malayi* were used to construct a distance-based phylogenetic tree using PHYLIP programs. Bootstrap values (1000 replicates) equal or higher than 80% are indicated (●) as cutoff values to support the family/subfamily classification. Functional classification is indicated individually in the protein labels. Protein families are highlighted by the vertical bars and include: JAK, Ack (Activated Cdc42-associated tyrosine kinase), Syk, FAK (Focal Adhesion Kinase), VKR, Fer, Sev, ALK (Anaplastic Lymphoma Kinase), InsR (Insulin Receptor), DDR (Discoidin Domain Receptor kinase), Musk (Muscle-Specific Kinase), Ror, Lmr (Lemur Kinase), Eph (Ephrin receptors), Axl (Also known as TAM (Tyro3, Axl, Mer)), Ryk, EGFR (Epidermal Growth Factor Receptor), FGFR (Fibroblast Growth Factor Receptor), PDGFR, VEGFR, Kin16, Jak (Janus Kinases), Tec, Abl (Abelson murine leukemia homolog), and Src.

Additional file 8: Phylogenetic analysis of the RGC group. Amino acid sequences of the catalytic domain of the RGC proteins of *S. mansoni* (blue labels), *C. elegans*, *D. melanogaster*, *S. cerevisiae*, *H. sapiens*, and *B. malayi* were used to construct a distance-based phylogenetic tree using PHYLIP programs. Bootstrap values (1000 replicates) equal or higher than 80% are indicated (●) as cutoff values to support the family/subfamily classification. Functional classification is indicated individually in the protein labels.

Additional file 9: Phylogenetic analysis of the TKL group. Amino acid sequences of the catalytic domain of the CK1 proteins of *S. mansoni* (blue labels), *C. elegans*, *D. melanogaster*, *S. cerevisiae*, *H. sapiens*, and *B. malayi* were used to construct a distance-based phylogenetic tree using PHYLIP programs. Bootstrap values (1000 replicates) equal or higher than 80% are indicated (●) as cutoff values to support the family/subfamily classification. Functional classification is indicated individually in the protein labels. Protein families are highlighted by the vertical bars and include: IRAK (IL1 Receptor Associated Kinase), RIPK (Receptor Interacting Protein Kinases), LISK (Family containing closely related TESK and LIMK sub-families), MLK (Miked Lineage Kinases), LRRK (Leucine Rich Repeat Kinase), RAF, STKR (Serine/Threonine Kinase Receptors - receptors for activin and TGFb ligands).

Abbreviations

ATP: adenosine triphosphate; cAMP: cyclic adenosine monophosphate; GTP: guanosine triphosphate; NCBI: National Center for Biotechnology Information; RNAi: RNA interference.

Acknowledgements

This work was funded by the National Institutes of Health - NIH/Fogarty International Center (PPM-00439-10 to GO), Secretaria de Ciência, Tecnologia e Ensino Superior de Minas Gerais, Fundação de Amparo à Pesquisa do Estado de Minas Gerais, SECTES/FAPEMIG, Brazil (CBB-1181/08 and PPM-00439-10 to GO), and Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (CNPq-Universal 476036/2010-0 to LAN, CNPq-INCT 573839/2008-5 to GO, and INCT-DT 573839/2008-5 to GO). LFA is a fellow of FIOCRUZ. The authors acknowledge the use of the computing resources of the Center for Excellence in Bioinformatics, CEBio/FIOCRUZ, Brazil.

Author details

[1]Genomics and Computational Biology Group, Instituto Nacional de Ciência e Tecnologia em Doenças Tropicais, Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz - FIOCRUZ, Belo Horizonte, MG- 30190-002, Brazil. [2]Centro de Excelência em Bioinformática, Fundação Oswaldo Cruz - FIOCRUZ, Belo Horizonte, MG- 30190-110, Brazil. [3]Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais - UFMG, Belo Horizonte, MG- 31270-910, Brazil.

Authors' contributions

LFA: carried out the functional annotation and phylogenetic studies, and drafted the manuscript. LAN: designed the phylogenetic studies, performed preliminary phylogenetic studies, coordinated the functional prediction, and co-wrote this manuscript. LGAV and LLS carried out the functional annotation based on sequence similarity search using HMMs. AZ wrote the Perl scripts for data manipulation and provided computational support for this study. JCR designed and performed the HMMs analysis. GO participated in the design and coordination of this study, and co-wrote the manuscript. All authors read and approved the final manuscript.

Received: 23 November 2010 Accepted: 6 May 2011
Published: 6 May 2011

References

1. Haas W, Schmitt R: **Characterization of chemical stimuli for the penetration of Schistosoma mansoni cercariae. I. Effective substances, host specificity.** *Z Parasitenkd* 1982, **66**:293-307.
2. Haas W, Diekhoff D, Koch K, Schmalfuss G, Loy C: **Schistosoma mansoni cercariae: stimulation of acetabular gland secretion is adapted to the chemical composition of mammalian skin.** *J Parasitol* 1997, **83**:1079-1085.
3. Han ZG, Brindley PJ, Wang SY, Chen Z: **Schistosoma genomics: new perspectives on schistosome biology and host-parasite interaction.** *Annu Rev Genomics Hum Genet* 2009, **10**:211-240.
4. Kunz W: **Schistosome male-female interaction: induction of germ-cell differentiation.** *Trends Parasitol* 2001, **17**:227-231.
5. LoVerde PT, Hirai H, Merrick JM, Lee NH, El-Sayed N: **Schistosoma mansoni genome project: an update.** *Parasitol Int* 2004, **53**:183-192.
6. Fitzpatrick JM, Johnston DA, Williams GW, Williams DJ, Freeman TC, Dunne DW, Hoffmann KF: **An oligonucleotide microarray for transcriptome analysis of Schistosoma mansoni and its application/use to investigate gender-associated gene expression.** *Mol Biochem Parasitol* 2005, **141**:1-13.
7. Waisberg M, Lobo FP, Cerqueira GC, Passos LK, Carvalho OS, Franco GR, El-Sayed NM: **Microarray analysis of gene expression induced by sexual contact in Schistosoma mansoni.** *BMC Genomics* 2007, **8**:181.

8. Waisberg M, Lobo FP, Cerqueira GC, Passos LK, Carvalho OS, El-Sayed NM, Franco GR: **Schistosoma mansoni: Microarray analysis of gene expression induced by host sex.** *Exp Parasitol* 2008, **120**:357-363.
9. Hanks SK, Quinn AM, Hunter T: **The protein kinase family: conserved features and deduced phylogeny of the catalytic domains.** *Science* 1988, **241**:42-52.
10. Manning G, Plowman GD, Hunter T, Sudarsanam S: **Evolution of protein kinase signaling from yeast to man.** *Trends Biochem Sci* 2002, **27**:514-520.
11. Hanks SK: **Genomic analysis of the eukaryotic protein kinase superfamily: a perspective.** *Genome Biol* 2003, **4**:111.
12. Hanks SK, Hunter T: **Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification.** *FASEB J* 1995, **9**:576-596.
13. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S: **The protein kinase complement of the human genome.** *Science* 2002, **298**:1912-1934.
14. Miranda-Saavedra D, Barton GJ: **Classification and functional annotation of eukaryotic protein kinases.** *Proteins* 2007, **68**:893-914.
15. KinBase: **Kinbase - The kinase database at Sugen/Salk.** 2010 [http://www.kinase.com/kinbase], 8-8-2009. Ref Type: Electronic Citation.
16. Boyle SN, Koleske AJ: **Dissecting kinase signaling pathways.** *Drug Discov Today* 2007, **12**:717-724.
17. Eglen RM, Reisine T: **The current status of drug discovery against the human kinome.** *Assay Drug Dev Technol* 2009, **7**:22-43.
18. Dissous C, Ahier A, Khayath N: **Protein tyrosine kinases as new potential targets against human schistosomiasis.** *Bioessays* 2007, **29**:1281-1288.
19. Naula C, Parsons M, Mottram JC: **Protein kinases as drug targets in trypanosomes and Leishmania.** *Biochim Biophys Acta* 2005, **1754**:151-159.
20. Ward P, Equinet L, Packer J, Doerig C: **Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote.** *BMC Genomics* 2004, **5**:79.
21. Beckmann S, Grevelding CG: **Imatinib has a fatal impact on morphology, pairing stability and survival of adult Schistosoma mansoni in vitro.** *Int J Parasitol* 2010, **40**:521-526.
22. Doenhoff MJ, Pica-Mattoccia L: **Praziquantel for the treatment of schistosomiasis: its use for control in areas with endemic disease and prospects for drug resistance.** *Expert Rev Anti Infect Ther* 2006, **4**:199-210.
23. Doenhoff MJ, Cioli D, Utzinger J: **Praziquantel: mechanisms of action, resistance and new derivatives for schistosomiasis.** *Curr Opin Infect Dis* 2008, **21**:659-667.
24. Melman SD, Steinauer ML, Cunningham C, Kubatko LS, Mwangi IN, Wynn NB, Mutuku MW, Karanja DM, Colley DG, Black CL, Secor WE, Mkoji GM, Loker ES: **Reduced susceptibility to praziquantel among naturally occurring Kenyan isolates of Schistosoma mansoni.** *PLoS Negl Trop Dis* 2009, **3**:e504.
25. Pica-Mattoccia L, Doenhoff MJ, Valle C, Basso A, Troiani AR, Liberti P, Festucci A, Guidi A, Cioli D: **Genetic analysis of decreased praziquantel sensitivity in a laboratory strain of Schistosoma mansoni.** *Acta Trop* 2009, **111**:82-85.
26. Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyama ST, Al-Lazikani B, Andrade LF, Ashton PD, Aslett MA, Bartholomeu DC, Blandin G, Caffrey CR, Coghlan A, Coulson R, Day TA, Delcher A, DeMarco R, Djikeng A, Eyre T, Gamble JA, Ghedin E, Gu Y, Hertz-Fowler C, Hirai H, Hirai Y, Houston R, Ivens A, Johnston DA, *et al*: **The genome of the blood fluke Schistosoma mansoni.** *Nature* 2009, **460**:352-358.
27. Verjovski-Almeida S, DeMarco R, Martins EA, Guimaraes PE, Ojopi EP, Paquola AC, Piazza JP, Nishiyama MY Jr, Kitajima JP, Adamson RE, Ashton PD, Bonaldo MF, Coulson PS, Dillon GP, Farias LP, Gregorio SP, Ho PL, Leite RA, Malaquias LC, Marques RC, Miyasato PA, Nascimento AL, Ohlweiler FP, Reis EM, Ribeiro MA, Sa RG, Stukart GC, Soares MB, Gargioni C, Kawano T, *et al*: **Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni.** *Nat Genet* 2003, **35**:148-157.
28. Webster JP, Oliviera G, Rollinson D, Gower CM: **Schistosome genomes: a wealth of information.** *Trends Parasitol* 2010, **26**:103-106.
29. Zerlotini A, Heiges M, Wang H, Moraes RL, Dominitini AJ, Ruiz JC, Kissinger JC, Oliveira G: **SchistoDB: a Schistosoma mansoni genome resource.** *Nucleic Acids Res* 2009, **37**:D579-D582.
30. Katoh K, Asimenos G, Toh H: **Multiple alignment of DNA sequences with MAFFT.** *Methods Mol Biol* 2009, **537**:39-64.
31. Felsenstein J: **Mathematics vs. Evolution: Mathematical Evolutionary Theory.** *Science* 1989, **246**:941-942.
32. Deshmukh K, Anamika K, Srinivasan N: **Evolution of domain combinations in protein kinases and its implications for functional diversity.** *Prog Biophys Mol Biol* 2010, **102**:1-15.
33. Aguero F, Al-Lazikani B, Aslett M, Berriman M, Buckner FS, Campbell RK, Carmona S, Carruthers IM, Chan AW, Chen F, Crowther GJ, Doyle MA, Hertz-Fowler C, Hopkins AL, McAllister G, Nwaka S, Overington JP, Pain A, Paolini GV, Pieper U, Ralph SA, Riechers A, Roos DS, Sali A, Shanmugam D, Suzuki T, Van Voorhis WC, Verlinde CL: **Genomic-scale prioritization of drug targets: the TDR Targets database.** *Nat Rev Drug Discov* 2008, **7**:900-907.
34. Boudeau J, Miranda-Saavedra D, Barton GJ, Alessi DR: **Emerging roles of pseudokinases.** *Trends Cell Biol* 2006, **16**:443-452.
35. Taylor SS, Buechler JA, Yonemoto W: **cAMP-dependent protein kinase: framework for a diverse family of regulatory enzymes.** *Annu Rev Biochem* 1990, **59**:971-1005.
36. Donald RG, Allocco J, Singh SB, Nare B, Salowe SP, Wiltsie J, Liberator PA: **Toxoplasma gondii cyclic GMP-dependent kinase: chemotherapeutic targeting of an essential parasite protein kinase.** *Eukaryot Cell* 2002, **1**:317-328.
37. Gurnett AM, Liberator PA, Dulski PM, Salowe SP, Donald RG, Anderson JW, Wiltsie J, Diaz CA, Harris G, Chang B, rkin-Rattray SJ, Nare B, Crumley T, Blum PS, Misura AS, Tamas T, Sardana MK, Yuan J, Biftu T, Schmatz DM: **Purification and molecular characterization of cGMP-dependent protein kinase from Apicomplexan parasites. A novel chemotherapeutic target.** *J Biol Chem* 2002, **277**:15913-15922.
38. Diaz CA, Allocco J, Powles MA, Yeung L, Donald RG, Anderson JW, Liberator PA: **Characterization of Plasmodium falciparum cGMP-dependent protein kinase (PfPKG): antiparasitic activity of a PKG inhibitor.** *Mol Biochem Parasitol* 2006, **146**:78-88.
39. Knockaert M, Gray N, Damiens E, Chang YT, Grellier P, Grant K, Fergusson D, Mottram J, Soete M, Dubremetz JF, Le RK, Doerig C, Schultz P, Meijer L: **Intracellular targets of cyclin-dependent kinase inhibitors: identification by affinity chromatography using immobilised inhibitors.** *Chem Biol* 2000, **7**:411-422.
40. Swierczewski BE, Davies SJ: **A schistosome cAMP-dependent protein kinase catalytic subunit is essential for parasite viability.** *PLoS Negl Trop Dis* 2009, **3**:e505.
41. Mellor H, Parker PJ: **The extended protein kinase C superfamily.** *Biochem J* 1998, **332(Pt 2)**:281-292.
42. Parker PJ, Murray-Rust J: **PKC at a glance.** *J Cell Sci* 2004, **117**:131-132.
43. Eckert JJ, McCallum A, Mears A, Rumsby MG, Cameron IT, Fleming TP: **Specific PKC isoforms regulate blastocoel formation during mouse preimplantation development.** *Dev Biol* 2004, **274**:384-401.
44. Gallicano GI, Yousef MC, Capco DG: **PKC–a pivotal regulator of early development.** *Bioessays* 1997, **19**:29-36.
45. Steinhart R, Kazimirsky G, Okhrimenko H, Ben-Hur T, Brodie C: **PKCepsilon induces astrocytic differentiation of multipotential neural precursor cells.** *Glia* 2007, **55**:224-232.
46. Bahia D, Avelar LG, Vigorosi F, Cioli D, Oliveira GC, Mortara RA: **The distribution of motor proteins in the muscles and flame cells of the Schistosoma mansoni miracidium and primary sporocyst.** *Parasitology* 2006, **133**:321-329.
47. Bahia D, Avelar L, Mortara RA, Khayath N, Yan Y, Noel C, Capron M, Dissous C, Pierce RJ, Oliveira G: **SmPKC1, a new protein kinase C identified in the platyhelminth parasite Schistosoma mansoni.** *Biochem Biophys Res Commun* 2006, **345**:1138-1148.
48. Ludtmann MH, Rollinson D, Emery AM, Walker AJ: **Protein kinase C signalling during miracidium to mother sporocyst development in the helminth parasite, Schistosoma mansoni.** *Int J Parasitol* 2009, **39**:1223-1233.
49. Gally C, Wissler F, Zahreddine H, Quintin S, Landmann F, Labouesse M: **Myosin II regulation during C. elegans embryonic elongation: LET-502/ROCK, MRCK-1 and PAK-1, three kinases with different roles.** *Development* 2009, **136**:3109-3119.
50. Swulius MT, Waxham MN: **Ca(2+)/calmodulin-dependent protein kinases.** *Cell Mol Life Sci* 2008, **65**:2637-2657.
51. Plowman GD, Sudarsanam S, Bingham J, Whyte D, Hunter T: **The protein kinases of Caenorhabditis elegans: a model for signal transduction in multicellular organisms.** *Proc Natl Acad Sci USA* 1999, **96**:13603-13610.
52. Capes-Davis A, Tolhurst O, Dunn JM, Jeffrey PL: **Expression of doublecortin (DCX) and doublecortin-like kinase (DCLK) within the developing chick brain.** *Dev Dyn* 2005, **232**:457-467.

53. Lin PT, Gleeson JG, Corbo JC, Flanagan L, Walsh CA: **DCAMKL1 encodes a protein kinase with homology to doublecortin that regulates microtubule polymerization.** *J Neurosci* 2000, **20**:9152-9161.
54. Lehtonen S, Ryan JJ, Kudlicka K, Iino N, Zhou H, Farquhar MG: **Cell junction-associated proteins IQGAP1, MAGI-2, CASK, spectrins, and alpha-actinin are components of the nephrin multiprotein complex.** *Proc Natl Acad Sci USA* 2005, **102**:9814-9819.
55. Caffrey CR, Rohwer A, Oellien F, Marhofer RJ, Braschi S, Oliveira G, McKerrow JH, Selzer PM: **A comparative chemogenomics strategy to predict potential drug targets in the metazoan pathogen, Schistosoma mansoni.** *PLoS One* 2009, **4**:e4413.
56. Bradham CA, Foltz KR, Beane WS, Arnone MI, Rizzo F, Coffman JA, Mushegian A, Goel M, Morales J, Geneviere AM, Lapraz F, Robertson AJ, Kelkar H, Loza-Coll M, Townley IK, Raisch M, Roux MM, Lepage T, Gache C, McClay DR, Manning G: **The sea urchin kinome: a first look.** *Dev Biol* 2006, **300**:180-193.
57. Kalis AK, Kroetz MB, Larson KM, Zarkower D: **Functional genomic identification of genes required for male gonadal differentiation in Caenorhabditis elegans.** *Genetics* 2010, **185**:523-535.
58. Dohrmann PR, Sclafani RA: **Novel role for checkpoint Rad53 protein kinase in the initiation of chromosomal DNA replication in Saccharomyces cerevisiae.** *Genetics* 2006, **174**:87-99.
59. Noselli S: **JNK signaling and morphogenesis in Drosophila.** *Trends Genet* 1998, **14**:33-38.
60. Stronach B, Perrimon N: **Activation of the JNK pathway during dorsal closure in Drosophila requires the mixed lineage kinase, slipper.** *Genes Dev* 2002, **16**:377-387.
61. Kondoh K, Nishida E: **Regulation of MAP kinases by MAP kinase phosphatases.** *Biochim Biophys Acta* 2007, **1773**:1227-1237.
62. Nishimoto S, Nishida E: **MAPK signalling: ERK5 versus ERK1/2.** *EMBO Rep* 2006, **7**:782-786.
63. Wilkinson MG, Millar JB: **Control of the eukaryotic cell cycle by MAP kinase signaling pathways.** *FASEB J* 2000, **14**:2147-2157.
64. Sundaram MV: **RTK/Ras/MAPK signaling.** *WormBook* 2006, 1-19.
65. Yan Y, Tulasne D, Browaeys E, Cailliau K, Khayath N, Pierce RJ, Trolet J, Fafeur V, Ben YA, Dissous C: **Molecular cloning and characterisation of SmSLK, a novel Ste20-like kinase in Schistosoma mansoni.** *Int J Parasitol* 2007, **37**:1539-1550.
66. Hu W, Brindley PJ, McManus DP, Feng Z, Han ZG: **Schistosome transcriptomes: new insights into the parasite and schistosomiasis.** *Trends Mol Med* 2004, **10**:217-225.
67. Salzet M, Capron A, Stefano GB: **Molecular crosstalk in host-parasite relationships: schistosome- and leech-host interactions.** *Parasitol Today* 2000, **16**:536-540.
68. Kemphues K: **Essential genes.** *WormBook* 2005, 1-7.
69. Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, Mao M, Li B, Cavet G, Linsley PS: **Expression profiling reveals off-target gene regulation by RNAi.** *Nat Biotechnol* 2003, **21**:635-637.
70. Moffat J, Reiling JH, Sabatini DM: **Off-target effects associated with long dsRNAs in Drosophila RNAi screens.** *Trends Pharmacol Sci* 2007, **28**:149-151.
71. Wilson RA, Ashton PD, Braschi S, Dillon GP, Berriman M, Ivens A: **'Oming in on schistosomes: prospects and limitations for post-genomics.** *Trends Parasitol* 2007, **23**:14-20.
72. Muhlrad PJ, Ward S: **Spermiogenesis initiation in Caenorhabditis elegans involves a casein kinase 1 encoded by the spe-6 gene.** *Genetics* 2002, **161**:143-155.
73. Reinke V, Smith HE, Nance J, Wang J, Van DC, Begley R, Jones SJ, Davis EB, Scherer S, Ward S, Kim SK: **A global profile of germline gene expression in C. elegans.** *Mol Cell* 2000, **6**:605-616.
74. Bahia D, Andrade LF, Ludolf F, Mortara RA, Oliveira G: **Protein tyrosine kinases in Schistosoma mansoni.** *Mem Inst Oswaldo Cruz* 2006, **101(Suppl 1)**:137-143.
75. Ahier A, Khayath N, Vicogne J, Dissous C: **Insulin receptors and glucose uptake in the human parasite Schistosoma mansoni.** *Parasite* 2008, **15**:573-579.
76. Khayath N, Vicogne J, Ahier A, BenYounes A, Konrad C, Trolet J, Viscogliosi E, Brehm K, Dissous C: **Diversification of the insulin receptor family in the helminth parasite Schistosoma mansoni.** *FEBS J* 2007, **274**:659-676.
77. Kim JJ, Accili D: **Signalling through IGF-I and insulin receptors: where is the specificity?** *Growth Horm IGF Res* 2002, **12**:84-90.
78. Van BC, Sternberg PW: **Epidermal growth factor signaling induces behavioral quiescence in Caenorhabditis elegans.** *Nat Neurosci* 2007, **10**:1300-1307.
79. Ramachandran H, Skelly PJ, Shoemaker CB: **The Schistosoma mansoni epidermal growth factor receptor homologue, SER, has tyrosine kinase activity and is localized in adult muscle.** *Mol Biochem Parasitol* 1996, **83**:1-10.
80. Vicogne J, Cailliau K, Tulasne D, Browaeys E, Yan YT, Fafeur V, Vilain JP, Legrand D, Trolet J, Dissous C: **Conservation of epidermal growth factor receptor function in the human parasitic helminth Schistosoma mansoni.** *J Biol Chem* 2004, **279**:37407-37414.
81. Ahier A, Rondard P, Gouignard N, Khayath N, Huang S, Trolet J, Donoghue DJ, Gauthier M, Pin JP, Dissous C: **A new family of receptor tyrosine kinases with a venus flytrap binding domain in insects and other invertebrates activated by aminoacids.** *PLoS One* 2009, **4**:e5651.
82. Kapp K, Schussler P, Kunz W, Grevelding CG: **Identification, isolation and characterization of a Fyn-like tyrosine kinase from Schistosoma mansoni.** *Parasitology* 2001, **122**:317-327.
83. Kapp K, Knobloch J, Schussler P, Sroka S, Lammers R, Kunz W, Grevelding CG: **The Schistosoma mansoni Src kinase TK3 is expressed in the gonads and likely involved in cytoskeletal organization.** *Mol Biochem Parasitol* 2004, **138**:171-182.
84. Knobloch J, Winnen R, Quack M, Kunz W, Grevelding CG: **A novel Syk-family tyrosine kinase from Schistosoma mansoni which is preferentially transcribed in reproductive organs.** *Gene* 2002, **294**:87-97.
85. Bahia D, Mortara RA, Kusel JR, Andrade LF, Ludolf F, Kuser PR, Avelar L, Trolet J, Dissous C, Pierce RJ, Oliveira G: **Schistosoma mansoni: expression of Fes-like tyrosine kinase SmFes in the tegument and terebratorium suggests its involvement in host penetration.** *Exp Parasitol* 2007, **116**:225-232.
86. Ludolf F, Bahia D, Andrade LF, Cousin A, Capron M, Dissous C, Pierce RJ, Oliveira G: **Molecular analysis of SmFes, a tyrosine kinase of Schistosoma mansoni orthologous to the members of the Fes/Fps/Fer family.** *Biochem Biophys Res Commun* 2007, **360**:163-172.
87. Manning G: **Genomic overview of protein kinases.** *WormBook* 2005, 1-19.
88. Vowels JJ, Thomas JH: **Genetic analysis of chemosensory control of dauer formation in Caenorhabditis elegans.** *Genetics* 1992, **130**:105-123.
89. Parsons M, Worthey EA, Ward PN, Mottram JC: **Comparative analysis of the kinomes of three pathogenic trypanosomatids: Leishmania major, Trypanosoma brucei and Trypanosoma cruzi.** *BMC Genomics* 2005, **6**:127.
90. Pils B, Schultz J: **Inactive enzyme-homologues find new function in regulatory processes.** *J Mol Biol* 2004, **340**:399-404.
91. Beall MJ, Pearce EJ: **Human transforming growth factor-beta activates a receptor serine/threonine kinase from the intravascular parasite Schistosoma mansoni.** *J Biol Chem* 2001, **276**:31613-31619.
92. Davies SJ, Shoemaker CB, Pearce EJ: **A divergent member of the transforming growth factor beta receptor family from Schistosoma mansoni is expressed on the parasite surface membrane.** *J Biol Chem* 1998, **273**:11234-11240.
93. Forrester SG, Warfel PW, Pearce EJ: **Tegumental expression of a novel type II receptor serine/threonine kinase (SmRK2) in Schistosoma mansoni.** *Mol Biochem Parasitol* 2004, **136**:149-156.
94. LoVerde PT, Osman A, Hinck A: **Schistosoma mansoni: TGF-beta signaling pathways.** *Exp Parasitol* 2007, **117**:304-317.
95. Osman A, Niles EG, Verjovski-Almeida S, LoVerde PT: **Schistosoma mansoni TGF-beta receptor II: role in host ligand-induced regulation of a schistosome target gene.** *PLoS Pathog* 2006, **2**:e54.
96. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38**:D211-D222.
97. Tanoue T, Yamamoto T, Nishida E: **Modular structure of a docking surface on MAPK phosphatases.** *J Biol Chem* 2002, **277**:22942-22949.
98. Zhou B, Wu L, Shen K, Zhang J, Lawrence DS, Zhang ZY: **Multiple regions of MAP kinase phosphatase 3 are involved in its recognition and activation by ERK2.** *J Biol Chem* 2001, **276**:6506-6515.
99. Tanoue T, Adachi M, Moriguchi T, Nishida E: **A conserved docking motif in MAP kinases common to substrates, activators and regulators.** *Nat Cell Biol* 2000, **2**:110-116.
100. KEGG: **Kyoto Encyclopedia of Genes and Genomes.** 2000 [http://www.genome.jp/kegg/], 6-10-2009. Ref Type: Electronic Citation.

101. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, *et al*: **InterPro: the integrative protein signature database.** *Nucleic Acids Res* 2009, **37**:D211-D215.
102. Bhagwat M, Aravind L: **PSI-BLAST.** In *methods in Molecular Biology* Edited by: Clifton NJ 2007, 86-177.
103. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16**:944-945.
104. Notredame C: **Decrease redundancy tool.** 2009 [http://www.expasy.ch/tools/redundancy], 20-11-2009. Ref Type: Electronic Citation.
105. Hall T: **Bioedit: a user-friendly bioçogical sequence alignment editor and analysis program for Windows 95/98/NT.** *Nucl Acids Symp Ser* 1999, **41**:95-98.
106. Rambaut A: **Tree Figure Drawing Tool, Version 1.3.1.** 2010 [http://tree.bio.ed.ac.uk/software/figtree], 1-3-2010. Ref Type: Electronic Citation.
107. Anamika K, Srinivasan N: **Comparative kinomics of Plasmodium organisms: unity in diversity.** *Protein Pept Lett* 2007, **14**:509-517.
108. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, Allen JE, Delcher AL, Guiliano DB, Miranda-Saavedra D, Angiuoli SV, Creasy T, Amedeo P, Haas B, El-Sayed NM, Wortman JR, Feldblyum T, Tallon L, Schatz M, Shumway M, Koo H, Salzberg SL, Schobel S, Pertea M, Pop M, White O, Barton GJ, Carlow CK, Crawford MJ, Daub J, *et al*: **Draft genome of the filarial nematode parasite Brugia malayi.** *Science* 2007, **317**:1756-1760.
109. Caenepeel S, Charydczak G, Sudarsanam S, Hunter T, Manning G: **The mouse kinome: discovery and comparative genomics of all mouse protein kinases.** *Proc Natl Acad Sci USA* 2004, **101**:11707-11712.
110. Zdobnov EM, von MC, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, Mueller HM, Dimopoulos G, Law JH, Wells MA, Birney E, Charlab R, Halpern AL, Kokoza E, Kraft CL, Lai Z, Lewis S, Louis C, Barillas-Mury C, Nusskern D, Rubin GM, Salzberg SL, Sutton GG, Topalis P, Wides R, Wincker P, *et al*: **Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster.** *Science* 2002, **298**:149-159.
111. Ren J, Wen L, Gao X, Jin C, Xue Y, Yao X: **DOG 1.0: illustrator of protein domain structures.** *Cell Res* 2009, **19**:271-273.

doi:10.1186/1471-2164-12-215
Cite this article as: Andrade *et al.*: Eukaryotic Protein Kinases (ePKs) of the Helminth Parasite *Schistosoma mansoni*. *BMC Genomics* 2011 **12**:215.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

3.4 - CAPÍTULO IV: Investigação do papel biológico de MAPKs

Andrade LF, Mourão MM, Geraldo JA, Coelho FS, Silva LL, Neves RH, Silva JRM, Pimenta R, Caffrey C, Oliveira G. Involvement of mitogen-activated protein kinase signaling pathway in *Schistosoma mansoni* development and reproduction during mammalian host infection

Neste estudo, em colaboração com o Dr. Conor Cafrey (Universidade da California) e Dr. José Roberto M. Silva (UFRJ), analisamos o papel biológico de membros da família MAPK (Mitogen-Activated Protein Kinase) da superfamília proteína quinases eucarióticas (ePKs). As ePKs de *S. mansoni*, *S. japonicum* e *S. haematobium* foram previamente identificadas e classificadas em grupos, famílias e subfamílias utilizando Modelos Ocultos de Markov (HMM) e abordagem filogenética (Capítulo III). Baseados nas informações obtidas neste estudo, foi possível selecionar um grupo de ePKs cujas funções foram preditas como sendo essenciais para a sobrevivência e reprodução desses parasitos. Dentre estas, destacam-se as ativadoras/efetoras da via de sinalização MAPK. Em virtude do exposto, proteínas chave da via MAPK (SmRas, SmERK1, SmERK2, SmJNK e SmCaMK2), foram selecionadas para caracterização experimental através do silenciamento da expressão gênica por RNA de interferência (RNAi). Após redução significativa no nível de transcritos dos genes selecionados, nenhuma alteração fenotípica visível foi observada na cultura de esquistossômulos. Ainda assim, esquistossômulos silenciados foram inoculados em camundongos com o objetivo de avaliar potenciais efeitos da diminuição nos níveis transcricionais durante a interação parasito-hospedeiro. Resultados bastante interessantes mostraram que a expressão de proteínas MAPK da subfamília JNK, quando silenciadas, causam efeitos devastadores no tegumento de vermes adultos de *S. mansoni*, levando à morte dos mesmos. Além disso, nossos resultados mostram que ePKs da subfamília ERK1 estão relacionadas com a produção de ovos, já que fêmeas com baixos níveis de transcritos SmERK1 e SmERK2 apresentam ovários pouco desenvolvidos e produção de ovos significativamente baixa. Portanto, as proteínas MAPK, SmERK e SmJNK, são alvos de interesse para o desenvolvimento de drogas contra a esquistossomose visto que um inibidor dessas proteínas provavelmente irá interromper o ciclo de vida de *Schistosoma* e impedir o progresso da doença.

Involvement of Mitogen-Activated Protein Kinase Signaling Pathway in *Schistosoma mansoni* Development and Reproduction During Mammalian Host Infection

***Luiza Freire de Andrade[1,2]; *Marina de Moraes Mourão[1,3]; Juliana Assis Geraldo[1,3]; Fernando Sales Coelho[1,4]; Larissa Lopes Silva [1,3,4]; Renata Heisler Neves [5,6]; José Roberto Machado-Silva[6]; Neusa Araujo[7]; Rafael Nacif-Pimenta[8]; Conor Caffrey[9]; §Guilherme Oliveira[1,3].**

[1]Grupo de Genômica e Biologia Computacional, Instituto Nacional de Ciência e Tecnologia em Doenças Tropicais, Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz – FIOCRUZ, Belo Horizonte, MG- 30190-002, Brazil;

[2]Centro de Oncologia Molecular, Instituto Sírio-Libanês de Ensino e Pesquisa, Hospital Sírio Libanês, São Paulo, SP 01308-060, Brasil

[3]Centro de Excelência em Bioinformática- CEBio, Fundação Oswaldo Cruz - FIOCRUZ, Belo Horizonte, MG- 30190-110, Brazil;

[4]Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais - UFMG, Belo Horizonte, MG- 31270-910, Brazil;

[5]Biologia Parasitária, Departamento de Ensino, Pavilhão Arthur Neiva, Instituto Oswaldo Cruz, Fiocruz, RJ- 21045-900, Brazil;

[6]Laboratório de Helmintologia Romero Lascasas Porto, Departamento de Patologia e Laboratórios, Faculdade de Ciências Médicas, Universidade do Estado do Rio de Janeiro, Brazil

[7]Laboratório de Esquistossomose Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz - FIOCRUZ, Belo Horizonte, MG- 30190-002, Brazil;

[8]Laboratório de Entomologia Médica, Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz - FIOCRUZ, Belo Horizonte, MG - 30190-002, Brazil;

[9]Sandler Center for Drug Discovery, California Institute for Quantitative Biosciences and the Department of Pathology, University of California, San Francisco, San Francisco, California

*These authors contributed equally to this work

§Corresponding author: Guilherme Oliveira. Centro de Pesquisas René Rachou, Centro de Excelência em Bioinformática, Fundação Oswaldo Cruz – FIOCRUZ. Avenida Augusto de Lima, 1715. Belo Horizonte, MG, 30190-002, Brazil. Tel: +55-31-3349-7781. Fax: +55-31-3295-3115. Email: oliveira@cpqrr.fiocruz.br

Email addresses:

LLF: luizafreiredeandrade@gmail.com
MMM: marinamm@cpqrr.fiocruz.br
JAG: julianaassis@cebio.org
FSC: fernandasc91@yahoo.com.br
LLS: larissa.silva@cebio.org
RHN: renataheisler@gmail.com
JRMS: machado@uerj.br
NA: araujon@cpqrr.fiocruz.br
RNP: rafanp@cpqrr.fiocruz.br
CC: conor.caffrey@ucsf.edu
GO: oliveira@cpqrr.fiocruz.br

Keywords: Schistosoma, Schistosomiasis, Kinase, MAPK pathway, RNAi, Gene suppression

Competing interests: The authors have declared that no competing interests exist.

ABSTRACT

Background: Protein kinases are good chemical and medical targets for drug development and an increasing number of ePK inhibitors have been approved. MAPKs are protein kinase family members that connect cell-surface receptors to regulatory targets within cells and influence a number of tissue-specific biological activities, as cell proliferation, differentiation, and survival. All these processes are essential for parasites with a complex life cycle as *Schistosoma*. However, the role of MAPK pathway for schistosome development remains unclear.

Methodology/Principal Findings: We used gene silencing by RNA interference to elucidate the functional role of five *S. mansoni* genes involved in MAPK signaling pathway. Mice were infected with the silenced schistosomulas and the development of adult worms was observed. The data demonstrate that SmJNK has an important role in transformation and survival of the parasites and that SmERK1/SmERK2 was related to egg production, as mice infected displayed significantly lower egg production and the recovered female worms had underdeveloped ovaries. Furthermore, it was showed that the c-fos transcription factor was overexpressed in parasites with low expression of SmERK1, SmJNK and SmCaMK2.

Conclusions: We conclude that MAPKs proteins, especially SmERK and SmJNK, are excellent targets for drug development as an inhibitor against these proteins will probably disrupt the life cycle of *Schistosoma* and prevent disease progression.

AUTHOR SUMMARY

To date, the lack of understanding functions of the vast majority of schistosomes and other parasitic helminths genes represent a significant barrier for investigators. In this study, we focus on the characterization of MAPKs roles in schistosomes, proteins that influence a number of biological activities, as cell proliferation, differentiation, and a variety of essential processes for cells survival in diverse organisms. In the present study, we used RNA interference (RNAi), which is a method employed for manipulating gene-specific expression in human blood flukes, to elucidate the functional role of five *Schistosoma mansoni* genes involved in MAPK signaling pathway. Additionally, mice infection with knockdown parasites allowed us to assess whether worms expressing lower levels of MAPK proteins would be successfully able to develop and infect its host. Results showed that SmJNK has a central role in parasite transformation and survival. In addition, SmERK1/SmERK2 seems to be involved in ovary maturation and/or oviposition. We also demonstrated a negative feedback in the regulation of a downstream transcription factor, called c-fos, showing its involvement in the MAPK pathway. These results indicate the potential of using MAPKs proteins as targets for drug development against schistosomiasis and improve our knowledge on Schistosoma pathways networks and characterization of its proteins function.

INTRODUCTION

Schistosomes are parasitic flatworms (Phylum Platyhelmithes) that can survive for years or decades in a potentially hostile environment, inured to host immune-mediated attack [1,2]. Besides the strategies to inhibit or modulate host immune responses, the maintenance of homeostasis and complex cellular adaptations in *Schistosoma*, parasites integrate specific extracellular signals to generate an appropriate cellular response [3]. Thus, signal transduction has essential functions in the cell control and involves non-linearly integrated networks that interact mostly by switching the activity status of proteins.

The mitogen-activated protein kinase (MAP kinase/MAPK) signaling pathway is activated by environmental stimuli in response to a variety of extracellular growth factor-receptor interactions at the cell surface and leads to immediate transcriptional activation of specific proteins [4]. For example, in mammals, activated ERK/JNK MAPKs can translocate into the nucleus and induce phosphorylation of specific transcription factors as ELK-1 [5]. ELK-1, on the other hand forms a complex with other transcription factor, SRF (serum response factor), increasing its ability to activate transcription. The ELK-1/SRF complex is able to bind the promoter of c-fos gene initiating its transcription [6]. Thereby, MAPKs influence a number of tissue-specific biological activities, as cell proliferation, survival, and differentiation through activation of other protein kinases, metabolic enzymes, or by phosphorylation of transcription factors and components of the cytoskeleton [7].

Recently, our group showed by *in silico* analyses that the MAPK signaling components are well conserved in the three main *Schistosoma* species infecting humans (*S. mansoni*, *S. japonicum*, and *S. haematobium*) and mammals, including representatives of the MAPK subfamilies ERK (extracellular signal-regulated kinase), p38, JNK (c-Jun N-terminal kinase) and nmo (nemo MAPK) [8]. However, the role of MAPK pathway in schistosome's development and survival remain to be elucidated.

ERK MAPKs are required for multiple developmental events in *C. elegans*, including induction of vulval, uterine, spicule cell fates, and promotion of germline meiosis [9]. Vicogne and colleagues (2004) [10] showed that the human epidermal growth factor (EGF) is able to activate Ras/ERK pathway in *S. mansoni* inducing meiosis initiation in oocytes. Moreover, oviposition is directly related to the pathogenesis of schistosomiasis. Females can release, in average, 300 highly immunoreactive eggs a day. About 50% of laid eggs leave the host body and, the remaining, gets trapped in host tissues. The retained eggs release proteolytic enzymes that typically incite eosinophilic and granulomatous inflammatory reactions, which are progressively followed by fibrotic deposits resulting in organ dysfunction that can lead to patient's death [11,12]. All these observations led to our hypothesis that ERK/MAPK pathway could be involved in *Schistosoma* reproduction and could be potentially good therapeutic target against schistosomiasis.

On the other hand, JNK proteins have evolutionary conserved roles in the control of cellular responses to stress stimuli induced by a range of intrinsic and environmental insults (e.g. UV irradiation, DNA damage, heat, bacterial antigens, and inflammatory cytokines) [13]. Our recent study showed that only one member of MAPK JNK subfamily is encoded by the *S. mansoni* genome, in contrast with five genes expressed in *Caenorhabditis elegans* and three genes in human [8]. This evolutionary context suggests that the JNK protein of *S. mansoni* could be a potential target for drug development, since drug's effectiveness can be increased when a single-copy gene is targeted [14].

In this context, here we aimed at elucidation of the function of ERK/JNK MAPK signaling pathways in the parasite *S. mansoni* using RNA interference (RNAi). In summary, our findings show that suppression of the SmJNK gene expression causes damages in adult worm tegument, resulting in death of parasites. Furthermore, the knockdown of SmERK gene expression causes low egg production by female worms recovered from murine host, which could be explained by the under developed ovary and immature oocytes observed in the phenotype of those female worms, suggesting the involvement of SmERK in parasite reproduction.

METHODS

Parasites

LE strains of *Schistosoma mansoni* was maintained at Centro de Pesquisas René Rachou – FIOCRUZ using *Biomphalaria glabrata* as intermediate snail host. Schistosomula were obtained by mechanical transformation of cercariae according to Howells *et al* (1974) [15] and cultured in MEM medium (*Minimum Essential Medium Eagle*) supplemented with 20 mM Hepes, 2 mM glutamate, $1x10^{-6}$ M serotonin, $5x10^{-7}$ M hypoxanthine, $2x10^{-7}$ M hydrocortisone, 0.5% MEM vitamin solution 100X, antibiotics (100 U/ml penicillin and 100 µg/ml streptomycin), and 2% fetal bovine serum (FBS).

ERK/JNK Phylogenetic Analysis

In order to establish the evolutionary relationships among ERK and JNK proteins, homologs from *S. mansoni* (NCBI TaxID: 6183), *S. haematobium* (NCBI TaxID: 6185), *S. japonicum* (NCBI TaxID: 6182), *Caenorhabditis elegans* (NCBI taxID: 6239), *Drosophila melanogaster* (NCBI TaxID: 7227), and *Homo sapiens* (NCBI taxID: 9606) were selected to perform phylogenetic analysis. Amino acid sequences corresponding to the conserved catalytic domain (PF00069) were aligned using MAFFT 7 with iterative refinement by the G-INS-i strategy [16] (Figure S1). The multiple sequence alignment comprising 34 sequences with 300 sites was manually refined using Jalview [17] and further used in phylogenetic analysis. To reconstruct the phylogenetic tree we used MrBayes (version 3.2.1), which performs Bayesian inference using a variant of Markov chain Monte Carlo [18]. MCMC analyses were run as four chains (one cold and three heated chains) for 10,000,000 generations and sampled every 100 generations. 25% 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.

Selected genes and primer design

The *S. mansoni* sequences were downloaded from *SchistoDB*, version 2.0 [19]. We selected five genes from MAPK signaling pathway to perform the inhibition experiments: SmCaMK2 (Smp_011660.2), SmJNK (Smp_172240), SmERK1 (Smp_142050), SmERK2 (Smp_047900), and SmRas (Smp_179910). In addition to the evaluation of transcript levels of those five genes, SmSRF (Smp_097730) and SmC-Fos (Smp_124600) transcription factors were also included in the analysis, in order to evaluate downstream interactions.

Primers were designed using the Primer 3 software (http://frodo.wi.mit.edu/) employing a 150-200 bp target product size for qPCR and 500-600 bp for templates of double stranded-RNA (dsRNA). A T7 promoter-tag was added at the 5’ end of all PCR primers designed for dsRNA template amplification (Table 1). A fragment of ~500 bp of GFP (from pCRII-GFP) was used as nonspecific control in the knockdown experiments. qPCR primers were designed to anneal outside the targeted region of dsRNAs. The *S. mansoni* Citocrome Oxidase – COX (GenBank AF216698) was used as endogenous normalization control in all qPCR tested samples. SmCaMK2 has two predicted alternative splicing products (Smp_011660.2 and Smp_011660.3). The region selected to design the dsRNA and the region selected to measure the transcription level by q-PCR are, respectively, a 100% similar in the three isoforms.

dsRNA synthesis and exposure

Following amplification, PCR products were separated on 1% agarose gels and purified using QIAquick Gel Extraction Kit (QIAGEN), following the manufacturer's protocol. dsRNAs targeting specific *S. mansoni* genes were synthesized using PCR products of approximately 500bp in size amplified from schistosomula cDNA and the T7 RiboMAX Express RNAi Kit (Promega) following the manufacturer's instructions. Briefly, dsRNAs synthesis reactions were allowed to incubate for 16 hr at 37°C prior to DNAse treatment. The expected sizes of amplicons were analyzed by electrophoresis in 1% agarose gels and identity was confirmed by DNA sequencing.

2,000 schistosomula were cultivated in 24-well polystyrene plates containing 2 mL supplemented MEM medium (plus 1% FBS, 100 U/ml penicillin and 100 µg/ml streptomycin). For each treatment, 100 nM of dsRNA were added in duplicate in three biological replicates. Incubations were continued for 2, 4 or 7 days at 37°C under 5% CO_2.

Gene expression analyses (qPCR)

For qPCR experiments, total RNA was extracted using RNeasy Mini Kit (Qiagen). 100 ng of RNA was used to synthesize cDNA with Superscript III cDNA Synthesis kit (Life Technologies). The residual DNA remaining was removed by DNase digestion using a Turbo DNA-free kit (Ambion, Life Technologies). Each cDNA sample was tested in three technical replicates per plate in a minimum of 3 biological replicates. Experiments were carried out in a 7500 Real Time PCR System (Life Technologies) using the Power SYBR Green Master mix (Life Technologies). Reactions were carried out in a final volume of 25 µl in 96 well plates. *S. mansoni* cytochrome C oxidase I (GenBank AF216698) was used as the sample normalizing transcript and GFP cDNA was used as endogenous control. Two internal controls assessing both possible genomic DNA contaminations (no reverse transcriptase) and purity of the reagents (no cDNA) was included. The results were analyzed using the ΔΔCt method [20]. ΔΔCt values were normalized to controls and expressed as percent folds of difference compared to unspecific GFP dsRNA followed by statistical analysis using the Mann-Whitney *U*-test ($p<0.05$).

In vivo experiments - Adult worms and eggs recovery

SWISS Webster mice were infected with 300 schistosomula 2 days after dsRNA treatment (3 independent experiments, 6 animals per group). After 37 days, mice were perfused according to Pellegrino and Siqueira (1956) [21] and the adult worms were recovered and counted. Livers from infected animals were weighed and counted after digestion with 10% KOH. The significance of the results was analyzed using Mann-Whitney test (Wilcoxon-Sum of Ranks, $p < 0.05$, N = 3).

Adult worms samples recovered after perfusion of mice infected with silenced schistosomula were analyzed by confocal microscopy in search of phenotypic changes. The parasites were fixed in AFA (2% acetic acid, 10% formaldehyde and 48% ethanol) and stored at room temperature for further use. Then the worms were stained with 2.5% hydrochloric carmine, dehydrated by passage through 70% ethanol, 90% and 100%, clarified with methyl salicylate and Canada balsam (1:2) and individually mounted on glass slides.

Morphometric analyzes were performed on male and female worms using computer images (Image Pro Plus - Media Cybernetics, USA) captured by a Sony camera coupled (640 x 480 pixels, RGB) in a light microscope (Olympus BX50). The following parameters were determined: number and area of testicular lobes, presence of tubercles, area of ovary, the presence of eggs and vitelline glands, and integrity of tegument. The significance of results was performed using Mann-Whitney test (Wilcoxon-Sum of Ranks, $p < 0.05$).

Confocal images (LSM-410, Zeiss) using 488 nm laser He / Ne and filter LP 585 in the reflected mode were used to study the reproductive system of males (testicular lobes) and female (ovary, uterus, ootype, and vitelline glands) and the integrity of tegument and shape of oral and ventral suckers of males and females.

The experiments with animals were carried out in accordance with the Ethics Commission for Animal Use (CEUA) of Fundação Oswaldo Cruz under the number P49/12-5.

RESULTS

ERK/JNK MAPKs conservation

To characterize the evolutionary relationships between ERK and JNK proteins encoded by parasites and free-living organisms, we selected homologs from six species including three Platyhelminthes, one Nematode, one Arthropod, and one Chordate - *S. mansoni*, *S. haematobium*, *S. japonicum*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*, respectively - to perform phylogenetic analyses. This taxon sampling covers important evolutionary innovations in processes which kinases are directly related as response to environmental stimuli, reproduction and development. As showed by Figure 1, gene duplication followed by divergence was probably the main evolutionary mechanism driving the evolution of ERK and JNK subfamily members. The tree topology shows two well-supported clades grouping ERK and JNK proteins, thus revealing that the catalytic domain (PF00069) is sufficiently divergent to enable the discrimination of these two protein subfamilies. The number of homologs in Schistosomes and other metazoans varies and the presence of sequence variants may implicate in structural and/or functional specialization. In most cases, when homologs were identified in the three *Schistosoma* species, the relationships among them reflect the current knowledge regarding the origin and evolution of the *Schistosoma* lineage. Altogether, those findings demonstrate that ERK and JNK proteins are evolutionarily conserved in metazoan species transducing signals from cell surface to nucleus.

Knockdown of MAPK members by RNAi

To characterize MAPK pathway members (SmERK1, SmERK2, SmJNK, SmCaMK2 and SmRas) we used synthetic double-strand RNAs (dsRNA) to expose schistosomula culture. To assess whether the presence of any dsRNA itself would induce some effect in Schistosoma, double-stranded GFP (dsGFP) was used as unspecific control for dsRNA treatments. A second control group was not exposed to dsRNA (negative control). The RNAi effects were analyzed by qPCR. As results we observed that all genes targeted are sensitive to RNAi experiments and were substantially suppressed after two, four or seven days (Figure 2). Transcript levels were reduced by up to 92% = 0,08 +/- 0,0079 (for SmERK1 after two days) to 42% = 0,58 +/- 0,04 (for SmRas after four days). Although SmERK1, SmERK2 and SmJNK transcripts were more efficiently reduced by day two after treatment, SmCaMK2 and SmRas were best suppressed on the fourth day post dsRNA exposure (Figure 2). Additionally, decreased transcript levels of SmERK2 were observed in parasites two days after SmERK1 dsRNA exposure (56% of inhibition: transcription levels relative to controls = 0,44 +/- 0,006), this could be due to the high sequence similarity between ERK1/2.

SmERK and SmJNK knockdown affects parasite viability

To investigate whether RNAi-mediated gene silencing of SmCaMK2, SmJNK and SmERK affects parasite viability when exposed to host immune system, groups of six mice were infected with schistosomula treated with dsRNA for two days. After 37 days, adult worms and eggs were recovered from the liver of infected mice. Due to the lack of treatment response, SmRas was not included in the in vivo test.

The knockdown of SmJNK expression in schistosomula resulted in death of 56% of parasites relative to mice infected with parasites from unspecific control (Figure 3). No significant changes in the number of ex-vivo adult worms were observed in mice infected with SmCaMK2 or SmERK-knockdown schistosomula.

Although the knockdown of SmERK does not seem to affect parasite transformation and growth, the decreased expression levels of this gene results in reduced parasite egg production (44% of reduction relative to unspecific control) (Figure 4). In addition, the number of eggs present in the liver of mice infected with SmJNK-knockdown schistosomula was also reduced

(59% of reduction relative to unspecific control), which could be a consequence of the decreased viability of worms.

After 37 days of dsRNA treatment, adult worm were analyzed by q-PCR to assess transcript levels of SmCaMK2, SmJNK and SmERK. No remaining gene suppression was observed.

The knockdown of SmERK and SmJNK leads to morphological changes in ex-vivo *S. mansoni* adult worms

In order to investigate whether morphological changes could be associated with viability and low egg production after SmJNK and SmERK knockdown, we analyzed the morphological structures of ex-vivo adult worms by confocal microscopy. We observed that the knockdown of SmJNK causes damages in male adult worm tegument (Figure 5). In those worms the number of tubercles was visibly reduced (Figure 5D) and the tegument of oral and ventral sucker areas were deformed (Figure 5B). Moreover, unusual dilations were observed in the tegument and digestive tract of male and female (Figure 5D/F and Figure 6F). In addition, females presented undifferentiated oocytes (Figure 6D).

The knockdown of SmERK does not cause changes in male worms (Figure 7) as they presented normal tegument, testicular lobes, and functional seminal vesicle with spermatozoids. However, the females presented small ovaries (Figure 8G) containing immature oocytes (Figure 8F). Some females also presented higher number of oocytes in the uterus (Figure 8B) and unfertilized eggs in the ootype (Figure 8D).

No phenotypic change was visible on adult worms recovered from mice infected with SmCaMK2-knockdown schistosomula. Those worms presented normal functional structures and morphology (Figure S2).

Effect of knockdown in the transcription regulation of MAPK target genes

To study the conservation of MAPK pathway in *S. mansoni* in comparison with other metazoans, we evaluated the transcription level of SRF transcription factor and c-fos gene after SmJNK, SmCaMK2, SmRas, and SmERK1 knockdown. Figure 9 shows that the knockdown of SmCaMK2, SmJNK and SmERK-1 by RNAi causes the over-expression of Smc-fos gene (1.62 +/- 0.28; 1.65 +/- 0.14; 1.47 +/- 0.06, respectively). In addition, as the MAPKs (SmJNK and SmERK-1) RNA expression increases, after seven days, the Smc-fos RNA level decreases (1.10 +/- 0.2 and 0.89 +/- 0.18, respectively) (Figure 9) receding progressively to the control level. SmSRF gene expression, in most cases, does not exhibit variation. A minor alteration in Smc-fos transcript level was observed after SmRAS inhibition (Figure 9B).

DISCUSSION

MAPKs connect cell-surface receptors to regulatory targets within cells to coordinate gene expression. Members of this family regulate essential cellular processes and are conserved in all eukaryotes [4]. It would be expected that this enzymes would also have important and essential functions in *Schistosoma* parasites. However, until the present study, little was known about MAPK pathway regulation in schistosomes.

Our findings show suppression in transcript levels of SmERK and SmJNK, as other members of the MAPK pathway, reached levels of up to 92% indicating that the RNAi approach is an efficient method to elucidate potential functions of *S. mansoni* proteins. The knockdown efficiency varied from gene to gene and also according to the exposure time for each target gene. Other authors [22,23] also reported variable profiles of transcript knockdown with success relying on selected target gene or specific dsRNA sequence. These results can be explained by the fact that some genes are expressed in cells/tissues that are inaccessible to dsRNA by the employed delivery method or due to accessibility of secondary structure of some mRNA to the RISC complex [24].

Parasites treated with SmRas dsRNA did not respond well to treatment in the various time points tested. As mentioned above, this transcript could have a rapid turnover or could be expressed in cells or tissues inaccessible to dsRNA by the soaking method. SmRas gene is highly similar to *C. elegans* let-60 protein (a Ras superfamily gene which the only known function is to activate the RTK/Ras/MAPK pathway) [9]. However, in *C. elegans* more than 56 members of Ras superfamily was identified.

The association of *in vitro* and *in vivo* studies allowed the elucidation of important roles of MAPK genes in *S. mansoni.* Moreover, it was possible to verify the knockdown effect in a system comparable to natural *Schistosoma* infection.

After 37 days, the mRNA levels of targeted genes in ex vivo worms were no longer suppressed. In the meantime, phenotypic alterations in those worms were visible, which could be due to the effect of gene knockdown at the beginning of development while the parasite was settling in the host. In addition, the recovered parasites could be individuals that were partially refractory to dsRNA treatment and, also, that the observed morphological changes were a mild phenotype, this variation in response to dsRNA exposure within populations has been previously described in *S. mansoni* (15).

Despite well conserved elements of JNK-MAPK pathway in *Schistosoma*, the number of JNK subfamily members is not evolutionarily conserved in comparison to other metazoans. Among the three *Schistosoma* species analyzed here, only one JNK protein seems to regulate cellular mechanisms, in contrast to five subfamily members in *C. elegans* and three in human. The existence of a unique JNK subfamily member in *Schistosoma* could be translated into an enormous potential for downstream signal specificity.

JNK signaling misregulation has been implicated in neurodegenerative diseases, diabetes, and cancer [25,26,27,28]. In *C. elegans,* JNK pathway could also be activated by Ca^{2+}/calmodulin-dependent protein kinase (CaMK, unc-43) in a cell specific signaling pathway [9,29]. As JNK, only one CaMK2 protein was found in the predicted proteome of *S. mansoni* and *S. haematobium.* Additionally, SmCaMK2 was recently identified as putative target for drug development after *in silico* comparative chemogenomics approach [30].

In the present study, we showed that the knockdown of SmJNK disturbs infective capability of worms once significantly low number of adult worms was recovered from mice. In addition, the recovered worms presented morphological changes in the tegument and digestive tract. These results suggest that the SmJNK proteins are essential for parasite transformation and survival. This fact is reinforced by previous knowledge of JNK signaling pathway influencing metabolism, growth, regeneration, and stress tolerance in *Drosophila* lifespan regulation [31]. Moreover, in flies, the JNK signaling pathway is also involved in midgut epithelial homeostasis and may be important in other contexts, such as oxidative stress for protection against gut infections [32]. Strong inhibition of JNK signaling activity in *Drosophila* shortens lifespan due to complete inhibition of intestinal steam cells proliferation [31].

Although SmCaMK2 was previously predicted as a potential protein target for drug development against schistosomiasis [30], our findings show that the SmCaMK2-knockdown does not produce a significant phenotypic alteration in *in vitro* schistosomula nor in *ex vivo* adult worms. Once CaMK2 (UNC-53) can activate JNK signaling pathway in *C. elegans*, similar results of SmJNK-knockdown would be expected. Few hypotheses could explain this outcome: i) either SmCaMK2 is not related to JNK signaling pathway or ii) SmCaMK2 is not the unique activator of JNK signaling pathway in *Schistosoma.* Additionally, we do not exclude the possibility that SmCaMK2 would regulate the JNK pathway only in particular cell type(s) or that the SmCaMK2 protein turnover is faster than its counterpart SmJNK and it would be necessary a gene knockout to inhibit signaling activity.

ERK signaling pathway plays essential role during oocyte maturation and egg activation in many animals [33,34]. As observed in this study, it seems to play the same role in *S. mansoni*. The low number of eggs in the liver and the distinct phenotypes observed in adult female worms recovered from mice infected with SmERK-knockdown schistosomula suggest that ERK pathway is related to reproductive organs development and/or breeding of parasite. In *Xenopus laevis*, the ERK protein is involved in the coordination of oocyte maturation [35]. In *C. elegans,* the knockout of ERK, affects the development of the vulva (necessary for egg-laying) and oocyte resulting in no egg production [36], while the inactivation of ERK-MAPK signaling pathway in mice can be associated with embryonic death caused by abnormal placental development [37]. Sandler and colleagues also demonstrated that ERK-MAPK is involved in starfish egg apoptosis [38]. All these data corroborate the results presented in this work and suggest a functional conservation of the ERK pathway in worms, flies, starfishes, frogs, mice, and could be also true for all eumetazoans.

In mammals, it has been previously known that, down-regulation of c-fos gene after inhibition of members of MAPK pathway [6]. In order to determine whether the MAPK pathway induces c-fos expression in *S. mansoni*, c-fos and SRF transcript levels were evaluated after transcript knockdown of SmRas, SmERK-1, SmJNK, and SmCaMK2. According to the literature, there are at least two main modes of gene regulation in eukaryotes. In the first one, the genes exhibit a variable rate of transcription and, for signaling transduction, proteins need to bind at a specific gene promoter to initiate its transcription. In the second one, the gene has a fixed transcription rate, presenting a stable conformation in active (On) or inactive (Off) modes, thus they are directly connected to fast signaling, such as most of signaling pathways in which PKs (protein kinases) are involved, including the MAPK pathway [5]. In this context, c-fos has a variable level of transcription that is dependent on elk-1/SRF binding, which in turn, has a less variable transcription rate [39].

Our results showed that the transcript level of SmSRF remained constant. On the other hand, the transcription of Smc-fos is highly activated after MAPK members' inhibition. This finding indicates that there is a relation between the activation of MAPK and Smc-fos transcription as in mammals. In this regard, a positive regulation was expected, in which the non-activation of SmERK-1 and SmJNK would prevent SRF activation which, in turn, would not be able to bind the c-fos promoter region. Thus, a negative correlation occurs, where low levels of SmERK-1, SmCaMK2, and SmJNK transcripts induce c-fos transcription.

Once the MAPK pathway is evolutionarily conserved, it could be hypothesized the signaling pathways that activate the expression of *S. mansoni* c-fos gene are probably similar to those of *C. elegans*, wherein elk-1 (LIN-31 in *C. elegans*) and SRF (LIN-1 in *C. elegans*) form a complex when MAPKs are not phosphorylated and this complex activates c-fos which inhibits the vulva development. When MAPK is phosphorylated, elk-1/SRF complex dissociates and LIN-31 promotes the vulva development in a signaling pathway activated by the epidermal growth factor (EGF) [40]. Moreover, it was recently reported that pathways involved in activation of c-fos gene expression may be activated by calcium influx through CaMK signaling pathway [30]. In this case, c-fos expression is induced by phosphorylation of CaMK2 which, in turn, phosphorylates SRF, without a direct relationship with ERK or JNK proteins [41].

Altogether, it is possible that high levels of Smc-fos transcripts, verified after the knockdown of SmERK, SmJNK, and SmCaMK2, could be related to the non-activation of the MAPK signaling pathway which led to induction of the elk-1/SRF complex formation (Figure 10B-C). In this case, SmJNK and SmCaMK2 may be involved in independent pathways or they simultaneously co-regulate the same gene in a particular cell type. Similarly, elk-1/SRF complex is target for different cascades and is involved in the regulation of c-fos. In the other hand, SmRas and SmERK would act in the same pathway, as occurs in *C. elegans*, mammals, and *Drosophila*, being directly involved in the *S. mansoni* egg development (Figure 10A).

ACKNOWLEDGMENTS

The authors acknowledge the use of the computing resources of the Center for Excellence in Bioinformatics, CEBio/FIOCRUZ, Brazil. The authors are grateful to Anderson Dominitini (Cebio), Adhemar Zerlotini (EMBRAPA), Eric Aguiar (CEBio) for bioinformatics technical support and to Mariana de Oliveira (CEBio) for help with illustrations in this work

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: LFA MMM GO CC. Carrie out the RNAi experiments: LFA MMM JAG FSC NA. Analyzed the data: LFA MMM. Microscopy analyses: LFA RRN JRMS RP. Contributed reagents/materials/analysis tools: GO CC. Designed and performed the phylogenetic studies: LLS. Wrote the manuscript: LFA MMM LLS GO. Reviewed and revised the manuscript: LFA GO CC. Coordination of this study: GO. All authors read and approved the final manuscript.

FINANTIAL DISCLOSURE: This work was funded by the National Institutes of Health - NIH/Fogarty International Center (2D43TW007012 to GO) and the National Council for Research and Development - CNPq (CNPq Research Fellowship 306879/2009-3 to GO, INCT-DT 573839/2008-5 to GO.

REFERENCES

1. Allen JE and Maizels RM. (2011) Diversity and dialogue in immunity to helminths. Nature Reviews Immunology 11: 375-388
2. Pearce EJ and Sher A. (1987) Mechanisms of immune evasion in schistosomiasis. Contrib Microbiol Immunol 8: 219-232
3. Han ZG, Brindley PJ, Wang SY, Chen Z. (2009) Schistosoma genomics: new perspectives on schistosome biology and host-parasite interaction. Annu.Rev.Genomics Hum.Genet. 10: 211-240
4. Wilkinson MG and Millar JB. (2000) Control of the eukaryotic cell cycle by MAP kinase signaling pathways. FASEB J. 14: 2147-2157
5. Hazzalin CA and Mahadevan LC. (2002) MAPK-regulated transcription: A continuously variable gene switch? Nature Reviews Molecular Cell Biology 3: 30-40
6. Cavigelli M, Dolfi F, Claret FX, Karin M. (1995) Induction of C-Fos Expression Through Jnk-Mediated Tcf/Elk-1 Phosphorylation. Embo Journal 14: 5957-5964
7. Pearson G, Robinson F, Gibson TB, Xu BE, Karandikar M et al. (2001) Mitogen-activated protein (MAP) kinase pathways: Regulation and physiological functions. Endocrine Reviews 22: 153-183
8. Andrade LF, Nahum LA, Avelar LG, Silva LL, Zerlotini A et al. (2011) Eukaryotic protein kinases (ePKs) of the helminth parasite Schistosoma mansoni. BMC.Genomics 12: 215-
9. Sundaram MV. (2006) RTK/Ras/MAPK signaling. WormBook. 1-19
10. Vicogne J, Cailliau K, Tulasne D, Browaeys E, Yan YT et al. (2004) Conservation of epidermal growth factor receptor function in the human parasitic helminth Schistosoma mansoni. J.Biol.Chem. 279: 37407-37414
11. Warren KS, Jordan P, Cook JA. (1972) Passive Transfer of Immunity in Human Schistosomiasis Mansoni - Effect of Hyperimmune Anti-Schistosome Gamma-Globulin on Early Established Infections. Transactions of the Royal Society of Tropical Medicine and Hygiene 66: 65-&
12. Wyler DJ. (1983) Regulation of Fibroblast Functions by Products of Schistosomal Egg Granulomas - Potential Role in the Pathogenesis of Hepatic-Fibrosis. Ciba Foundation Symposia 99: 190-204
13. Borsello T and Forloni G. (2007) JNK signalling: A possible target to prevent neurodegeneration. Current Pharmaceutical Design 13: 1875-1886
14. Emes RD and Yang ZH. (2008) Duplicated Paralogous Genes Subject to Positive Selection in the Genome of Trypanosoma brucei. Plos One 3:

15. Howells RE, Ramalhopinto FJ, Gazzinelli G, Deoliveira CC, Figueiredo EA et al. (1974) Schistosoma-Mansoni - Mechanism of Cercarial Tail Loss and Its Significance to Host Penetration. Experimental Parasitology 36: 373-385

16. Katoh K, Asimenos G, Toh H. (2009) Multiple alignment of DNA sequences with MAFFT. Methods Mol.Biol. 537: 39-64

17. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191

18. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A et al. (2012) MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Systematic Biology 61: 539-542

19. Zerlotini A, Aguiar ERGR, Yu FD, Xu HY, Li YX et al. (2013) SchistoDB: an updated genome resource for the three key schistosomes of humans. Nucleic Acids Research 41: D728-D731

20. Livak KJ and Schmittgen TD. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25: 402-408

21. Pellegrino J and SIQUEIRA AF. (1956) [A perfusion technic for recovery of Schistosoma mansoni from experimentally infected guinea pigs.]. Rev.Bras.Malariol.Doencas.Trop. 8: 589-597

22. Mourao MM, Dinguirard N, Franco GR, Yoshino TP. (2009) Phenotypic screen of early-developing larvae of the blood fluke, schistosoma mansoni, using RNA interference. PLoS.Negl.Trop.Dis. 3: e502-

23. Stefanic S, Dvorak J, Horn M, Braschi S, Sojka D et al. (2010) RNA interference in Schistosoma mansoni schistosomula: selectivity, sensitivity and operation for larger-scale screening. PLoS.Negl.Trop.Dis. 4: e850-

24. Rao MK and Wilkinson MF. (2006) Tissue-specific and cell type-specific RNA interference in vivo. Nature Protocols 1: 1494-1501

25. Hotamisligil GS. (2010) Endoplasmic reticulum stress and atherosclerosis. Nature Medicine 16: 396-399

26. Karin M and Gallagher E. (2005) From JNK to pay dirt: Jun kinases, their biochemistry, physiology and clinical importance. Iubmb Life 57: 283-295

27. Sabio G and Davis RJ. (2010) cJun NH2-terminal kinase 1 (JNK1): roles in metabolic regulation of insulin resistance. Trends in Biochemical Sciences 35: 490-496

28. Weston CR and Davis RJ. (2002) The JNK signal transduction pathway. Current Opinion in Genetics & Development 12: 14-21

29. Sagasti A, Hisamoto N, Hyodo J, Tanaka-Hino M, Matsumoto K et al. (2001) The CaMKII UNC-43 activates the MAPKKK NSY-1 to execute a lateral signaling decision required for asymmetric olfactory neuron fates. Cell 105: 221-232

30. Caffrey CR, Rohwer A, Oellien F, Marhofer RJ, Braschi S et al. (2009) A comparative chemogenomics strategy to predict potential drug targets in the metazoan pathogen, Schistosoma mansoni. PLoS.One. 4: e4413-

31. Biteau B, Karpac J, Supoyo S, DeGennaro M, Lehmann R et al. (2010) Lifespan Extension by Preserving Proliferative Homeostasis in Drosophila. Plos Genetics 6:

32. Ha EM, Oh CT, Ryu JH, Bae YS, Kang SW et al. (2005) An antioxidant system required for host protection against gut infection in Drosophila. Developmental Cell 8: 125-132

33. Fan HY and Sun QY. (2004) Involvement of mitogen-activated protein kinase cascade during oocyte maturation and fertilization in mammals. Biology of Reproduction 70: 535-547

34. Sackton KL, Buehner NA, Wolfner MF. (2007) Modulation of MAPK activities during egg activation in Drosophila. Fly 1: 222-227

35. Nebreda AR and Ferby I. (2000) Regulation of the meiotic cell cycle in oocytes. Curr.Opin.Cell Biol. 12: 666-675

36. Hajnal A and Berset T. (2002) The C.elegans MAPK phosphatase LIP-1 is required for the G(2)/M meiotic arrest of developing oocytes. EMBO J. 21: 4317-4326

37. Giroux S, Tremblay M, Bernard D, Cadrin-Girard JF, Aubry S et al. (1999) Embryonic death of Mek1-deficient mice reveals a role for this kinase in angiogenesis in the labyrinthine region of the placenta. Current Biology 9: 369-372

38. Sadler KC, Yuce O, Hamaratoglu F, Verge V, Peaucellier G et al. (2004) MAP kinases regulate unfertilized egg apoptosis and fertilization suppresses death via Ca2+ signaling. Molecular Reproduction and Development 67: 366-383

39. Johnson RS, Spiegelman BM, Papaioannou V. (1992) Pleiotropic Effects of A Null Mutation in the C-Fos Protooncogene. Cell 71: 577-586

40. Tan PB, Lackner MR, Kim SK. (1998) MAP kinase signaling specificity mediated by the LIN-1 Ets/LIN31 WH transcription factor complex during C-elegans vulval induction. Cell 93: 569-580

41. Ely HA, Mellon PL, Coss D. (2011) GnRH Induces the c-Fos Gene via Phosphorylation of SRF by the Calcium/Calmodulin Kinase II Pathway. Molecular Endocrinology 25: 669-680

FIGURES

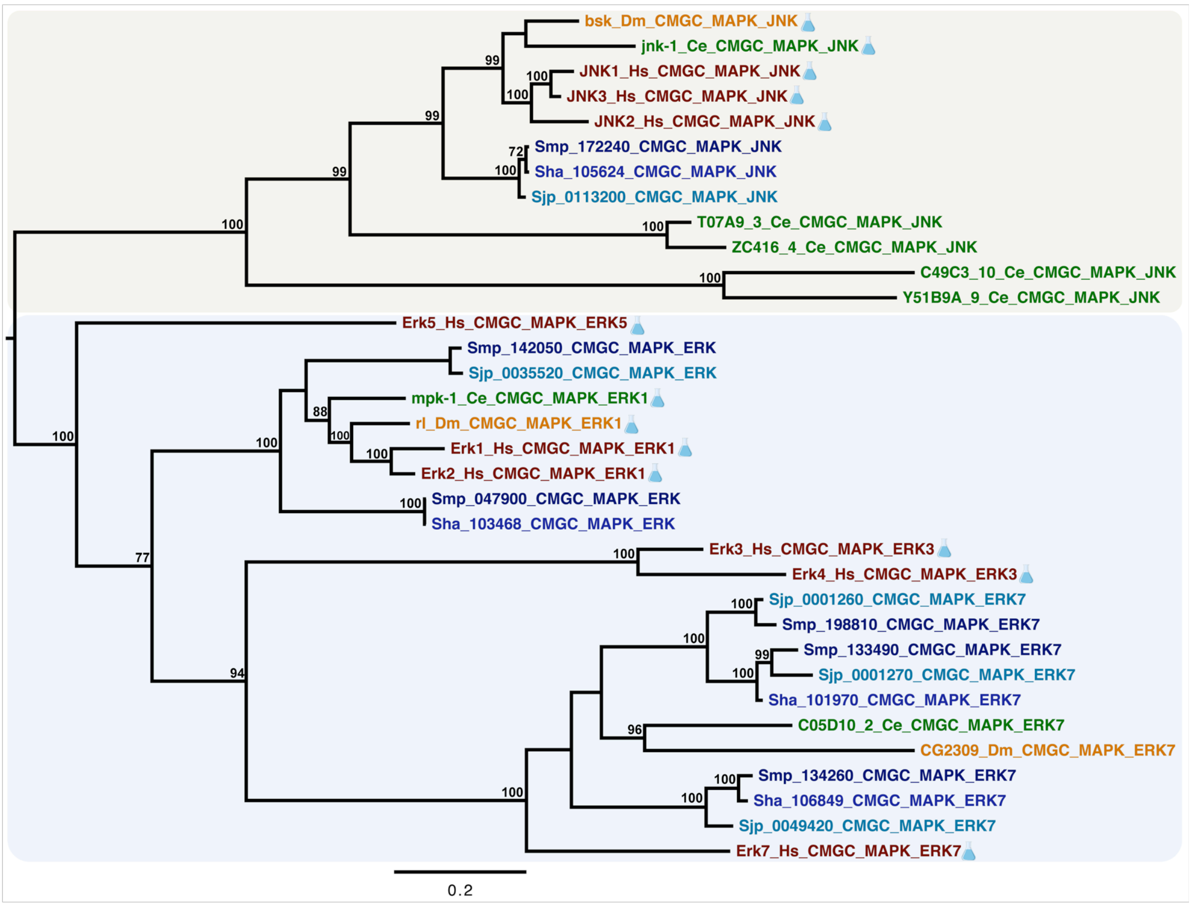


Figure 1 - Evolutionary relationships of ERK and JNK proteins. Evolutionary relationships of 34 ERK and JNK proteins encoded by parasites (*S. haematobium*, *S. japonicum*, and *S. mansoni*), and free-living organisms (*Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*) as inferred by Bayesian analysis. Experimentally characterized proteins are indicated by a symbol. Support values were computed by posterior probability. Two well-supported clades cluster ERK and JNK proteins. The analysis was performed with conserved amino acid sequences corresponding to the catalytic domain (PF00069). Mixed models were selected as implemented in MrBayes with 10 million generations sampled every 100 generations.

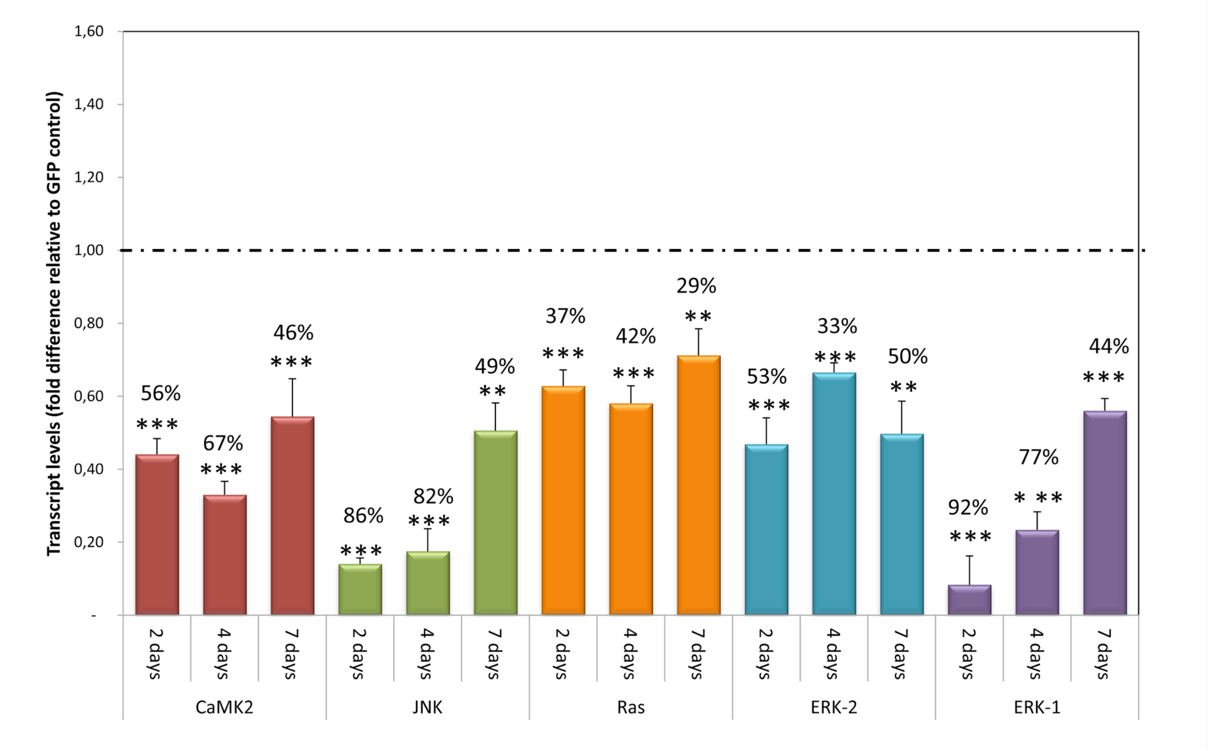


Figure 2 - Transcript levels of target genes after 2, 4, and 7 days after schistosomula dsRNA exposure. Bar graph indicating the relative steady-state transcript levels of SmCaMK2 (red), SmJNK (green), SmRas (orange), SmERK-2 (blue), and SmERK1(purple) genes after 2, 4, and 7 days after dsRNA exposure. For each dsRNA tested, data are represented as mean fold-differences (+/- SE) relative to GFP control (1.00 – dashed line). Transcript levels were determined by q-PCR and data analyzed using the ΔΔCt method followed by statistical analysis using the Mann-Whitney *U*-test. Significance levels (*) were set at *P*≤0.05. Data were generated from 3 independent experiments.

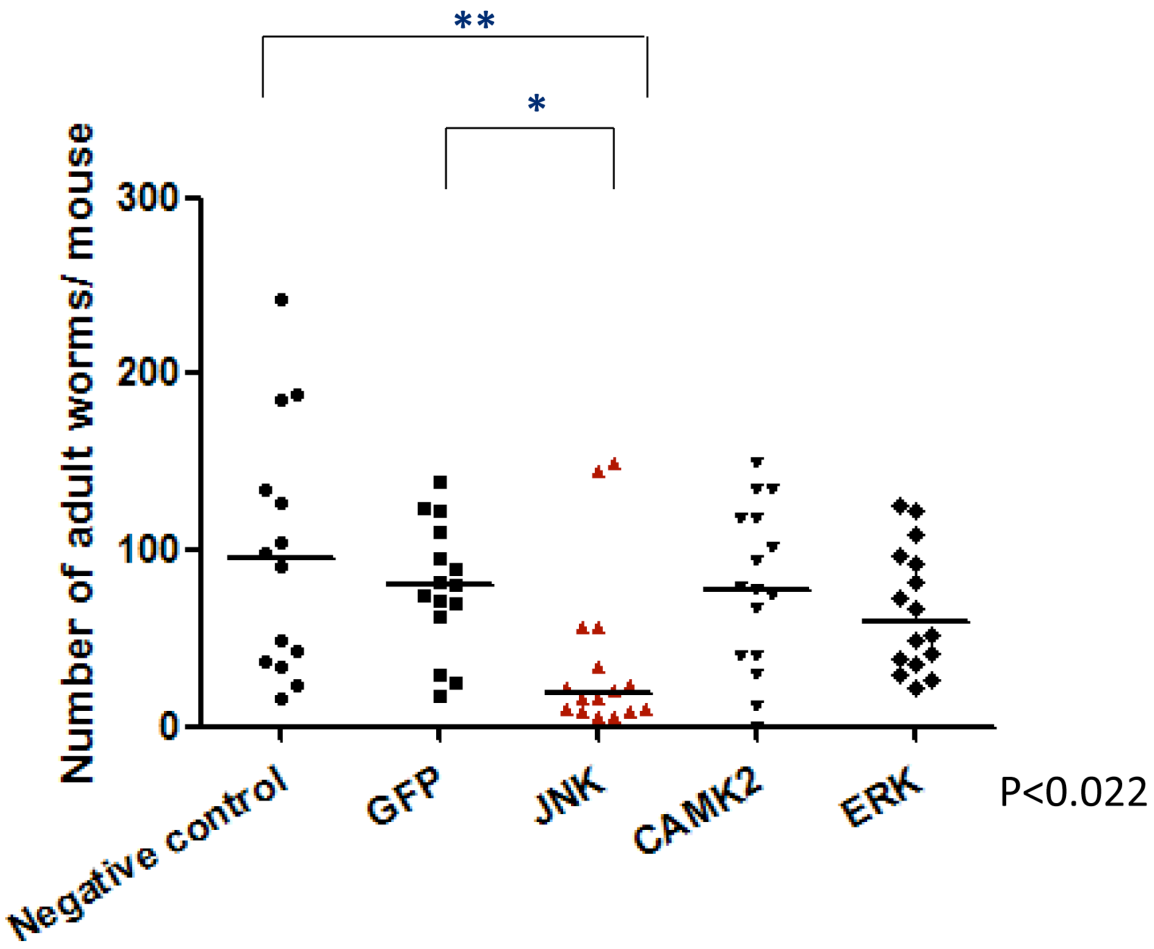


Figure 3 - Schistosome adult worm survival after 35 days knowdown-schistosomula infection. Schistosomula were treated with GFP, SmJNK, SmCaMK2, and SmERK1 dsRNAs and after two days were used to infect mice. Adult worm were recovered after 37 days from individual mouse represented by the dots. The lines indicated the median for each group. The (*) at the top represents SmJNK P values (*P*< 0,022) relative to negative control or GFP control dsRNA treatment. Data were generated from 3 independent experiments and all treatments were statistical analyzed using Mann-Whitney *U*-test within each experiment, *P*≤0.05.

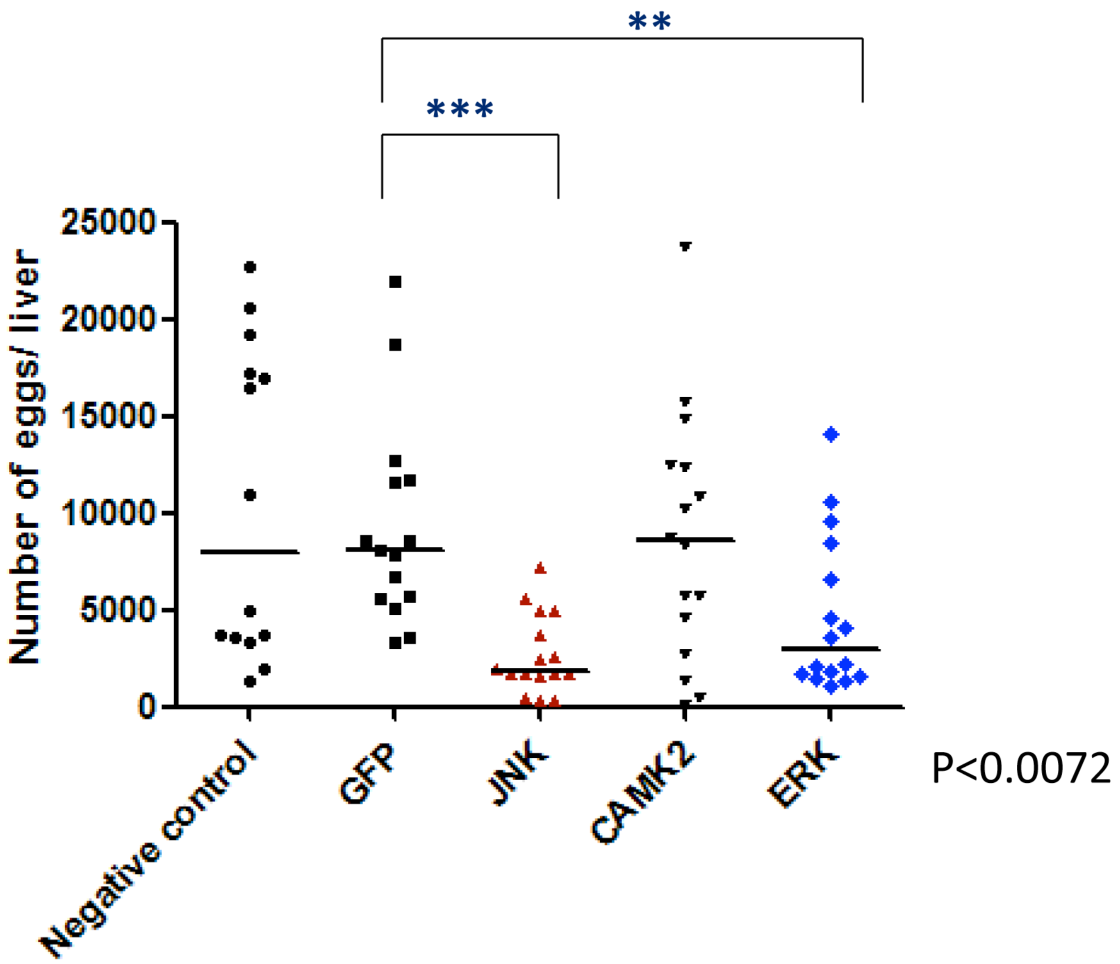


Figure 4 - Eggs recovered from infected mice with knockdown schistosomula. Schistosomula were treated with GFP, SmJNK, SmCaMK2, and SmERK1 dsRNAs and after two days were used to infect mice. Eggs of the liver were recovered after 37 days from individual mouse represented by the dots. The lines indicated the median for each group. The (*) at the top represents SmJNK and SmERK *P* values (*P*< 0,0072) relative to GFP control. Data were generated from 3 independent experiments and all treatments were statistical analyzed using Mann-Whitney *U*-test within each experiment, *P*≤0.05.

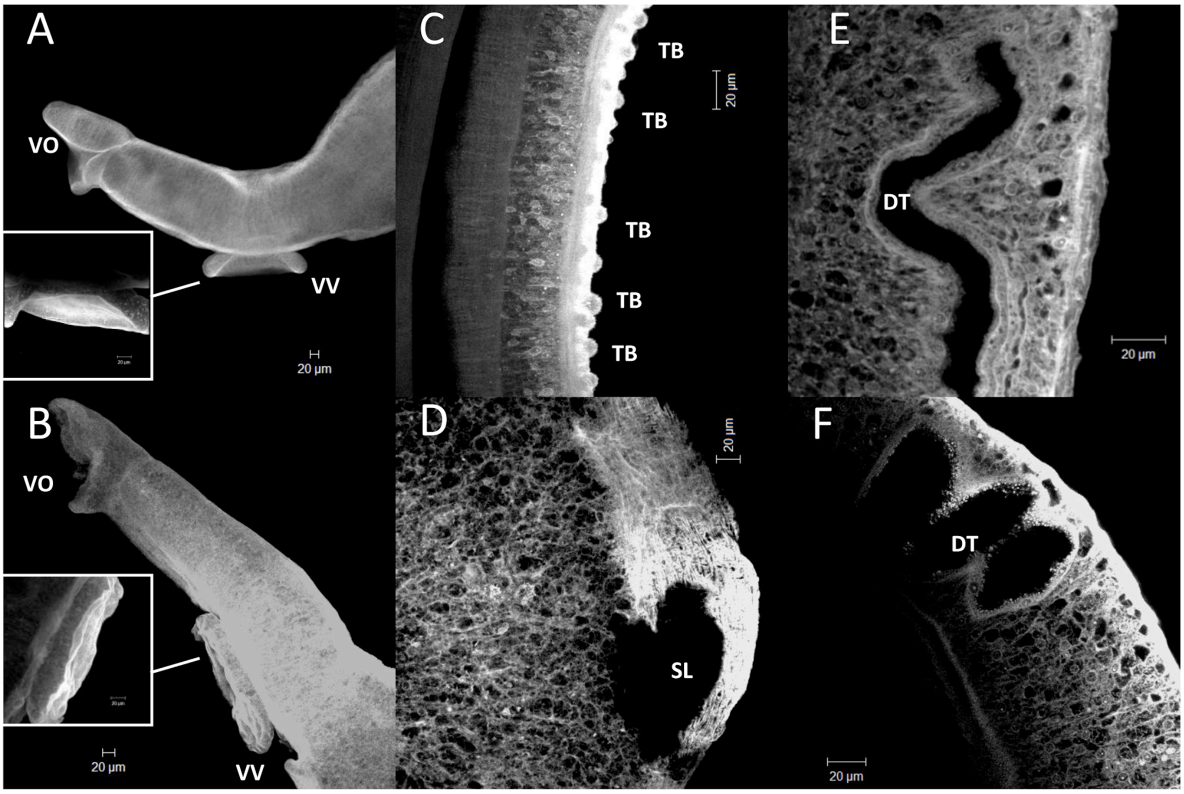


Figure 5 - Image of male adult worms recovered from mice infected with SmJNK-knockdown schistosomula. The 37-day-old adult worms were visualized by confocal microscopy. A, C and E show normal worms treated with GFP dsRNA, whereas B, D and F show morphological changes in worms treated with SmJNK dsRNA. The initials are VO: oral sucker; VV: ventral sucker; TB: tubercules; SL: subtegumentar lesion; DT: digestive tract. Scale bars: 20 µm.

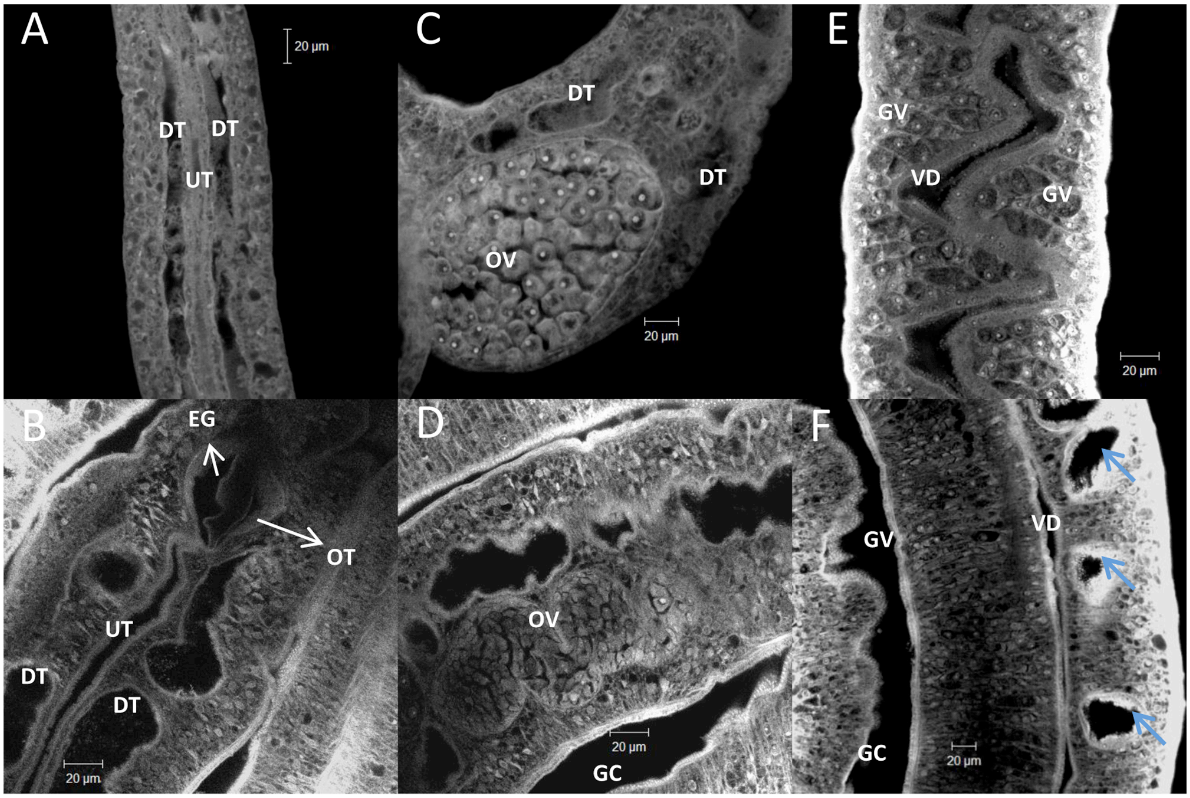


Figure 6 - Image of female adult worms recovered from mice infected with SmJNK-knockdown schistosomula. The 37-day-old adult worms were visualized by laser confocal microscopy. A, C and E show normal worms treated with GFP dsRNA, whereas B, D and F show morphological changes in worms treated with SmJNK dsRNA. The initials are DT: digestive tract; UT: uterus; OV: ovary; EG: egg; OT: ootype; GV: vitelline gland; GC: gynaecophoric canal; VD; vitelloduct. Scale bars: 20 µm.

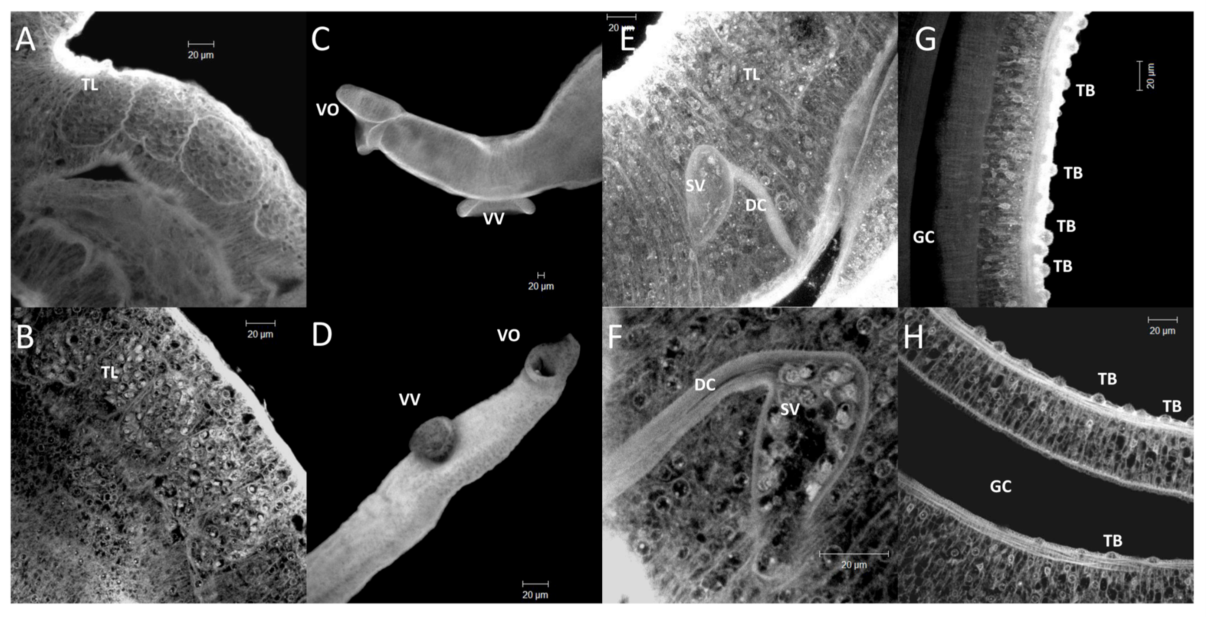


Figure 7 - Image of male adult worms recovered from mice infected with SmERK-knockdown schistosomula. The 37-day-old adult worms were visualized by confocal microscopy. A, C, E, and G show normal worms treated with GFP dsRNA, whereas B, D, F and H show worms treated with SmERK dsRNA. The initials are TL: testicular lobes; VO: oral sucker; VV: ventral sucker; SV: seminal vesicle; DC: duct for seminal pore; TB: tubercules; GC: gynaecophoric canal. Scale bars: 20 µm.

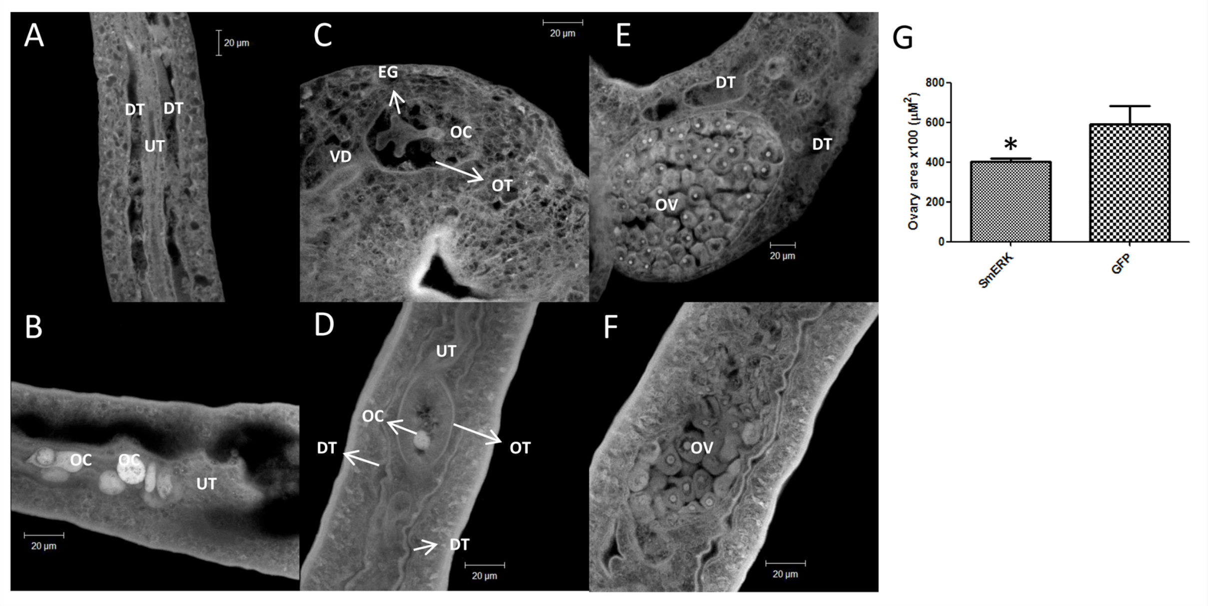


Figure 8 - Image of female adult worms recovered from mice infected with SmERK-knockdown schistosomula. The 37-day-old adult worms were visualized by laser confocal microscopy. A, C and E show normal worms treated with GFP dsRNA, whereas B, D and F show morphological changes in worms treated with SmERK dsRNA. The initials are DT: digestive tract; UT: uterus; VD; vitelloduct; OC: oocyte; EG: egg; OT: ootype; OV: ovary. Scale bars: 20 µm. The graph (G) shows the mean of females' ovary area (µM^2) of SmERK-knockdown and control. Statistical analyses were performed using Mann-Whitney *U*-test, $P \leq 0.05$; n= 5).

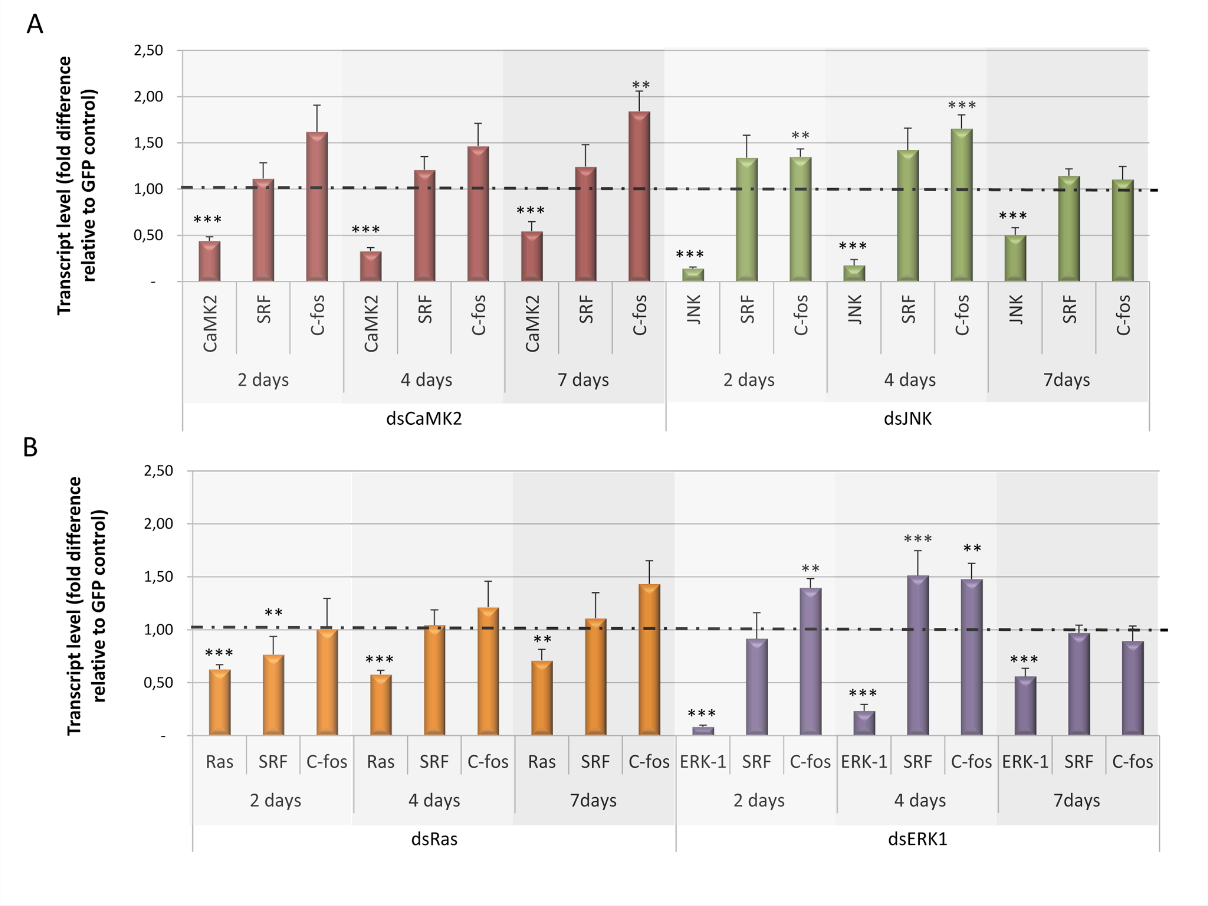


Figure 9 – Transcript levels of target dsRNA genes, SmSRF and Smc-fos after 2, 4, and 7 days after schistosomula dsRNA exposure. Bar graph indicating the relative steady-state transcript levels of (A) SmCaMK2 (red), SmJNK (green) and (B) SmRas (orange), SmERK1(purple) genes after 2, 4, and 7 days of dsRNA exposure. For each dsRNA tested, data are represented as mean fold-differences (+/- SE) relative to GFP control (1.00). Transcript levels were determined by q-PCR and data analyzed using the ΔΔCt method followed by statistical analysis using the Mann-Whitney *U*-test. Significance levels (*) were set at $P \leq 0.05$. Data were generated from 3 independent experiments.

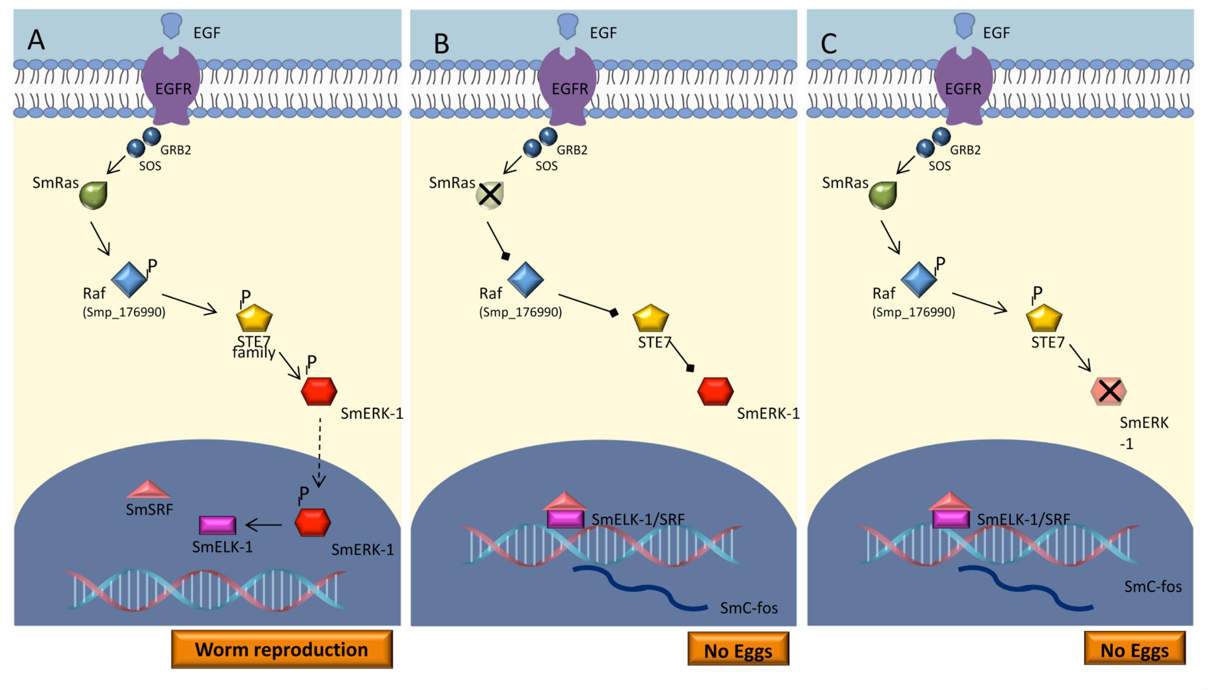


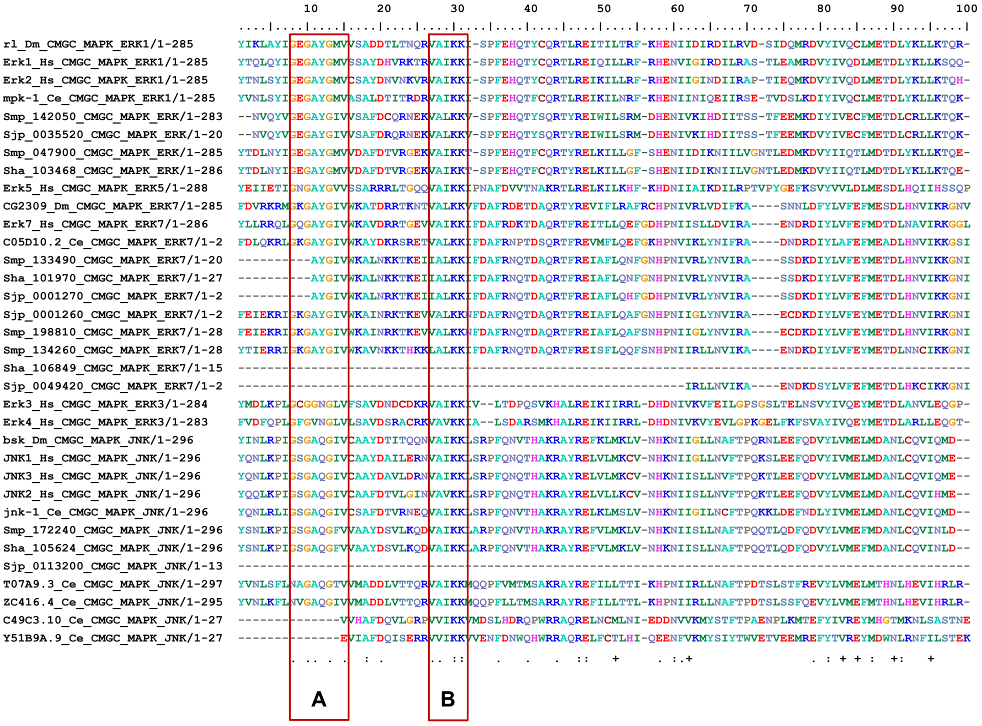
Figure 10 – *S. mansoni* hypothetical MAPK signaling pathway. In (A) EGF activates the Ras / ERK signaling pathway. EGFR transmits the signal to the intracellular environment through the activation of Ras and sequential phosphorylation of SmRAF (ePK of TKL group and raf family), SmSTE7 (ePK of STE group and STE7family) and SmERK (ePK of CMGC group, MAPK family and ERK1 / 2 subfamily). Activated ERK translocates to the nucleus and inhibits the formation of the elk1/SRF complex and, in this case, oviposition remains constant. In (B and C) SmRas or SmERK are silenced and the signal is not transmitted. elk1 forms a complex with SRF which binds to the c-Fos promoter and initiates its transcription that prevents the egg laying.

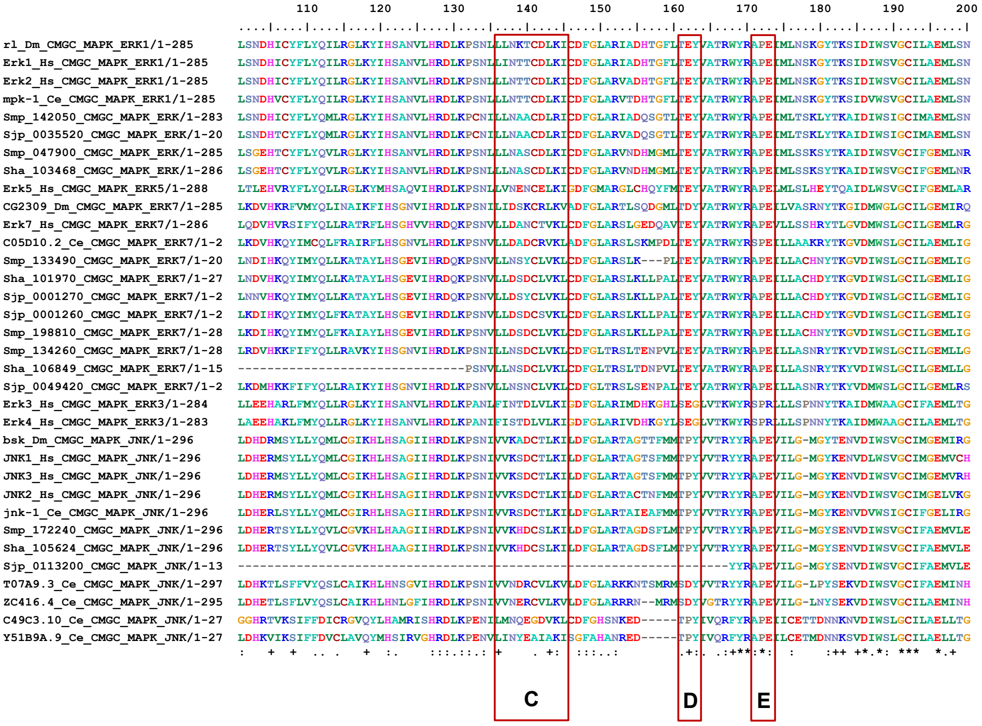
TABLES

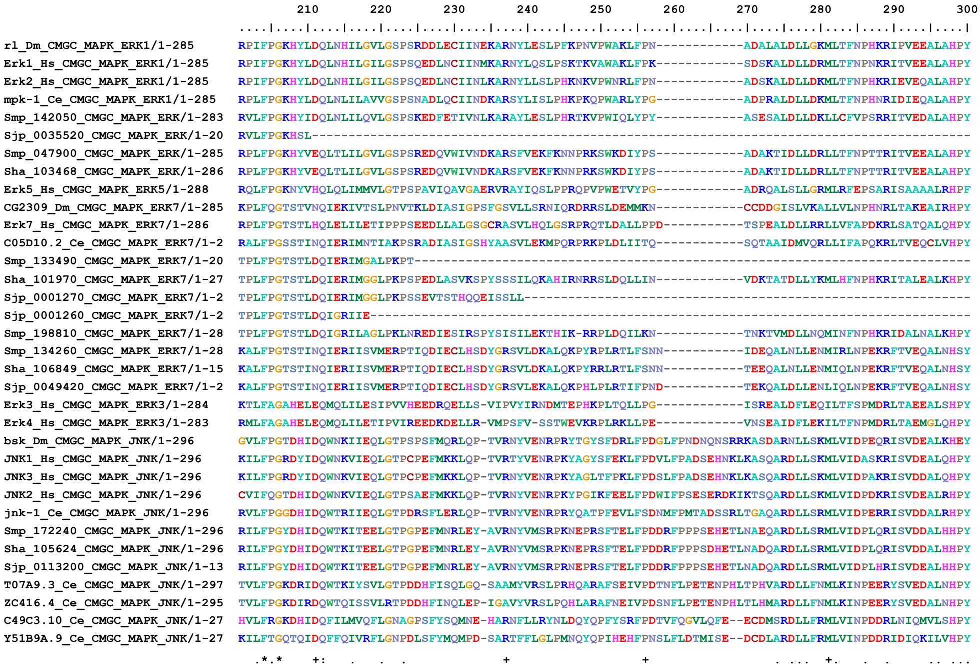
Table 1- Primers sequence

Protein ID	Primer name	dsRNA primers
Smp_142050	SmERK1	Fow:5'-taatacgactcactatagggTTGGTCAATTGGTTGTATTATGG-3' Rev:5'-taatacgactcactatagggGGAACAATGGCACCAGAAT-3'
Smp_047900	SmERK2	Fow:5'-taatacgactcactatagggTCTGCCAGCGAACATATCG-3' Rev:5'-taatacgactcactatagggGGATCACCAAGTCGTGAAGA-3'
Smp_011660.2	SmCaMK2	Fow:5'-taatacgactcactatagggGATGACATTCAGGACGAAGG-3' Rev:5'-taatacgactcactatagggTCGCAGGACTGACTGTTAG-3'
Smp_172240	SmJNK	Fow:5'-taatacgactcactatagggACATGCAGCCGGTATAATCC-3' Rev:5'-taatacgactcactatagggTTACTTCAGAGTCTTCATACCATACG-3'
Smp_179910	SmRas	Fow:5'-taatacgactcactatagggTGGCACCAGAACTTATCAGG-3' Rev:5'-taatacgactcactatagggGATATAGAGCAGTCATTGCATTCC-3'
pCRII-GFP	GFP	Fow:5'-taatacgactcactatagggTCTTCAAGTCCGCCATG-3' Rev:5'-taatacgactcactatagggTGCTCAGGTAGTGGTTGTC-3'
Protein ID	**Primer name**	**qPCR primers**
Smp_142050	qSmERK1	Fow:5'-TGCAACATCTTGTTGAATGC-3' Rev:5'-GCACGATACCAACGTGTACG-3'
Smp_047900	qSmERK2	Fow:5'-TTATCCTTCGGCGGATGC-3' Rev:5'-AGCAACAGGCTCATCACTAGG-3'
Smp_011660.2	qSmCaMK2	Fow:5'-ACGACTATGCTAGCCACACG-3' Rev:5'-CAGACGATTCCTTAATACCATCG-3'
Smp_172240	qSmJNK	Fow:5'-TCCTCCTGGGTATCATGTCG-3' Rev:5'-GCTACAACAAAGCCCTGAGC-3'
Smp_179910	qSmRas	Fow:5'-GACTGAGTACAAGTTAGTTGTTGTTGG-3' Rev:5'-TTCTATAAGAGTCCTCTATCGTTGG-3'
Smp_124600	qSmc-Fos	Fow:5'-GAGGCTGCAAGAGAATGTCG-3' Rev:5'-CAAAGTGCTTTAACTTTCTGAAGC-3'
Smp_097730	qSmSRF	Fow:5'-GATACCTATTGAATTTATTTCTGATCG-3' Rev:5'-CGGTTAATTCAGCCAATTCC-3'
AF216698.1	COX	Fow:5'-TACGGTTGGTGGTGTCACAG-3' Rev:5'-ACGGCCATCACCATACTAGC-3'

SUPPORTING INFORMATION







```
                                      210       220       230       240       250       260       270       280       290       300
                               ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
rl_Dm_CMGC_MAPK_ERK1/1-285     RPIFPGKHYLDQLNHILGVLGSPSRDDLECIINEKARNYLESLPFKPNVPWAKLFPN------------ADALALDLLGKMLTFNPHKRIPVEEALAHPY
Erk1_Hs_CMGC_MAPK_ERK1/1-285   RPIFPGKHYLDQLNHILGILGSPSQEDLNCIINMKARNYLQSLPSKTKVAWAKLFPK------------SDSKALDLLDRMLTFNPNKRITVEEALAHPY
Erk2_Hs_CMGC_MAPK_ERK1/1-285   RPIFPGKHYLDQLNHILGILGSPSQEDLNCIINLKARNYLLSLPHKNKVPWNRLFPN------------ADSKALDLLDKMLTFNPHKRIEVEQALAHPY
mpk-1_Ce_CMGC_MAPK_ERK1/1-285  RPLFPGKHYLDQLNLILAVVGSPSNADLQCIINDKARSYLISLPHKPKQPWARLYPG------------ADPRALDLLDKMLTFNPHNRIDIEQALAHPY
Smp_142050_CMGC_MAPK_ERK/1-283 RVLFPGKHYIDQLNLILQVLGSPSKEDFETIVNLKARAYLESLPHRTKVPWIQLYPY------------ASESALDLLDKLLCFVPSRRITVEDALAHPY
Sjp_0035520_CMGC_MAPK_ERK/1-20 RVLFPGKHSL------------------------------------------------------------------------------------------
Smp_047900_CMGC_MAPK_ERK/1-285 RPLFPGKHYVEQLTLILGVLGSPSREDQVWIVNDKARSFVEKFKNNPRKSWKDIYPS------------ADAKTIDLLDRLLTFNPTTRITVEEALAHPY
Sha_103468_CMGC_MAPK_ERK/1-286 RPLFPGKHYVEQLTLILGVLGSPSREDQVWIVNDKARSFVEKFKNNPRKSWKDIYPS------------ADAKTIDLLDRLLTFNPTTRITVEEALAHPY
Erk5_Hs_CMGC_MAPK_ERK5/1-288   RQLFPGKNYVHQLQLIMMVLGTPSPAVIQAVGAERVRAYIQSLPPRQPVPWETVYPG------------ADRQALSLLGRMLRFEPSARISAAAALRHPF
CG2309_Dm_CMGC_MAPK_ERK7/1-285 KPLFQGTSTVNQIEKIVTSLPNVTKLDIASIGPSFGSVLLSRNIQRDRRSLDEMMKN------------CCDDGISLVKALLVLNPHNRLTAKEAIRHPY
Erk7_Hs_CMGC_MAPK_ERK7/1-286   RPLFPGTSTLHQLELILETIPPPSEEDLLALGSGCRASVLHQLGSRPRQTLDALLPPD-----------TSPEALDLLRRLLVFAPDKRLSATQALQHPY
C05D10.2_Ce_CMGC_MAPK_ERK7/1-2 RALFPGSSTINQIERIMNTIAKPSRADIASIGSHYAASVLEKMPQRPRKPLDLIITQ------------SQTAAIDMVQRLLIFAPQKRLTVEQCLVHPY
Smp_133490_CMGC_MAPK_ERK7/1-20 TPLFPGTSTLDQIERIMGALPKPT----------------------------------------------------------------------------
Sha_101970_CMGC_MAPK_ERK7/1-27 TPLFPGTSTLDQIERIMGGLPKPSPEDLASVKSPYSSSILQKAHIRNRRSLDQLLIN------------VDKTATDLLYKMLHFNPHKRITALEALKHPY
Sjp_0001270_CMGC_MAPK_ERK7/1-2 TPLFPGTSTLDQIERIMGGLPKPSSEVTSTHQQEISSLL-------------------------------------------------------------
Sjp_0001260_CMGC_MAPK_ERK7/1-2 TPLFPGTSTLDQIGRIIE----------------------------------------------------------------------------------
Smp_198810_CMGC_MAPK_ERK7/1-28 TPLFPGTSTLDQIGRILAGLPKLNREDIESIRSPYSISILEKTHIK-RRPLDQILKN------------TNKTVMDLLNQMINFNPHKRIDALNALKHPY
Smp_134260_CMGC_MAPK_ERK7/1-28 KALFPGTSTINQIERIISVMERPTIQDIECLHSDYGRSVLDKALQKPYRPLRTLFSNN-----------IDEQALNLLENMIRLNPEKRFTVEQALNHSY
Sha_106849_CMGC_MAPK_ERK7/1-15 KALFPGTSTINQIERIISVMERPTIQDIECLHSDYGRSVLDKALQKPYRRLRTLFSNN-----------TEEQALNLLENMIQLNPEKRFTVEQALNHSY
Sjp_0049420_CMGC_MAPK_ERK7/1-2 KALFPGTSTINQIERIISVMERPTIQDIECLHSDYGRSVLEKALQKPHLPLRTIFPND-----------TEKQALDLLENLIQLNPEKRFTVEQALNHSY
Erk3_Hs_CMGC_MAPK_ERK3/1-284   KTLFAGAHELEQMQLILESIPVVHEEDRQELLS-VIPVYIRNDMTEPHKPLTQLLPG------------ISREALDFLEQILTFSPMDRLTAEEALSHPY
Erk4_Hs_CMGC_MAPK_ERK3/1-283   RMLFAGAHELEQMQLILETIPVIREEDKDELLR-VMPSFV-SSTWEVKRPLRKLLPE------------VNSEAIDFLEKILTFNPMDRLTAEMGLQHPY
bsk_Dm_CMGC_MAPK_JNK/1-296     GVLFPGTDHIDQWNKIIEQLGTPSPSFMQRLQP-TVRNYVENRPRYTGYSFDRLFPDGLFPNDNQNSRRKASDARNLLSKMLVIDPEQRISVDEALKHEY
JNK1_Hs_CMGC_MAPK_JNK/1-296    KILFPGRDYIDQWNKVIEQLGTPCPEFMKKLQP-TVRTYVENRPKYAGYSFEKLFPDVLFPADSEHNKLKASQARDLLSKMLVIDASKRISVDEALQHPY
JNK3_Hs_CMGC_MAPK_JNK/1-296    KILFPGRDYIDQWNKVIEQLGTPCPEFMKKLQP-TVRNYVENRPKYAGLTFPKLFPDSLFPADSEHNKLKASQARDLLSKMLVIDPAKRISVDDALQHPY
JNK2_Hs_CMGC_MAPK_JNK/1-296    CVIFQGTDHIDQWNKVIEQLGTPSAEFMKKLQP-TVRNYVENRPKYPGIKFEELFPDWIFPSESERDKIKTSQARDLLSKMLVIDPDKRISVDEALRHPY
jnk-1_Ce_CMGC_MAPK_JNK/1-296   RVLFPGGDHIDQWTRIIEQLGTPDRSFLERLQP-TVRNYVENRPRYQATPFEVLFSDNMFPMTADSSRLTGAQARDLLSRMLVIDPERRISVDDALRHPY
Smp_172240_CMGC_MAPK_JNK/1-296 RILFPGYDHIDQWTKITEELGTPGPEFMNRLEY-AVRNYVMSRPKNEPRSFTELFPDDRFPPPSEHETLNAEQARDLLSRMLVIDPLQRISVDDALHHPY
Sha_105624_CMGC_MAPK_JNK/1-296 RILFPGYDHIDQWTKITEELGTPGPEFMNRLEY-AVRNYVMSRPKNEPRSFTELFPDDRFPPPSDHETLNADQARDLLSRMLVIDPLQRISVDDALHHPY
Sjp_0113200_CMGC_MAPK_JNK/1-13 RILFPGYDHIDQWTKITEELGTPGPEFMNRLEY-AVRNYVMSRPRNEPRSFTELFPDDRFPPPSEHETLNADQARDLLSRMLVIDPLHRISVDEALHHPY
T07A9.3_Ce_CMGC_MAPK_JNK/1-297 TVLFPGKDRIDQWTKIYSVLGTPDDHFISQLGQ-SAAMYVRSLPRHQARAFSEIVPDTNFLPETENPHLTPHVARDLLFNMLKINPEERYSVEDALNHPY
ZC416.4_Ce_CMGC_MAPK_JNK/1-295 TVLFPGKDIRDQWTQISSVLRTPDDHFINQLEP-IGAVYVRSLPQHLARAFNEIVPDSNFLPETENPHLTLHMARDLLFNMLKINPEERYSVEDALNHPY
C49C3.10_Ce_CMGC_MAPK_JNK/1-27 HVLFRGKDHIDQFILMVQFLGNAGPSFYSQMNE-HARNFLLRYNLDQYQPFYSRFPDTVFQGVLQFE--ECDMSRDLLFRMLVINPDDRLNIQMVLSHPY
Y51B9A.9_Ce_CMGC_MAPK_JNK/1-27 KILFTGQTQIDQFFQIVRFLGNPDLSFYMQMPD-SARTFFLGLPMNQYQPIHEHFPNSLFLDTMISE--DCDLARDLLFRMLVINPDDRIDIQKILVHPY
                                 .*.*    +:   .   .  .             +                  +                 . ...  +.    .  .    .. ...
```

Figure S1 - Multiple Sequence Alignment of ERK and JNK proteins encoded by parasites and free-living organisms. Amino acid sequences of the conserved catalytic domain (PF00069) were aligned using MAFFT 7 with iterative refinement by the G-INS-i strategy [16]. The multiple sequence alignment comprising 34 sequences with 300 sites was manually refined using Jalview [17].

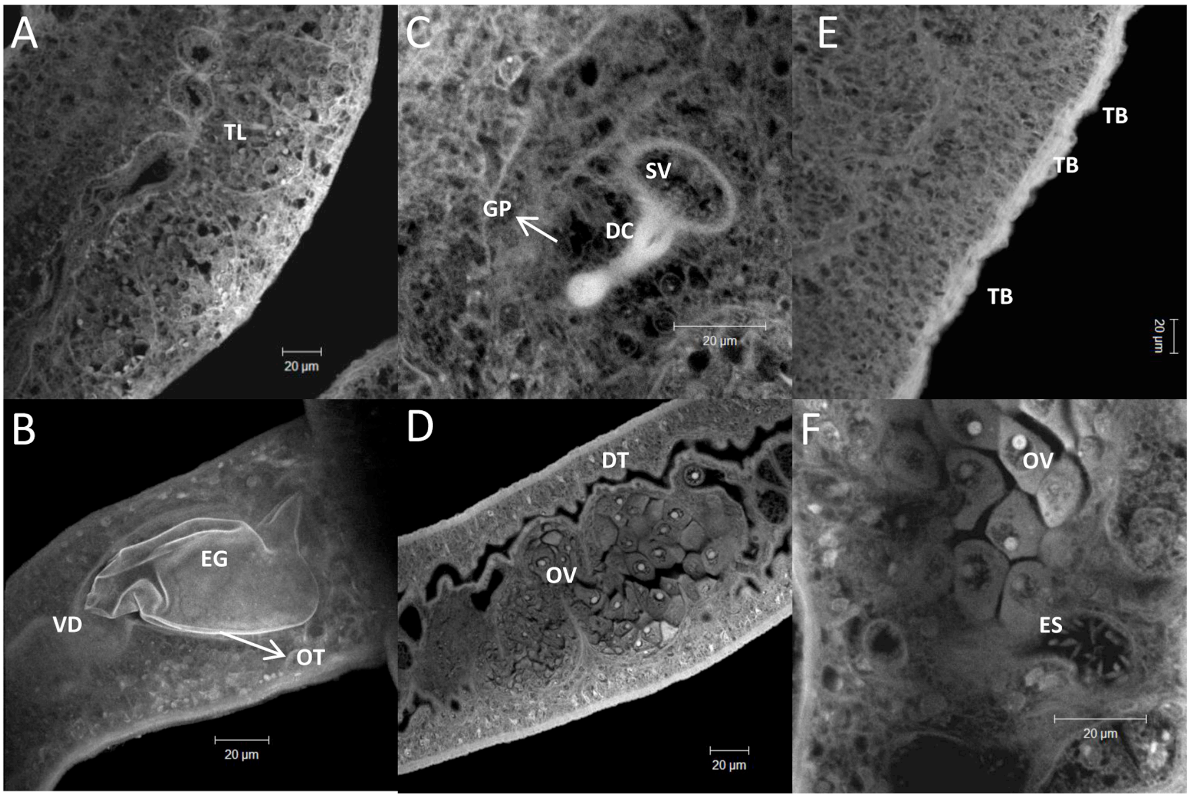


Figure S2 - Images of normal male and female adult worms recovered from mice infected with SmCaMK2-knockdown schistosomula. The 35-day-old adult worms were visualized by laser confocal microscopy. A, C and E show normal male worms treated with SmCaMK2 dsRNA, whereas B, D and F shows normal female worms treated with SmCaMK2 dsRNA. The initials are TL: testicular lobes; SV: seminal vesicle; GP: genital pore; DC: duct for seminal pore; TB: tubercles; EG: egg; OT: ootype; VD; vitelloduct; DT: digestive tract; OV: ovary; ES: spermathec. Scale bars: 20 µm.

IV - DISCUSSÃO

O termo evolução deriva do latim "evolutĭo" e faz referência ao efeito de evoluir. O termo é frequentemente aplicado ao movimento de objetos, pessoas, animais, da tecnologia, mas o contexto mais belo envolve, certamente, a evolução biológica - o processo, na maioria das vezes, lento e gradual de modificação de organismos vivos ao longo do tempo evolutivo. Portanto, evoluir significa mudar; de acordo com a hipótese da rainha vermelha, é correr o mais rápido possível para permanecer no mesmo lugar (Carrol, 1872; Van Valen, 1973). Na presente tese, estudamos a biodiversidade molecular de *S. mansoni* à luz da evolução com o objetivo de identificar os processos evolutivos que modelaram o genoma desse parasito, assim como características linhagem-específicas selecionadas pela seleção natural que permitiram a sobrevivência desses organismos em um ambiente hostil como o sistema circulatório do hospedeiro definitivo.

É sabido que estudos de genômica comparativa têm o potencial de revelar a biodiversidade molecular dentre e entre organismos, sendo uma abordagem relativamente recente que fornece novas informações para questões antigas e fundamentais em biologia. Os parasitas do gênero *Schistosoma* foram os primeiros platelmintos cujos genomas foram completamente sequenciados (Berriman et al., 2009; Zhou et al., 2009; Young et al., 2012). Logo, análises comparativas envolvendo estas espécies podem fornecer informações importantes no que diz respeito à evolução de parasitas lofotrocozoários.

Motivados pelas vantagens da associação de uma perspectiva evolutiva às análises biológicas comparativas, reconstruímos as relações evolutivas de todas as proteínas codificadas no genoma do *S. mansoni* (Capítulo I). O filoma resultante, obtido a partir da análise comparativa de 11.763 proteínas deste parasito com os potenciais homólogos em 12 outros organismos, permitiu uma melhoria na anotação funcional de 5.507 proteínas de *S. mansoni*, 946 das quais eram anotadas previamente como “proteínas hipotéticas”, cuja função era completamente desconhecida. Além de aprimorar a predição funcional do proteoma predito de *S. mansoni*, os resultados desta análise provêm informações importantes sobre a evolução do genoma deste parasito, como a identificação de duplicações gênicas que podem estar relacionadas a especificidades morfológicas, fisiológicas e à patogênese da esquistossomose (Capítulo II).

Dentre as famílias gênicas expandidas em *S. mansoni* com função previamente descrita destacam-se aquelas amplamente estudadas como alvos para drogas, vacina e diagnóstico contra a esquistossomose, como tetraspaninas, fucosiltransferases, leishmanolisinas, *venom allergen-like* (SmVAL ou Scp-like) e *tegumental allergen-like* (SmTAL) (Marques *et al.*, 2001; Tran *et al.*, 2006; McManus *et al.*, 2008; DeMarco & Verjovski-Almeida, 2009; Berriman *et al.,* 2009; Fitzpatrick *et al.*, 2009; Fitzsimmons et al., 2012). Nossos resultados sugerem que os membros dessas famílias proteicas se originaram por sucessivos eventos de duplicação gênica na linhagem ancestral após a divergência dos demais metazoários. Em geral, resíduos críticos estão conservados nesses genes/proteínas. Além disso, cada família proteica possui uma história evolutiva distinta. Com exceção de SmTAL, todas as demais proteínas acima mencionadas são peptidases. Análises proteômicas de parasitos do gênero *Schistosoma* apontam proteases como os principais fatores de virulências envolvidos na patogênese da esquistossomose (McKerrow et al. 2006, Kasný et al. 2009).

Em adição às proteínas que são secretadas, proteínas associadas ao tegumento de *Schistosoma* são frequentemente apontadas como alvo ideal para o desenvolvimento de novas drogas e vacina contra a esquistossomose (e.g. Mulvenna et al., 2010; Zhang et al., 2012). O tegumento é um sincício dinâmico e está diretamente envolvido na interação parasito-hospedeiro, desempenhando funções vitais como nutrição, excreção, osmorregulação, recepção sensorial, transdução de sinal e evasão do sistema imune do hospedeiro (Jones et al., 2004). Embora várias proteínas tenham sido identificadas como parte integrante do tegumento de *Schistosoma*, o conjunto completo de proteínas envolvidas na dinâmica da membrana permanece desconhecido (Mulvenna et al., 2010). Proteínas SmTAL além de estarem presentes em múltiplas cópias no genoma de *S. mansoni*, estão fortemente associadas com o tegumento e parecem ocorrer somente em platelmintos (Subpipattana et al., 2012; Zhang et al., 2012; Fitzsimmons et al., 2012). Em humanos infectados com *Schistosoma*, os membros da família TAL são os principais elementos desencadeadores da resposta IgE, a qual tem sido associada com resistência à reinfecção (Dune et al., 1997; Webster et al., 1996; Pinot de Moira et al., 2010). Dando continuidade a este trabalho, vamos analisar experimentalmente a expressão gênica de membros da família TAL codificados pelo genoma de *S.*

mansoni através da interferência por RNA (RNAi), um processo de silenciamento da expressão gênica desencadeado por pequenos RNAs (siRNA) (c.f anexo 7.1).

Um outro grupo de proteínas consideradas alvos em potencial para o desenvolvimento de drogas do ponto de vista médico e químico, são as proteína quinase eucarióticas (ePKs). Um crescente número de inibidores ePKs foram desenvolvidos e aprovados para o tratamento de diferentes doenças humanas (Boyle & Koleske, 2007). As proteínas quinase (PKs) pertencem a uma grande família de enzimas, muitas das quais mediam respostas a estímulos externos em células eucarióticas (Hanks et al., 1988). As PKs são proteínas de fosforilação reversível que desempenham um papel central na regulação de funções básicas de todos os eucariotas, tais como: replicação, controle do ciclo celular, transcrição gênica, rearranjo do citoesqueleto, apoptose, metabolismo energético e outras. Essas proteínas são também requeridas em funções mais finas em eucariotos complexos, como: diferenciação celular de tecidos e órgãos, comunicação entre células, interação da célula com o substrato e mediação de interações complexas com o ambiente externo (Plowman et al., 1999). Com o intuito de identificar, classificar e anotar funcionalmente as ePKs de *Schistosoma*, adotamos uma abordagem computacional mista baseada em modelos ocultos de Markov (*Hidden Markov Models* - HMM) e abordagem filogenética para anotação funcional de homólogos. Como resultado, um total 243 ePKs foram identificadas em *S. mansoni*, 247 em *S. haematobium* e 213 em *S. japonicum*, sendo que as três espécies possuem membros dos principais grupos (AGC, CaMK, CK1, CMGC, RGC, STE, TK e TKL). As três espécies de *Schistosoma* têm mais representantes no grupo CMGC, uma característica única desse gênero. A grande quantidade de ePKs do grupo CMGC pode estar relacionada ao fato de que as proteínas desse grupo são necessárias no controle da proliferação celular e para garantir a replicação e segregação correta de organelas. Em conjunto, esses processos são essenciais para parasitos que possuem um ciclo de vida complexo como é o caso dos agentes etiológicos da esquistossomose.

Após a identificação e análise do conjunto de proteínas quinase eucarióticas de *S. mansoni* (Capítulo III), baseados na importância funcional dos homólogos, ePKs que atuam como prováveis sinalizadores da via MAPK (Mitogen Activated Protein Kinase), preditas como tendo função essencial para o desenvolvimento e sobrevivência desse organismo, foram selecionadas para caracterização experimental através da interferência por RNA (RNAi) seguido de testes *in vivo* onde realizou-se a infecção de camundongos com esquistossômulos silenciados (Capítulo IV). Grande parte das ePKs selecionadas são representantes únicas de uma família/subfamília evitando assim a substituição de sua atividade nas vias das quais participam. A via de sinalização MAPK é umas das vias mais bem estudadas em diversos organismos e está envolvida em processos essenciais para a viabilidade celular em todos os eucariotos (Wilkinson & Millar, 2000). Os genes escolhidos para a incestigação do papel biológico foram: SmCaMK2 (Smp_011660.2), SmJNK (Smp_172240), SmERK1 (Smp_142050) e SmERK2 (Smp_047900).

Resultados muito interessantes mostraram que o silenciamento da expressão dos genes MAPK da subfamília JNK causam efeitos devastadores no tegumento de vermes adultos de *S. mansoni*, levando à morte dos mesmos. Além disso, os achados deste estudo mostram que ePKs da subfamília ERK1 estão relacionadas com a produção de ovos, já que fêmeas com baixos níveis de transcritos SmERK1 e SmERK2 apresentam ovários pouco desenvolvidos e produção de ovos significativamente baixa. Em conjunto, os resultados de caracterização experimental sugerem que as proteínas MAPK SmERK e SmJNK são alvos de interesse para o desenvolvimento de drogas contra a esquistossomose, visto que um inibidor dessas proteínas provavelmente irá interromper o ciclo de vida de *Schistosoma* e impedir o progresso da doença.

O racional dos estudos desenvolvidos nesta tese de doutorado é que características espécie-específicas refletem a biodiversidade molecular e as adaptações dos organismos a diferentes ambientes, o que pode auxiliar a identificação de novos alvos terapêuticos. Em síntese, as informações oriundas desta tese visam contribuir para o estudo da Biologia de Sistemas de parasitos do gênero *Schistosoma*, pois somente através do conhecimento do completo repertório proteico, dos elementos reguladores expressos em um organismo e como esses produtos interagem, seremos capazes de compreender a complexidade dos sistemas biológicos.

V - CONSIDERAÇÕES FINAIS

Ao longo das últimas décadas, o avanço das tecnologias de sequenciamento associadas ao aumento de poder computacional levou ao crescimento nunca antes observado no volume de dados genômicos, transcritômicos e proteômicos de diversos grupos taxonômicos, o que abriu novas fronteiras para as análises biológicas comparativas. Neste contexto, a filogenômica tem o potencial de conectar a diversidade genômica à biologia evolutiva, estabelecendo relações entre os domínios da vida e contribuindo para a aprimoração da anotação funcional de genes e proteínas. Esta abordagem tem impacto ainda maior no que diz respeito à análises genômicas de parasitos, visto que à luz da evolução é possível revelar características espécie-específicas com potencial aplicação em biomedicina e biotecnologia. Ao utilizar a abordagem filogenômica para analisar o proteoma predito de *S. mansoni*, em comparação com outros organismos parasitos e de vida livre, nós "iluminamos" os processos evolutivos que modelaram as interações parasito-hospedeiro ao longo do tempo evolutivo e identificamos e testamos experimentalmente potenciais alvos para o desenvolvimento de novas drogas contra a esquistossomose. Dando continuidade a este trabalho, análises comparativas envolvendo dados genômicos, transcritômicos e proteômicos de outros helmintos e vetores, assim como a caracterização experimental de outros alvos promissores, aumentarão nosso conhecimento sobre a biologia parasitária rumo à prevenção do progresso da doença.

VI - REFERÊNCIAS

- ✓ Andrade LF, Nahum LA, Avelar LG, Silva LL, Zerlotini A, Ruiz JC, Oliveira G. Eukaryotic protein kinases (ePKs) of the helminth parasite Schistosoma mansoni. BMC Genomics, v.12, 215, 2011.

- ✓ Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyama ST, Al-Lazikani B, Andrade LF, Ashton PD, Aslett MA, Bartholomeu DC, Blandin G, Caffrey CR, Coghlan A, Coulson R, Day TA, Delcher A, DeMarco R, Djikeng A, Eyre T, Gamble JA, Ghedin E, Gu Y, Hertz-Fowler C, Hirai H, Hirai Y, Houston R, Ivens A, Johnston DA, Lacerda D, Macedo CD, McVeigh P, Ning Z, Oliveira G, Overington JP, Parkhill J, Pertea M, Pierce RJ, Protasio AV, Quail MA, Rajandream MA, Rogers J, Sajid M, Salzberg SL, Stanke M, Tivey AR, White O, Williams DL, Wortman J, Wu W, Zamanian M, Zerlotini A, Fraser-Liggett CM, Barrell BG, El-Sayed NM. The genome of the blood fluke Schistosoma mansoni. Nature, v.460, p. 352–358, 2009.

- ✓ Bolton E, Wang Y, Thiessen PA, Bryant SH. PubChem: Integrated Platform of Small Molecules and Biological Activities. In Annual Reports in Computational Chemistry. American Chemical Society, Washington, DC, 2008. cap. 12.

- ✓ Botros SS, Bennett JL. Praziquantel resistance. Expert Opinion on Drug Discovery, v. 2, p. S35-S40, 2007.

- ✓ Boyle SN, Koleske AJ. Dissecting kinase signaling pathways. Drug Discov Today 2007; 12(17-18):717-724.

- ✓ Brown D, Sjolander K. Functional classification using phylogenomics inference. PLoS Computational Biology, v. 2, e. 77, 2006.

- ✓ Carroll L. Through the looking-glass. 1872.

- ✓ Criscione CD, Valentim CL, Hirai H, LoVerde PT, Anderson TJ. Genomic linkage map of the human blood fluke Schistosoma mansoni. Genome Biology, v. 10, R71, 2009.

- ✓ DeMarco R, Mathieson W, Dillon GP, Wilson RA. Schistosome albumin is of host, not parasite, origin. International Journal for Parasitology, v. 37, p. 1201-8, 2007.

- ✓ Dunne DW, Webster M, Smith P, Langley JG, Richardson BA, et al. The isolation of a 22 kDa band after SDS-PAGE of Schistosoma mansoni adult worms and its use to demonstrate that IgE responses against the antigen(s) it contains are associated with human resistance to reinfection. Parasite Immunol. v.19, p: 79–89, 1997.

- ✓ Eisen, JA. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Research, v. 8, p. 163-167, 1998.

- ✓ Eisen JA, Kaiser D, Myers RM. Gastrogenomic delights: a movable feast. Nature Medicine, v. 3, p. 1076-1078, 1997.

- ✓ Eisen JA, Wu M. Phylogenetic analysis and gene functional predictions: phylogenomics in action. Theoretical Population Biology, v. 61, p. 481–487, 2002.

- ✓ Felsenstein J. Phylogenies and the comparative method. Am Nat, v. 125, p. 1-15, 1985.

- ✓ Fitzpatrick JM, Peak E, Perally S, Chalmers IW, Barrett J, Yoshino TP, Ivens AC, Hoffmann KF. Anti-schistosomal intervention targets identified by lifecycle transcriptomic analyses. PLoS Neglected Tropical Diseases, v. 3, e. 543, 2009.

- ✓ Fitzsimmons CM, Jones FM, Stearn A, Chalmers IW, Hoffmann KF, Wawrzyniak J, Wilson S, Kabatereine NB, Dunne DW. The Schistosoma mansoni tegumental-allergen-like (TAL) protein family: influence of developmental expression on human IgE responses. PLoS Negl Trop Dis v.6, e1593, 2012.

- ✓ Gabaldón T. Comparative Genomics-Based Prediction of Protein Function. Genomics Protocols, v. 439, p. 387-401, 2008.

- ✓ Gabaldón T, Huynen MA. Prediction of protein function and pathways in the genome era. Cellular and Molecular Life Sciences, v.61, p. 930-44, 2004.

- ✓ Galperin MY, Koonin EV. Who's your neighbor? New computational approaches for functional genomics. Nature Biotechnology, v. 18, p. 609-13, 2000.

- ✓ Greenberg RM. New approaches for understanding mechanisms of drug resistance in schistosomes. Parasitology. 2013 Apr 3:1-13. [Epub ahead of print] PubMed PMID: 23552512.

- ✓ Gryseels B, Polman K, Clerinx J, Kestens L. Human schistosomiasis. Lancet, v. 368, p. 1106-18, 2006.

- ✓ Han ZG, Brindley PJ, Wang SY, Chen Z. Schistosoma genomics: new perspectives on schistosome biology and host-parasite interaction. Annu Rev Genomics Hum Genet, v. 10, p. 211-240, 2009.

- ✓ Hanks SK, Quinn AM, Hunter T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science 1988; 241(4861):42-52.

- ✓ Hardison RC. Comparative genomics. PLoS Biol, v. 1, e58, 2003.

- ✓ Hawkins T, Kihara D. Function prediction of uncharacterized proteins. Journal of Bioinformatics and Computacional Biology, v.5, p. 1-30, 2007.

- ✓ Hood L. Systems biology: integrating technology, biology, and computation. Mechanisms of Ageing and Development, v. 124, p. 9-16, 2003.

- ✓ Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T. The human phylome. Genome Biology, v. 8, R109, 2007.

- ✓ Huerta-Cepas J, Marcet-Houben M, Pignatelli M, Moya A, Gabaldón T. The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for Acyrthosiphon pisum genes. Insect Molecular Biology, v. 19, p. 13-21, 2010.

- ✓ Jones MK, Gobert GN, Zhang L, Sunderland P, McManus DP. The cytoskeleton and motor proteins of human schistosomes and their roles in surface maintenance and host-parasite interactions. Bioessays v.26, p: 752-65, 2004.

- ✓ Kasný M, Mikes L, Hampl V, Dvorák J, Caffrey CR, Dalton JP, Horák P. Chapter 4. Peptidases of trematodes. Adv Parasitol v. 69, p. 205-297, 2009.

- ✓ Lambertucci JR. Acute schistosomiasis mansoni: revisited and reconsidered. Memorias do Instituto Oswaldo Cruz, v. 105, p. 422–435, 2010.

- ✓ Lawton SP, Hirai H, Ironside JE, Johnston DA, Rollinson D. Genomes and geography: genomic insights into the evolution and phylogeography of the genus Schistosoma. Parasit Vectors, v. 7, 131, 2011.

- ✓ Liang YS, Dai JR, Zhu YC, Coles GC, Doenhoff MJ. Genetic analysis of praziquantel resistance in Schistosoma mansoni. The Southeast Asian Journal of Tropical Medicine and Public Health, v. 34, p. 274-80, 2003.

- ✓ LoVerde PT, Hirai H, Merrick JM, Lee NH, El-Sayed N. Schistosoma mansoni genome project: an update. Parasitology International, v. 53, p. 183-92, 2004.

- ✓ Marques Jr ETA, Ichikawa Y, Strand M, August JT, Hart GW, Schnaar RL. Fucosyltransferases in Schistosoma mansoni development. Glycobiology, v. 11, p. 249-59, 2001.

- ✓ McKerrow JH, Caffrey C, Kelly B, Loke P, Sajid M 2006. Proteases in parasitic diseases. Annu Rev Pathol 1: 497-536.

- ✓ McManus DP, Loukas A. Current status of vaccines for schistosomiasis. Clinical Microbiology Reviews, v. 21, p. 225-42, 2008.

- ✓ Medina M. Genomes, phylogeny, and evolutionary systems biology. Proceedings of the National Academy of Sciences, v. 102, p. 6630-6635, 2005.

- ✓ Melman SD, Steinauer ML, Cunningham C, Kubatko LS, Mwangi IN, Wynn NB, Mutuku MW, Karanja DM, Colley DG, Black CL, Secor WE, Mkoji GM, Loker ES. Reduced susceptibility to praziquantel among naturally occurring Kenyan isolates of Schistosoma mansoni. PLoS Neglected Tropical Disease, v. 3, e. 504, 2009.

- ✓ Mulvenna J, Moertel L, Jones MK, Nawaratna S, Lovas EM, Gobert GN, Colgrave M, Jones A, Loukas A, McManus DP. Exposed proteins of the Schistosoma japonicum tegument. Int J Parasitol. v.40, p: 543-54, 2010.

- ✓ Nahum LA, Goswami S, Serres MH: Protein families reflect the metabolic diversity of organisms and provide support for functional prediction. Physiol Genomics, v. 38, p:250–260, 2009.

- ✓ Nahum LA, Mourão MM, Oliveira G. New frontiers in schistosoma genomics and transcriptomics. J Parasitol Res, v. 2012, 849132, 2012.

- ✓ Nahum LA, Pereira SL. Phylogenomics, Protein Family Evolution, and the Tree of Life: An Integrated Approach between Molecular Evolution and Computational Intelligence. In: Smolinski TG, Milanova MG, Hassanien A-E (eds), Studies in Computational Intelligence (SCI) 122. Springer-Verlag Berlin Heidelberg, 2008. cap. 11, p. 259–279.

- ✓ Pica-Mattoccia L, Cioli D. Sex- and stage-related sensitivity of Schistosoma mansoni to in vivo and in vitro praziquantel treatment. International Journal for Parasitology, v. 34, p. 527-33, 2004.

- ✓ Pinot de Moira A, Fulford AJ, Kabatereine NB, Ouma JH, Booth M, et al. Analysis of complex patterns of human exposure and immunity to Schistosomiasis mansoni: the influence of age, sex, ethnicity and IgE. PLoS Negl Trop Dis v. 14, e820, 2010.

- ✓ Plowman GD, Sudarsanam S, Bingham J, Whyte D, Hunter T. The protein kinases of Caenorhabditis elegans: a model for signal transduction in multicellular organisms. Proc Natl Acad Sci U S A, v. 96, p. 13603-13610, 1999.

- ✓ Protasio AV, Tsai IJ, Babbage A, Nichol S, Hunt M, Aslett MA, De Silva N, Velarde GS, Anderson TJ, Clark RC, Davidson C, Dillon GP, Holroyd NE, LoVerde PT, Lloyd C, McQuillan J, Oliveira G, Otto TD, Parker-Manuel SJ, Quail MA, Wilson RA, Zerlotini A, Dunne DW, Berriman M: A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni. PLoS Negl Trop Dis, v. 6, e1455, 2012.

- ✓ Rokni MB: Schistosomiasis. Croatia: InTech, 2012. p. 322.

- ✓ Sjölander K. Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics, v. 20, p. 170-179, 2004.

✓ Steinmann P, Keiser J, Bos R, Tanner M, Utzinger J. Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. Lancet Infect Dis, v. 6, p. 411-25, 2006.

✓ Subpipattana P, Grams R, Vichasri-Grams S. Analysis of a calcium-*binding E*F-hand protein family in Fasciola gigantica. Exp Parasitol. v.130, p: 364-73, 2012.

✓ Swain MT, Larkin DM, Caffrey CR, Davies SJ, Loukas A, Skelly PJ, Hoffmann KF. Schistosoma comparative genomics: integrating genome structure, parasite biology and anthelmintic discovery. Trends Parasitol, v. 27, p. 555-64, 2011.
✓ Touchman, J. Comparative Genomics. Nature Education Knowledge, v. 3, p. 13

✓ Tran MH, Pearson MS, Bethony JM, Smyth DJ, Jones MK, Duke M, Don TA, McManus DP, Correa-Oliveira R, Loukas A. Tetraspanins on the surface of Schistosoma mansoni are protective antigens against schistosomiasis. Nature Medicine, v. 12, p. 835-40, 2006.

✓ Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL, Tracey A, Bobes RJ, Fragoso G, Sciutto E, Aslett M, Beasley H, Bennett HM, Cai J, Camicia F, Clark R, Cucher M, De Silva N, Day TA, Deplazes P, Estrada K, Fernández C, Holland PW, Hou J, Hu S, Huckvale T, Hung SS, Kamenetzky L, Keane JA, Kiss F, Koziol U, Lambert O, Liu K, Luo X, Luo Y, Macchiaroli N, Nichol S, Paps J, Parkinson J, Pouchkina-Stantcheva N, Riddiford N, Rosenzvit M, Salinas G, Wasmuth JD, Zamanian M, Zheng Y; Taenia solium Genome Consortium, Cai X, Soberón X, Olson PD, Laclette JP, Brehm K, Berriman M. The genomes of four tapeworm species reveal adaptations to parasitism. Nature, v. 496, p. 57-63, 2013.

✓ Van Valen L. A new evolutionary law. Evol. Theory, v. 1, p. 1-30, 1973.

✓ Waknine-Grinberg JH, Gold D, Ohayon A, Flescher E, Heyfets A, Doenhoff MJ, Schramm G, Haas H, Golenser J. Schistosoma mansoni infection reduces the incidence of murine cerebral malaria. Malaria Journal, v. 9, 2010.

✓ Webster JP, Oliviera G, Rollinson D, Gower CM. Schistosome genomes: a wealth of information. Trends in Parasitology, v. 26, p. 103-6, 2010.

✓ Wilkinson MG, Millar JB. Control of the eukaryotic cell cycle by MAP kinase signaling pathways. FASEB J, v. 14, p. 2147-2157, 2000.

✓ Webster M, Fulford AJ, Braun G, Ouma JH, Kariuki HC, et al. Human immunoglobulin E responses to a recombinant 22.6-kilodalton antigen from Schistosoma mansoni adult worms *are* associated with low intensities of reinfection after treatment. Infect Immun v.64, p: 4042–4046, 1996.

✓ Wei L, Liu Y, Dubchak I, Shon J, Park J. Comparative genomics approaches to study organism similarities and differences. J Biomed Inform, v. 35, p. 142-50, 2002.

- ✓ WHO - World Health Organization: Schistosomiasis: number of people treated in 2009. Wkly Epidemiol Rec, v. 86, p. 73–80, 2011.

- ✓ WHO - World Health Organization: Schistosomiasis: number of people treated in 2011. Wkly Epidemiol Rec, v. 88, p. 81-88, 2013.

- ✓ Young ND, Jex AR, Li B, Liu S, Yang L, Xiong Z, Li Y, Cantacessi C, Hall RS, Xu X, Chen F, Wu X, Zerlotini A, Oliveira G, Hofmann A, Zhang G, Fang X, Kang Y, Campbell BE, Loukas A, Ranganathan S, Rollinson D, Rinaldi G, Brindley PJ, Yang H, Wang J, Gasser RB: Whole-genome sequence of Schistosoma haematobium. Nature Genetics, v. 44, p:221–225, 2012

- ✓ Zerlotini A, Heiges M, Wang H, Moraes RL, Dominitini AJ, Ruiz JC, Kissinger JC, Oliveira G. SchistoDB: a Schistosoma mansoni genome resource. Nucleic Acids Research, v. 37, p. D579-82, 2009.

- ✓ Zhang Z, Xu H, Gan W, Zeng S, Hu X. Schistosoma japonicum calcium-binding tegumental protein SjTP22.4 immunization confers praziquantel schistosomulumicide and antifecundity effect in mice. Vaccine v. 30, p: 5141-50, 2012.

- ✓ Zhou Y, Zheng H, Chen Y, Zhang L, Wang K, Guo J, Huang Z, Zhang B, Huang W, Jin K, Dou T, Hasegawa M, Wang L, Zhang Y, Zhou J, Tao L, Cao Z, Li Y, Vinar T, Brejova B, Brown D, Li M, Miller DJ, Blair D, Zhong Y, Chen Z, Liu F, Hu W, Wang ZQ, Zhang QH, Song HD, Chen S, Xu X, Xu B, Ju C, Huang Y, Brindley PJ, McManus DP, Feng Z, Han ZG, Lu G, Ren S, Wang Y, Gu W, Kang H, Chen J, Chen X, Chen S, Wang L, Yan J, Wang B, Lv X, Jin L, Wang B, Pu S, Zhang X, Zhang W, Hu Q, Zhu G, Wang J, Yu J, Wang J, Yang H, Ning Z, Beriman M, Wei CL, Ruan Y, Zhao G, Wang S, Liu F, Zhou Y, Wang ZQ, Lu G, Zheng H, Brindley PJ, McManus DP, Blair D, Zhang QH, Zhong Y, Wang S, Han ZG, Chen Z, Wang S, Han ZG, Chen Z.. The Schistosoma japonicum genome reveals features of host-parasite interplay. Nature v.460, p: 345-351, 2009.

VII - ANEXOS

Desde o ingresso no Doutorado, além dos trabalhos já mencionados, atuo como colaboradora em outros estudos. Uma síntese destes projetos em andamento encontra-se a seguir.

7.1 - Evolutionary Genomics of Schistosome Tegumental-Allergen-Like (TAL) Proteins

Larissa Lopes Silva[1,2], Guilherme Oliveira[1,2], Laila Alves Nahum[1,2,3]

1 - Grupo de Genômica e Biologia Computacional, CPqRR/FIOCRUZ, Belo Horizonte, MG, Brazil. 2 - Centro de Excelência em Bioinformática, CPqRR/FIOCRUZ, Belo Horizonte, MG, Brazil. 3 - Faculdade Infórium de Tecnologia, Belo Horizonte, MG, 30130-180, Brazil.

The tegument of schistosomes is pointed as an ideal target for drug and vaccine development against schistosomiasis, since it is a very dynamic host-interactive layer involved key biological functions. Tegumental-Allergen-Like (TAL) proteins are strongly associated with the tegument and seem to occur solely in Platyhelminthes. In schistosome infections, TAL family members induce human IgE response that has been associated with resistance against parasite reinfection. The present study aimed at analyzing the *Schistosoma mansoni, S. haematobium* and *S. japonicum* predicted proteomes in order to identify the TAL proteins through combined computational approaches and to reconstruct the evolutionary history of these proteins and their homologs in other Platyhelminthes. Selected taxa included: *Clonorchis sinensis*, *Echinococcus multiloculares*, *Faciola gigantica*, *F. hepatica*, *Hymenolepis microstoma*, *Taenia solium*, *S. bovis*, and *S. mansoni*. Potential SmTAL homologs where retrieved from public databases by sequence similarity searches using BLAST. Additional sequences were identified from complete proteomes through HMMs (Hidden Markov Models). Whole amino acid sequences containing concomitant two conserved domains, EF-hand (IPR011992) and Dynein Light Chain (IPR001372), were aligned by MAFFT. Sequence alignments were further used in phylogenetic analysis using the maximum likelihood method as implemented in PhyML. The total number of TAL family members in the three *Schistosoma* species ranged from five (*S. japonicum*) to thirteen (*S. mansoni*). All *S. mansoni* TAL proteins have homologs in *S. haematobium*, with the exception of the putative dynein light chain 1 (cytoplasmic) protein (Smp_072620.1). Our findings suggest that TAL proteins originated from successive gene duplication events in the Platyhelminthes lineage after its diversification from other metazoans. In conclusion, this work provides an evolutionary view of TAL proteins allowing a deeper understanding of lineage specific adaptations potentially related to the Platyhelminthes parasitic lifestyle. Continuing this work, we aim at experimentally characterize these proteins through RNA interference (RNAi), a method employed for manipulating gene-specific expression, to elucidate the functional role of SmTAL proteins.

Este projeto encontra-se em andamento com previsão de publicação em 2014.

7.2 - Evolutionary History of Freshwater Cymothoids (Isopoda, Cymothoidae) from Lake Tanganyika, East Africa

Larissa Lopes Silva[1,2], Moritz Muschick[1], José Celso Malta[3], Walter Salzburger[1]

1 - University of Basel, Zoological Institute, Basel, Switzerland. 2 - Centro de Pesquisas René Rachou – FIOCRUZ, Belo Horizonte - MG, Brazil. 3 - Instituto Nacional de Pesquisas da Amazônia, Manaus – AM, Brazil

Cymothoids (Isopoda, Cymothoidae) are one of the most derived lineages of isopods and are among the largest fish parasites, causing considerable damage to their hosts. Around 325 cymothoid species from 42 genera are known and the vast majority is marine. Some species are found in freshwaters, though, e.g. in East African Lake Tanganyika, where cymothoids are known to parasite cichlid fishes. Little is known regarding the taxonomy and evolutionary history of cymothoids in freshwater environments. To explore this topic, the present study addresses two main questions: i) Where did Lake Tanganyika's cymothoid species originate? ii) Which is their evolutionary history? To answer these questions, we sequenced a 630-bp fragment of the Cytochrome Oxidase I (COI) gene from 14 marine and freshwater species. Phylogenetic analyses were performed using Maximum Likelihood. Preliminary results indicate that Tanganyikan freshwater cymothoids are monophyletic and derived from a marine ancestor. Congruency of host and parasite phylogeny also supports a co-radiation of freshwater cymothoids with cichlid hosts. To improve our understanding on these parasites, the inclusion of South American freshwater samples will allow us to determine if the colonization of freshwater by cymothoid species occurred before the split-up of Gondwana or if it happened as multiple events.

Este projeto encontra-se em andamento com previsão de publicação em 2014.

7.3 - First Report on the Occurrence of *Lymnaea cubensis* and *L. truncatula* in the State of Minas Gerais, Brazil: Intermediate Hosts of *Fasciola hepatica*

Larissa Lopes Silva[1,2,3], Paula Cristina Marques Cardoso[4], Camilla de Medeiros de Carvalho[4], Jean-Pierre Pointier[5], Alejandra Rumi[6], Omar dos Santos Carvalho[4], Roberta Lima Caldeira[4*]

1- Grupo de Genômica e Biologia Computacional, CPqRR/FIOCRUZ, Belo Horizonte, MG, Brazil. 2- Centro de Excelência em Bioinformática, CPqRR/FIOCRUZ, Belo Horizonte, MG, Brazil. 3- Instituto de Ciências Biológicas, UFMG 4- Laboratório de Helmintologia e Malacologia Médica, CPqRR/FIOCRUZ, Belo Horizonte, MG, Brazil. 5- Laboratoire Ecosystèmes Aquatiques Tropicaux et Méditerranéens, UMR 5244 CNRS-EPHE-UPVD, France. 6- División Zoologia Invertebrados Museo de La Plata, Facultad de Ciências Naturales y Museo, Universidad Nacional de La Plata, Argentina.

Molluscs belonging to the genus *Lymnaea* are vectors of fasciolosis, a parasitic disease of veterinary and medical importance. Morphological identification of these species is undertaken by means of comparison of characters of the shell, genital and renal systems. However, such identification is difficult due to interspecific similarities particularly between the cryptic species *L. viatrix*, *L. cubensis* and *L. truncatula*. In order to overcome these difficulties, we have assessed the use of polymerase chain reaction and restriction fragment length polymorphism (PCR-RFLP), directed to the internal transcribed spacer region II (ITS2) of rDNA. PCR-RFLP with the enzymes MboII, HpaII and AluI proved to be an effective taxonomic tool that allows molecular differentiation between *L. viatrix*, *L. cubensis* and *L. truncatula* when morphological characterization impaired species differentiation. Moreover, this technique enabled us to observe that populations of *Lymnaea* from Rio Acima (Minas Gerais, Brazil) and Belo Horizonte (Minas Gerais, Brazil), previously identified by morphology as *L. viatrix*, showed species-specific profiles corresponding respectively to *L. truncatula* and *L. cubensis*. To confirm the molecular identification, phylogenetic analyses were performed on partial mitochondrial (16S) gene sequences. The results obtained mirror the identification by PCR-RFLP. Altogether, our findings present an alternative for Lymneid's morphological identification due to their pronounced morphoanatomic uniformity, which makes a classification by traditional methods almost impossible, besides reporting for the first time the occurence of *L. truncatula* in Brazil and *L. cubensis* in the state of Minas Gerais, Brazil.

Este projeto encontra-se em fase final. O artigo está sendo redigido e a expectativa de submissão é outubro de 2013.

7.4 - Identification of Cystatin Homologs Across Different Helminth Species

Yesid Cuesta[1,2], Larissa Lopes Silva[1,2], Guilherme Oliveira[1,2], Laila Alves Nahum[1,2,3]

1 - Grupo de Genômica e Biologia Computacional, CPqRR/FIOCRUZ, Belo Horizonte, MG, Brazil. 2 - Centro de Excelência em Bioinformática, CPqRR/FIOCRUZ, Belo Horizonte, MG, Brazil. 3 - Faculdade Infórium de Tecnologia, Belo Horizonte, MG, 30130-180, Brazil.

Cystatins comprise a superfamily of cysteine protease inhibitors identified in a broad range of taxonomic groups including Platyhelminthes (flatworms) and Nematoda (roundworms). Some cystatins are known to be secreted. These proteins are involved in regulation of parasite cysteine protease activity and in modulation of host immune responses. Little is known about the origin and evolution of the cystatin superfamily, especially in respect to the functional diversification of these proteins in parasite helminths. The present study aimed at identifying cystatin homologs in the predicted proteome of different helminths and reconstructing the evolutionary relationships of these proteins through combined computational approaches. The predicted proteome data of 16 helminth species (4 Platyhelminthes and 12 Nematoda) was downloaded from public databases. Potential cystatin homologs were identified using Hidden Markov Models (HMMs). Proteins containing the conserved cystatin domain (PF00031), a signal peptide, and four conserved cysteine residues forming two disulphide bonds were selected for further analysis. With this approach, we expected to select type 2 cystatins, which are mainly extracellular secreted proteins as described elsewhere. Amino acid sequences corresponding to the cystatin conserved domain were aligned by MAFFT. The manually- edited alignment was further used in maximum likelihood evolutionary analysis as implemented in PhyML. We have identified potential cystatin homologs across different helminths including free-living and parasitic species. Our findings suggest that the number of homologs vary significantly across the helminth species analyzed. Additionally the evolutionary patterns observed in the tree showed that the cystatins are placed in different clades, reflecting their diversity at the molecular level. Trees also revealed post-speciation gene duplication events shaping the evolution of helminth cystatins. Our evolutionary analysis has provided insights into the evolution of cystatins improving our understanding of the diversity of such proteins at the molecular level. Moreover, our results have paved the way for future structural and evolutionary studies of the cystatin superfamily as well as other secretome protein families. In the future, we plan to perform protein-protein interaction network analysis to identify relationships among parasite secreted and host proteins. This approach will shed light on the diversity of interactions between helminths and their hosts in an evolutionary context.

Este projeto encontra-se em andamento com previsão de publicação em 2014.