Universidade Federal do Rio Grande do Sul

Programa de Pós-Graduação em Genética e Biologia Molecular

Análise evolutiva e funcional das proteínas de estresse universal em parasitos platelmintos

Sergio Martín Espínola

Tese submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da UFRGS como requisito parcial para a obtenção do grau de Doutor em Ciências (Genética e Biologia Molecular)

Orientador: Arnaldo Zaha
Co-Orientador: Martin Pablo Cancela Sehabiague

Porto Alegre, Dezembro de 2017

Este trabalho foi desenvolvido no Laboratório de Biologia Molecular de Cestódeos do Centro de Biologia (CBiot) da Universidade Federal do Rio Grande do Sul (UFRGS), e recebeu suporte financeiro das seguintes fontes de fomento à pesquisa: Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

O aluno recebeu bolsa de estudos concedida pela CAPES vinculada ao Programa de Pós-Graduação em Genética e Biologia Molecular (PPGBM) da UFRGS.

Dedicado a
Camila e Milena

Agradecimentos

Em primeiro lugar, ao Brasil, país que me acolheu da melhor forma possível. País cheio de cultura, pessoas, e costumes diferentes, que só contribuíram para o meu melhor. Espero um dia poder retribuir um pouco de tudo aquilo que me foi dado ao longo destes anos.

Ao corpo docente e não docente do CBiot e do PPGBM da UFRGS, pelo ensino superior de excelente qualidade.

Ao meu orientador Arnaldo Zaha, pelo apoio continuo, confiança, suporte, ensinamentos, e ótima orientação ao longo de todos esses anos. Estarei sempre agradecido por abrir as portas do seu laboratório.

Ao meu co-orientador Martin Cancela, pelos ensinamentos, companhia, e amizade dentro e fora do laboratório.

Aos membros da banca, por aceitar o convite e pelas contribuições a este trabalho.

Aos professores Henrique Ferreira e Karina Monteiro, pelos ensinamentos e acompanhamento ao longo desses anos.

Aos meus colegas dos laboratórios 204, 206, e 210, aqueles que já foram embora e aqueles que ainda estão trabalhando. Aprendi muito, mas muito junto com eles. Com certeza foram parte essencial da minha formação profissional e pessoal, pois nossa parceria não se limitava ao trabalho dentro do laboratório.

Aos meus amigos que fiz ao longo destes anos no Brasil e que vou levar comigo para sempre. André, Livia, Julio, Ricardo, Gleice, Priscila, obrigado pelos momentos incríveis que já compartilhamos e que ainda vamos viver.

Aos meus amigos e familiares da Argentina, pela parceria, suporte, amizade, companhia, e claro, muitíssimo amor envolvido.

A Tulipa, minha cachorrinha. Ela muitas vezes (para quem mora longe de casa), era minha única companhia, com aquele amor incondicional que os animais têm para dar.

Aos meus pais Rosa e Roberto, e irmãos Ariel e Gabriela, pela eterna parceria, amor, ensinamentos, suporte, e muito mais. Vocês sempre estiveram ao meu lado, e se algum dia estive mal, nesse momento pensei na incrível família que eu tenho, e isso já me fez feliz novamente.

E finalmente, às minhas sobrinhas Camila (1) e Milena (1), é para vocês que dedico este trabalho. Seu tio ama vocês e espera que no futuro vocês encontrem, assim como eu, a paixão por aquilo que faz.

Sumário

Lista de abreviaturas

°C: graus celsius
3D: tridimensional
aa: aminoácidos
ADP: adenosina difosfato
AMP: adenosina monofosfato
ATP: adenosina trifosfato
BI: inferência bayesiana
BSA: albumina de soro bovino
IPTG: isopropil-β-D-tiogalactopiranosideo
cDNA: DNA cópia
DNA: ácido desoxirribonucleico
GOLD: banco de dados *online* de genomas
GST: glutationa S transferase
GTP: guanosina trifosfato
HD: hospedeiro definitivo
HI: hospedeiro intermediário
HSP: proteínas de choque térmico
ID: número de identificação
Indel: inserção-deleção
Ka/Ks: razão entre o número de substituições não sinônimas e sinônimas
kDa: quilodalton(s)
LRT: teste da razão de verossimilhança
MEG: genes de micro éxons
ML: máxima verossimilhança
mRNA: RNA mensageiro
NTD: doenças tropicais desatendidas
OMS: organização mundial da saúde
pb: pares de base
PBS: tampão salino com fosfato
PCR: reação em cadeia da polimerase
PDB: banco de dados de proteínas
pH: potencial hidrogeniônico
Pi: fosfato inorgânico
PSC: protoescólices
PSS: sítios que apresentam seleção positiva
qPCR: PCR quantitativa
RNA: ácido ribonucleico
RNA-seq: sequenciamento de RNAs
RT-PCR: PCR a partir de produtos de transcrição reversa
SDS-PAGE: electroforese em gel de poliacrilamida em presença de SDS
SNP: polimorfismo de nucleotídeo único
TBS: tampão salino com tris
USP: Proteínas de estresse Universal
UV: Ultravioleta

Resumo

As proteínas de estresse universal (USPs) são codificadas por uma família gênica amplamente distribuída nos organismos, a exceção dos deuterostomados. As USPs possuem um motivo conservado de ligação a ATP e análogos, e participam na resposta a diversos tipos de estímulos como estresse oxidativo, choque térmico, falta de nutrientes, e radiação UV, o que sugere um papel importante destas proteínas na homeostasia celular. Devido à natureza intrínseca das relações parasito-hospedeiro, diversos tipos de estresses (bióticos e abióticos) estão presentes ao longo do ciclo de vida do parasito, o que indica que as USPs possam ser fundamentais para a sobrevivência e permanência no seu hospedeiro. Assim, o objetivo deste trabalho é realizar uma análise filogenética e de diversidade evolutiva através do uso da genômica comparativa e métodos para inferir divergência funcional em espécies do filo dos Platelmintos. Além disso, propomos a caracterização funcional das USPs através da produção de uma USP de *Echinococcus ortleppi* (classe Cestoda) como proteína recombinante (rEoUSP-1) em *Escherichia coli*, e posteriores análises de oligomerização e ligação e hidrólise de ATP. Nosso trabalho identificou que os genes das USPs se organizam em *clusters*, possuem pseudogenes, e apresentam um amplo grau de divergência funcional, especificamente em aminoácidos próximos ao domínio de interação com o ligante. Estes dados indicam que as USPs seguem o modelo de nascimento e morte (*birth and death*) de genes, onde os genes adotam caminhos evolutivos diferentes ao longo do tempo (pseudogenização, sub/neofuncionalização). Por outro lado, a proteína recombinante rEoUSP-1 mostrou sua capacidade de ligar ATP. A atividade ATPase foi detectada somente em altas concentrações da proteína, o que sugere que esta atividade precisaria de outros componentes celulares. A rEoUSP-1 foi capaz de formar dímeros, o que está de acordo com dados já descritos na literatura. Em conclusão, a divergência funcional das USPs pode ter implicações na resposta a estímulos diferentes, na interação com diversos tipos de ligante, ou numa expressão gênica específica no espaço e tempo. Além disso, as análises funcionais da rEoUSP-1 contribuem para o entendimento destas proteínas em organismos eucariotos, e na interface parasito-hospedeiro.

Abstract

Universal stress proteins (USPs) are encoded by a gene family with a wide distribution among organisms, except in deuterostomes. USPs exhibit a conserved motif capable to interact with ATP and analogs, and participate in response to a large variety of stimuli such as oxidative stress, heat shock, nutrient starvation, and UV radiation, suggesting an important role regarding the cellular homeostasis. Due to the intrinsic nature of the host-parasite relationship, different types of stresses (biotic and abiotic) are present during the parasite lifecycle, indicating that USPs could be critical for the survival and permanence in its host. Thus, the objective of this work is to perform a phylogenetic and evolutionary divergence analysis using comparative genomics and methods to infer functional divergence in species of the phylum Platyhelminthes. Moreover, we propose the functional characterization of USPs by producing an *Echinococcus ortleppi* (Cestoda, Platyhelminthes) USP as a recombinant protein (rEoUSP-1) in *Escherichia coli*, and performing oligomerization and ATP-binding and hydrolysis assays. Our work identified that USPs genes are organized in clusters, have pseudogenes, and exhibits large functional divergence, specifically in residues near the ATP-binding domain. These data indicates that USPs follow a birth and death model of evolution, where some genes carry deleterious mutations producing pseudogenes, and other maintain different functions (sub/neofunctionalization). On the other hand, we verified that rEoUSP-1 may bind ATP. The ATPase activity was detected only at higher protein concentrations, suggesting that other cellular components may be needed to achieve this function. The rEoUSP-1 forms dimers, a similar result also observed on literature. In conclusion, the functional divergence of USPs may have implications in the response to different stimuli, in the interaction with several types of ligand, or in the specific spatiotemporal gene expression. Additionally, functional analysis of rEoUSP-1 contributes to a better understanding of these proteins both in relation to eukaryotic species, and regarding the host-parasite interface.

Introdução

Platelmintos como modelos biológicos e agentes etiológicos de doenças parasitárias

Os platelmintos são organismos triblásticos, acelomados, e possuem simetria bilateral. A posição filogenética deste grupo tem sido motivo de muita controvérsia, justamente por serem membros basais dos Bilateria e dos Protostomados (ou derivados dos Protostomados) e possuírem características anatômicas chave (mesoderme, sistema nervoso central, etc.) que proporcionaram o molde para a evolução de órgãos e tecidos complexos e altamente organizados encontrados em organismos superiores (Newmark e Sánchez Alvarado, 2002).

De forma geral, o grupo pode ser dividido em organismos que apresentam um estilo de vida estritamente parasítico e aqueles de vida livre. As planárias, componentes deste último, têm sido alvo de muitos estudos principalmente pelas características do desenvolvimento (células tronco) e sua conhecida capacidade de regeneração através de células nomeadas de neoblastos (Nimeth *et al.*, 2007; Rink *et al.*, 2013; Collins *et al.*, 2013). Embora as espécies modelo como *Drosophila melanogaster* e *Caenorhabditis elegans* foram amplamente utilizadas para estudos de embriogênese, elas possuem uma limitada capacidade de regenerar seus tecidos. Outros organismos utilizados em estudos de regeneração, como os anfíbios urodelos, precisam de mais de um mês para a regeneração completa, possuem longos ciclos de vida, e apresentam genomas muito grandes que dificultam as análises moleculares. Assim, as planárias apresentam diversas vantagens para sua utilização como modelos do desenvolvimento e regeneração de tecidos em estudos que envolvem genômica funcional (Newmark e Sánchez Alvarado, 2002; Collins, 2017).

Apesar de que a utilização de parasitos platelmintos como espécies modelo para pesquisas em células tronco já tenham sido avaliadas (Collins e Newmark, 2013; Collins *et al.*, 2013; Wang *et al.*, 2013; Koziol *et al.*, 2014; Wendt e Collins, 2016), estes organismos são mais conhecidos pelo impacto negativo que causam na saúde pública e pecuária, sendo as classes Cestoda e Trematoda as principais representantes do grupo (Caira e Littlewood, 2013; Estatísticas da Saúde Mundial 2017, Organização Mundial da Saúde (OMS)). Os parasitos do gênero *Echinococcus* e *Taenia* (cestódeos) e *Schistosoma* (trematódeo) são os agentes etiológicos da equinococose, cisticercose, e esquistossomose, respectivamente, as quais formam parte da lista das 17 doenças tropicais negligenciadas (NTD, *neglected tropical disease*) que são priorizadas para o controle e eliminação, segundo a OMS. A esquistossomose afeta ao redor de 240 milhões de pessoas, a maioria na África (>90%), e mais de 700 milhões correm risco de infeção por viver em locais endêmicos. Por outro lado, a equinococose acontece em pouco mais de 1 milhão de pessoas, porém, em áreas hiperendêmicas (Argentina, Uruguai, China, Rio Grande do Sul no Brasil, etc.) o risco de contrair a doença é alto, com uma prevalência de 2-10% em humanos e 20-90% no gado, gerando um custo estimado de US$ 3 bilhões no tratamento dos casos e nas perdas na indústria pecuária (Budke *et al.*, 2006). Apesar da enorme diversidade de parasitos platelmintos, as drogas utilizadas para combater estas doenças são extremamente limitadas. Praziquantel, albendazol (ou mebendazol), ivermectina, ou a combinação de algum deles, são praticamente as únicas drogas disponíveis de forma comercial não só para o tratamento de cestodíases e trematodíases, senão também contra helmintos em geral (Estatísticas da Saúde Mundial 2017, OMS).

Na era genômica, a procura de alvos vacinais utilizando proteínas recombinantes tem sido bastante estudada. Atualmente, os antígenos recombinantes EG95 de *Echinococcus granulosus* e TSOL18 de *Taenia solium* tem dado resultados promissores para o controle das doenças transmitidas por estes organismos no seu hospedeiro intermediário (Jayashi *et al.*, 2012; Larrieu *et al.*, 2015). Porém, devido a possíveis mecanismos de resistência, a busca de alvos vacinais alternativos e de maior abrangência (hospedeiro definitivo, humanos), tanto para cestodíases como trematodíases, é extremamente necessária (Hewitson e Maizels, 2014). Considerando a relevância dos parasitos do gênero *Echinococcus* no sul do Brasil, aspectos relacionados ao número de espécies e ciclo de vida, serão aprofundados a seguir.

Ciclo de vida de E. granulosus e agentes de estresse

Até o momento, o gênero *Echinococcus* apresenta nove espécies: *Echinococcus granulosus* sensu stricto (G1-G3), *Echinococcus equinus* (G4), *Echinococcus ortleppi* (G5), e *Echinococcus canadensis* (G6-G10), os quais compõem o complexo *Echinococcus granulosus* sensu lato, e as espécies *Echinococcus multilocularis*, *Echinococcus vogeli*, *Echinococcus oligarthra*, *Echinococcus shiquicus*, e *Echinococcus felidis* (Nakao *et al.*, 2013). Em termos de relevância sanitária e impacto socioeconômico (Budke *et al.*, 2006; Torgerson *et al.*, 2010), *E. granulosus* sensu stricto (G1), causador da equinococose cística (distribuição mundial) e *E. multilocularis* responsável pela equinococose alveolar (restrita ao hemisfério norte) são as espécies mais importantes e mais estudadas do grupo. No Brasil, a grande prevalência da equinococose cística ocorre em áreas rurais do sul do país, especificamente aquelas destinadas à criação do gado (bovino, ovino, porcino, etc.) (Balbinotti *et al.*, 2012). O uso das vísceras do gado como fonte de alimento para canídeos, as condições de sanidade limitadas nestas áreas e a pouca preocupação do governo para atender este tipo de doenças, contribuem para a permanência do parasito, e assim, dificultam o controle e eliminação da doença nestas regiões.

E. granulosus possui um ciclo de vida que envolve dois hospedeiros (Figura 1). No hospedeiro definitivo (canídeos) ocorre a forma adulta do parasito (verme segmentado), que se localiza no intestino e possui inúmeros ovos na sua última proglótide (proglótide grávida). Junto com as fezes, os ovos são liberados ao meio externo. Quando ingerido por via oral pelo hospedeiro intermediário, a oncosfera eclode, atravessa a parede intestinal e é transportada pelo sistema circulatório, chegando preferencialmente no fígado e/ou pulmão onde se desenvolve no cisto hidático. Estes cistos são uniloculares, cheios de líquido, e compostos de três camadas. A camada mais externa, chamada adventícia, é produzida pelo hospedeiro como resposta à infeção. Seguidamente vem a camada laminar que é acelular, rica em carboidratos, e produzida pela camada germinativa (camada mais interna do parasito). Está última produz também os protoescólices, que são as formas pré-adultas do parasito. Quando as vísceras são ingeridas pelo hospedeiro definitivo, os protoescólices são ativados pelos sais biliares, baixo pH, entre outros estímulos e conduzidos até o intestino, onde começará a se desenvolver a forma adulta do parasito, fechando assim o ciclo (McManus e Smyth, 1986; Thompson, 1995) (Figura 1).

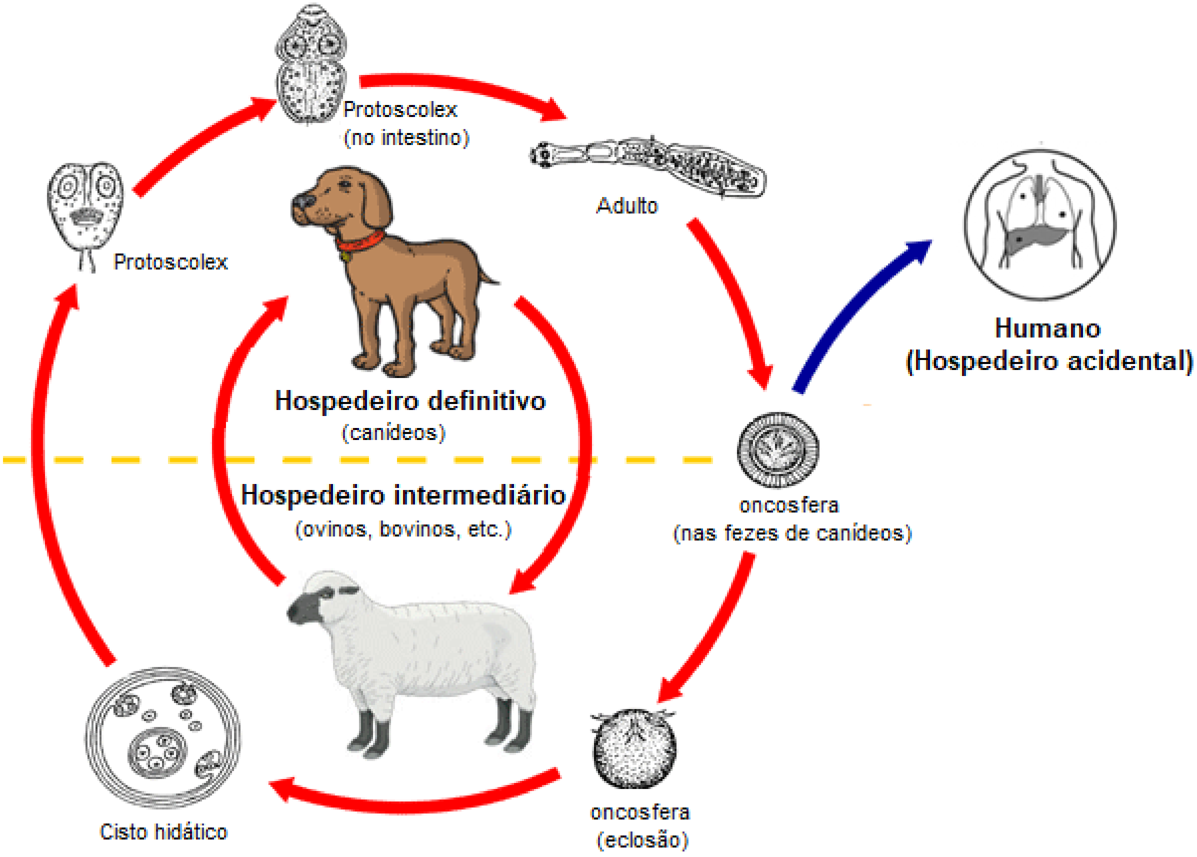


Figura 1. Ciclo de vida de *E. granulosus.* O verme adulto desenvolve-se no trato gastrointestinal do hospedeiro definitivo (HD, canídeos). Quando os ovos do verme adulto são eliminados pelas fezes do HD e são ingeridos pelos hospedeiros intermediários (HI, ungulados domésticos), pode ocorrer o desenvolvimento da fase larval ou metacestódeo. Nesta fase ocorre a formação do cisto hidático, onde as formas pré-adultas (protoescólices) do parasito começam a se desenvolver. Se as vísceras infectadas do HI são ingeridas pelo HD, começa o desenvolvimento da forma adulta, completando desta maneira o ciclo. O homem pode acidentalmente infectar-se ao entrar em contato com os ovos do parasito eliminados pelas fezes do HD. Do mesmo modo, o cisto hidático é formado. Modificado do Centro para o Controle e Prevenção de Doenças Infecciosas (https://www.cdc.gov/parasites/echinococcosis/biology.html)

Como exposto anteriormente, o ciclo de vida de *Echinococcus* envolve a interação com diferentes hospedeiros e com o meio ambiente externo, enfrentando diversos tipos de estresses tanto bióticos como abióticos. Vários estudos têm sido realizados sobre o impacto das variações do meio ambiente na viabilidade e sobrevivência de parasitos cestódeos e trematódeos. Mudanças na temperatura, luz, pressão osmótica, velocidade, e movimento da água, interferem tanto na eclosão dos ovos e infeção dos miracídios de *Schistosoma mansoni* no seu hospedeiro intermediário (caracóis), como à prevalência deste último nas águas (Upatham, 1973; Mccreesh e Booth, 2014). Da mesma forma, o impacto da temperatura e umidade relativa em ovos de *E. multilocularis*, *E. granulosus,* e de espécies do gênero *Taenia* também já tem sido analisado (Coman, 1975; Veit *et al.*, 1995; Sánchez Thevenet *et al.*, 2017). Independentemente da umidade relativa, temperaturas abaixo de -70 e acima de 40-60 °C têm um efeito letal nos ovos de *E. multilocularis* e impedem o desenvolvimento do metacestódeo no hospedeiro intermediário (Veit *et al.*, 1995). Porém, ovos mantidos a 2 °C em água por 2 anos e meio foram capazes de infectar com sucesso seu hospedeiro (Thomas e Babero, 1956), o que sugere a ampla resistência a baixas temperaturas neste estágio de vida livre do parasito.

O número de estresses bióticos parece ser consideravelmente maior devido às complexas interações com os hospedeiros. Uma vez ingerido pelos hospedeiros, a oncosfera e a forma pré-adulta do parasito devem enfrentar vários tipos de estresses (contato com sais biliares, enzimas digestivas, etc.), além dos mecanismos de defesa do hospedeiro, até atingir seu órgão alvo. A análise do transcritoma da oncosfera e do estágio inicial do desenvolvimento do metacestódeo de *E. multilocularis*, mostrou que os genes que codificam as proteínas HSP20, HSP70, e o antígeno II/3 possuem expressão constitutiva e estão expressos de forma abundante. Por outro lado, os antígenos gp50 e EG95 (alvo vacinal) foram mais expressos na oncosfera, enquanto o antígeno B esteve mais representado no metacestódeo (Huang *et al.*, 2016). Estudos de microarranjos realizados com o intestino do hospedeiro intermediário 4 dias após a infecção com ovos de *E. granulosus*, mostraram a regulação positiva de genes relacionados à resposta inata celular e molecular, por exemplo, KIR2DS1 e KLRJ1, os quais ativam a expressão de células *natural killer*; MCP1, MCP3, FcεRI, relacionados à ativação de mastócitos; e IgM produzida por células B1 (Hui *et al.*, 2015). A infecção pelo *E. granulosus* promove a diferenciação de monócitos em células dendríticas, onde os componentes do líquido hidático (entre eles, o antígeno B) apresentam um papel relevante neste processo (Riganò *et al.*, 2007). Na fase crônica da hidatidose cística, é abundante a expressão das imunoglobulinas IgG, IgM, e IgE, sendo que essa característica tem servido para o desenvolvimento de testes de sorodiagnósticos (Zhang *et al.*, 2012).

Embora algumas etapas do ciclo de vida do parasito podem ser consideradas como estímulos para o desenvolvimento e não explicitamente como agentes de estresse, por exemplo, o pH baixo e as enzimas digestivas presentes no trato digestivo do hospedeiro definitivo, fica claro que a grande quantidade de mecanismos de defesa imposta pelos hospedeiros e as condições do meio ambiente externo (principalmente no estágio de oncosfera), geram uma forte barreira que o parasito deve atravessar para atingir seus órgãos alvos e garantir sua sobrevivência.

Genomas de platelmintos parasitos e sua relação com o sistema adaptativo

As interações parasito-hospedeiro promovem o surgimento de novas estratégias de adaptação. Estas novidades adaptativas podem ser facilmente detectadas através do sequenciamento de genomas e/ou estudos de genômica comparativa.

Na década passada foram obtidos os primeiros genomas de platelmintos parasitos, representados pelos trematódeos *S. mansoni* e *Schistosoma japonicum* (Berriman *et al.*, 2009; Consortium, 2009). Em seguida, várias outras espécies parasíticas deste filo tiveram seus genomas sequenciados: *Clonorchis sinensis* (Wang *et al.*, 2011), *Schistosoma haematobium* (Young *et al.*, 2012), *E. multilocularis*, *E. granulosus*, *T. solium*, *Hymenolepis microstoma* (Tsai *et al.*, 2013; Zheng *et al.*, 2013), *Opistorchis viverrini* (Young *et al.*, 2014), *Fasciola hepática* (Cwiklinski *et al.*, 2015; McNulty *et al.*, 2017), *Taenia asiatica, Taenia saginata* (Wang *et al.*, 2016), e *Echinococcus canadensis* (Maldonado *et al.*, 2017). Os genomas destas espécies, entre outros platelmintos parasitos cujos genomas encontram-se parcialmente sequenciados (alguns ainda não publicados), estão disponíveis *online* no banco de dados da WormBase ParaSite (Howe *et al.*, 2016; Howe *et al.*, 2017), o que permite realizar diversas análises comparativas a partir de um grande número de sequências de diferentes organismos. Estes trabalhos permitiram conhecer diversos aspectos da estrutura e organização gênica, como também identificar um grande número de expansões e perdas de genes relacionados ao parasitismo.

De forma geral, os genomas possuem entre 10.000-12.000 genes, e apresentam um grande número de regiões repetitivas (10-40%) e micro-éxons (75% da região codificante em *S. mansoni*) (Berriman *et al.*, 2009; Consortium, 2009; Tsai *et al.*, 2013). As regiões repetitivas são mais abundantes em trematódeos, o que pode estar relacionado ao maior número de retrotransposons neste grupo. Por outro lado, os micro-éxons (MEG, *micro-exon genes*) dão origem a transcritos curtos (3-36nt), capazes de gerar diversas variantes através de *splicing* alternativo, que codificam pequenas proteínas que geralmente são secretadas (possuem peptídeo sinal). Além disso, MEGs apresentam regulação positiva quando localizados no hospedeiro mamífero, tornando-os alvos terapêuticos interessantes para o controle da esquistossomose (Berriman *et al.*, 2009; Consortium, 2009; Cwiklinski *et al.*, 2015).

Ao longo da história evolutiva, os platelmintos parasitos adquiriram diversas estratégias de adaptação aos seus hospedeiros. A seguir, são citadas algumas características mais relevantes em relação à biologia molecular desta interação.

Variabilidade genética

Uma das características mais importantes relacionadas à distribuição global dos parasitos é a sua capacidade de colonizar e se adaptar a diversos tipos de hospedeiros (espécies especialistas). Um exemplo claro destas relações são as espécies do gênero *Echinococcus*. Como descrito anteriormente, o gênero é composto, até o momento, de nove espécies diferentes. A pesar da estreita proximidade filogenética entre *E. canadensis* e *E. granulosus*, ambos pertencentes ao complexo *E. granulosus* sensu lato, estudos de genômica comparativa mostraram a presença de um grande número de polimorfismos de nucleotídeo único (SNP, *single nucleotide polymorphism*) entre estas espécies, inclusive maior do que entre *E. granulosus* e *E. multilocularis*, salientando a importância deste tipo de mutações na diferenciação genética das espécies (Maldonado *et al*., 2017). A relevância dos SNPs também se destaca em trematódeos. Em *F. hepatica* foi identificado um grande número de polimorfismos não sinônimos em genes relacionados aos processos de quimiotaxia e desenvolvimento neural, como caderinas, semaforinas, fascilinas, entre outros, os quais contribuem para a localização, colonização e migração dentro do tecido do hospedeiro (Cwiklinski *et al*., 2015).

Além dos SNPs, uma alta taxa de duplicações gênicas e diversificação funcional foi identificada entre espécies do gênero *Taenia*. A grande variabilidade do genoma e a rápida evolução adaptativa, principalmente em genes relevantes na interação parasito-hospedeiro (homeostasia, captação de nutrientes, etc.), contribuiu para a recente especiação entre *T. saginata* e *T. asiatica*, levando à conquista de um novo hospedeiro intermediário e a um novo sítio de infeção (hepatotropismo) nesta última (Wang *et al*., 2016). Entre outros mecanismos que promovem variabilidade, os processos de retrotransposição (de maior relevância em trematódeos) e a diversidade dos MEGs, são elementos que também podem estar relacionados ao sucesso do estilo de vida parasítico. Estes exemplos ilustram a importância do conhecimento da variação genética para a diferenciação das espécies, e seu possível impacto nas funções biológicas relacionadas ao parasitismo.

Metabolismo energético

Parasitos cestódeos e trematódeos apresentam as vias completas para os processos de glicólise, o ciclo do ácido tricarboxílico, e ciclo das pentoses fosfato. Porém, mecanismos como a síntese de esteróis e lipídios (como também de aminoácidos, purinas) estão ausentes (Berriman *et al*., 2009; Tsai *et al*., 2013). Assim, a captação destas moléculas do hospedeiro tornou-se uma necessidade básica para a sobrevivência do parasito. Proteínas de ligação a ácidos graxos (FABP, *fatty acid binding proteins*) e o antígeno B estão entre os genes mais expressos na interação parasito-hospedeiro (Obal *et al*., 2012). Embora tenham sido alvo de estudo por bastante tempo, ainda pouco se conhece sobre os mecanismos moleculares da interação, carregamento, e transporte dos lipídeos por estas proteínas. Trabalhos recentes utilizando linhagens celulares de pulmão e fígado de mamífero indicam que o antígeno B é internalizado através de vias de endocitose, principalmente através de *rafts* lipídicas (da Silva *et al.*, submetido). Além disso, a ausência de vários genes relacionados ao peroxissomo, local onde acontece a oxidação dos ácidos graxos, se correlaciona com a perda da síntese e processamento destas moléculas (Tsai *et al.*, 2013). A presença de enzimas chave na síntese *de novo* de lipídeos (e purinas) em *Schmidtea mediterranea*, platelminto de vida livre, indica que nos parasitos a captura destas moléculas do hospedeiro estaria associada à adoção do estilo de vida estritamente parasítico (Robb *et al.*, 2015).

Mecanismos de defesa

Como descrito anteriormente, a forma larval patogênica das espécies pertencentes ao gênero *Echinococcus* possui uma camada laminar rica em carboidratos, sintetizada pela camada germinativa adjacente. Além de manter a turgência do metacestódeo, a camada laminar é descrita como elemento chave na imunomodulação da resposta imune do hospedeiro (Díaz *et al.*, 2011). Esta camada é composta de proteínas glicosiladas (mucinas) que formam lâminas com diferentes graus de compactação e que pode chegar a medir até 3 mm de espessura (Díaz *et al.*, 2011; Díaz *et al.*, 2015). Além disso, a expansão gênica de enzimas que atuam nos processos de glicosilação (fucosil e xilosil-transferases), sugere um papel crucial na subversão do sistema imune em *Schistosoma*. Enzimas proteolíticas (~2,5% do proteoma) foram descritas, algumas delas como a catepsina B consideradas como potenciais alvos de droga. Proteínas kunitz, serpinas, e cistatinas, compõem os principais inibidores de proteases. Expansões gênicas também foram identificadas nas tetraspaninas, proteínas estruturais localizadas principalmente no tegumento e que atuam na resposta imune (Berriman *et al*., 2009; Tsai *et al*., 2013).

Desta maneira, fica evidenciado o amplo repertório de moléculas envolvidas na defesa contra o hospedeiro, algumas delas (antígeno B) agindo tanto no transporte de lipídeos como na imunomodulação da resposta imune.

Famílias gênicas como objeto de estudo

Uma família gênica pode ser definida como um grupo de genes que surgiu de um gene ancestral comum e que possuem sequências e, em geral, funções similares. Este conceito pode ser aplicado tanto a genes de um mesmo genoma originados por duplicação gênica (parálogos), como também a genes relacionados de diferentes genomas gerados por processos de especiação (ortólogos) (Ohno, 1970; Fitch, 1970). Um dos benefícios de agrupar genes com certo grau de similaridade é que estes *clusters* podem levar à similaridade estrutural do produto gênico, e subsequentemente, a uma similaridade funcional. Assim, os genes com sequência, estrutura e função conhecidas podem ser usados para inferir a função de outros genes de organismos diferentes através do uso da genômica comparativa. Com esta última ferramenta, também podem ser inferidos os processos evolutivos que moldaram a adaptação e diversidade dos genes envolvidos.

Um dos mecanismos mais importantes para o surgimento e expansão de famílias gênicas são os processos de duplicação. Uma nova cópia de um gene fornece a matéria prima para as novidades evolutivas, permitindo a geração de mutações sem que a função original do gene progenitor seja afetada (Ohno, 1970). Os processos de duplicação podem ser divididos com base no tamanho do fragmento duplicado, e se teve ou não a participação de um intermediário de RNA (Figura 2).

A retrotransposição acontece quando o transcrito (mRNA) do gene progenitor sofre transcrição reversa e a inserção no genoma. De forma geral, estas retrocópias carecem de íntrons e de regiões regulatórias, e por este último motivo poucas vezes são expressos (Figura 2A). A duplicação em tandem de segmentos de DNA pode ser consequência do cruzamento desigual das cromátides irmãs durante a meiose (Figura 2A) ou por mecanismos aleatórios que não envolvem homologia. Os elementos repetitivos como as sequências *Alu* em humanos, podem favorecer a proliferação das duplicações em tandem (Figura 2A). Menos frequentemente, os erros, mutações e outros tipos de rearranjos durante a replicação do DNA, também podem levar à duplicação continua de segmentos de DNA.

De forma geral, os genes duplicados podem seguir três destinos diferentes: pseudogenização, se surgirem mutações deletérias; neofuncionalização, se a função do novo gene for diferente à do gene progenitor; ou subfuncionalização, se a função do gene progenitor for dividida entre ele e o gene duplicado (Figura 2B). Pseudogenes são descritos como genes não funcionais, geralmente com ausência da expressão gênica, e originados de forma geral por mutações que afetam a transcrição (por exemplo, nos motivos do DNA nas regiões regulatórias, ou códons de parada na região codificadora). Genes com funções novas ou com características similares ao gene que deu origem (por exemplo, expressão num tecido especifico do organismo), geralmente estão associados a mutações nas regiões reguladoras ou mutações não sinônimas na região codificadora do gene que possam conferir alguma vantagem adaptativa para o organismo (aumento do *fitness*) (Hurles, 2004). Pelo fato de ser um mecanismo capaz de gerar novas e diversas funções, e que essas funções poderiam conferir uma vantagem adaptativa, o estudo de ganho e perda de genes dentro de uma família gênica constituem uma abordagem extremamente importante nas interações do tipo parasito-hospedeiro.

Proteínas de estresse universal

O sequenciamento de genomas permite conhecer o repertório completo de produtos gênicos gerados pelo organismo. Porém, uma porção considerável dessas identificações são classificadas como proteínas hipotéticas, ou seja, sem uma função determinada. Dessa forma, utilizando cristalografia de raio X, foi descrita pela primeira vez em 1998 a estrutura de uma proteína de estresse universal (USP, *universal stress protein*) na bactéria *Methanococcus jannaschii*, cujo genoma tinha sido sequenciado dois anos antes (Bult *et al.*, 1996; Zarembinski *et al.*, 1998). O estudo mostrou que o monômero USP possui uma estrutura secundária do tipo alfa/beta/alfa, com um motivo conservado de ligação a ATP (e análogos) que segue a sequência de aminoácidos [Gx2Gx9G(S/T)], e que é diferente à previamente descrita para outras proteínas que ligam nucleotídeos (Zarembinski *et al*., 1998). Nessa mesma década, uma proteína citoplasmática de *Escherichia coli* chamou a atenção dos pesquisadores pelo fato de ter sua expressão aumentada quando o crescimento da bactéria era inibido pela carência de qualquer um de diversos nutrientes (carbono, nitrogênio, fosfato, sulfato, amino ácidos essenciais) ou frente a agentes tóxicos (metais pesados, oxidantes, ácidos e antibióticos) (Van Bogelen *et al*., 1990; Nyström e Neidhardt, 1992).

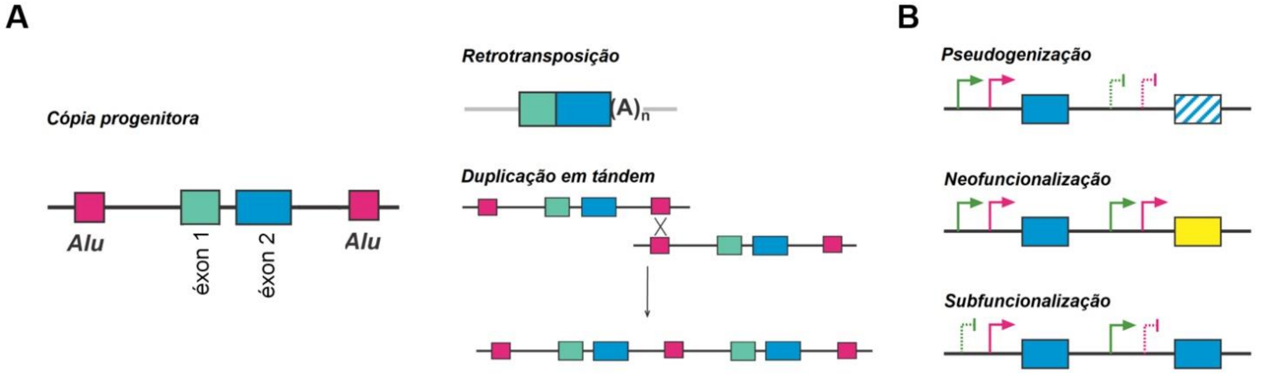


Figura 2. Mecanismos de duplicação gênica e destinos evolutivos dos genes duplicados. A) Um gene progenitor composto por dois éxons e flanqueado por elementos *Alu* pode sofrer retrotransposição do mRNA, o que leva a uma inserção do parálogo (sem íntrons) num local diferente no genoma. Por outro lado, a recombinação entre elementos *Alu*, assim como erros durante a replicação do DNA (não representado), conduz a eventos de duplicação em tandem. **B)** Um gene (caixa azul) com dois promotores (setas verde e rosa) que sofreu uma duplicação, pode ser inativado e/ou degradado (pseudofuncionalização), ou apresentar uma função nova (neofuncionalização). Também, o padrão de expressão do gene original pode ser particionado entre as cópias duplicadas devido ao silenciamento de um dos dois promotores de forma complementar (subfuncionalização). Modificado de Hurles, 2004.

A partir destes trabalhos, muitos outros estudos começaram a associar a presença de proteínas com o domínio USP (número pfam PF00582) a diversos tipos de estresse, como luz ultravioleta, mudanças de temperaturas, estresse oxidativo, carência de nutrientes, dano ao DNA, etc. A estrutura tridimensional, a capacidade de ligação e hidrólise de nucleotídeos, e os níveis de expressão gênica frente a múltiplos tipos de estresse, têm sido bastante analisadas na literatura, porém, a maior parte das pesquisas concentram-se em bactérias, archeas e plantas (Sousa e Mckay, 2001; Gustavsson *et al.*, 2002; Hingley-Wilson *et al.*, 2010). Apesar da evidente associação entre as

USPs e a resposta a estresses abióticos e bióticos (por exemplo, infecção por microrganismos) (Jung *et al.*, 2015; O Connor e Mcclean, 2017), os mecanismos de ação e as vias metabólicas em que as USPs estão envolvidas ainda são pouco conhecidas.

As USPs compõem uma família gênica, portanto, estão sujeitas à expansão e perda de genes através de processos de duplicação seguido de divergência funcional. Amplamente distribuídas nos organismos, o número de cópias é muito variável entre as espécies, desde 1-8 até 40 ou 80 cópias, estas últimas provavelmente por eventos de duplicação do genoma completo (Forêt *et al.*, 2011; Wasik *et al.*, 2015). A ausência de USPs na maioria dos deuterostomados e alguns protostomados (ecdisozoos), pode ser consequência de redundância funcional ou a presença de funções taxa-especifica (Fôret *et al*., 2011). Da mesma forma, mutantes *knock-out* de USPs em *Mycobacterium tuberculosis* não apresentaram diferenças no crescimento e sobrevivência entre bactérias crescidas em condições normais ou de hipóxia, o que sugere que as funções poderiam ser parciais ou completamente redundantes (Hingley-Wilson *et al*., 2010). Por outro lado, mutantes *knock-out* em *E. coli* mostraram sensibilidade ao estresse oxidativo (*uspA*, *uspD*) ou a participação dos genes na motilidade (*uspE*, *uspC*) e adesão celular (*uspF*, *uspG*), apresentando também uma superposição de funções (Nachin *et al.*, 2005). Assim, considerando a diversidade de funções, e a ausência desta família gênica em grandes grupos taxonômicos (além da redundância funcional obtida em testes experimentais), as USPs parecem ter seguido diferentes caminhos evolutivos ao longo da história. Compreender esta heterogeneidade através de sequências de DNA e aminoácidos, e as implicações que podem ter estas variações na função das proteínas, é um assunto pouco explorado nas USPs, e totalmente desconhecido em platelmintos.

Estruturalmente, as USPs se caracterizam pelo dobramento alternado de folha beta e alfa hélice e um motivo de ligação a nucleotídeos que segue a sequência consenso [G2xG9xG(S/T)]. Com base na homologia com a UspA de *E. coli*, as USPs podem ser classificadas como proteínas que interagem ou não com o ATP. A presença de modificações na sequência consenso do motivo proteico parece estar relacionada com a capacidade ou não de ligar, e eventualmente hidrolisar, o ATP (Zarembinski *et al*., 1998; Sousa e McKay, 2001; Weber e Jung, 2006). Embora estas associações ainda requerem de estudos mais aprofundados, os dois tipos de USPs parecem ter um papel relevante na resposta ao estresse (Gustavsson *et al*., 2002; Boes *et al.*, 2006; Jung *et al*., 2015). O monômero USP possui uma massa molecular de aproximadamente ~18 kDa, sendo capaz de formar complexos maiores (dímeros, trímeros, tetrâmeros) (Weber e Jung, 2006; Nachin *et al.*, 2008; Drumm *et al.*, 2009; Jung *et al.*, 2015). Além da formação de homo e heteroligômeros, as USPs podem-se associar a outras proteínas, como anexina, NADH oxidase, e kinase em plantas, ou o domínio *forkhead* do transportador ABC (*ATP-binding cassette*, ABC) em *M. tuberculosis*. Em plantas estas interações estão associadas ao estresse oxidativo ou à seca, enquanto em bactérias foi relacionada ao crescimento (Loukehaich *et al.*, 2012; Gonzali *et al.*, 2015; Glass *et al.*, 2017; Gutiérrez-Beltrán *et al.*, 2017). Até o momento, existem poucos trabalhos baseados na caracterização funcional de USPs em eucariotos, sendo quase restritos a plantas. Além disso, o repertório de proteínas associadas às USPs é uma área pouco explorada e que permitirá conhecer as diferentes rotas metabólicas nas quais estas proteínas estão envolvidas no contexto geral.

Na presente tese, através do uso da genômica comparativa, análises filogenéticas, e variabilidade genética (seleção positiva, divergência funcional), buscamos entender a estrutura e organização das USPs em platelmintos, procurando identificar prováveis destinos evolutivos para esta família gênica. Além disso, análises de interações proteína-ligante e proteína-proteína após a previa produção de uma proteína recombinante de *E. granulosus*, permitirá conhecer o repertório de proteínas que interagem com a USP e inferir sobre as funções celulares em escala global.

Objetivos

Objetivo Geral

Compreender a diversidade das USPs de platelmintos desde um ponto de vista evolutivo, e conhecer o papel biológico destas proteínas na interação parasito-hospedeiro utilizando *Echinococcus* como modelo experimental.

Objetivo Específicos

Analisar a estrutura e distribuição das *USPs* em espécies do filo platelmintos.

Estudar os mecanismos de divergência funcional e seu impacto na evolução da família gênica.

Realizar a clonagem, expressão, e purificação de uma USP de *E. ortleppi* para posterior caracterização funcional.

Verificar a interação da USP recombinante com os nucleotídeos AMP e ATP, e testar a capacidade da enzima de hidrolisar este último.

Conhecer o repertorio de proteínas que interagem com a USP recombinante através de ensaios de *cross-linking* químico e cromatografia de afinidade.

Capítulo I

Evolutionary fates of universal stress protein paralogs in Platyhelminthes[†]

Sergio Martin Espinola[1,2], Martin Pablo Cancela[2,3], Lauís Brisolara Corrêa[1], Arnaldo Zaha[2,3*]

[1] Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil

[2] Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil

[3] Programa de Pós-Graduação em Biologia Celular e Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil

E-mail: sergio@cbiot.ufrgs.br, Sergio Martin Espinola
martin@cbiot.ufrgs.br, Martin Pablo Cancela
lauisbrisolara@gmail.com, Lauís Brisolara Corrêa
zaha@cbiot.ufrgs.br, Arnaldo Zaha

[*]Author for correspondence

[†]*Artigo submetido à revista BMC Evolutionary Biology.*

Abstract

Background: Universal stress proteins (USPs) are present in all domains of life. Their expression is upregulated in response to a large variety of stress conditions. The functional diversity found in this protein family, paired with the sequence degeneration of the characteristic ATP-binding motif, suggests a complex evolutionary pattern for the paralogous USP-encoding genes. In this work, we investigated the origin, genomic organization, expression patterns and evolutionary history of the USP gene family in species of the phylum Platyhelminthes.

Results: Our data showed a cluster organization, a lineage-specific distribution, and the presence of several pseudogenes among the USP gene copies identified. The absence of a well conserved -*CCAATCA*- motif in the promoter region was positively correlated with low or null levels of gene expression, and with amino acid changes within the ligand binding motifs. Despite evidence of the pseudogenization of various USP genes, we detected an important functional divergence at several residues, mostly located near sites that are critical for ligand interaction.

Conclusions: Our results provide a broad framework for the evolution of the USP gene family, based on the emergence of new paralogs that face very contrasting fates, including pseudogenization, subfunctionalization or neofunctionalization. This framework aims to explain the sequence and functional diversity of this gene family, providing a foundation for future studies in other taxa in which USPs occur.

Keywords: stress responsive proteins, flatworms, evolutionary patterns, pseudogenes, functional divergence.

Background

The emergence of gene families is based on successive events of gene duplication. Duplicate copies can result from unequal crossing-over during meiosis or from retrotransposition processes [1]. While a crossing-over mismatch can generate a duplication of the entire gene structure, including promoter regions, introns and exons, retrotransposition events usually result in an intronless gene composed only of the exons of the ancestral gene, and giving rise to a single transcript. For each new duplicate copy, several outcomes are possible. First, a neofunctionalization process, where the new gene takes on a new function, different from that of the parental gene. Second, a subfunctionalization process, where the new copy preserves its function, but with a singular spatio-temporal regulation (e.g., expression in a specific tissue and at a specific developmental stage). Third, a pseudogenization process, where the duplicate copy accumulates deleterious mutations, leading to loss of function [2-4].

Members of the universal stress protein (USP) gene family are found in bacteria, archaea, and eukaryotes and are composed of a variable number of copies due to lineage-specific expansions [5]. These proteins are highly expressed in response to a large variety of stress conditions, such as oxidative stress, heat shock, and UV exposure [6-8]. In addition to stress resistance, they participate in the regulation of cell growth and host infection in *Mycobacterium tuberculosis [9]*, and contribute to cell adhesion and motility in *Escherichia coli* [7]. The USP protein domain exhibits a protein motif capable to interact with ATP, ADP, AMP, GTP, etc. [9, 10]. In some USPs, the amino acid sequence making up this motif is partially or completely degenerated [11]. In general, it was observed that almost all USPs crystals with the typical ATP-binding motif were solved with ATP or an ATP analog, while for those USPs where this motif is completely degenerated, neither ligand nor ion binding was observed [11]. Although different functions have been identified for USPs with typical and degenerated ATP-binding motifs [6, 8, 12], the functional impact of the amino acid substitutions at sites involved in ligand interaction remains poorly understood.

The Platyhelminthes include some harmful parasite species with considerable negative effect on public health, especially in developing countries [13] (World Health Organization, WHO, 2016). Several of the so-called neglected tropical diseases (NTDs), a diverse group of communicable diseases of the tropics, are caused by Platyhelminthes, including echinococcosis and schistosomiasis, which are responsible for 1200 and 11700 deaths per year worldwide, respectively [14]. The complex life cycle of parasitic Platyhelminthes involves their interaction with two or more hosts, accompanied by drastic physiological and morphological changes [15, 16]. This continuous change of microenvironments results in the exposure to a wide array of biotic and abiotic stressors [16-18]. In a recent review, the application of the USPs as novel anti-parasitic targets has been discussed [19]. USPs play an important role in the transition between different stages of the *Schistosoma* life cycle, including that between cercariae and schistosomula stages. Based on this and on the absence of this gene family in vertebrates, including humans, USPs could represent an interesting target for anti-schistosomal treatment [19].

Here, we use comparative genomics and the relationship between protein sequence variations and gene expression patterns to build a framework for the evolution of the USP gene family in the Platyhelminthes. This framework aims to explain the sequence and functional diversity of this gene family, providing a foundation for future studies in other taxa in which USPs occur.

Methods

Sample collection, genotyping and quantitative PCR

Bovine hydatid cysts were obtained from the Cooperleo Abattoir (São Leopoldo, Rio Grande do Sul, Brazil). The pre-adult stage (protoscoleces, PSC) of *Echinococcus ortleppi* was collected by hydatid cyst fluid aspiration and washed with phosphate buffered saline (PBS). Genotyping was performed on part of the cytochrome c oxidase subunit I (*cox1*) gene as previously described [20].

For quantitative PCR (qPCR) expression analysis, approximately 1.000 PSC were mixed with 0.5 mL of TRIzol reagent (Thermo Fisher Scientific) and immediately frozen in liquid nitrogen until RNA extraction. Total RNA was isolated using TRIzol according to the manufacturer's protocol. Isolated RNA was subsequently treated with RNase-free DNase I (Sigma-Aldrich) for 30 min at 25°C to remove all genomic DNA. Total RNA concentration was determined using a Qubit fluorometer (Thermo Fisher Scientific). The first strand of cDNA was synthesized from 200 ng of total RNA using M-MLV reverse transcriptase (Thermo Fischer Scientific) and an Oligo(dT)18 primer (0.5 μg/μL), following manufacturer's instructions. The final cDNA product was diluted 100-fold with nuclease-free water prior to use in qPCR experiments.

Real-time PCR was performed using an ABI Real-Time 7500 Fast PCR system (Applied Biosystems). Based on the genome of *Echinococcus granulosus*, specific primers were designed for six USP genes of *E. ortleppi* (table S1, Additional file 1), two of which are downregulated and four of which are upregulated in the pre-adult form, according to RNA-seq data [21, 22]. The reaction mixture and the qPCR cycling conditions were as described previously [23]. Control reactions without reverse transcriptase and without template were included to confirm the absence of genomic DNA and other PCR contaminants, respectively. All qPCR reactions were performed in technical and biological triplicates. The amplification efficiency was calculated using the LinRegPCR software [24]. The gene expression quantification was performed with the ΔΔCt method, using *EF-1α* as a normalizer gene [23].

USP sequence retrieval

The USP sequences of twelve Platyhelminthes species (*Gyrodactylus salaris, Macrostomum lignano, Schistosoma mansoni, Schistosoma haematobium, Opisthorchis viverrini, Clonorchis sinensis, Echinococcus granulosus, Echinococcus multilocularis, Echinococcus canadensis, Taenia solium, Hymenolepis microstoma,* and *Schmidtea mediterranea*) were extracted from the WormBase ParaSite and SmedGD databases [25, 26] using the USP Pfam code PF00582 and the keyword “universal stress protein”. Orthologs relationship were initially obtained by reciprocal BLASTn, and confirmed by the presence of monophyletic clades of each group of the orthologs in the phylogenetic trees. USP sequences were retrieved based on the following criteria: synteny, the presence of a single exon in most sequences (or the conserved position of the intron where present) [5], and phylogenetic relationships between orthologs. The identification of low homology sequences (probable pseudogenes) was achieved through a tBLASTn search (BLOSUM45). For each species, we blasted each USP protein sequence against the entire genome, applying an e-value threshold of 1e-1, allowing the alignment of low complexity regions, and using opening and extending gap penalties of 14 and 2, respectively. To avoid the recovery of spurious hits, we used similar criteria to those used in the search for orthologs, as follows: synteny, genes with a single intron or intronless, and the amino acid conservation in specific regions related to the interaction with ligands and belonging to the USP domain. The USP sequences of the molluscs *Lottia gigantea*, *Crassostrea gigas*, and *Octopus bimaculoides*, and the annelids *Helobdella robusta* and *Capitella teleta*, were retrieved from the Ensembl Genome and JGI databases [27, 28] using the Pfam code PF00582. These last species were used as outgroups in the phylogenetic analysis.

Phylogenetic trees

Phylogenetic analyses were performed using the Bayesian Inference (BI) and Maximum Likelihood (ML) probabilistic methods [29, 30]. Protein sequences were aligned with MAFFT v7 [31] using the FFT-NS-I method, and any columns containing more than 95% gaps were deleted using Gap Strip/Squeeze v2.1.0 [32]. The best substitution model for our data set was defined with the Smart Model Selection (SMS) tool incorporated in PhyML [33]. The ML tree was generated with PhyML v3.0 [29] using the aLRT-SH method for branch support. The BI tree was generated with BEAST v1.8.4 [30], using two independent runs of 50.000.000 chains and sampling at every 5.000 generations. The birth and death process [34], and the LG+G substitution model [35] with 4 gamma categories, were the priors for the analysis in BEAST v1.8.4. Other parameters (e.g. clock model) were used as default. The software TRACER v1.6 [36] was used to check the convergence of Monte Carlo Markov Chains (MCMC) and to ensure adequate effective sample sizes (ESS > 200) after the first 10% of generations were deleted as burn-in. The maximum clade credibility tree was estimated with TreeAnnotator, which is part of the BEAST v1.8.4 package, and the tree was visualized using Figtree v1.4.3 [37].

Positive selection analysis

Positive selection was tested using the codeml program incorporated in the PAML package [38] and the mechanistic empirical combination model (MEC), as implemented in Selecton 7 [39, 40]. Due to the high divergence between the USP sequences of the different Platyhelminthes classes (see Figure 1A), we limited the positive selection analysis to 63 coding sequences of the Cestoda class. Protein sequences were aligned with MAFFT v7 [31]. Subsequently, we use the program

Pal2Nal [41] to align the coding sequences corresponding the protein alignment obtained with MAFFT v7 . Considering codons (gaps in triplets), the ends of the sequences were removed manually, and columns containing more than 90% gaps were deleted using Gap Strip/Squeeze v2.1.0 [32]. The nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$) offers a sensitive measure of positively selected residues in proteins. Thus, a ω ratio greater than 1 suggest that nonsynonymous mutations would be adaptively advantageous in evolution and could be fixed in one or more populations. Thereby, lineages exposed to dissimilar pressures of selection on a given protein show differences in the ω ratio what could indicate these sites are on positive selection [42]. In order to verify whether the ω ratio deviated significantly from 1 for each alternative model considered, we performed a likelihood ratio test (LRT) on the results of the PAML run as follows: M1a (nearly neutral) versus M2a (positive selection); and M7 (beta) versus M8 (beta&ω).

DNA motif analysis of the promoter regions

In order to gain insights about the origin and regulation of the USPs, we searched for conserved patterns (DNA motifs) in the promoter region of these genes. DNA motif analysis was performed using the *mixture model by expectation maximization* (MEME) method, incorporated in the MEME suite [43]. 500 bp of the USP promoter region were extracted from the 5' end upstream of the start codon ATG. The *Saccharomyces cerevisiae database* was used to compare the identified motifs with others previously described (Tomtom) [44], and to find associations with genes linked to gene ontology terms (GOMo) [45]. *The motif search was executed with default parameters, considering a maximum width of 10 nucleotides and allowing any number of repetitions for the motifs in the sequence. All upstream sequences are available in the Additional file 2.*

Divergence analysis and USP protein modeling

Evolutionarily conserved amino acids are expected to have an important role in protein structure and function. Therefore, changes at these sites may be an indicator of functional divergence. We used the software Diverge v3.0 [46], to examine site-specific shifted evolutionary rates by calculating the coefficient of type I of divergence (Θ_I). Type I of divergence results in differing functional constraints (i.e., different evolutionary rates) between duplicated genes, regardless of the underlying evolutionary mechanisms. The null hypothesis (Θ_I=0) is assessed by the likelihood ratio test (LRT) [47], and its rejection indicates some level of functional divergence between the clusters compared. Because the output of Diverge v3.0 follows a chi-square distribution with one degree of freedom, the LRT values greater than or equal to 3.84 indicate functional divergence between clusters. Comparisons were performed between paralogs groups of approximately five sequences within the Cestoda and Trematoda classes. Using a cut-off value of

0.9 for the *a posteriori* probability, we identified amino acid sites under Type I of functional divergence.

For protein modeling, we chose the *E. granulosus* USP protein EgrG_08736, which exhibits the typical ATP-binding motif [Gx2G*x9*G(S/T)]. The 3D protein was modelled with a homologous template using Phyre v2 [48]. In addition to 3D modeling, Phyre v2 predicts ligand-binding sites and analyzes the effect of amino acid variants (Phyre Investigator). The obtained model was used to evaluate the effect of mutations at conserved sites, and to localize the amino acid residues found to be under functional divergence by Diverge v3.0. The quality of the model was assessed with ModFOLD v4.0 [49].

Results

USP gene organization in the Platyhelminthes

The high quality and completeness of several Platyhelminthes genomes (e.g. *Echinococcus* spp., *Schistosoma mansoni*) [21, 50], together with the criteria for the search for orthologs (see Methods), allowed us to locate and accurately retrieve all the DNA and protein sequences of the USP genes for each species. We found that the number of USP genes varied between Platyhelminthes species: 12 genes in *E. granulosus*, *E. multilocularis*, and *E. canadensis*, 13 in *T. solium*, 16 in *H. microstoma*, 10 in *S. mansoni* and *S. haematobium*, 18 in *C. sinensis* and *O. viverrini*, 17 in *S. mediterranea*, 6 in *G. salaris*, and 83 in *M. lignano* (Table S2, Additional file 1). The high number of USP sequences in this last species, including about 35 identical sequences, could be a consequence of an ancestral whole-genome duplication or recent large segmental duplications, as previously described [51]. To simplify, we removed the two zeros at the beginning and end of each USP identification number (ID) for the Cestoda species. Through reciprocal BLASTn, we detected orthologous relationships within the Cestoda and Trematoda classes; however, between classes, or when including the free-living flatworm *S. mediterranea* (class Turbellaria), the orthologous relationships between species become fuzzy and unrequited (Table S2, Additional file 1). In all Platyhelminthes species analyzed here, USP genes are distributed in clusters throughout the genome, with lineage-specific losses/expansions (Figure 1A). Clustering is more accentuated in the Cestoda and Trematoda than in Turbellaria (data available at the SmedGD database). The relaxed tBLASTn analysis detected three pseudogene candidates in the genus *Echinococcus*, two in *T. solium*, and one in *H. microstoma*. Synteny indicates that these pseudogenes may represent lineage-specific gene losses in *Echinococcus* spp. and *T. solium* compared with *H. microstoma*; and in *H. microstoma* compared with *H. diminuta* (Figures 1A and B). Pseudogenes are characterized by the presence of *indels* in their coding sequence, which lead to frameshift mutations and thereby generate stop codons (Figure 1B). Sequence differences between pseudogenes and their respective ortholog are highly variable for paralogous pseudogenes, reflecting an ancient pseudogenization process (Figure 1B). For the other species, there was no evidence of pseudogenes with our search strategy. Nevertheless, we identified twelve USP genes in the Trematoda, which could not be detected using the USP Pfam code. All of these were located in the vicinity of other USP genes. Two were not previously annotated (*Csin107892a*, *T265_02176a*), and one gene was re-annotated (*T265_02178*, corresponding to genes *T265_02178a* and *T265_02178b*). The other copies, which were annotated as "universal stress protein" or without description, were *Csin107891*, *Csin107892*, *Csin107893*, *Csin110039*, *Csin110041*, *T265_02177*, *T265_02179*, and *T265_02180*. All protein sequences reported in this work are available in the Additional file 2.

(Figure 1)

Phylogenetic trees and origin of the USP gene family

Phylogenetic trees (Figure 2; Figures S1 and S2 in Additional file 3) showed five genes to be shared across all Platyhelminthes species analyzed here (gene names are given for *E. granulosus*): *EgrG_09018*, *EgrG_7258*, *EgrG_10769*, *EgrG_06206*, *EgrG_20248*. Of these, *EgrG_09018* and its orthologs are the only ones with a single intron. While a few sequences from annelids (*HelroG188754*, *HelroG65703*, *HelroG194412*, *HelroG186168*, *HelroG184845*, *CapteP172559*, and *CapteP172328*) were shared with Platyhelminthes (*EgrG_09018*, *EgrG_7258*, *EgrG_10769*), no homology was found for the molluscs species (Figure 2; and Table S2, Additional file 1). The *EgrG_09018* gene probably gave rise to *EgrG_10769* and *EgrG_7258* (and by extension, their orthologs) by retrotransposition-mediated duplication. The absence of orthology between Cestoda and Trematoda retrocopies of *EgrG_09018* (and its orthologs) means that they likely represent a recent duplication event, which occurred after the split of the lineages of these classes. In the same way, many USP copies emerged independently from *EgrG_09018*, *EgrG_06206*, and *EgrG_20248*, resulting in class-specific clades for the Trematoda and Cestoda (Figure 2). Most of the USP genes that form local clusters in the genome, are also grouped together in the phylogenetic tree. This applies to *EgrG_08734*, *EgrG_08736*; *EgrG_08738*, and their orthologs in the Cestoda; *Csin107893*, *Csin107894*, *Csin107895*, and their orthologs in *O. viverrini*; and *Smp_13687* and *Smp_13689* and their orthologs in *S. haematobium* (Figures 1A and 2). Interestingly, *EgrG_08738* is phylogenetically close to *EgrG_20190* (named *EgrG_ps1* in Figure 1B). The latter one may therefore be a pseudogene that emerged from *EgrG_08738*. Along the same line, all other USP paralogs located in the same cluster (Figures 1 and 2) could have arisen from *EgrG_08738* or *EgrG_20190* by successive tandem duplications. This observation could be extended to the other Cestoda species. Based on the organization of USP genes within the genome and on their phylogenetic relationships, an array of tandem and retrotransposition duplication events might be inferred for the Trematoda and Cestoda classes. This latter mechanism seems to have played a pivotal role in the emergence of USP genes in the free-living flatworm *S. mediterranea*, whose USP gene tree exhibits a star-like topology with few ancestral genes and many locally isolated USP paralogs distributed throughout the genome.

(Figure 2)

DNA motif analysis

DNA motif analysis of the promoter regions detected the heptanucleotide -*CCAATCA*- between positions -200 and -40 upstream for almost all USP genes (Table S3, Additional file 1). This motif is a known DNA binding site for the mammalian nuclear transcription factor Y (NF-Y, HAP in *S. cerevisiae*), which promotes the initiation of gene transcription [52]. NF-Y consists of three subunits: NFA, NFB and NFC (HapB, HapC, and HapE orthologs in *S. cerevisiae*). In the presence of reactive oxygen species (ROS), oxidized HapC prevent the interaction with the HapE and HapB subunits. Consequently, the formation of the CCAAT-binding complex is abolished and their nuclear localization and regulation of target genes becomes affected [53]. Using the Pfam numbers PF02045 and PF00808, we identified the orthologs of NFYA, NFYB, and NFYC for the species studied (data available at the WormBase database). In line with this, the GOMo tool reports that the motif -*CCAATCA*- is involved in the oxidation-reduction processes as the ATP synthesis coupled to proton transport. In both Cestoda and Trematoda classes, some genes lack the conserved -*CCAATCA*-/-*CCAAT*- motif (e.g. *EgrG_02019*, *EgrG_08735*, *EgrG_09839*, and their orthologs). This could suggest a common origin for these genes, different regulation properties, or evidence of a pseudogenization process. In other genes (e.g. *EgrG_08736*, *EgrG_08738*, *EgrG_07258, EgrG_10769, EgrG_09018,* and orthologs), the -*CCAATCA*- motif occurs at the exact same position. These data provide insights about the functional diversification of the USP promoter regions and their origins by retrotransposition or tandem duplication events, as well as about the relationship between these last two.

3D protein modeling

The highest scoring template in the 3D structural analysis of EgrG_08736 was the USP MJ0577 from *Methanococcus jannaschii*, with a confidence of homology of 99.9%. Against this template, the alignment coverage was 84%, and the sequence identity between both proteins was 33%. Analysis with ModFOLD v4.0 returned a global quality model score of 0.7 and a p-value of 6.4E-4 (Figure S3A, Additional file 3). The presence of a large coil between the second beta strand and the second alpha helix (Figure S3B, Additional file 3) is due to the insertion of eleven amino acids in our query sequence relative to the template. This region is highly variable across USP paralogs [5] (Figure S4, Additional file 3); it is located on the outside of the protein pocket in the 3D model (Figure S3, Additional file 3). Using Phyre v2 Investigator, we predicted the likely functional sites in our model, as well as the effect of mutations at these specific sites (described below; see also Figure S3C in Additional file 3).

Gene expression analysis

We evaluated RNA-seq data from previous transcriptomic reports [21, 22, 50]. Although these data do not include all developmental stages for each species, they are representative of the entire parasitic life cycle, occurring in both the intermediate and the definitive host. USP gene expression is highly variable between different life cycle stages, and while some genes are expressed constitutively (*Smp_04312*; *EgrG_08738* and their orthologs in *E. multilocularis* and *H. microstoma*), others are expressed in a specific spatio-temporal manner (*Smp_07640* and *Smp_09793*; *EgrG_08734*, *EgrG_08111*, and their orthologs in *E. multilocularis* and *H. microstoma*) (Figure 3A). As described above, *EgrG_02019* contains an *indel* that generates a stop codon in *Echinococcus* spp. (Figure 1B). Because this gene also exhibits a null or very low expression in all life cycle stages, we consider it a pseudogene with residual transcriptional activity. Interestingly, the gene expression patterns of *EgrG_08734* and *EgrG_08735* are almost identical to those of *EgrG_02019* and its orthologs in the Cestoda, as they are expressed only in the oncosphere stage, and only at very low levels (Figure 3A). In addition, the translation product of *EgrG_02019* (after including the thymine at position 61 in the coding sequence, see Figure 1B) has an amino acid sequence variation at the ATP-binding motif (GS>DN; not shown). In the same manner, EgrG_08734 and EgrG_08735 show modifications at these sites, GS> GR for the former, and GS>DS for the latter (Figure 3B). Our protein modeling predicts that these changes are critical for ligand binding and may have a negative impact on protein function (Figure S3C, Additional file 3). Moreover, Smp_13687 and Smp_13689 are the only two USPs in *S. mansoni* showing mutations within the ATP-binding motif, which also had a very low or null gene expression in all life cycle stages (cercaria, schistosomula, adult; see Figures 3A and B). Although there are no gene expression data for the other species (e.g. *E. canadensis*, *C. sinensis*, *T. solium*, etc.), they shared several amino acid substitutions at sites predicted to bind the ATP molecule (Figure 3B). To validate the RNA-seq data for the genus *Echinococcus*, we performed real time PCR of six USP genes from the pre-adult form of *Echinococcus ortleppi*. As expected, *Eo_08734* and *Eo_08735* were expressed at a very low level (< 1/10 of *Eo_08736* expression and barely detectable in the qPCR curves; see Figure 3C). On the other hand, *Eo_10769*, *Eo_07797* and *Eo_08738* were expressed at medium levels, and *Eo_08736* was highly expressed (Figure 3C). These results indicate a positive correlation between amino acid substitutions at sites that are critical for the contact with ligands and gene expression levels in the different life cycle stages of Platyhelminthes species.

(Figure 3)

Positive selection and Divergence analysis

Using 63 coding sequences from species of the class Cestoda (codon alignment in Figure S5, Additional file 3), we detect several sites under positive selection with PAML (Table 1). At first, these residues do not correspond to specific sites for the interaction with ligands, however, they are located in the vicinity (e.g. sites 7K, 16E, 20T, 32K, 114K, 115I, 117E, 120G, and 151N) (Table 1). In four amino acids (20T, 32K, 80E, and 151N), the posterior probability were more than 99%, suggesting an important role as adaptive sites in the evolution of the USP genes. Using Selecton 7, which ranks the residues from 1 to 7 (values 1 and 2 means Ka/Ks ratio > 1, and values from 3 to 7 denotes Ka/Ks ratio > 1), there was no signal of positive selected sites. However, we detected several residues with the value of 3 (Ka/Ks ratio near 1) (Table S4, Additional file 1). Most of these sites correspond to the positive selected sites found with PAML (Table 1, and Table S4 Additional file 1). These results suggest the presence of highly divergent sequences with adaptive sites in the USP genes of species of class Cestoda.

At the same time, based on the protein sequences, we searched for functional divergence between paralogs in the Cestoda and Trematoda classes using Diverge v3.0. The results are summarized in Table 2. In general, we found large differences within both Cestoda and Trematoda clusters, from one or few sites to numerous sites under functional divergence (Table 2). Some clusters where genes are located in tandem (Ce8_08736 and Ce5_G08738; Tr3_CsinT265 and Tr4_CsinT265) have no or just one site showing functional divergence; however, we found the opposite for other cluster comparisons (Ce4_08735 and Ce8_08736; Tr4_CsinT265 and Tr5_CsinT265) (Table 2). Functionally divergent amino acids are not restricted to a specific site, but are instead spread over a number of sites near the ATP-binding motif (see Table 2 and Figure 3B for the reference sequence). However, several amino acid positions (e.g. 29, 35, 113) were more frequent than others. A minor number of clusters comparisons (6/66 for Cestoda, and 14/55 for Trematoda) did not show functional divergence. These results indicate the presence of distinct levels of functional divergence between several USP clusters within Cestoda and Trematoda classes.

Table 1. Positive selection analysis of USP genes for species of class Cestoda.

Model	Estimates of parameters	*L*	Positive selected sites (PSS) [a]
M0 (one-ratio)	ω=0.32480	-17568.832757	None
M1a (neutral)	ω_0=0.03822, ω_1=1, p_0=0.95404, p_1=0.04596	-16175.637618	Not allowed
M2a (selection)	ω_0=0.03823, ω_1=1., ω_2=1, p_0=0.95404, p_1=0.00224, p_2=0.04372	-16175.637618	151N [b]
M7 (beta)	p=0.04543, q=0.40703	-16106.206229	Not allowed
M8 (beta& ω)	p_0=0.97089, p=0.03908, q=0.98312, p_1=0.02911, ω=1	-16028.140026	7K 16E **20T** **32K** 45R 48K 49K 50R 51D 64K 65S 671N 72E 77L **80E** 83N 114K 115I 117E 120G **151N**

[a] Positive selected sites (Bayes Empirical Bayes, BEB) are inferred at a cutoff posterior probability $P \geq 95\%$. Values for $P \geq 99\%$ are shown in bold font. The underlined PSS indicate a value of 3 (range from 1, positive selection, to 7, purifying selection) obtained with SELECTON (see Table S4, Additional file 1). Amino acid sites correspond to the reference sequence MJ0577 from *Methanococcus jannaschii*.

[b] Despite the presence of positive selected sites (24 sites with $P \geq 50\%$, 1 site with $P \geq 95\%$), the LRT test was not significant when comparing the Log likelihood scores from the M1a and M2a models.

Table 2. Functional divergence analysis (Type I) within Cestoda and Trematoda species.

Cestoda comparisons [a]

Cluster 1	Cluster 2	Θ±SE	LRT	Sites (Qk>0.9) [c]
Ce11_07258	Ce12_09018	0.70±0.14	23.04	28,31,37,44,94,98,124
Ce10_10769	Ce12_09018	0.78±0.17	20.59	14,**29**,96,98,115,120,159
Ce11_07258	Ce10_10769	0.95±0.16	31.63	Almost all
Ce4_08735	Ce3_09839	0.97±0.14	45.46	Almost all
Ce4_08735	Ce5_08738	0.98±0.17	32.99	Almost all
Ce4_08735	Ce6_08734	0.91±0.14	40.82	Almost all
Ce4_08735	Ce8_08736	1.36±0.15	80.72	Almost all
Ce5_08738	Ce6_08734	0.80±0.18	18.96	7,**35**,**36**,39,45,59,98,100,108,**113**,121
Ce5_08738	Ce8_08736	0.64±0.14	18.81	**35**
Ce6_08734	Ce8_08736	0.79±0.14	29.00	19,**36**,72,100,**113**,121,159

Trematoda comparisons [b]

Cluster 1	Cluster 2	Θ±SE	LRT	Sites (Qk>0.9) [c]
Tr4_CsinT265	Tr3_CsinT265	0.64±0.20	9.84	None
Tr3_CsinT265	Tr5_CsinT265	0.75±0.21	12.58	134
Tr4_CsinT265	Tr5_CsinT265	0.99±0.24	16.21	Almost all
Tr2_SmpA	Tr3_CsinT265	0.94±0.13	45.93	Almost all
Tr2_SmpA	Tr5_CsinT265	0.68±0.17	15.36	**29**
Tr2_SmpA	Tr1_SmpA	0.69±0.18	14.17	12,26,102
Tr7_CsinT265SmpA	Tr1_SmpA	0.68±0.19	12.56	12,26,102
Tr6_CsinT265SmpA	Tr9_CsinT265SmpA	0.84±0.19	17.70	12,**29**,32,**35**,**36**,59,110,**113**,114,117,118,122,144,145,150

[a] Cestoda clusters are defined based on the *E. granulosus* IDs, e.g. the cluster Ce3_09839 is composed by the EgrG_09839, EmuJ_09839, Ecan_08199, TsM_09222, and HmN_01226 sequences.

[b] Trematoda clusters are described as follows: Tr1 (Smp_136870, A_04288, Smp_136890, A_06342), Tr2 (Smp_043120, A_03767, Smp_202690, A_04393), Tr3 (Csin107893, T265_02180, CsinSc585new, T265_02179), Tr4 (Csin107892, Csin110039, Csin107891, T265_02178a, T265_02178b,T265_02177), Tr5 (Csin107894, T265_02181, Csin107895, T265_02182), Tr6 (Smp_076400, A_07834, Csin112002, T265_05585), Tr7 (Smp_031300, A_04567, Csin112617, T265_03499), Tr9 (Smp_001000, A_07787, Smp_200240, A_05680).

[c] Amino acid sites correspond to the reference sequence MJ0577 from *Methanococcus jannaschii*. Amino acids shared by Cestoda and Trematoda species are indicated in bold font and plotted in the figure 3B. Underlined sites correspond to positive selected sites detected with PAML (Table 1).

Discussion

The expansion of gene families by gene duplication represents a successful strategy for the propagation of gene copies through the acquisition of specialized or novel functions (e.g. globin or homeobox gene families) [54, 55]. Although some genes may acquire adaptive novelties that are maintained from one generation to the next, others may follow a pseudogenization process through the accumulation of deleterious mutations. An understanding of when and how fast these duplications occur is key to our understanding of the duplicated genes' functional diversity.

Here, we explored the evolutionary fates of the USP gene family in Platyhelminthes species of medical relevance. We found that the USP genes of this phylum are mostly intronless, transcribed independently and encoding a single protein domain. A few USPs (*EgrG_09018* and orthologs) contain a single intron in a conserved position around amino acid 75. This is similar to what has been described for *Hydra*, where 22 out of 24 USP genes are intronless [5]. Based on a well-supported monophyletic clade, Forêt and colleagues consider that a single retrotransposition event had a pivotal role in the emergence of most intronless USP genes after the anthozoan/hydrozoan divergence [5]. Our results show that the same process could have been very important after the separation of the Cestoda and Trematoda classes. Nevertheless, the cluster organization of the USP genes in the Platyhelminthes (around 50% of USP genes occur in clusters in both Cestoda and Trematoda) revealed the importance of tandem duplications for the generation of new USP copies. This idea is supported by the presence of a well-conserved DNA motif occurring at the same position in tandemly organized genes (e.g. *EgrG_08736*, *EgrG_08738,* and orthologs in the Cestoda; *T265_02179* and *T265_02181*, *Smp_001000* and *Smp_200240*, and orthologs in *C. sinensis* and *S. haematobium*, respectively). Surprisingly, we found that the position of the -*CCAATCA*- motif was also preserved between isolated USP genes and their most closely related homologs (e.g. *EgrG_10769* and *EgrG_07258*, which probably emerged from *EgrG_09018*), suggesting a retrotransposition event that included both the coding sequence and promoter region. This might be due to the fact that the transcription start sites (TSS) tend to be interspersed rather than located at one specific site. If a TSS upstream of the promoter region is used, a large part of the core promoter may be transcribed [56, 57]. This mechanism could ensure the transcriptional activity of the newly retrotransposed genes and would consequently avoid the effects of neutral evolution, i.e., the accumulation of deleterious mutations.

USPs are often classified based on the presence or absence of the conserved ATP-binding motif residues [58]. A positive correlation between the conservation of the [Gx2G*x*9G(S/T)] protein motif and crystal solubility in the presence of ATP, or an ATP analog, has been previously described [11]. In addition, the same authors found a high level of conservation (~80%) at amino acid positions forming part of the motif across all crystals extracted from the PDB database, with exception of the second glycine (G130, which is preserved in 50% of crystals). Our protein model showed that amino acid alterations at specific ligand-binding sites have a negative effect on protein function, including modifications at P11, D13, V41, G127, G130, G140, and S/T141. The residue G130 was the most exchangeable amino acid, and could thus be most easily substituted non-synonymously (Figure S3C, Additional file 3). The USP gene expression data allowed us to associate the transcriptional activity with modifications at residues that are critical for ligand interaction. In general, we observed that alterations in the [Gx2G*x*9G(S/T)] motif and at other positions within the protein pocket (e,g, D13, V41) are associated with very low or null levels of gene expression in almost all life cycle stages of the parasites, probably as a result of functional redundancy [59]. Additionally, for several USPs, we observed amino acid insertions (*SMU15024145*, *EgrG_08735*, *HmN_05623*, etc.) and deletions (*T265_02176*, *HmN_00323*, *Csin110041*) within the [Gx2G*x*9G(S/T)] motif. These modifications could lead to a steric hindrance, thereby preventing the contact with the ligand. Based on this, we believe that several USP genes in the Cestoda (*EgrG_08734*, *EgrG_08735*, and orthologs; *HmN_05021* and *HmN_05022*; etc.) and Trematoda (*Smp_136870*, *Smp_136890*; *Smp_09793* and their orthologue *A_04226*; etc.) are in the process of pseudogenization. Like the pseudogenes described here (Figures 1A and B), the genes under pseudogenization lack the canonical and well conserved -*CCAATCA*- motif in the promoter region, which probably affects their transcriptional activity. Although less likely, the possibility that changes in the ATP-binding sites might expand the ligand repertoire should not be dismissed. Further functional studies will be necessary to clarify these findings.

Despite several gene losses in species of the *Taeniidae* family, we found a high functional divergence between the different USP paralogs. Positively selected sites identified by PAML (Table 1) included several residues near to the ligand binding sites, a similar result observed by Diverge v3.0 (Table 2). Several amino acids (e.g. 7K, 32K, 114K, 117E), were identified by the two approaches, suggesting an important role for these sites in the evolution of the USP genes. The number of sites under functional divergence was highly variable regarding cluster comparisons (Table 2). This dynamic behavior could indicate that some USP genes are evolving faster than others. Interestingly, no amino acid changes were observed in the ATP-binding motif of the USP genes shared by all species studied (“ancestral” USPs, see Figure 3). Moreover, the lack of the -*CCAATCA*- motif in promoter regions is restricted to the non-ancestral USPs. These observations may suggest the maintenance of ancient functions for the “ancestral” USPs, and the emergence of functional novelties (through recurrent mutations) for the “new” USP paralogs. Functional divergence is observed in a greater number of residues, but some amino acid positions (29, 35, 36, 113) are shared by several cluster comparisons. The repeated occurrence of these sites may represent an adaptive trait for the protein function. The high divergence agrees with the multifunctional behavior observed for several taxa. In *E. coli*, for example, the *uspC* and *uspE* both affect motility positively and adhesion negatively, while *uspA* and *uspD* are involved in the oxidative stress defense [7]. In addition to changes in the protein coding sequence, selective forces can lead to changes in the regulatory elements. For example, in *Arabidopsis thaliana*, a high divergence between duplicated copies was observed for the *cis*-regulatory elements and methylation patterns, leading to different expression profiles [60]. These findings suggest that different USP genes may be subjected to dissimilar types of selective pressure. Some genes will accumulate deleterious mutations both in the promoter regions (affecting their transcriptional activity) and in the protein coding sequence (generating a truncated and/or non-functional product). These genes will be under neutral evolution and become pseudogenized as a consequence of functional redundancy (Figure 4). Even so, there might be a possible role for the residual transcripts of genes under pseudogenization, including *EgrG_USPps1*, its orthologs in *E. multilocularis* and *E. canadensis,* and other genes (see above): in a novel mechanism of gene regulation by pseudogenes, these non-coding RNA (ncRNA) transcripts might act as gene expression regulators by promoting the degradation (or kidnapping) of functional USP mRNAs by hybridization (Figure 4) [61]. This mechanism might be of relevance in the context of the narrow spatiotemporal regulation of the USP genes in each life cycle stage.

(Figure 4)

In contrast to the process of pseudogenization, many USP paralogs may have acquired new functions, leading to functional diversification within the gene family. Since several amino acid modifications have occurred close to ligand-binding sites, this functional diversification may be associated to the interaction with different types of ligands. Furthermore, several USPs were found to occur as dimers or higher oligomeric complexes [8, 11], suggesting that substitutions involved in the protein oligomerization could increase the complexity of the protein-protein interactions. These sequence variations, and those found in promoter regions, could be considered adaptive traits that emerge as part of subfunctionalization or neofunctionalization processes (Figure 4). The publication of the genome sequences of several Platyhelminthes species [21, 22, 50, 62] revealed that gene expansion, such as in heat shock proteins, species-specific antigens, or proteases, is a widespread process related to the adaptation to parasitism. In this way, some of the USP paralogs could be considered adaptations to the parasitic lifestyle by increasing the repertoire of binding proteins, by establishing complex protein-protein interactions (homo- and heterodimers), or by being expressed in a specific tissue, life cycle stage, or in response to a particular stressor.

Conclusions

In the present work we found that the USP gene family has an ancient origin and follows a complex evolutionary pattern (pseudogenization and sub/neofunctionalization) for several Platyhelminth species. This scenario may result from different selective pressures acting on the USP genes. If these patterns are restricted to parasitic flatworms, or also include the free-living species, remains to be elucidated. Further studies associating functional diversity with the various sequence modifications will help deepen our knowledge about the patterns and regulation of USP gene expression. Additional analyses will be necessary to investigate the role of ncRNAs in the specific spatiotemporal regulation of the USP genes.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The datasets supporting the conclusions of this article are included within the article and its additional files.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior *(grant number 1278/2011*) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (grant number *472316/2013-3*), Ministry of Education, Brazil. The funders had no role in study design, data collection and analysis, decision to publish, interpretation of data, or preparation of the manuscript.

Authors' contributions

SME and AZ conceived and designed this study. SME, MPC, and LBC performed the experiments and analyzed the data. SME wrote the manuscript. MPC, LBC, and AZ helped to draft the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

We thank the following colleagues for helping us to improve our manuscript through their comments and suggestions: Claudia Thompson (Center for Biotechnology, UFRGS, Porto Alegre, Brazil), Loreta Freitas (Department of Genetics, UFGRS, Porto Alegre, Brazil), and Karen Haag (Department of Genetics, UFGRS, Porto Alegre, Brazil). We also thank to the editorial board and the anonymous reviewers for their comments and valuable contributions.

Figure legends

Fig. 1. Organization of the USP gene family in Platyhelminthes. (A) Platyhelminthes USP genes are distributed in clusters throughout the genome and show lineage-specific expansions and losses. For each species, the total number of USP genes are given in parentheses, followed by a scheme (boxes) showing the genomic organization for some genes. Boxes with blue frame correspond to syntenic genes in Cestoda species. The sequence alignment of the entire cluster in Cestoda parasites (50% of identity at baseline and using *E. granulosus* as the reference sequence) display high levels of sequence identity in the coding region. Gene identity of USP genes is lost when compared to the Trematoda and Turbellaria classes, suggesting a high divergence of the USP genes between groups. Some USP paralogs in the Cestoda were identified as pseudogenes (dotted, gripped, and striped boxes). Because of the synteny, pseudogenes in *T. solium* and *Echinococcus* spp. correspond to gene losses when compared to *H. microstoma*. The *EgrG_ps1* pseudogene refers to the *EgrG_2019* sequence (WormBase Parasite annotation). Asterisks indicate the same *USP* distribution for *E. canadensis* and *E. multilocularis* compared to *E. granulosus*, and for *S. haematobium* and *O. viverrini* compared to *S. mansoni* and *C. sinensis*, respectively. (B) The presence of *indels* (highlighted in red in the protein sequence) generates frameshift mutations and serves as evidence of a pseudogenization process. The sequences of some pseudogenes (*EgrG_ps1*, *HmN_ps1*) are very similar to that of their orthologs. Others (*EgrG_ps2*) are very different, and homologies are difficult to identify accurately by tBLASTn. In addition to the protein alignment, the *indels* are indicated in the coding sequence for three pseudogenes.

Fig. 2. Phylogenetic relationships between USPs in the Platyhelminthes. The maximum-likelihood phylogenetic tree (method aLRT-SH for branch support, see text for more details) shows several USP sequences shared by platyhelminthes and annelids, referred as ancestral USP genes (highlighted in yellow). On the other hand, the Trematoda and Cestoda (highlighted in green) classes show species-specific expansions and losses (one asterisk indicates losses in the *Taeniidae* family; while two asterisks represent losses in the genus *Echinococcus*). For simplification (and to facilitate associations with figure 1), we grouped the Cestoda sequences according to the IDs of *E. granulosus* based on the ortholog relationship (see Table S1 in Additional file 1). Gene names are in italic. Prefix species are as follow: CapteP for *C. teleta*, HelroG for *H. robusta*, Gsa for *G. salaris*, Mli for *M. lignano*, SMU for *S. mediterranea*, Smp for *S. mansoni*, Csin for *C. sinensis*, TsM for *T. solium*, HmN for *H. microstoma*, and EgrG for *E. granulosus*. The number in parenthesis beside SMU and Mli correspond to the number of collapsed sequences in *S. mediterranea* and *M. lignano*, respectively. The USPs clustered in the chromosome (see Figure 1) are also grouped together in the phylogenetic tree, suggesting an origin by subsequent tandem duplications. Identical sequences

from *M. lignano* (~35) were excluded in the analysis. Three molluscs (*L. gigantea*, *C. gigas*, and *O. bimaculoides*) and two annelids (*H. robusta* and *C. teleta*) were used as outgroups. A minor ID for *M. lignano* and *G. salaris* was used (Table S5). For an extended tree, see the Figures S1 and S2 in the Additional file 1. Branch support values obtained by Bayesian Inference are in bold font. Only values with a branch support greater than 0.7 are showed.

Fig. 3. Relationship between USP sequence modifications and gene expression patterns. (A) The gene expression profile is highly variable between the different life cycle stages of Cestoda and Trematoda species, with some USP genes with null or very low expression (*EgrG_08734*, *EgrG_08735* and orthologs), others expressed in specific manner (*EgrG_08111; EgrG_07797* and orthologs), and others constitutively expressed (*EgrG_08738* and orthologs). ID numbers in red color refer to USP proteins that exhibit modifications in the ATP binding motif. The asterisk refers to the "ancestral" USPs (see figure 2). (B) USP sequence variations in the Platyhelminthes. On top, sequence logo generated with all USP sequences without changes to the ATP binding motif [Gx2Gx9G(S/T)]. Amino acids interacting with ligands are shown in boxes. Below, alignment of USP protein sequences showing modifications in the protein motif. ID numbers (in red) refer to USP proteins for which RNA-seq data was available, to facilitate the comparison between sequence modification and gene expression patterns. Modifications in the [Gx2Gx9G(S/T)] motif and at other sites known to be involved in ligand interaction are highlighted in red and yellow, respectively. The sequence of the *Methanococcus jannaschii* USP MJ0577 USP was used as a reference (starting from position 6), with the ATP binding motif highlighted in green. The residues under functional divergence are indicated by arrows (black arrows, residues shared by Cestoda and Trematoda; gray arrows: residues specific to the Cestoda or Trematoda) (see Table 1). (C) qPCR gene expression analysis of USP genes in the pre-adult form of *E. ortleppi.* Some genes (*E0_08736*, *Eo_08738*, *Eo_07797*, and *Eo_10769*) showed higher levels of gene expression than others (*Eo_08734* and *Eo_08735*), in line with previously published RNA-seq data for the genus *Echinococcus* (see above). Asterisks indicate a p-value < 0.01 for the comparison of *Eo_08734* and *Eo_08735* with the other genes.

Fig. 4. Possible evolutionary fates for the USP paralogs in Platyhelminthes parasites. First, a new *USP* copy can accumulate deleterious mutations, leading to alterations in the protein sequence with a loss of function (pseudogenization). From this, some ncRNAs can be transcribed, and thus, regulate the gene expression of the other USP paralogs via mRNA degradation (regulation by ncRNAs). Second, the USP paralog could undergo several non-synonymous mutations, thereby acquiring a new function (neofunctionalization). Finally, some *USP* copies could maintain the

same function, but be expressed in a specific life cycle stage or in response to a specific stressor (subfunctionalization).

Additional files *

Additional file 1. Tables S1-S5. Table S1: Details of each primer designed for the six USP genes. Table S2: USP orthologues relationship between Platyhelminthes species. Table S3: Motif scanning in promoter regions of the USP genes. Table S4: SELECTON results for Cestoda species. Table S5: IDs used for *M. lignano* and *G. salaris* in the phylogenetic analysis. (.xls).

Additional file 2. Protein and upstream sequences. (.txt).

Additional file 3. Figures S1-S5. Figure S1: Phylogenetic tree generated using the aLRT-SH method with PhyML. Figure S2: Phylogenetic tree generated using Bayesian Inference with BEAST v1.8.4. Figure S3: EgrG_08736 3D protein modeling. Figure S4: Protein alignment used to generate the phylogenetic trees. Figure S5: Positive selection analysis in Cestoda species. (.pptx).

* *Os arquivos "Additional file 2" e "Additional file 3" foram disponibilizadas online.*

References

1. Hurles, M., Gene duplication: the genomic trade in spare parts. PLoS Biol, 2004. 2(7): p. E206.
2. Ohno, S., Evolution by Gene Duplication. 1970: Springer.
3. Hahn, M.W., Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered, 2009. 100(5): p. 605-17.
4. Innan, H. and F. Kondrashov, The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet, 2010. 11(2): p. 97-108.
5. Forêt, S., et al., Phylogenomics reveals an anomalous distribution of USP genes in metazoans. Mol Biol Evol, 2011. 28(1): p. 153-61.
6. Gustavsson, N., A. Diez, and T. Nyström, The universal stress protein paralogues of Escherichia coli are co-ordinately regulated and co-operate in the defence against DNA damage. Mol Microbiol, 2002. 43(1): p. 107-17.
7. Nachin, L., U. Nannmark, and T. Nyström, Differential roles of the universal stress proteins of Escherichia coli in oxidative stress resistance, adhesion, and motility. J Bacteriol, 2005. 187(18): p. 6265-72.
8. Jung, Y.J., et al., Universal Stress Protein Exhibits a Redox-Dependent Chaperone Function in Arabidopsis and Enhances Plant Tolerance to Heat Shock and Oxidative Stress. Front Plant Sci, 2015. 6: p. 1141.
9. Drumm, J.E., et al., Mycobacterium tuberculosis universal stress protein Rv2623 regulates bacillary growth by ATP-Binding: requirement for establishing chronic persistent infection. PLoS Pathog, 2009. 5(5): p. e1000460.
10. Zarembinski, T.I., et al., Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. Proc Natl Acad Sci U S A, 1998. 95(26): p. 15189-93.
11. Tkaczuk, K.L., et al., Structural and functional insight into the universal stress protein family. Evol Appl, 2013. 6(3): p. 434-49.
12. Boes, N., et al., The Pseudomonas aeruginosa universal stress protein PA4352 is essential for surviving anaerobic energy stress. J Bacteriol, 2006. 188(18): p. 6529-38.
13. World Health Organization. World Health Statistics 2016: monitoring health for the SDGs, sustainable development goals. Available from: http://www.who.int/en/ Accessed Dec 10, 2016.
14. Molyneux, D.H., L. Savioli, and D. Engels, Neglected tropical diseases: progress towards addressing the chronic pandemic. Lancet, 2016.
15. Thompson, R., Biology and systematics of Echinococcus. 1995: CAB International, Wallingford. 1-50.
16. Maeng, S., et al., Oxidative stress-mediated mouse liver lesions caused by Clonorchis sinensis infection. Int J Parasitol, 2016. 46(3): p. 195-204.
17. Negrão-Corrêa, D., et al., Interaction of Schistosoma mansoni Sporocysts and Hemocytes of Biomphalaria. J Parasitol Res, 2012. 2012: p. 743920.
18. Cheng, Z., et al., Identification and characterisation of Emp53, the homologue of human tumor suppressor p53, from Echinococcus multilocularis: its role in apoptosis and the oxidative stress response. Int J Parasitol, 2015. 45(8): p. 517-26.
19. Masamba, P., et al., Universal Stress Proteins as New Targets for Environmental and Therapeutic Interventions of Schistosomiasis. Int J Environ Res Public Health, 2016. 13(10).
20. Bowles, J., D. Blair, and D.P. McManus, Genetic variants within the genus Echinococcus identified by mitochondrial DNA sequencing. Mol Biochem Parasitol, 1992. 54(2): p. 165-73.
21. Tsai, I.J., et al., The genomes of four tapeworm species reveal adaptations to parasitism. Nature, 2013. 496(7443): p. 57-63.
22. Zheng, H., et al., The genome of the hydatid tapeworm Echinococcus granulosus. Nat Genet, 2013. 45(10): p. 1168-75.

23. Espinola, S.M., H.B. Ferreira, and A. Zaha, Validation of Suitable Reference Genes for Expression Normalization in Echinococcus spp. Larval Stages. Plos One, 2014. 9(7).
24. Ruijter, J.M., et al., Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Res, 2009. 37(6): p. e45.
25. Robb, S.M., et al., SmedGD 2.0: The Schmidtea mediterranea genome database. Genesis, 2015. 53(8): p. 535-46.
26. Howe, K.L., et al., WormBase 2016: expanding to enable helminth genomic research. Nucleic Acids Res, 2016. 44(D1): p. D774-80.
27. Aken, B.L., et al., Ensembl 2017. Nucleic Acids Res, 2017. 45(D1): p. D635-D642.
28. Nordberg, H., et al., The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res, 2014. 42(Database issue): p. D26-31.
29. Guindon, S., et al., New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol, 2010. 59(3): p. 307-21.
30. Drummond, A.J., et al., Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol, 2012. 29(8): p. 1969-73.
31. Katoh, K. and D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol, 2013. 30(4): p. 772-80.
32. HIV database: Gap Strip/Squeeze v2.1.0. Available from: https://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html. Accessed 10 September, 2017.
33. Lefort, V., J.E. Longueville, and O. Gascuel, SMS: Smart Model Selection in PhyML. Mol Biol Evol, 2017. 34(9): p. 2422-2424.
34. Gernhard, T., The conditioned reconstructed process. J Theor Biol, 2008. 253(4): p. 769-78.
35. Yang, Z., Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol, 1994. 39(3): p. 306-14.
36. Rambaut A, S.M., Xie D, Drummond AJ. Tracer v1.6. 2014, Available from http://tree.bio.ed.ac.uk/software/tracer/.
37. Rambaut, A. FigTree: Tree figure drawing tool. 2014. Available from http://tree.bio.ed.ac.uk/.
38. Zhang, J., R. Nielsen, and Z. Yang, Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol, 2005. 22(12): p. 2472-9.
39. Doron-Faigenboim, A. and T. Pupko, A combined empirical and mechanistic codon model. Mol Biol Evol, 2007. 24(2): p. 388-97.
40. Stern, A., et al., Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res, 2007. 35(Web Server issue): p. W506-11.
41. Suyama, M., D. Torrents, and P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res, 2006. 34(Web Server issue): p. W609-12.
42. Yang, Z. and R. Nielsen, Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol, 2002. 19(6): p. 908-17.
43. Bailey, T.L., et al., MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res, 2009. 37(Web Server issue): p. W202-8.
44. Gupta, S., et al., Quantifying similarity between motifs. Genome Biol, 2007. 8(2): p. R24.
45. Buske, F.A., et al., Assigning roles to DNA regulatory motifs using comparative genomics. Bioinformatics, 2010. 26(7): p. 860-6.
46. Gu, X., et al., An update of DIVERGE software for functional divergence analysis of protein family. Mol Biol Evol, 2013. 30(7): p. 1713-9.
47. Gu, X., Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol, 1999. 16(12): p. 1664-74.

48. Kelley, L.A., et al., The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc, 2015. 10(6): p. 845-58.
49. McGuffin, L.J., M.T. Buenavista, and D.B. Roche, The ModFOLD4 server for the quality assessment of 3D protein models. Nucleic Acids Res, 2013. 41(Web Server issue): p. W368-72.
50. Protasio, A.V., et al., A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni. PLoS Negl Trop Dis, 2012. 6(1): p. e1455.
51. Wasik, K., et al., Genome and transcriptome of the regeneration-competent flatworm, Macrostomum lignano. Proc Natl Acad Sci U S A, 2015. 112(40): p. 12462-7.
52. Mantovani, R., The molecular biology of the CCAAT-binding factor NF-Y. Gene, 1999. 239(1): p. 15-27.
53. Thön, M., et al., The CCAAT-binding complex coordinates the oxidative stress response in eukaryotes. Nucleic Acids Res, 2010. 38(4): p. 1098-113.
54. Holland, P.W., Evolution of homeobox genes. Wiley Interdiscip Rev Dev Biol, 2013. 2(1): p. 31-45.
55. Storz, J.F., Gene Duplication and Evolutionary Innovations in Hemoglobin-Oxygen Transport. Physiology (Bethesda), 2016. 31(3): p. 223-32.
56. Frith, M.C., et al., A code for transcription initiation in mammalian genomes. Genome Res, 2008. 18(1): p. 1-12.
57. Okamura, K. and K. Nakai, Retrotransposition as a source of new promoters. Mol Biol Evol, 2008. 25(6): p. 1231-8.
58. Sousa, M.C. and D.B. McKay, Structure of the universal stress protein of Haemophilus influenzae. Structure, 2001. 9(12): p. 1135-41.
59. Hingley-Wilson, S.M., et al., Individual Mycobacterium tuberculosis universal stress protein homologues are dispensable in vitro. Tuberculosis (Edinb), 2010. 90(4): p. 236-44.
60. Wang, J., N.C. Marowsky, and C. Fan, Divergent evolutionary and expression patterns between lineage specific new duplicate genes and their parental paralogs in Arabidopsis thaliana. PLoS One, 2013. 8(8): p. e72362.
61. Milligan, M.J. and L. Lipovich, Pseudogene-derived lncRNAs: emerging regulators of gene expression. Front Genet, 2014. 5: p. 476.
62. Berriman, M., et al., The genome of the blood fluke Schistosoma mansoni. Nature, 2009. 460(7253): p. 352-8

Figures

Figure 1

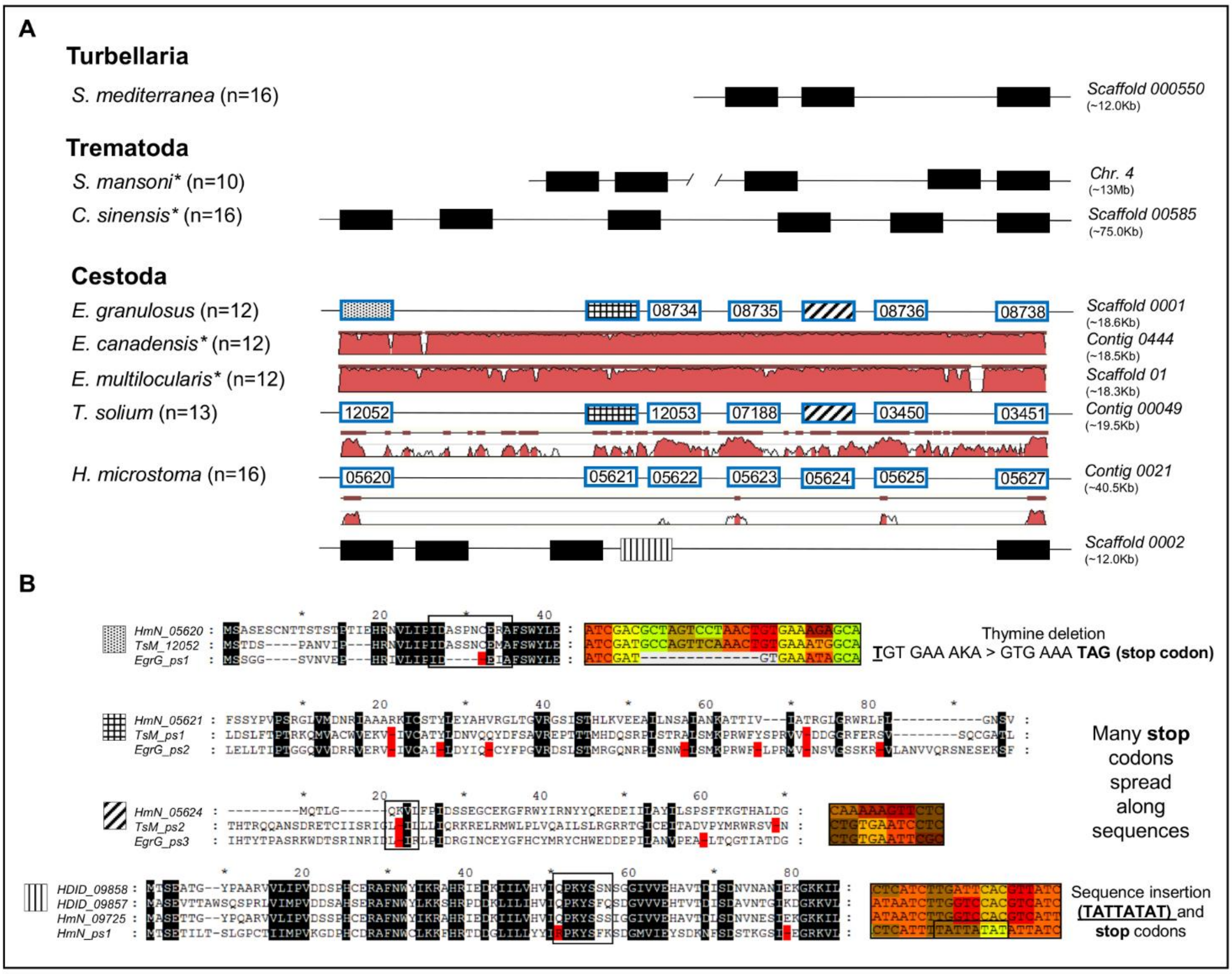

A
Turbellaria
S. mediterranea (n=16)
Scaffold 000550
(~12.0Kb)
Trematoda
S. mansoni* (n=10)
Chr. 4
(~13Mb)
C. sinensis* (n=16)
Scaffold 00585
(~75.0Kb)
Cestoda
E. granulosus (n=12)
08734
08735
08736
08738
Scaffold 0001
(~18.6Kb)
E. canadensis* (n=12)
Contig 0444
(~18.5Kb)
E. multilocularis* (n=12)
Scaffold 01
(~18.3Kb)
T. solium (n=13)
12052
12053
07188
03450
03451
Contig 00049
(~19.5Kb)
H. microstoma (n=16)
05620
05621
05622
05623
05624
05625
05627
Contig 0021
(~40.5Kb)
Scaffold 0002
(~12.0Kb)
B
HmN_05620 : MSASESCNTTSTSTPTIEHRNVLIPIDASPNCERAFSWYLE : ATCGACGCTAGTCCTAACTGTGAAAGAGCA
TsM_12052 : MSTDS----PANVIP---HRNVLIPIDASSNCEMAFSWYLE : ATCGATGCCAGTTCAAACTGTGAAATGGCA
EgrG_ps1 : MSSGG----SVNVEP---HRIVLIPID-----EIAFSWYLE : ATCGAT-------------GTGAAATAGCA
Thymine deletion
TGT GAA AKA > GTG AAA TAG (stop codon)
HmN_05621 : FSSYPVPSRGLVMDNRIAAARKICSTYLEYAHVRGLTGVRGSISTHLKVEEAILNSAIANKATTIV---IATRGLGRWRLFI-----------GNSV :
TsM_ps1 : LDSLFTPTRKQMVACWVEKV-IVCATYLDNVQQYDFSAVREPTTTMHDQSRPLSTRAISMKPRWFYSPRVV-DDGGRFERSV---------SQCGATL :
EgrG_ps2 : LELLTIPTGGQVVDRRVERV-IVCAI-LDYIQ-CYFPGVRDSLSTMRGQNRPISNW-ISMKPRWF-LPRMV-NSVGSSKR-VLANVVQRSNESEKSF :
Many stop codons spread along sequences
HmN_05624 : ---------MQTLG------QKVIFPIDSSEGCEKGFRWYIRNYYQKEDEIILAYILSPSFTKGTHALDG : CAAAAAGTTCTC
TsM_ps2 : THTRQQANSDRETCIISRIGL-IILLIQRKRELRMWLPLVQAILSLRGRRTGICEITADVPYMRWRSV-N : CTGTGAATCCTC
EgrG_ps3 : IHTYTPASRKWDTSRINRIDL-IRLPIDRGINCEYGFHCYMRYCHWEDDEPILANVPEA-LTQGTIATDG : CTGTGAATTCGC
HDID_09858 : MTSEATG--YPAARVVLIPVDDSPHCERAFNWYIKRAHRIEDKIILVHVIQPKYSSNSGGIVVEHAVTDISDNVNANIEKGKKIL : CTCATCTTGATTCACGTTATC
HDID_09857 : MASEVTTAWSQSPRLVIMPVDDSAHSERAFNWYLKKSHRPDDKLILIHVIQPKYSFQSDGVVVEHTVTDISDAVNTGIKDGKKVL : ATAATCTTGGTCCACGTCATC
HmN_09725 : MASETTG--YPQARVVLIPVDSSPHCERAFNWYIKRAHRIEDKIILVHVIQPKYSSSIGGIVVEHAVTDISDNVNESIEKGKKIL : ATAATCTTGGTCCACGTCATT
HmN_ps1 : MTSETILT-SLGPCTIIMPVKGDPHCDRAFNWCLKKFHRTDDGLILLYYIRPKYSFKSDGMVIEYSDKNFSDSTKGSI-EGRKVL : CTCATTTTATTATATATTATC
Sequence insertion
(TATTATAT) and
stop codons

Figure 2

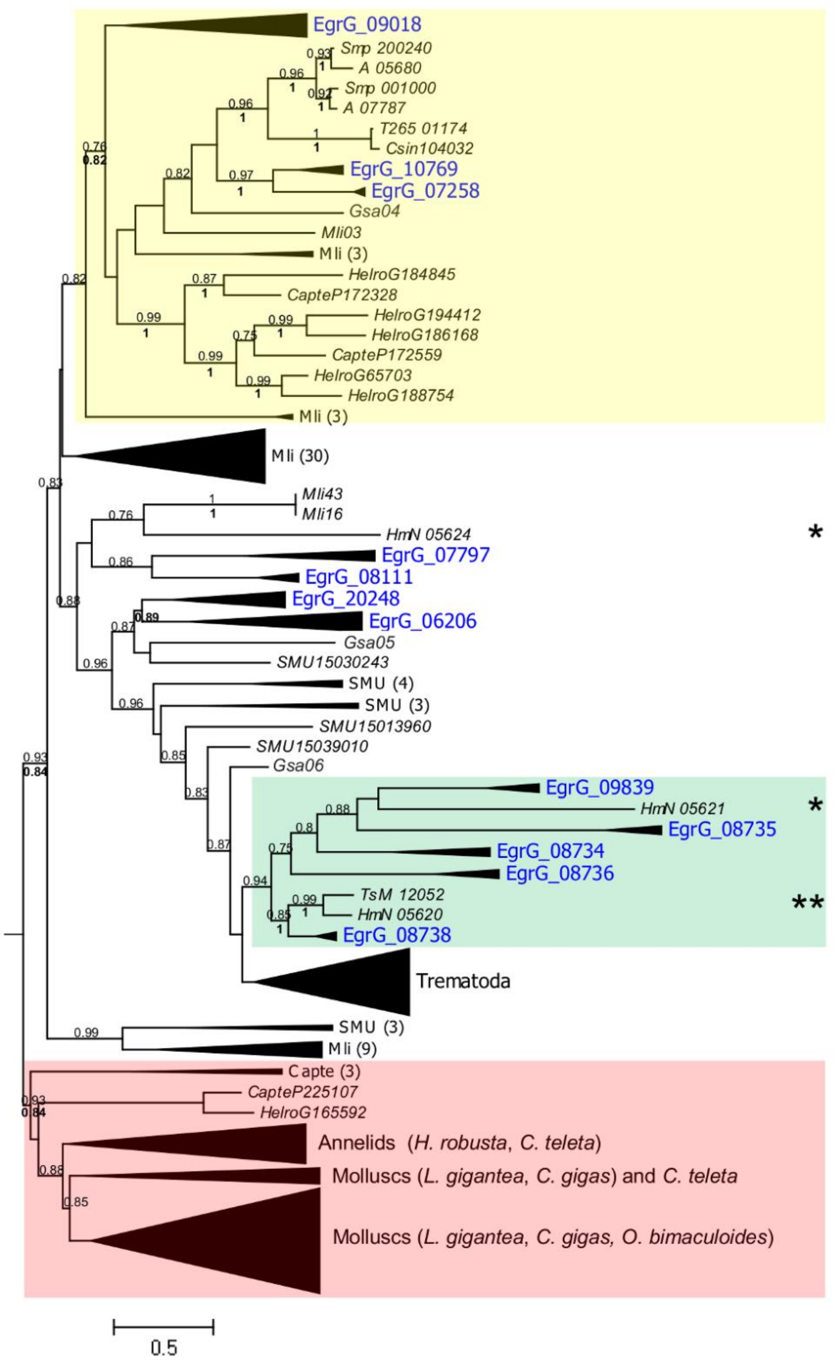

EgrG_09018
Smp 200240
A 05680
Smp 001000
A 07787
T265 01174
Csin104032
EgrG_10769
EgrG_07258
Gsa04
Mli03
Mli (3)
HelroG184845
CapteP172328
HelroG194412
HelroG186168
CapteP172559
HelroG65703
HelroG188754
Mli (3)
Mli (30)
Mli43
Mli16
HmN 05624
*
EgrG_07797
EgrG_08111
EgrG_20248
EgrG_06206
Gsa05
SMU15030243
SMU (4)
SMU (3)
SMU15013960
SMU15039010
Gsa06
EgrG_09839
HmN 05621
*
EgrG_08735
EgrG_08734
EgrG_08736
TsM 12052
HmN 05620
**
EgrG_08738
Trematoda
SMU (3)
Mli (9)
Capte (3)
CapteP225107
HelroG165592
Annelids (H. robusta, C. teleta)
Molluscs (L. gigantea, C. gigas) and C. teleta
Molluscs (L. gigantea, C. gigas, O. bimaculoides)
0.93
1
0.96
1
0.92
1
0.96
1
1
1
0.76
0.82
0.82
0.97
1
0.82
0.87
1
0.99
1
0.99
1
0.75
0.99
1
0.99
1
0.83
0.76
1
1
0.86
0.88
0.87
0.89
0.96
0.96
0.85
0.83
0.87
0.93
0.84
0.88
0.8
0.75
0.94
0.99
1
0.85
1
0.99
0.93
0.84
0.88
0.85
0.5

Figure 3

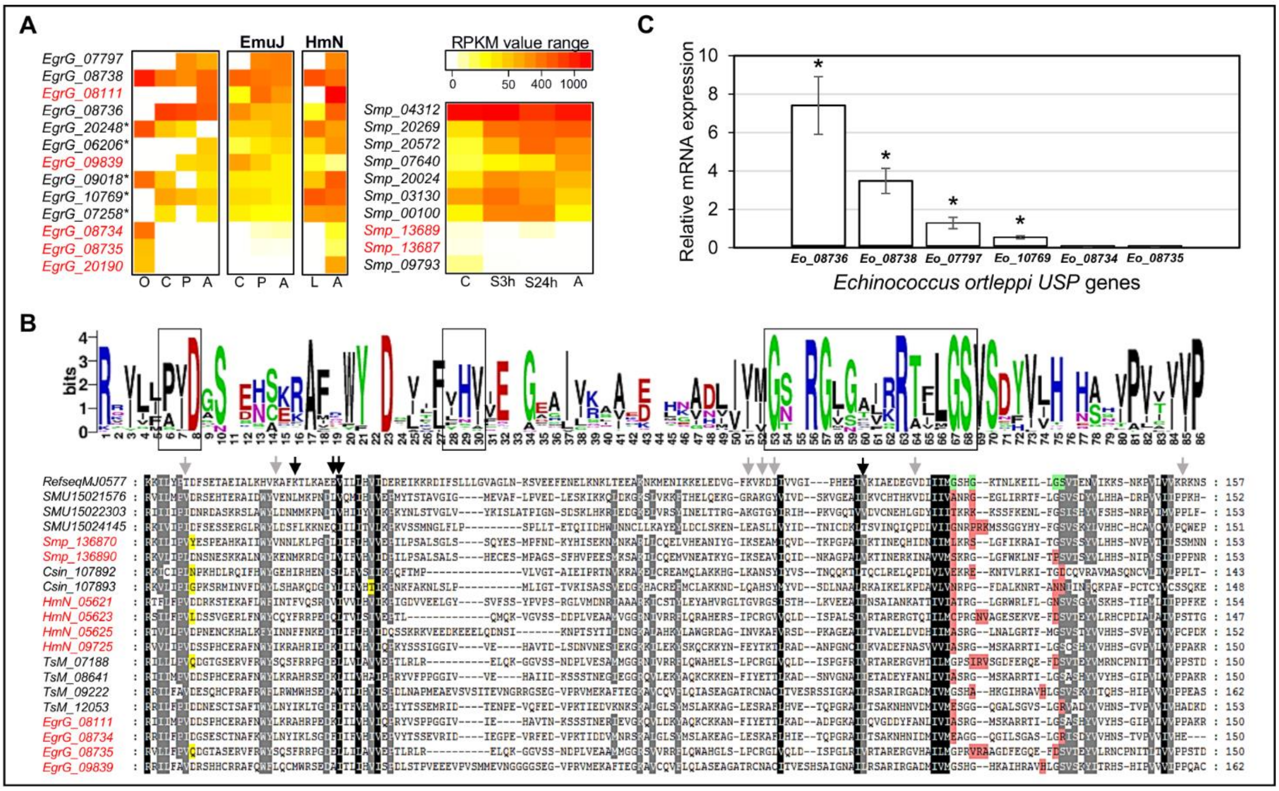
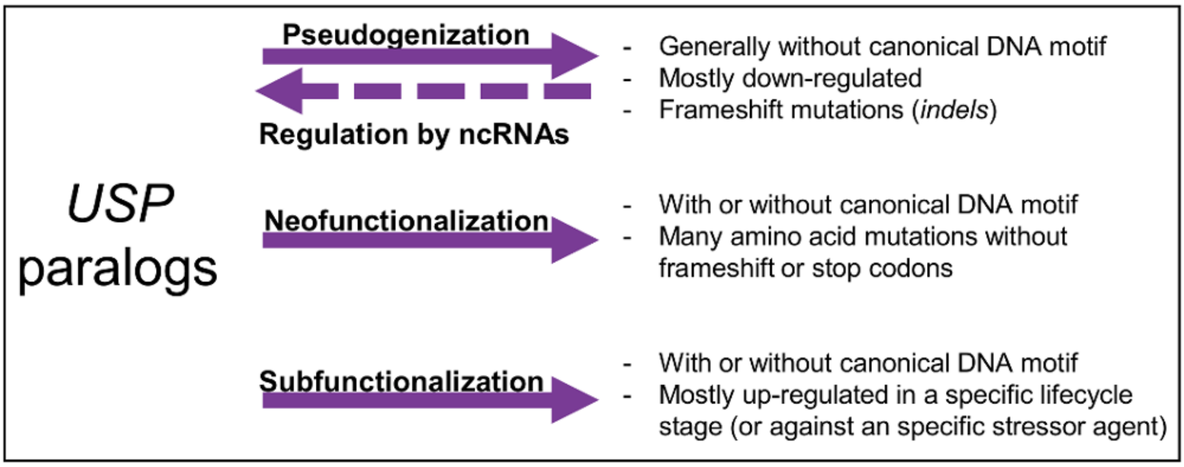

A
EmuJ
HmN
RPKM value range
0
50
400
1000
EgrG_07797
EgrG_08738
EgrG_08111
EgrG_08736
EgrG_20248*
EgrG_06206*
EgrG_09839
EgrG_09018*
EgrG_10769*
EgrG_07258*
EgrG_08734
EgrG_08735
EgrG_20190
O C P A
C P A
L A
Smp_04312
Smp_20269
Smp_20572
Smp_07640
Smp_20024
Smp_03130
Smp_00100
Smp_13689
Smp_13687
Smp_09793
C S3h S24h A
C
Relative mRNA expression
0
2
4
6
8
10
*
Eo_08736
Eo_08738
Eo_07797
Eo_10769
Eo_08734
Eo_08735
Echinococcus ortleppi USP genes
B
bits
RefseqMJ0577
SMU15021576
SMU15022303
SMU15024145
Smp_136870
Smp_136890
Csin_107892
Csin_107893
HmN_05621
HmN_05623
HmN_05625
HmN_09725
TsM_07188
TsM_08641
TsM_09222
TsM_12053
EgrG_08111
EgrG_08734
EgrG_08735
EgrG_09839

Figure 4


USP paralogs
Pseudogenization
Regulation by ncRNAs
- Generally without canonical DNA motif
- Mostly down-regulated
- Frameshift mutations (*indels*)
Neofunctionalization
- With or without canonical DNA motif
- Many amino acid mutations without frameshift or stop codons
Subfunctionalization
- With or without canonical DNA motif
- Mostly up-regulated in a specific lifecycle stage (or against an specific stressor agent)

Additional file 1

Table S1. Details of each primer designed for the six *USP* genes.

Amplification efficiency was calculated with LinRegPCR software. *EF-1α* (ampliffication efficiency of 90.4%) was used as a normalizer gene.

Gene ID	Primer sequence (Forward/Reverse) 5' -> 3'	Amplicon length (bp)	Tm	Amplification efficiency (%)
EgrG_001076900	GTAAAAGGCAAATCCTCCTCGCTG/CCCCAATCTTTTCTGACGGTTGAC	300	86.2	86.8
EgrG_000779700	CGGCTCAGTGAGTGTCGATGATG/TGAGAAGGTGGTGGGAGACGCTAC	227	88.6	86.8
EgrG_000873400	TGAAGTGCGAATCGACATAGAAGGC/TGCTCCAGACAAAATGCCCTGAC	253	85.5	81.7
EgrG_000873500	AGTTGGACATCAGTCCGGGCCT/GAAGTCGCCTGCAGCTCGTACA	101	87.8	87.7
EgrG_000873600	GGATTCGATGCGAAGGCTTTTG/CGGCTGCCAACGATGATATGGT	104	84.3	86.6
EgrG_000873800	GAACCTGTGTACACAACTCCAGCGA/AAGTCCACGGGAGCCAATAACGAT	255	86.3	86.9

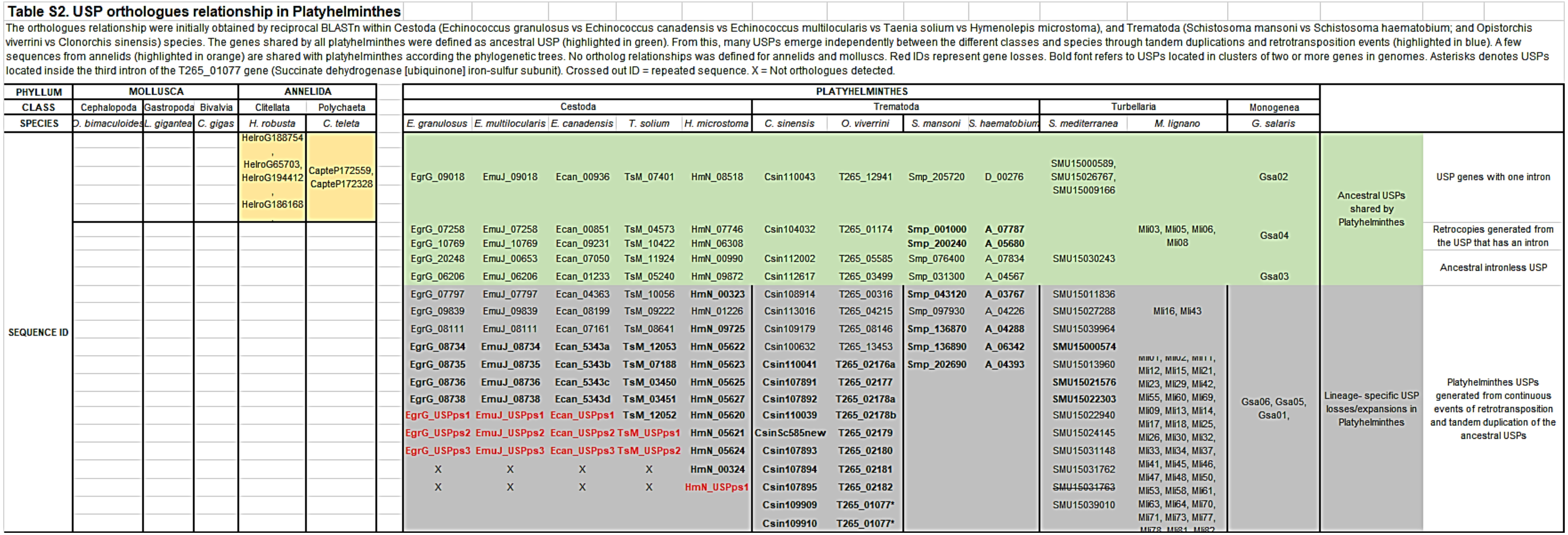


Table S2. USP orthologues relationship in Platyhelminthes

The orthologues relationship were initially obtained by reciprocal BLASTn within Cestoda (Echinococcus granulosus vs Echinococcus canadensis vs Echinococcus multilocularis vs Taenia solium vs Hymenolepis microstoma), and Trematoda (Schistosoma mansoni vs Schistosoma haematobium; and Opistorchis viverrini vs Clonorchis sinensis) species. The genes shared by all platyhelminthes were defined as ancestral USP (highlighted in green). From this, many USPs emerge independently between the different classes and species through tandem duplications and retrotransposition events (highlighted in blue). A few sequences from annelids (highlighted in orange) are shared with platyhelminthes according the phylogenetic trees. No ortholog relationships was defined for annelids and molluscs. Red IDs represent gene losses. Bold font refers to USPs located in clusters of two or more genes in genomes. Asterisks denotes USPs located inside the third intron of the T265_01077 gene (Succinate dehydrogenase [ubiquinone] iron-sulfur subunit). Crossed out ID = repeated sequence. X = Not orthologues detected.

PHYLLUM	**MOLLUSCA**			**ANNELIDA**		**PLATYHELMINTHES**													
CLASS	Cephalopoda	Gastropoda	Bivalvia	Clitellata	Polychaeta	Cestoda					Trematoda				Turbellaria		Monogenea		
SPECIES	*D. bimaculoides*	*L. gigantea*	*C. gigas*	*H. robusta*	*C. teleta*	*E. granulosus*	*E. multilocularis*	*E. canadensis*	*T. solium*	*H. microstoma*	*C. sinensis*	*O. viverrini*	*S. mansoni*	*S. haematobium*	*S. mediterranea*	*M. lignano*	*G. salaris*		
SEQUENCE ID				HelroG188754, HelroG65703, HelroG194412, HelroG186168	CapteP172559, CapteP172328	EgrG_09018	EmuJ_09018	Ecan_00936	TsM_07401	HmN_08518	Csin110043	T265_12941	Smp_205720	D_00276	SMU15000589, SMU15026767, SMU15009166		Gsa02	Ancestral USPs shared by Platyhelminthes	USP genes with one intron
						EgrG_07258	EmuJ_07258	Ecan_00851	TsM_04573	HmN_07746	Csin104032	T265_01174	**Smp_001000**	**A_07787**		Mli03, Mli05, Mli06, Mli08	Gsa04		Retrocopies generated from the USP that has an intron
						EgrG_10769	EmuJ_10769	Ecan_09231	TsM_10422	HmN_06308			**Smp_200240**	**A_05680**					
						EgrG_20248	EmuJ_00653	Ecan_07050	TsM_11924	HmN_00990	Csin112002	T265_05585	Smp_076400	A_07834	SMU15030243				Ancestral intronless USP
						EgrG_06206	EmuJ_06206	Ecan_01233	TsM_05240	HmN_09872	Csin112617	T265_03499	Smp_031300	A_04567			Gsa03		
						EgrG_07797	EmuJ_07797	Ecan_04363	TsM_10056	**HmN_00323**	Csin108914	T265_00316	**Smp_043120**	**A_03767**	SMU15011836		Gsa06, Gsa05, Gsa01,	Lineage- specific USP losses/expansions in Platyhelminthes	Platyhelminthes USPs generated from continuous events of retrotransposition and tandem duplication of the ancestral USPs
						EgrG_09839	EmuJ_09839	Ecan_08199	TsM_09222	HmN_01226	Csin113016	T265_04215	Smp_097930	A_04226	SMU15027288	Mli16, Mli43			
						EgrG_08111	EmuJ_08111	Ecan_07161	TsM_08641	**HmN_09725**	Csin109179	T265_08146	**Smp_136870**	**A_04288**	SMU15039964				
						EgrG_08734	**EmuJ_08734**	**Ecan_5343a**	**TsM_12053**	**HmN_05622**	Csin100632	T265_13453	**Smp_136890**	**A_06342**	**SMU15000574**				
						EgrG_08735	**EmuJ_08735**	**Ecan_5343b**	**TsM_07188**	**HmN_05623**	**Csin110041**	**T265_02176a**	**Smp_202690**	**A_04393**	SMU15013960	Mli01, Mli02, Mli11, Mli12, Mli15, Mli21, Mli23, Mli29, Mli42, Mli55, Mli60, Mli69, Mli09, Mli13, Mli14, Mli17, Mli18, Mli25, Mli26, Mli30, Mli32, Mli33, Mli34, Mli37, Mli41, Mli45, Mli46, Mli47, Mli48, Mli50, Mli53, Mli58, Mli61, Mli63, Mli64, Mli70, Mli71, Mli73, Mli77, Mli78, Mli81, Mli82			
						EgrG_08736	**EmuJ_08736**	**Ecan_5343c**	**TsM_03450**	**HmN_05625**	**Csin107891**	**T265_02177**			**SMU15021576**				
						EgrG_08738	**EmuJ_08738**	**Ecan_5343d**	**TsM_03451**	**HmN_05627**	**Csin107892**	**T265_02178a**			**SMU15022303**				
						EgrG_USPps1	**EmuJ_USPps1**	**Ecan_USPps1**	**TsM_12052**	**HmN_05620**	**Csin110039**	**T265_02178b**			SMU15022940				
						EgrG_USPps2	**EmuJ_USPps2**	**Ecan_USPps2**	**TsM_USPps1**	**HmN_05621**	**CsinSc585new**	**T265_02179**			SMU15024145				
						EgrG_USPps3	**EmuJ_USPps3**	**Ecan_USPps3**	**TsM_USPps2**	**HmN_05624**	**Csin107893**	**T265_02180**			SMU15031148				
						X	X	X	X	**HmN_00324**	**Csin107894**	**T265_02181**			SMU15031762				
						X	X	X	X	**HmN_USPps1**	**Csin107895**	**T265_02182**			~~SMU15031763~~				
											Csin109909	**T265_01077***			SMU15039010				
											Csin109910	**T265_01077***							

Table S3. Motif scanning in promoter regions of the *USP* genes.

The motif discovery was performed with the MEME suite tool. The typical heptanucleotide -***CCAATCA***- motif is represented by the color red. The other motif, where only the sequence -***CCAAT***- is conserved, is denoted in color orange. Except for TsM_09222 in Cestoda, and Csin109179, Csin107892, Smp_205720, Smp_043120, Smp_136890, D_00276, and A_06342 in Trematoda, all the DNA motifs are located between positions -50 and -200 (300-450 in the table) before the ATG start codon. *E. canadensis* and *E. multilocularis* (not shown) represent the same motifs than *E. granulosus* , mostly at the same position. The same is observed when comparing orthologues from *C. sinensis* vs *O. viverrini* and *S. mansoni* vs *S. haematobium* . Because the USP upstream region were not accurately retrieved for EgrG_USPps2 (and orthologues in *E. canadensis* and *E. multilocularis,* not shown), and TsM_USPps1, these sequences must be considered carefully. For all the motif found, the p-value were minor than 0.00001.

Cestoda class					Trematoda class				
Gene ID	Motif		Strand	Start/End	Gene ID	Motif		Strand	Start/End
EgrG_09018	🟥	🟧	Plus/Minus	394-403/418-427	Csin110043	🟥	🟧	Plus/Minus	363-372/394-403
EgrG_07258	🟥	🟧	Plus/Minus	427-436/405-414	Csin104032	🟥	🟧	Plus/Minus/Minus	337-346/369-378/403-412
EgrG_10769	🟥		Minus/Plus	406-415/423-432	Csin112002	🟥		Minus	411-420
EgrG_20248	🟥		Plus/Plus	348-357/414-423	Csin112617	🟥		Minus	379-388
EgrG_06206	Not detected				Csin108914	🟥		Plus	398-407
EgrG_07797	🟥		Minus	313-322	Csin113016	🟥		Minus	392-401
EgrG_09839	Not detected				Csin109179	🟥		Plus	238-247
EgrG_08111		🟧	Minus	395-404	Csin100632	🟥		Minus	368-377
EgrG_08734		🟧	Plus	439-448	Csin110041	🟥	🟧	Plus/Minus	383-392/407-416
EgrG_08735	Not detected				Csin107891	🟥		Plus	415-424
EgrG_08736	🟥	🟧	Plus/Plus	427-436/395-404	Csin107892	🟥		Plus	92-101
EgrG_08738	🟥	🟧	Plus/Plus	427-436/399-408	Csin110039	🟥		Plus	417-426
EgrG_USPps1	Not detected				CsinSc585new	🟥		Plus/Plus	383-392/406-415
EgrG_USPps2	Not detected				Csin107893		🟧	Plus	381-390
EgrG_USPps3	Not detected				Csin107894	🟥		Plus	424-433
TsM_10422	🟥		Minus/Plus	403-412/421-430	Csin107895	Not detected			
TsM_07401	🟥	🟧	Plus/Minus	392-401/418-427	Csin109909	🟥		Minus	426-435
TsM_04573	🟥	🟧	Minus/Plus	404-413/427-436	Csin109910	🟥		Minus	381-390
TsM_11924	🟥		Minus/Plus/Plus	291-300/345-354/409-418	T265_12941	🟥	🟧	Plus/Minus	378-387/409-418
TsM_05240		🟧	Minus/Minus	342-351/372-381	T265_01174	🟥		Plus/Minus	337-346/403-412
TsM_10056	Not detected				T265_05585	🟥		Minus	411-420
TsM_09222	🟥		Plus	141-150	T265_03499/12260	🟥	🟧	Plus/Minus/Plus	368-377/400-409/439-448
TsM_08641		🟧	Minus	395-404	T265_00316	🟥		Plus	398-407
TsM_12053	🟥		Plus	438-447	T265_04215	🟥		Minus	446-455
TsM_07188	Not detected				T265_08146		🟧	Plus	418-427
TsM_03450	🟥	🟧	Plus/Plus	426-435/394-403	T265_13453	🟥		Minus	367-376
TsM_03451	🟥	🟧	Plus/Plus	427-436/397-406	T265_02176a	🟥		Plus	383-392
TsM_12052		🟧	Plus/Plus	155-164/381-390	T265_02177		🟧	Plus	425-434
TsM_USPps1	Not detected				T265_02178a	🟥		Plus	416-425
TsM_USPps2	Not detected				T265_02178b	🟥		Plus	415-424
HmN_08518	🟥	🟧	Plus/Minus	387-396/413-422	T265_02179	🟥		Plus	383-392
HmN_07746	🟥	🟧	Minus/Plus	412-421/368-377	T265_02180		🟧	Plus	380-389
HmN_06308	🟥		Plus	419-428	T265_02181	🟥		Plus	424-433
HmN_00990		🟧	Plus/Minus	355-364/383-392	T265_02182	🟥		Plus	417-426
HmN_09872	🟥		Plus/Minus/Minus	309-318/346-355/401-410	Smp_205720		🟧	Plus	248-257
HmN_00324	Not detected				Smp_001000	🟥		Plus/Plus/Plus/Minus	301-310/326-335/356-365/394-403
HmN_01226	Not detected				Smp_200240	🟥	🟧	Plus/Plus/Plus	328-337/356-365/385-394
HmN_09725	Not detected				Smp_076400	🟥	🟧	Minus/Plus	293-302/330-339
HmN_05622		🟧	Minus	447-456	Smp_031300	🟥		Plus	376-385
HmN_05623	Not detected				Smp_043120		🟧	Minus	258-267
HmN_05625	🟥	🟧	Plus/Plus	391-400/427-436	Smp_097930	Not detected			
HmN_05627	🟥	🟧	Plus/Plus	399-408/427-436	Smp_136870	🟥		Plus/Plus	384-393/411-420
HmN_05620		🟧	Minus	323-332	Smp_136890	🟥		Plus	235-244
HmN_05621	Not detected				Smp_202690	🟥		Plus	404-413
HmN_05624	Not detected				D_00276		🟧	Plus/Plus	285-294/319-328
HmN_00323	🟥		Minus	392-401	A_07787	🟥		Plus/Minus	370-379/407-416
HmN_USPps1	Not detected				A_05680		🟧	Plus	335-344
					A_07834	🟥	🟧	Minus/Plus	294-303/333-342
					A_04567	🟥		Plus	376-385
					A_03767		🟧	Minus/Minus	278-287/369-378
					A_04226	Not detected			
					A_04288	🟥		Plus/Plus	373-382/400-409
					A_06342		🟧	Plus	231-240
					A_04393	🟥		Plus	404-413

Table S4. SELECTON results for Cestoda species

SELECTON results for 63 codon sequences from Cestoda species. Colors (from orange to pink) were adapted from the SELECTON server program. EgrG_08736 was used as reference sequence. Amino acids with positive selection in PAML (M7 vs. M8) are indicated with one (probability > 95%) and two asterisks (probability > 99%). EgrG_08736 was used as reference sequence.

Selecton color range	
7	Purifying selection
6	
5	
4	
3	
2	
1	Positive selection

Site	Amino acid
1	R
2	K*
3	Y
4	L
5	L
6	P
7	I
8	D
9	T
10	S*
11	E
12	N
13	C
14	K
15	R**
16	A
17	A
18	K
19	F
20	Y
21	S
22	E
23	N
24	L
25	H
26	R
27	G**
28	D
29	D
30	T
31	I
32	I
33	F
34	L
35	H
36	V
37	I
38	E
39	P
40	N*
41	L
42	K
43	A*
44	N*
45	S*
46	K*
47	N
48	I
49	A
50	R
51	K
52	E
53	D
54	S*
55	E*
56	I
57	H
58	E
59	T
60	T
61	S*
62	Q*
63	L
64	Q
65	K
66	T*
67	V
68	D
69	M**
70	G
71	K
72	A*
73	L
74	G
75	H
76	K
77	Y
78	L
79	A
80	W
81	G
82	R
83	E
84	A
85	G
86	F
87	D
88	A
89	K
90	A
91	F
92	V
93	R
94	S
95	D
96	S
97	K
98	P
99	G
100	V
101	A
102	I
103	I
104	K*
105	A*
106	A
107	K*
108	E
109	L
110	E*
111	V
112	D
113	H
114	I
115	I
116	V
117	G
118	S
119	R
120	G
121	L
122	N
123	A
124	L
125	G
126	R
127	T
128	L
129	L
130	G
131	S
132	V
133	S
134	S
135	Y
136	V
137	I
138	H
139	H
140	S
141	Q**
142	V
143	P
144	V
145	T
146	V
147	V
148	P
149	C

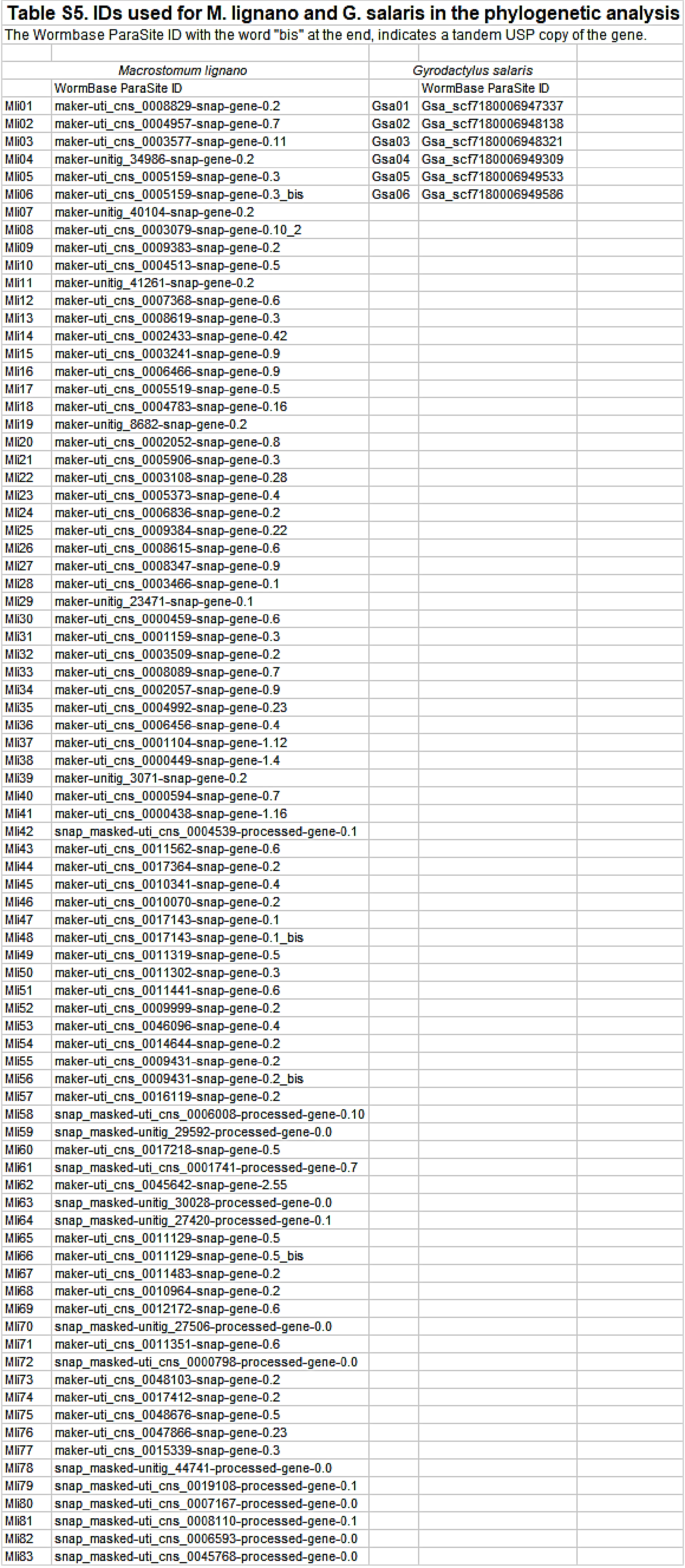


Table S5. IDs used for M. lignano and G. salaris in the phylogenetic analysis

The Wormbase ParaSite ID with the word "bis" at the end, indicates a tandem USP copy of the gene.

Macrostomum lignano		*Gyrodactylus salaris*		
	WormBase ParaSite ID		WormBase ParaSite ID	
Mli01	maker-uti_cns_0008829-snap-gene-0.2	Gsa01	Gsa_scf7180006947337	
Mli02	maker-uti_cns_0004957-snap-gene-0.7	Gsa02	Gsa_scf7180006948138	
Mli03	maker-uti_cns_0003577-snap-gene-0.11	Gsa03	Gsa_scf7180006948321	
Mli04	maker-unitig_34986-snap-gene-0.2	Gsa04	Gsa_scf7180006949309	
Mli05	maker-uti_cns_0005159-snap-gene-0.3	Gsa05	Gsa_scf7180006949533	
Mli06	maker-uti_cns_0005159-snap-gene-0.3_bis	Gsa06	Gsa_scf7180006949586	
Mli07	maker-unitig_40104-snap-gene-0.2			
Mli08	maker-uti_cns_0003079-snap-gene-0.10_2			
Mli09	maker-uti_cns_0009383-snap-gene-0.2			
Mli10	maker-uti_cns_0004513-snap-gene-0.5			
Mli11	maker-unitig_41261-snap-gene-0.2			
Mli12	maker-uti_cns_0007368-snap-gene-0.6			
Mli13	maker-uti_cns_0008619-snap-gene-0.3			
Mli14	maker-uti_cns_0002433-snap-gene-0.42			
Mli15	maker-uti_cns_0003241-snap-gene-0.9			
Mli16	maker-uti_cns_0006466-snap-gene-0.9			
Mli17	maker-uti_cns_0005519-snap-gene-0.5			
Mli18	maker-uti_cns_0004783-snap-gene-0.16			
Mli19	maker-unitig_8682-snap-gene-0.2			
Mli20	maker-uti_cns_0002052-snap-gene-0.8			
Mli21	maker-uti_cns_0005906-snap-gene-0.3			
Mli22	maker-uti_cns_0003108-snap-gene-0.28			
Mli23	maker-uti_cns_0005373-snap-gene-0.4			
Mli24	maker-uti_cns_0006836-snap-gene-0.2			
Mli25	maker-uti_cns_0009384-snap-gene-0.22			
Mli26	maker-uti_cns_0008615-snap-gene-0.6			
Mli27	maker-uti_cns_0008347-snap-gene-0.9			
Mli28	maker-uti_cns_0003466-snap-gene-0.1			
Mli29	maker-unitig_23471-snap-gene-0.1			
Mli30	maker-uti_cns_0000459-snap-gene-0.6			
Mli31	maker-uti_cns_0001159-snap-gene-0.3			
Mli32	maker-uti_cns_0003509-snap-gene-0.2			
Mli33	maker-uti_cns_0008089-snap-gene-0.7			
Mli34	maker-uti_cns_0002057-snap-gene-0.9			
Mli35	maker-uti_cns_0004992-snap-gene-0.23			
Mli36	maker-uti_cns_0006456-snap-gene-0.4			
Mli37	maker-uti_cns_0001104-snap-gene-1.12			
Mli38	maker-uti_cns_0000449-snap-gene-1.4			
Mli39	maker-unitig_3071-snap-gene-0.2			
Mli40	maker-uti_cns_0000594-snap-gene-0.7			
Mli41	maker-uti_cns_0000438-snap-gene-1.16			
Mli42	snap_masked-uti_cns_0004539-processed-gene-0.1			
Mli43	maker-uti_cns_0011562-snap-gene-0.6			
Mli44	maker-uti_cns_0017364-snap-gene-0.2			
Mli45	maker-uti_cns_0010341-snap-gene-0.4			
Mli46	maker-uti_cns_0010070-snap-gene-0.2			
Mli47	maker-uti_cns_0017143-snap-gene-0.1			
Mli48	maker-uti_cns_0017143-snap-gene-0.1_bis			
Mli49	maker-uti_cns_0011319-snap-gene-0.5			
Mli50	maker-uti_cns_0011302-snap-gene-0.3			
Mli51	maker-uti_cns_0011441-snap-gene-0.6			
Mli52	maker-uti_cns_0009999-snap-gene-0.2			
Mli53	maker-uti_cns_0046096-snap-gene-0.4			
Mli54	maker-uti_cns_0014644-snap-gene-0.2			
Mli55	maker-uti_cns_0009431-snap-gene-0.2			
Mli56	maker-uti_cns_0009431-snap-gene-0.2_bis			
Mli57	maker-uti_cns_0016119-snap-gene-0.2			
Mli58	snap_masked-uti_cns_0006008-processed-gene-0.10			
Mli59	snap_masked-unitig_29592-processed-gene-0.0			
Mli60	maker-uti_cns_0017218-snap-gene-0.5			
Mli61	snap_masked-uti_cns_0001741-processed-gene-0.7			
Mli62	maker-uti_cns_0045642-snap-gene-2.55			
Mli63	snap_masked-unitig_30028-processed-gene-0.0			
Mli64	snap_masked-unitig_27420-processed-gene-0.1			
Mli65	maker-uti_cns_0011129-snap-gene-0.5			
Mli66	maker-uti_cns_0011129-snap-gene-0.5_bis			
Mli67	maker-uti_cns_0011483-snap-gene-0.2			
Mli68	maker-uti_cns_0010964-snap-gene-0.2			
Mli69	maker-uti_cns_0012172-snap-gene-0.6			
Mli70	snap_masked-unitig_27506-processed-gene-0.0			
Mli71	maker-uti_cns_0011351-snap-gene-0.6			
Mli72	snap_masked-uti_cns_0000798-processed-gene-0.0			
Mli73	maker-uti_cns_0048103-snap-gene-0.2			
Mli74	maker-uti_cns_0017412-snap-gene-0.2			
Mli75	maker-uti_cns_0048676-snap-gene-0.5			
Mli76	maker-uti_cns_0047866-snap-gene-0.23			
Mli77	maker-uti_cns_0015339-snap-gene-0.3			
Mli78	snap_masked-unitig_44741-processed-gene-0.0			
Mli79	snap_masked-uti_cns_0019108-processed-gene-0.1			
Mli80	snap_masked-uti_cns_0007167-processed-gene-0.0			
Mli81	snap_masked-uti_cns_0008110-processed-gene-0.1			
Mli82	snap_masked-uti_cns_0006593-processed-gene-0.0			
Mli83	snap_masked-uti_cns_0045768-processed-gene-0.0			

Capítulo II

Heterologous expression and functional analysis of a Universal Stress Protein from *Echinococcus ortleppi.*

Sergio Martin Espinola[1,2], Maria Eduarda Matos Marques[2], Martin Pablo Cancela[2,3], Liziane Raquel Beckenkamp[4], Márcia Rosângela Wink[4], Henrique Bunselmeyer Ferreira[2,3], Arnaldo Zaha[1,2,3*]

[1] Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil

[2] Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil

[3] Programa de Pós-Graduação em Biologia Celular e Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil

[4] Departamento de Ciências Básicas da Saúde e Laboratório de Biologia Celular, Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre, RS, Brasil

E-mail: s.espinola@gmail.com, Sergio Martin Espinola
mariaeduarda.matos@gmail.com, Maria Eduarda Matos Marques
martin@cbiot.ufrgs.br, Martin Pablo Cancela
lizianeraquel@yahoo.com.br, Liziane Raquel Beckenkamp
mwink@ufcspa.edu.br, Márcia Rosângela Wink
henrique@cbiot.ufrgs.br, Henrique Bunselmeyer Ferreira
zaha@cbiot.ufrgs.br, Arnaldo Zaha

[*] Author for correspondence

Abstract

Universal stress proteins (USPs) belongs to a gene family with a critical role in the maintenance of the cellular homeostasis. They help the organisms to face diverse types of biotic and abiotic stresses, including heat shock, oxidative damage, and nutrient starvation. It is known that USPs are capable to bind ATP and analogs, however, the different pathways in which USPs are involved, and the molecular mechanisms that control and trigger their expression against stressor agents, are still unclear. In this work, we perform the cloning, expression, and purification of a recombinant USP from *Echinococcus ortleppi*, causative of the cyst echinococcosis. We assessed the nucleotide binding capacity and the ATPase activity of rEoUSP-1, as well as its ability to form oligomers. Our results show that rEoUSP-1 is unable to interact with AMP, but may bind ATP. The ATPase activity was detected only at high concentrations of the recombinant protein, suggesting that other cellular compounds may be needed for the catalytic function. Structurally, monomers of rEgUSP36 form dimers rapidly and stably along time, which agree with previous results obtained in the literature. We conclude that rEoUSP-1 probably bind and hydrolyze ATP, and is capable to form the oligomeric state that is suggested to achieve their functions. Further analysis will help in a better characterization of this protein. In animals, this is the first functional characterization of a recombinant USP, highlighting their relevance not only in eukaryotes, but also regarding the host-parasite interactions.

Introduction

Universal Stress Proteins (USPs) belong to a gene family with a multifunctional behavior against stressor agents. Their upregulation in response to oxidative damage, heat shock, UV irradiation, nutrient starvation, pathogen infection, among others, has been analyzed in a wide range of organisms, particularly in bacteria and plants (Gustavsson *et al.*, 2002; Liu *et al.*, 2007; Drumm *et al.*, 2009; Udawat *et al.*, 2014; Jung *et al.*, 2015; Soufi *et al.*, 2015). Besides the high levels of gene expression upon stress induction, they appear to be essential for survival, growth, and the persistence of chronic infection (Liu *et al.*, 2007; Drumm *et al.*, 2009; Soufi *et al.*, 2015; Glass *et al.*, 2017). The involvement of USPs in such diverse biological processes and environmental conditions, suggest a significant role for these genes in the maintenance of the cellular homeostasis.

The typical USP structure display a Rossman-like α/β-fold with five parallel β-strands and four α-helices, which is capable to form dimers, trimers, and higher complexes (Weber and Jung, 2006; Nachin *et al.*, 2008; Drumm *et al.*, 2009; Jung *et al.*, 2015; Gutiérrez-Beltrán *et al.*, 2017). In general, based in the homology with the *E. coli* UspA, two Usp subclasses can be distinguished: the non-ATP and the ATP-binding UspA homologues (Zarembinski *et al.*, 1998; Sousa and Mckay, 2001; Weber and Jung, 2006; Drumm *et al.*, 2009). The last contains a conserved ATP binding motif [Gx2Gx9G(S/T)] and is capable of interacting with several nucleotides, such as the USP MJ0577 of *Metchanococcus jannaschii* (PDB ID 1MJH) (Zarembinski *et al.*, 1998). The other, exhibits modifications in their ATP-binding motif, and thus, lack the ability to bind and hydrolyze ATP (e.g. PDB ID 1JMV from *Haemophilus influenza*) (Sousa and Mckay, 2001). The ATP-binding pocket is structurally different from other commonly present in ATP binding proteins (Traut, 1994; Zarembinski *et al.*, 1998) and the interaction with nucleotides appear to be relevant in several biological process. Structure-guided mutagenesis of the USP Rv2623 in *Mycobacterium tuberculosis* impaired its ability to bind ATP, leading to an attenuated growth phenotype (Drumm *et al.*, 2009; Glass *et al.*, 2017). In the same species, the USP Rv1636 binds specifically and with high affinity to cAMP compared to ATP. The authors suggest that Rv1636 can be acting as a “sink” for this second messenger, having an important role in the signal transduction cascades (Banerjee *et al.*, 2015). In contrast, there are USPs that lack the capability to bind and hydrolyze nucleotides, but maintain their function as stress responsive proteins. Although a chaperone function was suggested (Bochkareva *et al.*, 2002; Jung *et al.*, 2015) the molecular mechanisms that control and trigger their expression against stress agents are still unknown.

Echinoccocosis is a neglected tropical disease caused by species of the genus *Echinococcus* (Cestoda, Platyhelminthes), with a high negative impact regarding public health and livestock industry. *Echinococcus granulosus*, with a global distribution and responsible for cyst echinococcosis, and *Echinococcus multilocularis*, restricted to the north hemisphere and causative of alveolar echinococcosis, are the two major species of the genus, leading to more than 1 million of people infected worldwide (World Health Statistics 2017, World Health Organization). *Echinococcus ortleppi* belongs to the *E. granulosus* sensu lato complex, and is becoming a threat because of its high frequency in the cattle of southern Brazil, and human cases around the world (de la Rue *et al.*, 2011; Balbinotti *et al.*, 2012; Grenouillet *et al.*, 2014). Like other zoonosis, *Echinococcus* species must infect two hosts to complete its lifecycle. The adult worms develop inside the intestine of canids (definitive host) and infective eggs are released together with feces into the external environmental. The infective eggs can be ingested by the intermediate hosts (mammals, and accidentally, humans), starting the larval stage, with the formation of the hydatid cyst. If viscera of slaughtered animals are ingested by canids, the development of the protoscoleces to the adult form takes place, thus, concluding the lifecycle (Thompson, 1995). Due to the intrinsic nature of the host-parasite relationships, the number of biotic and abiotic stressor agents during the parasite lifecycle seems to be very high and diverse. The effect of the variation of the external environment against the parasite eggs (temperature, relative humidity), and the defense mechanisms in the larval stage (parasite layers, immunomodulatory proteins, redox system, etc.), is well described and widely analyzed in the literature (Upatham, 1973; Coman, 1975; Veit *et al.*, 1995; McCreesh and Booth, 2014; Hui *et al.*, 2015; Huang *et al.*, 2016; Sánchez Thevenet *et al.*, 2017). In a previous work, we have shown that USPs follows different evolutionary fates, which leads to a functional diversification of this gene family in species of the phylum Platyhelminthes (Espinola *et al.*, submitted). In this manner, the functional characterization of the USPs based on the structure and the ligand-binding properties, will allow expanding our knowledge of these important group of genes. Here, we cloned and produced a recombinant USP from *Echinococcus ortleppi* and analyzed its ATP-binding and hydrolysis activity, and its ability to self-assembly in higher oligomeric complexes.

Methods

Parasite material

Hydatid cysts were obtained from bovines infected with *Echinococcus* in the endemic region of Rio Grande do Sul (RS), Brazil, and kindly provided by the Cooperleo Abbatoir (São Leopoldo, RS). Protoescoleces were recovered by punction and aspiration of hydatid liquid, and washed several times with PBS 1X. Viability of parasites was assessed by optical microscope observation. Species determination was performed by the amplification of the mitochondrial *cox1* gene, followed by *Alu1* digestion, as previously described (Bowles *et al.*, 1992; Avila *et al.*, 2017).

RNA extraction, cDNA synthesis, and molecular cloning

Total RNA from *E. ortleppi* protoscoleces was isolated with the TRIzol reagent (Thermo Fischer Scientific). RNA samples were treated with DNase I (Sigma-Aldrich), and the first cDNA strand were synthetized using the RevertAid transcriptase (Thermo Fischer Scientific) with an oligo (dT) primer, following the manufacture instructions. The complete coding sequence of the *E. ortleppi* gene (named as *EoUSP-1*) corresponding to *EgrG_000873600* in *E. granulosus* (WormBase ParaSite website, http://parasite.wormbase.org/index.html) was amplified by PCR using specific primers (FI 5’-TATTTTCAGGGAGAATTCCCGGGTATGGACGCAAATACCTTTTGCCAA-3’ and RI 5’-GCGAGGCAGATCGTCAGTCAGTCATTAGTAGAGCGACAGCCGTCGCATG-3’), with an additional sequence (underlined nucleotides) matching the cloning vector pGEX-TEV. The PCR product was used as template for a second PCR reaction where a 26-nt fragment (underlined nucleotides) was included (FII 5′-TGGTTCCGCGTGGATCTGAAAACCTGTATTTTCAGGGAGAATTCCCGGGT-3′ and RII (5′-GGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGATCGTCAGTCAGTCA-3′) to their 5’ ends to increase the homology with the vector. Molecular cloning was achieved by *in vivo* homologous recombination with pGEX-TEV using *Escherichia coli* KC8 cells.

Protein expression and purification

Recombinant plasmids were used for transformation into *E. coli* BL21 Star (DE3). GST-rEoUSP-1 expression was induced with isopropyl β-D-1-thiogalactopyranoside (IPTG) 0.1 mM for 3 hours at 37 °C. Cells were pelleted by centrifugation (6000 x g), resuspended in PBS 1X, and lysed by sonication. GST-rEoUSP-1 were recovered from soluble fraction by affinity chromatography in Glutathione Sepharose 4B (GE Healthcare), followed by cleavage with the protease TEV for 16 h at 37 °C. Induced and purified proteins were observed in 12% SDS-PAGE stained with comassie blue G-250 (Sigma). Protein quantification was performed with Qubit (Quant-iT Protein Assay Kit, Invitrogen) and samples were stored in PBS 1X at -20 °C.

Nucleotide binding assays

The affinity of the rEoUSP-1 for adenosine triphosphate (ATP) and adenosine monophosphate (AMP) was tested using these nucleotides immobilized in an agarose matrix (Sigma). The resins were equilibrated with buffer A (25 mM Hepes pH 7.6, 50 mM KCl, 2.5 mM $MgCl_2$, 10 mM 2-mercaptoethanol, and 1 mM EDTA) and incubated with 200µg of rUSP36 for 16 hours at 4 ^{o}C. After one wash with buffer A, three washes with buffer B (buffer A + 0.5 M NaCl), and other wash with buffer A, proteins were eluted with buffer C (buffer A plus 20mM ATP for ATP agarose or 10mM NADP+ for AMP agarose) (Roth *et al.*, 2002; Wu *et al.*, 2004). Nonspecifically bound proteins were eluted with 7 M urea. Glutathione S-transferase (GST) and bovine serum albumin were used as negative controls, and followed the same steps as the rEoUSP-1.

Measurement of ATP hydrolysis

The purified protein was dialyzed against TBS 1X to remove any trace of phosphate from the buffer. To determine the ATP hydrolysis, the reaction mixture containing 2 mM ATP, 2 mM $CaCl_2$, 120 mM NaCl, 5 mM KCl, 10 mM glucose, 20 mM Hepes, pH 7.4, was incubated with different concentrations of rEoUSP-1 (0-50 µg) at 37ºC or 45ºC for 30 and 60 minutes in a final volume of 0.2 mL. The reaction was stopped by the addition of 0.2 mL of trichloroacetic acid (TCA) to a final concentration of 5% (w/v). The samples were chilled on ice and the amount of inorganic phosphate (Pi) liberated was measured by the method of Chan *et al*. (1986), with KH_2PO_4 as the Pi standard. In order to correct non-enzymatic hydrolysis, we performed controls by adding the protein after the reaction was stopped with TCA. All samples were assayed in duplicate in at least three independent experiments. Enzyme activities were expressed as nanomoles of Pi released. GST was used as negative control (non-ATP hydrolyzing enzyme).

Cross-linking assay

Cross-linking experiments were performed as described previously (Darawshe *et al.*, 1987; Monteiro *et al.*, 2007), with minor modifications. Briefly, 70µL of rEoUSP-1 (1mg/mL, PBS 1X buffer) was mixed with 0,1% of glutaraldehyde (v/v) at room temperature. Seven aliquots of 10µL were collected at different times (0.5, 1, 5, 10, 20, 30, and 60 minutes) and reactions were stopped with SDS and boiling for 5 minutes. Cross-linked products were analyzed in SDS-PAGE 15% and visualized by coomassie blue staining.

Results and Discussion

Protein expression and purification

The coding sequence (501 nt) of the *EgrG_000873600* ortholog gene in *E. ortleppi* (*EoUSP-1*) was cloned into the pGEX-TEV vector and expressed in *E. coli* BL21 Star (DE3). After induction with IPTG, we obtain a ~44.5 kDa product that correspond to the rEoUSP-1 protein (18.5 kDa, predicted by ExPASy) fused to GST (26 kDa) (Figure 1A). The cleavage with the TEV protease allow separating this complex, and thus, the isolation of the rEoUSP-1, with a yield of 5 mg per liter of culture (Figure 1B). Similar results were described for other recombinant USP proteins expressed in *E. coli* cells, with a molecular weights ranging from 17 to 18 kDa (Weber and Jung, 2006; Sagurthi *et al.*, 2007; Udawat *et al.*, 2014; De Souza *et al.*, 2016).

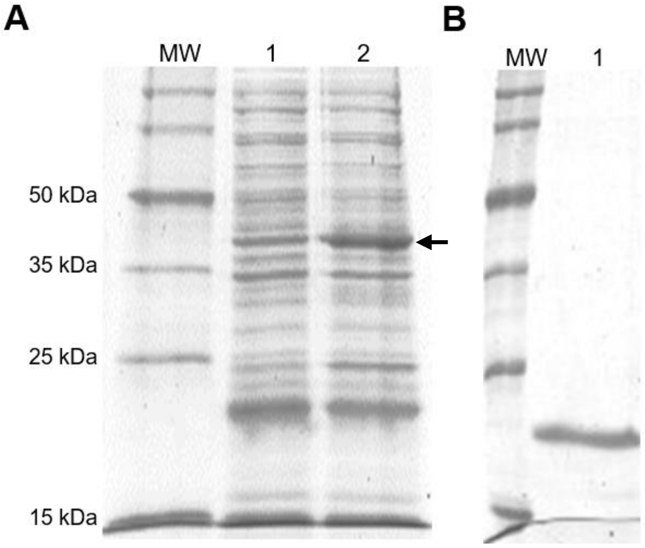


Figure 1. Expression and purification of rEgrUSP36. A) Expression of rEoUSP-1 fused to GST (~44.5 kDa, black arrow) was induced with 0.1 mM of IPTG using *E. coli* BL21 Star (DE3) cells. Line 1, not induced cells, line 2, induced cells. **B)**. Purified rEoUSP-1 obtained after TEV cleavage. Line 1, purified rEoUSP-1 with a predicted molecular weight of 18.8 kDa. MW, molecular weight marker.

Nucleotide binding and ATP hydrolyze

The affinity of rEoUSP-1 by AMP or ATP was analyzed by incubation of these nucleotides immobilized in a resin, and visualized by SDS-PAGE 12% gel, after. No interaction was detected with AMP (Figure S1, Additional information). On the other hand, although rEoUSP-1 bound to the ATP resin (Figure 2 A), the protein was not eluted in the presence of 20 mM of ATP, neither at higher concentrations of this nucleotide (100 mM, data not shown). This result shows that rEoUSP-1 binds ATP, however, this interaction seems to be nonspecific. Perhaps if the ligand-protein affinity is high, ATP molecules may already be bound to the recombinant protein when expressed in *E coli* cells. The ATP binding pocket will be blocked and unavailable to interact with the ATP from resin, and thus, any contact between rEoUSP-1 and immobilized ATP (or agarose matrix) will result in a nonspecific interaction. In order to identify the possibility of nucleotide hydrolysis, we analyze the ATPase activity using the malachite green reagent, a colorimetric assay based in the quantification of the Pi released (Chan *et al.*, 1986). Using GST as control and different concentrations of proteins, no ATPase activity was detected at 30 minutes of incubation at 37 ^{o}C for almost all protein concentrations (Figure 2 B). However, a significant difference was observed at the highest rEoUSP-1 concentration (50 µg, Figure 2 A), suggesting that rEoUSP-1 may require other cellular components to accomplish its ATPase activity (Zarembinski *et al.*, 1998). Increasing the incubation time (1 hour) or temperature (45 ^{o}C), generated the same results (Figure S2, Additional information). Because EgrG_000873600 (EoUSP-1 ortholog in *E. granulosus*) preserves the specific amino acids making the ATP-binding motif, and a proposed protein model suggests its interaction with ATP (Espinola *et al.*, submitted), the nucleotide binding ability remains unclear. Moreover, posttranslational modifications were described for USPs (Weber and Jung, 2006; Glass *et al.*, 2017). The USP Rv2623 from *M. tuberculosis* interact with a forkhead-associated (FHA) domain of the ATP-binding cassette transporter. The absence of phosphorylation in a threonine of Rv2623 (T237) leads to a diminished interaction with the FHA domain, also affecting the growth-regulatory capacity (Glass *et al.*, 2017). If rEoUSP-1 requires other cellular components (cooperation) or specific posttranslational eukaryotic tags to interact or enhance its affinity by nucleotides, the protein-ligand contact may be affected.

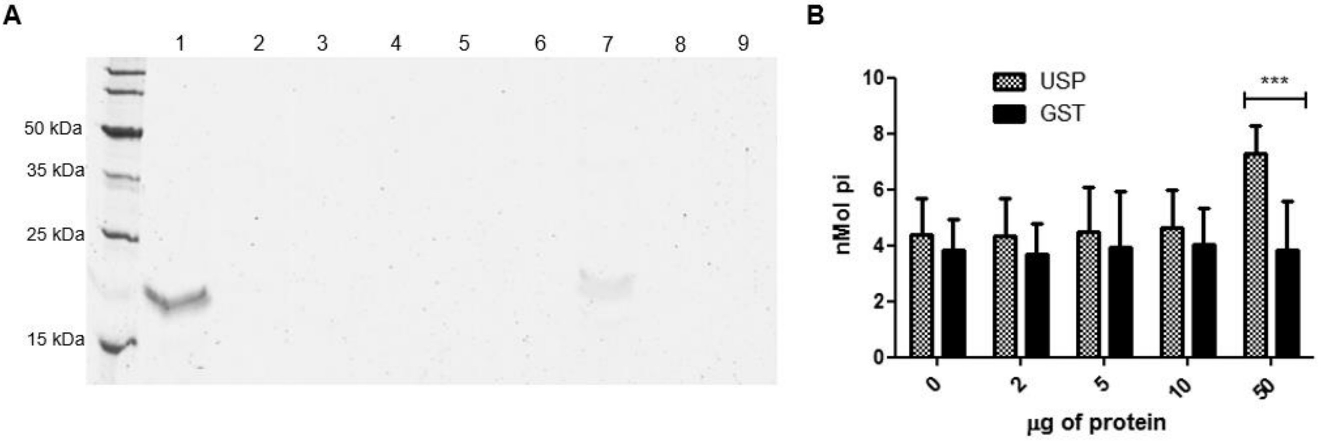


Figure 2. ATP binding and ATPase activity. A) The ATP-binding assay show that rEoUSP-1 is interacting with resin (line 1), contrary to GST or BSA (lines 2, 3). No protein was eluted using 20 mM of ATP (lines 4 for rEoUSP-1, line 5 for GST, and 6 for BSA), however rEoUSP-1 was finally eluted with urea 7 M (line 7; GST and BSA in lines 8 and 9, respectively). **B)** The rEoUSP-1 ATPase activity shows differences in relation to the control at high protein concentration at 37 ^{o}C for 30 minutes, suggesting that hydrolytic activity may need other molecular partners to achieve its function (Zarembinski *et al*., 1998). The increase of the incubation time (1 hour) or temperature (45 ^{o}C), yielded similar results (Figure S2, Additional information).

Glutaraldehyde cross-linking

To gain insights regarding the oligomerization states, the self-association of rEoUSP-1 was assessed by glutaraldehyde cross-linking. We identified the formation of dimers after 0.5 minutes of incubation, the products remaining stable until the end of the experiment (60 minutes) (Figure 3). Both monomers and dimers showed a faster migration in gel after the treatment with glutaraldehyde. The presence of two cysteine residues in the EoUSP-1, could be affecting the mobility (Griffith, 1972), which generates a band of ~16 kDa for the monomer and ~32 kDa for the dimer, when compared with the molecular marker. We also observed the presence of protein complexes higher than 50 kDa, particularly after 10 minutes (Figure 3). The shorter the time of cross-linking, the closer the interactions between oligomers (Monteiro *et al.*, 2007). These results show the rapid establishment of protein dimers of rEoUSP-1, suggesting a tight proximity between monomers with stable interactions along time. Generally, when crystallized, USPs are in the dimeric form, most commonly as homodimers (Zarembinski *et al.*, 1998; Sousa and Mckay, 2001; Kim *et al.*, 2015). However, heterodimers and higher oligomeric states have already been described (Nachin *et al.*, 2008; Jung *et al.*, 2015). The ability to produce these versatile associations can have an important functional effect, allowing the response against a specific stressor, or, at the same time, promoting overlapping functions (Nachin *et al.*, 2005).

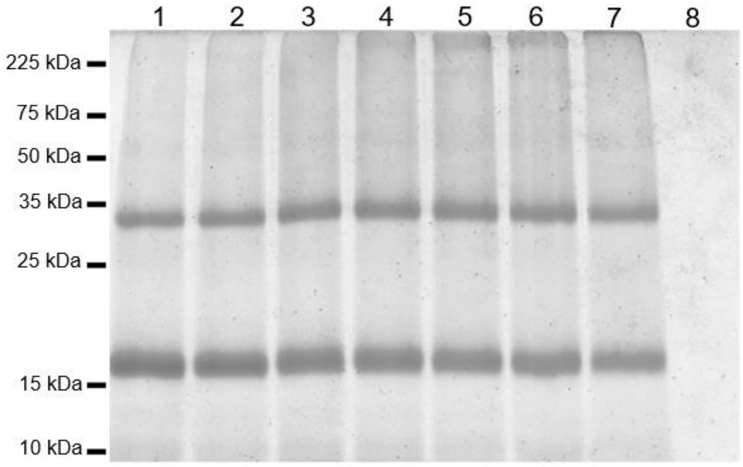


Figure 3. Cross-linking assay. The self-assembly of rEoUSP-1 was assessed with glutaraldehyde at different times: lines 1-7, samples incubated at 0.5, 1, 5, 10, 20, 30, and 60 minutes, respectively. Note the ~32 kDa band corresponding a protein dimer. Line 8, sample without rEoUSP-1. Molecular weight markers are shown to the left.

Conclusions

The Universal Stress Proteins are a relatively “new” gene family that was discovered in the last two decades. Their relevance in the maintenance of the cellular homeostasis, specifically against a large variety of stressor agents, make them an interesting target for study. However, the information available at this moment is almost restricted to bacteria and plant species. The exploration of their function in other eukaryotic species warrant considerable attention, particularly in the host-parasite interactions, where the contact with stressor agents is an intrinsic characteristic of this type of association. In this work, we have cloned and expressed a recombinant USP from *E. ortleppi*, a species that belongs to the *Echinococcus granulosus* sensu lato complex, causative of cyst echinococcosis. We conclude that rEoUSP-1 probably binds to ATP, and is capable to hydrolyze this molecule. However, the ATPase activity occurs at high protein concentration, suggesting that rEoUSP-1 may need the association to other cellular compounds to catalyze this reaction. Moreover, this protein is capable of forming dimers, with tight proximity and stable interactions. Further functional experiments will be needed for a better characterization of this rEoUSP-1.

References

Avila, H. G., G. B. Santos, M. A. Cucher, N. Macchiaroli, M. G. Pérez, G. Baldi, O. Jensen, V. Pérez, R. López, P. Negro, E. Scialfa, A. Zaha, H. B. Ferreira, M. Rosenzvit and L. Kamenetzky (2017). "Implementation of new tools in molecular epidemiology studies of Echinococcus granulosus sensu lato in South America." Parasitol Int **66**(3): 250-257.

Balbinotti, H., G. B. Santos, J. Badaraco, A. C. Arend, D. Graichen, K. L. Haag and A. Zaha (2012). "Echinococcus ortleppi (G5) and Echinococcus granulosus sensu stricto (G1) loads in cattle from Southern Brazil." Vet Parasitol **188**(3-4): 255-260.

Banerjee, A., R. S. Adolph, J. Gopalakrishnapai, S. Kleinboelting, C. Emmerich, C. Steegborn and S. S. Visweswariah (2015). "A universal stress protein (USP) in mycobacteria binds cAMP." J Biol Chem **290**(20): 12731-12743.

Bochkareva, E. S., A. S. Girshovich and E. Bibi (2002). "Identification and characterization of the Escherichia coli stress protein UP12, a putative in vivo substrate of GroEL." Eur J Biochem **269**(12): 3032-3040.

Bowles, J., D. Blair and D. P. McManus (1992). "Genetic variants within the genus Echinococcus identified by mitochondrial DNA sequencing." Mol Biochem Parasitol **54**(2): 165-173.

Chan, K. M., D. Delfert and K. D. Junger (1986). "A direct colorimetric assay for Ca2+ -stimulated ATPase activity." Anal Biochem **157**(2): 375-380.

Coman, B. J. (1975). "The survival of Taenia pisiformis eggs under laboratory conditions and in the field environment." Aust Vet J **51**(12): 560-565.

Darawshe, S., Y. Tsafadyah and E. Daniel (1987). "Quaternary structure of erythrocruorin from the nematode Ascaris suum. Evidence for unsaturated haem-binding sites." Biochem J **242**(3): 689-694.

de la Rue, M. L., K. Takano, J. F. Brochado, C. V. Costa, A. G. Soares, K. Yamano, K. Yagi, Y. Katoh and K. Takahashi (2011). "Infection of humans and animals with Echinococcus granulosus (G1 and G3 strains) and E. ortleppi in Southern Brazil." Vet Parasitol **177**(1-2): 97-103.

de Souza, C. S., A. G. Torres, A. Caravelli, A. Silva, J. M. Polatto and R. M. Piazza (2016). "Characterization of the universal stress protein F from atypical enteropathogenic Escherichia coli and its prevalence in Enterobacteriaceae." Protein Sci **25**(12): 2142-2151.

Drumm, J. E., K. Mi, P. Bilder, M. Sun, J. Lim, H. Bielefeldt-Ohmann, R. Basaraba, M. So, G. Zhu, J. M. Tufariello, A. A. Izzo, I. M. Orme, S. C. Almo, T. S. Leyh and J. Chan (2009). "Mycobacterium tuberculosis universal stress protein Rv2623 regulates bacillary growth by ATP-Binding: requirement for establishing chronic persistent infection." PLoS Pathog **5**(5): e1000460.

Glass, L. N., G. Swapna, S. S. Chavadi, J. M. Tufariello, K. Mi, J. E. Drumm, T. T. Lam, G. Zhu, C. Zhan, C. Vilchéze, J. Arcos, Y. Chen, L. Bi, S. Mehta, S. A. Porcelli, S. C. Almo, S. R. Yeh, W. R. Jacobs, J. B. Torrelles and J. Chan (2017). "Mycobacterium tuberculosis universal stress protein Rv2623 interacts with the putative ATP binding cassette (ABC) transporter Rv1747 to regulate mycobacterial growth." PLoS Pathog **13**(7): e1006515.

Grenouillet, F., G. Umhang, F. Arbez-Gindre, G. Mantion, E. Delabrousse, L. Millon and F. Boué (2014). "Echinococcus ortleppi infections in humans and cattle, France." Emerg Infect Dis **20**(12): 2100-2102.

Griffith, I. P. (1972). "The effect of cross-links on the mobility of proteins in dodecyl sulphate-polyacrylamide gels." Biochem J **126**(3): 553-560.

Gustavsson, N., A. Diez and T. Nyström (2002). "The universal stress protein paralogues of Escherichia coli are co-ordinately regulated and co-operate in the defence against DNA damage." Mol Microbiol **43**(1): 107-117.

Gutiérrez-Beltrán, E., J. M. Personat, F. de la Torre and O. Del Pozo (2017). "A Universal Stress Protein Involved in Oxidative Stress Is a Phosphorylation Target for Protein Kinase CIPK6." Plant Physiol **173**(1): 836-852.

Huang, F., Z. Dang, Y. Suzuki, T. Horiuchi, K. Yagi, H. Kouguchi, T. Irie, K. Kim and Y. Oku (2016). "Analysis on Gene Expression Profile in Oncospheres and Early Stage Metacestodes from Echinococcus multilocularis." PLoS Negl Trop Dis **10**(4): e0004634.

Hui, W., S. Jiang, J. Tang, H. Hou, S. Chen, B. Jia and Q. Ban (2015). "An Immediate Innate Immune Response Occurred In the Early Stage of E. granulosus Eggs Infection in Sheep: Evidence from Microarray Analysis." PLoS One **10**(8): e0135096.

Jung, Y. J., S. M. Melencion, E. S. Lee, J. H. Park, C. V. Alinapon, H. T. Oh, D. J. Yun, Y. H. Chi and S. Y. Lee (2015). "Universal Stress Protein Exhibits a Redox-Dependent Chaperone Function in Arabidopsis and Enhances Plant Tolerance to Heat Shock and Oxidative Stress." Front Plant Sci **6**: 1141.

Kim, D. J., E. Bitto, C. A. Bingman, H. J. Kim, B. W. Han and G. N. Phillips (2015). "Crystal structure of the protein At3g01520, a eukaryotic universal stress protein-like protein from Arabidopsis thaliana in complex with AMP." Proteins **83**(7): 1368-1373.

Liu, W. T., M. H. Karavolos, D. M. Bulmer, A. Allaoui, R. D. Hormaeche, J. J. Lee and C. M. Khan (2007). "Role of the universal stress protein UspA of Salmonella in growth arrest, stress and virulence." Microb Pathog **42**(1): 2-10.

McCreesh, N. and M. Booth (2014). "The effect of increasing water temperatures on Schistosoma mansoni transmission and Biomphalaria pfeifferi population dynamics: an agent-based modelling study." PLoS One **9**(7): e101462.

Monteiro, K. M., S. M. Scapin, M. V. Navarro, N. I. Zanchin, M. B. Cardoso, N. P. da Silveira, P. F. Gonçalves, H. K. Stassen, A. Zaha and H. B. Ferreira (2007). "Self-assembly and structural characterization of Echinococcus granulosus antigen B recombinant subunit oligomers." Biochim Biophys Acta **1774**(2): 278-285.

Nachin, L., L. Brive, K. C. Persson, P. Svensson and T. Nyström (2008). "Heterodimer formation within universal stress protein classes revealed by an in silico and experimental approach." J Mol Biol **380**(2): 340-350.

Nachin, L., U. Nannmark and T. Nyström (2005). "Differential roles of the universal stress proteins of Escherichia coli in oxidative stress resistance, adhesion, and motility." J Bacteriol **187**(18): 6265-6272.

Roth, S., N. Willcox, R. Rzepka, M. P. Mayer and I. Melchers (2002). "Major differences in antigen-processing correlate with a single Arg71<-->Lys substitution in HLA-DR molecules predisposing to rheumatoid arthritis and with their selective interactions with 70-kDa heat shock protein chaperones." J Immunol **169**(6): 3015-3020.

Sagurthi, S. R., R. R. Panigrahi, G. Gowda, H. S. Savithri and M. R. Murthy (2007). "Cloning, expression, purification, crystallization and preliminary X-ray diffraction analysis of universal stress protein F (YnaF) from Salmonella typhimurium." Acta Crystallogr Sect F Struct Biol Cryst Commun **63**(Pt 11): 957-960.

Soufi, B., K. Krug, A. Harst and B. Macek (2015). "Characterization of the E. coli proteome and its modifications during growth and ethanol stress." Front Microbiol **6**: 103.

Sousa, M. C. and D. B. McKay (2001). "Structure of the universal stress protein of Haemophilus influenzae." Structure **9**(12): 1135-1141.

Sánchez Thevenet, P., H. M. Alvarez and J. A. Basualdo (2017). "Survival, physical and physiological changes of Taenia hydatigena eggs under different conditions of water stress." Exp Parasitol **177**: 47-56.

Thompson, R. (1995). Biology and systematics of *Echinococcus*, CAB International, Wallingford.

Traut, T. W. (1994). "The functions and consensus motifs of nine types of peptide segments that form different types of nucleotide-binding sites." Eur J Biochem **222**(1): 9-19.

Udawat, P., A. Mishra and B. Jha (2014). "Heterologous expression of an uncharacterized universal stress protein gene (SbUSP) from the extreme halophyte, Salicornia brachiata, which confers salt and osmotic tolerance to E. coli." Gene **536**(1): 163-170.

Upatham, E. S. (1973). "Location of Biomphalaria glabrata (Say) by miracidia of Schistosoma mansoni Sambon in natural standing and running waters on the West Indian Island of St. Lucia." Int J Parasitol **3**(3): 289-297.

Veit, P., B. Bilger, V. Schad, J. Schäfer, W. Frank and R. Lucius (1995). "Influence of environmental factors on the infectivity of Echinococcus multilocularis eggs." Parasitology **110 (Pt 1)**: 79-86.

Weber, A. and K. Jung (2006). "Biochemical properties of UspG, a universal stress protein of Escherichia coli." Biochemistry **45**(6): 1620-1628.

World health statistics 2017: monitoring health for the SDGs, Sustainable Development Goals. Geneva: World Health Organization; 2017

Wu, X., M. Yano, H. Washida and H. Kido (2004). "The second metal-binding site of 70 kDa heat-shock protein is essential for ADP binding, ATP hydrolysis and ATP synthesis." Biochem J **378**(Pt 3): 793-799.

Zarembinski, T. I., L. W. Hung, H. J. Mueller-Dieckmann, K. K. Kim, H. Yokota, R. Kim and S. H. Kim (1998). "Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics." Proc Natl Acad Sci U S A **95**(26): 15189-15193.

Additional information

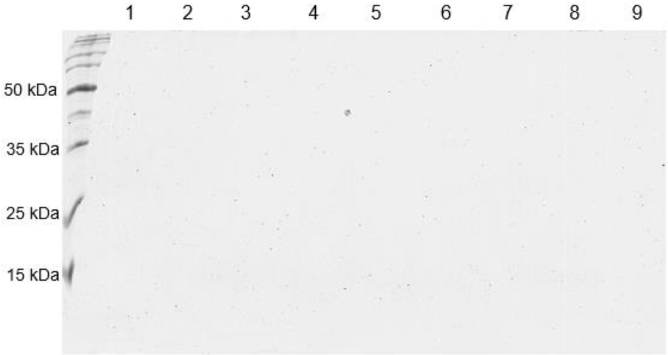


Figure S1. AMP binding assay. The AMP resin (AMP immobilized on an agarose matrix) is unable to interact with rEoUSP-1, GST, or BSA (lines 1, 2, 3, respectively). No protein was eluted using 10 mM of NADP+ (lines 4, 5, 6) or urea 7 M (line 7; 8, 9).

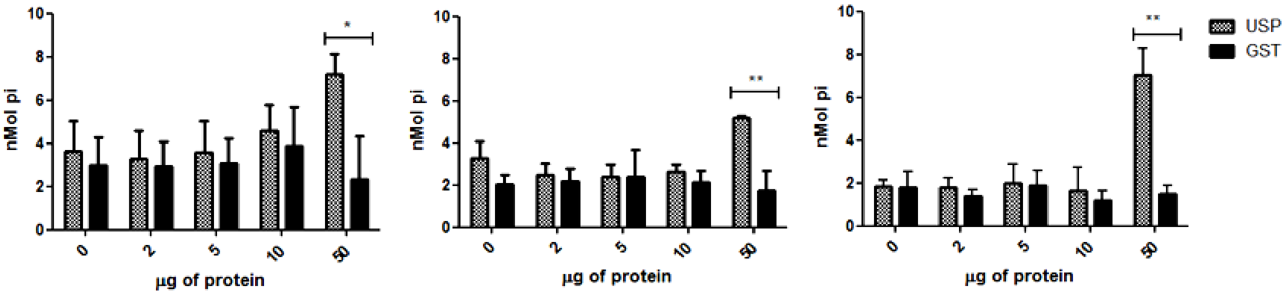


Figure S2. ATPase activity. The rEgrUSP36 ATPase activity shows differences in relation to the control in high protein high protein concentration at 37 °C for 60 minutes (left), or at 45 °C for 30 minutes (middle) or 60 minutes (right), As described previously, our ATPase experiments suggest that hydrolytic activity may need other molecular partners to achieve its function (Zarembinski *et al*., 1998).

Discussão

Há 40 anos atrás, tinha sido publicado a primeira sequência completa de nucleotídeos de DNA do bacteriófago ΦX174 através do método *plus and minus*, atualmente conhecido como método de sequenciamento de Sanger (Sanger *et al.*, 1977). Embora naquela época poderia ser considerada como uma estratégia simples e rápida (Sanger and Coulson, 1975; Sanger *et al.*, 1977), a automatização e a marcação e detecção fluorescentes trouxeram inúmeras vantagens para o sequenciamento mais rápido e preciso do DNA (Smith *et al.*, 1986).

Atualmente, existem 3320 genomas completos e 11105 genomas parciais (*draft*), e outros milhares de projetos em andamento (GOLD, *Genomes OnLine Database*). Entre estes, se encontra o projeto dos 50 genomas de helmintos, cujo objetivo é obter os genomas parciais de espécies de relevância sanitária, veterinária e pecuária (50 *Helminth Genomes Initiative*, www.sanger.ac.uk/science/collaboration/50HGP). Desta forma, diversos genomas de platelmintos parasitos já foram obtidos, alguns deles com uma qualidade comparável às de espécies modelo como *C. elegans* e *D. melanogaster* (Protasio *et al.*, 2012; Tsai *et al*., 2013). A partir da obtenção dos genomas, estudos filogenéticos e de genômica comparativa são fundamentais para inferir as relações evolutivas entre as espécies como assim também as diferentes adaptações (por exemplo, ganho e perda de genes) que foram surgindo ao longo da evolução.

O presente trabalho, teve como foco de estudo uma família gênica nomeada como proteínas de estresse universal, ou USPs, em espécies do filo dos Platelmintos. O uso da genômica comparativa e de metodologias para avaliar o grau de divergência evolutiva, permitiu traçar diferentes caminhos evolutivos dentro das USPs, os quais possuem um grande impacto em relação à função destes genes. Por outro lado, ensaios experimentais permitiram a produção de uma USP recombinante de *Echinococcus ortleppi*, que teve como objetivo conhecer a estrutura e a função destes genes numa espécie eucariota. Ensaios adicionais (ver Perspectivas) estão sendo realizados para uma melhor caracterização funcional desta USP na relação parasito-hospedeiro.

As USPs apresentaram um número variável em Platelmintos, entre 10 e 16 *S. mansoni* e *S. mediterranea*, respectivamente, até 83 em *M. lignano*. A análise filogenômica das USPs também mostrou uma ampla variação, de 1 ou 2 em alguns procariotas e fungos, até 48 em *Arabidopsis* (Foret *et al*., 2010). Trabalhos anteriores em *S. mansoni* identificaram 9 genes e não 10 USPs como nosso estudo, provavelmente devido a estarem baseados numa versão não atualizada do genoma (Berriman *et al*., 2009), diferente à disponível atualmente na base de dados (Protasio *et al*., 2012). O grande número de USPs em *M. lignano* estaria associado a duplicações de grandes segmentos de DNA ou até de duplicações do cromossomo inteiro (Wasik *et al*., 2015). Expansões gênicas na família Argonauta também resultou no grande número de copias em *M. lignano* em comparação a outros Platelmintos, provavelmente devido aos níveis de ploidia que adquiriu esta espécie (Fontenla *et al.*, 2017). Foret *et al.* (2010) identificaram uma distribuição desigual das USPs em metazoários, as quais estariam restritas ao gênero *Ciona* em deuterostomados, e ausentes em ecdisozoos (protostomados). Através de uma busca pelo domínio Pfam PF00582, identificamos 22 USPs em várias espécies de nematódeos: *Caenorhaditis angaria*, *Brugia timori Cylicostephanus goldi*, *Globodera pallida*, *Necator americanos*, *Trichuris trichiura*, e em espécies do gênero *Steinernema* (Tabela 1, Anexos). Os genomas parciais destas espécies foram obtidos depois do trabalho de filogenomica das USPs (Foret *et al.*, 2010), o que sugere que a distribuição desigual inferida previamente, implica somente o grupo dos deuterostomados, pois estaria representada nos protostomados, tanto em lophotrocozoos como em ecdisozoos.

Nossas análises permitiram identificar que várias USPs (50 %) encontram-se agrupadas, provavelmente devido a duplicações em tandem. A formação destes *clusters* é mais evidente em Cestódeos e Trematódeos, em comparação com *S. mediterranea* ou *M. lignano*. Por outro lado, o número de íntrons em Moluscos e Anelídeos (média de 3, dados não mostrados) foi maior do que em Platelmintos (1 íntron ou nenhum). Estes resultados sugerem que as duplicações em tandem foram mais frequentes em parasitos do que em espécies de vida livre, e que provavelmente os mecanismos de retrotransposição foram mais significativos nos Platelmintos do que em Moluscos e Anelídeos. Os genomas de parasitos platelmintos apresentam diversas expansões gênicas, muitas delas resultado da interação permanente com seu hospedeiro (tetraspaninas, antígeno B, etc.) (Berriman *et al.*, 2009; Tsai *et al.*, 2013). Portanto, os eventos de retrotransposição e duplicações em tandem podem estar sendo favorecidos em associações interespecíficas do tipo parasito-hospedeiro, promovendo a multiplicação de genes relevantes para a resposta ao estresse.

Além das diferenças na estrutura gênica, algumas USPs não possuíam os seus ortólogos correspondentes em espécies relacionadas ou não estavam identificadas na anotação automática dos genomas (Howe *et al.*, 2016; Howe *et al.*, 2017). Assim, identificamos prováveis pseudogenes dentro desta família gênica. O processo de pseudogenização se caracteriza pelo acumulo de mutações sem sentido (*nonsense*) que originam códons de terminação prematuros, mutações que modificam a fase aberta de leitura (*frameshift*), e também inserções e deleções (*indels*) aleatórias. Estas alterações, no conjunto, afetam os processos subsequentes de transcrição e tradução, e culminam com a perda de função do produto gênico. A análise da expressão gênica (RNA-seq, qPCR) mostrou a baixa ou nula expressão em quase todos os estágios do desenvolvimento dos parasitos, o que sugere, em princípio, a perda de função.

A presença de pseudogenes e ao mesmo tempo de genes funcionais (inferida pela presença de transcritos) adere com o processo de nascimento e morte (*birth and death*) de genes, onde a partir de duplicações gênicas, algumas cópias se mantem funcionais no genoma, enquanto outras são perdidas pelo surgimento de mutações deletérias, sendo desta forma, pseudogenizados. As imunoglobulinas (Ig), o complexo maior de histocompatibilidade (MHC), e as histonas, são alguns exemplos bem conhecidos de famílias gênicas que seguem este modelo evolutivo. Nessas famílias foram identificados tanto pseudogenes como genes duplicados com alto grau de polimorfismo genético (Nei *et al.*, 1997; Nei e Rooney, 2005; Minias *et al.*, 2016). Embora não foram encontradas evidencias de pseudogenes em Trematódeos, nosso trabalho identificou vários genes com grande similaridade e adjacentes às USPs (provavelmente como resultado de duplicações em tandem), embora sem a presença do domínio USP quando analisados pelo Pfam. Se estes genes são geneticamente próximos às USPs, deveriam agrupar no mesmo *cluster* numa árvore filogenética, e foi isto que foi encontrado. Eles apresentam diversas modificações na sequência de aminoácidos (por exemplo, genes *Csin_107892* e *Csin_107893*, Figura 3, Capitulo I), algumas delas em sítios que são preditos como relevantes para a interação com o ligante, e que provavelmente por este motivo, impede sua identificação pelo Pfam. A ausência de dados de RNA-seq para estes parasitos não permite inferir em relação aos níveis de expressão nos tecidos ou à função destes genes nos diferentes estágios do ciclo de vida. Nos pseudogenes de Cestódeos e alguns genes de Trematódeos, não foi possível identificar o motivo -*CCAATCA*- na região 5' reguladora. Este motivo de DNA ocorre entre as posições -200 a -40 e parece estar muito conservado nas USPs de Platelmintos. A influência das regiões reguladoras na transcrição das USPs frente a condições de estresse foi estudada em *Arabidopsis thaliana* e *Gossypium arboreum* utilizando o gene repórter *GUS* e diferentes combinações do promotor (Loukehaich *et al.*, 2012; Bhuria *et al.*, 2016). Os efeitos foram significativos somente quando removidos mais do que 200 pb *upstream*. Porém, várias repetições correspondentes aos sítios de ligação do *CCAAT box* estão presentes antes da posição -200, e, portanto, teriam um papel importante na transcrição desses genes. Em conjunto, nossos achados sugerem que tanto Cestódeos como Trematódeos possuem *USPs* em vias de pseudogenização.

Além da pseudogenização, existem outros caminhos evolutivos que os genes recentemente duplicados podem seguir. Na sub- e neofuncionalização as mutações também são uma importante fonte de variação, porém, elas tendem a não ser deletérias e levam à aparição de funções complementares ou diferentes às do gene progenitor (Ohno, 1970). Nossas análises de seleção positiva e divergência funcional, identificaram vários resíduos onde a variabilidade e divergência genética estão sendo favorecidas.

Na comparação de *clusters* realizada no Diverge 3.0, alguns grupos de USPs localizadas em tandem apresentaram poucos, e outras, muitos sítios divergentes, indicando que genes contíguos podem seguir caminhos evolutivos diferentes. Os níveis variáveis de expressão gênica nas diferentes etapas do ciclo de vida, e a conservação ou não do motivo proteico [G2xG9xG(S/T)], são exemplos das diferentes características que podem adotar estes genes. Além da nossa análise dos dados de RNA-seq, a expressão de algumas USPs de *S. mansoni* (*Smp_001000*, *Smp_043120*, *Smp_031300, e Smp_001010* (*Smp_200240* neste trabalho, Protasio *et al*., 2012)) já foram previamente analisados por RT-PCR (Isokpehi *et al.*, 2011). É provável que devido às limitações da técnica em relação à sensibilidade na detecção, só foram detectados transcritos para *Smp_001000* (adultos, macho e fêmea) e *Smp_043120* (schistosomula, e adultos fêmea) (Isokpehi *et al*., 2011). Ambas as metodologias confirmam a variabilidade na expressão das USPs nesta espécie, porém, em genes contíguos (*Smp_001000* e *Smp_200240*, ou *Smp_202690* e *Smp_043120*), é possível observar somente a partir dos dados de RNA-seq. Por outro lado, alguns genes em tandem como *EgrG_08736* e *EgrG_08738* apresentam padrões de expressão similares e quase não possuem sítios divergentes.

Devido ao fato de que as USPs são capazes de formar tanto homo como heterodímeros, a expressão simultânea de USPs ao longo do ciclo de vida do parasito poderia favorecer a formação destas estruturas. Assim, os distintos níveis de oligomerização e as diferentes combinações de USPs teriam potencial de aumentar o número de ligantes. A interação de um monômero com seu substrato pode produzir mudanças conformacionais na subunidade, induzindo ao mesmo tempo mudanças conformacionais em outras monômeros de forma cooperativa. Assim, associações diméricas, tetraméricas, etc. poderiam aumentar a afinidade da proteína com o seu ligante (Eisenstein e Schachman, 1989; Marianayagam *et al.*, 2004; Griffin e Gerrard, 2012). Ocultar domínios ou resíduos hidrofóbicos, também justifica a formação de oligômeros (Eisenstein e Schachman, 1989; Ali e Imperiali, 2005).

As USPs são proteínas citoplasmáticas e as suas funções até o momento descritas estariam relacionadas a atividade chaperona, ou também, envolvidas na transdução de sinais (Bochkareva *et al.*, 2002; Banerjee *et al.*, 2015; Jung *et al.*, 2015). A presença de resíduos hidrofóbicos expostos nas, poderia impedir o seu funcionamento das USPs. Por este motivo, é provável que a associação em dímeros tenha como objetivo esconder os aminoácidos hidrofóbicos localizados principalmente entre as folhas beta 4 e 5 (Kim *et al.*, 2015). Entre os resíduos compartilhados em ambas as metodologias utilizadas para a análise de divergência funcional (PAML e Diverge 3.0) encontram-se 7K, 32K, 45R, 72E, 114K, 115I, 117E, e 120G. Embora esses aminoácidos não são preditos como de interação direta com o ligante, eles se encontram bem próximos destes (P11, D13, V41, G127, G130, G140, e S/T141). Devido à presença de *gaps* no alinhamento de códons e à grande variabilidade de aminoácidos na região central da proteína, os sítios com seleção positiva e os sítios com divergência funcional que se encontram entre as posições ~40 e ~70, deveriam ser analisados com cuidado.

De forma geral, nossos achados sugerem que as USPs possuem caminhos evolutivos opostos em platelmintos parasitos. As diferenças entre genes parálogos e, provavelmente, as distintas associações oligoméricas que as proteínas podem adquirir, contribuem para a diversidade funcional que caracteriza esta família genica. Estudos direcionados ao conhecimento dos parceiros moleculares das USPs vão permitir expandir nosso conhecimento sobre as diversas vias em que estas proteínas estão envolvidas. A continuação, segue uma serie de analogias entre as USPs e as proteínas de choque térmico (HSPs). As HSPs são proteínas bastante estudadas e amplamente distribuídas nos organismos, e talvez possam direcionar futuras pesquisas em relação as USPs.

Relações entre USPs e HSPs

Os organismos devem ser capazes de controlar e regular suas funções celulares e se adaptar ao meio externo, mantendo sua estabilidade e evitando a morte. Ao longo da evolução, as espécies adquiriram diversas ferramentas moleculares para lidar com as variações do meio externo, sejam elas do tipo abiótico ou biótico (temperatura, pH, infecções por microrganismos, resistência a antibióticos, etc.). Um exemplo bastante conhecido são as proteínas de choque térmico (HSP, *heat shock proteins*).

As HSPs se apresentam em abundancia dentro da célula e estão envolvidas no processo de dobramento de proteínas, direcionamento destas para degradação, regulação das vias de transdução de sinais, e na resposta imune através da sua atividade chaperona sobre elementos antigênicos (Srivastava, 2002; Mayer, 2013; Kelly *et al.*, 2017). Da mesma forma que as HSPs, as USPs possuem um domínio conservado e estão amplamente distribuídas nos organismos (Foret *et al.*, 2010). Mudanças na temperatura, pH, e a exposição à radiação UV, levam a um aumento nos níveis de expressão gênica, indicando um papel importante na resposta às variações do meio externo para as duas famílias (Park e Seo, 2015).

Estruturalmente, as HSP possuem vários domínios, por exemplo, a HSP90 possui um domínio N-terminal de ligação a nucleotídeos (NBD, *nucleotide binding domain*), um domínio de ligação ao polipeptídio, e um domínio de homo dimerização, enquanto a HSP70 possui um NBD e um domínio de interação com substrato (SBD, *substrate binding domain*) (Javid *et al.*, 2007; Mayer, 2013). USPs são proteínas mais pequenas (140-160 aa, ~18 kDa), mas também possuem um domínio de ligação a nucleotídeos, sendo capazes de hidrolisar ATP. O ciclo da atividade chaperona da HSP é controlado pela ligação a nucleotídeos que possuem a função de interruptor (*switch*), diferenciando dois estados: HSP-ATP, com alta afinidade por polipeptídios; e HSP-ADP (depois da hidrolise do ATP), com uma afinidade reduzida de ligação a polipeptídios. O domínio NBD da HSP70 possui atividade ATPase, e para atingir a hidrólise de forma eficiente, ela precisa se associar a outras proteínas (co-chaperonas), como as proteínas de domínio J (JDPs) e os fatores de intercambio de nucleotídeos (NEFs) (Mayer, 2013). Alguns estudos sugerem que as USPs podem funcionar como interruptores moleculares (através da ligação e/ou hidrolise de nucleotídeos), tendo uma atividade no dobramento de proteínas. Embora os mecanismos moleculares responsáveis pela função das USPs ainda estão na fase inicial, a semelhança destas com outras proteínas já bem estudadas como as HSPs pode servir de exemplo para direcionar futuras pesquisas.

Perspectivas

As perspectivas desse trabalho, atualmente em andamento, estão direcionadas para a finalização dos experimentos, cujos resultados farão parte do Artigo 2. A seguir, é apresentada uma breve descrição de cada um dos experimentos que estão sendo realizados.

Interação da rEoUSP-1 com o análogo fluorescente do ATP

Com a finalidade de corroborar nossos resultados de interação da rEoUSP-1 com ATP (resina com ATP imobilizado) utilizando outra metodologia, a afinidade da proteína com este nucleotídeo será avaliada com o uso do reagente trinitrofenil adenosina 5'-trifosfato (TNT-ATP), um análogo fluorescente do ATP. A titulação será realizada tanto com a proteína purificada (1-10 µM) como com o reagente TNP-ATP (1-100 µM). O reagente TNP-ATP e a proteína rEoUSP-1 serão misturados e incubados por 10 minutos num volume final de 200 µL no buffer TBS 1X. Ensaios de inibição serão realizados mediante a incubação adicional de 10 minutos com 20 mM de ATP. A leitura da fluorescência será realizada no aparelho SpectraMax M3 com longitudes de onda de emissão e excitação de 403 e 535 nanômetros, respectivamente, em placas de 96 poços pretas de fundo claro. O ajuste da curva e obtenção dos gráficos será realizada com o programa GraphPad Prism.

Mutações sítio dirigidas no motivo de ligação ao ATP da rEoUSP-1

Com o intuito de conhecer os aminoácidos da envolvidos na interação com o ligante, quatro *primers* foram desenhados para serem utilizados com o kit *QuikChange XL* Site-Directed Mutagenesis. As mutações sítio dirigidas escolhidas envolvem a modificação de um só nucleotídeo (mutação de ponto), a seguir: a29c (D13A), t113a (V41E), g394c (G140R), g398t (S141I). Até o momento, foi obtido e confirmado por sequenciamento somente o primeiro mutante. Essas proteínas modificadas serão caracterizadas comparativamente à proteína selvagem (*rEoUSP-1*). As proteínas serão expressadas, purificadas, e utilizadas nos ensaios de interação usando a resina com nucleotídeos imobilizados, e o análogo fluorescente de ATP.

Expressão gênica em resposta a agentes de estresse

A resposta das USPs a diversos tipos de estresse é uma característica funcional já descrita para esta família gênica. Utilizando *primers* específicos desenhados para seis USPs de *Echinococcus* spp. (Artigo 1), serão avaliados os níveis de expressão gênica quantificados por PCR em tempo real, em resposta a dois agentes de estresse. O estresse oxidativo será analisado em protoscólices mantidos em cultura *in vitro* com a presença de peróxido de hidrogênio (0-2.5 mM).

Por outro lado, a resposta dos protoescólices ao choque térmico será avaliada mantendo os indivíduos a 42 ^{o}C em meio de cultura.

Cromatografia de afinidade e Cross-linking químico (Sulfo-SBED)

Até o momento, não existem relatos sobre quais são as proteínas que interagem de forma direta com as USPs. A identificação destes parceiros moleculares, vai permitir conhecer as vias metabólicas nas quais as USPs estão envolvidas e revelar as funções específicas destas proteínas no contexto global. Desta maneira, interações do tipo *bona fide* serão obtidas a partir de experimentos de cromatografia de afinidade usando a coluna Sepharose-CH ativada 4B (Sigma-Aldrich). Brevemente, a proteína rEoUSP-1 será imobilizada na resina, e subsequentemente, o extrato proteico obtido a partir de protoscólices de *Echinococcus* spp. será misturado e incubado por 16 hs. Após sucessivas lavagens e remoção de proteínas com interações inespecíficas, as proteínas serão eluídas com 7 M de ureia. Por outro lado, o uso do reagente Sulfo-SBED permitirá identificar interações do tipo transitórias, além daquelas mais estáveis. O Sulfo-SBED é um reagente de *cross-linking* que possui três grupos funcionais: um grupo N-hidroxisuccinimida (NHS) que iria interagir de forma covalente com a proteína recombinante; um grupo aril azida foto ativável que promove a ligação covalente dos parceiros moleculares próximos à proteína de interesse; e biotina, que facilitaria a posterior purificação através do uso, por exemplo, de uma coluna de avidina monomérica. As amostras obtidas por ambas as metodologias serão analisadas na Unidade de Química de Proteínas e Espectrometria de Massas (Uniprote-MS) do Centro de Biotecnología da UFRGS. As duas metodologias descritas acima já foram realizadas pelo nosso grupo (Teichmann *et al.*, 2015), o que facilitará a realização das mesmas.

Referências bibliográficas

Ali, M. H. and B. Imperiali (2005). "Protein oligomerization: how and why." Bioorg Med Chem **13**(17): 5013-5020.

Balbinotti, H., G. B. Santos, J. Badaraco, A. C. Arend, D. Graichen, K. L. Haag and A. Zaha (2012). "Echinococcus ortleppi (G5) and Echinococcus granulosus sensu stricto (G1) loads in cattle from Southern Brazil." Vet Parasitol **188**(3-4): 255-260.

Banerjee, A., R. S. Adolph, J. Gopalakrishnapai, S. Kleinboelting, C. Emmerich, C. Steegborn and S. S. Visweswariah (2015). "A universal stress protein (USP) in mycobacteria binds cAMP." J Biol Chem **290**(20): 12731-12743.

Berriman, M., B. J. Haas, P. T. LoVerde, R. A. Wilson, G. P. Dillon, G. C. Cerqueira, S. T. Mashiyama, B. Al-Lazikani, L. F. Andrade, P. D. Ashton, M. A. Aslett, D. C. Bartholomeu, G. Blandin, C. R. Caffrey, A. Coghlan, R. Coulson, T. A. Day, A. Delcher, R. DeMarco, A. Djikeng, T. Eyre, J. A. Gamble, E. Ghedin, Y. Gu, C. Hertz-Fowler, H. Hirai, Y. Hirai, R. Houston, A. Ivens, D. A. Johnston, D. Lacerda, C. D. Macedo, P. McVeigh, Z. Ning, G. Oliveira, J. P. Overington, J. Parkhill, M. Pertea, R. J. Pierce, A. V. Protasio, M. A. Quail, M. A. Rajandream, J. Rogers, M. Sajid, S. L. Salzberg, M. Stanke, A. R. Tivey, O. White, D. L. Williams, J. Wortman, W. Wu, M. Zamanian, A. Zerlotini, C. M. Fraser-Liggett, B. G. Barrell and N. M. El-Sayed (2009). "The genome of the blood fluke Schistosoma mansoni." Nature **460**(7253): 352-358.

Bhuria, M., P. Goel, S. Kumar and A. K. Singh (2016). "The Promoter of AtUSP Is Co-regulated by Phytohormones and Abiotic Stresses in Arabidopsis thaliana." Front Plant Sci **7**: 1957.

Bochkareva, E. S., A. S. Girshovich and E. Bibi (2002). "Identification and characterization of the Escherichia coli stress protein UP12, a putative in vivo substrate of GroEL." Eur J Biochem **269**(12): 3032-3040.

Boes, N., K. Schreiber, E. Härtig, L. Jaensch and M. Schobert (2006). "The Pseudomonas aeruginosa universal stress protein PA4352 is essential for surviving anaerobic energy stress." J Bacteriol **188**(18): 6529-6538.

Budke, C. M., P. Deplazes and P. R. Torgerson (2006). "Global socioeconomic impact of cystic echinococcosis." Emerg Infect Dis **12**(2): 296-303.

Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J. F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. Geoghagen and J. C. Venter (1996). "Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii." Science **273**(5278): 1058-1073.

Caira JN, Littlewood DTJ. 2013. Worms, Platyhelminthes. In: Levin SA, ed. Encyclopedia of biodiversity, 2nd edn. Waltham, MA: Academic Press, 437–469.

Collins, J. J. (2017). "Platyhelminthes." Curr Biol **27**(7): R252-R256.

Collins, J. J. and P. A. Newmark (2013). "It's no fluke: the planarian as a model for understanding schistosomes." PLoS Pathog **9**(7): e1003396.

Collins, J. J., B. Wang, B. G. Lambrus, M. E. Tharp, H. Iyer and P. A. Newmark (2013). "Adult somatic stem cells in the human parasite Schistosoma mansoni." Nature **494**(7438): 476-479.

Coman, B. J. (1975). "The survival of Taenia pisiformis eggs under laboratory conditions and in the field environment." Aust Vet J **51**(12): 560-565.

Consortium, S. j. G. S. a. F. A. (2009). "The Schistosoma japonicum genome reveals features of host-parasite interplay." Nature **460**(7253): 345-351.

Cwiklinski, K., J. P. Dalton, P. J. Dufresne, J. La Course, D. J. Williams, J. Hodgkinson and S. Paterson (2015). "The Fasciola hepatica genome: gene duplication and polymorphism reveals adaptation to the host environment and the capacity for rapid evolution." Genome Biol **16**: 71.

Drumm, J. E., K. Mi, P. Bilder, M. Sun, J. Lim, H. Bielefeldt-Ohmann, R. Basaraba, M. So, G. Zhu, J. M. Tufariello, A. A. Izzo, I. M. Orme, S. C. Almo, T. S. Leyh and J. Chan (2009). "Mycobacterium tuberculosis universal stress protein Rv2623 regulates bacillary growth by

ATP-Binding: requirement for establishing chronic persistent infection." PLoS Pathog **5**(5): e1000460.

Díaz, A., C. Casaravilla, J. E. Allen, R. B. Sim and A. M. Ferreira (2011). "Understanding the laminated layer of larval Echinococcus II: immunology." Trends Parasitol **27**(6): 264-273.

Díaz, A., C. Casaravilla, F. Irigoín, G. Lin, J. O. Previato and F. Ferreira (2011). "Understanding the laminated layer of larval Echinococcus I: structure." Trends Parasitol **27**(5): 204-213.

Díaz, Á., C. Fernández, Á. Pittini, P. I. Seoane, J. E. Allen and C. Casaravilla (2015). "The laminated layer: Recent advances and insights into Echinococcus biology and evolution." Exp Parasitol **158**: 23-30.

Eisenstein E, Schachman HK (1989). "Determining the roles of subunits in protein function". In: Creighton TE, ed. Protein Function. Oxford: IRL Press

Fitch, W. M. (1970). "Distinguishing homologous from analogous proteins." Syst Zool **19**(2): 99-113.

Fontenla, S., G. Rinaldi, P. Smircich and J. F. Tort (2017). "Conservation and diversification of small RNA pathways within flatworms." BMC Evol Biol **17**(1): 215.

Forêt, S., F. Seneca, D. de Jong, A. Bieller, G. Hemmrich, R. Augustin, D. C. Hayward, E. E. Ball, T. C. Bosch, K. Agata, M. Hassel and D. J. Miller (2011). "Phylogenomics reveals an anomalous distribution of USP genes in metazoans." Mol Biol Evol **28**(1): 153-161.

Glass, L. N., G. Swapna, S. S. Chavadi, J. M. Tufariello, K. Mi, J. E. Drumm, T. T. Lam, G. Zhu, C. Zhan, C. Vilchéze, J. Arcos, Y. Chen, L. Bi, S. Mehta, S. A. Porcelli, S. C. Almo, S. R. Yeh, W. R. Jacobs, J. B. Torrelles and J. Chan (2017). "Mycobacterium tuberculosis universal stress protein Rv2623 interacts with the putative ATP binding cassette (ABC) transporter Rv1747 to regulate mycobacterial growth." PLoS Pathog **13**(7): e1006515.

Gonzali, S., E. Loreti, F. Cardarelli, G. Novi, S. Parlanti, C. Pucciariello, L. Bassolino, V. Banti, F. Licausi and P. Perata (2015). "Universal stress protein HRU1 mediates ROS homeostasis under anoxia." Nat Plants **1**: 15151.

Griffin, M. D. and J. A. Gerrard (2012). "The relationship between oligomeric state and protein function." Adv Exp Med Biol **747**: 74-90.

Gustavsson, N., A. Diez and T. Nyström (2002). "The universal stress protein paralogues of Escherichia coli are co-ordinately regulated and co-operate in the defence against DNA damage." Mol Microbiol **43**(1): 107-117.

Gutiérrez-Beltrán, E., J. M. Personat, F. de la Torre and O. Del Pozo (2017). "A Universal Stress Protein Involved in Oxidative Stress Is a Phosphorylation Target for Protein Kinase CIPK6." Plant Physiol **173**(1): 836-852.

Hewitson, J. P. and R. M. Maizels (2014). "Vaccination against helminth parasite infections." Expert Rev Vaccines **13**(4): 473-487.

Hingley-Wilson, S. M., K. E. Lougheed, K. Ferguson, S. Leiva and H. D. Williams (2010). "Individual Mycobacterium tuberculosis universal stress protein homologues are dispensable in vitro." Tuberculosis (Edinb) **90**(4): 236-244.

Howe, K. L., B. J. Bolt, S. Cain, J. Chan, W. J. Chen, P. Davis, J. Done, T. Down, S. Gao, C. Grove, T. W. Harris, R. Kishore, R. Lee, J. Lomax, Y. Li, H. M. Muller, C. Nakamura, P. Nuin, M. Paulini, D. Raciti, G. Schindelman, E. Stanley, M. A. Tuli, K. Van Auken, D. Wang, X. Wang, G. Williams, A. Wright, K. Yook, M. Berriman, P. Kersey, T. Schedl, L. Stein and P. W. Sternberg (2016). "WormBase 2016: expanding to enable helminth genomic research." Nucleic Acids Res **44**(D1): D774-780.

Howe, K. L., B. J. Bolt, M. Shafie, P. Kersey and M. Berriman (2017). "WormBase ParaSite - a comprehensive resource for helminth genomics." Mol Biochem Parasitol **215**: 2-10.

Huang, F., Z. Dang, Y. Suzuki, T. Horiuchi, K. Yagi, H. Kouguchi, T. Irie, K. Kim and Y. Oku (2016). "Analysis on Gene Expression Profile in Oncospheres and Early Stage Metacestodes from Echinococcus multilocularis." PLoS Negl Trop Dis **10**(4): e0004634.

Hurles, M. (2004). "Gene duplication: the genomic trade in spare parts." PLoS Biol **2**(7): E206.

Isokpehi, R. D., O. Mahmud, A. N. Mbah, S. S. Simmons, L. Avelar, R. V. Rajnarayanan, U. K. Udensi, W. K. Ayensu, H. H. Cohly, S. D. Brown, C. R. Dates, S. D. Hentz, S. J. Hughes, D. R. Smith-McInnis, C. O. Patterson, J. N. Sims, K. T. Turner, B. S. Williams, M. O. Johnson, T. Adubi, J. V. Mbuh, C. I. Anumudu, G. O. Adeoye, B. N. Thomas, O. Nashiru and G. Oliveira (2011). "Developmental Regulation of Genes Encoding Universal Stress Proteins in Schistosoma mansoni." Gene Regul Syst Bio **5**: 61-74.

Javid, B., P. A. MacAry and P. J. Lehner (2007). "Structure and function: heat shock proteins and adaptive immunity." J Immunol **179**(4): 2035-2040.

Jayashi, C. M., C. T. Kyngdon, C. G. Gauci, A. E. Gonzalez and M. W. Lightowlers (2012). "Successful immunization of naturally reared pigs against porcine cysticercosis with a recombinant oncosphere antigen vaccine." Vet Parasitol **188**(3-4): 261-267.

Jung, Y. J., S. M. Melencion, E. S. Lee, J. H. Park, C. V. Alinapon, H. T. Oh, D. J. Yun, Y. H. Chi and S. Y. Lee (2015). "Universal Stress Protein Exhibits a Redox-Dependent Chaperone Function in Arabidopsis and Enhances Plant Tolerance to Heat Shock and Oxidative Stress." Front Plant Sci **6**: 1141.

Kelly, M., D. McNeel, P. Fisch and M. Malkovsky (2017). "Immunological considerations underlying heat shock protein-mediated cancer vaccine strategies." Immunol Lett **193**: 1-10.

Kim, D. J., E. Bitto, C. A. Bingman, H. J. Kim, B. W. Han and G. N. Phillips (2015). "Crystal structure of the protein At3g01520, a eukaryotic universal stress protein-like protein from Arabidopsis thaliana in complex with AMP." Proteins **83**(7): 1368-1373.

Koziol, U., T. Rauschendorfer, L. Zanon Rodríguez, G. Krohne and K. Brehm (2014). "The unique stem cell system of the immortal larva of the human parasite Echinococcus multilocularis." Evodevo **5**(1): 10.

Larrieu, E., G. Mujica, C. G. Gauci, K. Vizcaychipi, M. Seleiman, E. Herrero, J. L. Labanchi, D. Araya, L. Sepúlveda, C. Grizmado, A. Calabro, G. Talmon, T. V. Poggio, P. Crowley, G. Cespedes, G. Santillán, M. García Cachau, R. Lamberti, L. Gino, M. Donadeu and M. W. Lightowlers (2015). "Pilot Field Trial of the EG95 Vaccine Against Ovine Cystic Echinococcosis in Rio Negro, Argentina: Second Study of Impact." PLoS Negl Trop Dis **9**(10): e0004134.

Loukehaich, R., T. Wang, B. Ouyang, K. Ziaf, H. Li, J. Zhang, Y. Lu and Z. Ye (2012). "SpUSP, an annexin-interacting universal stress protein, enhances drought tolerance in tomato." J Exp Bot **63**(15): 5593-5606.

Maldonado, L. L., J. Assis, F. M. Araújo, A. C. Salim, N. Macchiaroli, M. Cucher, F. Camicia, A. Fox, M. Rosenzvit, G. Oliveira and L. Kamenetzky (2017). "The Echinococcus canadensis (G7) genome: a key knowledge of parasitic platyhelminth human diseases." BMC Genomics **18**(1): 204.

Marianayagam, N. J., M. Sunde and J. M. Matthews (2004). "The power of two: protein dimerization in biology." Trends Biochem Sci **29**(11): 618-625.

Mayer, M. P. (2013). "Hsp70 chaperone dynamics and molecular mechanism." Trends Biochem Sci **38**(10): 507-514.

McCreesh, N. and M. Booth (2014). "The effect of increasing water temperatures on Schistosoma mansoni transmission and Biomphalaria pfeifferi population dynamics: an agent-based modelling study." PLoS One **9**(7): e101462.

McManus, D. P. and J. D. Smyth (1986). "Hydatidosis: changing concepts in epidemiology and speciation." Parasitol Today **2**(6): 163-168.

McNulty, S. N., J. F. Tort, G. Rinaldi, K. Fischer, B. A. Rosa, P. Smircich, S. Fontenla, Y. J. Choi, R. Tyagi, K. Hallsworth-Pepin, V. H. Mann, L. Kammili, P. S. Latham, N. Dell'Oca, F. Dominguez, C. Carmona, P. U. Fischer, P. J. Brindley and M. Mitreva (2017). "Genomes of Fasciola hepatica from the Americas Reveal Colonization with Neorickettsia Endobacteria Related to the Agents of Potomac Horse and Human Sennetsu Fevers." PLoS Genet **13**(1): e1006537.

Minias, P., Z. W. Bateson, L. A. Whittingham, J. A. Johnson, S. Oyler-McCance and P. O. Dunn (2016). "Contrasting evolutionary histories of MHC class I and class II loci in grouse--effects of selection and gene conversion." Heredity (Edinb) **116**(5): 466-476.

Nachin, L., L. Brive, K. C. Persson, P. Svensson and T. Nyström (2008). "Heterodimer formation within universal stress protein classes revealed by an in silico and experimental approach." J Mol Biol **380**(2): 340-350.

Nachin, L., U. Nannmark and T. Nyström (2005). "Differential roles of the universal stress proteins of Escherichia coli in oxidative stress resistance, adhesion, and motility." J Bacteriol **187**(18): 6265-6272.

Nakao, M., A. Lavikainen, T. Yanagida and A. Ito (2013). "Phylogenetic systematics of the genus Echinococcus (Cestoda: Taeniidae)." Int J Parasitol **43**(12-13): 1017-1029.

Nei, M., X. Gu and T. Sitnikova (1997). "Evolution by the birth-and-death process in multigene families of the vertebrate immune system." Proc Natl Acad Sci U S A **94**(15): 7799-7806.

Nei, M. and A. P. Rooney (2005). "Concerted and birth-and-death evolution of multigene families." Annu Rev Genet **39**: 121-152.

Newmark, P. A. and A. Sánchez Alvarado (2002). "Not your father's planarian: a classic model enters the era of functional genomics." Nat Rev Genet **3**(3): 210-219.

Nimeth, K. T., B. Egger, R. Rieger, W. Salvenmoser, R. Peter and R. Gschwentner (2007). "Regeneration in Macrostomum lignano (Platyhelminthes): cellular dynamics in the neoblast stem cell system." Cell Tissue Res **327**(3): 637-646.

Nyström, T. and F. C. Neidhardt (1992). "Cloning, mapping and nucleotide sequencing of a gene encoding a universal stress protein in Escherichia coli." Mol Microbiol **6**(21): 3187-3198.

O Connor, A. and S. McClean (2017). "The role of Universal Stress Proteins in Bacterial Infections." Curr Med Chem.

Obal, G., A. L. Ramos, V. Silva, A. Lima, C. Batthyany, M. I. Bessio, F. Ferreira, G. Salinas and A. M. Ferreira (2012). "Characterisation of the native lipid moiety of Echinococcus granulosus antigen B." PLoS Negl Trop Dis **6**(5): e1642.

Ohno, S., Evolution by Gene Duplication. 1970: Springer.

Park, C. J. and Y. S. Seo (2015). "Heat Shock Proteins: A Review of the Molecular Chaperones for Plant Immunity." Plant Pathol J **31**(4): 323-333.

Protasio, A. V., I. J. Tsai, A. Babbage, S. Nichol, M. Hunt, M. A. Aslett, N. De Silva, G. S. Velarde, T. J. Anderson, R. C. Clark, C. Davidson, G. P. Dillon, N. E. Holroyd, P. T. LoVerde, C. Lloyd, J. McQuillan, G. Oliveira, T. D. Otto, S. J. Parker-Manuel, M. A. Quail, R. A. Wilson, A. Zerlotini, D. W. Dunne and M. Berriman (2012). "A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni." PLoS Negl Trop Dis **6**(1): e1455.

Riganò, R., B. Buttari, E. Profumo, E. Ortona, F. Delunardo, P. Margutti, V. Mattei, A. Teggi, M. Sorice and A. Siracusano (2007). "Echinococcus granulosus antigen B impairs human dendritic cell differentiation and polarizes immature dendritic cell maturation towards a Th2 cell response." Infect Immun **75**(4): 1667-1678.

Rink, J. C. (2013). "Stem cell systems and regeneration in planaria." Dev Genes Evol **223**(1-2): 67-84.

Robb, S. M., K. Gotting, E. Ross and A. Sánchez Alvarado (2015). "SmedGD 2.0: The Schmidtea mediterranea genome database." Genesis **53**(8): 535-546.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe and M. Smith (1977). "Nucleotide sequence of bacteriophage phi X174 DNA." Nature **265**(5596): 687-695.

Sanger, F. and A. R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase." J Mol Biol **94**(3): 441-448.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent and L. E. Hood (1986). "Fluorescence detection in automated DNA sequence analysis." Nature **321**(6071): 674-679.

Sousa, M. C. and D. B. McKay (2001). "Structure of the universal stress protein of Haemophilus influenzae." Structure **9**(12): 1135-1141.

Srivastava, P. (2002). "Interaction of heat shock proteins with peptides and antigen presenting cells: chaperoning of the innate and adaptive immune responses." Annu Rev Immunol **20**: 395-425.

Sánchez Thevenet, P., H. M. Alvarez and J. A. Basualdo (2017). "Survival, physical and physiological changes of Taenia hydatigena eggs under different conditions of water stress." Exp Parasitol **177**: 47-56.

Teichmann, A., D. M. Vargas, K. M. Monteiro, B. V. Meneghetti, C. S. Dutra, R. Paredes, N. Galanti, A. Zaha and H. B. Ferreira (2015). "Characterization of 14-3-3 isoforms expressed in the Echinococcus granulosus pathogenic larval stage." J Proteome Res **14**(4): 1700-1715.

Thomas, L. and Babero, B. B. (1956). Observations on the infectivity of Echinococcus eggs obtained from foxes (Alopex lagopus Linn.) on St. Lawrence Island, Alaska. Journal of Parasitology 42, 659.Thompson, R. (1995). Biology and systematics of *Echinococcus*, CAB International, Wallingford.

Thompson, R. (1995). Biology and systematics of *Echinococcus*, CAB International, Wallingford.

Torgerson, P. R., K. Keller, M. Magnotta and N. Ragland (2010). "The global burden of alveolar echinococcosis." PLoS Negl Trop Dis **4**(6): e722.

Tsai, I. J., M. Zarowiecki, N. Holroyd, A. Garciarrubio, A. Sanchez-Flores, K. L. Brooks, A. Tracey, R. J. Bobes, G. Fragoso, E. Sciutto, M. Aslett, H. Beasley, H. M. Bennett, J. Cai, F. Camicia, R. Clark, M. Cucher, N. De Silva, T. A. Day, P. Deplazes, K. Estrada, C. Fernández, P. W. Holland, J. Hou, S. Hu, T. Huckvale, S. S. Hung, L. Kamenetzky, J. A. Keane, F. Kiss, U. Koziol, O. Lambert, K. Liu, X. Luo, Y. Luo, N. Macchiaroli, S. Nichol, J. Paps, J. Parkinson, N. Pouchkina-Stantcheva, N. Riddiford, M. Rosenzvit, G. Salinas, J. D. Wasmuth, M. Zamanian, Y. Zheng, X. Cai, X. Soberón, P. D. Olson, J. P. Laclette, K. Brehm, M. Berriman and T. s. G. Consortium (2013). "The genomes of four tapeworm species reveal adaptations to parasitism." Nature **496**(7443): 57-63.

Upatham, E. S. (1973). "Location of Biomphalaria glabrata (Say) by miracidia of Schistosoma mansoni Sambon in natural standing and running waters on the West Indian Island of St. Lucia." Int J Parasitol **3**(3): 289-297.

Van Bogelen RA, Hutton ME, Neidhardt FC (1990). "Gene-protein database of Escherichia coli K-12" edition 3. Electrophoresis, 11:1131-1166.

Veit, P., B. Bilger, V. Schad, J. Schäfer, W. Frank and R. Lucius (1995). "Influence of environmental factors on the infectivity of Echinococcus multilocularis eggs." Parasitology **110 (Pt 1)**: 79-86.

Wang, B., J. J. Collins and P. A. Newmark (2013). "Functional genomic characterization of neoblast-like stem cells in larval Schistosoma mansoni." Elife **2**: e00768.

Wang, S., Y. Luo, L. Xiao, X. Luo, S. Gao, Y. Dou, H. Zhang, A. Guo, Q. Meng, J. Hou, B. Zhang, S. Zhang, M. Yang, X. Meng, H. Mei, H. Li, Z. He, X. Zhu, X. Tan, X. Q. Zhu, J. Yu, J. Cai, G. Zhu, S. Hu and X. Cai (2016). "Comparative genomics reveals adaptive evolution of Asian tapeworm in switching to a new intermediate host." Nat Commun **7**: 12845.

Wang, X., W. Chen, Y. Huang, J. Sun, J. Men, H. Liu, F. Luo, L. Guo, X. Lv, C. Deng, C. Zhou, Y. Fan, X. Li, L. Huang, Y. Hu, C. Liang, X. Hu, J. Xu and X. Yu (2011). "The draft genome of the carcinogenic human liver fluke Clonorchis sinensis." Genome Biol **12**(10): R107.

Wasik, K., J. Gurtowski, X. Zhou, O. M. Ramos, M. J. Delás, G. Battistoni, O. El Demerdash, I. Falciatori, D. B. Vizoso, A. D. Smith, P. Ladurner, L. Schärer, W. R. McCombie, G. J. Hannon and M. Schatz (2015). "Genome and transcriptome of the regeneration-competent flatworm, Macrostomum lignano." Proc Natl Acad Sci U S A **112**(40): 12462-12467.

Weber, A. and K. Jung (2006). "Biochemical properties of UspG, a universal stress protein of Escherichia coli." Biochemistry **45**(6): 1620-1628.

Wendt, G. R. and J. J. Collins (2016). "Schistosomiasis as a disease of stem cells." Curr Opin Genet Dev **40**: 95-102.

World health statistics 2017: monitoring health for the SDGs, Sustainable Development Goals. Geneva: World Health Organization; 2017.

Young, N. D., A. R. Jex, B. Li, S. Liu, L. Yang, Z. Xiong, Y. Li, C. Cantacessi, R. S. Hall, X. Xu, F. Chen, X. Wu, A. Zerlotini, G. Oliveira, A. Hofmann, G. Zhang, X. Fang, Y. Kang, B. E. Campbell, A. Loukas, S. Ranganathan, D. Rollinson, G. Rinaldi, P. J. Brindley, H. Yang, J. Wang and R. B. Gasser (2012). "Whole-genome sequence of Schistosoma haematobium." Nat Genet **44**(2): 221-225.

Young, N. D., N. Nagarajan, S. J. Lin, P. K. Korhonen, A. R. Jex, R. S. Hall, H. Safavi-Hemami, W. Kaewkong, D. Bertrand, S. Gao, Q. Seet, S. Wongkham, B. T. Teh, C. Wongkham, P. M. Intapan, W. Maleewong, X. Yang, M. Hu, Z. Wang, A. Hofmann, P. W. Sternberg, P. Tan, J. Wang and R. B. Gasser (2014). "The Opisthorchis viverrini genome provides insights into life in the bile duct." Nat Commun **5**: 4378.

Zarembinski, T. I., L. W. Hung, H. J. Mueller-Dieckmann, K. K. Kim, H. Yokota, R. Kim and S. H. Kim (1998). "Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics." Proc Natl Acad Sci U S A **95**(26): 15189-15193.

Zhang, W., H. Wen, J. Li, R. Lin and D. P. McManus (2012). "Immunology and immunodiagnosis of cystic echinococcosis: an update." Clin Dev Immunol **2012**: 101895.

Zheng, H., W. Zhang, L. Zhang, Z. Zhang, J. Li, G. Lu, Y. Zhu, Y. Wang, Y. Huang, J. Liu, H. Kang, J. Chen, L. Wang, A. Chen, S. Yu, Z. Gao, L. Jin, W. Gu, Z. Wang, L. Zhao, B. Shi, H. Wen, R. Lin, M. K. Jones, B. Brejova, T. Vinar, G. Zhao, D. P. McManus, Z. Chen, Y. Zhou and S. Wang (2013). "The genome of the hydatid tapeworm Echinococcus granulosus." Nat Genet **45**(10): 1168-1175.

Anexos

Tabela 1. USPs de Nematódeos obtidos da WormBase ParaSite através do código PF00582.

Espécie (Nome do projeto)	Clado	Nome do Gene (ID)
Trichuris trichiura (PRJEB535)	I	TTRE_0000930401
Brugia timori (PRJEB4663)	III	BTMF_0000213201
Brugia timori (PRJEB4663)	III	BTMF_0000565801
Steinernema glaseri (PRJNA204943)	IV	L893_g16284
Steinernema glaseri (PRJNA204943)	IV	L893_g3599
Steinernema glaseri (PRJNA204943)	IV	L893_g3853
Steinernema monticolum (PRJNA205067)	IV	L898_g15708
Steinernema monticolum (PRJNA205067)	IV	L898_g29194
Steinernema monticolum (PRJNA205067)	IV	L898_g5960
Steinernema monticolum (PRJNA205067)	IV	L898_g9545
Globodera pallida (PRJEB123)	IV	GPLIN_000966100
Steinernema scapterisci (PRJNA204942)	IV	L892_g140
Steinernema scapterisci (PRJNA204942)	IV	L892_g29459
Steinernema scapterisci (PRJNA204942)	IV	L892_g8385
Necator americanus (PRJNA72135)	V	NECAME_18230
Caenorhabditis angaria (PRJNA51225)	V	Cang_2012_03_13_01596.g16570
Caenorhabditis angaria (PRJNA51225)	V	Cang_2012_03_13_01596.g16570
Caenorhabditis angaria (PRJNA51225)	V	Cang_2012_03_13_03068.g17765
Caenorhabditis angaria (PRJNA51225)	V	Cang_2012_03_13_05125.g19371
Caenorhabditis angaria (PRJNA51225)	V	Cang_2012_03_13_12618.g23848
Cylicostephanus goldi (PRJEB498)	V	CGOC_0001002201
Cylicostephanus goldi (PRJEB498)	V	CGOC_0001002201