Universidade Federal do Rio Grande do Sul
Centro de Biotecnologia
Programa de Pós-Graduação em Biologia Celular e Molecular

Estudo comparativo *in silico* dos produtos de excreção ou secreção de *Echinococcus granulosus* e *Echinococcus multilocularis*

Dissertação de Mestrado

Tiago Minuzzi Freire da Fontoura Gomes

Porto Alegre, Abril de 2018

Universidade Federal do Rio Grande do Sul
Centro de Biotecnologia
Programa de Pós-Graduação em Biologia Celular e Molecular

Estudo comparativo *in silico* dos produtos de excreção ou secreção de *Echinococcus granulosus* e *Echinococcus multilocularis*

Dissertação submetida ao programa de Pós-Graduação em Biologia Celular e Molecular do Centro de Biotecnologia da UFRGS como requisito parcial para a obtenção do grau de Mestre.

Tiago Minuzzi Freire da Fontoura Gomes

Prof. Dr. Henrique Bunselmeyer Ferreira - Orientador

Porto Alegre, Abril de 2018

Este trabalho foi desenvolvido no Laboratório de Genômica Estrutural e Funcional do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul (Cbiot/UFRGS) e no Laboratório Nacional de Computação Científica (LNCC) em Petrópolis/RJ, e contou com apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

“Nunca é demais dizer: todas as formas de vida têm algo em comum. Essa é, e suspeito sempre será, a afirmação mais profundamente verdadeira que existe.”

Bill Bryson

Sumário

Lista de abreviaturas, símbolos e unidades

AAR: região de abundância antigênica (de *Antigenic Abundant Region*).

DC: célula dendrítica (de *dendritic cell*).

ES: excreção ou secreção.

EST: marcadores de sequência expressa (de *expressed sequence tags*).

GO: ontologia gênica (de *gene ontology*).

GPI: glicofosfatidilinositol (de *glycophosphatidylinositol*).

ML: aprendizagem de máquina (de *machine learning*).

NO: óxido nítrico (de *nitric oxide*).

PSNS: pares secretados/não-secretados.

RBH: melhores resultados recíprocos (de *reciprocal best hits*).

RNA-seq: sequenciamento de RNA (de *RNA sequencing*).

TM: transmembrana.

Lista de Figuras

Lista de Tabelas

Resumo

As fases larvais (metacestódeos) de *Echinococcus granulosus* e *Echinococcus multilocularis* causam diferentes formas de equinococose em diferentes espécies de hospedeiros intermediários, incluindo o homem. Os metacestódeos são capazes de sobreviver por anos no hospedeiro humano muito devido a proteínas secretadas, as quais possuem atividades imunomoduladoras e proteolíticas, por exemplo. O presente trabalho apresenta predições *in silico* dos conjuntos de proteínas secretáveis (secretoma) de *E. granulosus* e *E. multilocularis* baseadas em dados genômicos, realiza comparações entre estes secretomas preditos, identifica possíveis problemas de predição e apresenta alternativas para obter com predições *in silico* dados representativos dos secretomas reais destas espécies. A predição inicial dos secretomas de *E. granulosus* e *E. multilocularis* (662, 669 proteínas, respectivamente) apresentou valores semelhantes (~6,4% do proteoma predito), o esperado entre duas espécies próximas. Porém, a análise comparativa entre os pares de ortólogos secretados/não-secretados (PSNS) indicou possíveis problemas de predição de secreção por via não-clássica e de anotação de sequências genômicas. O *software* WoLF PSORT foi utilizado para implementar a predição de secreção por via não-clássica, diminuindo o número de inconsistências de 214 para 114 PSNS. Refinou-se a anotação dos genomas de *E. granulosus* e *E. multilocularis* usando-se estratégia conjunta entre dados de bibliotecas de RNA-seq, ESTs e com os dados das predições *ab initio* do sequenciamento original. O secretoma de *E. granulosus* predito com os dados da reanotação reduziu de 662 proteínas para 658, destas, 43% mantiveram as sequências originais. Em *E. multilocularis*, reduziu de 669 proteínas para 581, 48% destas mantiveram as sequências originais. A qualidade das sequências melhorou com o refinamento da anotação, porém as inconsistências na predição de secreção por via não-clássica mantiveram proporções semelhantes às das sequências sem o refinamento, demonstrando a importância de algoritmos treinados adequadamente para os dados analisados ou de alternativas de *workflows* de análise com critérios bem delineados visando à obtenção de dados mais representativos do conjunto real de proteínas secretadas.

Abstract

The larval stages (metacestode) of *Echinococcus granulosus* and *Echinococcus multilocularis* cause different forms of echinococcosis in different species of intermediate hosts, including humans. Metacestodes are able to survive for years in the human host, much due to its secreted proteins, which have immunomodulatory and proteolytic activities, for example. The present work presents in silico predictions of the secretable protein sets (secretome) of *E. granulosus* and *E. multilocularis* based on genomic data, makes comparisons between these predicted secretomes, identifies possible mispredictios and presents alternatives to obtain *in silico* predictions more representative of the real secretomes of these species. The initial prediction of *E. granulosus* and *E. multilocularis* (662, 669 proteins, respectively) presented similar values (6.4% of predicted proteome), expected between two species. However, the comparative analysis between the secreted/non-secreted orthopairs (PSNS) indicated possible problems of non-classical secretion prediction and annotation of genomic sequences. The WoLF PSORT software was used to implement the prediction of non-classical secretion, reducing the number of inconsistencies from 214 to 114 PSNS. The genomes of *E. granulosus* and *E. multilocularis* were annotated using a combined strategy between data from RNA-seq libraries, ESTs and the *ab initio* prediction data from the original genomic sequencing. The predicted *E. granulosus* secretome with the reannotation reduced from 662 proteins to 658, 43% kept the original sequences. In *E. multilocularis*, reduced from 669 proteins to 581, 48% of these kept the original sequences. The quality of the sequences improved with the refinement of the annotation, but the inconsistencies in the prediction of non-classical secretion maintained proportions similar to the sequences without the refinement, demonstrating the importance of having adequately trained algorithms for the analyzed data or alternatives of workflows with well delineated criteria aiming to obtain more representative data of the real set of secreted proteins.

1. INTRODUÇÃO

1.1. Platelmintos da classe Cestoda

Os membros do filo Platyhelminthes (Gr. *platys*, achatado, + *helmins*, verme), comumente denominados vermes achatados ou platelmintos, variam de tamanho entre 1 mm, ou menos, e vários metros (como algumas tênias), mas a maioria tem de 1 a 3 cm em suas formas adultas (HICKMAN JR. *et al*., 2013). Seus corpos podem ser finos e com forma foliácea ou alongados e com forma de fita. O filo contém formas de vida livre, a exemplo da planaria comum, e espécies parasitas, como os trematódeos e as tênias. Estão divididos em quatro classes (Turbellaria, Trematoda, Monogenea e Cestoda), sendo todos os membros das classes Monogenea, Trematoda e Cestoda, parasitos (HICKMAN JR. *et al*., 2013).

Os platelmintos, juntamente com os nematelmintos, estão entre os agentes infeciosos mais comuns de seres humanos nos países em desenvolvimento (BRINDLEY et al., 2009). A classe Cestoda de platelmintos compreende mais de 5000 espécies descritas, incluindo os agentes etiológicos das equinococoses (hidatidoses) e da cisticercose – *Echinococcus* spp. e *Taenia* spp., respectivamente (WAESCHENBACH et al., 2012). Os cestódeos, quando adultos, são parasitos entéricos de todas as classes de vertebrados e, em diversos casos, utilizam um artrópode como primeiro hospedeiro intermediário (OLSON; TKACH, 2005). Estes parasitos não possuem órgãos respiratório, circulatório ou digestivo e são monoicos, produzindo ovos diploides que originam as oncosferas (ZHENG, 2013). Pertencentes à ordem Cyclophyliidae, o ciclo de vida típico dos parasitos dos gêneros *Echinococcus* e *Taenia* é indireto, com um ou mais hospedeiros intermediários. Com poucas exceções, o verme adulto encontra-se na parte final do intestino delgado do hospedeiro definitivo, com os seus segmentos e ovos chegando ao exterior com as fezes do hospedeiro (TAYLOR *et al*., 2009). A estrobilização é uma característica notável da biologia dos cestódeos. Neste processo de desenvolvimento, há diferenciação distal progressiva a partir do escoléx anterior, resultando na produção em tandem de unidades reprodutivas (proglótides), exibindo graus crescentes de maturação (TSAI et al., 2013). A enorme capacidade reprodutiva e o potencial de crescimento metastático das formas larvais de alguns

organismos desta classe pode produzir consequências patológicas sérias (OLSON et al., 2012). Portanto, as doenças causadas por cestódeos continuam sendo uma ameaça relevante para a saúde pública humana e animal, estando entre as principais doenças tropicais negligenciadas segundo a Organização Mundial de Saúde (WHO, 2013).

1.1.1. *Echinococcus* spp.

*A*s espécies do gênero *Echinococcus* são de importância médica e veterinária, pois a infecção do hospedeiro intermediário por metacestódeos pode acarretar doença severa e morte do hospedeiro, sendo as espécies *E. granulosus* e *E. multilocularis* as mais importantes dentro do gênero, concernente a suas relevâncias em saúde pública e quanto a distribuição geográfica (YANG; RANNALA, 2012; TORGERSON et al., 2015). A espécie *E. granulosus* contém um grande número de variantes genotípicas com diferenças de morfologia, distribuição geográfica especificidade do hospedeiro entre outras características (ROMIG et al., 2015). O *E. granulosus* sensu lato é dividido em genótipos de G1-G8 e G10, porém estudos sugerem não haver diferenças entre os genótipos G1 e G3, devendo estes pertencer à mesma espécie, denominada *E. granulosus* senso stricto (KINKAR et al., 2017). O *E. granulosus* ocorre mundialmente em todos os continentes, incluindo as zonas circumpolar, temperada, subtropical e tropical, enquanto o *E. multilocularis* tem maior ocorrência nas regiões central e nordeste da Europa e Ásia, e nordeste da América do Norte (WHO/OIE, 2001). A região andina e o Cone Sul da América do Sul são áreas altamente endêmicas de equinococose cística; esta, uma doença considerada subnotificada, afeta mais de 1 milhão de pessoas mundialmente e aparenta estar reemergindo (WHO, 2013), podendo-se justificar este fato por não ser uma doença de notificação compulsória em muitos países (ROSSI et al., 2016).

E. granulosus, uma tênia de cães, é o agente causador da equinococose cística. O verme adulto, uma pequena tênia, desenvolve-se em canídeos, e a forma juvenil (metacestódeo), em mamíferos de mais de 40 espécies, incluindo seres humanos, macacos, ovelhas, renas e gado (HICKMAN JR. *et al*., 2013). O ciclo de vida desse parasito é indireto, requerendo dois hospedeiros mamíferos (SOUTO et al., 2016). O verme adulto vive no intestino de cães e outros canídeos, hospedeiros

definitivos do parasito, onde deposita ovos que são liberados para o meio ambiente com as fezes do animal infectado (ROMIG et al., 2017). Ungulados domésticos ou selvagens são hospedeiros intermediários e adquirem a infecção através da ingestão acidental dos ovos, os quais se desenvolvem para o estágio larval de metacestódeo (cisto hidático) nos órgãos internos, por fim causando a patogenia associada à equinococose cística (CARDONA; CARMENA, 2013). O ciclo se completa quando o hospedeiro definitivo ingere esses órgãos infectados (TAYLOR et al., 2009). Seres humanos são considerados hospedeiros intermediários acidentais, pois, apesar de infectáveis pela forma larval, em geral não proporcionam a continuidade do ciclo de vida do parasito (MCMANUS, 2010). Os ovos eclodem no trato gastrointestinal, tornando-se larvas ativas que penetram na parede intestinal e entram na circulação sanguínea, eventualmente localizando-se em órgãos internos, normalmente fígado e pulmões, onde desenvolvem-se em um cisto hidático (AZIZ et al., 2011). O desenvolvimento do cisto hidático no hospedeiro é lento e a maturidade é alcançada em 6-12 meses, atingindo, no fígado e nos pulmões, 5-10 cm em média (TESSELE et al., 2013).

Agente causador da equinococose alveolar, o *E. multilocularis* possui ciclo de vida principalmente silvestre, tendo, em geral, um roedor como hospedeiro intermediário e um canídeo selvagem como hospedeiro definitivo (ROMIG et al., 2015). O hospedeiro intermediário se infecta após ingerir os ovos do parasito contendo oncosferas. Após a eclosão, as oncosferas colonizam com maior frequência o fígado e, diferentemente da infecção causada por *E. granulosus*, não ocorre a formação de cistos, mas a formação de vesículas com proliferação infiltrativa (BRUNETTI et al., 2010), e o metacestódeo de *E. multilocularis* possui uma matriz semisólida ao invés de um fluido como em *E. granulosus* (BROOVÁ et al., 2017). O ciclo do parasito se completa quando o hospedeiro definitivo, via de regra um canídeo selvagem, preda um hospedeiro intermediário infectado pelo metacestódeo do parasito e ingere as vísceras infectadas (TAYLOR et al., 2009). Os seres humanos são hospedeiros acidentais do ciclo de vida do parasito, assim como outros primatas, castores e ratos-almiscarados também o são (GOTTSTEIN et al., 2014). A equinococose alveolar possui um período de incubação, no qual a doença é assintomática, em torno de 10 a 15 anos e, após esse período, os sinais clínicos

iniciais são dor abdominal e icterícia colestática, originando lesões neste órgão, podendo causar séria injúria, progredindo de dores abdominais até falência hepática, e mesmo a morte do hospedeiro (TORGERSON et al., 2010; KNAPP et al., 2015). Por este crescimento infiltrativo e características de disseminação metastática do tecido dos metacestódeos, com comportamento semelhante a um tumor, a equinococose alveolar pode causar morte prematura em estágios avançados, especialmente se continuar sem tratamento ou for tratada de forma inadequada (DU et al., 2016). Os mecanismos exatos de como o metacestódeo consegue se infiltrar no tecido do hospedeiro e evadir do sistema imune permanecem desconhecidos (SASAKI; SAKO, 2017).

1.2. Produtos de excreção ou secreção

O secretoma se refere ao conjunto de proteínas que são excretadas ou secretadas por células, tecidos ou organismos, em um dado momento, sob particular condição fisiológica, patológica ou experimental (INAL et al., 2013; GOMEZ et al., 2015). Ele inclui as proteínas da matriz extracelular, proteínas vesiculares e proteínas lançadas pela membrana celular. Estes produtos de excreção ou secreção (ES) podem ser secretados por via clássica (quando há a presença de peptídeo-sinal) ou via não-clássica (sem a presença de peptídeo sinal) (MAKRIDAKIS; VLAHOU, 2010). Os produtos de ES são responsáveis por diversas funções, abrangendo desde a adesão celular até a ação como neurotransmissores no sistema nervoso (CHOI et al., 2010). Em organismos patogênicos, produtos de ES podem atuar na modulação imunológica de espécies hospedeiras ou como efetores que destroem células ou viabilizam a infiltração do patógeno em tecidos do hospedeiro (WANG et al., 2017).

Os produtos de ES têm um papel importante na interação parasito-hospedeiro já que podem agir como fatores de virulência ou reguladores imunológicos para o sistema imune do hospedeiro, além do mais, esses produtos são cruciais para a sobrevivência do parasito dentro e fora do hospedeiro (SCHICHT et al., 2013; GOMEZ et al., 2015). Outrossim, os produtos de ES podem interferir com demais processos da interação parasito-hospedeiro como a degradação da matriz extracelular em patologias relacionadas ao remodelamento tecidual, manutenção da

capacidade progressiva de crescimento dos metacestódeos e proteção contra danos oxidativos (DITGEN et al., 2014).

Diversas pesquisas recentes têm se focado nos produtos de ES liberados pelos helmintos por sua capacidade de, com esses produtos, causarem alteração no organismo hospedeiro pela modulação do sistema imune (DITGEN et al., 2014). Descobriu-se, por exemplo, que os produtos de ES de *E. granulosus* podem diminuir as defesas imunes por impedirem a maturação e prejudicarem a função de células dendríticas (DC) e por induzirem a geração de linfócitos T $CD4^+$, $CD25^+$ e $FoxP3^+$ (PAN et al., 2014). A indução do aumento das células $FoxP3^+$ e a falta de responsividade das DC, por ação dos produtos de ES, sugere que estes são importantes para o estabelecimento e persistência do parasito no hospedeiro (NONO et al., 2012; VENDELOVA et al., 2016).

Muitas proteínas preditas de ES apontam também para domínios e famílias de peptidases, as quais são conhecidas por estarem envolvidas com a virulência do parasito (GARG; RANGANATHAN, 2012). As proteases secretadas têm sido descritas como enzimas-chave para a degradação do tecido do hospedeiro, excistamento/encistamento, invasão tecidual e migração larval, entre outras funções. Ademais, proteases como as peptidases, por serem notavelmente imunogênicas, podem ser exploradas como marcadores serodiagnósticos ideais e elementos de escolha para o desenvolvimento de vacinas (WANG et al., 2015a).

1.3. Principais métodos de predição *in silico* de produtos de ES

A predição *in silico* dos produtos de ES auxilia na obtenção de um provável perfil secretório de uma célula de forma rápida, permitindo o estudo direcionado a proteínas identificadas como de interesse. Os principais métodos utilizados por *softwares* de predição de proteínas secretáveis são baseados em matrizes de peso, alinhamento de sequências e algoritmos de aprendizagem de máquina (do inglês, *machine learning*) (CACCIA et al., 2013).

As matrizes de peso medem a probabilidade de se encontrar um aminoácido em certa frequência em uma dada posição de acordo com um limiar mínimo estabelecido (peso) (CHEN et al., 2007). A matriz é gerada com base em um

conjunto de sequências, que servirão de parâmetro para se calcular a confiabilidade do resultado.

O alinhamento de sequências compara a sequência inquirida com outra ou outras sequências que possuem a característica a ser identificada (OROBITG et al., 2015). Há casos como o de identificação de petídeos-sinal para os quais esta abordagem pode não ser de grande valia, devido ao fato dos petídeos-sinal não possuírem sequências conservadas e apresentarem grande discrepância em seus tamanhos (PEARSON et al., 2005; LAI et al., 2012). Isso prejudica o real valor preditivo de algoritmos baseados em alinhamento de sequências para a identificação de peptídeos-sinal.

Abordagens mais atuais usam algoritmos de aprendizagem de máquina, um campo da ciência que busca desenvolver algoritmos computacionais que possam se adaptar e aprender por experiência (CILINGIR; BROSCHAT, 2015). A aprendizagem de máquina é usualmente classificada em três categorias gerais: aprendizagem supervisionada, aprendizagem não-supervisionada e aprendizagem por reforço. *Softwares* baseados em aprendizagem de máquina obtiveram resultados superiores de confiabilidade para a predição de produtos de ES quando comparados com outros métodos (CHOO et al., 2009), sendo os algoritmos baseados em aprendizagem de máquina os mais usados nos *softwares* que compõem as *pipelines* mais atuais de predição de produtos de ES, a exemplo dos *softwares* SignalP 4.1, SecretomeP e TargetP.

1.4. Justificativas

Os metacestódeos de *E. granulosus* e *E. multilocularis* são capazes de sobreviver e proliferar por muitos anos em um hospedeiro intermediário, mesmo em indivíduos imunologicamente competentes (AHN et al., 2017). Esta capacidade de se manter viável e fértil por longos períodos de tempo é devida, em parte, a proteínas secretadas pelo parasito no fluido (líquido hidático) que preenche cistos ou vesículas hidáticas. As proteínas parasitárias presentes no líquido hidático agem como uma linha de frente para combater as células do sistema imune do hospedeiro, possibilitando a sobrevivência do parasito e o subsequente desenvolvimento da doença (SILVA-ÁLVAREZ et al., 2016). Além disso, outras proteínas secretadas dos

metacestódeos de *Echinococcus* spp. podem participar de processos de captação e assimilação de lipídeos do hospedeiro, não sintetizáveis pelo parasito, e podem ter ação antioxidante contra o estresse oxidativo ou podem ter atividades proteolíticas, importantes para mecanismos de defesa e de assimilação (TSAI et al., 2013; SILVA-ÁLVAREZ et al., 2015).

As manifestações clínicas das equinococoses, especialmente no caso da equinococose cística, normalmente só ocorrem anos após o início da infecção pelo parasito, quando o hospedeiro já está com um nível alto de infecção e cistos de grande volume, que acabam por afetar os órgãos atingidos e tecidos adjacentes (PETRONE et al., 2017). Neste cenário, o combate à doença torna-se mais complicado, necessitando-se estratégias mais agressivas para sua eliminação, podendo deixar com sequelas o hospedeiro afetado (AKBULUT et al., 2018). Uma cura segura e eficiente para equinococoses e outras cestodíases ainda precisa ser descoberta (PENSEL et al., 2017). Há uma necessidade urgente, devido à relevância destas doenças, de novos alvos para o desenvolvimento de testes diagnósticos mais eficientes, de novas drogas anti-helmínticas e de vacinas (KERN, 2010; BREHM; KOZIOL, 2016), imprescindíveis para o controle, tratamento e prevenção de cestodíases.

Devido à urgência por novas formas de prevenção e combate às equinococoses (e de outras cestodíases), os produtos de ES dos parasitos, em razão de seus altos potenciais antigênicos, de virulência e a suas características imunomodulatórias (NAZ et al., 2015), têm sido apontados como biomarcadores e como candidatos adequados ao desenvolvimento de vacinas (DITGEN et al., 2014). Muitos parasitos apresentam mecanismos bastante similares para se evadir das defesas imunes do hospedeiro, particularmente, da imunidade inata (HEWITSON et al., 2009). Sendo assim, o estudo e a investigação dos produtos de ES de *Echinococcus* spp., com ferramentas e métodos identificados como confiáveis, pode ser um ponto de partida para a identificação de produtos de ES comuns a outras espécies de parasitos que possuem dados genômicos de baixa qualidade ou que ainda estejam em fase de anotação, por exemplo.

Para a identificação de produtos de ES, é interessante a utilização de uma abordagem *in silico*, pois, enquanto o processo de identificação de produtos de ES

através de métodos experimentais é oneroso e demanda muito tempo, uma abordagem bioinformática é mais custo-efetiva, visto que, a partir das predições *in silico*, pode-se priorizar, na análise experimental, alvos selecionados com base na sua drogabilidade, por exemplo (GAHOI; GAUTAM, 2017). A predição *in silico* de produtos de ES e os dados por ela gerados são também importantes para a confirmação de dados experimentais proteômicos (por espectrometria de massas) voltados à identificação e quantificação de produtos de ES (AYALEW et al., 2017). A predição *in silico* também pode ser utilizada para direcionar a busca de alvos para o desenvolvimento de vacinas pelo método de "vacinologia reversa" (SERRUTO; RAPPUOLI, 2006; RAPPUOLI, 2001), através do qual, a partir das sequências genômicas, busca-se alvos mais adequados para tal fim (BRUNO et al., 2015; BAMBINI; RAPPUOLI, 2009).

Os algoritmos mais utilizados para a predição *in silico* de produtos de ES são baseados em aprendizagem de máquina, método este que exige treinamento do algoritmo com sequências que possuam as características adequadas para se obter resultados confiáveis (MIN et al., 2016). Todavia, não há ainda *softwares* treinados com sequências de organismos próximos a *E. granulosus* e *E. multilocularis*, o que afeta a confiabilidade dos resultados obtidos para estas e outras espécies de cestódeos com os *softwares* hoje disponíveis. Isto é ainda mais significativo ao se fazer a predição de proteínas secretadas pela via de secreção não-clássica (sem presença de peptídeo-sinal), pois um achado comum em estudos proteômicos de líquido hidático de metacestódeos de *Echinococcus* spp. é a presença de proteínas sem sequências de sinalização por via clássica (BREHM; KOZIOL, 2017; MONTEIRO et al., 2017). É, portanto, essencial o desenvolvimento de algoritmos ou estratégias que proporcionem a maior confiabilidade possível às predições de produtos de ES.

2. OBJETIVOS

2.1. Objetivo geral

Realizar um estudo comparativo *in silico* dos produtos de excreção ou secreção preditos de helmintos parasitos das espécies *Echinococcus multilocularis* e *Echinococcus granulosus*, identificar inconsistências e propor meios mais confiáveis para obtenção de produtos de ES preditos.

2.2. Objetivos específicos

2.2.1. Predizer os secretomas e as vias de secreção para *E. granulosus*. e *E. multilocularis.*

2.2.2. Comparar e validar reciprocamente os conjuntos de proteínas secretáveis preditas para *E. granulosus* e *E. multilocularis*.

2.2.3. Identificar predições inconsistentes de proteínas secretáveis de *E. granulosus* e *E. multilocularis* e, com base nelas, aprimorar as estratégias ou ferramentas de predição.

3. Manuscrito – *Echinococcus* spp. in silico comparative secretomics calls into question: can we rely on current protein secretion predictors for non-model organisms?

3.1. Apresentação

O artigo que compõe esta seção foi elaborado conforme o formato exigido para submissão à revista *GigaScience* (https://academic.oup.com/gigascience). Todos os experimentos descritos foram realizados pelo aluno Tiago Minuzzi F. F. Gomes, sendo os demais autores responsáveis pela sua orientação.

Echinococcus spp. *in silico* comparative secretomics calls into question: can we rely on current protein secretion predictors for non-model organisms?

Tiago Minuzzi F. F. Gomes[1,2], Gabriela P. Paludo[1,2], Luis William Pacheco Arge[3], Ana Tereza Ribeiro de Vasconcelos[3], Arnaldo Zaha[2], Henrique B. Ferreira[1,2]*.

1. Structural and Functional Genomics Laboratory, Center of Biotechnology, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil.
2. Cestodes Molecular Biology Laboratory, Center of Biotechnology, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil.
3. Laboratorio Nacional de Computação Cientifica (LNCC), Petrópolis, RJ, Brazil.

*Corresponding author:
Henrique B. Ferreira
Universidade Federal do Rio Grande do Sul, Centro de Biotecnologia.
Av. Bento Gonçalves, 9500 – Prédio 43421 – Sala 210.
Campus do Vale/UFRGS.
91501970 – Porto Alegre/RS, Brasil – Caixa Postal: 15005
Phone number (+55 51) 33087768
e-mail: henrique@cbiot.ufrgs.br

Abstract

The advances in the “omics” studies generated an enormous amount of data that needs to be analyzed by proper tools and methods to obtain reliable results. However, many softwares are still not strongly reliable when applied to datasets less related to the ones used for training the algorithm. This is specially noticeable when analysing datasets from non-model organisms. Here, the secretomes of parasite cestodes *Echinococcus granulosus* and *Echinococcus multilocularis* were predicted by a standard workflow and compared in order to identify inconsistencies between predictions for ortholog proteins. A high proportion (~25%) of sequences predicted as secreted for one species had a ortholog not predicted as secreted in the other species, which would not be expected in such extent for closely related species. In

order to identify possible causes for that, it was found that from this ~25% of inconsistent predictions, ~75% were generated in the workflow at the SecretomeP level. The ortholog sequences were also aligned to identify possible genome misannotations. To minimize inconsistencies in the predictions of secretion for ortholog proteins, the two sequenced genomes were reannotated, based on transcriptomic data, and WoLF PSORT was proposed as a new step in the secretion prediction workflow, which led to improved secretome predictions for both species. It was the first time that a cross-validation of secretome predictions by comparative analysis using closely related species was made. Applications of the modified workflow and implications of the revision of predicted *Echinococcus* spp. secretomes are discussed.

Introduction

The post-genomic era brought us new challenges on the way towards the elucidation of the secrets hidden inside genomes. The solving of genomic riddles are essential for understanding how living beings function and behave at the molecular level. In this scenario, bioinformatics became a highly important discipline due to the huge amounts of genomic data that has been generated. Moreover, additional big data brought by different kinds of “omics” studies, such as transcriptomics and proteomics, required the development of novel computational approaches to be adequately interpreted and explored [1,2]. To get reliable conclusions from *in silico* analysis, data of high quality are necessary, along with appropriate tools and methods to analyse the data [3].

Several bioinformatic softwares use machine learning (ML) algorithms to analyse the different sorts of biological data [4–6]. The major resources for training the ML algorithms are public databases, which are continuously updated. However, these data are not usually incorporated in already published machine learning algorithms, and, therefore, they rapidly become outdated [7,8]. The use of softwares with outdated ML algorithms may be a major issue when performing *in silico* biological analyses, such as predictions of secreted proteins and its secretion pathways.

Secretomics, a subfield of proteomic analyses, is the global study of secreted proteins, and the secretome is the set of proteins secreted by a given cell, at a given time, under a particular physiological, pathological or experimental condition [9,10]. Proteins can be secreted either by so called “classical pathway” (proteins with a signal peptide), or by a “non-classical pathway” (proteins without a signal peptide) [11]. Parasite secretomics is a blooming field, very important to elucidate parasitic strategies that allow the persistent infection of suitable hosts, and to develop new forms to prevent or treat parasitic diseases [12,13].

Helminth parasites are quite interesting subjects to investigate the roles of ES products in parasite-host relationships, as they tipically establish intimate and long-term contacts with specific hosts species. In this sense, *Echinococcus granulosus* and *Echinococcus multilocularis*, two closely related species belonging to the Cestoda class of flatworms (Plathyhelminthes) are of major interest. Beside being relevant for their impact in both human and veterinary medicine, as etiological agents of different forms of echinococcosis [14], they are also attractive models to address the role of ES products for parasite survival in the context of long term infections of intermediate hosts. The pathogenic metacestode larvae of *Echinococcus* spp. are cysts or vesicles filled with a fluid (hydatid fluid) rich in ES products from both parasite and host origin [15]. ES products from *Echinococcus* spp. metacestodes are known to mediate host immunomodulation and other mechanisms important for parasite’s survival and development [16], and the content of ES proteins in *E. granulosus* and *E. multilocularis* hydatid fluid have been characterized by previous proteomic studies conducted by our group [15,17,18].

The proteomic approaches used for the identification of ES products in *Echinococcus* hydatid fluid are comprehensive, but present sensitivity limitations. More represented proteins in the analysed samples impair detection of minor components, which, despite their low representativity, may also be of biological relevance [19]. Therefore, *in silico* predictions based on whole genome sequences are important for the description of the full set of secreted proteins (the secretome) of an organism. Moreover, the prediction of the whole secretome helps to select proteins of interest for wet lab functional studies or for their characterization as diagnostic antigens or potential drug targets, for example [20]. The published *in silico*

secretome surveys of *Echinococcus* and other helminths were all performed using essentialy the same workflow for predictions of protein secretion and secretion pathways [13,20,21]. Although comprehensive and useful, these surveys, like others, performed with non-model organisms, faced some limitations. For instance, most of the assessed helminth genomes are available only as draft versions, still with potential problems in sequence assembly and/or annotation [22,23]. Moreover, software updates do not undergo the same constant updating as sequence databases [19], and this means that softwares used for helminth secretome assessment were based on algorithms developed and validated with sequences of distantly related organisms (e.g. vertebrates or bacteria). These limitations reinforce the need for developing new *in silico* strategies to lower as much as possible the eventual inconsistences in the predictions of secretion proteins and secretion pathways for non-model organisms.

Herein, is presented a comparative *in silico* study of secretory products from *E. granulosus* and *E. multilocularis.* Inconsistences found in secretion predictions for orthologous proteins led to the revision of the genome annotations and to the customization of the workflow of analysis. With that, more reliable secretome predictions were generated for both species. The applicability of the modified strategy for more accurate secretome predictions to non-model organisms in general is discussed.

Results

E. granulosus and *E. multilocularis* secretome prediction

The *E. granulosus* and *E. multilocularis* predicted proteomes were analysed *in silico* using the standard workflow for predictions of secretion products. The generated results are shown in Table 1.

From the 10,274 proteins of the deducted *E. granulosus* proteome TMHMM identified 8173 as transmembrane (TM)-free proteins and 904 proteins as containing only one TM region. These 904 proteins containing a single TM region were then further analysed by Phobius and 158 of them were predicted as TM-free. The set of 8331 TM-free sequences was submitted to SignalP, resulting in 507 sequences predicted as classical secreted proteins. Out of the 7824 sequences without a signal

peptide, SecretomeP classified 253 as non-classical secreted proteins. The set of 760 proteins including those predicted as classical secreted (507) or non-classical secreted (253) was then analysed by TargetP, identifying in 28 mitochondrial targeted proteins which were excluded, along with 6 proteins containing the endoplasmic reticulum retention motif PS00014, identified by Scan Prosite, and 64 proteins containing a GPI-anchor predicted by PredGPI. Overall, 662 *E. granulosus* proteins were predicted as secreted, which corresponds to ~6.4% of the whole predicted proteome.

From the 10,552 proteins deducted *E. multilocularis* proteome, TMHMM identified 8402 as TM-free sequences and 924 proteins as containig a single TM region. The 924 sequences with only one TM region were then further analysed by Phobius and 162 sequences were predicted as TM-free sequences. The set of 8564 TM-free sequences were submitted to SignalP, resulting in 552 sequences predicted as classical secreted proteins. Of the 8011 sequences without a signal peptide, SecretomeP classified 228 as non-classical secreted proteins. The set of 780 sequences predicted as classical secreted and non-classical secreted proteins were then analysed by TargetP, resulting in 26 mitochondrial targeted proteins, which were excluded along with 9 proteins containing the endoplasmic reticulum retention motif PS00014, identified by Scan ProSite, and 74 proteins containing a GPI-anchor, predicted by PredGPI. Overall, 669 *E. multilocularis* proteins were predicted as secreted, which correspond to ~6.3% of the whole predicted proteome.

Table 1 Summary of the *E. granulosus* (left) and *E. multilocularis* (right) predictions using the standard workflow of secretome analysis.

E. granulosus	Nº	%PT
Proteome (PT)	10274	100,00
TMHMM ≥ 1	2101	20,45
∟TMHMM = 1	904	8,80
TMHMM = 0	8173	79,55
Phobius = 0	158	1,54
TM free	8331	81,09
Signal P = Y	507	4,93
Signal P = N	7824	76,15
Secretome P ≥ 0.9	253	2,46
class. & non-class. proteins	760	7,40
Target P ≠ M	732	7,12
Target P = M	28	0,27
Scan Prosite ≠ ER	726	7,07
Scan Prosite = ER (ps00014)	6	0,06
PredGPI	662	6,44
PredGPI (anchored)	64	0,62
ES proteins	**662**	**6,44**

E. multilocularis	Nº	%PT
Proteome (PT)	10552	100,00
TMHMM ≥ 1	1226	11,62
∟TMHMM = 1	924	8,76
TMHMM = 0	8402	79,62
Phobius = 0	162	1,54
TM free	8564	81,16
Signal P = Y	552	5,23
Signal P = N	8011	75,92
Secretome P ≥ 0.9	228	2,16
class. & non-class. proteins	780	7,39
Target P ≠ M	752	7,13
Target P = M	26	0,25
Scan Prosite ≠ ER	237	2,25
Scan Prosite = ER (ps00014)	9	0,09
PredGPI	669	6,34
PredGPI (anchored)	73	0,69
ES proteins	**669**	**6,34**

Comparison between the *E. granulosus* and *E. multilocularis* predicted sets of secreted proteins

Despite the similarities in overall numbers between the *E. granulosus* and *E. multilocularis* sets of secreted proteins predicted by the standard workflow, they were further compared in order to identify possible inconsistencies between predictions for ortholog proteins. Prior to that, the Reciprocal Best Blast Hits (RBH) method was used to find the orthologs of the *E. granulosus* proteins predicted as secreted in the predicted *E. multilocularis* proteome, and the *E. multilocularis* proteins predicted as secreted in the predicted *E. granulosus* proteome.

According to the RBH results, 414 ortholog proteins were predicted as secreted both in *E. granulosus* and *E. multilocularis* (Figure 1). However, 111 proteins predicted as secreted in *E. granulosus* had a non-secreted ortholog in *E. multilocularis* and 103 proteins predicted as secreted in *E. multilocularis* had a non-secreted ortholog in *E. granulosus*. Therefore, relatively large proportions of the secretion predictions were inconsistent between *E. granulosus* and *E. multilocularis* (~17% and ~15%, respectively). Moreover, in the performed reciprocal searches, no orthologs were found for 137 *E. granulosus* and 152 *E. multilocularis* proteins predicted as secreted (Supplementary table 1). These proteins might constitute sets of species-specific secretion products.

Assuming that most of the observed inconsistencies in secretion predictions for *E. granulosus* and *E. multilocularis* orthologs might be the result of the lack of customization of the used softwares for these two species, we initially investigated in which point of the standard workflow the differences between orthologs were generated. As shown in Supplementary tables 2A and 2B, most inconsistencies were generated at the SecretomeP step. There, from the 111 proteins not predicted as secreted in *E. multilocularis* with a predicted secreted ortholog in *E. granulosus*, 85 (76%) were eliminated. Likewise, from the 103 proteins not predicted as secreted in *E. granulosus* with a predicted secreted ortholog in *E. multilocularis*, 77 (74%) were eliminated. The other inconsistencies in the secretions predictions for *E. granulosus* and *E. multilocularis* were generated, respectively, at the TMHMM (3.6% and 5.8%), Phobius (8.1% and 11.65%), TargetP (3.6% and 4.85%) or PredGPI (6.3% and 2.91%) steps.

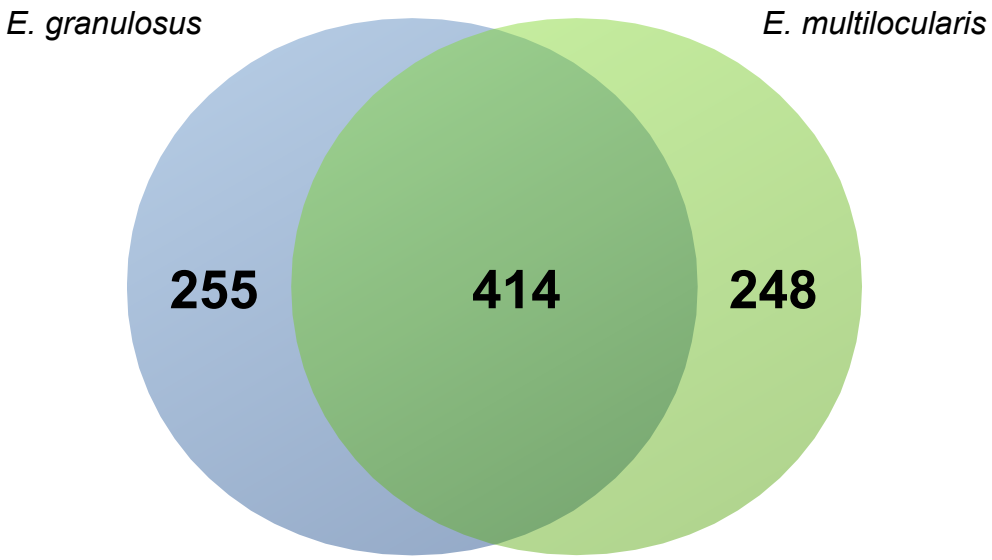

Fig. 1 Overall numbers of exclusive and shared secreted proteins in the secretomes of *E. granulosus* and *E. multilocularis*.

Workflow refinement to improve secretion predictions for *E. granulosus* and *E. multilocularis* orthologs

In order to improve the reliabilty of the secretome predictions for *E. granulosus* and *E. multilocularis,* WoLF PSORT was used to analyse the 85 *E. multilocularis* and the 77 *E. granulosus* proteins eliminated by SecretomeP. As can be seen in Supplementary tables 2A and 2B, the WoLF PSORT analyses allowed to classify 51 (60%) of the 85 *E. multilocularis* SecretomeP-eliminated proteins and 49 (64%) of the 77 *E. granulosus* SecretomeP-eliminated proteins as non-classical secreted proteins. With the WoLF PSORT analyses, the *E. granulosus* and *E. multilocularis* predicted numbers of secreted proteins increased from 662 to 711 proteins and from 669 to 720 proteins, respectively, and the number of *E. granulosus* and *E. multilocularis* orthologs predicted as secreted increased ~20%, from 414 to 514. On the other hand, the number of inconsistencies in the predictions of secretion for *E. granulosus* and *E.* multilocularis orthologs decreased from 214 to 114 proteins (54 proteins not predicted as secreted in *E. granulosus* with a secreted ortholog in *E. multilocularis*,

and 60 proteins not predicted as secreted in *E. multilocularis* with a secreted ortholog in *E. granulosus*).

Reannotation of *E. granulosus* and *E. multilocularis* genome sequences to improve secretion predictions

Many discrepancies were detected when the secreted/non-secreted ortho-pairs from *E. granulosus* and *E. multilocularis* were aligned, including indels and amino acid mismatches (data not shown), which could explain at least part of the inconsistencies between secretion predictions for ortholog proteins. Thus, to further refine the secretome predictions, the corresponding sequenced *E. granulosus* and *E. multilocularis* genomes were reannotated based on transcriptomic data. First, *E.* granulosus and *E. multilocularis* transcriptomic data from EST and RNA-seq libraries were assembled using CAP3 and Trinity, respectively, and then combined. The genomic reannotation was performed using the combined transcriptomic assemblies in the PASA pipeline.

Cap3 assembled 28834 *E. granulosus* EST libraries into 10082 contigs, and 1168 *E. multilocularis* EST libraries into 774 contigs. Trinity assembled the *E. granulosus* SRR1508667 and SRR1508668 RNA-seq libraries into 119936 contigs, and the *E. multilocularis* SRR1508669 and SRR1508670 RNA-seq libraries into 143555 contigs. Using these transcriptomic data, PASA reannotation resulted in updated *E. granulosus* predicted proteomes comprehending 10,082 and 10,554 proteins respectively. Only 3398 (~33%) of the *E. granulosus* proteins and 4093 (~39%) of the *E. multilocularis* proteins kept the same amino acid sequence of the original annotation.

Based on the PASA revised proteomes, the *E.* granulosus and *E. multilocularis* secretome predictions were updated using the standard workflow. Their updated secretomes comprehended 658 proteins (6.5% of the proteome) and 581 proteins (5.5% of the proteome), respectively. In the *E. granulosus* and *E. multilocularis* updated secretomes, 283 proteins (~43%) and 280 proteins (~48%), respectively, kept the original sequence.

The *E. granulosus* and *E. multilocularis* reannotated secretomes share 325 orthologs. There was still 106 proteins not predicted as secreted in *E. granulosus*

with a *E. multilocularis* secreted ortholog, and 186 proteins not predicted as secreted in *E. multilocularis* with a *E. granulosus* secreted ortholog. From the 186 proteins not predicted as secreted in *E. multilocularis* with a predicted secreted ortholog in *E. granulosus*, 122 (~66%) were eliminated at SecretomeP level. Likewise, from the 106 proteins not predicted as secreted in *E. granulosus* with a predicted secreted ortholog in *E. multilocularis*, 84 (~79%) were eliminated at SecretomeP level.

Of the 106 *E. granulosus* proteins not predicted as secreted with a secreted ortholog in *E. multilocularis*, just a single one had its amino acid sequece altered in the reannotation and this changed its prediction to 'secreted'. This same protein was predicted as an extracellular one by WoLF PSORT prior to the reannotation and this WoLF PSORT prediction was mantained with the new annotation.

Of the 186 the *E. multilocularis* proteins not predicted as secreted with a secreted ortholog in *E. granulosus*, 48 had their amino acid sequence altered in the reannotation, but only 2 of them had their prediction changed to 'secreted' due to the revised annotation. These 2 proteins were predicted as a extracellular ones by WoLF PSORT prior to the reannotation and remained predicted as such after their annotation revision.

Functional comparative analyses and antigenic regions prediction of the *E. granulosus* and *E. multilocularis* predicted secretomes

Functional analyses were performed with the revised *E. granulosus* and *E. multilocularis* secretomes, in order to identify differences that could be related to the distinct biological features of these two related species. From the set of 658 *E. granulosus* predicted secreted proteins, 438 proteins were predicted as classical secreted proteins, and 220 as non-classicaly secreted (Supplementary table 3A). From the set of *E. multilocularis* 581 proteins predicted as secreted, 388 were predicted as classicaly secreted proteins, and 193 as non-classicaly secreted proteins (Supplementary table 3B).

Moreover, the proteins comprehended in the predicted secretomes of *E. granulosus* and *E. multilocularis* were classified according to GO terms to infer possible differences between these two species in functions involving secreted proteins. From the totals of 658 *E. granulosus* secreted proteins, and 581 *E.*

multilocularis secreted proteins, 289, and 270, respectively, were successfully categorized according to GO terms into ‘biological process’ (BP), ‘molecular function’ (MF) and ‘cellular component’ (CC) categories (Supplementary table 4 A-E). No annotations were retrieved for 369 *E. granulosus* proteins, and 309 *E. multilocularis* proteins. Table 2 shows the top 10 most represented exclusive GO terms for *E. granulosus* and *E. multilocularis* secretomes.

The functional enrichment analysis of *E. granulosus* and *E. multilocularis* showed several functions related to BP, MF CC categories that are represented in both the *E. granulosus* and *E. multilocularis* predicted secretomes such as 'lipid transport', 'peptidase activity', 'cell adhesion', 'granular secretion', 'extracellular space' and 'immune response'. Other functions were enriched only in the *E. granulosus* or in the *E.* multilocularis secretome. The *E. granulosus* secretome showed specific enrichment in BP subcategories like 'protein dephosphorylation', 'protein glycosylation' and 'protein arginylation'; in MF subcategories like 'ubiquitin phosphorylase activity', 'dipeptidyl peptidase activity' and 'lipoate synthase activity'; and in CC subcategories like 'RES complex', 'P-body' and 'glycerol-3-phosphatase complex dehydrogenase'. The *E. multilocularis* secretome, in turn, showed specific enrichment in BP subcategories like 'cell proliferation', 'vesicular organization' and 'endocytosis'; in MF subcategories like 'peroxidase activity', 'ferroxidase activity' and 'GTPase activity'; and in CC subcategories like 'focal adhesion', 'endosome' and 'extrinsic membrane component'.

Table 2 Top 10 most represented exclusive GO terms found in the functional analysis of the *E. granulosus* (A) and *E. multilocularis* (B) revised secretomes.

GO_ID	Number_of_ocurrences	Description	Categorie
GO:0005840	4	ribosome	CC
GO:0006486	4	protein glycosylation	BP
GO:0008417	3	fucosyltransferase activity	MF
GO:0000398	2	mRNA splicing	BP
GO:0003824	2	catalytic activity	MF
GO:0005763	2	mitochondrial small ribosomal subunit	CC
GO:0006470	2	protein dephosphorylation	BP
GO:0045944	2	positive regulation of transcription from RNA polymerase II promoter	BP
GO:0071011	2	precatalytic spliceosome	CC
GO:0071013	2	catalytic step 2 spliceosome	CC

(A)

GO_ID	Number_of_ocurrences	Description	Categorie
GO:0005622	3	intracellular	CC
GO:0003924	2	GTPase activity	MF
GO:0005525	2	GTP binding	MF
GO:0005578	2	proteinaceous extracellular matrix	CC
GO:0008283	2	cell proliferation	BP
GO:0015994	2	chlorophyll metabolic process	BP
GO:0030206	2	chondroitin sulfate biosynthetic process	BP
GO:0001104	1	RNA polymerase II transcription cofactor activity	MF
GO:0001575	1	globoside metabolic process	BP
GO:0002143	1	tRNA wobble position uridine thiolation	BP

(B)

Two methods were independently used to predict antigenicity of *in silico* predicted repertoires of the *E. granulosus* and *E. multilocularis* secreted proteins. These results are shown in Supplementary table 5 A-D. A total of 650 *E. granulosus* predicted secreted proteins had antigenicity predicted by methods. For these proteins, values of abundance of antigenic regions (AARs) ranging from 14.2 to 104 were calculated based on the Kolaskar-Tongaonkar method predictions, and AAR values ranging from 19 to 2211 were calculated based on the BepiPred-2.0 predictions, with overall AAR means of 26.20 and 56.47, respectively.

For *E. multilocularis,* 576 predicted secreted proteins had antigenicity predicted by both methods. For these proteins, Aar values ranging from 13 to 155.25 were calculated based on the Kolaskar-Tongaonkar results, and AAR values ranging

from 18 to 1008 were calculated based on the BepiPred-2.0 results, with overall AAR means of 25.75 and 57.71, respectively.

Discussion

Parasitic helminth secretomics is a field that has benefited from the huge amount of data generated by 'omics' studies. With the increasing number of sequenced helminth genomes, it is of upmost importance to have reliable *in silico* predictions of the encoded secreted proteins. Such proteins are key elements in interactions at the host-parasite interface [21,24]. Moreover, parasite secretomes are major sources of novel antigens for diagnosis and vaccine development, as well as of target molecules for the development of novel anti-parasitic drugs [20].

Some softwares used in workflows of secretome prediction, like SignalP, TargetP, TMHMM, have been trained using data from different eukaryotic or more related organisms. This is a factor that helps to obtain more reliable results for data from a wide range of species. Indeed, the results generated for *E. granulosus* and *E. multilocularis* with these softwares in the used standard workflow were mostly coincident for the ortholog proteins, assumed to be of the same secretory natures. SecretomeP, on the other hand, has been trained using only data from mammals and gram positive and gram negative bacteria [25], leading to outcomes not as accurate for less related organisms, like plants [26] or, as presented here, helminths.

In our study, the high number of proteins predicted as secreted in one *Echinococcus* species, but with an ortholog not predicted as secreted in another species of the same genus was higher than expected, and it was taken as a clear evidence of inconsistencies in the predictions provided by the standard workflow. Also, in recent proteomic studies performed by our group [17,18], many proteins detected in excretion/secretion compartments for both *E. granulosus* and *E. multilocularis*, were predicted as secreted for one species, but not to the other (data not shown). Moreover, based in proteomic analyses of *in vivo* and *in vitro* ES products of *E. granulosus* [17], *E. multilocularis* [18] and *Mesocestoides corti* [27], 37,3%, 52,7% and 56,4% of the proteins detected by mass spectrometry, respectively, could not have their secretory pathways identified *in silico*.

In the standard workflow of secretion analysis used here, the SecretomeP step was the major responsible for the observed inconsistencies in the secretion predictions, as problems caused by genome misannotations were relatively minor. Moreover, *in silico* secretomic studies carried out with other parasitic platyhelminths [28], fungi [29], and plants [26] also pointed out to the limitations of using a software trained with organisms not closely related to the query ones.

SecretomeP needs several other softwares to function. One of them is are PSORTII [30], which predicts the subcellular localization of proteins. PSORTII, however, is already outdated [31]. WoLF PSORT, on the other hand, is a more updated PSORTII extension, which is much more accurate for predicting subcellular localization. It classifies proteins into more than 10 localization sites, with an estimated sensitivity and specificity around 70% for nucleus, mitochondria, cytosol, plasma membrane and extracellular space predictions [32]. WolF PSORT utilizes Uniprot database anntotations to score the probability of each subcellular localization [33], and is known as one of the most reliable softwares of protein secretion prediction [29,32]. With the use of WoLF PSORT as a complementary step to predict non-classical secreted proteins in our customized workflow, we were able to achieve more consistent predictions of secretion or not for *E. granulosus* and *E. multilocularis* ortholog proteins, mostly coincident.

In our analyses, we considered proteins classified as cytoplasmic as secreted proteins because proteins can be secreted via vesicles like exososomes, exososomes-like vesicles, lysossomes and microvesicles in the non-classical secretory pathway [12,34]. Proteins classified as transmembrane proteins by WoLF PSORT were also considered as secreted, since TMHMM and Phobius were previously used in the workflow to discard TM proteins and because WoLF PSORT may misinterpret a signal peptide as a TM domain, due to the presence of the hydrophobic moieties present in both signal peptide and TM domains [29,35].

Workflows for *in silico* secretome predictions are widely used, and mostly correspond the standard one initially used in our analyses [9,20,36]. They are usually used for one or few species, and comparative analysis in general based on the overall numbers of proteins secreted by the different secretion pathways. As most studies do not compare predictions for ortholog proteins of closely related species,

they are not able to detect the types of inconsistencies observed here when comparing the secretomes of two *Echinococcus* species. Therefore, many predictions inconsistencies may be overlooked in *in silico* secretome survey that do not indiviadually check the coincidence or not of predictions for ortholog proteins.

Even the complete genome reannotation for *E. granulosus* and *E. multilocularis*, did not significantly improved the coincidence between the predictions for ortholog proteins, as the use of the outdated PSORTII used by the SecretomeP step in the standard workflow was the major determinant of inconsistencies. This reinforces the need of PSORTII replacement by, or complementation with, more updated softwares, like WoLF PSORT, in the workflow. More desirable than that would be the laborious and time consuming training of the algorithms with more related sets of data, prior to the analysis of the query data. When not possible, the parallel and comparative analysis of closely related species and the cross-check of coincidence between predictions for ortholog proteins would be a good alternative to verify the degree of accuracy of the predictions.

The revised predicted secretomes from *E. granulosus* and *E. multilocularis* provided more dependable datasets for comparative analyses. The performed GO functional analysis of predicted secretomes showed that *E. granulosus* and *E. multilocularis* presented overall similarities, as expected for closely related species. For both species, it was observed a large number of secreted proteins involved in proteolysis and peptidase inhibition activities, functions of major importance for host invasion and establishment of the parasite in a target organ [27]. It has been suggested that, during chronic stages of the infection, *Echinococcus* spp. metacestodes can survive despite the strong host immune responses at least in part due to immunomodulation mediated by parasite's proteases and/or peptidases [37]. Also secreted proteases are related to host-tissue degradation, tissue invasion and larval migration [38]. Proteins with GO terms assigned to carbohydrates metabolizing activties, like GO:0015018, GO:0030247, GO:0006013, were also observed in both species. The carbohydrate-metabolizing proteins assigned to these GO terms are usually related to nutrient uptake, but some studies indicated that they may have moonlighting functions, playing roles like antioxidant detoxification, IgA immunomodulation, cellular adhesion and invasion [39].

Moreover, different arginase-related GO terms were assigned exclusively to either *E. granulosus* or *E. multilocularis*. Arginases contribute to the protection of the parasite from the host's NO compounds produced on a attemptive to kill the parasites [40] by competing with the host nitric-oxide-synthase-2 for the same substrate L-arginine, reducing the production of NO through arginine depletion [41].

The revised predicted secretomes of *E. granulosus* and *E. multilocularis* were still further analysed for antigenicity predictions by two independent methods. The derived AAR values were similar ranges and also similar in average for both secretomes, pointing out to several secreted proteins with high epitope densities, and, therefore, with correspondingly high probabilities of being antigenic and immunogenic [9]. Based on that, our data provided several proteins with potential for use as immunodiagnostic antigens or as components of recombinant vaccine formulations. Immunodiagnosis is the more specific diagnostic approach available for echinococcoses, although improvements are still required in order to provide better sensitivities and specificities for the available tests [42], which depends on the characterization of novel antigenic proteins, especially secreted ones. Moreover, it is well known that the ES products of parasitic helminths may provide antigenic stimulation, which may elicit protective responses in mammal hosts [43,44]. Also, it was shown for *Fasciola hepatica*, another cestode parasite, that ES products can be useful to develop monoclonal antibodies capable to discriminate the acute and invasive periods of the infection [45]. Therefore, a rationale based on AAR values of *E. granulosus* and *E. multilocularis* predicted ES products is feasiable and reliable for the initial selection of potential target antigens for immunodiagnosis and/or vaccination. As observed by the results of the functional enrichment, functional predictions may also help in target selection, by assigning antigenic/immunogenic proteins to functions important for parasite's survival and development [46].

The different types of echinococcosis are still burdensome diseases that require innovative strategies for prevention, diagnosis and treatment. By using complementary *in silico* approaches of secretome prediction, functional enrichment and antigenicity prediction, many novel molecules with functional importance and antigenicity/immunogenicity potential could be selected for wet lab experiments. This is expected to shorten the time necessary to discover and select new and better

Echinococcus spp. diagnostic molecules, and drug and vaccine targets. In the long run, such data will contribute for a healthier life in endemic Countries for echinococcosis, particularly the poverty-stricken ones.

Methods

In silico secretome prediciton

E. granulosus and *E. multilocularis* secretomes were predicted using the workflow pictured in figure 2, and briefly described here. TMHMM 2.0 [47,48] was used to predict TM domains: proteins with more than one TM domain were discarded. Phobius [49,50] was used to differentiate between a TM domain and a signal peptide in proteins predicted as having a single TM domain by TMHMM. SignalP 4.1 [35] was used to predict proteins secreted by the classical pathway using the eukaryote option and default settings for further options. SecretomeP 2.0 [25] was used for prediction of non-classicaly secreted proteins, considering a NN-score ≥ 0.9 to minimize false-positives. TargetP 1.1 [51] was used for mitochondrial protein predictions. ScanProsite [52] was used to look for the presence of the endoplasmic reticulum retention motif PS00014, and PredGPI [53] was used to predict GPI membrane achored proteins. Proteins predicted as mitochondrial, or as bearing PS00014 motif or a GPI anchor were discarded.

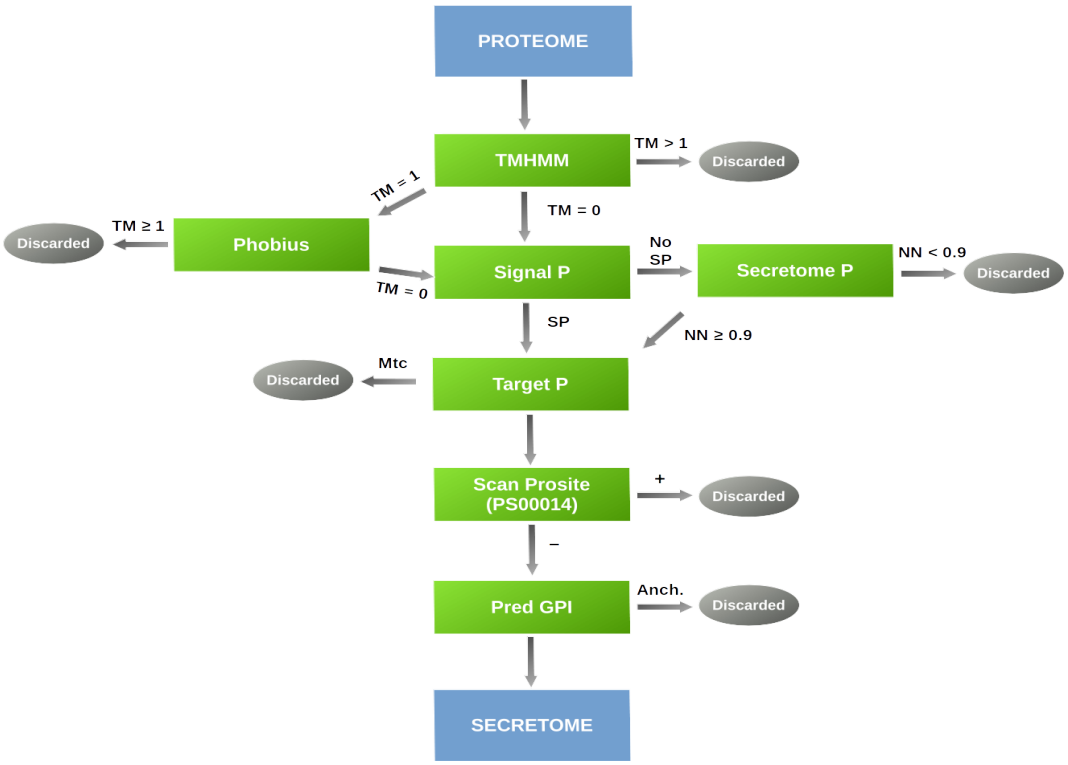

Fig. 2 Standard workflow used for initial *E. granulosus* and *E. multilocularis* secretome predictions.

Ortholog searches

E. granulosus and *E.* multilocularis ortholog proteins were defined based on the RBH method [54,55], using the BLAST+ tool with the following parameters for ortholog searches: e-value of $1x10^{-06}$, query coverage of 50%, bit score equal or greater than 50, max target seqs and max hsps options both equal to 1. The *E.* granulosus predicted secretome was BLASTed against the *E. multilocularis* predicted proteome, and the *E. multilocularis* predicted proteome was BLASTed against the *E.* granulosus predicted secretome. The same approach was used to compare the *E. multilocularis* secretome and the *E. granulosus* proteome. A Python script was used to look for ortholog pairs present both in the secretome against proteome search and in the proteome against secretome search. Ortholog pairs present only in one search were not included in the final result.

Customized workflow for *in silico* secretome prediction

An analysis by WoLF PSORT [32] was added as a new step in the standard workflow depicted in Figure 2, to analyse proteins not predicted as secreted but that

have a secreted ortholog in the other species. For these proteins, the ones predicted by WoLF PSORT as located in cytoplasm, plasma membrane or in the extracellular space were classified as 'secreted', thus, included in the final secretome.

Transcripts assembly and annotation refinement

FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC (http://multiqc.info/) were used to check the quality of RNA-seq reads, which had high quality scores. CAP3 [56] was used to assemble the EST reads. Trinity [57] was used to assemble RNA-seq reads. Transcript assembling was carried out by both CAP3 and Trinity using default parameters. The Program to Assemble Spliced Alignments (PASA) was used to refine the previous *ab initio* genome annotations of *E. granulosus* and *E. multilocularis* published by Tsai et al. (2013).

Functional annotation and antigen prediction

GO term enrichment analysis was performed locally by Blast2GO [58] by Fisher's Exact Test with multiple test correction of FDR (FDR < 0.05) using the entire database of Blast2GO. REVIGO [59] was used to summarize the enriched GO-terms list and to remove redundant GO terms. Pandas library [60] from Python was used to check common GO terms between *E. granulosus* and *E. multilocularis* and the exclusive GO terms for each species.

The antigenicity potential of *E. granulosus* and *E. multilocularis* predicted secreted proteins was independently evaluated using the Kolaskar-Tongaonkar method [61] with a threshold of 1.0, and BepiPred-2.0 standalone software [62] with default options. In both cases, only antigenic segments of at least 6 amino acids were considered. For each protein, the AAR value was calculated as the ratio between the sequence length and the number of antigenic regions predicted by the Kolaskar-Tongaonkar method or by the BepiPred-2.0 software. For each species (*E. granulosus* and *E. multilocularis*), only proteins with antigenicity predictions by both the Kolaskar-Tongaonkar method and by the BepiPred-2.0 were considered for AAR calculation. AAR values and associated statistics were calculated using the Pandas library from Python.

Data access

The predicted proteomes fasta files deducted from the published genome sequences of *E. granulosus* and *E. multilocularis* Tsai et al. (2013) were downloaded from the Welcome Trust Sanger Institute database in March, 2016, and are available at ftp://ftp.sanger.ac.uk/pub/pathogens/Echinococcus/granulosus/genome/ and ftp://ftp.sanger.ac.uk/pub/pathogens/Echinococcus/multilocularis/genome/ , under accession numbers PRJEB121 and PRJEB122, respectively.

All *E. granulosus* and *E. multilocularis* EST libraries available in the National Center for Biotechnology Information (NCBI) EST database (https://www.ncbi.nlm.nih.gov/nucest) were downloaded in July, 2017, comprising 28834 *E. granulosus* EST libraries and 1169 *E. multilocularis* EST libraries.

RNA-seq libraries, generated from the protoscoleces of hydatid cysts, were downloaded in July, 2017 from https://trace.ncbi.nlm.nih.gov/Traces/sra/ using the NCBI's SRA Toolkit for Ubuntu Linux 64 bit. Sequence run accession codes are SRR1508667 and SRR1508668, for *E. granulosus* RNA-seq libraries, and SRR1508669 and SRR1508670 for *E. multilocularis* RNA-seq libraries.

Acknowledgments

This study was supported by Conselho Nacional de Desenvolvimento Científico e Tecnologico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil. T.M.F.F.G was a recipient of CNPq M.Sc. Fellowship. G.P.P. and L.W.P.A are recipients of CAPES Ph.D. and post-doctoral fellowships, respectively.

References

1. Mcfadden B, Heitzman-powell L. Elucidation of the CHO Super-Ome (CHO-SO) by Proteoinformatics. J Proteome Res. 2015;8:1699–712.

2. Shu L, Arneson D, Yang X. Bioinformatics Principles for Deciphering Cardiovascular Diseases. Ref Modul Biomed Sci. Elsevier; 2017.

3. Kass RE, Caffo BS, Davidian M, Meng X-L, Yu B, Reid N, et al. Ten Simple Rules for Effective Statistical Practice. PLOS Comput Biol. 2016;12:e1003858.

4. Dumancas G, Adrianto I, Bello G, Dozmorov M. Current Developments in Machine Learning Techniques in Biological Data Mining. Bioinform Biol Insights. Libertas Academica; 2017;11:1177932216687545.

5. Min S, Lee B, Yoon S. Deep Learning in Bioinformatics. Brief Bioinform. 2016;1–19.

6. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. Brief Bioinform. 2006;7:86–112.

7. Cilingir G, Broschat SL. Automated training for algorithms that learn from genomic data. Biomed Res Int. Hindawi Publishing Corporation; 2015;2015:234236.

8. Leung MKK, Delong A, Alipanahi B, Frey BJ. Machine learning in genomic medicine: A review of computational problems and data sets [Internet]. Proc. IEEE. 2016. p. 176–97.

9. Gomez S, Adalid-Peralta L, Palafox-Fonseca H, Cantu-Robles VA, Soberón X, Sciutto E, et al. Genome analysis of Excretory/Secretory proteins in Taenia solium reveals their Abundance of Antigenic Regions (AAR). Sci Rep. 2015;5:9683.

10. Inal JM, Kosgodage U, Azam S, Stratton D, Antwi-Baffour S, Lange S. Blood/plasma secretome and microvesicles. Biochim Biophys Acta - Proteins Proteomics. 2013;1834:2317–25.

11. Makridakis M, Vlahou A. Secretome proteomics for discovery of cancer biomarkers. J Proteomics. Elsevier B.V.; 2010;73:2291–305.

12. Ditgen D, Anandarajah EM, Meissner KA, Brattig N, Wrenger C, Liebau E. Harnessing the Helminth Secretome for Therapeutic Immunomodulators. Biomed Res Int. 2014;2014.

13. Robinson MW, Dalton JP, O'Brien BA, Donnelly S. Fasciola hepatica: The therapeutic potential of a worm secretome. Int. J. Parasitol. 2013. p. 283–91.

14. Broová A, Jankovská I, Bejcek V, Nechybová S, Peřinková P, Horáková B, et al. Echinococcus spp.: Tapeworms that Pose a Danger to Both Animals and Humans - A Review. Sci Agric Bohem. De Gruyter Open; 2017;48:193–201.

15. Monteiro KM, De Carvalho MO, Zaha A, Ferreira HB. Proteomic analysis of the Echinococcus granulosus metacestode during infection of its intermediate host. Proteomics. 2010;10:1985–99.

16. Pan W, Hao WT, Shen YJ, Li XY, Wang YJ, Sun FF, et al. The excretory-secretory products of Echinococcus granulosus protoscoleces directly regulate the differentiation of B10, B17 and Th17 cells. Parasites and Vectors. BioMed Central; 2017;10:348.

17. Santos GB do., Monteiro KM, da Silva ED, Battistella ME, Ferreira HB, Zaha A. Excretory/secretory products in the Echinococcus granulosus metacestode: is the intermediate host complacent with infection caused by the larval form of the parasite? Int J Parasitol. 2016;46:843–56.

18. Monteiro KM, Lorenzatto KR, de Lima JC, dos Santos GB, Förster S, Paludo GP, et al. Comparative proteomics of hydatid fluids from two Echinococcus multilocularis isolates. J Proteomics. 2017;162:40–51.

19. Lindoso RS, Sandim V, Collino F, Carvalho AB, Dias J, da Costa MR, et al. Proteomics of cell-cell interactions in health and disease. Proteomics. 2016;16:328–44.

20. Gahoi S, Gautam B. Genome-wide analysis of Excretory/Secretory proteins in root-knot nematode, Meloidogyne incognita provides potential targets for parasite control. Comput Biol Chem. 2017;67:225–33.

21. Cuesta-Astroz Y, Oliveira FS de, Nahum LA, Oliveira G. Helminth secretomes reflect different lifestyles and parasitized hosts. Int J Parasitol. 2017;

22. Punta M, Ofran Y. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. Lewitter F, editor. PLoS Comput Biol. CRC Press; 2008;4:e1000160.

23. Pellegrin C, Morin E, Martin FM, Veneault-Fourrey C. Comparative analysis of secretomes from ectomycorrhizal fungi with an emphasis on small-secreted proteins. Front Microbiol. Frontiers Media SA; 2015;6:1278.

24. Conraths FJ, Deplazes P. Echinococcus multilocularis: Epidemiology, surveillance and state-of-the-art diagnostics from a veterinary public health perspective. Vet Parasitol. 2015;213:149–61.

25. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S. Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel. 2004;17:349–56.

26. Lonsdale A, Davis MJ, Doblin MS, Bacic A. Better Than Nothing? Limitations of the Prediction Tool SecretomeP in the Search for Leaderless Secretory Proteins (LSPs) in Plants. Front Plant Sci. Frontiers Media SA; 2016;7:1451.

27. Vendelova E, Camargo de Lima J, Lorenzatto KR, Monteiro KM, Mueller T, Veepaschit J, et al. Proteomic Analysis of Excretory-Secretory Products of Mesocestoides corti Metacestodes Reveals Potential Suppressors of Dendritic Cell Functions. PLoS Negl Trop Dis. Public Library of Science; 2016;10:e0005061.

28. Sotillo J, Pearson M, Potriquet J, Becker L, Pickering D, Mulvenna J, et al. Extracellular vesicles secreted by Schistosoma mansoni contain protein vaccine candidates. Int J Parasitol. 2016;46:1–5.

29. Sperschneider J, Williams AH, Hane JK, Singh KB, Taylor JM. Evaluation of Secretion Prediction Highlights Differing Approaches Needed for Oomycete and Fungal Effectors. Front Plant Sci. Frontiers Media SA; 2015;6:1–14.

30. Horton P, Nakai K. Better prediction of protein cellular localization sites with the k nearest neighbors classifier. Proc Int Conf Intell Syst Mol Biol. 1997;5:147–52.

31. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. Nature Publishing Group; 2007;2:953–71.

32. Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res. Oxford University Press; 2007;35:W585-7.

33. Caccia D, Dugo M, Callari M, Bongarzone I. Bioinformatics tools for secretome analysis. Biochim Biophys Acta - Proteins Proteomics. 2013;1834:2442–53.

34. Siles-Lucas M, Sánchez-Ovejero C, González-Sánchez M, González E, Falcón-Pérez JM, Boufana B, et al. Isolation and characterization of exosomes derived from fertile sheep hydatid cysts. Vet Parasitol. 2017;236:22–33.

35. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. Nature Research; 2011;8:785–6.

36. Ayalew S, Confer AW, Hartson SD, Canaan PJ, Payton M, Couger B. Proteomic and bioinformatic analyses of putative Mannheimia haemolytica secretome by liquid chromatography and tandem mass spectrometry. Vet Microbiol. 2017;203:73–80.

37. Siracusano A, Delunardo F, Teggi A, Ortona E. Host-parasite relationship in cystic echinococcosis: An evolving story [Internet]. Clin. Dev. Immunol. Hindawi Limited; 2012. p. 639362.

38. Hewitson JP, Grainger JR, Maizels RM. Helminth immunoregulation: The role of parasite secreted proteins in modulating host immunity. Mol Biochem Parasitol. Elsevier B.V.; 2009;167:1–11.

39. Ahn C-S, Kim J-G, Han X, Kang I, Kong Y. Comparison of Echinococcus multilocularis and Echinococcus granulosus hydatid fluid proteome provides molecular strategies for specialized host-parasite interactions. Oncotarget. Impact Journals, LLC; 2017;8:97009–24.

40. Abd Ellah MR. Involvement of free radicals in parasitic infestations. J Appl Anim Res. Taylor & Francis Group ; 2013;41:69–76.

41. Amri M, Touil-Boukoffa C. A protective effect of the laminated layer on Echinococcus granulosus survival dependent on upregulation of host arginase. Acta Trop. Elsevier; 2015;149:186–94.

42. Soria-Guerra RE, Nieto-Gomez R, Govea-Alonso DO, Rosales-Mendoza S. An overview of bioinformatics tools for epitope prediction: Implications on vaccine development. J Biomed Inform. Elsevier Inc.; 2015;53:405–14.

43. Lightowlers MW, Rickard MD. Excretory-secretory products of helminth parasites: effects on host immune responses. Parasitology. 1988;96 Suppl:S123-66.

44. Ranasinghe SL, Duke M, Harvie M, McManus DP. Kunitz-type protease inhibitor as a vaccine candidate against schistosomiasis mansoni. Int J Infect Dis. Elsevier; 2018;66:26–32.

45. Abdolahi Khabisi S, Sarkari B, Moshfe A, Jalali S. Production of Monoclonal Antibody Against Excretory-Secretory Antigen of Fasciola hepatica and Evaluation of Its Efficacy in the Diagnosis of Fascioliasis. Monoclon Antib Immunodiagn Immunother. 2017;36:8–14.

46. Wang S, Wei W, Cai X. Genome-wide analysis of excretory/secretory proteins in Echinococcus multilocularis: insights into functional characteristics of the tapeworm secretome. Parasit Vectors. BioMed Central; 2015;8:666.

47. Möller S, Croning MD, Apweiler R. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics. Oxford University Press; 2001;17:646–53.

48. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J Mol Biol. 2001;305:567–80.

49. Käll L, Krogh A, Sonnhammer EL. L. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 2004;338:1027–36.

50. Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. Nucleic Acids Res. Oxford University Press; 2007;35:429–32.

51. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. 2000;300:1005–16.

52. de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, et al. ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 2006;34:362–5.

53. Pierleoni A, Martelli PL, Casadio R. PredGPI: a GPI-anchor predictor. BMC Bioinformatics. 2008;9:392.

54. Salichos L, Rokas A. Evaluating ortholog prediction algorithms in a yeast model clade. PLoS One. Public Library of Science; 2011;6:e18755.

55. Kuzniar A, van Ham RCHJ, Pongor S, Leunissen JAM. The quest for orthologs: finding the corresponding gene across genomes. Trends Genet. 2008;24:539–51.

56. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. Cold Spring Harbor Laboratory Press; 1999;9:868–77.

57. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. Nature Publishing Group; 2011;29:644–52.

58. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. Oxford University Press; 2008;36:3420–35.

59. Supek F, Bošnjak M, Škunca N, Šmuc T. Revigo summarizes and visualizes long lists of gene ontology terms. Gibas C, editor. PLoS One. Public Library of Science; 2011;6:e21800.

60. McKinney W. Data Structures for Statistical Computing in Python. In: van der Walt S, Millman J, editors. Proc 9th Python Sci Conf. 2010. p. 51–6.

61. Kolaskar AS, Tongaonkar PC. A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett. 1990;276:172–4.

62. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: Improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. 2017;45:W24–9.

63. Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL, et al. The genomes of four tapeworm species reveal adaptations to parasitism. Nature. Nature Research; 2013;496:57–63.

4. DISCUSSÃO

Dentre os seres vivos, basicamente todos podem ser parasitados. A evolução da vida, desde seu nível molecular, tem no parasitismo uma característica impulsionadora das mudanças que levam às adaptações dos seres vivos frente aos obstáculos encontrados na batalha travada entre parasitos e hospedeiros pela sobrevivência. A interação entre parasito e hospedeiro é um sistema complexo e dinâmico. Parasitos bem adaptados são capazes de infectar o hospedeiro, usufruindo dos recursos energéticos e conseguindo ludibriar as defesas do hospedeiro para reproduzir-se e perpetuar-se.

Os parasitos *E. granulosus* e *E. multilocularis*, pertencentes à classe Cestoda, são capazes de, na sua fase larval, infectar seus hospedeiros intermediários por anos antes que qualquer sinal clínico seja notado. Esses parasitos são causadores, respectivamente, da equinococose cística e da equinococose alveolar, doenças responsáveis por muitos prejuízos à saúde humana e animal. Entender o que ocorre na interação parasito-hospedeiro e como ela ocorre, é ponto-chave para o desenvolvimento de estratégias de prevenção e combate às equinococoses e outras doenças parasitárias.

Algumas das estratégias usadas pelos parasitos *E. granulosus* e *E. multilocularis* para evadirem-se da resposta do sistema imune do hospedeiro intermediário e causarem infecções crônicas são baseadas em produtos de ES. No contexto de infecção por metacestódeos de *Echinococcus* spp., produtos de ES podem regular negativamente as funções dos macrófagos, induzem apoptose de células dendríticas (NONO et al., 2012; VIRGINIO et al., 2012). Além disso, os produtos de ES podem estar envolvidos em outras funções importantes para a sobrevivência do parasito como a captação de metabólitos e facilitam a migração, penetração e estabelecimento no hospedeiro (SOTILLO et al., 2017).

A Identificação e caracterização de produtos de ES de parasitos é um trabalho complexo e laborioso, necessitando de estratégias de ação diversificadas e conjuntas para este intuito. A disponibilidade de material biológico é muitas vezes limitada e a identificação de produtos por técnicas proteômicas baseadas em espectrômetria de massas exige emprego de muito tempo e tem um alto custo pecuniário (BREHM; SPILIOTIS, 2008; LINDOSO et al., 2016). Assim sendo, o uso

de ferramentas de bioinformática para predição *in silico* de produtos de ES baseadas em sequências genômicas tornou-se uma estratégia fundamental e de escolha para estudos iniciais. Uma vez feitas estas predições iniciais, pode-se direcionar os estudos experimentais de bancada, mais demorados e de maior custo, para proteínas com maior potencial como alvos para o desenvolvimento de drogas anti-parasitárias ou como antígenos para o desenvolvimento de testes imunodiagnósticos e vacinas.

Para que possa ser obtido um secretoma predito confiável, é essencial o uso de ferramentas bioinformáticas com algoritmos capazes de identificar os produtos de ES de maneira eficiente e adaptáveis a diferentes organismos. Os algoritmos de aprendizagem de máquina têm sido amplamente utilizados, pois, são capazes de, mediante treinamento, identificarem as características comuns entre os dados analisados, classificando-os de forma autômata (LEUNG et al., 2016).

O *workflow* empregado para a predição de proteínas secretadas segue, de forma geral, um mesmo padrão, normalmente utilizando-se os mesmos *softwares* para os mais diversos organismos. A predição dos produtos é comumente feita para apenas uma espécie ou, quando aplicada para mais espécies em um mesmo estudo, sem uma comparação par a par das proteínas ortólogas preditas como secretadas para espécies próximas. Comparações mais gerais, considerando apenas números/frações globais de proteínas secretadas por uma ou outra via de secreção não são capazes de evidenciar eventuais inconsistências pontuais nas predições de secreção ou de não secreção.

Considerando as possíveis limitações dos *softwares* e a falta de estudos comparativos entre espécies relacionadas, não se possui exata ciência do quão representativo do secretoma real pode ser o secretoma predito. Estas limitações dizem respeito principalmente a organismos não tradicionalmente usados como organismos-modelo, para os quais os *softwares* não são customizados e dados experimentais confirmatórios nem sempre estão disponíveis.

No manuscrito que constitui o corpo principal desta dissertação, foi apresentado um estudo comparativo *in silico* dos produtos proteicos de ES preditos de *E. granulosus* e *E. multilocularis*, duas espécies próximas filogeneticamente. O estudo comparativo de espécies geneticamente similares possibilitou a validação

cruzada das predições e a identificação dos vieses que podem ser encontrados usando-se do *workflow* comumente usado na predição de proteínas secretadas. Foram evidenciadas falhas nas predições de secreção para proteínas de espécies para as quais o algoritmo de ML não está treinado, decorrentes da falta de atualização dos *softwares* com dados genômicos atualizados e de limitações de desenvolvimento de *software* devido à falta de conhecimento completo sobre a via não-clássica de secreção de proteínas.

Tais vieses não poderiam ser evidenciados apenas predizendo-se o secretoma de cada espécie individualmente e comparando-se os números gerais de proteínas secretadas por cada via nas duas espécies, os quais foram muito similares. Com a identificação das limitações nos *workflows* normalmente utilizados na predição de proteinas secretadas, foi possível propor diferentes estratégias de ação para aumentar a confiabilidade do secretoma predito como representação mais próxima possível do secretoma real.

Foi possível verificar que, dentre os *softwares* usados no *workflow*-padrão para predição, as maiores inconsistências deram-se no passo de predição de secreção pela via não-clássica pelo *software* SecretomeP. Isto ocorre porque ainda não há um entendimento completo dos mecanismos de secreção por via não-clássica, dificultando o desenvolvimento de um algoritmo preciso para a identificação de proteínas secretadas por esta via. Além disso, o *software* SecretomeP foi até agora treinado apenas com dados de bactérias gram-negativas, bactérias gram-positivas e mamíferos, não permitindo resultados mais acurados para dados de outros grupos de organismos, como vegetais, fungos e helmintos (SPERSCHNEIDER et al., 2015; LONSDALE et al., 2016; SOTILLO et al., 2016).

Como estratégia para verificação de predições de produtos de ES, acrescentou-se ao *workflow*-padrão o *software* WoLF PSORT, que prediz a localização subcelular das proteínas. Os resultados obtidos depois deste acréscimo mostraram que muitas das proteínas não preditas como secretadas e com um par ortólogo secretado, apresentavam localizações indicativas de proteínas secretadas.

Apesar de se ter identificado o problema de predição de proteínas pelo *software* SecretomeP, verificou-se que, no alinhamento dos pares de ortólogos secretados/não-secretados, muitos deles apresentavam discrepâncias entre suas

sequências. Pelo menos parte destas discrepâncias entre sequências ortólogas poderiam ser decorrentes de problemas de sequenciamento ou anotação, fatores que também poderiam resultar em predições de secreção ou não diferentes para proteínas ortólogas.

Para se testar tal hipótese, fez-se o refinamento da anotação dos genomas de *E. granulosus* e *E. multilocularis* com o uso de dados transcritômicos de ESTs e RNA-seq destas espécies, aliados aos dados de anotação *ab initio* do sequenciamento genômico original. Combinar dados de diferentes fontes é uma forma de se obter uma anotação mais confiável, visto que a predição gênica computacional automática é apenas um primeiro passo de uma anotação, estando, por exemplo, sujeita a erros por possíveis limitações de *software* ao confrontar-se com dados novos (SINGH et al., 2017).

Apesar da qualidade das sequências ter melhorado com o refinamento da anotação, o seu impacto no secretoma predito foi maior em relação ao número total de proteínas do secretoma de cada espécie, sendo menos relevante para a diferenciação da secreção ou não dos pares de ortólogos secretados/não-secretados. Os números de proteínas preditas como secretadas para *E. granulosus* e *E. multilocularis* mantiveram praticamente as mesmas proporções dos secretomas anteriores em relação aos proteomas totais preditos para estas espécies (~6% para ambas). As proporções de proteínas não-secretadas em uma espécie com um ortólogo secretado na outra também se mantiveram aproximadamente as mesmas após o refinamento da anotação. Também, após a reanotação das sequências genômicas, a maioria das proteínas foi eliminada ao nível do SecretomeP (~70%), reforçando a ideia de que o problema maior nas predições de ES está relacionado ao algoritmo de predição de secreção de via não-clássica.

Para que se possa realizar predições mais confiáveis, além de softwares adequados, a qualidade das sequências é importante, afinal, sequências que não estejam sob constantes melhorias baseadas em dados computacionais e funcionais, dificultam a predição, por exemplo, funcional de proteínas. Este é um grande desafio enfrentado nas ciências “ômicas” de helmintos, pois, em média, metade dos genes codificadores de proteínas não estão caracterizados no que diz respeito a funcionalidade (PALEVICH et al., 2017). Isso pode ser notado pelo grande número

de proteínas dos secretomas preditos de *E. granulosus* e *E. multilocularis* que não obtiveram nenhum termo GO associado (~56% e ~53% do secretoma predito, respectivamente).

Apesar da maior parte das proteínas dos secretomas preditos não ter sido associada a nenhum termo GO, os resultados da análise de enriquecimento funcional comuns a *E. granulosus* e *E. multilocularis* sugerem várias funções essenciais para a invasão e sobrevivência do parasito no hospedeiro intermediário. Por exemplo, funções correspondentes a proteínas relacionadas a atividades hormonais e resposta a estímulos (GO:0005179 e GO:0050896, respectivamente) podem estar ligadas a vias de sinalização de insulina, essenciais para desenvolvimento larval do parasito, (HEMER et al., 2014). Isso é corroborado pelo fato de que metacestódeos de *E. granulosus* e *E. multilocularis* têm como um de seus órgãos-alvo no hospedeiro intermediário o fígado, local onde há presença de altas concentrações de insulina. Outros termos, como GO:0005520 e GO:0016942, são relacionados ao fator de crescimento semelhante à insulina (IGF), fator responsável por proliferação celular e inibição de apoptose (SALGADO et al., 2010). Tais funções também podem ser consideradas como essenciais para a sobrevivência do metacestódeo no hospedeiro intermediário. Os termos GO:0005576, GO:0005615, também enriquecidos tanto para *E. granulosus* como para *E. multilocularis*, estão relacionados ao espaço extracelular e evidenciam a correta predição das proteínas como produtos de ES. *E. multilocularis* apresenta ainda o enriquecimento do termo GO:0070062, correspondente a exossomos extracelulares, portanto, indicativo de secreção por via não-clássica. Além disso, muitos outros termos relacionados a proteases e peptidases, proteínas já notórias por sua ação para penetração e sobrevivência do parasito no hospedeiro intermediário (HEWITSON et al., 2009), estão entre as mais frequentes em ambas as espécies. Além disso, proteases e peptidases podem hidrolisar anticorpos, protegendo assim o parasito da resposta imune do hospedeiro (HEIZER et al., 2013).

As principais características que diferenciam *E. granulosus* e *E. multilocularis* no que diz respeito à infecção do hospedeiro intermediário são os modos de proliferação e invasão dos tecidos do hospedeiro intermediário. O metacestódeo de

E. granulosus é um cisto unilocular, cujo crescimento ocorre de forma expansiva, por alargamento concêntrico (THOMPSON, 2017). De acordo com essas características, foi observado, dentre os produtos de ES preditos para esta espécie, o enriquecimento dos termos GO:0050794 e GO:0051673, correspondentes à regulação de processos celulares e ao rompimento de membrana em outro organismo, respectivamente.

O metacestódeo de *E. multilocularis*, por sua vez, é uma estrutura multivesicular, consistindo de numerosas pequenas vesículas envoltas em denso tecido conjuntivo que são infiltrativas, pois possuem um comportamento metastático semelhante a um tumor cancerígeno (YANG et al., 2012; BREHM; KOZIOL, 2017). Estas características de crescimento e invasividade de metacestódeos de *E. multilocularis* podem ser associadas a termos GO para produtos de ES preditos como enriquecidos apenas nesta espécie, como GO:0008283, referente à proliferação celular; GO:0007154 e GO:0007267, referentes à comunicação celular e à sinalização entre células; GO:0016050, referente à organização em vesículas; e GO:0040018, referente à regulação positiva de crescimento de organismo multicelular.

Com o uso de ferramentas *in silico* para predição de antigenicidade de proteínas, foi possível verificar que os secretomas preditos de *E. granulosus* e *E. multilocularis* possuem muitas proteínas com alta densidade de epitopos. Existem diversos métodos para a predição de epitopos. Os mais clássicos, baseados em características como hidrofobicidade e acessibilidade de superfície, possuem acurácia de aproximadamente 57%. Há também métodos baseados em *machine learning*, com acurácia estimada em, aproximadamente, 66% (SORIA-GUERRA et al., 2015).

Nenhum desses métodos, porém, leva em consideração o tamanho das sequências para normalização da densidade de epitopos. Ao se usar da estratégia da contagem de AARs, que considera o tamanho das sequências, é possível se obter resultados mais confiáveis de predição de antigenicidade (GOMEZ et al., 2015). Esta estratégia de predição de antigenicidade é de grande auxilio para a seleção de alvos vacinais e diagnósticos, pois a abordagem convencional de

clonagem e expressão heteróloga de proteínas recombinantes para caracterização imunológica é muito trabalhosa para aplicação prática em escala genômica.

Cruzando-se os dados de predição de enriquecimento funcional e de antigenicidade, foi possível notar que proteínas assinaladas como peptidases e proteases possuem baixos valores de AAR, significando alta densidade de epitopos, a exemplo da proteina de *E. granulosus* EgrG_000575400, que possui pelos métodos Kolaskar-Tongaonkar e BepiPred 2.0, os valores de AAR de 16.17 e 19.4, respectivamente, e da proteina de *E. multilocularis* EmuJ_000158900, que possui pelos métodos Kolaskar-Tongaonkar e BepiPred 2.0, os valores de AAR de 16 e 22.4, respectivamente.

As discrepâncias de valores de AAR para uma mesma proteína quando são calculados pelo método de Kolaskar-Tongaonkar ou pelo software BepiPred 2.0 são devidas às diferentes formas utilizadas por cada algoritmo para predizer regiões antigênicas. O método de Kolaskar-Tongaonkar realiza uma abordagem clássica de predição, baseando as predições de antigenicidade nas propriedades físico-químicas dos aminoácidos (KOLASKAR; TONGAONKAR, 1990). Já o *software* BepiPred 2.0 combina propriedades de hidropaticidade dos aminoácidos e o uso de algoritmos de ML de modelos ocultos de Markov (do inglês, *Hidden Markov Models*) para predizer epitopos de células B (JESPERSEN et al., 2017). O uso de métodos clássicos, como o de Kolaskar-Tongaonkar, aliados a métodos modernos como algoritmos de ML, como o do *software* BepiPred 2.0, ajuda a se ter mais confiabilidade nas predições de antigenicidade (SORIA-GUERRA et al., 2015).

Os valores de AAR encontrados para os secretomas preditos de *E. granulosus* e *E. multilocularis* são muito semelhantes aos encontrados em diferentes secretomas de outros helmintos, como *Taenia solium*, *Fasciola gigantica*, *Mesocestoides corti* e *Caenorhabditis brenneri*, valores de AAR menores que os de proteínas transmembrana (GOMEZ et al., 2015), as quais são comumente consideradas com bom potencial para alvo de vacinas (WANG et al., 2015a; MEHLA; RAMANA, 2016), demonstrando o potencial do uso de proteinas de ES como possíveis alvos para o desenvolvimento de novas estratégias de diagnóstico, prevenção e tratamento de helmintíases, como as equinococoses.

A caracterização funcional das proteínas de ES de *E. granulosus* e *E. multilocularis* também auxilia, no direcionamento da escolha de proteínas que possam auxiliar no desenvolvimento de estratégias de combate e prevenção às equinococoses. Afinal, graças às suas propriedades imunomoduladoras, as proteínas secretadas por helmintos há muito são estudadas como potenciais moléculas para utilização no desenvolvimento de vacinas para a prevenção de infecções por estes parasitos (LIGHTOWLERS; RICKARD, 1988). As proteínas de ES têm também potencial como marcadores imunodiagnósticos para helmintíases e podem também ser alvos para o desenvolvimento de novas drogas anti-parasitárias (WANG et al., 2015b; CUESTA-ASTROZ et al., 2017). Além disso, o uso destes produtos de ES pode ser estendido para o tratamento de doenças imunomediadas humanas (HARNETT, 2014; NASCIMENTO SANTOS et al., 2016), devido às suas propriedades anti-inflamatórias, regulando a ação de macrófagos, por exemplo. Evidências apontam o sucesso do uso destas proteínas de ES ao conferirem proteção, ao menos parcial, contra parasitos como *Ancylostoma duodenale* (BETHONY, 2005), *Schistosoma mansoni* (SOTILLO et al., 2016; RANASINGHE et al., 2018) e *Heligmosomoides polygyrus* (COAKLEY et al., 2017), por exemplo.

Poucas abordagens imunoterapêuticas têm sido aplicadas pra tentar combater as equinococoses, sendo mais comum o uso de drogas anti-helmínticas (POURSEIF et al., 2017), que apresentam inconvenientes como a presença de resíduos no leite ou na carne dos animais de produção tratados e a possibilidade do desenvolvimento de resistência dos parasitos às drogas. Por isso, as descrições dos repertórios de proteínas secretadas por *E. granulosus* e *E. multilocularis* constituem um grande passo rumo a novas formas de prevenção e combate às equinococoses e outras helmintíases. Neste contexto, a identificação *in silico* das proteínas de ES é uma forma prática e eficiente para o direcionamento destes estudos.

5. PERSPECTIVAS

- Estabelecimento de critérios otimizados para o uso do WoLF PSORT no *workflow* de predição de produtos de ES de *Echinococcus* spp. e de outros cestódeos;

- Desenvolvimento de um novo algoritmo de *machine learning* para a predição de proteínas secretadas por via não-clássica de parasitos cestódeos.

- Predições mais confiáveis de produtos de ES preditos *in silico* com o estabelecimento do *workflow* melhorado para *Echinococcus* spp. e outros cestódeos.

- Definição de produtos de ES de cestódeos preditos *in silico* como potenciais antígenos diagnósticos ou vacinais ou como alvos para o desenvolvimento de novas drogas anti-helmínticas.

REFERÊNCIAS BIBLIOGRÁFICAS

AHN, Chun-Seob et al. Comparison of Echinococcus multilocularis and Echinococcus granulosus hydatid fluid proteome provides molecular strategies for specialized host-parasite interactions. **Oncotarget**, v. 8, n. 57, p. 97009–97024, 2017.

AKBULUT, Sami et al. Associating liver partition and portal vein ligation for staged hepatectomy for extensive alveolar echinococcosis: First case report in the literature. **World journal of gastrointestinal surgery**, v. 10, n. 1, p. 1–5, 2018.

ASSIS, L. M. et al. B-cell epitopes of antigenic proteins in Leishmania infantum: An in silico analysis. **Parasite Immunology**, v. 36, n. 7, p. 313–323, 2014.

AYALEW, Sahlu et al. Proteomic and bioinformatic analyses of putative Mannheimia haemolytica secretome by liquid chromatography and tandem mass spectrometry. **Veterinary Microbiology**, v. 203, p. 73–80, 2017.

AZIZ, Ammar et al. Proteomic characterisation of Echinococcus granulosus hydatid cyst fluid from sheep, cattle and humans. **Journal of Proteomics**, v. 74, n. 9, p. 1560–1572, 2011.

BETHONY, Jeffrey. Antibodies against a secreted protein from hookworm larvae reduce the intensity of hookworm infection in humans and vaccinated laboratory animals. **The FASEB Journal**, 2005.

BREHM, K.; KOZIOL, U. Echinococcus–Host Interactions at Cellular and Molecular Levels. **Advances in Parasitology**, 2016.

BREHM, K.; KOZIOL, U. **Echinococcus-Host Interactions at Cellular and Molecular LevelsAdvances in parasitology**Academic Press, , 2017.

BREHM, Klaus; SPILIOTIS, Markus. Recent advances in the in vitro cultivation and genetic manipulation of Echinococcus multilocularis metacestodes and germinal cells. **Experimental Parasitology**, v. 119, n. 4, p. 506–515, 2008.

BRINDLEY, Paul J. et al. **Helminth genomics: The implications for human health** (Matty Knight, Ed.)**PLoS Neglected Tropical Diseases**Public Library of Science, , 2009.

BROOVÁ, A. et al. Echinococcus spp.: Tapeworms that Pose a Danger to Both Animals and Humans - A Review. **Scientia Agriculturae Bohemica**, v. 48, n. 4, p. 193–201, 2017.

BRUNETTI, Enrico; KERN, Peter; VUITTON, Dominique Angèle. Expert consensus for the diagnosis and treatment of cystic and alveolar echinococcosis in humans. **Acta Tropica**, v. 114, n. 1, p. 1–16, 2010.

BRUNO, Luca et al. Lessons from Reverse Vaccinology for viral vaccine design. **Current Opinion in Virology**, v. 11, p. 89–97, 2015.

CACCIA, Dario et al. Bioinformatics tools for secretome analysis. **Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics**, v. 1834, n. 11, p. 2442–2453, 2013.

CARDONA, Guillermo A.; CARMENA, David. A review of the global prevalence, molecular epidemiology and economics of cystic echinococcosis in production animals. **Veterinary Parasitology**, v. 192, n. 1–3, p. 10–32, 2013.

CHEN, Xin; GUO, Lingqiong; FAN, Zhaocheng. Learning Position Weight Matrices from Sequence and Expression Data. **Comput Syst Bioinform Conf.**, v. 6, p. 249–260, 2007.

CHOI, Jaeyoung et al. Fungal Secretome Database: Integrated platform for annotation of fungal secretomes. **BMC Genomics**, v. 11, n. 1, p. 105, 2010.

CHOO, Khar Heng; TAN, Tin Wee; RANGANATHAN, Shoba. A comprehensive assessment of N-terminal signal peptides prediction methods. **BMC bioinformatics**, v. 10 Suppl 1, n. Suppl 15, p. S2, 2009.

CILINGIR, Gokcen; BROSCHAT, Shira L. Automated training for algorithms that learn from genomic data. **BioMed Research International**, v. 2015, p. 234236, 2015.

COAKLEY, Gillian et al. Extracellular Vesicles from a Helminth Parasite Suppress Macrophage Activation and Constitute an Effective Vaccine for Protective Immunity. **Cell Reports**, v. 19, n. 8, p. 1545–1557, 2017.

CUESTA-ASTROZ, Yesid et al. Helminth secretomes reflect different lifestyles and parasitized hosts. **International Journal for Parasitology**, 2017.

DITGEN, Dana et al. Harnessing the Helminth Secretome for Therapeutic Immunomodulators. **BioMed Research International**, v. 2014, 2014.

DU, Chengsong et al. Hepatectomy for patients with alveolar echinococcosis: Long-term follow-up observations of 144 cases. **International Journal of Surgery**, v. 35, p. 147–152, 2016.

GAHOI, Shachi; GAUTAM, Budhayash. Genome-wide analysis of Excretory/Secretory proteins in root-knot nematode, Meloidogyne incognita provides potential targets for parasite control. **Computational Biology and Chemistry**, v. 67, p. 225–233, 2017.

GARG, Gagan; RANGANATHAN, Shoba. Helminth secretome database (HSD): a collection of helminth excretory/secretory proteins predicted from expressed sequence tags (ESTs). **BMC genomics**, v. 13 Suppl 7, n. Suppl 7, p. S8, 2012.

GOMEZ, Sandra et al. Genome analysis of Excretory/Secretory proteins in Taenia solium reveals their Abundance of Antigenic Regions (AAR). **Scientific reports**, v. 5, p. 9683, 2015.

GOTTSTEIN, B. et al. Immunoblotting for the serodiagnosis of alveolar echinococcosis in alive and dead Eurasian beavers (Castor fiber). **Veterinary Parasitology**, v. 205, n. 1, p. 113–118, 2014.

GOTTSTEIN, Bruno et al. **Threat of alveolar echinococcosis to public health - a challenge for EuropeTrends in Parasitology**, 2015.

HARNETT, William. **Secretory products of helminth parasites as immunomodulatorsMolecular and Biochemical Parasitology**Elsevier, , 2014.

HEIZER, Esley et al. Transcriptome analyses reveal protein and domain families that delineate stage-related development in the economically important parasitic nematodes, Ostertagia ostertagi and Cooperia oncophora. **BMC Genomics**, v. 14, n. 1, p. 118, 2013.

HEMER, Sarah et al. Host insulin stimulates Echinococcus multilocularis insulin signalling pathways and larval development. **BMC Biology**, v. 12, p. 5, 2014.

HEWITSON, James P.; GRAINGER, John R.; MAIZELS, Rick M. Helminth immunoregulation: The role of parasite secreted proteins in modulating host immunity. **Molecular and Biochemical Parasitology**, v. 167, n. 1, p. 1–11, 2009.

INAL, Jameel M. et al. Blood/plasma secretome and microvesicles. **Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics**, v. 1834, n. 11, p. 2317–2325, 2013.

JESPERSEN, Martin Closter et al. BepiPred-2.0: Improving sequence-based B-cell epitope prediction using conformational epitopes. **Nucleic Acids Research**, v. 45, n. W1, p. W24–W29, 2017.

KERN, Peter. Clinical features and treatment of alveolar echinococcosis. **Current Opinion in Infectious Diseases**, v. 23, n. 5, p. 505–512, 2010.

KINKAR, Liina et al. New mitogenome and nuclear evidence on the phylogeny and taxonomy of the highly zoonotic tapeworm Echinococcus granulosus sensu stricto. **Infection, Genetics and Evolution**, v. 52, p. 52–58, 2017.

KNAPP, Jenny et al. Taxonomy, phylogeny and molecular epidemiology of Echinococcus multilocularis: From fundamental knowledge to health ecology. **Veterinary Parasitology**, v. 213, n. 3–4, p. 85–91, 2015.

KOLASKAR, A. S.; TONGAONKAR, P. C. A semi-empirical method for prediction of antigenic determinants on protein antigens. **FEBS letters**, v. 276, n. 1–2, p. 172–4, 1990.

LAI, Jhih-Siang et al. Computational Comparative Study of Tuberculosis Proteomes Using a Model Learned from Signal Peptide Structures. **PLoS ONE**, v. 7, n. 4, p. e35018, 2012.

LEUNG, Michael K. K. et al. **Machine learning in genomic medicine: A review of computational problems and data setsProceedings of the IEEE**, 2016.

LIGHTOWLERS, M. W.; RICKARD, M. D. Excretory-secretory products of helminth parasites: effects on host immune responses. **Parasitology**, v. 96 Suppl, p. S123-66, 1988.

LINDOSO, Rafael S. et al. Proteomics of cell-cell interactions in health and disease. **Proteomics**, v. 16, n. 2, p. 328–344, 2016.

LONSDALE, Andrew et al. Better Than Nothing? Limitations of the Prediction Tool SecretomeP in the Search for Leaderless Secretory Proteins (LSPs) in Plants. **Frontiers in plant science**, v. 7, p. 1451, 2016.

MAKRIDAKIS, Manousos; VLAHOU, Antonia. Secretome proteomics for discovery of cancer biomarkers. **Journal of Proteomics**, v. 73, n. 12, p. 2291–2305, 2010.

MCMANUS, Donald P. Echinococcosis with Particular Reference to Southeast Asia. In: **Advances in Parasitology**. [s.l: s.n.]. v. 72p. 267–303.

MEHLA, Kusum; RAMANA, Jayashree. Identification of epitope-based peptide vaccine candidates against enterotoxigenic Escherichia coli: a comparative genomics and immunoinformatics approach. **Mol. BioSyst.**, v. 12, n. 3, p. 890–901, 2016.

MIN, Seonwoo; LEE, Byunghan; YOON, Sungroh. Deep Learning in Bioinformatics. **Briefings in Bioinformatics**, n. March, p. 1–19, 2016.

MONTEIRO, Karina M. et al. Comparative proteomics of hydatid fluids from two Echinococcus multilocularis isolates. **Journal of Proteomics**, v. 162, p. 40–51, 2017.

NASCIMENTO SANTOS, Leonardo et al. **Recombinant proteins of helminths with immunoregulatory properties and their possible therapeutic useActa Tropica**, 2016.

NAZ, Anam et al. Identification of putative vaccine candidates against Helicobacter pylori exploiting exoproteome and secretome: A reverse vaccinology based approach. **Infection, Genetics and Evolution**, v. 32, p. 280–291, 2015.

NONO, Justin Komguep et al. Excretory/secretory-products of echinococcus multilocularis larvae induce apoptosis and tolerogenic properties in dendritic cells in vitro. **PLoS Neglected Tropical Diseases**, v. 6, n. 2, p. e1516, 2012.

OLSON, P. D. et al. Cestode genomics - progress and prospects for advancing basic and applied aspects of flatworm biology. **Parasite Immunology**, v. 34, n. 2–3, p. 130–150, 2012.

OLSON, Peter D.; TKACH, Vasyl V. Advances and Trends in the Molecular Systematics of the Parasitic Platyhelminthes. **Advances in Parasitology**, v. 60, p. 165–243, 2005.

OROBITG, Miquel et al. High Performance computing improvements on bioinformatics consistency-based multiple sequence alignment tools. **Parallel Computing**, v. 42, p. 18–34, 2015.

PALEVICH, Nikola et al. Tackling Hypotheticals in Helminth Genomes. **Trends in Parasitology**, v. 34, n. 3, p. 179–183, 2017.

PAN, Wei et al. Transcriptome Profiles of the Protoscoleces of Echinococcus granulosus Reveal that Excretory-Secretory Products Are Essential to Metabolic Adaptation. **PLoS Neglected Tropical Diseases**, v. 8, n. 12, p. 1–15, 2014.

PEARSON, Mark S. et al. In vitro and in silico analysis of signal peptides from the human blood fluke, Schistosoma mansoni. **FEMS Immunology and Medical Microbiology**, v. 45, n. 2, p. 201–211, 2005.

PENSEL, Patricia E. et al. Experimental cystic echinococcosis therapy: In vitro and in vivo combined 5-fluorouracil/albendazole treatment. **Veterinary Parasitology**, v. 245, p. 62–70, 2017.

PETRONE, L. et al. A T-cell diagnostic test for cystic echinococcosis based on Antigen B peptides. **Parasite Immunology**, v. 39, n. 12, p. e12499, 2017.

POURSEIF, Mohammad Mostafa et al. Current status and future prospective of vaccine development against Echinococcus granulosus. **Biologicals**, v. 51, p. 1–11, 2017.

RANASINGHE, Shiwanthi L. et al. Kunitz-type protease inhibitor as a vaccine candidate against schistosomiasis mansoni. **International Journal of Infectious Diseases**, v. 66, p. 26–32, 2018.

RAPPUOLI, Rino. **Reverse vaccinology, a genome-based approach to vaccine developmentVaccine**. [s.l: s.n.].

ROMIG, T. et al. Chapter Five – Ecology and Life Cycle Patterns of Echinococcus Species. **Advances in Parasitology**, v. 95, p. 213–314, 2017.

ROMIG, T.; EBI, D.; WASSERMANN, M. Taxonomy and molecular epidemiology of Echinococcus granulosus sensu lato. **Veterinary Parasitology**, v. 213, n. 3–4, p. 76–84, 2015.

ROSSI, Patrizia et al. The first meeting of the European Register of Cystic Echinococcosis (ERCE). **Parasites and Vectors**, v. 9, n. 1, p. 243, 2016.

SALGADO, António J. Braga Osório Gomes et al. Adipose tissue derived stem cells secretome: soluble factors and their roles in regenerative medicine. **Current stem cell research & therapy**, v. 5, n. 2, p. 103–110, 2010.

SASAKI, Mizuki; SAKO, Yasuhito. The putative serine protease inhibitor (serpin) genes encoded on Echinococcus multilocularis genome and their expressions in metacestodal stage. **Veterinary Parasitology**, v. 233, p. 20–24, 2017.

SCHICHT, Sabine et al. The predicted secretome and transmembranome of the poultry red mite Dermanyssus gallinae. **Parasites & vectors**, v. 6, n. 1, p. 259, 2013.

SERRUTO, Davide; RAPPUOLI, Rino. Post-genomic vaccine development. **FEBS Letters**, v. 580, n. 12, p. 2985–2992, 2006.

SILVA-ÁLVAREZ, Valeria et al. Echinococcus granulosus antigen B: A Hydrophobic Ligand Binding Protein at the host-parasite interface. **Prostaglandins Leukotrienes and Essential Fatty Acids**, v. 93, p. 17–23, 2015.

SILVA-ÁLVAREZ, Valeria et al. Echinococcus granulosus Antigen B binds to monocytes and macrophages modulating cell response to inflammation. **Parasites & Vectors**, v. 9, n. 1, p. 69, 2016.

SINGH, Reema et al. Improved annotation with de novo transcriptome assembly in four social amoeba species. **BMC Genomics**, v. 18, n. 1, p. 120, 2017.

SORIA-GUERRA, Ruth E. et al. An overview of bioinformatics tools for epitope prediction: Implications on vaccine development. **Journal of Biomedical Informatics**, v. 53, p. 405–414, 2015.

SOTILLO, Javier et al. Extracellular vesicles secreted by Schistosoma mansoni contain protein vaccine candidates. **International Journal for Parasitology**, v. 46, n. 1, p. 1–5, 2016.

SOTILLO, Javier et al. Exploiting Helminth-Host Interactomes through Big Data. **Trends in Parasitology**, v. 33, n. 11, p. 875–888, 2017.

SOUTO, M. G.; SANCHEZ THEVENET, P.; BASUALDO FARJAT, J. Evaluation of the presence of Echinococcus granulosus sensu lato in the environment and in hosts in a region endemic for hydatidosis in the province of Chubut (Argentina). **Veterinary Parasitology: Regional Studies and Reports**, v. 6, p. 42–46, 2016.

SPERSCHNEIDER, Jana et al. Evaluation of Secretion Prediction Highlights Differing Approaches Needed for Oomycete and Fungal Effectors. **Frontiers in Plant Science**, v. 6, n. December, p. 1–14, 2015.

TESSELE, Bianca; BRUM, Juliana S.; BARROS, Claudio S. L. Lesões parasitárias encontradas em bovinos abatidos para consumo humano. **Pesquisa Veterinária Brasileira**, v. 33, n. 7, p. 873–889, 2013.

THOMPSON, R. C. A. **Biology and Systematics of EchinococcusAdvances in parasitology**, 2017.

TORGERSON, Paul R. et al. The Global Burden of Alveolar Echinococcosis. **PLoS Neglected Tropical Diseases**, v. 4, n. 6, p. e722, 2010.

TORGERSON, Paul R. et al. World Health Organization Estimates of the Global and Regional Disease Burden of 11 Foodborne Parasitic Diseases, 2010: A Data Synthesis. **PLoS Medicine**, v. 12, n. 12, p. e1001920, 2015.

TSAI, Isheng J. et al. The genomes of four tapeworm species reveal adaptations to parasitism. **Nature**, v. 496, n. 7443, p. 57–63, 2013.

VENDELOVA, Emilia et al. Proteomic Analysis of Excretory-Secretory Products of Mesocestoides corti Metacestodes Reveals Potential Suppressors of Dendritic Cell Functions. **PLoS Neglected Tropical Diseases**, v. 10, n. 10, p. e0005061, 2016.

VIRGINIO, Veridiana G. et al. Excretory/secretory products from in vitro-cultured Echinococcus granulosus protoscoleces. **Molecular and Biochemical Parasitology**, v. 183, n. 1, p. 15–22, 2012.

WAESCHENBACH, Andrea; WEBSTER, B. L.; LITTLEWOOD, D. T. J. Adding resolution to ordinal level relationships of tapeworms (Platyhelminthes: Cestoda) with large fragments of mtDNA. **Molecular Phylogenetics and Evolution**, v. 63, n. 3, p. 834–847, 2012.

WANG, Shuai; WEI, Wei; CAI, Xuepeng. Genome-wide analysis of excretory/secretory proteins in Echinococcus multilocularis: insights into functional characteristics of the tapeworm secretome. **Parasites & Vectors**, v. 8, n. 1, p. 666, 2015. a.

WANG, Ying et al. Proteomic analysis of the excretory/secretory products and antigenic proteins of Echinococcus granulosus adult worms from infected dogs. **BMC Veterinary Research**, v. 11, n. 1, p. 1–7, 2015. b.

WANG, Yujian et al. Modulation of goat monocyte function by HCcyst-2, a secreted cystatin from Haemonchus contortus. **Oncotarget**, p. 1–13, 2017.

YANG, YuRong; ELLIS, Magda K.; MCMANUS, Donald P. **Immunogenetics of human echinococcosisTrends in Parasitology**, 2012.

YANG, Ziheng; RANNALA, Bruce. Molecular phylogenetics: principles and practice. **Nature Reviews Genetics**, v. 13, n. 5, p. 303–314, 2012.

ZHENG, Yadong. Strategies of Echinococcus species responses to immune attacks: Implications for therapeutic tool development. **International Immunopharmacology**, v. 17, n. 3, p. 495–501, 2013.

Curriculum Vitae Resumido

Gomes, Tiago M. F. F; Gomes, T. M. F. F.

1. Dados Pessoais

Nome:

Tiago Minuzzi Freire da Fontoura Gomes

Local e data de nascimento:

Foz do Iguaçu, Paraná, Brasil, 13/11/1987

Endereço Profissional:

Universidade Federal do Rio Grande do Sul, Centro de Biotecnologia
Avenida Bento Gonçalves, 9500, Prédio 43421, sala 210
91501-970, Porto Alegre, RS, Brasileira
Telefone: (051) 33087769

E-mail:

minuzzitiago@hotmail.com
tiago.minuzzi87@gmail.com
tiago.minuzzi@ufrgs.br

2. Formação

2016 – Atual

Mestrado em Biologia Celular e Molecular
Universidade Federal do Rio Grande do Sul, UFRGS, Porto Alegre, RS, Brasil
Orientador: Henrique Bunselmeyer Ferreira
Bolsista: Conselho Nacional de Desenvolvimento Científico e Tecnológico

2015 – 2016

Graduação incompleta em Ciências Biológicas
Universidade Federal de Santa Maria, UFSM, Santa Maria, RS, Brasil

2007 – 2011

Graduação em Medicina Veterinária

Universidade Federal de Santa Maria, UFSM, Santa Maria, RS, Brasil

3. Estágios

2015 - 2016

Estágio Curricular

Enquadramento Funcional: Estagiário – Iniciação Científica

Carga horária: 20h

Laboratório de Biologia Molecular e Sequenciamento Genético, LabDros (UFSM)

Orientador: Dr. Elgion Lucio da Silva Loreto

4. Artigos Completos Publicados

OLIVEIRA, DANIEL S. ; GOMES, TIAGO M.F.F. ; LORETO, ELGION L.S. The rearranged mitochondrial genome of Leptopilina boulardi (Hymenoptera: Figitidae), a parasitoid wasp of Drosophila. Genetics and Molecular Biology (online version), v. 39, p. 611-615, 2016.

Apêndices

Apêndice 1. Orthologs found between *E. granulosus* e *E. multilocularis* predicted secretomes by the RBH method

Arquivo: Supplementary_table_1.xlsx

Acesso via midia digital.

Apêndice 2. *E. granulosus* and *E. multilocularis* secreted/non-secreted orthopairs and WoLF PSORT predictions

Arquivo: Supplementary_table_2.xlsx

Acesso via midia digital.

Apêndice 3. *E. granulosus* and *E. multilocularis* revised predicted secretomes secretion pathways

Arquivo: Supplementary_table_3.xlsx

Acesso via midia digital.

Apêndice 4. Functional enrichment prediction results for *E. granulosus* and *E. multilocularis* revised predicted secretomes

Arquivo: Supplementary_table_4.xlsx

Acesso via midia digital.

Apêndice 5. Antigenicity predictions for *E. granulosus* and *E. multilocularis* revised predicted secretomes

Arquivo: Supplementary_table_5.xlsx

Acesso via midia digital.