JENIFFER NOVAES

Análise da expressão diferencial entre merozoítos e esporozoítos de *Eimeria tenella* empregando a técnica de LongSAGE

Tese apresentada ao Instituto de Ciências Biomédicas da Universidade de São Paulo, para obtenção do Título de Doutor em Ciências.

Área de Concentração: Biologia da Relação Patógeno-Hospedeiro.

São Paulo

2009

JENIFFER NOVAES

Análise da expressão diferencial entre merozoítos e esporozoítos de *Eimeria tenella* empregando a técnica de LongSAGE

Tese apresentada ao Instituto de Ciências Biomédicas da Universidade de São Paulo, para obtenção do Título de Doutor em Ciências.

Área de Concentração: Biologia da Relação Patógeno-Hospedeiro.

Orientadora:

Profa. Dra. Alda Maria Backx N. Madeira

São Paulo

2009

DADOS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
Serviço de Biblioteca e Informação Biomédica do
Instituto de Ciências Biomédicas da Universidade de São Paulo

Novaes, Jeniffer.
Análise da expressão diferencial entre merozoítos e esporozoítos de *Eimeria tenella* empregando a técnica de LongSAGE / Jeniffer Novaes. -- São Paulo, 2009.

Orientador: Alda Maria Backx Noronha Madeira.

Tese (Doutorado) – Universidade de São Paulo. Instituto de Ciências Biomédicas. Departamento de Parasitologia. Área de concentração: Biologia da Relação Patógeno-Hospedeiro. Linha de pesquisa: Biologia molecular de *Eimeria spp*.

Versão do título para o inglês: Differential expression analysis between merozoites and sporozoites of *Eimeria tenella* using LongSAGE.

Descritores: 1. Coccidiose 2. *Eimeria tenella* 3. Expressão Gênica 4. Merozoítos 5. Esporozoítos 6. Etiquetas de seqüências expressas I. Madeira, Alda Maria Backx Noronha II. Universidade de São Paulo. Instituto de Ciências Biomédicas. Programa de Pós-Graduação em Biologia da Relação Patógeno-Hospedeiro III. Título.

ICB/SBIB0207/2009

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Candidato(a): Jeniffer Novaes.

Título da Tese: Análise da expressão diferencial entre merozoítos e esporozoítos de *Eimeria tenella* empregando a técnica de LongSAGE.

Orientador(a): Alda Maria Backx Noronha Madeira.

A Comissão Julgadora dos trabalhos de Defesa da Tese de Doutorado, em sessão pública realizada a ...../...../....., considerou

( ) Aprovado(a) **( ) Reprovado(a)**

Examinador(a): Assinatura: .....
Nome: .....
Instituição: .....

Examinador(a): Assinatura: .....
Nome: .....
Instituição: .....

Examinador(a): Assinatura: .....
Nome: .....
Instituição: .....

Examinador(a): Assinatura: .....
Nome: .....
Instituição: .....

Presidente: Assinatura: .....
Nome: .....
Instituição: .....

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Cidade Universitária "Armando de Salles Oliveira"
Av. Prof. Lineu Prestes, 2415 – CEP. 05508-000 São Paulo, SP – Brasil
Telefone :(55) (011) 3091.7733 – telefax : (55) (011) 3091.7438
e-mail: cep@icb.usp.br

CERTIFICADO

Certificamos que o protocolo registrado sob nº **124** nas fls. **10** do livro **2** para uso de animais em experimentação, sob a responsabilidade da Profa. Dra. Alda Maria Backx Noronha Madeira, Coordenadora da Linha de Pesquisa "***Biologia Molecular e celular de parasitas de galinha doméstica do gênero eimeria***" do qual participou(aram) o(s) aluno(s): **Jeniffer Novaes Gonçalves Dias, Alessandra Popov dos Santos Manha** e o pesquisador **Arthur Gruber**, está de acordo com os Princípios Éticos de Experimentação Animal adotado pelo Colégio Brasileiro de Experimentação Animal (COBEA) e foi aprovado pela *COMISSÃO DE ÉTICA EM EXPERIMENTAÇÃO ANIMAL* (CEEA) em **17.02.2005.**

São Paulo, 18 de fevereiro de 2005.

Profa. Dra. Marília C. Leite Seelaender
Coordenadora da CEEA

Prof. Dr. Francisco Carlos Pereira
Secretário da CEEA

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Cidade Universitária "Armando de Salles Oliveira"
Av. Prof. Lineu Prestes, 2415 - cep. 05508-000 São Paulo, SP - Brasil
Telefone :(55) (011) 3091.7733 - telefax : (55) (011) 3091 7438
e-mail: cep@icb.usp.br

Of.CEEA.017.08
WTL/mcgn

São Paulo, 11 de novembro de 2008.

REF.: Protocolo nº 124/04.

"Biologia Molecular e celular de parasitas de galinha doméstica do gênero Eimeria"

Prezada Professora,

Informo que a sua licença para uso de animais em experimentação, constante no protocolo em epígrafe, **foi prorrogada até 17.02.2011.**

Reitero que havendo alteração de metodologia e inserção de novos alunos ao projeto de pesquisa vinculado à referida licença a CEEA/ICB deverá ser informada.

Cordialmente,

Prof. Dr. WOTHAN TAVARES DE LIMA
Coordenador da Comissão de Ética em
Experimentação Animal - ICB /USP

Ilma. Sra.
Profa Dra. ALDA MARIA BACKX NORONHA MADEIRA
Departamento de Parasitologia
Instituto de Ciências Biomédicas -USP

Dedico esta tese a todas as pessoas especiais que
já fizeram e ainda fazem parte da minha vida,
tornando cada um dos meus dias mais felizes.
De forma especial à minha família:
Meus avôs e minha avó (in memoriam),
minha querida Mãe e Pai,
minhas irmãs especiais,
e meus amigos (incluindo os chefes)
por sempre se orgulharem
e acreditarem em mim!

AGRADECIMENTOS

À Profa. Dra. Alda Backx Noronha Madeira, primeiramente, por me aceitar como sua aluna, apesar do meu jeito peculiar. Pela formação científica e pelo apoio durante todos estes anos, incluindo a sua disponibilidade (e muitas vezes excesso de atenção), em resolver questões técnicas, científicas e até mesmo financeiras dentro da minha formação e desenvolvimento do projeto. Por ser aberta ao diálogo e compreensiva, o que fez com que nossa relação fosse melhorando cada vez mais com o passar dos anos, tornando-se parte da minha vida e uma pessoa de alta estima e amizade; além de fornecer uma boa estrutura de trabalho e um ambiente agradável de trabalho, assim como pelas diversas oportunidades;

Ao Prof. Dr. Arthur Gruber, por aceitar me orientar (evento que aconteceu por osmose, mas de forma muito natural). Apesar de não oficial, para mim você sempre será meu orientador também (apesar de às vezes você parecer não gostar muito disso). Sempre disposto a desenvolver novas ferramentas e facilidades para o progresso dos projetos do laboratório, sempre cobrando duro, mas ao mesmo tempo acreditando em minha capacidade e estendendo à mão para encontrar soluções. Fazendo parte da minha formação científica e também no desenvolvimento de um bom gosto por cerveja, entre outras coisas, além de aturar a minha extrema ansiedade. Agradeço também à suas broncas, que me ajudaram muito a amadurecer, e ainda estão ajudando. Por todos estes anos e oportunidades oferecidas, por fazer parte da minha vida e tornar-se um amigo, e também por comprar uma grande briga, que mudou radicalmente muitas vidas, mas acredito que nos fez crescer;

À minha primeira orientadora, Isabela (Isa) que me aceitou muito jovem (iniciação científica) e permitiu a minha familiaridade com o laboratório, ensinando muito bem a metodologia científica, e também à Judite que participou deste processo de crescimento;

Aos outros orientadores que fizeram parte deste trabalho, o Prof. Dr. Alan Durham, por estar sempre aberto a discussões e por sua disponibilidade e atenção

dedicada a mim e a este projeto. Ao Prof. Dr. Carlos Alberto de Bragança Pereira, pelos ensinamentos e desenvolvimento das análises estatísticas;

À equipe de bioinformática do Coccilab, os ex-membros Leonardo Varuzza (sempre engraçado e disponível, uma fonte de boas idéias), ao Thiago Sobreira pelo desenvolvimento de softwares simples e geniais, à Milene Ferro, André Kashiwabara e Ricardo Yamamoto pelo desenvolvimento e aprimoramento do Egene, entre outros favores. Ao Samuel pelo desenvolvimento dos programas de SAGE (SAGE Suite e SAGE Analysis), plataformas super práticas, arrojadas e essenciais para o desenvolvimento desta Tese. Aos novos membros, André, Marcelo e Thibério, por povoarem novamente o laboratório e alegrarem o nosso cotidiano, especialmente ao Eimério pelo desenvolvimento do módulo de análise de KOG;

Aos alunos companheiros de bancada, Itácio (novo membro), Alê, Laure e à Ursa. Alê e Ursa, obrigado por me agüentarem nos momentos de estresse e ruins e por me acompanharem em inúmeros momentos felizes e engraçados. E até mesmo muitas ciladas. Vocês são parte da minha família, afinal, convivemos mais no laboratório do que em casa.;

Aos alunos companheiros de departamento, à Alessandra Fratus (comédia), Catarina e todas as meninas do laboratório, à Bianca (companheira desde o Mackenzie), Andrezinho e seus companheiros de laboratório, e a todos os outros alunos que tornaram o convívio cotidiano agradável e divertido;

Agradeço a todos os Professores do Departamento, pelo convívio agradável e pelos ensinamentos, tanto no dia a dia quanto nas disciplinas, seminários e outras atividades. Especialmente ao Prof. Gerhard, pela atenção e também por compartilhar o *Realplex* com o departamento todo e, também a todos os funcionários do Departamento;

Aos antigos alunos do laboratório: Sandra, Jane e Joana, pelos primeiros ensinamentos em *Eimeria.* E às antigas técnicas que participaram na execução da parte experimental: Lívia e Lú, assim como os diversos estudantes que conviveram no laboratório durante estes anos que estive lá;

A todos meus professores desde a formação básica até a pós-graduação, que incentivaram meu desenvolvimento, pensamento crítico e interesse pelos estudos, especialmente às professoras Alaíde e Patty, que me incentivaram a ser Bióloga;

À minha família, a qual eu dedico esta tese, que sempre foram compreensivos e forneceram todo apoio necessário para meu desenvolvimento, nunca questionando às minhas escolhas e sempre acreditando em mim. Pelo apóio psicológico, de estrutura e financeiro. Apesar de pequena, nossa família é muito feliz e unida pelo amor. Obrigado Pai, Mãe e queridas irmãs (Samantha e Vanessa), tão presentes e essenciais em minha vida. E também aos meus queridos sobrinhos (Victor e Bruno) e ao meu cunhado André, e meu tio Gui, por fazerem parte da minha vida, além dos meus gatos amados (Michel, Lancelot e Roger (*in memoriam)*), queridos conpanheiros peludos;

Aos meus amigos especiais que me acompanham e me apóiam até hoje: Dani e Thé (de longos anos), Tita, Ferna, Lú, Panda (e suas queridas amigas) que fizeram e fazem parte da minha vida, em inúmeros momentos especiais desde o Mackenzie, e especialmente à Marix, por me entender completamente (tarefa extremamente difícil), e a Ursa por confiar e compartilhar muitos momentos de sua vida comigo. Ao Fê da Ursa e ao Fábio da Marix, grandes companheiros;

Ao Instituto de Ciências Biomédicas da Universidade de São Paulo, pela formação durante a pós-graduação;

Ao Conselho Nacional de Pesquisa (CNPq) pelo suporte financeiro e pelas bolsas de estudos concedidas (desde a iniciação científica) e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo apoio financeiro ao laboratório e ao presente projeto;

"A sabedoria não nos é dada; temos de descobri-la sozinhos depois de fazer uma jornada que ninguém pode fazer por nós ou nos poupar de fazer."

Marcel Proust

RESUMO

Novaes J. Análise da expressão diferencial entre merozoítos e esporozoítos de *Eimeria tenella* empregando a técnica de LongSAGE [Tese]. São Paulo: Instituto de Ciências Biomédicas da Universidade de São Paulo; 2009.

A coccidiose aviária é causada por protozoários do gênero *Eimeria* e é responsável por grandes prejuízos à indústria avícola mundial. O nosso grupo gerou e anotou mais de 15.000 seqüências expressas do tipo ORESTES das três principais espécies de *Eimeria* que acometem a galinha doméstica. Dentre estas, a *Eimeria tenella* é a mais estudada devido a sua alta freqüência e virulência no campo, sendo empregada como modelo de estudo para coccidiose. Com o objetivo de se estudar o perfil de expressão quantitativo de dois estágios invasivos destes parasitas, foram construídas bibliotecas de LongSAGE a partir de merozoítos de segunda geração e esporozoítos. As seqüências de DNA foram submetidas a um processo seriado de múltiplos passos utilizando a plataforma Egene e, apenas seqüências apresentando alta qualidade foram consideradas para as etapas posteriores. A extração e contagem das *tags* foram realizadas utilizando o SAGE Analysis, um pacote desenvolvido localmente. Para análise estatística foi empregado o Kemp, uma implementação de um teste freqüentista. No total, obtivemos mais de 35.000 *tags*, que corresponde a 9.516 *tags* únicas. Deste conjunto, 270 *tags* foram classificadas como diferencialmente expressas e utilizadas como sementes para a reconstrução dos cDNAs correspondentes utilizando cerca de 48.000 leituras de ORESTES/ESTs e o programa GenSeed. Um total de 199 seqüências de cDNA foram reconstruídas com sucesso e submetidas a um *pipeline* de anotação automática empregando o Egene. O processo de anotação consistiu na identificação das seqüências codificadoras, e a caracterização dos produtos correspondentes utilizando buscas de similaridade, buscas de motivos protéicos, detecção de ortólogos e mapeamento de ontologia gênica. Estes resultados mostraram que, apesar de invasivos, estes estágios apresentam genes de expressão estágio-específica e um perfil transcricional distinto. Para merozoítos, grande parte dos produtos protéicos diferencialmente expressos estavam relacionados à tradução, modificação e manutenção da conformação das proteínas e processos de ligação. Enquanto que em esporozoítos, o perfil transcricional obtido foi distinto, com poucos resultados positivos de BLAST. Alguns dos produtos protéicos identificados foram as histonas,

proteínas associadas a transporte e atividade catalítica. Uma validação experimental preliminar utilizando um pequeno conjunto de genes foi realizada e uma boa correlação entre estas técnicas foi obtida ao se comparar os dados quantitativos de expressão gênica. A análise comparativa entre os dados diferencialmente expressos obtidos por LongSAGE e *Northern Digital* de ESTs/ORESTES de *E. tenella* mostrou que há coerência entre os dados obtidos, já que muitos genes diferencialmente expressos foram identificados por ambas as técnicas.

Palavras-chave: Coccidiose. *Eimeria tenella*. Expressão gênica. Merozoítos. Esporozoítos. Etiquetas de seqüências expressas.

ABSTRACT

Novaes J. Differential expression analysis between merozoites and sporozoites of *Eimeria tenella* using LongSAGE [Thesis]. São Paulo: Instituto de Ciências Biomédicas da Universidade de São Paulo; 2009.

Coccidiosis of the domestic fowl is caused by protozoan parasites of the genus *Eimeria*. The disease is world widely distributed and is responsible for relevant economic losses to the poultry industry. Our group generated and annotated more than 15.000 expressed sequences (ORESTES) for each one of the three most important *Eimeria* species. *Eimeria tenella* is the most studied species due to its high prevalence and virulence, and because is more amenable to laboratory manipulation. For these reasons, *E. tenella* became the model species on coccidiosis research. Aiming at obtaining a quantitative expression profile of two invasive stages of these parasites, we constructed LongSAGE libraries of sporozoites and second-generation merozoites. DNA reads were submitted to a multistep processing pipeline using EGene platform, and only sequence stretches showing high quality were considered for downstream analyses. Tag extraction and counting were performed using SAGE Analysis, a locally developed package. For the statistical analysis we used Kemp, an implementation of a frequentist test. In total we have obtained more than 35,000 tags, which correspond to 9,5126 unique tags. From this set, 270 tags were considered differentially expressed and utilized as seeds to reconstruct the corresponding cDNAs using circa 48,000 ORESTES/EST reads and the program GenSeed. A total of 199 cDNAs sequences were successfully reconstructed and submitted to an automated annotation pipeline using EGene. The annotation process consisted on the identification of the coding sequences, and the characterization of the corresponding products using similarity searches, protein motif finding, orthology detection and gene ontology mapping. These results showed that besides the invasive nature of both *stages*, we were able to detect stage-specific expressed genes and a distinct transcriptional profile. In merozoites, most of the differentially expressed genes were associated to translation, protein modification and folding, and binding processes. Conversely, sporozoites showed a very distinct transcription profile, with a low number of BLAST hits. Some of the identified products included histones and proteins associated with transport and catalytic activity. A preliminary experimental validation using real-time PCR was carried out with a small group of

genes and the expression ratio observed between the tested developmental stages has been experimentally confirmed with a good correlation. A comparison analysis of the differential expressed genes obtained by LongSAGE and Northern Digital of ESTs/ORESTES also showed a good agreement, and many differentially expressed genes have been mutually identified by both techniques.

Keywords: Coccidiosis. *Eimeria tenella*. Gene Expression. Merozoites. Sporozoites. Expressed sequence *tags*.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

A - absorbância

Água Milli-Q - água purificada por osmose reversa, deionizada e tratada com luz UV (Sistema RiOs e Milli-Q Synthesis da Millipore)

ATP - adenosina trifosfato

BeT – Best hiT (melhor alinhamento)

BLAST - *Basic Local Alignment Search Tool* (Ferramenta de busca de alinhamento local básica)

BBSRC - *Biotechnology and Biological Sciences Research* (Comitê de Pesquisa em Ciências Biotecnológicas e Biológicas - Reino Unido)

CAP3 - *Contig Assembly Program* (Programa de Mon*tag*em de Contigs)

cDNA - *complementary DNA* (DNA complementar)

COG – *Clusters of Orthologous Groups of proteins* (Grupos de agrupamentos de proteínas ortólogas)

C-terminal - carboxi-terminal

C_t – *Cycle Threshold* (o ponto C_t indica o momento a partir do qual o produto amplificado é quantificado).

DAG - Grafico Acíclico Direto

DD - *Differential Display*

DDBJ - *DNA Data Bank of Japan* (banco de dados de DNA do Japão)

DEPC - dietilpirocarbonato de sódio

DNA - ácido desoxirribonucleico

DNAse - desoxirribonuclease

dNTP -abreviação coletiva para desoxinucleosídios trifosfato (dATP, dCTP, dGTP e dTTP)

DTT ditiotreitol

EDTA - ácido etilenodiaminotetracético

EMBL - *European Molecular Biology Laboratory* (Laboratório Europeu de Biologia Molecular)

EST - E*xpressed Sequence Tags* (etiquetas de seqüências expressas)

FT - *feature table* (tabela de características)

G+C - guanina + timina

GFF - *Generic Feature Format* (Formato Genérico de Características)

GLGI - *Generation of Longer cDNA fragments from serial analysis of gene expression tags for Gene Identification* (Geração de fragmentos maiores de cDNA a partir de *tags* de SAGE para a identificação de genes)

GO - *Gene Ontology* (Ontologia Gênica)

HIV - *Human Imunodeficiency Vírus* (Vírus da Imunodeficiência humana)

HTML - *HyperText Markup Language* (Linguagem de Marcação de Hipertexto)

KOG – euKariotic Orthologous Groups (Grupos de Ortólogos de Eucariotos)

IAH - *Institute for Animal Health*, Compton, Inglaterra (Instituto de Saúde Animal)

ITS1 - *Internal transcribed spacer* 1 (Espaçador Ribossômico Interno 1)

MOPS - acido 3-[N-morfilino] propanosulfônico

MPSS - *Massive Pararell Signature Sequencing* (Sequenciamento Paralelo Maciço de Assinaturas)

NCBI - *National Center for Biotechnology Information* (Centro Nacional de Informação Biotecnológica - EUA)

N-terminal -amino-terminal

ORF - O*pen Reading Frame* (fase aberta de leitura)

ORESTES - ORF ESTs (*Open Reading Frame Expressed Sequence Tags*)

PBS – Solução Salina Tamponada

PCR - *polymerase chain reaction* (reação em cadeia da polimerase)

pH - -log $[H^+]$

PNA – Ácido nucléico peptídico neutro

ppm - partes por milhão

p/v -peso por volume

qPCR – PCR quantitativo, refere-se a tecnologia de PCR em tempo rela (*Real Time PCR)*

RACE - *Rapid Amplification of cDNA Ends* (Amplificação rápida das extremidades do cDNA)

RAP_PCR - RNA *fingerprinting by arbitrary primed* PCR (Impressão digital de RNA utilizando *primers* arbitrário

RAPD - *Random Amplified Polymorphic DNA* (polimorfismo de DNA amplificado randomicamente)

RNA - ácido ribonucléico

RNAm – RNA mensageiro

RNAr – RNA Ribossômico

RNAse – ribonuclease

RNAt - RNA transportador

RT -*reverse transcription* (transcrição reversa)

SAGE - *Serial Analysis of Gene Expression* (Análise Serial de Expressão Gênica)

SCARs – *Sequence Characterized Amplified Regions* (Regiões amplificadas caracterizadas por seqüenciamento)

SDS - dodecilsulfato de sódio

SDS-PAGE -eletroforese em gel de poliacrilamida contendo SDS

SSH - *Supression Subtractive Hybridization* (Hibridização Subtrativa por Supressão)

SNPs -*Single nucleotide polymorphism* (Polimorfismo de nucleotídeo único)

spp. - espécie

TBE -solução tampão Tris-borato-EDTA

TE -solução tampão Tris-EDTA

Tm -temperatura de fusão (de um *primer*)

TRF - *Tandem Repeats Finder* (Localizador de repetições seriadas)

Tris -Tris[hidroxmetil]aminometano

URL - *Uniform Resource Locator* (Localizador Uniforme de Recursos)

USDA - *United States Departament of Agriculture* (Departamento de Agricultura dos Estados Unidos)

UTR -*Untranslated Region* (região não traduzida)

UV - ultravioleta

v/v - volume por volume

WTSI (*Wellcome Trust Sanger Institute*, Cambridge, Inglaterra)

XML - *eXtensible Markup Language* (Linguagem de Marcação Estendida)

LISTA DE SÍMBOLOS

^{o}C - grau Celsius

g - aceleração da gravidade

g -grama

h -hora

Kb -quilobase

L -litro

M -molar

mg - miligrama

mL - mililitro

mm -milímetro

mM -milimolar

N -normal

ng - nanogramas

pb - pares de bases

pmoles -picomoles

rpm - rotações por minuto

U - unidade

V - Volts

Ω - Ohm

µF - microfarad

µg - micrograma

µL – microlitro

µM - micromolar

SUMÁRIO

1 INTRODUÇÃO

1.1 *Eimeria* spp. e o Filo Apicomplexa

Protistas do gênero *Eimeria* (Classe Coccidia, Filo Apicomplexa) foram de acordo com a história, os primeiros protozoários parasitas a serem visualizados. Em 1674, Antony van Leeuwenhoek, considerado um dos criadores do microscópio, descreveu a presença de corpúsculos nos dutos biliares de um coelho, os quais provavelmente tratavam-se de oocistos de *Eimeria stiedae*. Somente 150 anos mais tarde o primeiro membro deste filo foi taxonomicamente classificado (Levine, 1973).

O Filo Apicomplexa é composto por mais de 5.000 espécies de protozoários intracelulares obrigatórios são considerados os parasitas eucarióticos mais bem sucedidos do planeta. Dentro deste grupo, patógenos de importância médica e veterinária são responsáveis por grandes ameaças à saúde mundial causando prejuízos à economia global (Sibley, 2004; Tomley, 2009)

Dentre os parasitas de importância médica, podemos destacar o gênero *Plasmodium*, responsável pela malária, o *Toxoplasma gondii* que pode causar abortos e danos neurológicos congênitos e também os gêneros *Cryptosporidium* e *Cyclospora* que estão associados a infecções oportunísticas em indivíduos imunossuprimidos, particularmente os infectados pelo HIV (Roos, 2005; Striepen *et al.*, 2002). Com relação aos patógenos de importância médico-veterinária, podemos citar os gêneros *Babesia, Neospora*, *Theileria* e *Eimeria*, que são capazes de infectar animais domésticos e de produção, acarretando em grandes prejuízos econômicos (Belli *et al.*, 2005; Hu *et al.*, 2006).

Na figura 1 podemos observar uma árvore filogenética de alguns gêneros do Filo Apicomplexa (Wu, 2008). Os organismos deste filo são caracterizados pela presença de uma combinação única de organelas na extremidade apical das formas “zoítas” (infectantes) destes parasitas, denominada de complexo apical (Levine, 1973; Tomley e Soldati, 2001).

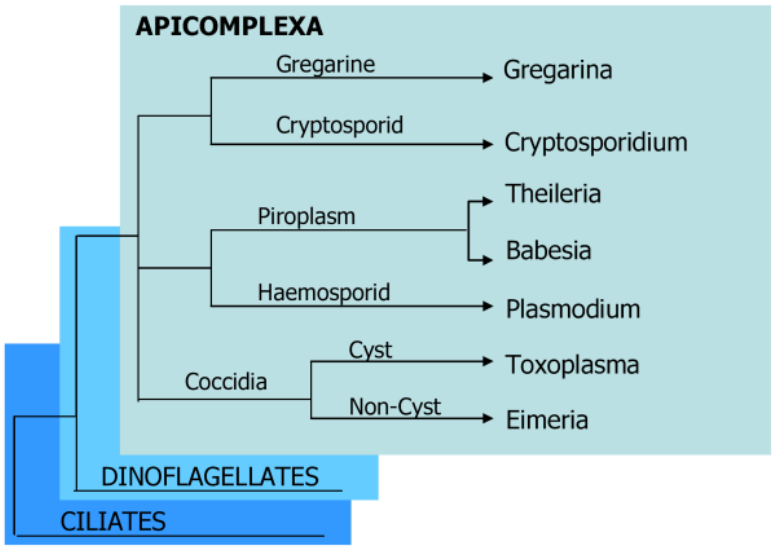


Figura 1 – Árvore filogenética de organismos do Filo Apicomplexa baseada nas seqüências da subunidade menor do RNA ribossômico.
Fonte: Wu, 2008.

O complexo apical é composto por um conóide, anéis polares, microtúbulos subpeliculares, roptrias, micronemas e grânulos densos. (Dubremetz *et al.*, 1998; Santos *et al.*, 2009). Estas estruturas estão associadas à motilidade, adesão e penetração dos parasitas nas células hospedeiras e formação do vacúolo parasitóforo (Bumstead e Tomley, 2000; Morrissette e Sibley, 2002). Os micronemas, roptrias e grânulos densos constituem um grupo de organelas secretoras que liberam seu conteúdo de forma seqüencial durante o processo de invasão. Inicialmente as micronemas secretam suas proteínas que estão envolvidas na motilidade, adesão e reconhecimento da célula hospedeira (Bumstead e Tomley, 2000; Tomley e Soldati, 2001). Em seguida há liberação das proteínas das roptrias, que auxiliam na movimentação do parasita para dentro da célula hospedeira e na formação da membrana do vacúolo parasitóforo (Sibley, 2004). Finalmente, os grânulos densos, que tem a função de remodelar o vacúolo parasitóforo tornando este compartimento metabolicamente ativo, liberam seu conteúdo (Bromley *et al.*, 2003; Dubremetz *et al.*, 1998; Entzeroth *et al.*, 1998). O conóide (presente apenas nos coccídias), os anéis polares, e os microtúbulos subpeliculares, fazem parte do

citoesqueleto celular, responsável por mediar o processo de invasão celular. O conóide é uma estrutura pequena na forma de cone, composta por filamentos espiralados e, acredita-se que esteja envolvida na função mecânica da invasão. Os anéis polares constituem um dos centros de organização dos microtúbulos subpeliculares, sendo ambos importantes para a manutenção da forma, polaridade e tráfico de organelas nestes organismos (Morrissette e Sibley, 2002; Santos *et al.*, 2009).

Dentro deste filo há a classe Coccidia, grupo variado de parasitas cujas espécies podem ter ciclo de vida monoxênico ou heteroxênico, podendo ou não acometer uma grande gama de hospedeiros (Tenter *et al.*, 2002). Dentro desta classe, o gênero *Eimeria* é o mais representativo do Filo Apicomplexa, com mais de 1700 espécies descritas (Levine, 1988), as quais podem ser encontradas em diversos hospedeiros, desde organismos invertebrados como anelídeos e insetos, até vertebrados, incluindo anfíbios, répteis, aves e mamíferos, que representam a grande maioria dos hospedeiros.

Os protozoários do gênero *Eimeria* apresentam um ciclo de vida direto (monoxênico) e grande parte destes organismos infectam as células epiteliais da mucosa intestinal. O ciclo de vida é dividido em esporogonia, que ocorre fora do hospedeiro sendo responsável pela formação do estágio infectante do parasita, a esquizogonia (reprodução assexuada) e a gametogonia (reprodução sexuada) as quais ocorrem dentro da célula de um hospedeiro espécie-específico (Kogut, 1990; Shirley e Harvey, 1996, 2000). Tyzzer (1929) foi o pioneiro na caracterização das espécies deste gênero e estabeleceu a base do estudo em coccídias, trabalhando com espécies de *Eimeria* que acometem as aves (Ball *et al.*, 1989).

1.2 Coccidiose aviária

1.2.1 A doença e importância econômica

Os protozoários do gênero *Eimeria* são responsáveis pela coccidiose aviária, uma doença entérica que acomete galinhas domésticas e é causada por sete espécies deste gênero: *E. acervulina*, *E. maxima*, *E. tenella*, *E. necatrix*, *E. brunetti*, *E. praecox* e *E. mitis* (Fernando, 1990). Economicamente, as espécies capazes de infectar a galinha doméstica são consideradas as mais relevantes do gênero *Eimeria* (Shirley *et al.*, 2005).

Estes parasitas são cosmopolitas e, acometem principalmente frangos de corte e matrizes reprodutoras (Williams, 1998). Esta doença resulta em aumento da conversão alimentar, menor capacidade de absorção de nutrientes pelo trato intestinal, menor ganho de peso e, em alguns casos, aumento da mortalidade (Min *et al.*, 2004).

Cada espécie se desenvolve em uma localidade específica do trato intestinal das aves (sítio-específica) e pode induzir diferentes graus de patogenicidade, variando de moderado até severo (Long e Joyner, 1984). A imunidade é espécie-específica, ou seja, uma ave imune a uma espécie é suscetível às outras (Dalloul e Lillehoj, 2005) e, além disso, um único hospedeiro pode ser infectado por mais de uma espécie simultaneamente.

Apesar da disponibilidade de drogas anticoccidianas e de vacinas vivas para o controle desta doença, os prejuízos causados à indústria avícola são enormes. Grande parte dos custos está relacionada à profilaxia e tratamento, além de perdas com o aumento de mortalidade e queda da produção. Calcula-se que os gastos mundiais relacionados ao controle desta doença variem de 800 milhões (Allen e Fetterer, 2002) a 3 bilhões de dólares por ano (Shirley *et al.*, 2004b).

A indústria avícola tem um papel fundamental na economia brasileira. O Brasil é o primeiro maior exportador e o terceiro maior produtor de carne de frango do mundo (Figura 2). O frango constitui o 3º produto das exportações na área de agronegócios e 5º produto brasileiro na pauta de exportação. Sendo exportado para

mais de 150 países. Em 2008, a produção de carne de frango chegou a 10,9 milhões de toneladas das quais 67% foi consumida pelo mercado interno e as exportações resultaram em uma receita cambial de quase 7 bilhões de dólares.

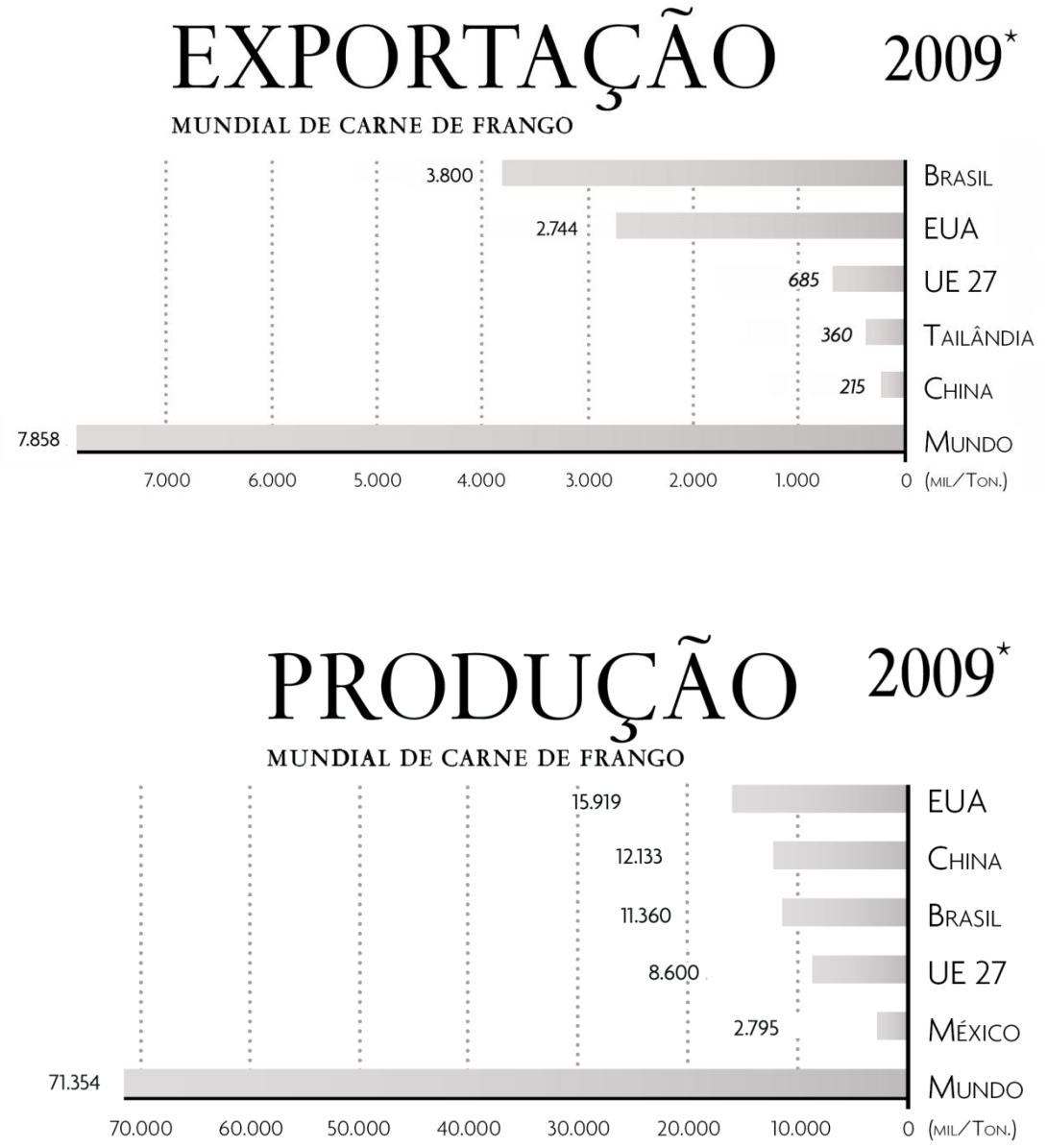


Figura 2 – Exportação (acima) e produção (abaixo) mundial de carne de frango, (*) projeção de 2009 em milhões de toneladas.
Fonte*:* Associação Brasileira dos Produtores e Exportadores de Frangos, Relatório Anual (ABEF) 2008-09.

1.2.2 Diagnóstico

Tradicionalmente, as espécies de *Eimeria* são identificadas com base em critérios como dimensão e morfologia dos oocistos, especificidade do hospedeiro, sítios intestinais de colonização, características macro e microscópicas das lesões intestinais, localização tecidual e morfologia das formas evolutivas, período pré-patente, especificidade imunológica e tempo mínimo de esporulação dos oocistos (Long e Joyner, 1984; Long *et al.*, 1976). Estes parâmetros analisados de forma conjunta permitem a distinção entre as espécies, entretanto, exigem pessoal altamente treinado. Nem sempre este diagnóstico é confiável, uma vez que pode ocorrer sobreposição de características entre espécies distintas, principalmente em infecções mistas. Além disso, o uso de vacinas de cepas precoces tem limitado a utilização do período de pré-patência como critério de discriminação.

Uma vez que as características biológicas não são suficientemente precisas para se fazer a discriminação entre as espécies de *Eimeria*, novas técnicas baseadas em métodos moleculares têm sido desenvolvidas.

Os primeiros ensaios diagnósticos baseados em PCR (*Polymerase Chain Reaction*) foram realizados utilizando a técnica de RAPD (*Random Amplified Polymorphic DNA*) (Johnston e Fernando, 1995; MacPherson e Gajadhar, 1993; Shirley e Bumstead, 1994). Além de não ser adequado para diagnóstico de amostras mistas, a baixa confiabilidade e reprodutibilidade deste método, restringiram severamente seu amplo uso entre diferentes laboratórios.

Schnitzler *et al.* (1998; 1999) desenvolveram um teste diagnóstico utilizando como alvos de amplificação a região do espaçador ribossômico interno transcrito 1 (ITS1).

Nosso grupo também desenvolveu um conjunto de marcadores moleculares denominados SCARs (*Sequence-Characterized Amplified Regions*), os quais permitem o diagnostico de espécies por PCR, empregando *primers* específicos. Desenvolvido na forma de um PCR multiplex, este teste tem a grande vantagem de permitir o diagnostico das sete espécies simultaneamente em uma única reação (Fernandez *et al.*, 2003).

Mais recentemente, alguns grupos desenvolveram diagnósticos moleculares

baseados na técnica de PCR em tempo real (Blake *et al.*, 2008; Kawahara *et al.*, 2008; Morgan *et al.*, 2009).

1.2.3 Controle da doença

As estratégias convencionais de controle desta doença baseiam-se principalmente no uso de drogas anticoccidianas, adicionadas na ração das aves de forma profilática ou quimioterápica (Williams, 1998). No entanto, o constante surgimento de cepas resistentes às drogas, aliada à preocupação do mercado consumidor com a contaminação dos alimentos com resíduos químicos, têm influenciado negativamente o uso de quimioterápicos (Dalloul e Lillehoj, 2005; Shirley *et al.*, 2007).

As restrições legislativas à utilização destes fármacos têm crescido, limitando a quantidade, tempo de exposição e gama de medicamentos permitidos, principalmente na Europa e Estados Unidos, onde algumas drogas já foram totalmente banidas (Shirley *et al.*, 2007). No Reino Unido, por exemplo, o número de anticoccidianos disponíveis comercialmente caiu de 17 em 2002, para apenas 10 em 2004. Em decorrência destas restrições, a última nova droga introduzida no mercado foi há mais de 10 anos.

Em detrimento ao uso de drogas, as vacinas vivas, compostas de cepas virulentas ou atenuadas de *Eimeria* são a única alternativa eficaz para o controle da coccidiose aviária (Williams, 1998). Estas vacinas são multivalentes (contém mais de uma espécie) já que a imunidade é espécie-específica e, geralmente, são compostas por cepas sensíveis às drogas. A primeira vacina lançada no mercado foi a Coccivac® (Schering-Plough) em meados de 1950, e permaneceu a única durante praticamente 35 anos. Atualmente, mais de 13 tipos de vacinas estão disponíveis no mercado (Chapman *et al.*, 2002).

Em relação às vacinas compostas por cepas virulentas, para que não ocorra queda nos índices de produção e ocasionalmente, desenvolvimento de sintomas clínicos da doença (Shirley *et al.*, 2005), a aplicação destas vacinas requer um controle rígido e a utilização de baixas doses.

O uso de vacinas de cepas atenuadas oferece uma alternativa mais segura. Os períodos de pré-patência do ciclo destes parasitas são mais curtos, há redução de um ou mais ciclos de reprodução assexuada acarretando em uma diminuição da capacidade reprodutiva sem prejuízo da imunogenicidade (Allen e Fetterer, 2002; Shirley *et al.*, 2007). No entanto, devido à baixa produção de oocistos, os custos de produção destas vacinas são muito maiores (Dalloul e Lillehoj, 2005).

Um sério problema relativo ao uso de vacinas vivas refere-se à imunovariabilidade entre cepas de uma mesma espécie o que muitas vezes não confere proteção cruzada, como é o caso de algumas cepas de *E. maxima* (Shirley *et al.,* 2004a*; Smith et al.,* 2002). Por exemplo, a vacina Paracox® (Schering-Plough) possui duas cepas de *E. maxima* em sua formulação.

1.3 A *Eimeria tenella* como modelo de estudo da coccidiose aviária

1.3.1 Caracterização

Desde o trabalho pioneiro de Tyzzer (1929), a *E. tenella* tem sido considerada como modelo de estudo para o entendimento da coccidiose aviária (Chapman e Shirley, 2003). Esta espécie apresenta alta prevalência no campo, alta virulência, facilidade de obtenção de formas invasivas (zoítas) e de oocistos diretamente dos cecos, alta taxa de replicação e de esporulação, além de ser a única espécie capaz de ser propagada *in vitro* (Chapman e Shirley, 2003; Ling *et al.*, 2007). A cepa Houghton (H), isolada no Reino Unido em 1949, cuja linhagem de origem clonal foi obtida em 1986, é uma das poucas cepas de referência usadas para estudos experimentais (Chapman e Shirley, 2003; Ling *et al.*, 2007). Recentemente, alguns artigos relatando a transfecção estável em *Eimeria tenella* foram publicados, no entanto, para seleção de transfectantes, ainda é necessária a utilização de passagens *in vivo* (Clark *et al.*, 2008; Shi *et al.*, 2009; Yan *et al.*, 2009; Zou *et al.*, 2009).

E. tenella é altamente patogênica, provocando perda de peso, diarréia com sangue e até mortalidade das aves (de Venevelles *et al.*, 2004). Coloniza prioritariamente os cecos (Figura 3A), onde há o desenvolvimento de lesões hemorrágicas e, em casos severos há necrose tecidual associada à formação de coágulos sanguíneos. As lesões também podem ser observadas na superfície serosa dos cecos (Figura 3B).

Este parasita é caracterizado por um ciclo de vida direto (monoxênico), com vários ciclos intestinais endógenos envolvendo estágios assexuados (merogonia ou esquizogonia), seguidos de um estágio sexual (gamogonia ou gametogonia) que resulta na formação de um oocisto não esporulado (Shirley e Harvey, 2000). Os oocistos esporulados (Figura 3C), assim como em todos os organismos do gênero *Eimeria*, apresentam 4 esporocistos, com 2 esporozoítos cada.

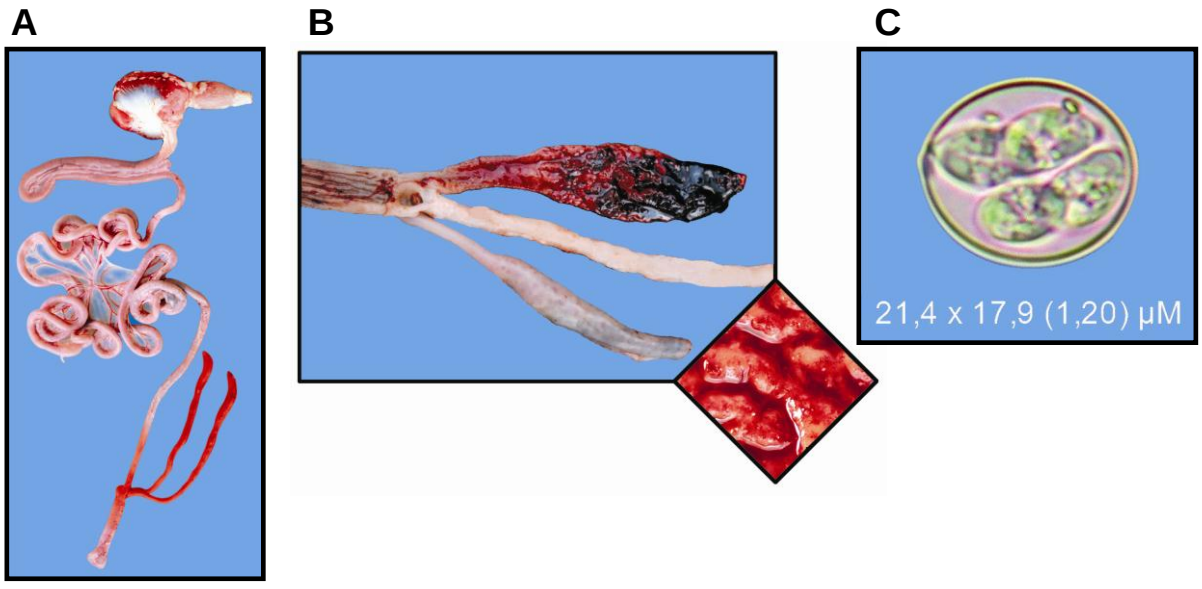


Figura 3 - *Eimeria tenella*: sítio da lesão intestinal, lesão macroscópica e oocisto. (A) Os cecos, destacados em vermelho constituem o sítio de colonização desta espécie, (B) lesões hemorrágicas nos cecos podem ser visualizadas tanto na mucosa quanto na serosa. (C) Oocisto de *E. tenella*, com as medidas de tamanho referentes ao diâmetro maior e menor e a razão entre eles (entre parênteses).

1.3.2 Ciclo de vida

O ciclo de vida de *Eimeria tenella* (Figura 4) é iniciado quando uma ave suscetível ingere um oocisto esporulado (Etapa 1). Este oocisto é rompido fisicamente pela ação da moela liberando os esporocistos. Os esporocistos, no intestino delgado, são submetidos à ação de sais biliares e tripsina que digerem o corpúsculo de Stieda (localizado em uma das extremidades do esporocisto), abrindo um pequeno orifício por onde os esporozoítos saem ativamente (Etapa 2), processo este denominado de excistação.

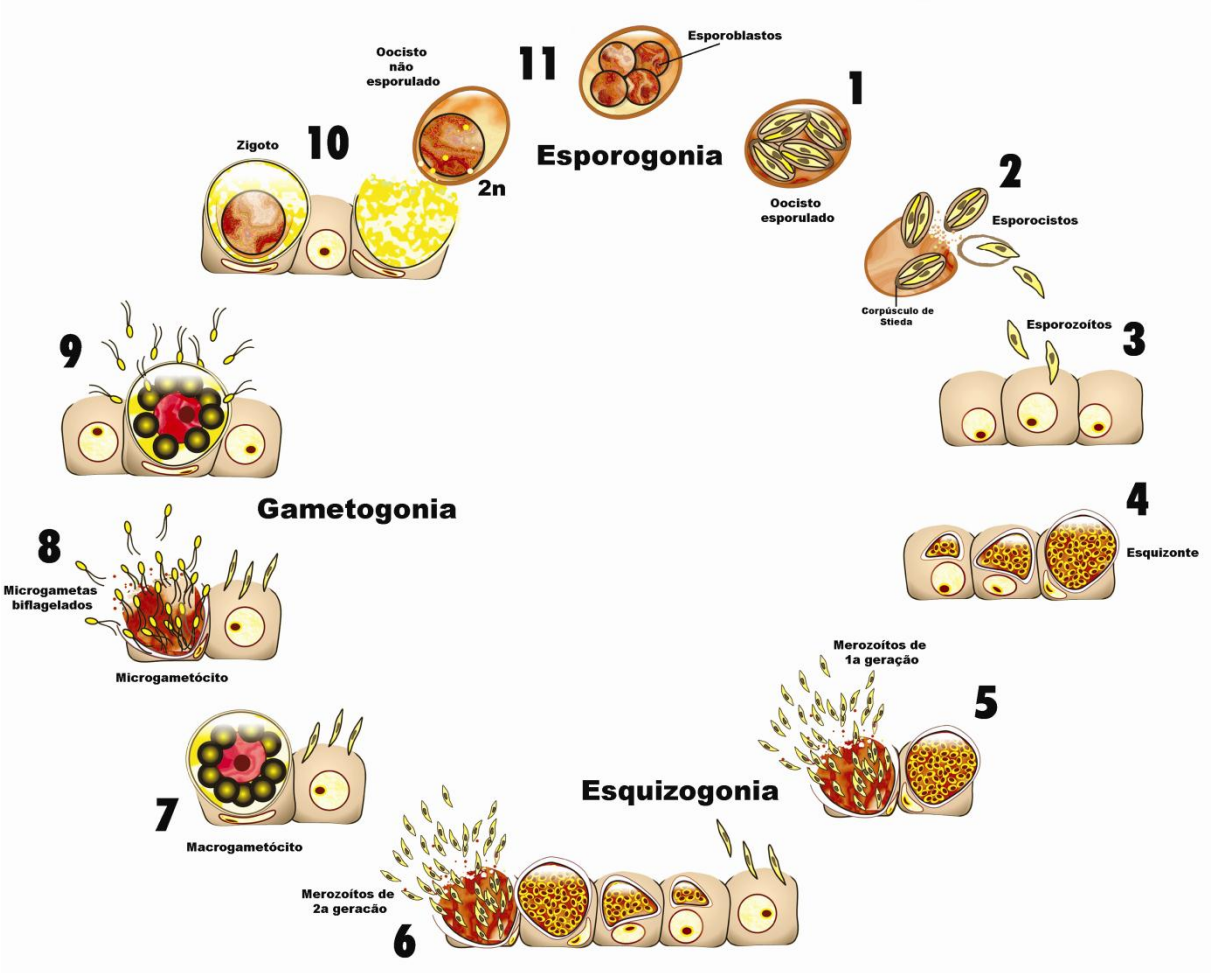


Figura 4 - Ciclo de vida de *Eimeria tenella* (Arte Gráfica: Helton Barreiro).

Os esporozoítos, em seguida, são carregados até as criptas cecais por linfócitos intra-epiteliais, onde aderem e invadem as células epiteliais e da submucosa (Etapa 3), diferenciando-se em seguida em formas esquizontes.

Os esquizontes passam por um ciclo de reprodução assexuada denominada de esquizogonia (merogonia), que consiste em múltiplas divisões mitóticas sem citocinese que geram esquizontes multinucleados (Etapa 4) os quais se diferenciam em merozoítos uninucleados de 1ª geração. Com o rompimento destas células, os merozoítos de 1ª geração são liberados (Etapa 5), invadem novas células (Etapa 6), formando em seguida o esquizonte de 2ª geração.

Após 3 ou 4 ciclos de reprodução assexuada (esquizogonia), cada merozoíto pode se diferenciar em macrogametócito (Etapa 7) ou microgametócito (Etapa 8). Os microgametócitos se rompem liberando os microgametas biflagelados que fertilizam os macrogametas (Etapa 9), resultando na formação de um zigoto (único estágio diplóide do parasita). O zigoto é coberto por camadas multilamelares que irão constituir a parede do oocisto. Uma vez maduro o oocisto não esporulado é liberado no ambiente juntamente com as fezes (Etapa 10). Sob condições ideais de temperatura (entre 26-28ºC), umidade e oxigenação, o oocisto passa por uma etapa de reprodução assexuada (esporogonia), também conhecida como esporulação. Durante este processo, o oocisto não esporulado (diplóide) sofre uma meiose, gerando 4 esporoblastos haplóides (Etapa 11), cada um, após uma divisão mitótica, dá origem aos esporocistos com dois esporozoítos cada (Current *et al.*, 1990; Fernando, 1990; Hammod, 1973; Kinnaird *et al.*, 2004; Schmatz, 1997).

1.3.3 Caracterização do genoma

O genoma nuclear de *Eimeria* spp. é haplóide durante a maior parte do ciclo de vida deste parasita, com exceção do zigoto. A cepa H de *Eimeria tenella* tem sido utilizada amplamente para os estudos de caracterização molecular e determinação do genoma destes organismos.

O material genômico haplóide está organizado em 14 cromossomos, que variam de tamanho entre 1,0 a mais de 7,0 Mpb, determinado através de

eletroforese de campo-pulsado (PFGE – *pulsed-field gel electrophoresis*) (Shirley, 2000, 1994). O genoma possui uma complexidade de cerca de 55 milhões de pares de bases (Ling *et al.*, 2007) e conteúdo GC de 53% (Shirley, 2000).

O genoma da cepa H de *E. tenella* foi seqüenciado empregando a abordagem *shotgun* com uma cobertura de 8,4 vezes e os dados estão publicamente disponíveis no sítio do Instituto Sanger (http://www.sanger.ac.uk/Projects/E_tenella/). Esta iniciativa foi patrocinada pelo BBSRC (*Biotechnology and Biological Sciences Research*) e realizada pelo WTSI (*Wellcome Trust Sanger Institute*, Cambridge, Inglaterra), em colaboração com o IAH (*Institute for Animal Health*, Compton, Inglaterra).

Além do genoma, o cromossomo 1 desta espécie foi totalmente seqüenciado (Ling *et al.*, 2007). Este cromossomo possui uma organização genômica incomum segmentada em regiões R (*feature-rich*), que são ricas em repetições, apresentando elementos similares a transposons e repetições teloméricas, confirmando estudos preliminares que apontavam para uma distribuição freqüente de regiões repetitivas no genoma de *Eimeria tenella* (Shirley, 2000; Shirley *et al.*, 2004b). Também apresenta segmentos livre de repetições P (*feature-poor*) (Ling *et al.*, 2007).

Grande parte das repetições seriadas do trinucleotídeo CAG e do heptâmero AGGGTTT estão presentes nas 3 regiões de segmentos R, que são separadas por 4 regiões de segmentos P. Os microssatélites CAG são encontrados preferencialmente em regiões codificadoras, enquanto que as seqüências do heptâmero, identificadas como unidades repetitivas teloméricas em *Plasmodium*, são encontradas nas pontas da seqüência consenso, assim como nas regiões intrônicas e intergênicas do segmento R. Além destes, há também o octâmero palíndromo TGCATGCA, aparentemente único em apicomplexas, que está presente 157 vezes das quais grande parte estão em regiões intergênicas do segmento R. As duplicações gênicas em sua grande maioria estão associadas aos segmentos P e podem conter genes, enquanto que as duplicações detectadas nos fragmentos R são geralmente pequenas e associadas às regiões similares a telômeros (Ling *et al.*, 2007).

Além do genoma nuclear, estão presentes nestes parasitas, dois genomas extracromossômicos: o mitocondrial e o do apicoplasto. Estes protozoários apresentam uma única mitocôndria de formato tubular e o genoma é constituído de

concatâmeros lineares contendo unidades repetitivas de 6kb (Chapman e Shirley, 2003; Romano, 2004). O apicoplasto é uma organela exclusiva dos organismos do Filo Apicomplexa, homóloga aos cloroplastos das plantas, não possuindo atividade fotossintética (Waller e McFadden, 2005). O genoma desta organela em *Eimeria tenella* é circular, rico em conteúdo AT e contém cerca de 35 kb (Cai *et al.*, 2003).

1.4 Métodos de análise da expressão gênica

A geração de seqüências de DNA genômico tem sido de grande utilidade para a caracterização molecular de organismos. No entanto, o conhecimento destas seqüências não é suficiente para a caracterização inequívoca dos genes expressos. Para um melhor entendimento da biologia dos parasitas, diversos estudos sobre expressão gênica têm sido realizados (Ojopi *et al.*, 2007; Skuce *et al.*, 2005).

O modelo de genes eucarióticos é extremamente complexo conforme ilustrado e descrito na Figura 5.

Considerando que as proteínas constituem o produto final da função gênica, o estudo destas representa a forma mais direta de avaliação dos genes expressos. Como existem várias estudos que indicam que a quantidade de proteína produzida pode ser diretamente proporcional à quantidade de RNAm (RNA mensageiro) que a codifica (Yamamoto *et al.*, 2001) a medição dos níveis de RNAm pode também ser utilizada para o estudo de função e anotação dos genes (Sun *et al.*, 2004).

Os métodos para a análise de expressão gênica evoluíram rapidamente em poucos anos. Tradicionalmente, a avaliação do nível de um RNAm transcrito era realizada empregando-se métodos como o *Northern blot* e RT-PCR (*Reverse Transcriptase* PCR) semi-quantitativo, utilizando-se seqüências de DNAs complementares (cDNAs) transcritas reversamente a partir das amostras de RNA mensageiros, as quais representam uma fonte direta para a descoberta de regiões codificadoras de genes.

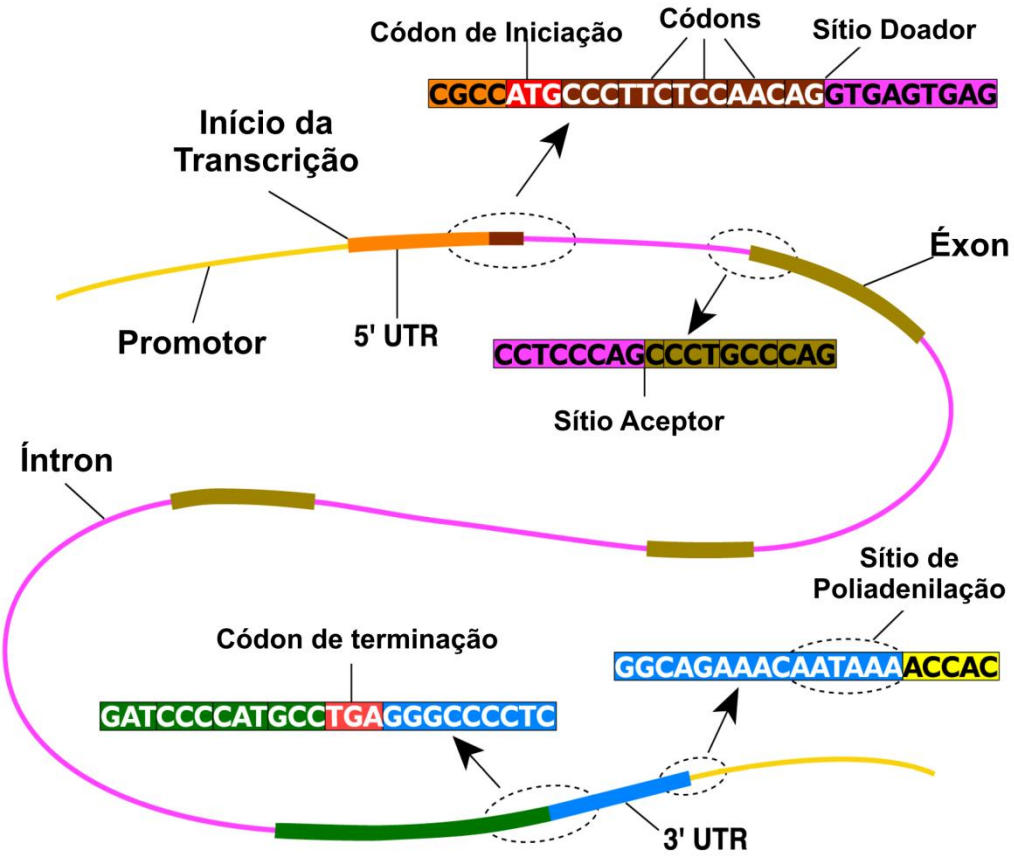


Figura 5 – Esquema do modelo gênico de eucariotos. A linha contínua corresponde à seqüência de DNA. As regiões intergênicas estão nas extremidades em amarelo. O gene começa com o sítio de início da transcrição, e o final é representado pelo sítio de poliadenilação. As barras que estão na linha representam os éxons do gene, enquanto que as linhas em rosa são os íntrons (não transcritos). As junções entre éxons e íntrons são sítios de *splicing* denominados de sítio doador e sítio aceptor. A tradução do transcrito é iniciada no códon de iniciação e finalizada no códon de terminação originando o produto gênico. As barras em laranja e azul correspondem às regiões não traduzidas (UTRs).
Fonte: Adaptado de Wu, 2008.

Em 1991 Adams *et al.* descreveram uma abordagem para obtenção e sequenciamento de pequenas seqüências expressas, denominadas de ESTs (*Expressed Sequence Tags*) que permitem avaliar o perfil transcricional de um subconjunto de genes expressos. Milhares de seqüências de ESTs foram geradas para vários organismos em diferentes tecidos ou estágios (Nagaraj *et al.*, 2007).

Basicamente, para a geração de ESTs, clones de bibliotecas de cDNA são selecionados randomicamente e submetidos a um sequenciamento de uma única fita (*single-pass*) gerando seqüências curtas de aproximadamente 300-600 pb que podem ser obtidas a partir de um tecido, ou estágio específico de desenvolvimento de um organismo. As bibliotecas convencionais podem ser construídas a partir da região 5' ou 3' dos transcritos, utilizando-se adaptadores que permitem a clonagem unidirecional. As bibliotecas direcionais 5' apresentam maior cobertura de regiões codificadoras, facilitando a identificação de genes desconhecidos. No entanto, as bibliotecas 3', que geralmente identificam as porções não traduzidas (3'UTR – *Untranslated Regions)* apresentam maior variabilidade do que as regiões codificadoras e permitem a identificação inequívoca destes transcritos, facilitando a obtenção de um perfil de transcrição quantitativo (Gruber, 2007). Bibliotecas de ESTs também podem ser construídas de maneira alternativa empregando a técnica ORESTES (*Open Reading Frame EST*). Este método baseia-se em PCR de baixa estringência utilizando *primers* arbitrários para amplificação dos cDNA (Dias Neto *et al.*, 2000), resultando desta forma, em uma maior cobertura da região central da sequência dos RNAs mensageiros, a qual contém a maior parte da porção codificadora dos genes. Como estas bibliotecas são normalizadas, aumenta-se a probabilidade de detecção de transcritos raros. Importante salientar que apenas bibliotecas não normalizadas podem refletir quantitativamente os níveis de expressão dos transcritos (Moody, 2001). Entretanto, a detecção de transcritos de baixa abundância é limitada, principalmente nas bibliotecas não normalizadas (Alba *et al.*, 2004; Sun *et al.*, 2004). Além disso, a cobertura dos transcritos em sua maioria é parcial, sujeita a erros e, muitas vezes, devido à freqüência heterogênea dos transcritos, pode ser altamente redundante e enviesada (Gruber, 2007; Li *et al.*, 2003; Nagaraj *et al.*, 2007).

A geração de ESTs tem contribuído de maneira significativa para a identificação de novos genes, determinação das regiões codificadoras das sequências genômicas (fronteiras íntron-exon), mapeamento físico, assim como análise de perfil de expressão (Sterky e Lundeberg, 2000). Além disso, ESTs que compartilham trechos em comum podem ser agrupadas e montadas, permitindo muitas vezes a identificação de sequências completas de RNAs mensageiros (Kozian e Kirschbaum, 1999).

Técnicas de hibridização subtrativa de bibliotecas de cDNA permitiram um avanço na análise dos padrões de expressão gênica, constituindo um dos primeiros métodos a serem utilizados para detectar genes diferencialmente expressos em larga escala. Basicamente, os genes diferencialmente expressos são isolados pela hibridização de duas amostras de cDNAs: uma amostra em excesso (teste) e uma segunda amostra (controle). Os transcritos expressos em ambas as amostras formam uma molécula híbrida, enquanto que as moléculas exclusivamente expressas na amostra teste permanecem em forma de fita simples, estas últimas, são isoladas podendo ser utilizadas para a construção de bibliotecas de cDNA (Moody, 2001). A técnica de hibridização subtrativa de cDNAs começou a ser utilizada na década de 80 sendo aprimorada anos mais tarde de forma a incluir uma etapa de PCR, a qual foi denominada de *Supression Subtractive Hybridization* (SSH) (Diatchenko *et al.*, 1996). Nesta variação, a etapa de amplificação é direcionada de forma a amplificar apenas as seqüencias de genes diferencialmente expressos.

Além dos métodos subtrativos, a geração de *fingerprinting* de RNA também foi uma das primeiras abordagens de seleção de genes de interesse (diferencialmente expressos) e incluem o *Differential Display* (DD) (Liang e Pardee, 1992) e o RNA *fingerprinting by arbitrary primed* PCR (RAP_PCR) (Welsh *et al.*, 1992). O DD é baseado na conversão do RNAm em cDNA (duas ou mais amostras a serem comparadas), utilizando oligo dT *primers,* em seguida, o cDNA é amplificado empregando-se oligo dT *primers* e *primers* arbitrários. As amostras são separadas por tamanho em gel de poliacrilamida avaliando-se a presença ou ausência de bandas, assim como bandas de diferentes intensidades as quais podem representar potenciais transcritos diferencialmente expressos. Estas bandas são excisadas, os transcritos são clonados e seqüenciados (Moody, 2001). Apesar de largamente utilizada, esta técnica apresenta um alto índice de falsos positivos (Green *et al.*, 2001). O RAP-PCR é uma abordagem muito semelhante ao DD, que ao invés de usar oligo dT *primers* para síntese de cDNA utiliza *primers* arbitrários de 10 pb, o que permite a utilização de RNA total nos ensaios experimentais (Avison, 2008). Estes métodos apresentam bons resultados e ainda são usualmente empregados, no entanto, possuem uma limitação quanto ao número de transcritos que podem ser analisados simultaneamente, além disso, são capazes de gerar

apenas um retrato parcial sem informação direta a respeito da abundância destes genes.

Recentemente, outros métodos de análise de expressão gênica foram desenvolvidos incluindo SAGE (*Serial Analysis of Gene Expression*), MPSS (*Massive Pararell Signature Sequencing*), e *Microarrays* (van Ruissen *et al.*, 2008). Atualmente, os métodos de análise de expressão gênica são divididos em dois grupos: sistemas de plataformas abertas ou fechadas.

Uma plataforma é considerada fechada, quando os genes que determinam o espaço da investigação são finitos, e, portanto, para utilização destas técnicas é necessário o conhecimento prévio das seqüências a serem utilizadas o que depende da disponibilidade de informações do organismo a ser estudado (Green *et al.*, 2001). Os métodos mais amplamente utilizados são os de *microarrays,* classificados como plataformas fechadas de larga escala baseadas em hibridização. Os mais conhecidos são os *microarrays* de cDNA (Schena *et al.*, 1995), de oligonucleotídeos (Pease *et al.*, 1994) e os *Affymetrix chips* comercialmente disponíveis que sintetizam centenas de sondas de oligonucleotídeos *in situ* (diretamente nos chips) utilizando um método fotolitográfico (http://www.affymetrix.com/index.affx). O PCR em tempo real (*Real time PCR* ou qPCR, *quantitative* PCR) (Heid *et al.*, 1996), é também considerado como uma plataforma fechada, porém de baixa escala.

Um *microarray* é tipicamente definido como um arranjo pré-definido de moléculas (fragmentos de DNA, cDNAs ou oligonucleotídeos) quimicamente ligadas (ligações covalentes) a uma superfície sólida ou a uma membrana (carregada positivamente). A superfície sólida pode ser de vidro ou de silicone (*Affy chip - Affymetrix*), ou conter orifícios microscópicos na superfície (*Illumina*- http://www.illumina.com/pages.ilmn?ID=5). O princípio deste método baseia-se na hibridização entre as sondas que compõem estes microarranjos com o gene ou molécula de interesse. A ligação dos alvos marcados com fluoróforos permite a detecção e quantificação das moléculas hibridizadas. Em seguida, uma imagem de hibridação é gerada e detectada por meio de leitores (*scanners*) (Dufva, 2009; Jaluria *et al.*, 2007).

Centenas a milhares de sondas (genes) podem ser testadas em um único experimento, o que torna a tecnologia de *microarray* bastante atraente e amplamente utilizada. No entanto, como qualquer outra técnica, apresenta

limitações. Por se tratar de um método de hibridização, são necessárias replicatas experimentais (normalmente duas), além disso, os dados de fluorescência somente detectam a presença ou ausência das moléculas hibridizadas, não fornecendo nenhuma outra informação a respeito do tamanho ou composição dos alvos capturados. Problemas relativos à hibridização cruzada (famílias de genes contendo seqüências muito semelhantes) ou inespecífica são inerentes à técnica, já que moléculas de diferentes tamanhos e composições são simultaneamente submetidas ao processo. Assim, a escolha correta das sondas é de fundamental importância para o sucesso do experimento, bem como a utilização de análise estatística complexa para interpretação dos dados de fluorescência. Como diferentes marcadores, métodos de hibridização e construção dos arrays podem ser empregados, a comparação inter-plataformas é uma tarefa difícil. Medições de expressão gênica absolutas somente são obtidas através de microarrays de oligonucleotídeos (*single-channel*), enquanto que as outras tecnologias, em geral, são relativas (*dual-channel*) (Dufva, 2009; Jaluria *et al.*, 2007).

O PCR em tempo real tornou-se um procedimento muito bem estabelecido para quantificação de níveis de expressão gênica. Esta técnica baseia-se na detecção da quantidade de produto amplificado em cada ciclo da reação empregando-se fluoróforos (Ponchel, 2006). Para tal, diferentes estratégias podem ser utilizadas como substâncias fluorescentes intercalantes de DNA (SYBR® Green) ou uso de sondas marcadas. O SYBR® Green possui a propriedade de emitir fluorescência quando ligado às moléculas de DNA dupla fita e é amplamente utilizado, devido ao seu baixo custo, facilidade de uso e sensibilidade. Como este corante se liga a qualquer molécula dupla-fita de DNA, como dímeros de *primers* ou produtos inespecíficos, a concentração do fragmento alvo pode ser superestimada (Kubista *et al.*, 2006; Ponchel, 2006). Este problema pode ser minimizado empregando-se *primers* que não formem dímeros e que sejam mais específicos. O TaqMan® (sondas de hidrólise) utiliza sondas que apresentam em sua extremidade um fluoróforo repórter, e na outra extremidade um fluoróforo aceptor (*quencher*), que quando ligados na mesma molécula não emitem fluorescência, pois o fluoróforo aceptor absorve o sinal emitido pelo repórter. Durante a amplificação e sob a atividade de exonuclease da *Taq* DNA polimerase, a sonda de oligonucleotídeos é clivada, o repórter e o aceptor são separados, permitindo a detecção da

fluorescência emitida pelo fluoróforo repórter. O princípio da detecção da fluorescência é similar ao do SYBR® Green no qual os produtos amplificados são mensurados a cada ciclo de amplificação. Como somente a fluorescência emitida pelas seqüências alvo é mensurada, a tecnologia TaqMan® é mais específica (Bell e Ranford-Cartwright, 2002; Valasek e Repa, 2005). Existem outras variações da utilização de um repórter e um *quencher*, baseados em sondas de hidrólise, incluindo *molecular beacons, sunrise primers* e *scorpion primers*, entretanto os custos são maiores. Há também sondas baseadas em um único corante, na qual a fluorescência é detectada de acordo com a ligação ao DNA, tais como as sondas *LightUp*, que possuem em sua estrutura um ácido nucléico peptídico neutro (PNA) que se liga ao DNA com maior afinidade do que os oligonucleotídeos comuns (Kubista *et al.*, 2006; Valaseq e Repa, 2005).

Na técnica de *real time PCR,* os métodos de quantificação podem ser classificados como absolutos ou relativos. Na quantificação absoluta há a construção de uma curva padrão baseada na diluição de uma molécula de concentração conhecida, enquanto que nas abordagens relativas à determinação da diferença de expressão é calculada em relação a um grupo experimental de amostras controle (Peirson *et al.*, 2003). Apesar da alta sensibilidade, a ampla utilização da técnica de PCR em tempo real ainda é limitada devido ao alto custo e baixo *throughput.*

Um dos grandes desafios da chamada era pós-genômica está na elucidação do transcriptoma dos organismos o qual depende de predições computacionais e de seqüências previamente caracterizadas (que podem estar incompletas ou mal anotadas). A utilização das plataformas abertas de larga escala pode acelerar estes estudos e melhorar a qualidade do conhecimento obtido até o momento (Harbers e Carninci, 2005). Estas plataformas não necessitam de um conhecimento *a priori* do transcriptoma a ser estudado permitindo desta forma, a descoberta de novos genes. Além disso, não são limitadas a uma possível complexidade inerente destes transcritos, como por exemplo, a presença de SNPs (*Single Nucleotide Polymorphims)*, de *splicing* alternativo, entre outros (Green *et al.*, 2001).

Dentre as plataformas abertas de larga escala destacam-se o SAGE (*Serial Analysis of Gene Expression*) e suas variações (microSAGE, miniSAGE, LongSAGE, superSAGE, etc) e o MPSS (*Massive Pararell Signature Sequencing*), considerados

como métodos de perfil de expressão digital, pois os dados de expressão gênica obtidos são analisados utilizando-se a contagem dos transcritos (freqüência). A contagem é uma medida simples e estatisticamente fácil de modelar. A expressão dos transcritos é medida em termos absolutos, permitindo uma análise global da expressão gênica.

O MPSS (Brenner *et al.*, 2000; Reinartz *et al.*, 2002) baseia-se na produção de sequências (*signature sequences*) de 16-20 bases adjacentes ao sítio de restrição da enzima *Dpn*II mais próximo à extremidade 3' do RNAm. Este método, no entanto, possui peculiaridades quanto à manipulação bioquímica e estratégia de sequenciamento, combina clonagem e amplificação de cDNA *in vitro* na superfície de milhares de *microbeads* a uma alta capacidade de sequenciamento de DNA não baseada em gel. Em cada *microbead* há a ligação de apenas uma molécula de cDNA. Após a hibridização, as seqüências alvo são clonadas obtendo-se cerca de 100.000 fragmentos de cDNA idênticos. Para o sequenciamento, os *microbeads* são posicionados em uma plataforma de análise e ligados a um adaptador (contendo o sítio de reconhecimento para a enzima do tipo IIs). Cada *bead* é digerido de forma seqüencial com esta enzima que cliva o DNA de forma a deixar uma extremidade coesiva de 4 nucleotídeos. Adaptadores fluorescentes (possuindo todas as combinações possíveis dos 4 nucleotídeos) são ligados as extremidade coesivas. A imagem é captada por uma câmera e analisada por um software capaz de processar os sinais de fluorescência durante as etapas de hibridização, ligação e clivagem determinando a seqüência destes 4 nucleotídeos. O processo todo é repetido algumas vezes, até a obtenção das sequências de 16-20 nucleotídeos. Cada experimento pode gerar cerca de 250.000 – 400.000 sequências. Como esta tecnologia é baseada em instrumentos sofisticados e protegida por direitos autorais, seu custo é elevado, sendo disponível apenas comercialmente inicialmente pela *Lynxgxgen Therapeutica, Inc*., e posteriormente pela *Solexa Inc.*, que encerrou esta prestação de serviço em 2006.

SAGE (Velculescu, 1999; Velculescu *et al.*, 2000; Velculescu *et al.*, 1995; Velculescu *et al.*, 1997) é uma técnica que envolve o isolamento de pequenas seqüências de nucleotídeos, denominadas *tags* (etiquetas), que são concatenadas, seqüenciadas e quantificadas (Tuteja e Tuteja, 2004). Cada *tag* consiste de 4 pb da seqüência de reconhecimento (CATG) da enzima de restrição *Nla*III e 10 pb de

seqüência adjacente na direção da cauda poliA de uma molécula de RNAm (Patankar *et al.*, 2001). Cada *tag* representa uma seqüência expressa e a contagem de cada *tag* é diretamente proporcional à sua freqüência em uma dada população de RNAm (Dinel *et al.*, 2005). Além de permitir a descoberta de novos transcritos (Rivals *et al.*, 2007), SAGE permite a identificação de transcritos de baixa abundância (Chen *et al.*, 2002; Kim *et al.*, 2006), com uma sensibilidade 26 vezes maior do que ESTs (Sun *et al.*, 2004).

A caracterização de transcritos diferenciais, resultantes de *splicing* alternativo (Khattra *et al.*, 2007; Kuo *et al.*, 2006; Wahl *et al.*, 2005a), poliadenilação heterogênea (Keime *et al.*, 2007; Ojopi *et al.*, 2007), SNPs (*Single nucleotide polymorphisms*) (Radke *et al.*, 2005; Silva *et al.*, 2004), transcritos *antisense* (Patankar *et al.*, 2001; Quére *et al.*, 2004; Wahl *et al.*, 2005b; Williams *et al.*, 2007) e, até mesmo microRNAs (Ge *et al.*, 2006) também pode ser obtida através da análise dos dados de SAGE.

O SAGE tem sido utilizado com sucesso em diversos organismos, tanto para a obtenção de perfis de expressão gênica, assim como na identificação de genes diferencialmente expressos, incluindo estudos em humanos (Keime *et al.*, 2007), principalmente relacionados a câncer (Boon *et al.*, 2002; Kirschbaum-Slager *et al.*, 2005; Silveira *et al.*, 2008), em organismos modelos como ratos (Wahl *et al.*, 2005a, 2005b), *Saccharomyces cerevisiae* (Velculescu *et al.*, 1997), *Caenorhabditis elegans* (Jones *et al.*, 2001; Wang *et al.*, 2009), *Danio rerio* (*zebrafish*) (Knoll-Gellida *et al.*, 2006), *Arabidopsis* spp. (Fizames *et al.*, 2004; Robinson e Parkin, 2008). Esta técnica também foi aplicada em estudos de parasitas, como *Schistossoma mansoni* (Ojopi *et al.*, 2007; Taft *et al.*, 2009; Williams *et al.*, 2007), *Leishmania* spp. (Guerfali *et al.*, 2008; Li *et al.*, 2008), *Haemonchus contortus* (Skuce *et al.*, 2005) e *Giardia lamblia* (Palm *et al.*, 2005).

Dentro do Filo Apicomplexa estudos em *Plasmodium falciparum* (Munasinghe *et al.*, 2001; Patankar *et al.*, 2001), esporozoítos de *Plasmodium berguei* encontrados na glândula salivar de mosquito (Rosinski-Chupin *et al.*, 2007) e *Toxoplasma gondii* (Radke *et al.*, 2005) também utilizaram a plataforma SAGE. Estes estudos revelaram alta expressão de transcritos estágio específicos, detecção de genes mitocondriais, *tags antisense*, poliadenilação alternativa, além da caracterização de novos genes. Em *Toxoplasma*, foram construídas bibliotecas de

SAGE utilizando taquizoítos das 3 cepas de referência deste organismo (tipo 1 - RH, tipo 2 - Me49B7 e tipo 3 – VEGmsj). A partir da cepa VEG foi também realizado um estudo detalhado de seis fases do ciclo de vida, como oocistos, e várias fases de reprodução assexuada até a diferenciação em bradizoítos. Foram encontrados diferentes conjuntos de RNAm quando comparados a eucariotos superiores ou eucariotos unicelulares, como *Saccharomyces*. Foi observado que menos de 5% dos genes representam cerca de 75% dos transcritos totais. Diversos genes apicomplexa específicos foram encontrados nas classes de genes mais expressos. Segundo os autores, a transcrição nestes parasitas parece ser bastante dinâmica, com um alto número de genes expressos exclusivamente em um único estágio de desenvolvimento, demonstrando a importância da regulação transcricional.

Um aprimoramento desta técnica é o LongSAGE que ao invés de *tags* de 14pb, gera *tags* com 21pb, facilitando desta forma a predição *in silico* e a anotação gênica (Bianchetti *et al.*, 2007; Saha *et al.*, 2002). Por serem mais longas, o mapeamento das *tags* no transcriptoma e no genoma é mais específico, (Akmaev e Wang, 2004; Wahl *et al.*, 2004; Wahl *et al.*, 2005a) quando comparado ao SAGE convencional (Li *et al.*, 2006; Lu *et al.*, 2004), possui maior sensibilidade e cobertura gênica do que MPSS (Hene *et al.*, 2007), além de ser a técnica de larga escala como menor índice de viés composicional GC (Margulies *et al.*, 2001; Siddiqui *et al.*, 2006).

1.5 Base de dados de GO (*Gene Ontology*) e KOG (*euKaryotic Orthologous Groups*)

O conhecimento da função biológica de uma determinada proteína em um dado organismo pode ajudar a inferir a sua função em outros organismos. Genes ortólogos são genes presentes em diferentes espécies derivados de um ancestral comum, e que normalmente retém essa mesma função no decorrer da evolução, enquanto que parálogos estão relacionados a eventos de duplicação (Koonin, 2005; Tatusov *et al.*, 1997). A comparação de novos dados com proteínas conhecidas é importante para anotação, e, conseqüentemente, atribuição da função de produtos

gênicos ainda não caracterizados. As bases de dados de GO e KOG são atualmente muito utilizadas para este tipo de busca e possuem a grande vantagem de serem curadas, o que aumenta a confiabilidade de suas informações, no entanto, apresentam um menor volume de dados.

Com o objetivo de estabelecer um vocabulário controlado comum, estruturado, dinâmico, precisamente definido e padronizado para a descrição das funções dos produtos gênicos em quaisquer organismos, em 2000, foi criado o consórcio de Ontologia Gênica (GO) (Ashburner *et al.*, 2000). A classificação de GO é estruturada em três ontologias de termos para descrever: a função molecular, o processo biológico e o componente celular dos produtos protéicos. A função molecular descreve a função normalmente exercida pelo produto gênico, como por exemplo, *structural molecule activity* (atividade molecular estrutural), o processo biológico classifica o processo em que o produto protéico está envolvido, como *metabolic process* (processo metabólico) e o componente celular, descreve a localização celular onde o produto ocorre, por exemplo, *organelle* (organela) (Dimmer *et al.*, 2008). Uma mesma proteína pode estar classificada dentro das três ontologias, refletindo a realidade biológica desta em particular, que pode funcionar em diferentes processos, conter domínios que carregam diversas funções moleculares e participar em interações múltiplas com outras proteínas, organelas ou localizações celulares. A estrutura do GO é representada com um gráfico acíclico direto conhecido como DAG (*directed acyclic graph*), onde as relações são diretas, como por exemplo, a mitocôndria é uma organela, mas uma organela não é uma mitocôndria. A organização deste gráfico consiste em termos *child* (filhos), que são mais especializados e *parents* (pais), que são menos específicos, no entanto, cada termo pode possuir mais de um termo *parent*, e também termos *sibling* (irmãos) (Martin *et al.*, 2004). Os termos são ligados por relações do tipo *is a* (é um) e *part of* (é parte de) conforme exemplificado na Figura 6.

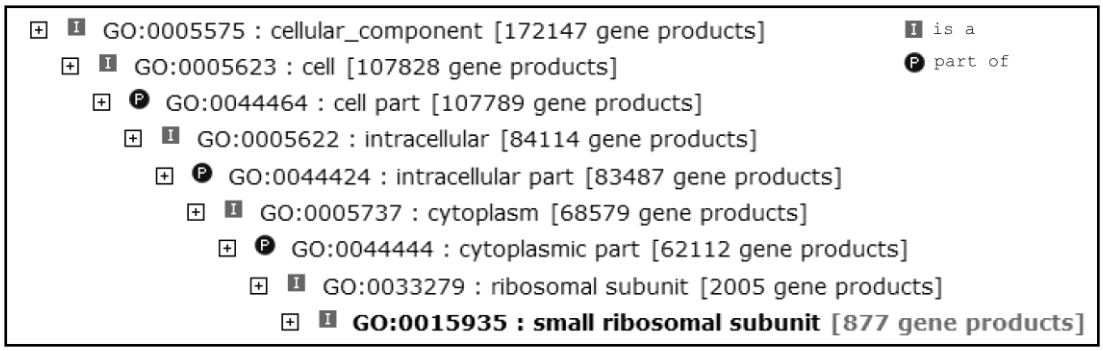


Figura 6 – Exemplo de gráfico acíclico direto (DAG) de termos GO, mostrando a estrutura hierárquica e as relações entre os termos: *is a* (é um) e *part of* (é parte de).

O COG (*Clusters of Orthologous Groups of proteins*) (Tatusov *et al.*, 2000) é uma das bases de dados de ortológos mais antigas sendo composta atualmente por 66 organismos unicelulares e a base de dados equivalente para eucariotos é denominada de KOG (*euKaryotic Orthologous Groups*). Apesar de não serem atualizadas desde 2003, são consideradas por muitos autores como a base de dados de ortólogos padrão (Altenhoff e Dessimoz, 2009). A base de dados de KOG inclui proteínas de 7 genomas de eucariotos: três de animais (*Caenorhabditis elegans*, *Drosophila melanogaster* e *Homo sapiens),* uma planta *Arabidopsis thaliana*, dois fungos (*Saccharomyces cerevisiae* e *Schizosaccharomyces pombe*) e o parasita intracelular microsporídeo *Encephalitozoon cuniculi*, que atualmente pertencente ao Reino Fungi. O conjunto atual de KOG consiste de 4.852 *clusters* de ortólogos, que incluem 59.838 proteínas (Tatusov *et al.*, 2003).

Cada KOG consiste de um gene ortólogo individual ou de grupos de ortólogos parálogos de 3 ou mais linhagens filogenéticas (cada genoma é considerado uma linhagem filogenética). Em outras palavras, quaisquer duas proteínas provenientes de diferentes linhagens, e que pertençam ao mesmo KOG são consideradas ortólogas. Assume-se que cada COG/KOG evoluiu de um ancestral comum através de uma série de eventos de especiação e duplicação. Para definição dos KOGs, as seqüências codificadoras de proteínas dos sete genomas eucarióticos foram comparadas entre si, e para cada proteína, o melhor *hit* (*best hit – BeT*), ou seja, o melhor alinhamento encontrado contra cada um dos outros genomas foi computado.

Se um dos genomas possui *BeTs* em outros dois genomas, considera-se que estas seqüências sejam ortólogas, e, portanto, este é o critério utilizado para definir uma unidade mínima de KOG (Tatusov *et al.*, 1997).

1.6 Estudo da expressão gênica em *Eimeria* spp.

1.6.1 Transcriptoma

Vários projetos para geração de ESTs de *Eimeria* de galinha doméstica, envolvendo diferentes estágios do ciclo de vida destes parasitas foram realizados. Até o momento foram depositadas no *GenBank* cerca de 35.000 sequências ESTs (*Expressed Sequence Tags*) de *E. tenella*, das quais 80% destas foram obtidas a partir de estágios esporozoítos e merozoítos de segunda geração. Grande parte destas seqüências depositadas foram resultantes do projeto WashU-Merck *E. tenella* estabelecido entre a *Washington University* e a Merck, que produziu 27,500 ESTs. Cerca de 1.000 seqüências foram produzidas pela *University Kebangsann Malaysia* (Ng *et al.*, 2002; Wan *et al.*, 1999), 499 ESTs foram gerados a partir de oocistos esporulados e não esporulados pelo USDA (*United States Departamento of Agriculture*) (Miska *et al.*, 2004). *Welcome Trust Sanger Institute* gerou mais de 8.000 ESTs incluindo oocistos não esporulados, esporozoítos e merozoítos de primeira geração (dados não publicados), destas aproximadamente 2.500 seqüências já estão disponíveis publicamente no *Genbank*.

Além destas ESTs convencionais, o nosso grupo, como parte integrante do Consórcio Internacional do Genoma de *E. tenella* (http://www.sanger.ac.uk/Projects/E_tenella/consortium.shtml) foi responsável pelo sequenciamento de mais de 15.000 ESTs do tipo ORESTES (*Open Reading Frame EST*) para cada uma das três principais espécies de *Eimeria* de galinha doméstica: *E. acervulina*, *E. maxima* e *E. tenella* (Shirley *et al.*, 2004b). Em *E. tenella*, as seqüências geradas pelo nosso grupo, foram obtidas a partir de diferentes estágios

evolutivos deste parasita, incluindo: oocistos esporulados, oocistos parcialmente esporulados, oocistos não esporulados, merozoítos de 2ª geração e esporozoítos.

Protocolos de anotação automática de sequências ORESTES de *Eimeria* spp. de galinha doméstica foram desenvolvidos pela nossa equipe de bioinformática e aplicados aos conjuntos de dados utilizando-se componentes recém-desenvolvidos para a plataforma EGene (Durham *et al.*, 2005; Ferro, 2008). No caso de *E. tenella*, além dos dados de ORESTES, foram utilizados os ESTs convencionais disponíveis publicamente, o que permitiu uma cobertura tanto das regiões centrais dos transcritos, obtidas pelos ORESTES, quanto das regiões terminais. Para *E. maxima* e *E. acervulina*, foram usadas somente as sequências ORESTES produzidas pelo nosso grupo.

Os dados relativos à anotação automática estão disponíveis no sítio www.coccidia.icb.usp.br/eimeria/annotation/. O sítio *web* encontra-se provisoriamente protegido por senha (*username* consortium, *senha* etgc2005). Esta página *web* disponibiliza atualmente os produtos da anotação para estas três espécies de *Eimeria*, a montagem das seqüências, a classificação dos termos de ontologia gênica (GO), anotação na base de dados KOG, além de uma base de dados relacional.

1.6.2 Expressão gênica diferencial

A *Eimeria* apresenta um ciclo de vida complexo, envolvendo vários estágios de desenvolvimento, e pouco se sabe quais são os conjuntos de genes mais importantes e expressos em cada um dos estágios de desenvolvimento da *Eimeria*. Por exemplo, as formas zoítas como esporozoítos, merozoítos de primeira geração e merozoítos de segunda geração apresentam uma morfologia relativamente semelhante, e compartilham funcionalmente a utilização do complexo apical para a adesão e penetração dos parasitas nas células do hospedeiro. As fases de proliferação são críticas para a patogênese no hospedeiro, enquanto o ciclo sexual é de suma importância para a transmissão, dispersão e geração da diversidade

genética. Apesar dos oocistos não esporulados não serem infectantes, no ambiente sob condições adequadas, o processo de esporulação é disparado. Os oocistos de *Eimeria* responsáveis pela dispersão da doença para novos hospedeiros são capazes de permanecer no ambiente por longos períodos devido à proteção conferida pela parede dos oocistos (Belli *et al.*, 2005; Schmatz, 1997).

Em *Toxoplasma gondii* e *Plasmodium falciparum,* a análise do perfil de transcrição em diferentes estágios evolutivos, revelou a predominância de grupos de genes diferencialmente expressos e predominantemente estágio-específicos (Bozdech *et al.*, 2003; Cleary *et al.*, 2002; Duncan, 2004; Llinas e DeRisi, 2004). Estudos dos níveis protéicos também indicaram um perfil de expressão semelhante (Hall *et al.*, 2005; Xia *et al.*, 2008).

Estudos comparativos entre diferentes estágios evolutivos de *Eimeria* já foram descritos, no entanto, grande parte destes utilizou um número de transcritos reduzido, ou a caracterização estágio-específica foi direcionada a um determinado gene ou a uma família de genes específicos.

Wan *et al.* (1999), analisando 500 ESTs provenientes de merozoítos de segunda geração, observou que 47,7% destas ESTs tiveram resultados positivos em buscas de similaridade empregando o programa BLAST, incluindo proteínas de ribossomo (correspondentes a quase um terço dos resultados positivos), um grupo de proteínas com funções enzimáticas, tais como quinases, oxidases, desidrogenases, sintetases, entre outras, e um grupo de ESTs com identidades putativas representando moléculas com uma vasta gama de funções. Dentro destes três grupos, 14,3% representaram genes previamente identificados para *E. tenella*. Um estudo muito similar realizado pelo mesmo grupo (Ng *et al.*, 2002) analisou 556 ESTs provenientes de esporozoítos. Após a busca por similaridade por BLAST foi observado um padrão distinto ao observado em merozoítos. Apenas 27,3% das seqüências apresentaram resultados positivos, sendo que destes, 22,5% representaram genes nunca estudados em *E. tenella.* ESTs com funções putativas foram classificadas em 10 grupos baseado em suas funções biológicas, para tanto, foram utilizadas 125 seqüências de esporozoítos e 179 seqüências de merozoítos. Esporozoítos mostraram um maior número de transcritos associados com crescimento celular, divisão celular e síntese de DNA, enquanto que merozoítos apresentaram transcritos envolvidos com a expressão de genes e proteínas.

Uma análise de mais de 55.000 ESTs de organismos do Filo Apicomplexa incluindo cerca de 13.500 ESTs provenientes de *Eimeria tenella* foi realizada por Li *et al.* (2003). Quando comparadas, 80% das seqüências provenientes de esporozoítos e merozoítos de *Eimeria tenella*, foram classificadas como estágio-específicos. e, interessantemente, o mesmo índice de transcritos estágio-específicos foi encontrado quando formas assexuais de *Plasmodium falciparum* foram analisadas.

Um conjunto de 499 ESTs, 225 clones específicos de oocistos esporulados e 274 de oocistos não esporulados foram obtidos empregando o método de hibridização subtrativa (Miska *et al.*, 2004). A montagem destas seqüências gerou 162 *contigs* dos quais 64% não apresentam resultados em buscas por similaridade empregando o programa BLAST. Em seguida, os dados provenientes dos oocistos foram comparados com seqüências disponíveis de outros estágios de *Eimeria,* metade dos transcritos únicos de oocistos esporulados foi também encontrada em merozoítos e esporozoítos, enquanto que 79% dos transcritos isolados de oocistos não esporulados não foram detectados em outros estágios de desenvolvimento. Não é surpresa que oocistos esporulados compartilhem semelhanças aos estágios invasivos, já que cada oocisto esporulado contém 8 esporozoítos no seu interior, entretanto, a outra metade dos transcritos não foi identificada e, segundo os autores, poderiam fazer parte de um grupo de genes com expressão limitada a oocistos esporulados (Miska *et al.*, 2004).

Avaliando-se a presença de proteínas de micronema durante o processo de esporulação, Ryan *et al.* (2000) demonstraram que a expressão de genes destas proteínas é altamente coordenada. No início da esporulação muitas dessas proteínas foram detectadas em níveis baixos, mas, a partir de 22 horas, quando ocorre a maturação dos esporozoítos, cinco tipos de proteínas de micronema foram observadas em altos níveis. Entretanto, os RNAs mensageiros que codificam para estas proteínas foram detectados cerca de 10 a 12 horas antes, indicando um controle tanto transcricional, quanto pós-transcricional. Tomley *et al.*, (2001) ao estudar e caracterizar a EtMIC4 (*E. tenella* Proteína de Micronema 4) verificaram que esta proteína também apresenta localização diferencial durante a esquizogonia. No início desta fase, esta proteína estava presente na extremidade apical dos esporozoítos, desaparecendo durante o progresso desse processo. Na etapa final,

foi detectada novamente sendo observada tanto no citoplasma, quanto na periferia do esquizonte. Nos esquizontes totalmente maduros, a EtMIC4 foi localizada apenas na extremidade apical dos merozoítos.

A regulação da expressão gênica nos diferentes estágios de desenvolvimento de *Eimeria tenella* também foi relatada para outros genes. A EtCRK2, uma quinase dependente de ciclina, é transcrita tanto nas fases sexuais quanto assexuais do parasita, sendo diferencialmente expressa durante o processo de esporulação. Segundo Kinnaird *et al.* (2004) o nível do RNAm que codifica para esta proteína mostrou-se reduzido entre 6 a 12 horas de esporulação, e no final do processo, entre 40-48 horas, elevou-se novamente. Schaap *et al.* (2005) estudaram transcritos que codificam proteínas de ribossomo. Foi observado que estes transcritos são expressos de forma diferencial, representando cerca de 10% dos transcritos totais encontrados em merozoítos e praticamente não sendo detectados em oocistos e esporozoítos. Transcritos que codificam as proteínas de ribossomo L5 e L23 foram detectados nos esquizontes não sendo observados em oocistos esporulados.

Portanto, assim como já relatado para outros organismos do Filo Apicomplexa, a regulação da expressão gênica em *Eimeria* spp também parece ser estágio-específica. Considerando a importância da espécie *Eimeria tenella* tanto economicamente quanto como modelo de estudo para coccidiose, um aprofundamento do conhecimento da expressão de genes diferencialmente expressos nos estágios de desenvolvimento deste parasita pode ser de grande importância para a caracterização de genes fundamentais para que o ciclo biológico deste parasita se complete.

As fases invasivas têm um papel fundamental na patogênese da doença, os esporozoítos são os responsáveis pelo início da replicação endógena do parasita e são o maior alvo da resposta imune protetiva do hospedeiro (Lillehoj e Lillehoj, 2000), enquanto que os merozoítos de segunda geração estão relacionados à patogenia da doença propriamente dita (Ng *et al.*, 2002), sendo responsáveis pelos maiores danos causados à mucosa e submucosa intestinais (Schmatz, 1997).

Estudos comparativos realizados pelo nosso grupo a partir de seqüências ORESTES provenientes de diversos estágios de desenvolvimento do ciclo de *Eimeria tenella* sugerem que um pequeno conjunto de genes é altamente expresso em cada fase do ciclo de vida, e que menos de 20% dos genes são compartilhados

entre os estágios estudados (Novaes *et al.*, 2005). Estes dados foram obtidos a partir de bibliotecas normalizadas e, portanto, não refletem um quadro quantitativo, além disso, as ESTs convencionais disponíveis publicamente (em sua grande parte provenientes de esporozoítos e merozoítos) também não refletem completamente um perfil quantitativo, devido às limitações inerentes à técnica.

Assim, com o objetivo de se obter um perfil de expressão gênica em larga escala, tanto quantitativo quanto qualitativo, decidimos empregar a técnica de LongSAGE (Saha *et al.*, 2002) para construção de bibliotecas a partir esporozoítos e merozoítos de segunda geração de *E. tenella*, duas fases invasivas de grande importância no ciclo de vida destes organismos.

Os dados gerados poderão fornecer novas informações em relação aos genes expressos nestes organismos, incluindo dados quantitativos mais acurados. Além disso, estes dados poderão complementar as informações atualmente disponíveis (ORESTES e ESTs), possibilitando um melhor conhecimento do transcriptoma destes parasitas e a elucidação de vários mecanismos moleculares ainda desconhecidos, como também os envolvidos na regulação das principais etapas do ciclo de vida destes parasitas.

O estudo de genes diferencialmente expressos entre estas duas fases não somente poderá contribuir para um melhor entendimento do processo de adesão e interiorização do parasita nas células do hospedeiro, como também no futuro poderão ser utilizados como alvos terapêuticos, permitindo assim o desenvolvimento de novas estratégias de controle da doença.

2 OBJETIVOS

2.1 Objetivos Gerais

- Analisar comparativamente o perfil de expressão de esporozoítos e merozoítos de segunda geração de *Eimeria tenella* empregando a técnica de LongSAGE.

2.2 Objetivos Específicos

- Construção de duas bibliotecas de LongSAGE de cada estágio evolutivo (merozoítos de segunda geração e esporozoítos);
- Pré-processamento das seqüências dos concatâmeros, extração e contagem das *tags;*
- Processamento dos dados *in silico* e análise estatística;
- Mapeamento das *tags* contra o genoma de *E. tenella*;
- Identificação dos genes diferencialmente expressos entre merozoítos e esporozoítos;
    - Mapeamento das *tags* diferencialmente expressas contra as montagens de ESTs / ORESTES;
    - Anotação automática dos genes diferencialmente expressos e curagem manual;
    - Análise de ontologia gênica utilizando a classificação GO (*Gene Ontology*) e KOG (*Eukaryotic Orthologous Groups*) para os genes diferencialmente expressos;
- Validação das análises de LongSAGE utilizando um pequeno conjunto de genes para quantificar os dados de expressão empregando-se a técnica de Real-time PCR*;*

- Comparação entre os dados de SAGE e de *Northern* eletrônico (ESTs/ORESTES) para os estágios de esporozoítos e merozoítos de segunda geração de *E. tenella;*
- Publicação dos resultados da anotação automática na *web*, para consulta e análise dos resultados obtidos.

3 MATERIAL E MÉTODOS

3.1 Preparo e tratamento das soluções

Todas as soluções foram preparadas utilizando-se água purificada por osmose reversa, deionizada e tratada com luz UV pelo Sistema RiOs e Milli-Q Synthesis da Millipore (água Milli-Q). As soluções foram autoclavadas por 15 minutos para a eliminação de possíveis contaminações por ácidos nucléicos e enzimas. Para reagentes termossensíveis, utilizou-se água pré-autoclavada.

Para a purificação de RNA mensageiro, todo o material utilizado foi adquirido ou previamente tratado de forma a não conter RNAses. Para isto, todas as soluções foram preparadas com água previamente tratada com dietilpirocarbonato de sódio (DEPC) 0,1% por 2 horas, a 37 ^{o}C e posteriormente autoclavadas. Toda a vidraria utilizada foi incubada em Forno Pasteur a 180 ^{o}C por 8 horas e o material plástico foi tratado com solução 0,1% DEPC por 2 horas a 37 ºC e em seguida autoclavado. Bancadas e fluxos utilizados nas etapas de manipulação de RNA, foram limpos com solução de dodecil sulfato de sódio (SDS) 0,5%.

3.2 Condições de criação e manutenção das aves

Machos de postura da linhagem Bovans White foram gentilmente cedidos pela granja Kunitomo (Mogi das Cruzes, SP) com um dia de idade. As aves foram criadas e mantidas no biotério de aves do Laboratório de Biologia Molecular de Coccídias, Departamento de Parasitologia, Instituto de Ciências Biomédicas, Universidade de São Paulo.

Considerando-se que várias espécies de *Eimeria* podem infectar simultaneamente um mesmo hospedeiro, e como no nosso laboratório, algumas das espécies de *Eimeria* de galinha doméstica, são manipuladas e propagadas com certa freqüência, é imprescindível que as condições de criação e manutenção das aves sejam severamente controladas.

O nosso biotério possui um sistema de filtração do ar de entrada e de saída. A sala de criação tem pressão positiva enquanto que as salas de infecção pressão

negativa e, além disto, as salas de infecção possuem uma rede de gás amônia instalada, para esterilização do ambiente e eliminação de oocistos residuais.

Além destas medidas, durante o período de criação, as fezes das aves foram monitoradas quinzenalmente quanto à presença de oocistos de *Eimeria* que acometem a galinha doméstica.

3.3 Propagação dos Parasitas

3.3.1 Infecção das Aves

Para a propagação dos parasitas, foram utilizados machos de postura da linhagem Bovans White com idade entre 3 a 5 semanas. Os animais foram criados em ambiente livre de coccídias e alimentados *ad libitum* com ração sem adição de antibióticos e anticoccidianos. Para infecção com *Eimeria tenella* cepa H, utilizou-se uma dose infectante de $2x10^3$ e $5x10^5$ oocistos esporulados/ave para obtenção de oocistos e merozoítos de segunda geração, respectivamente. A inoculação foi feita por via oral com o auxílio de uma seringa de plástico conectada a uma sonda de polietileno. A cepa de referência H de *E. tenella,* utilizada neste estudo, foi gentilmente cedida pelo Dr. Martin W. Shirley (*Institute for Animal Health*, Reino Unido).

3.3.2 Purificação dos oocistos de *Eimeria tenella*

Os oocistos foram purificados empregando-se uma variação do método utilizado por Long *et al.* (1976). Após 7 dias de infecção os cecos das aves foram coletados e abertos longitudinalmente. A mucosa dos cecos foi raspada, com auxílio de uma lamínula de vidro e lavada com água destilada. O raspado foi peneirado em uma tela de aço inoxidável de 40 *mesh* e, posteriormente centrifugado a 2.500 *g* por 5 minutos. O sobrenadante foi descartado e o sedimento foi submetido a um

tratamento com uma solução 10% (v/v) de hipoclorito de sódio 10-12% (cloro ativo). Após incubação por 10 minutos no gelo, a amostra foi diluída em água destilada, novas centrifugações foram realizadas, até a completa remoção do hipoclorito de sódio (cerca de 2 a 3 centrifugações).

3.3.3 Esporulação dos oocistos

Os oocistos obtidos de *Eimeria tenella* foram ressuspensos em solução de dicromato de potássio 2% (p/v) na concentração de 250.000 oocistos/mL e submetidos à esporulação (Shirley, 1995). A esporulação dos oocistos foi realizada durante 72 horas à temperatura de 28 °C (Long *et al.*, 1976) em uma câmara de incubação B.O.D. (Fanem) com controle automático da temperatura. Os oocistos foram mantidos sob oxigenação forçada através de borbulhamento de ar.

3.3.4 Tratamento dos oocistos com hipoclorito de sódio

Para purificação dos esporozoítos, os oocistos esporulados foram submetidos a um tratamento com hipoclorito de sódio visando a remoção de impurezas e possíveis bactérias contaminantes. Inicialmente, a suspensão de oocistos foi centrifugada a 2.500 *g* por 5 minutos e o sedimento lavado duas vezes em água destilada para a remoção completa do dicromato de potássio. Em seguida, o sedimento foi ressuspenso em hipoclorito de sódio 10-12% (cloro ativo) e incubado por 10 minutos em gelo. Após a incubação, o volume foi completado com água destilada na proporção (9:1) e centrifugado nas mesmas condições acima. A centrifugação foi repetida por 2 a 3 vezes lavando-se com água destilada para completa remoção do hipoclorito de sódio.

3.3.5 Isolamento e purificação de esporozoítos

Os esporozoítos de *E. tenella* foram obtidos a partir de oocistos esporulados previamente tratados com hipoclorito de sódio, conforme previamente descrito (ver item 3.3.4).

Primeiramente, os oocistos foram ressuspensos em meio Hanks, transferidos para um *erlenmeyer* e um volume de pérolas de vidro de 3 mm de diâmetro foi adicionado à amostra. Para uma ruptura parcial dos oocistos visando a obtenção de esporocistos, os oocistos foram incubados à temperatura ambiente sob agitação (200 rpm). O progresso da quebra foi cuidadosamente monitorado por microscopia óptica. Esta etapa de quebra parcial *in vitro* é equivalente ao processo de ruptura mecânica dos oocistos que ocorre na moela das aves, culminando na liberação dos esporocistos.

Após a quebra parcial, a amostra contendo os esporocistos, foi transferida para outro *erlenmeyer*, foi ressuspensa em solução de excistação (Hanks acrescido de 0,25% de tripsina; 1% de ácido cólico e 10 mM de cloreto de magnésio; pH 7,6) e incubada a 41 ºC sob leve agitação durante 3 horas, ou até total excistação dos esporozoítos. Após a excistação *in vitro* a amostra foi lavada duas vezes em solução salina tamponada (PBS) (pH 8,0; 1% glicose) centrifugando-se a 1.500 *g* durante 5 minutos.

Para a purificação de esporozoítos é necessário remover os oocistos e esporocistos provenientes das etapas anteriores. Para tanto, utilizou-se uma coluna híbrida, que contém duas camadas: uma de fibras de nylon para filtração do material e uma de resina DE-52 para cromatografia por troca iônica (Figura 7) (Shirley, 1995). Esta coluna é capaz de reter os oocistos, cascas de oocistos e esporocistos, e permite a passagem dos esporozoítos. A amostra obtida após a excistação *in vitro* foi adicionada à esta coluna, diversas lavagens com PBS (pH 8,0; 1% glicose) foram efetuadas, e a passagem dos esporozoítos foi monitorada por microscopia óptica. Os esporozoítos obtidos foram centrifugados, e posteriormente, submetidos à extração de RNA mensageiro.

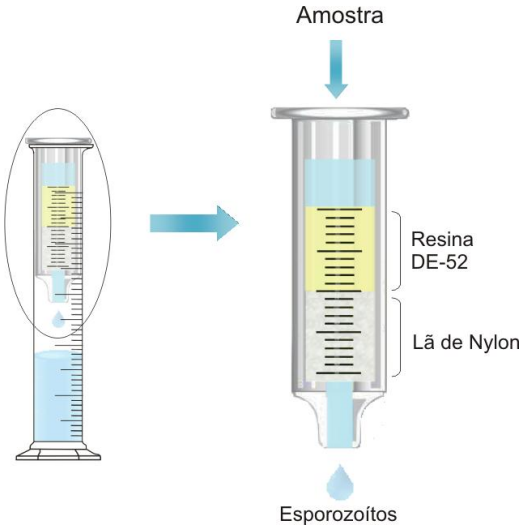


Figura 7 – Coluna para a purificação de esporozoítos. Método híbrido envolvendo a utilização de lã de nylon para filtração e resina DE-52 para cromatografia por troca iônica.

3.3.6 Purificação de merozoítos de segunda geração de *E. tenella*

A purificação de merozoítos de *E. tenella* foi realizada de acordo com Shirley (1995), com pequenas modificações. Após 112 horas de infecção, os cecos das aves foram coletados, mantidos e lavados em PBS gelado (pH 7,6) e cortados em pequenos pedaços (2 mm). Os fragmentos de tecido foram adicionados a uma solução de excistação (Hanks acrescido de 0,25% de tripsina; 1% de ácido cólico e 10 mM de cloreto de magnésio; pH 7,6) e incubados a 41 ^{o}C sob agitação de 120 rpm por 45 minutos. Após a incubação, os fragmentos foram peneirados em telas de aço inoxidável de 50 *mesh* e 140 *mesh*, e posteriormente, a amostra foi lavada em PBS (pH 7,6) centrifugando-se 2 a 3 vezes a 1500 g por 5 minutos. Os merozoítos obtidos foram submetidos à extração de RNA mensageiro.

3.3.7 Verificação de contaminação inter-específica

Além da realização dos exames parasitológicos das fezes das aves mantidas no nosso biotério (ver item 3.2), todas as amostras de oocistos purificados foram submetidas a um ensaio de PCR multiplex que é capaz de diagnosticar

simultaneamente as sete espécies de *Eimeria* que infectam a galinha doméstica. Este teste baseia-se na utilização de marcadores SCAR (*Sequence Characterized Amplified Regions*) espécie-específicos (Fernandez *et al.*, 2003) e foi empregado com o objetivo de monitorar uma possível contaminação inter-específica no material purificado.

Para tanto, o DNA foi extraído a partir de $5x10^6$ oocistos previamente tratados com hipoclorito de sódio conforme descrito no item 3.3.4, e ressuspensos em tampão de extração (Tris-Cl 10 mM pH 8,0; EDTA 50 mM pH 8,0). Os oocistos foram rompidos por meio de agitações com pérolas de vidro de 0,5 mm de diâmetro, em velocidade máxima por cerca de 6 minutos. Após uma centrifugação a 20.000 *g* durante 10 minutos a 4 ºC, o sobrenadante foi recolhido e tratado com RNAse A (20µg/mL) a 37 ºC por 30 minutos. A seguir foram adicionados SDS (0,5% p/v) e Proteinase K (100µg/mL) e a amostra foi incubada a 50 ºC durante 1 hora. O DNA foi purificado pelo método fenol/clorofórmio (Sambrook e Russell, 2001). Os ensaios de PCR foram realizados de acordo com Fernandez *et al.* (2003) e, para todas as reações foram adicionados controles positivo e negativo. Os produtos amplificados foram submetidos à eletroforese em gel de agarose 1,5% em tampão 1x Tris-borato-EDTA (TBE) e visualizado sob luz UV.

3.4 Extração de RNA mensageiro

O RNA mensageiro (RNAm) de esporozoítos e merozoítos de segunda geração foi extraído utilizando-se o kit µMACS mRNA isolation (Miltenyi Biotech), conforme instruções do fabricante. Este kit baseia-se no uso de oligo dT ligado a partículas paramagnéticas permitindo o isolamento do RNAm a partir de células em uma única etapa. Para avaliação quantitativa e qualitativa, um décimo do material purificado foi submetida à eletroforese em gel de agarose 1% em tampão 1x MOPS (0,2 M de *ácido* 3-(N-morfolino propano sulfônico), 20 mM acetato de Sódio; 10mM EDTA, pH 8,0), na ausência de brometo de etídeo. Após a corrida, o gel foi incubado por 10 minutos em MOPS 1x contendo brometo de etídeo (0,5 µg/mL) e visualizado sob luz UV. Para a construção de bibliotecas de LongSAGE o RNA mensageiro foi

quantificado pela medição da absorbância das amostras em espectrofotômetro (GeneQuant™, Amersham Pharmacia Biotech).

3.4.1 Tratamento com DNAse

Para eliminação de possíveis contaminações com DNA genômico, as amostras de RNA mensageiro purificado, foram submetidas a tratamento com DNAse. Para o ensaio utilizou-se DNAse RQ1 (Promega) livre de RNAse (1U/10µL de reação), na presença de tampão de DNAse 1x concentrado, DTT (2 mM final) e inibidor de RNAse, RNAsin (Promega) (0,5 U/µL de reação). A amostra foi incubada a 37 ^{o}C por 30 minutos e a reação de digestão foi interrompida pela adição de *Stop Solution* e incubação a 65 ^{o}C por 5 minutos. O RNA mensageiro após a digestão foi devidamente dividido em alíquotas e congelado a -80 ºC.

3.4.2 Verificação de contaminação do RNA mensageiro com DNA genômico

Além do tratamento com DNAse (ver item 3.4.1), a presença de DNA genômico contaminante nas amostras de RNA mensageiro foi descartada pela realização de ensaio de PCR empregando *primers* baseados em marcadores SCAR espécie-específicos para *E. tenella* (Fernandez *et al.*, 2003). Os testes de PCR foram realizados em um volume final de 25 µL contendo 200 µM de dNTP, 1x de tampão, 1 U de BIOLASE^{TM} DNA polimerase (Bioline, UK), 1,5 mM de $MgCl_2$, 1 µM de cada *primer* específico para *E. tenella*, empregando as condições de ciclagem segundo Fernandez *et al.* (2003). Para todas as reações foram realizados controles positivo e negativo. Os produtos amplificados foram submetidos à eletroforese em gel de agarose 1,5%.

3.4.3 Verificação da integridade do RNA mensageiro

Com o objetivo de verificar a integridade do RNA mensageiro purificado, foi realizada uma reação de transcrição reversa (RT-PCR) empregando *primers* específicos para uma região do gene de Proteína de Micronema 4 (ETMIC-4 - *accession number* AJ306453) de *Eimeria tenella.* A EtMIC4 é expressa em esporozoítos e merozoítos, e está associada à locomoção e invasão do parasita na célula hospedeira (Tomley *et al.*, 2001). A amplificação desta região gera um fragmento de 986 pares de bases (pb). Por ser de grande tamanho, a amplificação deste fragmento indica de forma indireta que o RNAm das amostras está integro. A síntese de cDNA foi realizada a partir de *primer* específico, utilizando-se a enzima SuperScript™ II *Reverse Transcriptase* (Invitrogen), conforme instruções do fabricante. A amplificação do cDNA foi realizada utilizando-se 10% do volume da reação de síntese de cDNA, 1 U de Platinum *Taq* DNA Polymerase (Invitrogen), 1x tampão da enzima, 1,5 mM de Cloreto de Magnésio, 100 µM de dNTP e 1 µM de cada primer (*Foward* (F) - TATAGACGAGTGCCAAGACCCG; *Reverse* (R) – CCGTCACCTGAATAGCCAGCTA). As condições da reação consistiram em uma desnaturação inicial de 95 ºC por 5 minutos; 30 ciclos de 95 ºC por 1 minuto, 56 ºC por 1 minuto e 72 ºC por 1 minuto e uma etapa final de extensão a 72 ºC por 5 minutos. Para verificação da amplificação, um terço de cada amostra amplificada foi submetido à eletroforese em gel de agarose.

3.5 Construção de bibliotecas de LongSAGE

As bibliotecas de LongSAGE foram obtidas a partir de esporozoítos e merozoítos de segunda geração de *Eimeria tenella* cepa H. Para tal, foi empregado o kit I-SAGE Long (Invitrogen), de acordo com as instruções do fabricante (Figura 8):

a) o RNA mensageiro extraído foi ligado às esferas magnéticas contendo oligo dT;
b) em seguida, o RNAm foi submetido à síntese de cDNA dupla fita;

c) o cDNA foi submetido à digestão utilizando-se a enzima de restrição sítio-específica *Nla*III. Como o sítio de restrição desta enzima ocorre em média a cada 256 pb, grande parte dos transcritos pode ser clivada ao menos uma vez. Controles relativos à síntese de cDNA e digestão com a enzima *Nla*III foram realizados conforme descrito no item 3.5.1;

d) após a digestão, o cDNA foi dividido em duas partes e cada uma destas partes foi ligada a um adaptador A e B (oligo com cerca de 40pb). Os adaptadores contem 4 bases complementares ao cDNA digerido com *Nla*III, um sítio de reconhecimento para a endonuclease *Mme*I na extremidade 3' e sítios de pareamento de *primers* para a etapa posterior de amplificação;

e) para a liberação das *tags* da cauda poliA, o cDNA foi digerido com a enzima *Mme*I. Esta enzima de restrição tipo IIS (enzima de *tagging*) tem a propriedade de clivar numa distância definida, de aproximadamente 21 pb *downstream* ao sítio de reconhecimento, localizado no adaptador. A digestão libera então *tags* de aproximadamente 60 pb com extremidades coesivas de 2 bases. Cada *tag* consiste em 40 pb de bases do adaptador e aproximadamente 21 pb de uma sequência única de cada transcrito;

f) as duas partes de cDNA foram misturadas e submetidas à reação de ligação para a formação das *ditags* de cerca de 130pb. As *ditags* foram amplificadas utilizando-se *primers* complementares às seqüências dos adaptadores;

g) após a amplificação, as *ditags* de 130 pb foram isoladas e purificadas e, posteriormente, digeridas com a enzima *Nla*III para liberação dos adaptadores. As *ditags* de 34 pb, livre dos adaptadores, são resultantes de seqüências derivadas exclusivamente dos cDNAs transcritos. Cada *ditag* contém em suas extremidades sítios de pontuação, correspondentes aos sítios de clivagem da enzima *Nla*III;

h) as *ditags* de aproximadamente 34 pb foram então purificadas, e posteriormente concatenadas por reação de ligação.

i) os concatâmeros foram clonados em vetor pZErO®-1 (item 3.5.2) , submetidos a PCR de colônia (3.5.3), e posteriormente seqüenciados (item 3.5.4).

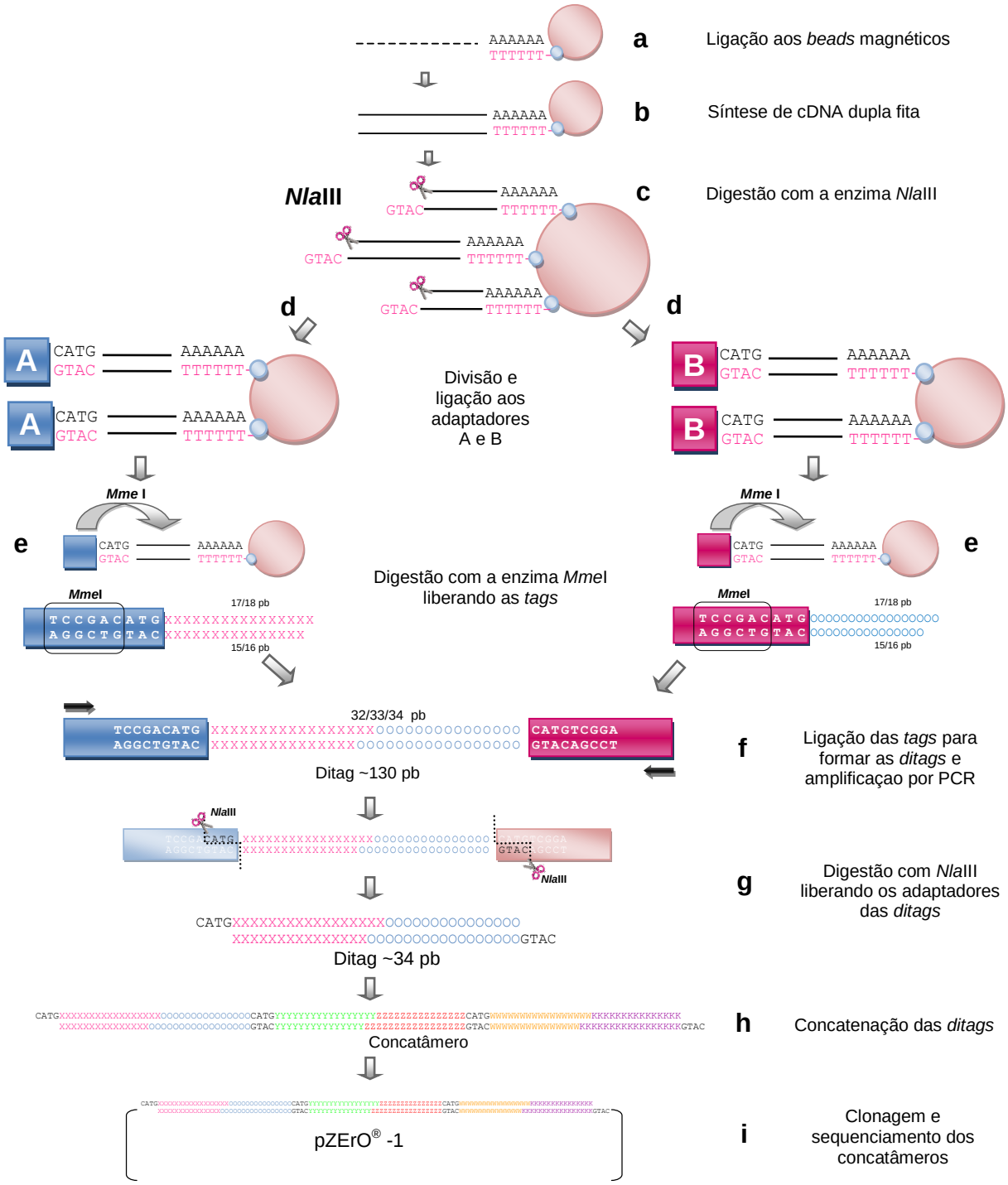


Figura 8 – Representação esquemática da construção das bibliotecas de LongSAGE. Fonte: Adaptado de Velculescu *et al.* (1995).

Para a construção das bibliotecas de LongSAGE poucas modificações foram realizadas em relação ao protocolo original do kit. , Ao invés de 100 ng conforme indicado pelo kit, para a construção das bibliotecas, foi utilizado cerca de 2.000 ng de RNAm por biblioteca. A quantidade de RNAm foi modificada pois com 100ng não obtivemos material suficiente para a etapa de clonagem e sequenciamento dos concatâmeros. Os tempos de incubação, digestão, centrifugação entre outros, foram realizados empregando-se os tempos mais longos recomendados pelo kit. Para as etapas de digestão do cDNA dupla fita com a enzima de restrição *Nla*III e para a liberação dos adaptadores, as amostras foram incubadas respectivamente, por 3 horas (ao invés de 1 hora) e por 12 horas (ao invés de 3 horas). Para a linearização do vetor pZero, 50 mM de espermidina foi acrescentado à reação de digestão.

3.5.1 Verificação da síntese e da digestão do cDNA com enzima de restrição *Nla*III

O kit comercial I-SAGE Long (Invitrogen), foi originalmente concebido para análise seriada da expressão gênica em humanos. Como o kit foi empregado para a construção de bibliotecas de *Eimeria tenella,* para verificação da síntese do cDNA e digestão do cDNA com a enzima *Nla*III (etapas controle), *primers* específicos para esta espécie foram desenhados.

Para a verificação da síntese de cDNA, foi desenhado um par de *primers* a partir de uma sequência de cDNA *full lenght* (AY508221.1) do gene EtCRK2, expresso em merozoítos e esporozoítos (Kinnaird *et al.*, 2004). Esta seqüência é desprovida de sítios de restrição para enzima *Nla*III, permitindo a amplificação deste produto a partir de cDNA digerido ou não com a enzima *Nla*III.

Para a verificação da etapa de digestão com a enzima *Nla*III, um segundo par de *primers* foi desenhado de forma a conter um sítio de restrição para esta enzima. Para tanto, foi utilizada uma seqüência de *Eimeria tenella* (Contig Eten_3204, montagem E_tenella_ORESTES/EST_assembled cDNAs – USP_v.2008_10-02 - http://www.coccidia.icb.usp.br), que tem alta similaridade com o gene de GAPDH de *Toxoplasma gondii* (AF265361). O fragmento amplificado foi escolhido de forma a

conter o primeiro sítio de restrição para a enzima *Nla*III mais próximo à cauda poliA, permitindo avaliar indiretamente a etapa de digestão.

Os *primers* foram desenhados utilizando-se o programa Oligo – Primer Analysis Software 4.0 (autoria de Wojciech Rychlik). As reações de amplificação foram feitas seguindo o protocolo do kit, utilizando-se os primers da Tabela 1.

Tabela 1 - *Primers* utilizados para verificação da síntese e digestão do cDNA com enzima de restrição *Nla*III.

Nome do *primer*	**Sequência do *primer***
EtCRK2 F	TCCTCAAAGAACTCCACCACCC
R	CCCTAAACCCTAAACCCTTGCG
GAPDH F	AAGCTTACTGGGATGGCTTTCCG
R	AGTTGCCGTCTTTCTGGGACATG

3.5.2 Clonagem dos concatâmeros e seleção

Os concatâmeros foram clonados em vetor pZErO®-1, de acordo com as instruções do fabricante. Este vetor permite a seleção direta por conter gene letal *ccdB* na extremidade C-terminal do LacZa. Além disso, contém gene de resistência para zeocina e sítio de pareamento para os primers M13F e M13R.

O vetor foi inserido por eletroporação em bactérias eletrocompetentes Shot® TOP10 eletrocomp™ *E coli*, utilizando o eletroporador Gene Pulser II™ (Bio-RAD), calibrado para um pulso de 2.500 V, com capacitância de 50 µF e resistência de 200 Ω (parâmetros ideais para a cubeta de 0,2 mm).

Após a transformação, o material foi semeado em placas de *petri* contendo meio LB agar com zeocina e incubado a 37 ºC durante 18 a 24 horas. As colônias isoladas foram picadas com palitos estéreis e inoculadas nos orifícios de microplacas de 96 cavidades contendo meio TB adicionado de zeocina. Em seguida, as bactérias foram incubadas a 37 ºC por 14 horas e estocadas a -80 ^{o}C com adição de glicerol 30% estéril.

3.5.3 PCR de colônia

Para a obtenção dos DNA moldes, os insertos presentes nas bactérias transformadas foram amplificados por reação de PCR de colônia. Esta técnica se baseia no principio de que a temperatura inicial de desnaturação da reação de amplificação é suficientemente alta para promover a lise da bactéria e conseqüente liberação do DNA. As reações de amplificação foram realizadas em um volume final de 15 µL, utilizando-se 1 µL de cada suspensão bacteriana transferida das microplacas de cultura com o auxílio de um replicador de 96 agulhas. Para cada reação de PCR foram utilizados 2,3 µM de cada primer M13 forward (TGTAAAACGACGGCCAGT) e M13 reverse (CAGGAAACAGCTATGACC), 125 µM de dNTP, 1 U de enzima BIOLASE^{TM} DNA polimerase (Bioline) e tampão da enzima 1x. As condições de amplificação consistiram em uma desnaturação inicial de 95 ºC por 4 minutos; 40 ciclos de 95 ºC por 45 segundos, 55 ºC por 45 segundos e 72 ºC por 1 minuto; e uma etapa final de extensão a 72 ºC por 5 minutos. Para verificação da amplificação, 3µL dos produtos de PCR foram submetidos à eletroforese em gel de agarose 1,2%.

3.5.4 Sequenciamento de DNA

A reação de sequenciamento foi realizada pelo método de terminação de cadeia de Sanger (Sanger *et al.*, 1977), utilizando-se o ABI PRISM BigDye Terminator Cycle Sequencing Ready Reaction Kit *version* 3. Para a realização do seqüenciamento, o DNA resultante do PCR de colônia foi diluído 1:10 em água.

As reações de sequenciamento foram realizadas em um volume final de 20 µL, utilizando-se 1 µL de cada suspensão bacteriana transferida das microplacas de cultura com o auxílio de um replicador de 96 agulhas. Para cada reação de PCR foram utilizados 30 µM de primer M13, 2 µL de ABI BigDye Terminator e tampão da enzima 1x. A condição de reação empregada foi de 40 ciclos de 96 ºC por 10 segundos, 52 ºC por 20 segundos e 60 ºC por 4 minutos. Após a reação, os

terminadores não incorporados foram removidos por precipitação diferencial do DNA. Em seguida, as amostras foram submetidas à eletroforese em um seqüenciador automático modelo ABI 3700 de 96 capilares (Applied Biosystems).

3.6 Pré-processamento das seqüências

Os cromatogramas gerados pelo seqüenciador de DNA foram submetidos a um processamento seriado de múltiplas etapas ("*pipeline*") no sistema EGene (Durham *et al.*, 2005), representado graficamente na Figura 9. Basicamente, o processamento consistiu das seguintes etapas:

a) **atribuição de bases ("base calling") e avaliação da qualidade.** Para essa etapa utilizou-se o componente `upload_traces.pl` do EGene, acoplado ao programa Phred (Ewing e Green, 1998; Ewing *et al.*, 1998);
b) **mascaramento de bases de vetor**. Utilizou-se o componente do EGene `mask_cross_match.pl` acoplado ao programa Cross_match (Phil Green, não publicado) do pacote Phred/Phrap/Consed, para o mascaramento contra a seqüência nucleotídica do vetor pZero-1 (Invitrogen) utilizado nas construções das bibliotecas de LongSAGE;
c) **mascaramento das seqüências dos adaptadores**. Foi feito o mascaramento contra os dois adaptadores utilizados na construção das bibliotecas, utilizando-se o componente `mask_cross_match.pl` do EGene e o programa Cross_match;
d) **mascaramento de seqüências de concatâmeros similares**. Os concatâmeros foram comparadas todos contra todos, e as seqüências comuns de comprimento equivalente a duas *ditags* (84 pb) foram mascaradas. Esta etapa tem o objetivo de eliminar clones duplicados derivados da replicação de clones bacterianos durante a etapa de cultivo das bactérias no meio SOC (protocolo de transformação bacteriana) logo antes do plaqueamento (Dinel *et al.*, 2005). Para tanto, foi empregado o componente

`mask_cross_match_concatamers.pl` do EGene, o qual trabalha de forma conjunta com o programa Cross_match;

e) **filtragem por qualidade**. Nesta etapa procurou-se eliminar todas as seqüências que não atendiam critério de qualidade Phred. Para ser aceita, uma leitura deveria apresentar pelo menos uma seqüência contínua de 42 pb com 100% das bases com qualidade Phred igual ou maior do que 30. Para isso, utilizou-se uma abordagem de janela deslizante de 42 pb, equivalente ao tamanho de uma *ditag* (incluindo o sítio de pontuação CATG 5' terminal). O racional desse critério é que cada leitura deveria apresentar um trecho de seqüência de alta qualidade com comprimento equivalente a pelo menos o de uma *ditag* de LongSAGE. O programa utilizado para essa etapa foi o `filter_quality.pl` do sistema EGene;

f) **aparamento das pontas**. Após a filtragem por qualidade, as seqüências foram submetidas ao programa `trimming.pl` do EGene para remover os trechos de seqüência com qualidade inferior ao conjunto de parâmetros definidos. O processamento é realizado em duas etapas. No primeiro passo usa-se uma janela deslizante de 42 pb. Dentro dessa janela, a seqüência, para ser aceita, tem que apresentar todas as bases com índice Phred igual ou maior do que 32. No segundo passo, uma janela de verificação de 21 pb faz a varredura da seqüência e, toda vez que a mesma encontra um trecho que não atende o critério de 100% de bases com Phred igual ou superior a 32, a seqüência 5' é salva e a janela prossegue até encontrar um novo trecho de seqüência que atenda o critério. Desta forma, ao final do processo, obtém-se um ou múltiplos trechos de seqüências que apresentam 100% das bases com esta qualidade;

g) **filtragem por tamanho.** Após o processamento e aparamento de pontas, todas as seqüências foram submetidas ao programa `filter_size.pl` do EGene. Somente foram aceitas seqüências que continham pelo menos 42 pb.

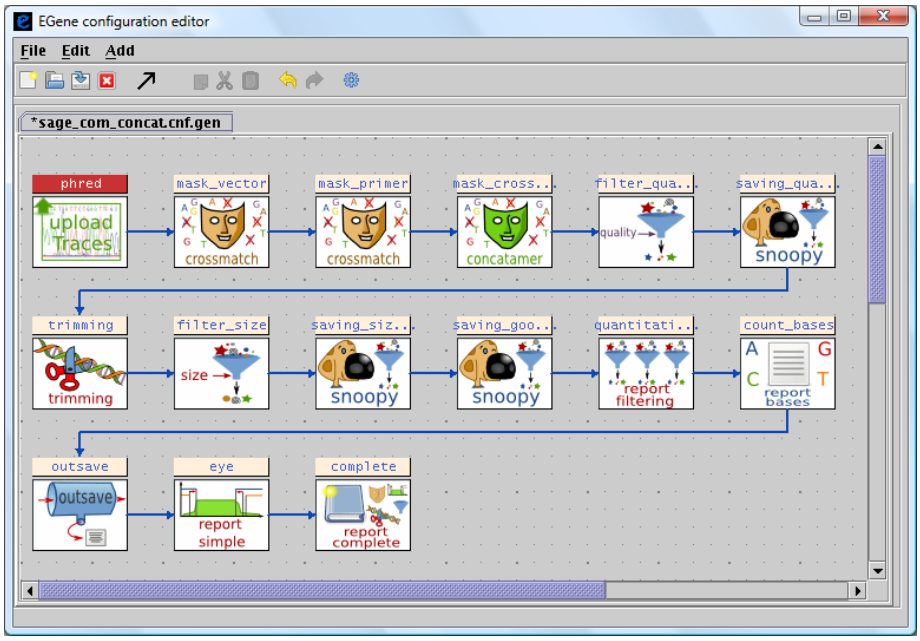


Figura 9 - Captura de tela do editor gráfico (CoEd) do sistema EGene, mostrando o *pipeline* de pré-processamento dos cromatogramas gerados pelo seqüenciador de DNA. Os ícones representam os componentes do *pipeline*.

3.7 Extração e contagem das *tags* de LongSAGE

As etapas de extração e contagem de *tags* foram realizadas com um pacote de programas desenvolvido pelo nosso grupo de bioinformática, intitulado SAGE Analysis. Este programa foi desenvolvido de forma genérica, possibilitando a extração de *tags* de SAGE convencional, LongSAGE, MPSS, entre outras. O pacote ainda possui uma versão gráfica compatível com os sistemas operacionais Linux, Windows e Macintosh. O processo de extração das *tags* consistiu em duas etapas, como descrito a seguir:

a) extração das *ditags*. Nesta etapa, o programa utilizou um arquivo de múltiplas seqüências em formato FASTA (concatâmeros) (Figura 10A) e, com base na existência dos sítios CATG (da enzima de ancoragem *Nla*III), determinou o posicionamento das *ditags* e sua extração Para a técnica de SAGE convencional (Velculescu *et al.*, 1995), utiliza-se como enzima de *tagging* a endonuclease de restrição de tipo IIS *Bsm*FI, que cliva o DNA 10 pares de bases *downstream* ao sítio de reconhecimento, resultando em uma *tag* de 14 pb (incluído o CATG). No caso do LongSAGE (Saha *et al.*, 2002), emprega-se a enzima *Mme*I, que cliva o DNA gerando uma extremidade coesiva, com sítios de corte localizados a 21 e 19 bases *downstream* ao sítio de reconhecimento. A enzima, entretanto, muito freqüentemente também apresenta clivagens nos sítios 20 e 18 bases *downstream* ao sítio de reconhecimento (Emmersen *et al.*, 2007; Harbers e Carninci, 2005). Para efeito de cálculo do tamanho das respectivas *tags* geradas, quando o corte é feito nas posições 21/19, obtém-se *tags* de 18 pb, enquanto que no corte nas posições 20/18 as *tags* resultantes têm 17 pb, sem incluir a seqüência CATG. Como esses cortes ocorrem simultaneamente no processo de digestão, ao se fazer a junção de *tags* para formação de *ditags* durante a construção das bibliotecas de LongSAGE, é comum se observar *ditags* apresentando 32 pb (duas *tags* de 17 pb subtraindo-se a sobreposição de duas bases das extremidades coesivas), 33 pb (uma *tag* de 17 pb e outra de 18 pb, menos a sobreposição) e, finalmente, *ditags* de 34 pb (duas *tags* de 18 pb). O programa foi concebido de forma a permitir a escolha do tamanho mínimo e máximo das *ditags,* no nosso caso foram escolhidos *ditags* de 32 a 34 pb. O programa possibilita ainda a extração das *tags* das pontas dos concatâmeros (*monotags*), permitindo uma recuperação maior do número de *tags*. Após a extração das *monotags*, a seqüência da *tag* foi completada com Xs, formando uma *ditag*. Assim, o programa gera uma lista de *ditags* a serem utilizadas para posterior extração das *tags*;

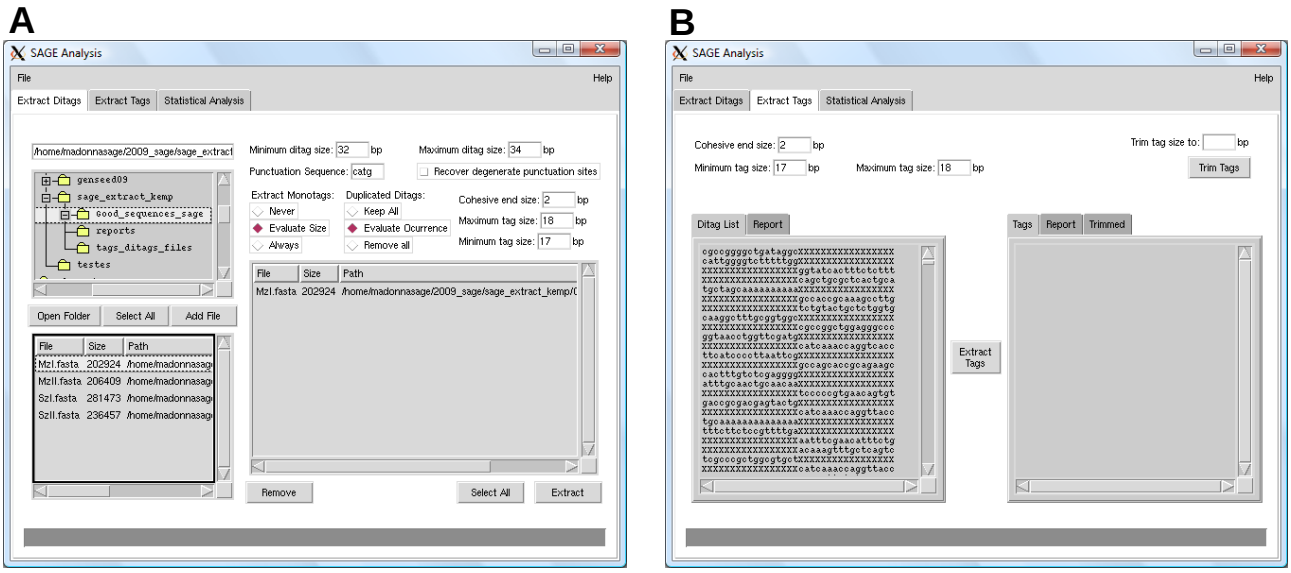


Figura 10 - Captura de tela da interface gráfica do programa SAGE Analysis. (A) Etapa de extração das *ditags* a partir de arquivo com a lista de concatâmeros. (B) Lista de *ditags* extraídas.

b) extração e contagem das *tags*. A partir da lista de *ditags* (Figura 10B), o programa gera uma lista de *tags* únicas e suas respectivas contagens (Figura 11a).

Figura 11 - Captura de tela da interface gráfica do programa SAGE Analysis, durante a etapa de extração das *tags*. (A) extração das *tags* (à direita) a partir de uma lista de *ditags* (à esquerda), (B) aparamento das *tags* de acordo com o tamanho selecionado.

Nesta etapa é possível definir quantas bases compõem a sobreposição entre *tags.* Por exemplo, na técnica de SAGE convencional, a enzima de *tagging Bsm*FI gera sítios de extremidades cegas e, portanto, as *tags* unidas para formar *ditags* não apresentam sobreposições. Já na técnica de LongSAGE (Saha *et al.*, 2002), a enzima *Mme*I gera extremidades coesivas de duas bases, de tal forma que as *tags* ligadas entre si em cada *ditag* apresentam duas bases de sobreposição. Assim, o programa permite estabelecer através de um parâmetro o tamanho esperado da sobreposição (valor zero se não houver). É possível também definir o tamanho máximo e mínimo das *tags* a serem extraídas e quando necessário, realizar o aparamento das extremidades das *tags* para um tamanho definido pelo usuário (Figura 11B), ferramenta esta importante para análise estatística das *tags*.

3.8 Mapeamento de *tags* contra uma base de *tags* genômicas virtuais

O conjunto de *tags* únicas obtidas a partir das bibliotecas de SAGE foi submetido a um mapeamento global contra *tags* virtuais extraídas a partir da montagem do genoma de *E. tenella* cepa H (versão de 08/05/2007), (ftp://ftp.sanger.ac.uk/pub/pathogens/Eimeria/tenella/genome/assemblies/assembly_2007_05_08.gz), disponível no sítio do Sanger Institute. Para tanto, foi utilizado um conjunto de *scripts* em linguagem Perl que gera um banco de *tags* virtuais hipotéticas de 17 pares de bases *downstream e upstream* ao sítio da enzima de restrição *Nla*III (CATG). Todos os sítios de restrição da enzima encontrados no genoma foram utilizados. As *tags* experimentais foram comparadas ao banco de *tags* virtuais, mapeadas e classificadas de acordo com sua freqüência.

3.9 Mapeamento das *tags* contra o genoma mitocondrial de *E. tenella*

O conjunto de *tags* únicas, obtidas a partir das bibliotecas de SAGE, foram mapeadas contra o genoma de mitocôndria de *E. tenella.* Para tanto, buscas de similaridade por BLASTN utilizando parâmetro *Wordsize* 21 foram realizadas.

Paralelamente, as *tags* foram reconstruídas a partir de leituras do genoma de *E. tenella*, utilizando-se o programa Genseed (Sobreira e Gruber, 2008). Uma vez reconstruídas, as seqüências dos cDNAs foram submetidas a um *pipeline* de anotação automática (ver item 3.11). Estes dados não foram mostrados no presente trabalho, mas foram utilizados para a caracterização das *tags* encontradas no genoma mitocondrial.

3.10 Análise estatística das bibliotecas de LongSAGE

3.10.1 Análises pelo programa Kemp

Os dados foram submetidos a uma ferramenta de análise estatística denominada Kemp desenvolvida por Varuzza *et al.*, 2008 (em fase de elaboração)[1] que é baseada em um teste frequentista, desenvolvido para a comparação de perfis de expressão digital. Esta ferramenta foi acoplada ao programa SAGE Analysis, permitindo desta forma, a extração, contagem e análise estatística das *tags* em um único pacote (Figura 12).

Para classificar se uma determinada *tag* é expressa diferentemente em um conjunto de bibliotecas empregou-se o *p-value* (p) que representa uma medida numérica da coerência dos dados observados em relação à hipótese nula (Cox, 1975; Kempthorn, 1976). Neste teste os valores de p foram calculados utilizando como hipótese nula a não diferenciação das *tags.*

[1] Varuzza L, Gruber A, Pereira, CAB. Significance tests for comparing digital gene expression profile, 2008 (em fase de elaboração). Nature Precedings: hdl:10101/npre.2008.2002.3).

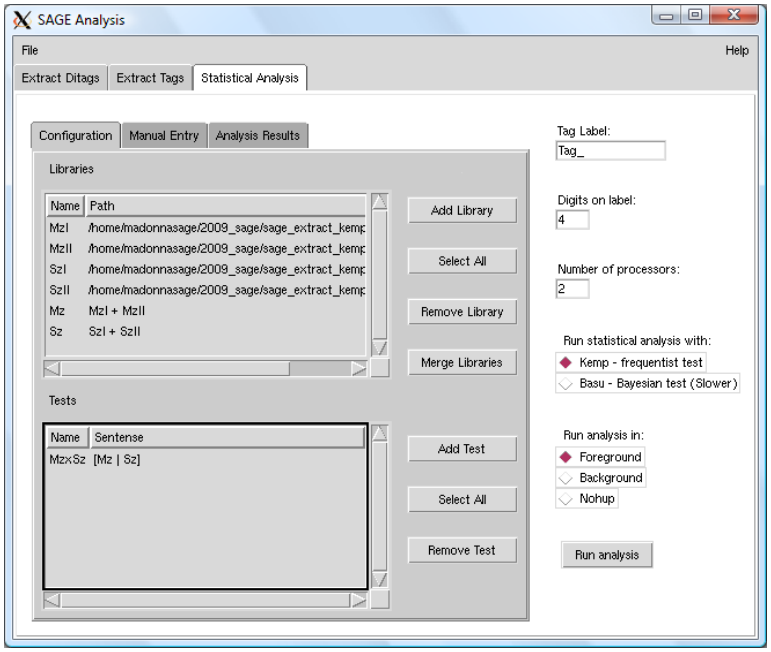


Figura 12 - Captura de tela da interface gráfica do programa SAGE Analysis, na etapa de análise estatística.

Além do cálculo dos valores de p, o Kemp calcula automaticamente o nível crítico (α) para cada *tag* em particular. Este valor é utilizado para se comparar os valores de p, e assim, aceitar ou rejeitar a hipótese nula. A escolha do nível crítico pode gerar dois tipos de erro:

Erro do tipo I: Rejeitar a hipótese nula quando ela é verdadeira. Este erro também é chamado de falso positivo;

Erro do tipo II: Aceitar a hipótese nula quando ela é falsa. Também chamado de falso negativo.

Para tanto, o programa utiliza um procedimento de decisão teórica que minimize os erros do tipo I (α') e do tipo II (β) da função aα' + bβ. Neste caso foi escolhido a=4 e b=1, o que caracteriza um nível crítico mais estringente em relação à escolha das *tags* diferencialmente expressas, minimizando o número de resultados falso positivos, pois o valor de α', que considera a hipótese nula verdadeira (não diferenciação das *tags*) possui um peso maior do que o valor de β, onde a hipótese nula é considerada falsa (diferenciação das *tags*).

O programa Kemp gera um arquivo de saída do tipo csv (*comma-separated values*), ou seja, de valores separados por vírgulas, o qual pode ser importado em qualquer programa de planilhas. O arquivo de saída consiste em uma lista na qual cada linha do arquivo contém as informações relativas a uma *tag* distinta. Assim, cada linha possui um identificador da *tag*, a seqüência nucleotídica, a freqüência (contagem) da *tag* em cada biblioteca, os valores de p (valor de significância), de α (nível crítico), o *score* da *tag* (10 x (1-*pvalue*/α), e a categoria (U para não diferencial e D para diferencial). O *score* serve como um parâmetro de ordenação das *tags*, ou seja, quanto maior *score*, mais diferencialmente expressa é a *tag.* A categoria foi decidida de acordo com o *p-value* e o valor de α. Se o *p-value* for menor do que α, a *tag* é classificada como diferencialmente expressa. A Figura 13 mostra um exemplo de algumas linhas da saída do programa Kemp.

```
tag name,sequence,MzI,MzII,SzI,SzII,MzxSz:pvalue,alpha,score,category
Tag_0001,aaaaaaaaaacaaaaaa,0,0,1,0,1,0.0629925,0,U
Tag_0002,aaaaaaaataaaaaaag,0,0,1,0,1,0.0629925,0,U
Tag_0003,aaaaaaaattggcttgg,0,0,1,0,1,0.0629925,0,U
Tag_0004,aaaaaaacaaaaaaaaa,1,0,0,0,0.423842,0.0629925,0,U
Tag_0005,aaaaaaaggtcaaagaa,0,1,0,0,0.424371,0.0629925,0,U
Tag_0006,aaaaacaaaaaaaaaac,0,1,0,0,0.424371,0.0629925,0,U
Tag_0007,aaaaactggaccactgt,0,0,0,1,1,0.0629925,0,U
Tag_0008,aaaaacttaatagaagg,1,0,0,0,0.423842,0.0629925,0,U
Tag_0009,aaaaagaaactggaagc,4,0,0,1,0.170166,0.030645,0,U
Tag_0010,aaaaagagaaaatgaag,0,0,0,1,1,0.0629925,0,U
Tag_0011,aaaaagagatcttttat,0,0,0,1,1,0.0629925,0,U
Tag_0012,aaaaagaggcaaaaacc,1,0,1,0,1,0.0459066,0,U
Tag_0013,aaaaagcccagtatgaa,0,1,1,0,1,0.0459066,0,U
Tag_0014,aaaaagctaaacggaga,0,5,0,0,0.01372,0.030645,5.52293,D
Tag_0015,aaaaaggagacagaaaa,1,0,0,0,0.424371,0.0629925,0,U
```

Figura 13 – Saída em modo texto no formato CVS (valores separados por vírgulas) gerado pelo programa Kemp. A primeira linha representa um cabeçalho contendo informações relativas a cada um dos valores: nome da *tag*, seqüência da *tag*, con*tag*ens absolutas nas bibliotecas MzI, MzII, SzI e SzII, valores de p da comparação entre soma das bibliotecas Mz contra a soma das bibliotecas Sz, valor de α, *score* e categoria.Cada linha apresenta os dados relativos a uma *tag* específica.

Visando identificar as *tags* de expressão diferencial, os arquivos de contagens de *tags* das bibliotecas de formas merozoítas (MzI e MzII) foram comparados com os de esporozoítos (SzI e SzII).

3.10.2 Análise dos dados no Microsoft Excel

O arquivo produzido pelo `kemp_shell.pl`, formato CSV foi exportado para o programa Excel. Neste programa, os valores de contagens absolutas de *tags* foram normalizados, uma vez que cada biblioteca possuía um número total de *tags* distinto. Para a normalização, a con*tag*em de cada *tag* foi dividida pela con*tag*em total de *tags* de sua biblioteca e multiplicada pela con*tag*em total da biblioteca de menor representatividade numérica. Vamos assumir como exemplo, que a *Tag*_0235 obteve uma con*tag*em de 100 na biblioteca A e 150 na biblioteca B, e que as contagens totais das duas bibliotecas foram de respectivamente 5.000 e 6.000. Assim, a contagem normalizada da *Tag*_0235 na biblioteca A foi de [(100/5.000) x 5.000], e na biblioteca B de [(150/6.000) x 5.000], o que corresponde respectivamente a 100 e 125. A normalização visa facilitar a comparação visual de con*tag*ens a partir de bibliotecas de representatividades distintas. Após a normalização, procedeu-se ao cálculo do que denominamos de “poder de expressão” (PE), que corresponde para cada *tag* individual, à soma das contagens das *tags* em todas as bibliotecas. Cálculos referentes à diferença de expressão entre as bibliotecas também foram realizados.

3.11 Reconstrução dos cDNAs das *tags* diferencialmente expressas

As *tags* de 21 pb foram submetidas a um programa de reconstrução de seqüências denominado GenSeed desenvolvido em nosso laboratório (Sobreira e Gruber, 2008). Este programa utiliza seqüências denominadas sementes para realizar uma montagem progressiva através de múltiplos ciclos de busca de similaridade, recuperação de seqüências e montagem. Como base de dados, foi utilizado um conjunto de aproximadamente 48.000 leituras de cDNAs de ESTs de *E. tenella*, incluindo cerca de 14.000 ORESTES gerados pelo nosso grupo. Uma vez reconstruídas, as seqüências dos cDNAs foram submetidas a um *pipeline* de anotação automática.

3.12 Anotação automática

O nosso grupo de pesquisa em Bioinformática desenvolveu um conjunto de componentes do sistema EGene (Durham *et al.*, 2005) para a anotação automática de seqüências (Ferro, 2008). A partir daí, foi construído um *pipeline* de anotação automática dos transcritos consistindo nas seguintes etapas (Figura 14):

a) **busca de similaridade de *tags*.** O objetivo desta etapa, foi de verificar se *tags* diferentes à *tag* utilizada para reconstrução dos transcritos, estavam presentes nestas seqüências. Para tanto, utilizou-se o componente `annotation_blast.pl` do EGene acoplado ao programa BLAST (Altschul *et al.*, 1997), utilizando-se o BLASTN e como parâmetro na busca de similaridade *Wordsize* 21. Foram considerados alinhamentos com *e-values* inferiores 1e-03 e porcentagem de identidade igual a 100%;
b) **determinação das fases de leitura abertas (ORFs) e tradução protéica.** Para essa etapa utilizou-se o componente `annotation_orf` do EGene. Foram escolhidas as fases de leitura abertas codificando pelo menos 50 resíduos de aminoácidos, sem requerimento de códon de iniciação. Todas as ORFs foram traduzidas conceitualmente utilizando-se o código genético universal e as seqüências nucleotídicas e protéicas correspondentes foram armazenadas em relatórios;
c) **busca de similaridade.** Para as buscas de similaridade foi utilizado o componente `annotation_blast.pl` do EGene, acoplado ao programa BLAST (Altschul *et al.*, 1997). Como base de dados utilizou-se o nr do GenBank. Os cinco melhores alinhamentos com *e-values* inferiores 1e-04 e mínima porcentagem de identidade igual a 45% foram anotados. Além disso, os resultados das buscas foram armazenados em arquivos texto assim como em arquivos HTML, os quais podem ser visualizados utilizando qualquer navegador web;
d) **busca de domínios conservados.** As seqüências protéicas geradas pelo componente `annotation_orf.pl` foram submetidas a uma busca de domínios conservados contra a base de dados CDD (Marchler-Bauer *et al.*,

2007). Para isso, utilizou-se o componente `annotation_rpsblast.pl` acoplado ao programa RPS-BLAST (Marchler-Bauer *et al.*, 2002). O RPS-BLAST é um programa que compara uma seqüência protéica contra uma base de dados de matrizes de escores posição-específica (*Position Specific Scoring Matrices* – PSSMs). Somente foram considerados os alinhamentos com *e-value*s inferiores a 5e-04. Os resultados completos foram armazenados em arquivos texto;

e) **busca de domínios protéicos.** As seqüências protéicas foram submetidas a uma busca de motivos protéicos utilizando-se o componente do EGene `annotation_interpro.pl` acoplado ao programa InterproScan (Mulder e Apweiler, 2007). O InterProScan é um programa que identifica assinaturas a partir de bases de dados pertencentes ao InterPro: Pfam, PROSITE, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D, e PANTHER. Todos os resultados foram armazenados em arquivos texto e HTML;

f) **busca de seqüências de peptídeo-sinal.** Para essa busca foram utilizados os componentes `annotation_signalP.pl` e `annotation_phobius.pl`, acoplados respectivamente aos programas SignalP (Bendtsen *et al.*, 2004) e Phobius (Kall *et al.*, 2004);

g) **busca de domínios transmembranares.** Nessa busca foram utilizados os componentes `annotation_tmhmm.pl` e `annotation_phobius.pl`, acoplados respectivamente aos programas TMHMM (Krogh *et al.*, 2001) e Phobius (Kall *et al.*, 2004);

h) **busca de seqüências de sítios de âncoras GPI.** Nessa etapa utilizou-se o componente `annotation_dgpi.pl` acoplado ao programa DGPI (Kronegg e Buloz, 1999; http://129.194.185.165/dgpi/);

i) **busca de seqüências repetitivas seriadas**. Foi utilizado o componente `annotation_trf.pl` acoplado ao programa Tandem Repeats Finder (Benson, 1999), que realiza a busca por repetições seriadas em seqüências;

j) **seleção das ORFs anotadas.** Foi utilizado o componente `report_ORF.pl` para selecionar as ORFs com maior número de evidências. O conjunto de evidências a ser considerado para a seleção das ORF é livremente definido pelo usuário e nesse caso incluiu resultados de similaridade por BLAST,

considerando resultados com melhor *e-value*, presença de domínios conservados detectados pelo RPS-BLAST e motivos protéicos encontrados pelo InterproScan. Quando ocorre um empate do número de evidências, o componente seleciona a ORF de maior comprimento;

k) geração de anotação automática em formato *Feature Table*. Foram utilizados os componentes do EGene `report_feature_table_submission.pl` e `report_feature_table_artemis.pl` para gerar arquivos de anotação completa em formato *feature table* no padrão de submissão de seqüências para bancos internacionais e na versão ampliada para visualização e edição no Artemis. Essa última versão contém algumas etiquetas que somente são válidas dentro do programa Artemis, como, por exemplo, a visualização de cor nos quadros que representam as ORFs;

l) geração de anotação automática em formato *Generic Feature Format*. Foi utilizado o componente do EGene `report_gff.pl` para gerar arquivos de anotação completa em formato GFF3, que está se tornando um formato universal de intercâmbio de anotações genômicas;

m) mapeamento dos termos GO (*Gene Ontology*). Uma vez concluído o *pipeline* de anotação automática, foi usado o componente do EGene `mapping_GO.pl` para realizar o mapeamento e quantificação dos termos GO nas três ontologias gênicas para cada seqüência anotada. O mapeamento foi originalmente realizado pelo próprio programa InterproScan, que utiliza uma tabela de conversão denominada interpro2go, e guarda os termos encontrados num arquivo de saída. O componente do EGene `mapping_GO.pl`, gera a partir desses termos um arquivo denominado *Gene Association File*, o qual é utilizado pelo *script* `map2slim.pl` para quantificar o número de seqüências com termos GO dentro do subconjunto de ontologias *GO Slim*. O componente gera no final um conjunto de arquivos HTML contendo uma tabela dos termos de *GO Slim*, o número de seqüências positivas para cada termo GO e os *links* para as respectivas seqüências nucleotídica e protéica. Também é gerada uma versão XML dessa saída, mas, nesse caso as ontologias são representadas de forma hierárquica, com uma árvore na qual os *links* podem ser expandidos ou colapsados. As seqüências

que não apresentaram resultados positivos de Interpro receberam números de GO referentes às três ontologias gênicas, já que não puderam ser classificadas.

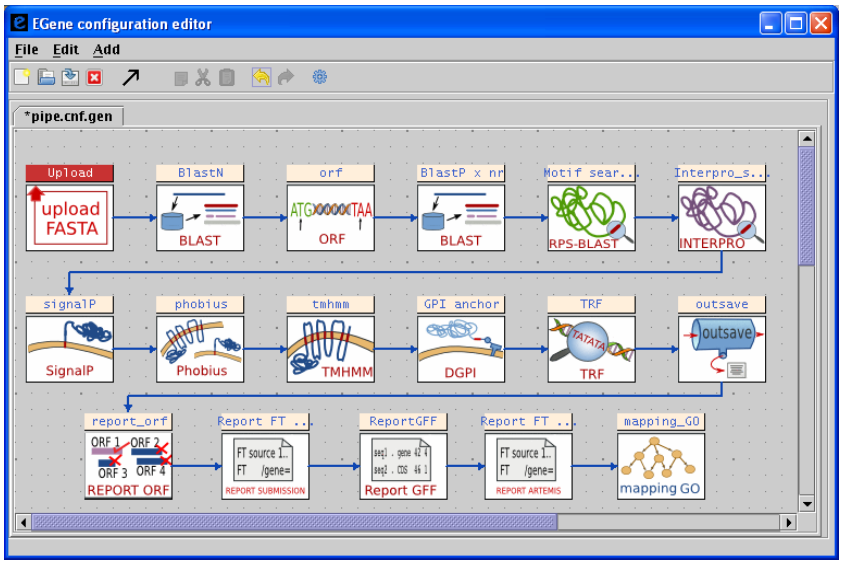


Figura 14 - Tela do editor gráfico (CoEd) do sistema EGene mostrando um *pipeline* com os componentes de anotação automática utilizados para o processamento de seqüências de cDNA (vide ícones na tela).

3.12.1 Anotação utilizando a base de dados KOG (*euKaryotic Orthologous Groups*)

Para a anotação dos produtos protéicos na base de dados KOG foi utilizado um *script* em Python desenvolvido pelo nosso grupo de bioinformática (`thibinitor`) que é uma modificação da versão *stand-alone* `dignitor` disponível no sítio (ftp://ftp.ncbi.nih.gov/pub/tatusov/dignitor/old/SGI/). Este *script* cria uma página HTML com os resultados da anotação, um gráfico de distribuição das categorias de KOG e links para os resultados individuais e alinhamentos. Para utilização deste *script* são necessários dois arquivos: o primeiro (utilizado nas buscas de similaridade por BLASTP) é denominado de kyva

(ftp://ftp.ncbi.nih.gov/pub/COG/KOG/), um arquivo do tipo FASTA contendo todas as proteínas (ou domínios) dos 7 genomas completos de eucariotos, e o segundo, denominado de euCogan (kog no FTP), que contém a informação de cada KOG, incluindo o nome, a classificação funcional (código de uma letra), a descrição funcional e a lista de proteínas (ou domínios).

3.12.2. Geração de páginas *web* com resultados da anotação

Foi utilizado o componente do EGene `report_html.pl` para gerar dinamicamente páginas *web* para a visualização e consulta dos resultados de anotação automática das seqüências reconstruídas a partir das *tags* diferencialmente expressas de *Eimeria tenella*.

3.12.3 Curadoria manual da anotação automática

Após a construção do protocolo de anotação, implementação na forma de um *pipeline* no sistema EGene e execução, procedeu-se a curadoria das anotações das seqüências. Basicamente, os dados gerados automaticamente, correspondentes aos arquivos de anotação na versão ampliada, foram abertos no programa Artemis e as seqüências foram analisadas quanto à presença de uma ou mais *tags* por transcrito, à posição e direção (*sense* ou *antisense*) da *tag* assim como a direção do transcrito.

Baseando-se na seqüência e nos alinhamentos obtidos por BLAST (quando positivo), foram verificados a presença de códons de iniciação e terminação, a correta identificação do produto protéico da anotação e a presença ou ausência de cauda poliA (Figura 15).

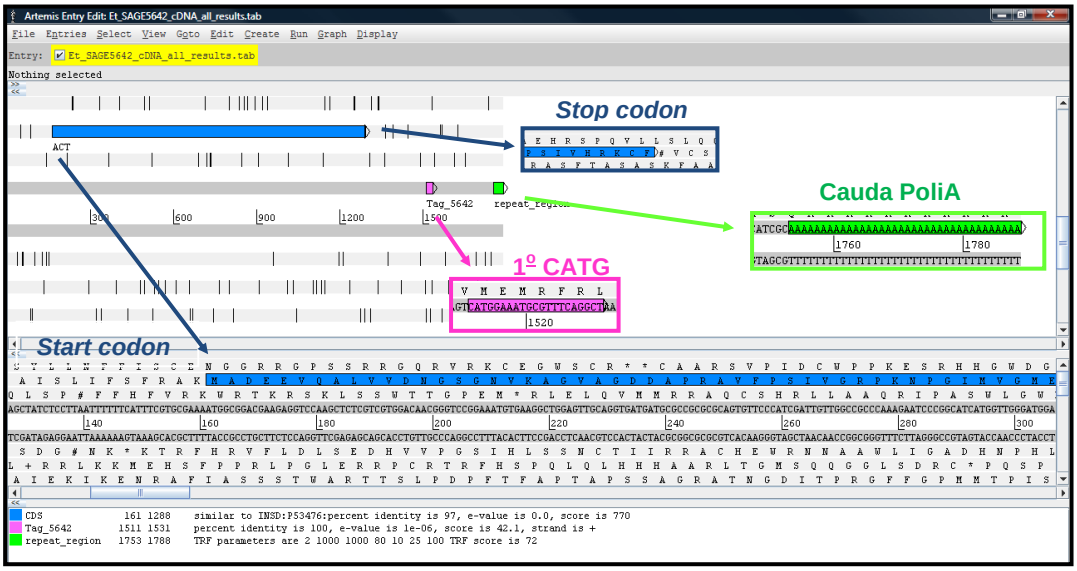


Figura 15 - Tela do programa Artemis mostrando um exemplo de anotação de uma sequência reconstruída no formato *feature table* estendido. Em destaque: a identificação de *start* e *stop codon,* o sítio de restrição CATG mais próximo à extremidade 3', a respectiva *tag* e a cauda PoliA.

3.13 PCR em tempo real

Experimentos de PCR em tempo real (qPCR – PCR quantitativo) foram executados com o objetivo de realizar uma validação preliminar dos resultados obtidos através da técnica de LongSAGE. A partir das seqüências reconstruídas (ver item 3.10), *primers* (Tabela 2) foram desenhados para um pequeno conjunto de genes com auxílio do programa Oligo – Primer Analysis Software 4.0 (autoria de Wojciech Rychlik), de forma criteriosa. Este programa, além do desenho dos *primers*, analisa características dos oligonucleotídeos quanto à Tm (temperatura de *melting*), complementaridade entre os *primers* (dímeros), falsos sítios de pareamento e formação de *hairpins*. O tamanho das seqüências a serem amplificadas foi limitado a 350 pares de bases.

Tabela 2 - *Primers* utilizados para a verificação de expressão gênica diferencial em experimentos de PCR em tempo real.

Nome do *primer* (*tag*)	Produto da Anotação	Seqüência do *primer*
0263_Mz F	40 S *Ribosomal Protein S8* (RPS8)	CGACAGAAGGCACAAGCACC
0263_Mz R		GCACTGGGCGGACTCTTTTG
1591_Mz F	*Actin Depolymerizing factor* (ADF)	AGCGGAATGCCAGTCAACG
1591_Mz R		ACGGTTCCCTCCAGTTTCTTC
2126_Sz F	*Putative Uncharacterized Protein* (PUP)	TGTTTAGGGGTGGCGATG
2126_Sz R		CAAACGAAAGTATGTCTGTGCAT
5173_Sz F	*Eukaryotic translation initiation factor 5*	CGGAGTTGCTGGAGGAGAA
5173_Sz R		GCGGGGGCTGTTGTAGTAAG
5642_Mz F	*Actin* (ACT)	CGAGCATCGTTCACCGCA
5642_Mz R		CTAGCCTCATCTTCCCACCAGG
5992_Sz F	*Glut/Leu/Phe/Val dehydrogenase family protein*	TCTTCGAGCAGCAAACGGG
5992_Sz R	(ELFV_dehydrog).	CGAACTCTTCCTCAGTCAGCGA

A extração de RNAm foi realizada de acordo com o item 3.4. Para quantificação deste material foi utilizado o fluorômetro Qubit™ (Invitrogen), o qual utiliza um fluoróforo que se torna fluorescente somente quando se liga especificamente à moléculas de RNA. Este tipo de medição é mais acurado que a medida da absorbância em luz UV, já que a quantificação reporta somente a concentração da molécula de interesse.

Para a síntese de cDNA utilizou-se em cada reação 10 ng de RNAm. Os ensaios de transcrição reversa foram realizados a partir de *primers* específicos, utilizando-se a enzima RevertAid™ H Minus M-MuLV (Fermentas), conforme instruções do fabricante. Para a detecção de amplificação em tempo real utilizou-se o termociclador “Realplex2/2” (Eppendorf).

As reações de qPCR foram feitas em triplicata, em um volume final de 20 µL por reação. O kit ABsolute™ QPCR SYBR Green Mix (Thermo Scientific), contendo dNTPs, enzima *Taq* polimerase (Thermo-Start) e cloreto de magnésio 3mM foi utilizado, acrescido de 10% da reação de cDNA (produto da transcrição reversa) e 0,3 mM de cada *primer*. Como controle negativo, para cada gene testado, foram realizadas reações sem a adição de cDNA. Para a avaliação da expressão gênica relativa, reações empregando o gene controle também foram realizadas.

A seleção de genes controle (Tabela 3) foi feita a partir de transcritos reconstruídos que apresentaram *p-value* igual a 1 (hipótese nula verdadeira), e que em buscas de similaridade por BLAST, tiveram resultados positivos. Além disso, para verificar se o perfil de expressão gênica constitutivo também era observado em espécies relacionadas, estes genes foram comparados com dados de expressão de ortólogos do Filo Apicomplexa.

Os experimentos de validação do gene controle, foram realizados empregando-se amostras de RNAs mensageiros de duas extrações independentes purificadas a partir de esporozoítos e merozoítos de segunda geração.

Tabela 3 - *Primers* testados em experimentos de PCR em tempo real para a escolha do gene controle (*housekeeping gene* constitutivamente expresso).

Nome do *primer* (*Tag)*	**Produto da Anotação**	**Seqüência do *primer***
0247H F 0247H R	*Ribosomal Protein 22* (RP22)	CCAGATGCCCCAGGACGA CTGTCTGCGAAAAGCCAAGTTA
1360H F 1360H R	*GTP-binding nuclear protein*	ACGTAACGAGTCGCATCACCTA AACTCCCTCTCCGCTTGCTG
7041H F 7041H R	*Histone H2A*	CCGCCAGAGACCACAAGAAAAC GCATGACGCCGCCAGAGG
7924H F 7924H R	*Ubiquitin*	CGTCCAGTTCAAGCGTTCGG ATCTGCATTGTAGCGCGGC
8019H F 8019H R	*Zinc Finger*	TTGCCACAAGTGCAACAAGA CATCGGGGTAGCGGTGC

Para o programa de ciclagem da reação foi empregada uma etapa inicial de 15 minutos a 95 ºC para ativação da *Taq* DNA polimerase. A amplificação foi feita em 35 ciclos consistindo cada um em 15 segundos de desnaturação a 95 ºC, 30 segundos para o anelamento a 55 ºC e 30 segundos de extensão a 72 ºC. Para a determinação da especificidade dos produtos de PCR e verificação da formação de dímeros de *primers* as curvas de dissociação (curvas de *melting*) dos produtos amplificados foram analisadas. A análise da temperatura de dissociação consistiu em um ciclo de 95 ºC por 30 segundos seguido de uma etapa a 60 ºC durante 30 segundos e um ciclo de 20 minutos de aquecimento gradual, partindo de 60 ºC ate

95 ºC. Os sinais de fluorescência foram capturados de forma constante. Todos os ensaios, bem como todas as amostras foram avaliados pelo menos duas vezes.

A análise dos dados foi realizada utilizando o software Realplex (Eppendorf). Para a determinação de um único $C_{t,}$ os dados de fluorescência foram coletados na fase exponencial da reação. Genericamente, o C_t (*Cycle Threshold*) corresponde ao ciclo de qPCR no qual a fluorescência emitida ultrapassa o *threshold,* ponto da reação onde um número suficiente de *amplicons* já pode ser detectado. O *threshold* foi automaticamente definido pelo software Realplex seguindo os parâmetros *default* do programa.

Para a quantificação da expressão, optamos pelo método de análise relativa, onde a expressão dos genes selecionados foi normalizada em relação a um gene controle. Para tanto, os valores de C_t foram exportados para o programa Q-Gene (Muller *et al.*, 2002). Neste programa, a eficiência de todas as reações foi considerada igual a 2 (100%) e para o cálculo de $2^{-\Delta}C_t$ foi utilizado o parâmetro 2 do programa (cálculo da média de expressão gênica normalizada utilizando concomitantemente os 3 valores de C_t referente às triplicatas). Em seguida, estes dados foram exportados para uma planilha Excel para cálculo da média dos valores de C_t obtidos nos diferentes experimentos (duplicatas ou triplicatas experimentais) e valores de $2^{-\Delta\Delta}C_t$ no formato padrão.

Para a comparação dos dados de expressão entre LongSAGE e qRT-PCR, os valores de $2^{-\Delta\Delta}C_t$ e as diferenças de expressão detectadas por LongSAGE foram transformados em valores de log na base (2). Para avaliar a concordância dos dados, o coeficiente de correlação de Pearson e o valor de R^2 (curva de regressão) foram calculados.

3.14 Análise comparativa de expressão gênica em merozoítos e esporozoítos de *E. tenella* obtida por LongSAGE e *Northern* digital (ESTs/ORESTES)

3.14.1 *Northern* Digital de ESTs/ORESTES de merozoítos de segunda geração e esporozoítos de *E. tenella*

Protocolos de anotação automática de sequências ORESTES de *Eimeria* spp. de galinha doméstica foram desenvolvidos pela equipe de bioinformática do nosso grupo e, aplicados aos nossos conjuntos de dados utilizando-se componentes recém-desenvolvidos para a plataforma EGene (Durham *et al.*, 2005; Ferro, 2008).

Para *E. tenella*, além dos dados de ORESTES, ESTs convencionais disponíveis publicamente também foram utilizados permitindo uma cobertura tanto das regiões centrais dos transcritos, obtidas pelos ORESTES, quanto das regiões terminais. Após o pré-processamento das seqüências no sistema EGene (Durham *et al.*, 2005), 48.361 seqüências de *E. tenella* (Tabela 4), provenientes de diferentes estágios evolutivos (merozoítos de primeira e segunda geração, esporozoítos, oocistos não esporulados, parcialmente esporulados e oocistos esporulados) foram submetidas à reconstrução (agrupamento e montagem de ESTs/ORESTES) utilizando-se o programa CAP3 (Huang e Madan, 1999). Os cDNAs reconstruídos foram submetidos ao protocolo de anotação automática (Ferro, 2008).

Tabela 4 - Origem, tipo de biblioteca e quantidade de leituras utilizadas para a *clusterização* das seqüências de *E. tenella*.

Origem	Tipo de biblioteca	Nº de leituras	Formato	Fonte
Sanger Institute, Reino Unido	ESTs convencionais	5.939	FASTA e SCF	Sanger e NCBI[1]
Universiti Kebangsaan Malaysia	ESTs convencionais	1.028	FASTA	NCBI[1]
Washington University, EUA	ESTs convencionais	26.249	Cromatogramas	WUSTL[2]
USDA[3], EUA	ESTs convencionais	1.022	Cromatogramas	USDA[3]
USP[4], Brasil	ORESTES	14.123	Cromatogramas	USP[4]
Total	-	**48.361**	-	-

[1] National Center of Biotechnology Information (http://www.ncbi.nlm.nih.gov/), Bethesda, EUA
[2] Washington University (http://www.washington.edu/), Seatlle, EUA
[3] United States Department of Agriculture (http://www.usda.gov/wps/portal/usdahome), Baltimore, EUA
[4] Universidade de São Paulo, São Paulo, Brasil

Considerando que cada cDNA reconstruído pode estar representado por seqüências de diferentes estágios evolutivos, a estimativa da freqüência de leituras de cada estágio pode ser estimada permitindo a identificação de genes diferencialmente expressos através de uma abordagem denominada *Northern* eletrônico ou digital. O *Northern* eletrônico foi realizado a partir das contagens das leituras presentes nos cDNAs reconstruídos.

A partir do arquivo de montagem dos cDNAs, a contagem do número de leituras respectivas a cada estágio evolutivo que compõe cada sequência foi realizada empregando-se um *script* em Perl. Os dados foram submetidos à ferramenta de análise estatística Kemp (ver item 3.8.1). Em seguida, os genes diferencialmente expressos foram identificados.

3.14.2 Comparação dos dados de expressão diferencial obtida por LongSAGE e *Northern* digital

Os genes provenientes de formas merozoítas e esporozoítas identificados nas técnicas de LongSAGE e *Northern* digital dos ESTs/ORESTES como sendo de expressão diferencial foram comparados entre si, através de buscas de similaridade

por BLASTN (1e-40) e análise dos dados individualmente, observando os alinhamentos e os resultados da anotação automática.

3.14.3 Análise comparativa entre os transcritos mais expressos em LongSAGE e *Northern* digital (ESTs/ORESTES)

Os genes provenientes de formas merozoítas e esporozoítas identificados nas técnicas de SAGE e *Northern* digital como sendo de expressão maior (numericamente mais representativos) foram comparados entre si. A partir do conjunto de genes mais expressos por LongSAGE foi avaliado quantos desses genes confirmaram o caráter de maior expressão por *Northern* digital.

Em ambas as plataformas foram selecionadas as 100 sequências numericamente mais representativas em cada estágio evolutivo (merozoítos de segunda geração e esporozoítos). A comparação foi feita através de buscas de similaridade por BLASTN (1e-40) e análise dos dados individualmente, observando os alinhamentos e os resultados da anotação automática.

4 RESULTADOS

4.1 Propagação dos parasitas

Oocistos: após a infecção e purificação dos oocistos foi obtido um rendimento médio de $3x10^7$ oocistos/ave. A taxa de esporulação média foi de 85-90%.

Esporozoítos: após a quebra parcial dos oocistos, excistação *in vitro* e purificação dos esporozoítos em coluna híbrida de lã de nylon (filtração) e resina DE-52 (cromatografia por troca-iônica), o número de esporozoítos obtidos foi o equivalente a 5-10% do número inicial de esporozoítos. Dados semelhantes foram também obtidos por Tomley, 2004 (comunicação pessoal)[2]. Para a construção de cada uma das bibliotecas de LongSAGE, cerca de 10^9 esporozoítos foram utilizados.

Merozoítos de segunda geração: o rendimento médio obtido nas infecções foi de 10^8 merozoítos/ave. Para a construção de cada uma das bibliotecas de LongSAGE, cerca de 10^9 merozoítos de segunda geração foram utilizados.

4.2 Verificação de contaminação inter-específica

O DNA dos oocistos utilizados para infecção das aves, assim como o DNA dos oocistos recém purificados e utilizados em todos os experimentos, foram submetidos a um ensaio de PCR multiplex capaz de detectar simultaneamente as sete espécies de *Eimeria* que infectam a galinha doméstica.

[2]Tomley F. São Paulo; 2004 (comunicação pessoal).

Em todos os ensaios de PCR realizados, apenas uma única banda de 539 pares de bases compatível com o tamanho esperado para o marcador SCAR específico de *E. tenella* (Fernandez *et al.*, 2003) foi observada, conforme exemplificado na Figura 16. O fato de não terem sido observadas bandas de outros tamanhos compatíveis com as demais espécies de *Eimeria* é indicativo de que não ocorreu contaminação inter-específica das amostras de *Eimeria tenella* utilizadas.

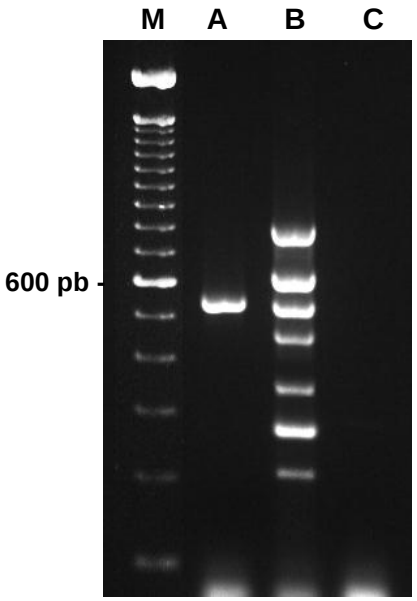


Figura 16 - Resultado do ensaio de PCR *multiplex* para detecção das sete espécies de *Eimeria* de galinha doméstica. (A) DNA de oocistos de *E. tenella* (B) amostra mista de DNA contendo as sete espécies de *Eimeria*, (C) controle sem adição de DNA (M) Marcador de peso molecular do tipo escada de 100 pb.

4.3 Avaliação da qualidade das amostras de RNA mensageiro

Para a construção de bibliotecas de LongSAGE, assim como, para os ensaios de qPCR, foram utilizadas amostras de RNA mensageiro de alta qualidade e livre de

contaminantes. Conforme exemplificado na Figura 17, os RNAs mensageiros purificados e digeridos com DNAse apresentaram-se íntegros e livres de bandas ribossômicas.

Após a digestão com DNAse, as amostras de RNA mensageiro foram também submetidas a uma reação de PCR capaz de detectar a presença de contaminação com DNA genômico de *E. tenella*. Em cada uma das amostras utilizadas para o desenvolvimento deste trabalho, não foi possível observar a presença de banda compatível com o fragmento de 539 pb, do marcador SCAR espécie-específico para DNA de *E. tenella* (Fernandez *et al.*, 2003), conforme exemplificado na Figura 18. Estes resultados indicam que as amostras de RNA mensageiro utilizadas, apresentaram-se livres de contaminação com DNA genômico.

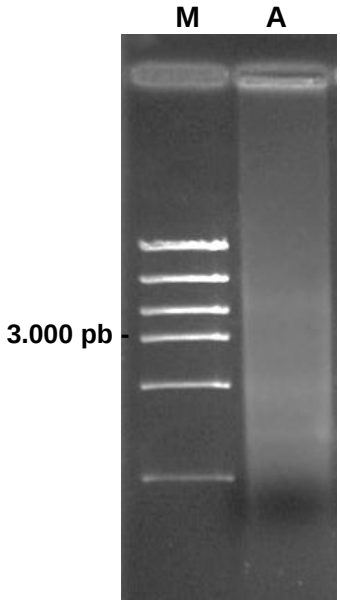


Figura 17 - Eletroforese em gel de agarose do RNA mensageiro purificado e tratado com DNAse RQ1. (A) RNA mensageiro extraído a partir de esporozoítos de *Eimeria tenella*. (M) marcador de peso molecular *High Mass Ladder* (Invitrogen).

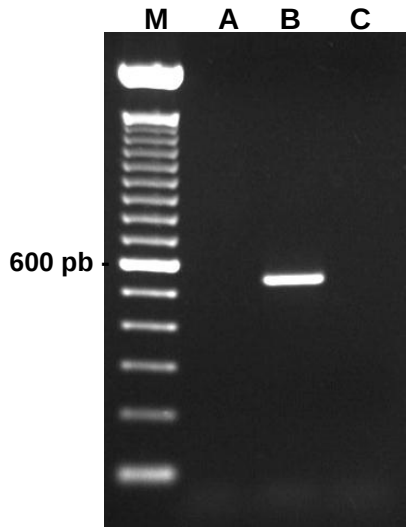


Figura 18 - Resultado do ensaio de PCR para verificação de contaminação do RNA mensageiro com DNA genômico de *E. tenella.* (A) RNA mensageiro de esporozoítos, (B) Controle positivo (DNA de *E tenella*), (C) controle negativo da reação (M) Marcador de peso molecular do tipo escada de 100 pb.

Além da visualização das amostras de RNA mensageiro em gel de agarose, a integridade do RNA mensageiro foi também verificada pela amplificação de um transcrito de alto peso molecular proveniente do gene de Proteína de Micronema 4 (AJ306453) de *Eimeria tenella*. Para todos os RNAs utilizados, após o ensaio de RT-PCR, foi observada uma banda de 986 pb compatível com o fragmento do gene de Proteína Micronema 4, conforme exemplificado na Figura 19. Estes resultados indicam de forma indireta que as amostras de RNA mensageiro utilizadas apresentavam-se íntegras e, portanto, puderam ser utilizadas nas demais etapas deste trabalho.

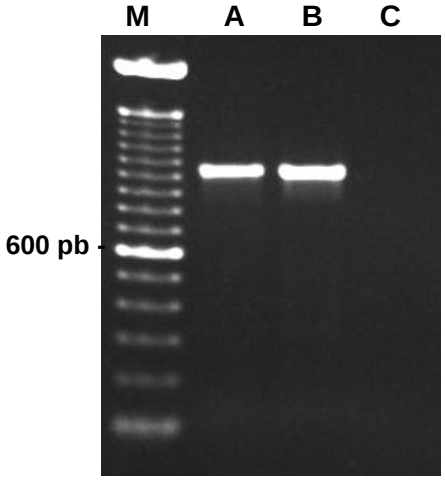


Figura 19 - Resultado do ensaio de PCR para controle de integridade do RNA mensageiro. (A) amostra de cDNA obtida a partir de RNA mensageiro de *E. tenella*, (B) controle positivo da reação (C) controle negativo, (M) Marcador de peso molecular do tipo escada de 100 pb.

4.4 Construção das bibliotecas de LongSAGE de *Eimeria tenella*

Foram construídas com sucesso um total de quatro bibliotecas de LongSAGE a partir de RNA mensageiro de dois estágios evolutivos de *Eimeria tenella*: duas de merozoítos de segunda geração e outras duas a partir de esporozoítos.

4.4.1. Síntese de cDNA e digestão com a enzima de restrição *Nla*III

A síntese de cDNA dupla-fita e a posterior digestão com a enzima de restrição *Nla*III são as etapas iniciais e de fundamental importância para a construção das bibliotecas de LongSAGE. Desta forma, ensaios de PCR foram realizados com o objetivo de verificar se os cDNA provenientes dos RNA mensageiros purificados a partir de esporozoítos e merozoítos haviam sido sintetizados, e posteriormente, adequadamente digeridos pela enzima *Nla*III. Para os ensaios, regiões dos genes constitutivos EtCRK2 e GAPDH foram amplificadas.

Amostras de cDNA recém sintetizado e amostras de cDNA digeridas com a enzima de restrição *Nla*III foram utilizadas como alvo de amplificação. Controles

positivos utilizando DNA genômico de *E. tenella* e, controles negativos sem adição de ácidos nucléicos foram também realizados.

Como podemos observar na Figura 20, a amplificação do fragmento de 247 pb de EtCRK2 tanto a partir de cDNA recém sintetizado (Figura 20A), quanto a partir de cDNA digerido com a enzima *Nla*III (Figura 20B), indica que ocorreu síntese de cDNA. Como a seqüência que codifica para o gene EtCRK2 não possui sítio de restrição para a enzima *Nla*III, a amplificação do produto pós digestão foi utilizada como controle positivo do PCR da reação da digestão com a enzima *Nla*III. O controle positivo utilizando DNA genômico apresentou 564 pb (Figura 20C) devido à presença de íntrons.

Por outro lado, como a seqüência do gene de GAPDH selecionada para amplificação possui sítio de restrição para a enzima *Nla*III, somente foi observada amplificação de um fragmento de 334 pb a partir de cDNA não digerido (Figura 20E). É importante verificar se há amplificação deste fragmento, pois a presença de um produto pós-digestão pode ser indicativo de que a digestão não ocorreu adequadamente.

Analisando todos os dados em conjunto podemos concluir que as etapas de síntese e digestão do cDNA com enzima *Nla*III foram realizadas com sucesso.

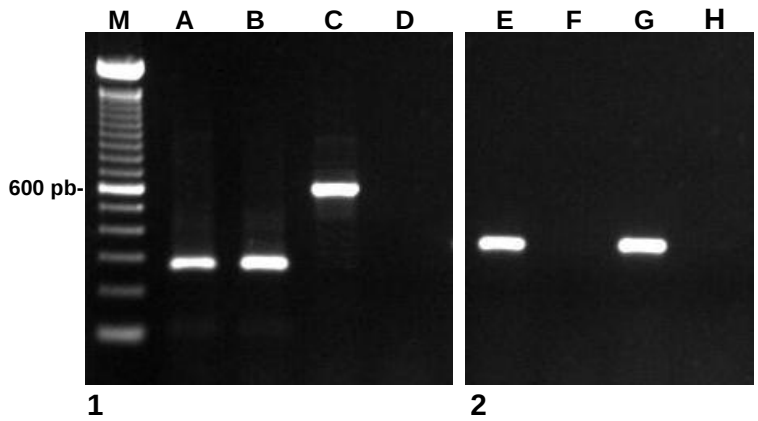


Figura 20 - Verificação da síntese de cDNA e digestão do cDNA com a enzima *Nla*III: (**1**) Produto da amplificação do gene EtCRK2 a partir de: (A) cDNA recém sintetizado; (B) cDNA digerido com *Nla*III*;* (C) DNA genômico; (D) controle negativo. (**2**) Produto da amplificação do gene GAPDH a partir de: (E) cDNA recém sintetizado; (F) cDNA digerido com *Nla*III*;* (G) DNA genômico; (H) controle negativo. (M) Marcador de peso molecular do tipo escada de 100 pb.

4.4.2 Ligação dos Adaptadores, clivagem com a enzima *Mme*I e ligação das *tags* para formação das *ditags*

Para todas as quatro bibliotecas, as etapas de ligação dos adaptadores, clivagem com a enzima *Mme*I e ligação das *tags* para formação de *ditags* foram realizadas adequadamente. As *ditags* de aproximadamente 130 pb foram então amplificadas por PCR. Previamente à amplificação em larga escala por PCR, uma etapa preliminar foi realizada para a otimização da construção das bibliotecas. Três diluições do cDNA das *ditags* foram realizadas e após amplificação e comparação com o controle positivo do kit, a diluição de 1:40 foi escolhida para construção das quatro bibliotecas.

A partir desta diluição, cerca de 300 reações de PCR foram realizadas para cada uma das bibliotecas. O produto total da amplificação foi submetido à eletroforese em gel de poliacrilamida 12% (Figura 21) e as *ditags* de ~130 pb foram excisadas do gel e purificadas eliminando-se assim os produtos da digestão parcial e os adaptadores não incorporados.

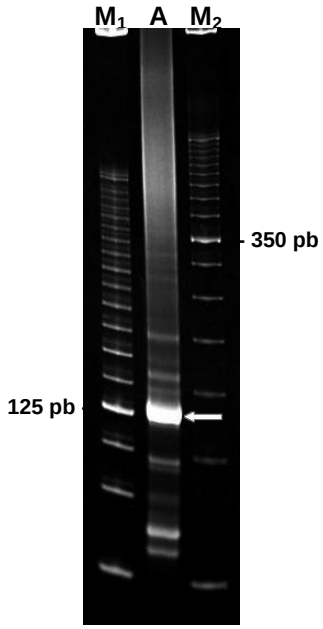


Figura 21 – Eletroforese em gel de poliacrilamida 12% para purificação das *ditags* de ~130 pb. (A) Produto de amplificação das *ditags* de ~130 pb (a seta branca indica o fragmento de ~130 pb que foi excisado do gel), (M_1) Marcador de peso molecular do tipo escada de 25 pb, (M_2) Marcador de peso molecular do tipo escada de 50 pb.

4.4.3 Clivagem com a enzima *Nla*III, purificação das *ditags* de 34 pb e concatenação

Após a purificação das *ditags* de ~130 pb, estas foram submetidas à digestão com a enzima *Nla*III, para a liberação dos adaptadores, resultando nas *ditags* de ~34 pb. Para eliminação de produtos de digestão parcial, assim como as seqüências provenientes dos adaptadores (~40 pb), as *ditags* foram purificadas em gel de poliacrilamida 12%, conforme o exemplo da Figura 22.

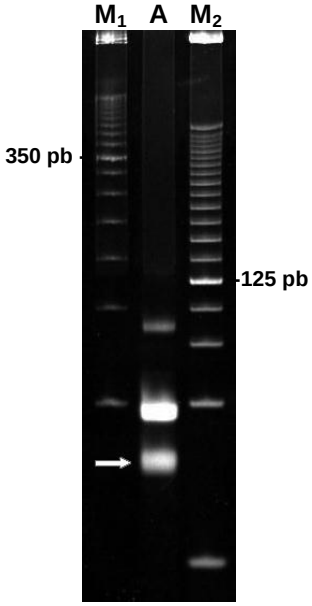


Figura 22 - Eletroforese em gel de poliacrilamida 12% para purificação das *ditags* de ~34 pb. (A) Produto da digestão das *ditags* de ~130 pb com a enzima *Nla*III (a seta branca indica a banda de ~34 pb excisada do gel), (M_1) Marcador de peso molecular do tipo escada de 50 pb, (M_2) Marcador de peso molecular do tipo escada de 25 pb.

Em seguida, as *ditags* de ~34 pb excisadas do gel, foram submetidas a uma reação de ligação para formação dos concatâmeros. Os concatâmeros foram submetidos à eletroforese em gel de poliacrilamida 8%, para seleção por tamanho, (G) 800 – 1000 pb, (M) 500 -800 pb, (P) 400 – 500 pb, e (PP) 300 – 400 pb conforme o exemplo da Figura 23.

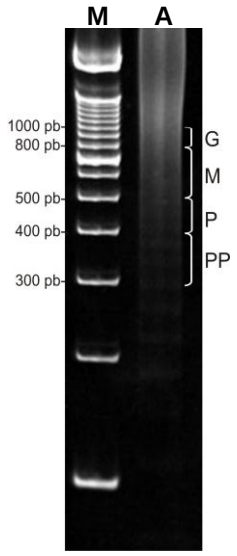


Figura 23 - Eletroforese em gel de poliacrilamida 8% para seleção dos concatâmeros por tamanho. (A) concatâmeros de *ditags* de ~34 pb. As letras em destaque, representam a seleção dos concatâmeros de acordo com seu tamanho: (G) 800 – 1000 pb, (M) 500 -800 pb, (P) 400 – 500 pb, e (PP) 300 – 400 pb, (M) Marcador de peso molecular do tipo escada de 100 pb.

4.4.4 Clonagem, transformação, PCR de colônia e reação de sequenciamento

Os concatâmeros foram ligados ao vetor pZero-1, clonados, amplificados (PCR de colônia) e posteriormente seqüenciados. Para todas as bibliotecas, o grupo de concatâmeros de tamanho entre 500 – 800 pb (M) apresentou maior eficiência na ligação e rendimento na etapa de clonagem sendo o escolhido para a geração de grande parte das seqüências. A Figura 24 mostra o resultado de algumas reações de PCR de colônia obtidas utilizando-se concatâmeros de tamanho entre 500 – 800 pb (M).

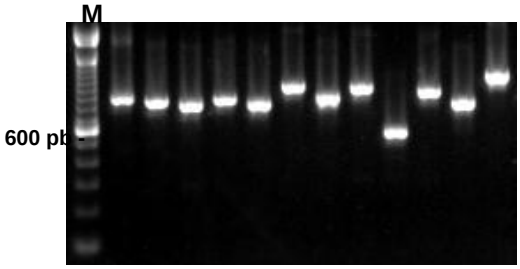


Figura 24 – Produto da reação de PCR de colônia a partir de seqüências clonadas de concatâmeros de tamanho M (500-800 pb). (M) Marcador de peso molecular do tipo escada de 100 pb.

4.5 Pré-processamento das seqüências provenientes das bibliotecas de LongSage

Foram geradas quatro bibliotecas de LongSAGE, sendo duas de estágios merozoítos (MzI e MzII) e duas de esporozoítos (SzI e SzII).

Os cromatogramas gerados foram submetidos a um *pipeline* de processamento seriado automático empregando-se o sistema EGene, como descrito no item 3.6. A Tabela 5 mostra os resultados obtidos discriminados pelas respectivas etapas do *pipeline*. Conforme se pode verificar, a taxa de aceitação de seqüências foi de 63 a 81%, o que, em vista da alta qualidade exigida no *pipeline* (qualidade igual ou superior a 32 em todas as bases), pode ser considerado um bom resultado. A escolha da qualidade 32 foi feita com base nos resultados de extração de *tags*, descritos no próximo item.

Tabela 5 – Resultados do pré-processamento das leituras de LongSAGE de *E. tenella* submetidas a um *pipeline* no sistema EGene.

Discriminação	Biblioteca							
	MzI		MzII		SzI		SzII	
Total inicial de seqüências	2.360	(100%)	1.367	(100%)	1.462	(100%)	1.469	(100%)
Filtradas por tamanho	194	(8,2%)	38	(2,8%)	74	(5%)	79	(5,4%)
Filtradas por qualidade	681	(28,9%)	451	(33%)	207	(14,2%)	235	(16%)
Seqüências aceitas	1.485	(62,9%)	878	(64,2%)	1.181	(80,8%)	1.155	(78,6%)

4.6 Extração e quantificação das *tags*

As *tags* foram extraídas com o conjunto de programas desenvolvido pelo nosso grupo, conforme descrito no item 3.7, utilizando o índice de qualidade Phred 32. Este parâmetro tão estringente poderia resultar no descarte de muitas seqüências, ocorrendo assim uma queda no número final de *tags* extraídas. Para tanto, foi realizado um teste de comparação a partir de seqüências processadas com filtros de qualidade com valores de Phred 20 e 32 (Tabela 6). O uso do valor de Phred 32 eliminou em média 24% das *tags* extraídas. Embora tenhamos perdido parte significativa das *tags*, as seqüências obtidas apresentaram uma altíssima confiabilidade, já que cerca de 20% das *tags* com qualidade Phred 20 poderiam apresentar erros de seqüenciamento.

Tabela 6 – Comparação entre valores de Phred utilizados no filtro de qualidade e seu impacto na obtenção de dados das bibliotecas de LongSAGE.

Critério de avaliação	Biblioteca			
	MzI	MzII	SzI	SzII
Total de bases (Phred 20)	228.314	230.785	304.250	270.550
Total de bases (Phred 32)	155.292	179.004	242.821	197.007
Seqüências aceitas (Phred 20)	1.613	914	1.217	1.310
Seqüências aceitas (Phred 32)	1.485	878	1.181	1.155
Média de bases/seqüência (Phred 20)	141,5	252,5	250	206,52
Média de bases/seqüência (Phred 32)	104,6	203,9	205,6	170,56
Tags extraídas (Phred 20)	9.510	10.374	13.928	12.246
Tags extraídas (Phred 32)	6.550	8.391	11.247	9.060

Como a enzima *Mme*I freqüentemente cliva de forma irregular (Emmersen *et al.*, 2007; Nakonieczna *et al.*, 2009), o nosso programa extraiu *tags* de 17 ou 18 pb (ver item 3.7b). A Figura 25 mostra um histograma de distribuição das *ditags* em relação ao seu tamanho utilizando o conjunto de todas as 4 bibliotecas. Para construção deste histograma utilizou-se *ditags* de 20 a 40 bases. Como podemos observar, a grande concentração de *ditags* encontra-se entre 32 a 34 bases, exatamente o tamanho escolhido no processo de extração das *ditags*. Se assumirmos que apenas *tags* de 17 e 18 pb compõe este grupo de *ditags* e, analisando-se nossos dados, podemos dizer que 41,2% das *tags* possuem 17 pb e 58,8% 18 pb. As *tags* extraídas das pontas dos concatâmeros (*monotags*) não foram utilizadas para esta análise.

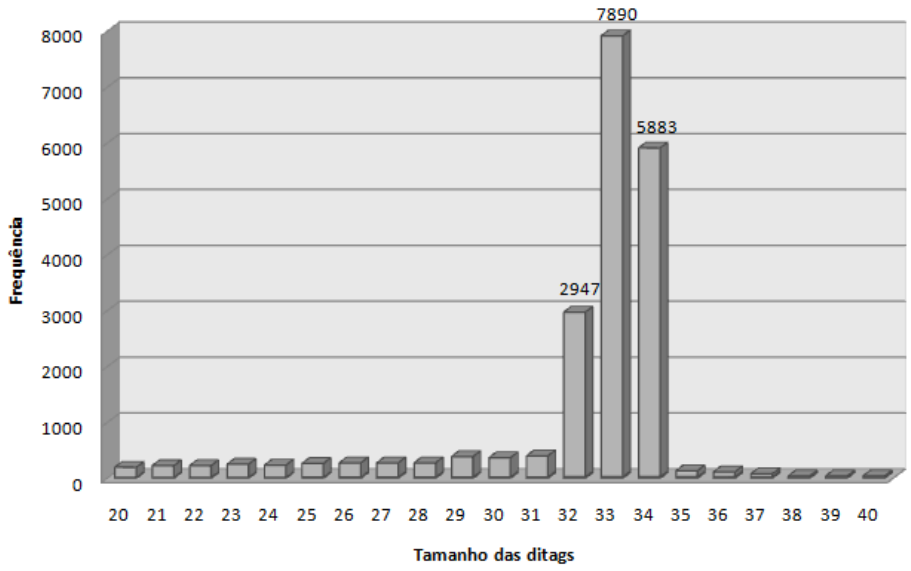


Figura 25 – Histograma de distribuição do tamanho das *ditags* em relação à freqüência nas 4 bibliotecas de LongSAGE de *E. tenella*.

Akmaev e Wang (2004) já haviam descrito que as *ditags* de LongSAGE possuíam uma variação de tamanho entre 32 e 34 nucleotídeos, das quais mais de 50% apresentavam tamanho de 33 pb, conforme observado em nossas bibliotecas (Figura 23).

Para a obtenção de um conjunto uniforme de *tags* de 17 pb as *tags* maiores foram aparadas em uma base. O número de *tags* únicas de cada biblioteca foi determinado e está apresentado na Tabela 7. Conforme pode se ver, cada biblioteca resultou em um número de *tags* variando de 2.633 a 3.951. A determinação do conjunto total de *tags* únicas foi obtida através da junção das *tags* de todas as bibliotecas. Isso foi feito para eliminar a redundância global, visto que há *tags* únicas que estão presentes em mais de uma biblioteca. Após essa junção, obtivemos um total de 9.516 *tags* únicas. É importante ressaltar, que além dos controles relacionados à digestão com a enzima *Nla*III (ver item 4.4.1), nenhuma das 9.516 *tags* únicas obtidas possuem no meio de sua sequência o sítio de reconhecimento CATG, o que é um forte indicativo de que não ocorreu digestão parcial com a enzima *Nla*III.

Tabela 7 – Número de *tags* totais e únicas extraídas das bibliotecas de LongSAGE de *E.tenella*.

Biblioteca	*Tags* totais	*Tags* únicas (17pb)
MzI	6.550	2.633
MzII	8.391	3.489
SzI	11.247	3.951
SzII	9.060	3.457
Total	35.248	-

Das 9.516 *tags* únicas totais, 5.401 *tags* foram detectadas nas bibliotecas de merozoítos, 6.090 *tags* em esporozoítos e apenas 1.975 (20,75%) são compartilhadas entre ambos os estágios evolutivos (Figura 26). As bibliotecas de merozoítos compartilham 1.145 *tags* entre si, enquanto que as esporozoítos compartilham 830. O conjunto total de *tags* (9.516) foi em seguida processado pelo programa Kemp de análise estatística.

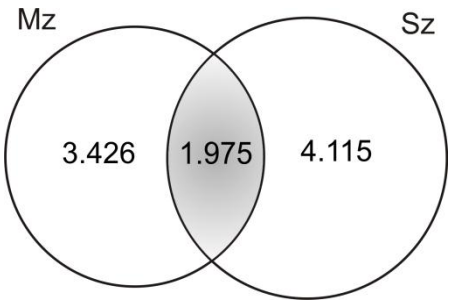


Figura 26 – Diagrama de Venn mostrando a distribuição das 9.516 *tags* únicas de acordo com estágio evolutivo: Mz (merozoítos de segunda geração) e Sz (esporozoítos).

4.7 Análise da freqüência das *tags*

Analisando-se a distribuição das *tags* totais (35.248) podemos observar na Tabela 8 e Figura 26 que 66,5% das *tags* geradas (6.328) ocorrem somente uma vez e representam 17,9% das *tags* únicas totais, além disso, mais de 88% das *tags* únicas apresentam uma con*tag*em inferior a 5. As *tags* com freqüência entre 5 e 50 representam 11,1% do total. Analisando-se as *tags* que aparecem em maior freqüência, de 51 a 1.147 vezes, observa-se que nestas bibliotecas, um pequeno grupo (<1%) de *tags* é altamente expresso.

Se considerarmos que cada *tag* corresponde a um transcrito (o que não é necessariamente correto), esses dados sugerem que 88% dos transcritos apresentam baixa expressão (contagem de até 4). Apenas 11,1% das *tags* apresentaram uma expressão média a alta enquanto somente 0,9% dos genes mostraram uma expressão muito alta (51 a 1.147 vezes), correspondente a 30,2% de todos os transcritos encontrados no parasita.

Tabela 8 - Freqüência das *tags* provenientes das bibliotecas de LongSAGE de *Eimeria tenella.*

Freqüência	*tags* únicas	total de *tags*
1	6.328	6.328
2	1.211	2.422
3	544	1.632
4	298	1.192
5 a 50	1.054	13.019
51 a 1.147	81	10.655
Total	**9.516**	**35.248**

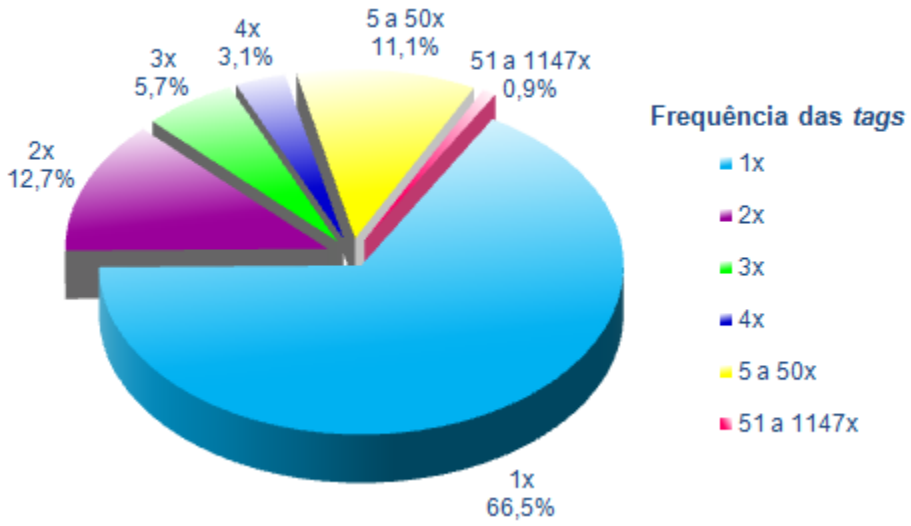


Figura 27 – Porcentagem da freqüência das *tags* únicas provenientes das bibliotecas de LongSAGE de *Eimeria tenella.*

4.8 Mapeamento das *tags* contra o banco de *tags* genômicas virtuais

O conjunto das 9.516 *tags* foi submetido a um mapeamento contra um banco de *tags* virtuais do genoma completo de *E. tenella*, conforme descrito no item 3.8. O banco virtual, obtido a partir do genoma, resultou em aproximadamente 141.000

sítios de reconhecimento (CATG) para a enzima *Nla*III gerando uma lista de 281.862 *tags virtuais (*uma *tag sense* e uma *antisense* para cada sítio CATG). Considerando que o tamanho da mon*tag*em do genoma utilizada apresentava 47.745.645 bases, o sítio de restrição para esta enzima foi encontrado em média a cada 339 pb.

Após a comparação com o banco virtual, um total de 5.626 *tags* (59%) foi mapeado (Figura 28A). Das *tags* mapeadas (5.626), 5.337 foram encontradas apenas uma vez (95%), 172 duas vezes (3%), e 117 várias vezes (2%) (Figura 28B).

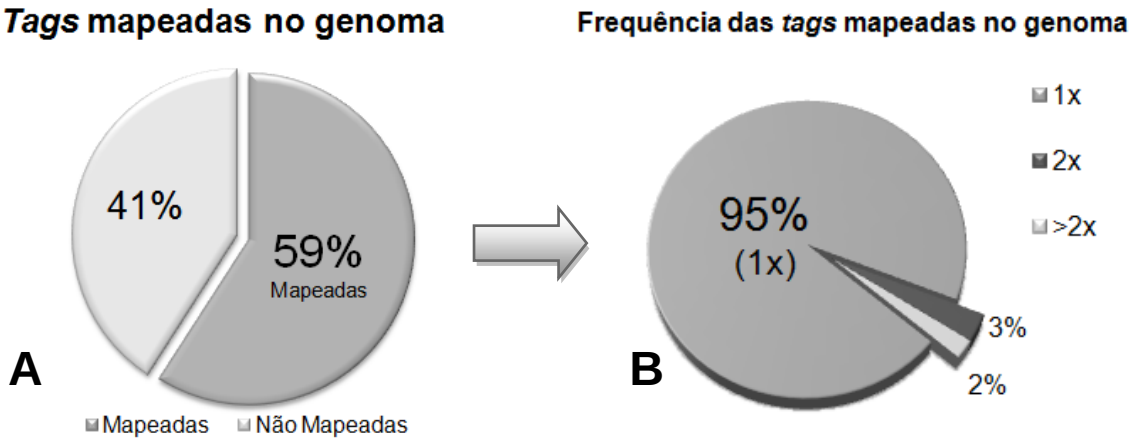
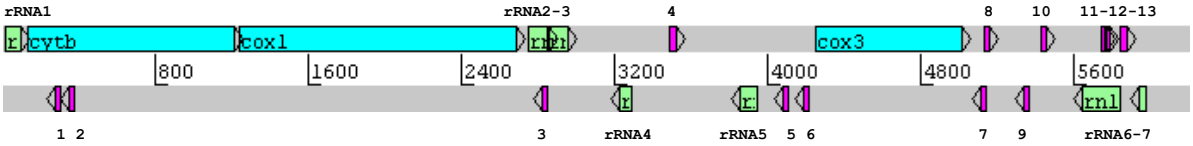


Figura 28 – Mapeamento das *tags* no genoma (A) Porcen*tag*em de *tags* totais mapeadas, (B) freqüência de ocorrência das *tags* mapeadas no genoma.

4.9 Mapeamento das *tags* contra o genoma mitocondrial de *E. tenella*

Um total de 13 *tags* foi mapeado no genoma mitocondrial de *E. tenella*, das quais 8 encontram-se na fita *antisense* (Figura 29). Apenas 4 *tags* apresentaram resultados positivos em buscas de similaridade por BLAST.



1	rRNA	14	82		similar to rRNA fragment 1 of P. falciparum mitochondrial genome (M76611)
	CDS	127	1206		
1	Tag_3928	277	297	c	
2	Tag_1896	343	363	c	
	CDS	1239	2681		
2	rRNA	2748	2845		similar to fragment 10 of P. falciparum mitochondrial genome (M76611)
3	Tag_8293	2823	2843	c	Similar to cytochrome c oxidase subunit I
3	rRNA	2847	2949		similar to fragment 11 of P. falciparum mitochondrial genome (M76611)
4	rRNA	3223	3283	c	similar to fragment number 7 of P. falciparum mitochondrial genome (M76611)
4	Tag_4318	3492	3512		
5	rRNA	3853	3939	c	similar to fragment number 2 of P. falciparum mitochondrial genome (M76611)
5	Tag_1459	4082	4102	c	Similar to cytochrome oxidase I
6	Tag_7422	4188	4208	c	
	CDS	4250	5005		precise initiation codon uncertain
7	Tag_6043	5116	5136	c	
8	Tag_7549	5133	5153		Similar to cytochrome c oxidase subunit III
9	Tag_0934	5343	5363	c	
10	Tag_8523	5433	5453		
6	rRNA	5648	5833	c	similar to fragment number 8 of P. falciparum mitochondrial genome (M76611)
11	Tag_0556	5747	5767		Similar to CYB OXG0 Cytochrome b (b-c1 complex subunit 3)
12	Tag_3030	5764	5784		
13	Tag_9222	5849	5869		
7	rRNA	5944	5971	c	similar to fragment 6 of P. falciparum mitochondrial genome

Figura 29 - Representação esquemática da anotação do genoma mitocondrial de 6.213 pb de *Eimeria tenella*. Genes de citocromo *b* e das subunidades 1 e 3 do citocromo *c* (quadros azuis) dispostos ao longo do genoma mitocondrial, fragmentos das subunidades pequena (rns) e grande (rnl) de rRNA (quadros verdes), e mapeamento das *tags* de LongSAGE (quadros magenta). Abaixo estão relacionados os genes, os rRNAs e as *tags*, as coordenadas e os produtos da anotação.

4.10 Análise estatística e seleção das *tags* diferencialmente expressas

O programa Kemp, além de gerar os dados de *p-value*, possui um procedimento automático para o cálculo do nível crítico (α) para cada *tag* em particular, que neste caso foi escolhido o peso 4 para a hipótese nula ser verdadeira e peso 1 para hipótese nula ser falsa, desta forma, se o *p-value* for menor do que α, a *tag* é classificada como diferencialmente expressa. Portanto a seleção de *tags* diferencialmente expressas foi feita de forma automática. No conjunto de 9.516 *tags* únicas, 270 *tags* foram classificadas como diferencialmente expressas. Dentro deste conjunto, todas as *tags* apresentaram poder de expressão (soma das contagens absolutas das *tags*) maior ou igual a 4.

É interessante observar, que após a análise de freqüência das *tags* (item 4.7), dentro do grupo das *tags* únicas altamente expressas (81 *tags* únicas - 30,2% do total), 45 destas (55,5%) foram classificadas como diferencialmente expressas (dados não mostrados). Em *Toxoplasma gondii*, das 412 *tags* classificadas como altamente expressas, apenas 37% são constitutivamente expressas nos estágios estudados (Radke *et al.*, 2005), indicando que em ambos os parasitas, os genes mais expressos não são representados apenas por *housekeeping genes*, perfil este geralmente encontrado em células animais.

É importante ressaltar ainda que o fato de não termos analisado um conjunto de cerca de 9.000 *tags* não significa que as mesmas não tenham utilidade. De fato, o conjunto de todas as *tags* pode nos fornecer várias outras informações altamente valiosas como, por exemplo, o perfil de expressão. Além disso, esse conjunto de *tags* permitirá, através do mapeamento, a definição do transcriptoma do parasita e a identificação de novos genes. Essas aplicações deverão ser levadas adiante pelo nosso grupo em breve.

4.11 Mapeamento das *tags* diferencialmente expressas contra ESTs/ORESTES e reconstrução dos cDNAs

As 270 *tags* classificadas como diferencialmente expressas, foram submetidas a um processo de reconstrução de seqüências usando-se o programa GenSeed, conforme descrito no item 3.10. Utilizando-se como bases de dados os ESTs/ORESTES, das 270 *tags* iniciais, 199 *tags* foram mapeadas e reconstruídas (73,7%) com uma média de 1.132 pb por seqüência. As 199 seqüências de cDNAs reconstruídas foram então submetidas ao *pipeline* de anotação automática. Quando comparamos esta estratégia de mapeamento nos ESTs com o mapeamento no genoma (item 4.8), que foi feito de forma bem global (apenas localizando os sítios CATG), das 270 *tags* diferencialmente expressas, apenas 211 foram localizadas no genoma. E das 199 *tags* reconstruídas, apenas 160 foram mapeadas no genoma através desta abordagem. Diversas razões podem explicar este resultado, o genoma de *E. tenella* ainda é apenas um rascunho, não finalizado e, não anotado completamente e possui muitos *contigs* (4.707). Talvez estas *tags* poderiam estar

nas fronteiras entre os diversos *contigs*, além disso, algumas *tags* virtuais também poderiam conter SNPs.

4.12 Anotação automática e curagem manual dos cDNAs reconstruídos a aprtir de *tags* diferencialmente expressas em merozoítos e esporozoítos

A anotação foi realizada automaticamente para o conjunto de 199 seqüências de cDNA de *E. tenella* reconstruídas. Do total dessas seqüências, 197 (99%) apresentaram pelo menos uma ORF (*open reading frame* – fase aberta de leitura) codificando um produto protéico contendo no mínimo 50 resíduos de aminoácidos. Apenas 2 seqüências reconstruídas (1%) não foram aceitas pelo critério utilizado no *pipeline* (http://www.coccidia.icb.usp.br/sage_eimeria).

Das seqüências anotadas automaticamente, 102 (51,7%) produtos protéicos não apresentaram similaridades significativas (1e-04) em buscas de similaridade por BLASTP contra a base de dados nr (*GenBank*), enquanto que 95 produtos protéicos (48,3%) apresentaram similaridade contra proteínas de função conhecida ou hipotéticas conservadas.

Os resultados de anotação dos reconstruídos a partir das *tags* diferencialmente expressas entre merozoítos e esporozoítos (197) podem ser observados nas Tabelas 9 a 12. Conforme pode se observar, foram identificadas 130 *tags* com expressão diferencial (*up-regulated*) em merozoítos (Tabelas 9 e 10) e 67 em esporozoítos (Tabelas 11e 12).

Para as 130 seqüências reconstruídas a partir de *tags* diferencialmente expressas em merozoítos, 79 (60,7%) apresentaram resultados positivos de BLAST e estão representados na Tabela 9. É interessante observar que 70 destas seqüências (88,6%) apresentaram produtos protéicos similares aos de organismos do Filo Apicomplexa (Figura 30A), destes, 48 (68,6%) pertencem à classe Coccidia, 28 (40%) ao gênero *Eimeria* e 25 à espécie *Eimeria tenella* (Figura 30B). Apenas 9 seqüências (11,4%) apresentaram resultados positivos contra outros organismos diversos, incluindo fungos, bactérias, plantas, dinoflagelados, entre outros, sendo que 5 destes produtos correspondem a proteínas de ribossomo (normalmente conservadas entre organismos de espécies distintas).

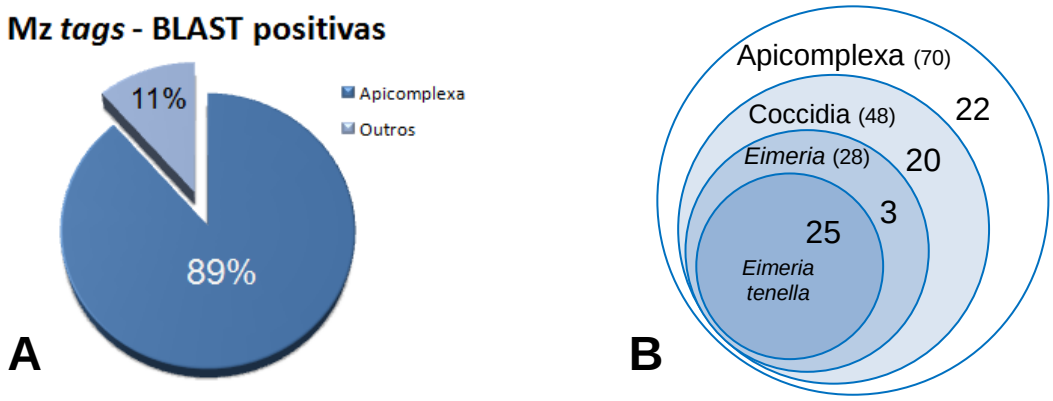


Figura 30 – Distribuição dos 79 produtos protéicos diferencialmente expressos em merozoítos (*up-regulated*), de acordo com as espécies encontradas nas buscas de similaridade por BLAST (A) Porcen*tag*em de produtos protéicos similares à sequências de organismos do Filo Apicomplexa, (B) Diagrama de Venn mostrando a distribuição das seqüências em relação ao Filo Apicomplexa, à classe Coccidia, ao Gênero *Eimeria* e à espécie *E. tenella*.

Ainda analisando os 79 produtos protéicos das seqüências reconstruídas a partir das *tags* de merozoítos (Figura 31 e Tabela 9), podemos observar que 38 destes correspondem a proteínas de ribossomo (48,1%), 20 a antígenos (25,3%), sendo que destes 17 são similares a antígenos de superfície e outros 3 produtos (3,8%) correspondem a proteínas de micronema, os 18 produtos restantes correspondem a outros genes de funções diversas, incluindo actina, calmodulina, tiorredoxina, proteína wx, entre outras.

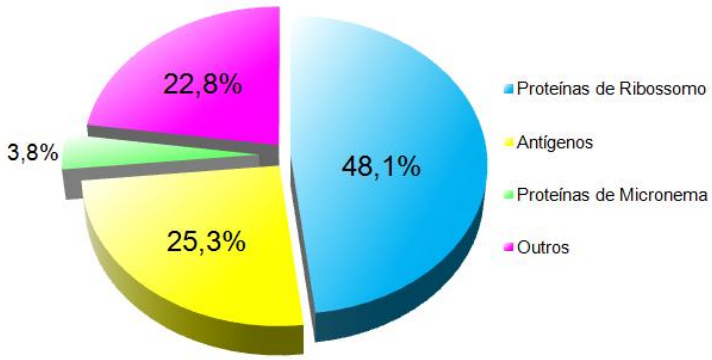


Figura 31 – Distribuição dos 79 produtos protéicos resultantes da anotação dos cDNAs reconstruídos a partir das *tags* diferencialmente expressas em merozoítos (*up-regulated*) que apresentaram resultados positivos de BLAST.

Tabela 9 – Seqüências de cDNAs reconstruídas a partir de *tags* diferencialmente expressas em merozoítos *(up-regulated)* que apresentaram resultado de BLAST positivo. Estão apresentados na tabela os valores de p (*p-value*) após a comparação entre os estágios de merozoítos e esporozoítos (Mz_vs_Sz), alpha (nível critico), poder de expressão (PE), o número de vezes que as *tags* foram mais expressos em relação a esporozoítos (Mz↑), os resultados da busca de similaridade por BLAST contra a base de dados nr, os códigos de acesso das seqüências e o organismo.

Id do cDNA	Seqüência da *tag*	*p-value*	alpha	PE	Mz↑	Resultado de BLAST	# Acesso	Organismo
Et_SAGE0093	aaagcacttttcgatgc	1,1E-03	2,5E-02	8	3,1	Similar to ribosomal protein S5	XP_727402	*Plasmodium yoelii*
Et_SAGE0125	aaatgcactgcgcttcc	1,9E-03	2,1E-02	12	15,1	Similar to 60S ribosomal protein L7a, putative	EEA05430	*Cryptosporidium muris*
Et_SAGE0135	aaatttacacttgcggc	3,1E-03	2,0E-02	13	7,4	Similar to wx protein	AAO60053	*Toxoplasma gondii*
Et_SAGE0161	aacagaggcacagaaga	1,1E-03	2,5E-02	8	3,7	Similar to surface antigen 16	CAE52303	*Eimeria tenella*
Et_SAGE0263	aactcggcaggcccgca	2,1E-04	1,8E-02	17	10,2	Similar to 40S ribosomal protein S8, putative	EEA08343	*Cryptosporidium muris*
Et_SAGE0357	aagccctgcttccagtc	3,2E-02	3,4E-02	4	1,6	Similar to AF353165_1 microneme protein 8	AAK19757	*Toxoplasma gondii*
Et_SAGE0459	aagtgagttgactgcgc	1,8E-04	2,3E-02	10	3,9	Similar to surface antigen 23	CAE52308	*Eimeria tenella*
Et_SAGE0478	aagtggcgaaagccctg	1,3E-02	2,4E-02	9	10,7	Similar to 40S ribosomal protein S12	XP_762841	*Theileria parva*
Et_SAGE0675	acccaaaccctaaaccc	5,5E-04	1,7E-02	21	5,5	Similar to 40S ribosomal protein S13	XP_001350306	*Plasmodium falciparum*
Et_SAGE0689	accggcggggcctggac	1,3E-02	2,4E-02	9	10,7	Similar to ribosomal L37ae protein family	AAQ23712	*Toxoplasma gondii*
Et_SAGE1072	agctgctgctgctgctg	3,2E-02	3,4E-02	4	1,6	Similar to peptidyl-prolyl cis-trans isomerase	XP_001941476	*Pyrenophora tritici-repentis*
Et_SAGE1328	agttgcagttttcaaag	0	1,5E-02	29	11,3	Similar to surface antigen 7	CAE52294	*Eimeria tenella*
Et_SAGE1431	atgaacaacgccgccag	3,0E-06	1,3E-02	43	5,0	Similar to 19 kDa sporozoite antigen	AAW31899	*Eimeria tenella*
Et_SAGE1463	atgattttctacctccc	0	5,2E-03	230	2,1	Similar to microneme protein 4	CAC34726	*Eimeria tenella*
Et_SAGE1591	atttactcttcatcgaa	0	1,6E-02	22	8,6	Similar to actin depolymerizing factor	ABM89551	*Eimeria tenella*
Et_SAGE1740	caaaggactagtggcgc	0	8,4E-03	92	4,5	Similar to ribosomal protein L11	BAF98658	*Solea senegalensis*
Et_SAGE1929	caagactgaactcagtc	3,0E-06	1,3E-02	40	4,7	Similar to ubiquitin/s27a 40s ribosomal protein	AAW56553	*Nicotiana benthamiana*
Et_SAGE1938	caagcaaacaacaccgc	0	1,6E-02	22	9,2	Similar to surface antigen 22	CAE52312	*Eimeria tenella*
Et_SAGE2112	cacacaaaagacgtgcg	1,9E-03	2,1E-02	12	15,5	Similar to putative 60S ribosomal protein RPL31	XP_001610775	*Babesia bovis*
Et_SAGE2794	cagccatttgcaaagaa	1,1E-03	2,5E-02	8	3,7	Similar to heat shock protein 90	AAS17969	*Eimeria acervulina*
Et_SAGE2865	cagcgcccgggaggcag	9,0E-06	1,5E-02	28	7,8	Similar to 60S ribosomal protein L19, putative	XP_001608966	*Babesia bovis*
Et_SAGE2948	cagctcggcagcagctt	1,4E-02	3,1E-02	5	2,0	Similar to nascent polypeptide associated complex alpha chain protein, putative	NP_001078369	*Arabidopsis thaliana*
Et_SAGE2994	cagctgggaggccgcgg	3,2E-02	3,4E-02	4	1,6	Similar to 60S ribosomal protein L38	Q4PMD1	*Ixodes scapularis*

continua

Tabela 9 – (continuação)

Id do cDNA	Seqüência da tag	*p-value*	alpha	PE	Mz↑	Resultado de Blast	# Acesso	Organismo
Et_SAGE3089	caggcttataggaaacg	6,9E-05	2,2E-02	11	4,3	Similar to surface antigen 17	CAE52302	*Eimeria tenella*
Et_SAGE3542	ccaaacctccaactctc	1,0E-06	1,9E-02	15	6,5	Similar to surface antigen 16	CAE52303	*Eimeria tenella*
Et_SAGE3560	ccaagcctccaactgtc	1,2E-04	2,0E-02	14	17,4	Similar to surface antigen 17	CAE52302	*Eimeria tenella*
Et_SAGE3624	ccattgactcgaagtaa	2,5E-02	2,5E-02	8	9,4	Similar to surface antigen 19	CAE52305	*Eimeria tenella*
Et_SAGE3646	cccagcgtgcgggccaa	2,5E-02	2,5E-02	8	7,6	Similar to 60S ribosomal protein L15	XP_762879	*Theileria parva*
Et_SAGE3730	ccgaacgaaagacacct	3,2E-02	3,4E-02	4	2,2	Similar to 40S ribosomal protein S5	XP_762985	*Theileria parva*
Et_SAGE3839	cctcacattagcacgcc	7,0E-06	1,4E-02	36	5,0	Similar to microneme protein 5	CAB52368	*Eimeria tenella*
Et_SAGE3957	cgaacagccaggtggag	3,2E-02	3,4E-02	4	1,6	Similar to 60S ribosomal protein L24	XP_739533	*Plasmodium chabaudi*
Et_SAGE3991	cgacagaaacttatggt	2,3E-05	1,7E-02	21	12,9	Similar to hypothetical protein, conserved	XP_001238614	*Eimeria tenella*
Et_SAGE4101	cgcactgcgcggaagct	2,5E-02	2,5E-02	8	9,4	Similar to ribosomal protein S23	CAD98683	*Cryptosporidium parvum*
Et_SAGE4143	cgcaggccccgctgcag	0	1,7E-02	20	25,5	Similar to 60S ribosomal protein L38	Q4PMD1	*Ixodes scapularis*
Et_SAGE4275	cgcgcctgagcccgctg	5,7E-04	1,6E-02	24	5,1	Similar to 40s ribosomal protein s17, putative	CAJ20516	*Toxoplasma gondii*
Et_SAGE4421	cgcttggcgttccgcct	0	7,0E-03	132	28,3	Similar to surface antigen	AAA16457	*Eimeria tenella*
Et_SAGE4592	cgggaagcttttccgcg	1,2E-02	2,2E-02	11	6,0	Similar to 60S ribosomal protein L30, putative	XP_626935	*Cryptosporidium parvum*
Et_SAGE4673	cggttagtaactcttct	0	8,9E-03	83	12,6	Similar to ribosomal protein S7e family protein	XP_001612059	*Babesia bovis*
Et_SAGE4699	cgtaggggcatcgtgaa	2,6E-05	1,7E-02	20	12,1	Similar to 40S ribosomal protein S15a	XP_764633	*Theileria parva*
Et_SAGE4919	ctaactcaaccggatac	2,5E-03	2,7E-02	7	2,7	Similar to surface antigen	AAA16457	*Eimeria tenella*
Et_SAGE5007	ctcacgaaccagatcac	1,1E-03	2,5E-02	8	3,1	Similar to putative ribosomal protein S2	CAD43146	*Toxoplasma gondii*
Et_SAGE5210	ctgcgactaataggcag	6,0E-06	1,6E-02	23	14,3	Similar to RL5_EIMTE 60S ribosomal protein L5	Q5EY89	*Eimeria tenella*
Et_SAGE5438	ctgtgggacgtgagtga	0	1,6E-02	23	29,5	Similar to receptor for activated C kinase 1	AAT11121	*Toxoplasma gondii*
Et_SAGE5444	ctgttgatgcacaacac	1,2E-02	2,2E-02	11	6,0	Similar to 60S ribosomal protein L32	XP_627679	*Cryptosporidium parvum*
Et_SAGE5459	cttacggcattgtcggg	3,2E-02	3,4E-02	4	1,6	Similar to surface antigen 9	CAE52300	*Eimeria tenella*
Et_SAGE5610	gaaacagttactgctga	1,4E-02	3,1E-02	5	2,0	Similar to unknown	ABX80198	*Prorocentrum minimum*
Et_SAGE5642	gaaatgcgtttcaggct	0	1,5E-02	28	36,6	Similar to ACT_TOXGO Actin	P53476	*Toxoplasma gondii*
Et_SAGE5789	gacgccctcgagccgta	3,2E-02	3,4E-02	4	1,6	Similar to superoxide dismutase	YP_002151128	*Proteus mirabilis*
Et_SAGE5846	gagaaacgcacagtgct	0	8,0E-03	101	13,2	Similar to 60S ribosomal protein L24	XP_739533	*Plasmodium chabaudi*
Et_SAGE6008	gagtacctgcgtgacta	1,4E-02	3,1E-02	5	2,6	Similar to 40S ribosomal protein S10	XP_764473	*Theileria parva*
Et_SAGE6107	gcaaaccgggcattggc	0	7,4E-03	117	2,7	Similar to 40s ribosomal protein s15, putative	XP_952742	*Theileria annulata*

continua

Tabela 9 – (continuação)

Id do cDNA	Seqüência da tag	*p-value*	alpha	PE	Mz↑	Resultado de BLAST	# Acesso	Organismo
Et_SAGE6121	gcaacacagctggcctc	0	5,4E-03	216	4,8	Similar to antigen	AAA29080	*Eimeria acervulina*
Et_SAGE6140	gcacaaggcgcagtagt	3,3E-03	1,8E-02	18	4,4	Similar to myosin light chain TgMLC1	AAL08211	*Toxoplasma gondii*
Et_SAGE6144	gcacagacttaactttg	3,7E-03	2,2E-02	11	11,4	Similar to antigen	AAA29080	*Eimeria acervulina*
Et_SAGE6284	gccacttgcattacgac	3,2E-02	3,4E-02	4	1,6	Similar to acid phosphatase, putative	CAQ42925	*Plasmodium knowlesi*
Et_SAGE6364	gccgcccccgtcgtcag	6,8E-03	2,3E-02	10	9,7	Similar to possible apicomplexan-specific protein	XP_625389	*Cryptosporidium parvum*
Et_SAGE6426	gccttctgtctgttgca	3,2E-02	3,4E-02	4	1,6	Similar to vacuolar ATP synthase subunit G1, putative	EEA06526	*Cryptosporidium muris*
Et_SAGE6639	gctgcctgtttggaggt	0	1,7E-02	20	26,2	Similar to surface antigen 2	CAE52301	*Eimeria tenella*
Et_SAGE6890	ggcagggccggaacggc	1,4E-02	3,1E-02	5	3,2	Similar to 60S ribosomal protein L8	XP_667075	*Cryptosporidium hominis*
Et_SAGE7411	ggtttctctctctttga	0	1,4E-02	32	8,9	Similar to ribosomal protein P1	BAD67181	*Neospora caninum*
Et_SAGE7414	ggtttgtctctctttga	8,1E-04	1,9E-02	15	7,9	Similar to AF353514_1 ribosomal protein P2	AAK38885	*Eimeria tenella*
Et_SAGE7450	gtatatatcctagactc	0	1,2E-02	49	15,5	Similar to surface antigen 18	CAE52309	*Eimeria tenella*
Et_SAGE7507	gtctcgctgcccaagga	1,2E-02	2,0E-02	14	5,4	Similar to 40s ribosomal protein s4, putative	CAJ20484	*Toxoplasma gondii*
Et_SAGE7601	gtggcactgtgacgcgt	3,2E-02	3,4E-02	4	2,2	Similar to surface antigen 9	CAE52300	*Eimeria tenella*
Et_SAGE7668	gtgttgcttttctcctc	3,2E-02	3,4E-02	4	1,6	Similar to thioredoxin 1	AAF34541	*Plasmodium falciparum*
Et_SAGE7689	gttctcgggacctgtcc	3,2E-02	3,4E-02	4	1,6	Similar to 60S ribosomal protein l35a, putative	XP_955277	*Theileria annulata*
Et_SAGE7722	gttttcccttctggtcg	4,7E-03	1,3E-02	38	2,7	Similar to 60S ribosomal protein L12, putative	CAQ40420	*Plasmodium knowlesi*
Et_SAGE7808	tacaaagaagtccgcga	4,5E-04	2,4E-02	9	3,5	Similar to ribosomal protein L18A	CAJ17297	*Georissus sp.*
Et_SAGE7853	tacctaacttaacagtt	1,0E-06	1,9E-02	15	6,5	Similar to calmodulin	XP_765284	*Theileria parva*
Et_SAGE8085	tatttggctcaaactaa	1,9E-03	2,1E-02	12	14,7	Similar to surface antigen 19	CAE52305	*Eimeria tenella*
Et_SAGE8093	tcaaactgtccatcgag	1,4E-02	3,1E-02	5	2,0	Similar to guanylyl cyclase	CAC00546	*Plasmodium falciparum*
Et_SAGE8099	tcaaccccgtcaggtgc	0	1,6E-02	22	8,6	Similar to 40S ribosomal protein S26e	XP_766501	*Theileria parva*
Et_SAGE8495	tgagctgctgctgcagc	9,0E-06	1,5E-02	28	7,6	Similar to 40S ribosomal protein S3, putative	EEA07044	*Cryptosporidium muris*
Et_SAGE8907	tggtacgaaggagcttc	5,8E-03	2,8E-02	6	3,0	Similar to 1433_EIMTE 14-3-3 protein	O96436	*Eimeria tenella*
Et_SAGE8984	tgtgaataagtcgctaa	4,5E-04	2,4E-02	9	3,5	Similar to surface antigen 19	CAE52305	*Eimeria tenella*
Et_SAGE9037	tgtttcccacggggcac	1,8E-04	1,1E-02	58	2,6	Similar to 60S ribosomal protein L8	XP_667075	*Cryptosporidium hominis*
Et_SAGE9363	tttacgggggtgacgaa	1,1E-03	2,5E-02	8	3,1	Similar to ribosomal protein L6 homologue, putative	CAQ41377	*Plasmodium knowlesi*
Et_SAGE9393	tttcggatttttaggaa	1,8E-04	2,3E-02	10	3,9	Similar to surface antigen 21	CAE52307	*Eimeria tenella*
Et_SAGE9512	tttttgctagttttttc	3,2E-02	3,4E-02	4	1,6	Similar to 60s ribosomal protein L36, putative	EEA07466	*Cryptosporidium muris*

conclusão

Analisando ainda os resultados da Tabela 9, podemos observar que alguns cDNAs reconstruídos a partir de *tags* distintas apresentaram produtos de anotação idênticos, ou seja, o mesmo número de acesso no *GenBank.*

Por exemplo, as *tags* 4421 e 4919 apresentam como produto da anotação um antígeno de *E. tenella* (estes genes de antígenos são muito semelhantes entre si, formando famílias gênicas). Ambas as *tags* estão no mesmo transcrito reconstruído, a *tag* 4421 representa a *tag* mais próxima da cauda poliA (1º CATG), enquanto que a *tag* 4919, está presente no segundo sítio de reconhecimento para a enzima *Nla*III (2º CATG). O mesmo acontece para as *tags* 5846 (1º CATG) e 3957 (2º CATG), similares a uma RPL24 (*ribosomal protein* L24 - *Plasmodium chabaudi*). Nestes dois casos, este resultado pode ser decorrente de digestão parcial com a enzima *Nla*III, uma falha inerente à técnica, apesar de todos nossos controles apontarem para uma digestão completa. Além disso, podem também representar de fato variantes de transcritos, resultados de um *splicing* alternativo na porção 3' UTR, ou, sítios de poliadenilação heterogênea. A seqüência contendo as *tags* 4421 e 4919 possui dois sítios (ATAAA) de sinal de poliadenilação, mas que estão muito distantes da cauda poliA (cerca de 500-600pb), o que a princípio não explicaria uma poliadenilação heterogênea. A seqüência reconstruída a partir das *tags* 5846 e 3957 não possui sinal de poliadenilação, mas é dotada de uma cauda poliA de cerca de 170 nucleotídeos. A região codificadora predita de 462 pb similar a RPL24 quando submetida à busca de similaridade por BLASTN no genoma de *E. tenella*, alinha em 3 porções distintas do mesmo *contig* do genoma separados por menos de 400 pb cada. Estes fragmentos poderiam corresponder à 3 exons distintos, e, portanto seria razoável supor que estes transcritos diferenciais possam ser derivados de *splicing* alternativo.

Os reconstruídos a partir das *tags* 6890 e 9037 foram identificados como uma proteína de ribossomo (RPL8 - *Cryptosporidium hominis*). Ambas as *tags* estão na mesma seqüência reconstruída, a *tag* 9037 está posicionada no 1º sítio de restrição para a enzima *Nla*III mais próximo da cauda poliA, no sentido *sense*, enquanto que a *tag* 6890, encontra-se na mesma posição no sentido *antisense*. Neste caso podemos supor que exista um controle pós-transcricional utilizando seqüências *antisense*. O mesmo foi encontrado para as *tags* 2994 (*sense*) e 4143 (*antisense*), correspondentes a uma RPL38 e para as *tags* 4421 (*sense*) (RPL24- *Plasmodium chabaudi*) e 8276 (*antisense*), anotada como uma proteína hipotética (Tabela 9).

Os cDNAs reconstruídos a partir das *tags* 7601 e 5459 são semelhantes a um antígeno de superfície de *Eimeria tenella*, no entanto, as *tags* estão apresentadas separadamente em cada uma destas seqüências reconstruídas. Os produtos possuem uma região de similaridade, mas as seqüências reconstruídas são diferentes entre si, sendo que cada uma delas quando submetidas a buscas de similaridade na base de ESTs/ORESTES montada de *E. tenella*, são encontradas em *contigs* diferentes, e quando comparadas ao genoma de *E. tenella* alinham no mesmo *contig*, mas em regiões diferentes, o que parece indicar a presença de um mesmo gene com duas isoformas.

Na Tabela 10, estão apresentados os 51 (39,3%) cDNAs reconstruídos a partir de *tags* diferencialmente expressas em merozoítos (*up-regulated*) em relação a esporozoítos que não apresentaram resultados positivos de BLAST, e portanto foram caracterizadas como proteínas hipotéticas. Analisando-se a diferença de expressão entre merozoítos e esporozoítos, 9 *tags* possuem uma diferença considerável de 10 a 31,6 vezes maior do que em esporozoítos.

Tabela 10 – Seqüências de cDNAs reconstruídas a partir de *tags* diferencialmente expressas em merozoítos *(up-regulated)* que não apresentaram resultado de BLAST. Estão apresentados na tabela os valores de p após a comparação entre os estágios de merozoítos e esporozoítos (Mz_vs_Sz), alpha (nível critico), poder de expressão (PE) e o número de vezes que as *tags* foram mais expressas em relação a esporozoítos (Mz↑).

Id do cDNA	Seqüência da *tag*	*p-value*	alpha	PE	Mz↑
Et_SAGE0014	aaaaagctaaacggaga	1,4E-02	3,1E-02	5	2,0
Et_SAGE0038	aaaagtcttagcggcag	3,2E-02	3,4E-02	4	1,6
Et_SAGE1282	agtgaggctcgagcaac	1,3E-02	2,4E-02	9	10,7
Et_SAGE1401	atcgccgtttactacgt	3,2E-02	3,4E-02	4	1,6
Et_SAGE1655	caaacactgagggacta	0	1,2E-02	50	12,1
Et_SAGE1891	caactgcgaaaagagct	1,4E-02	3,1E-02	5	2,0
Et_SAGE1942	caagcagcagcccactg	0	1,4E-02	32	20,5
Et_SAGE2117	cacacacttggcgtttg	3,2E-02	3,4E-02	4	1,6
Et_SAGE3072	caggcgatggtctcgtg	0	6,7E-03	144	3,1
Et_SAGE3288	cagtttgagctgatatg	4,0E-06	6,9E-03	135	2,5
Et_SAGE3440	cattcggagccccagag	2,2E-02	2,3E-02	10	4,8
Et_SAGE3664	ccccacgtctgcaccct	2,5E-03	2,7E-02	7	2,7
Et_SAGE3830	cctactagggctaaccg	5,8E-03	2,8E-02	6	2,3
Et_SAGE3903	ccttacattctacagac	3,2E-02	3,4E-02	4	1,6

continua

Tabela 10 – (continuação)

Id do cDNA	Seqüência da *tag*	*p-value*	alpha	PE	Mz↑
Et_SAGE3924	cctttgactcaaaccaa	5,0E-06	1,8E-02	18	23,2
Et_SAGE3927	cctttgattcaatgtaa	5,8E-03	2,8E-02	6	2,3
Et_SAGE4064	cgattcgggccgtccaa	3,2E-02	3,4E-02	4	1,6
Et_SAGE4267	cgcgcagctgcgggcgg	0	1,6E-02	24	31,6
Et_SAGE4407	cgcttccgagcctccgt	2,2E-02	2,3E-02	10	5,1
Et_SAGE4467	cggagcagaagcagcac	0	1,9E-02	16	6,2
Et_SAGE4492	cggatttagtctaacag	2,5E-02	2,5E-02	8	8,5
Et_SAGE4590	cggctttgctgctgcgc	1,0E-06	1,9E-02	15	5,9
Et_SAGE4819	cgttcgtgtctgtaaag	0	1,6E-02	25	13,8
Et_SAGE4913	ctaaaggcttcaaaaaa	1,4E-02	3,1E-02	5	2,0
Et_SAGE5162	ctgcagcagcagctgca	0	1,5E-02	30	7,9
Et_SAGE5268	ctgctgcaaggcatcgc	6,9E-05	2,2E-02	11	4,9
Et_SAGE5549	cttgttttctctgatgc	3,2E-02	3,4E-02	4	2,2
Et_SAGE5599	gaaaagacttctctgaa	5,8E-03	2,8E-02	6	2,3
Et_SAGE5780	gacgacctcgttcccga	5,8E-03	2,8E-02	6	2,3
Et_SAGE6150	gcacatcatcaatcctt	3,0E-06	1,6E-02	24	13,3
Et_SAGE6409	gcctgcggtgtattgac	1,4E-02	3,1E-02	5	2,6
Et_SAGE6515	gcgggaaacgggcgtga	1,1E-03	2,5E-02	8	3,1
Et_SAGE6540	gcgtctcgaccttgctt	1,4E-02	3,1E-02	5	2,0
Et_SAGE6767	ggacgccaaagcaaagt	1,8E-04	2,3E-02	10	3,9
Et_SAGE6835	ggattctctctcttcga	1,2E-02	2,2E-02	11	6,0
Et_SAGE6962	ggcggaggggagggttc	3,7E-03	2,2E-02	11	13,8
Et_SAGE7025	ggctttgtgggggtttc	4,2E-04	1,9E-02	16	9,8
Et_SAGE7075	gggcagcgcaggagtca	5,6E-03	1,4E-02	31	2,5
Et_SAGE7269	gggggccccgcaggtgc	3,2E-02	3,4E-02	4	1,6
Et_SAGE7647	gtggtgctatttaattt	4,5E-04	2,4E-02	9	3,5
Et_SAGE7737	taaaaggcttgttgaat	2,6E-05	2,1E-02	12	5,3
Et_SAGE7798	taattgctgcagctggg	1,0E-06	1,1E-02	57	4,3
Et_SAGE8276	tcgcggcacaggcaagc	2,3E-05	1,5E-02	26	7,1
Et_SAGE8514	tgataccgcgacaaatt	1,1E-03	2,5E-02	8	3,1
Et_SAGE8631	tgccgcttttacaaaag	0	1,2E-02	47	5,4
Et_SAGE8795	tggataccagcgcgatt	3,2E-02	3,4E-02	4	1,6
Et_SAGE9103	ttatgaggcagttaagt	1,1E-03	2,5E-02	8	3,1
Et_SAGE9283	ttgctgcatctttcacc	1,1E-03	2,5E-02	8	3,1
Et_SAGE9340	ttgtttattactgttcc	1,0E-06	1,9E-02	15	5,9
Et_SAGE9461	tttgttcctggaacaga	2,6E-05	2,1E-02	12	4,7
Et_SAGE9497	ttttgttttgaataaac	0	1,6E-02	24	25,1

conclusão

Podemos observar que dentro deste conjunto de produtos protéicos classificados como hipotéticos, 3 *tags* (5,9%) são de alta expressão (> 50x), 40 *tags* (78,4%) de média expressão, e apenas 8 *tags* (15,7%) são de baixa expressão, conforme destacado na Figura 32.

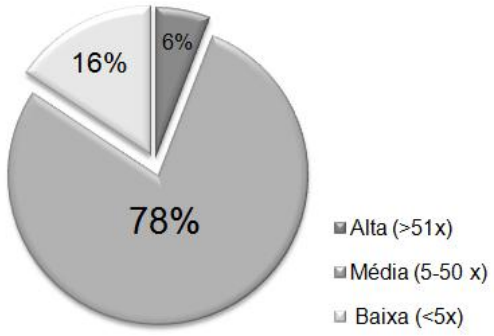


Figura 32 – Distribuição das 51 proteínas hipotéticas resultantes da anotação dos cDNAs reconstruídos a partir das *tags* diferencialmente expressas em merozoítos (*up-regulated*), de acordo com a freqüência de contagem.

Analisando-se os reconstruídos diferencialmente expressos em esporozoítos em relação à merozoítos (Tabela 11 e 12), apenas 16 *tags* reconstruídas (23,9%) apresentaram similaridade significativa por BLAST (Tabela 11), sendo que destas, 5 (31%) correspondem a proteínas de *Eimeria* tenella de função desconhecida. Para esporozoítos não temos grandes grupos de genes relacionados mas sim produtos protéicos diversos de grande importância como histona, catepsina, facilitador do transporte de glicose entre outros. Novamente grande parte das seqüências anotadas apresentou similaridade com seqüências pertencentes a organismos do Filo Apicomplexa, no total 13 (81,2%), sendo que destas, 7 (43,4%) são da própria espécie *Eimeria tenella* (Figura 33).

Tabela 11 – Seqüências de cDNAs reconstruídas a partir de *tags* diferencialmente expressas em esporozoítos *(up-regulated)* que apresentaram resultado de BLAST positivo. Estão apresentados na tabela os valores de p (*p-value*) após a comparação entre os estágios de merozoítos e esporozoítos (Mz_vs_Sz), alpha (nível critico), poder de expressão (PE), o número de vezes que as *tags* foram mais expressos em relação a merozoítos (Sz↑), os resultados da busca de similaridade por BLAST contra a base de dados nr, os códigos de acesso das seqüências e o organismo.

Id do cDNA	Seqüência da *tag*	*p-value*	alpha	PE	Sz↑	Resultado de BLAST	# Acesso	Organismo
Et_SAGE0488	aagttcacgaaatatgg	0	1,2E-02	48	6,9	Similar to unknown	ABQ41432	*Eimeria tenella*
Et_SAGE1028	agcgaataagcagggcc	1,0E-02	2,0E-02	13	10,0	Similar to AF361949_1 histone 3	AAM00267	*Eimeria tenella*
Et_SAGE1236	agggtttggaggccctg	0	7,3E-03	122	5,6	Similar to AF361949_1 histone 3	AAM00267	*Eimeria tenella*
Et_SAGE1279	agtctttcggggctcaa	0	1,3E-02	41	8,5	Similar to ribosome associated membrane protein, putative	CAQ38380	*Plasmodium knowlesi*
Et_SAGE1668	caaacctcggcagaaac	7,9E-03	1,7E-02	21	3,9	Similar to ATP-binding cassette transporter sub-family A	XP_001377138	*Monodelphis domestica*
Et_SAGE1996	caaggctttgcggtggc	1,0E-05	3,2E-03	565	1,3	Similar to AF518411_1 facilitative glucose transporter; GT1	AAM69350	*Toxoplasma gondii*
Et_SAGE2219	cacctcggcggctgctg	1,4E-04	8,3E-03	94	2,1	Similar to non-transmembrane antigen	AAO65977	*Toxoplasma gondii*
Et_SAGE3171	cagtatgaggcatacac	2,3E-02	2,7E-02	7	3,2	Similar to ribosome associated membrane protein, putative	CAQ38380	*Plasmodium knowlesi*
Et_SAGE3792	ccgtacacggcaatgga	2,0E-03	1,9E-02	16	10,1	Similar to cathepsin C1	AAZ15654	*Toxoplasma gondii*
Et_SAGE4769	cgtcttgggtatgagag	1,0E-06	7,8E-03	108	2,7	Similar to unknown	ABQ41433	*Eimeria tenella*
Et_SAGE5402	ctggttgtcggcgcctt	1,8E-02	2,1E-02	12	9,6	Similar to hypothetical protein BAL199_03464	ZP_02187509	*alpha proteobacterium*
Et_SAGE5992	gaggtcttggccgcgct	2,0E-06	1,3E-02	40	10,4	Similar to Glutamate/Leu/Phe/Val dehydrogenase family protein	XP_001612092	*Babesia bovis*
Et_SAGE6544	gcgtcttacaggggttt	7,3E-03	2,4E-02	9	4,1	Similar to GK21745 (membrane associated phosphatidic acid phosphatases)	XP_002063107	*Drosophila willistoni*
Et_SAGE7435	gtagacagctctccact	1,5E-05	8,2E-03	98	2,3	Similar to unknown	ABQ41434	*Eimeria tenella*
Et_SAGE7752	taaagttttaacgcaaa	0	6,2E-03	163	4,2	Similar to unknown	ABQ41432	*Eimeria tenella*
Et_SAGE7755	taaagttttaacgcaat	0	9,6E-03	72	10,9	Similar to unknown	ABQ41432	*Eimeria tenella*

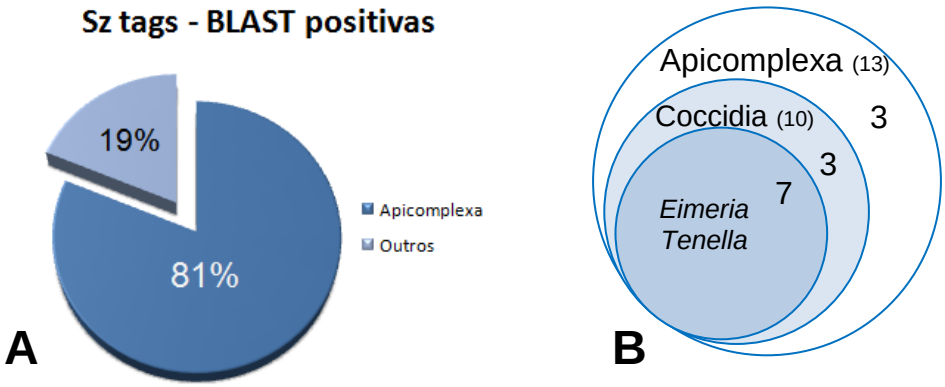


Figura 33 – Distribuição dos 16 produtos protéicos diferencialmente expressos em esporozoítos (*up-regulated*) de acordo com as espécies encontradas nas buscas de similaridade por BLAST. (A) Porcen*tag*em de produtos protéicos similares a seqüências de organismos do Filo Apicomplexa, (B) Diagrama de Venn mostrando a distribuição das seqüências em relação ao Filo Apicomplexa, à classe Coccidia e espécie *E. tenella*.

Assim como em merozoítos, alguns cDNAs reconstruídos a partir de *tags* distintas classificadas como diferencialmente expressas em esporozoítos apresentaram produtos de anotação idênticos. As *tags* 1236 (*sense*) e 1028 (*antisense*) têm como produto de anotação uma Histona 3 de *Eimeria tenella*.

As *tags* 0488, 7752 e 7755, similares a uma proteína não caracterizada de *Eimeria tenella* constituem um exemplo interessante, as *tags* 7752 e 7755 possuem apenas uma base de diferença entre si, correspondente ao último nucleotídeo, devido a este SNP (polimorfismo de um único nucleotídeo) cada uma delas reconstruiu uma seqüência distinta. A *tag* 7752 reconstruiu uma seqüência que possui 2133 pb e uma cauda poliA de mais de 80 nucleotídeos, enquanto que a *tag* 7755 resultou em uma seqüência de 213 pb, não apresentando cauda poliA. Porém, o produto da anotação é idêntico, mas as regiões 3' UTR não, podendo indicar a presença de duas isoformas. De fato, quando estas duas seqüências são alinhadas do nucleotídeo 1 ao 2043 há um alinhamento perfeito, sendo que os 90 nucleotídeos que restaram da seqüência 7752 são diferentes dos 170 nucleotídeos da porção 3' terminal da seqüência 7755. Ambas as isoformas apresentam a *tag* 0488 como *tag antisense*, o que pode indicar a presença de controle pós-transcricional nas duas isoformas deste transcrito.

Grande parte das *tags* diferencialmente expressas em esporozoítos em relação à merozoítos (51 *tags* - 76,1%) não apresentou resultados positivos de

BLAST, conforme mostrado na Tabela 12. Destas *tags* 14, apresentaram diferenças de expressão consideráveis em relação aos merozoítos de 7,2 a 85,9 vezes mais expressas.

Tabela 12 – Seqüências de cDNAs reconstruídas a partir de *tags* diferencialmente expressas em esporozoítos (*up-regulated*), que não apresentaram resultado de BLAST. Estão apresentados na tabela os valores de p após a comparação entre os estágios de merozoítos e esporozoítos (Mz_vs_Sz), alpha (nível critico), poder de expressão (PE) e o número de vezes que as *tags* foram mais expressas em relação a merozoítos (Sz↑).

Id do cDNA	Seqüência da *tag*	*p-value*	alpha	PE	Sz↑
Et_SAGE0411	aaggccgttacccccgc	2,3E-02	2,7E-02	7	3,0
Et_SAGE0620	acactgttcacgggggga	0	2,5E-03	880	1,5
Et_SAGE1067	agctgatttagtgttac	7,9E-03	1,7E-02	21	4,1
Et_SAGE1125	aggactttctcggcgcg	1,0E-02	2,0E-02	13	7,8
Et_SAGE1143	aggcaccggcgaaacgt	4,2E-03	2,3E-02	10	4,9
Et_SAGE1177	aggctttggctgcttct	1,0E-02	2,0E-02	13	7,6
Et_SAGE1503	atgcttttaaaaaaaaa	2,3E-02	2,7E-02	7	3,0
Et_SAGE1504	atgcttttagaaaaaaa	0	9,7E-03	71	6,1
Et_SAGE1508	atgcttttagaattacg	1,2E-02	1,7E-02	20	3,9
Et_SAGE1717	caaagccccaaaggagg	2,7E-05	8,2E-03	97	2,3
Et_SAGE1764	caaaggtgttttcaagg	0	4,8E-03	263	2,4
Et_SAGE1787	caaatcaaacaggaaac	2,4E-03	2,2E-02	11	4,0
Et_SAGE1893	caactgctgattcaatt	0	6,4E-03	157	2,6
Et_SAGE1917	caagaagtgctgaacga	7,2E-04	1,2E-02	45	3,4
Et_SAGE2091	caatttggacggacagc	2,2E-03	1,2E-02	45	2,7
Et_SAGE2198	caccgatctgcaagagt	8,0E-06	1,5E-02	26	16,1
Et_SAGE2320	cactcaagcttggcttg	4,5E-03	1,4E-02	33	3,0
Et_SAGE2443	cagaactcgcctgcggc	4,9E-03	1,6E-02	23	4,4
Et_SAGE2505	cagactaaggcttcttt	1,7E-05	1,2E-02	47	4,6
Et_SAGE2507	cagactcaccttgtcaa	7,3E-03	2,4E-02	9	3,5
Et_SAGE2685	cagcagcagcaaacttg	5,1E-03	1,0E-02	64	1,9
Et_SAGE2977	cagctgctctttatttt	0	9,7E-03	71	5,9

continua

Tabela 12 – (continuação)

Id do cDNA	Seqüência da *tag*	*p-value*	alpha	PE	Sz↑
Et_SAGE3008	cagcttcgcccttctat	1,2E-04	1,5E-02	27	8,4
Et_SAGE3110	cagggcggaggaagcga	4,3E-03	1,2E-02	46	2,6
Et_SAGE3233	cagtgggtggcgcgaaa	2,3E-02	2,7E-02	7	3,6
Et_SAGE3237	cagtgtacctacacctg	1,1E-05	7,4E-03	119	2,2
Et_SAGE3777	ccggaaacaaatgaaca	7,0E-03	1,0E-02	63	2,0
Et_SAGE3786	ccgggtgtctagacagt	2,1E-04	1,7E-02	21	13,1
Et_SAGE3813	cctaaagtggatgtaaa	0	8,6E-03	88	3,8
Et_SAGE4085	cgcaatggcgcagagcc	9,5E-05	1,8E-02	17	7,2
Et_SAGE4103	cgcacttgcgtatataa	2,3E-02	2,7E-02	7	3,0
Et_SAGE4228	cgcctccgcctcaagat	3,5E-05	1,5E-02	29	8,6
Et_SAGE4544	cggcgcattagttgaat	0,0E+00	7,1E-03	127	4,0
Et_SAGE4629	cgggggcgtgctgcgtc	3,4E-04	1,6E-02	24	7,2
Et_SAGE4895	cgttgctttgatgtttt	0	1,0E-02	64	85,9
Et_SAGE5236	ctgcgtatatggaccca	0	4,0E-03	372	1,6
Et_SAGE5322	ctgctggcagaccctac	1,3E-02	2,5E-02	8	3,6
Et_SAGE5372	ctggctcagactctgga	1,2E-02	1,8E-02	17	6,0
Et_SAGE6340	gcccgctctgggtgcct	1,3E-03	1,4E-02	33	3,6
Et_SAGE6672	gctgttcaaaaagaaac	2,3E-02	2,7E-02	7	3,0
Et_SAGE7169	ggggcccacctcaaggg	2,2E-04	1,3E-02	41	3,7
Et_SAGE7226	ggggccctcaagggccc	7,0E-06	1,3E-02	38	7,2
Et_SAGE7398	ggtgtagcagcagaaaa	2,8E-05	1,8E-02	19	8,0
Et_SAGE7490	gtcctcagctaacccac	1,2E-04	1,5E-02	27	9,8
Et_SAGE7559	gtgcataggcgcgaacc	2,8E-05	6,1E-03	168	1,8
Et_SAGE7620	gtgggcaaaagagaagc	0	5,0E-03	251	2,3
Et_SAGE8175	tcccaaagaaggacaat	0	9,1E-03	80	8,2
Et_SAGE8608	tgcatcttgaatggttt	7,9E-03	1,7E-02	21	4,8
Et_SAGE8654	tgccttgtcttcaagga	1,2E-02	1,7E-02	20	4,0
Et_SAGE8703	tgctcctgcgcgggtgc	1,0E-03	7,9E-03	105	1,8
Et_SAGE8801	tggcaacccatctgtga	1,7E-05	1,4E-02	35	7,7

conclusão

Destes produtos protéicos classificados como hipotéticos, 17 *tags* (33,3%) são de alta expressão e as 34 *tags* restantes (66,7%) representam *tags* de média expressão, conforme demonstrado na Figura 34.

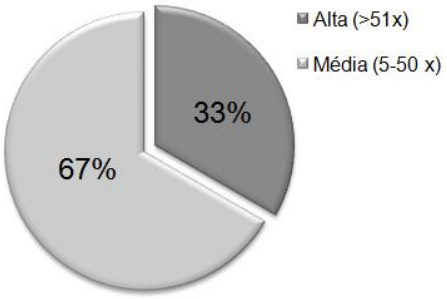


Figura 34 – Distribuição das 51 proteínas hipotéticas resultantes da anotação dos cDNAs reconstruídos a partir das *tags* diferencialmente expressas em esporozoítos (*up-regulated*), de acordo com a freqüência de contagem.

Analisando ainda o conjunto total de 197 *tags* diferencialmente expressas reconstruídas, anotadas e curadas foi possível verificar a presença de *tags antisense* e transcritos *antisense*. Estes casos poderiam simplesmente se tratar de transcritos *sense* apresentados de forma invertida. Para verificar a correta direção destes reconstruídos, estas seqüências reconstruídas foram submetidas a buscas de similaridade por BLASTN contra ESTs de *Eimeria tenella* depositados no *Genbank*. Quando comparadas a ESTs direcionais 5', 18 destes reconstruídos apresentaram alinhamentos com a fita oposta em mais de um *hit* sugerindo que poderiam representar transcritos *antisense*. Além destas, foram encontrados 9 casos de *tag sense* e *antisense* na mesma pontuação (sítio CATG), totalizando 27 *tags antisense* putativas, o que representa 13,7% das *tags* diferencialmente expressas.

4.12.1 Mapeamento de termos de ontologia gênica (GO)

Todas as 197 seqüências analisadas tiveram atribuição de termos GO. De acordo com o nosso protocolo de anotação automática, quando o produto protéico é desconhecido (Interpro), a seqüência é classificada nas três ontologias gênicas (processo biológico, componente celular e função molecular), pois poderia pertencer a qualquer uma das categorias (Tabela 13). Na Figura 35 podemos verificar os termos GO representados na forma de gráfico acíclico direto, o que permite

Tabela 13 – Freqüências de termos GO para as três principais ontologias gênicas (processo biológico, componente celular e função molecular) dos produtos protéicos reconstruídos a partir de *tags* diferencialmente expressas em merozoítos e esporozoítos de *E. tenella*.

Termo GO - Processo Biológico	**# de contagens**
GO:0008150: biological_process	197 (100.00%)
GO:0008152: metabolic process	52 (26.40%)
GO:0044238: primary metabolic process	51 (25.89%)
GO:0019538: protein metabolic process	47 (23.86%)
GO:0009058: biosynthetic process	39 (19.80%)
GO:0006412: translation	38 (19.29%)
GO:0006464: protein modification process	5 (2.54%)
GO:0016043: cellular component organization and biogenesis	4 (2.03%)
GO:0006996: organelle organization and biogenesis	4 (2.03%)
GO:0006810: transport	3 (1.52%)
GO:0006139: nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	3 (1.52%)
GO:0006259: DNA metabolic process	2 (1.02%)
GO:0019725: cell homeostasis	1 (0.51%)
GO:0009605: response to external stimulus	1 (0.51%)
GO:0007165: signal transduction	1 (0.51%)
GO:0007154: cell communication	1 (0.51%)
GO:0007010: cytoskeleton organization and biogenesis	1 (0.51%)
GO:0006950: response to stress	1 (0.51%)
GO:0006519: amino acid and derivative metabolic process	1 (0.51%)
Termo GO - Componente Celular	**# de contagens**
GO:0005575: cellular_component	197 (100.00%)
GO:0005623: cell	48 (24.37%)
GO:0005622: intracellular	44 (22.34%)
GO:0043226: organelle	41 (20.81%)
GO:0005840: ribosome	38 (19.29%)
GO:0005737: cytoplasm	38 (19.29%)
GO:0043234: protein complex	9 (4.57%)
GO:0005694: chromosome	2 (1.02%)
GO:0005634: nucleus	2 (1.02%)
GO:0005856: cytoskeleton	1 (0.51%)
GO:0005576: extracellular region	1 (0.51%)
Termo GO - Função Molecular	**# de contagens**
GO:0003674: molecular_function	197 (100.00%)
GO:0005198: structural molecule activity	38 (19.29%)
GO:0005488: binding	25 (12.69%)
GO:0003824: catalytic activity	15 (7.61%)
GO:0003676: nucleic acid binding	9 (4.57%)
GO:0000166: nucleotide binding	7 (3.55%)
GO:0016787: hydrolase activity	5 (2.54%)
GO:0016740: transferase activity	5 (2.54%)
GO:0005515: protein binding	5 (2.54%)
GO:0016301: kinase activity	4 (2.03%)
GO:0005509: calcium ion binding	4 (2.03%)
GO:0005215: transporter activity	4 (2.03%)
GO:0004672: protein kinase activity	4 (2.03%)
GO:0003723: RNA binding	4 (2.03%)
GO:0003677: DNA binding	4 (2.03%)
GO:0008233: peptidase activity	3 (1.52%)
GO:0008092: cytoskeletal protein binding	2 (1.02%)
GO:0003779: actin binding	2 (1.02%)
GO:0030528: transcription regulator activity	1 (0.51%)

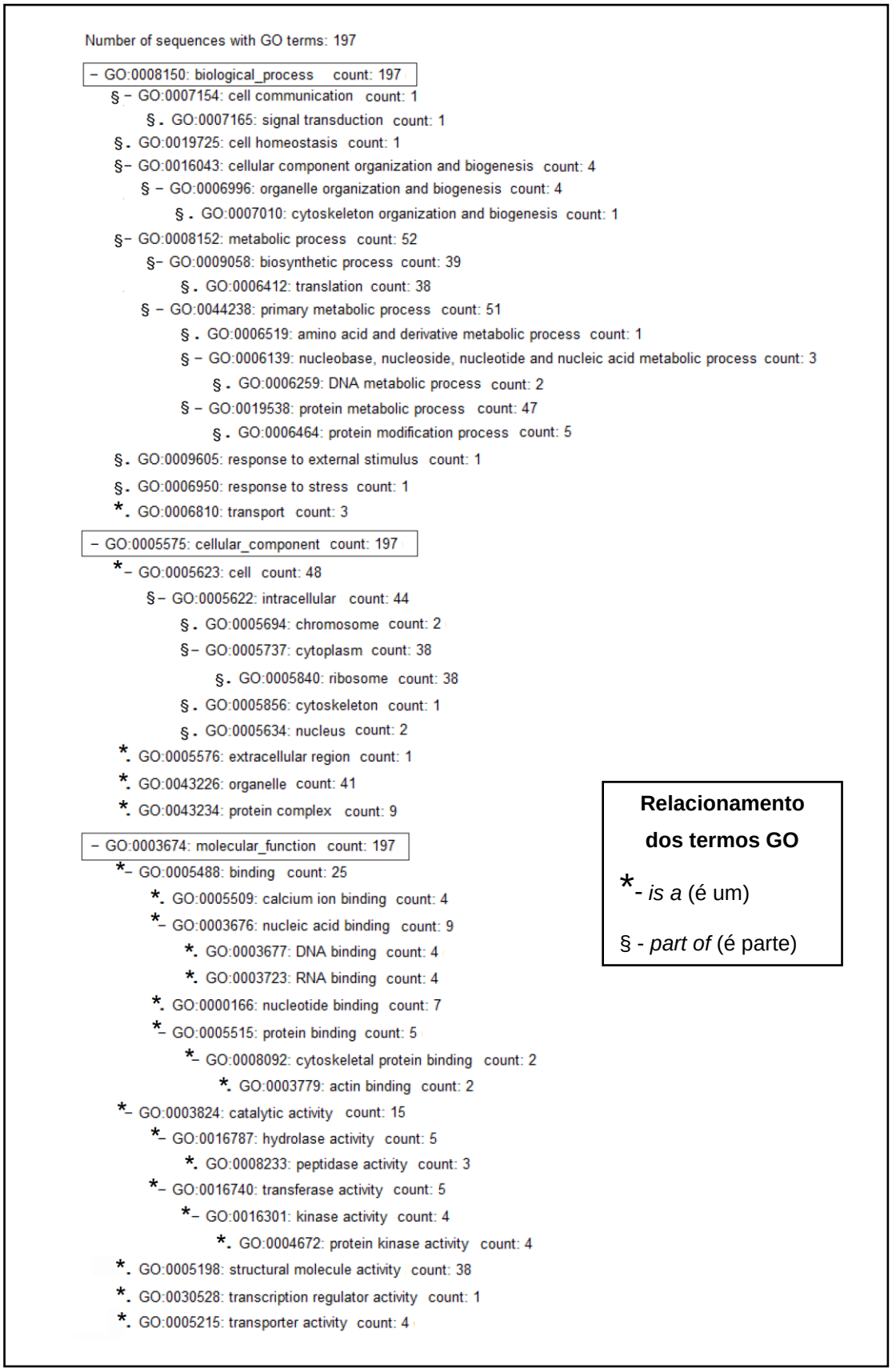


Figura 35 – Gráfico acíclico direto de termos GO das seqüências protéicas codificadas pelos cDNAs reconstruídos a partir de *tags* diferencialmente expressas de merozoítos e esporozoítos de *E. tenella*.

visualizar a relação entre os termos e a hierarquia, possibilitando a distinção entre termos *parents* e *child*. Este gráfico foi gerado utilizando um GO *Slim* genérico e também pode ser visualizado no sítio (http://www.coccidia.icb.usp.br/sage_eimeria).

Do conjunto das 197 seqüências, um total de 72 (36,5%) apresentou resultados positivos nas buscas de similaridade contra as bases de dados do Interpro sendo mapeadas em pelo menos uma ontologia. Destas 55 (76,4%) e 17 (22,5%) são provenientes de seqüências reconstruídas a partir de *tags* diferencialmente expressas em merozoítos e esporozoítos, respectivamente. Como uma seqüência protéica pode apresentar um único termo GO e/ou pode ser mapeada em mais de um termo GO dentro das três ontologias gênicas, para as 72 seqüências foram atribuídos 260 termos de GO, 64 (24,6%) na ontologia de processo biológico, 100 (38,4%) para componente celular e 96 (37%) para função molecular.

O número de sequências que tiveram resultados positivos em cada estágio evolutivo (merozoítos *versus* esporozoítos) foi normalizado para permitir a comparação em termos de porcentagem de seqüências que tiveram termos de GO atribuídos em cada categoria analisada. Na Figura 36, podemos observar a porcentagem de seqüências (normalizadas) as quais foram atribuídos termos GO para cada uma das três ontologias gênicas de acordo com o estágio evolutivo.

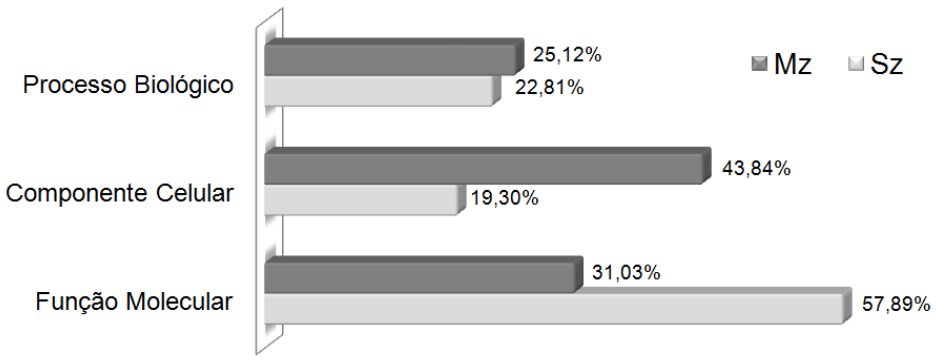


Figura 36 – Porcentagem de seqüências (normalizadas) com termos GO atribuídos para as três ontologias gênicas (processo biológico, componente celular e função molecular), de acordo com os produtos protéicos reconstruídos a partir de *tags* diferencialmente expressas em merozoítos (Mz) ou esporozoítos (Sz).

As seqüências reconstruídas a partir de *tags* diferencialmente expressas em merozoítos apresentaram uma maior porcentagem de produtos protéicos relacionados à categoria componente celular (43,84%), enquanto que para esporozoítos, os produtos protéicos apresentaram uma maior porcentagem de seqüências na categoria função molecular (57,89%).

Nas Figuras 37, 38 e 39 podemos observar a porcentagem de seqüências (normalizadas) reconstruídas a partir de *tags* diferencialmente expressas em merozoítos ou esporozoítos as quais foram atribuídos termos GO para processo biológico, componente celular e função molecular, respectivamente.

Para os termos relacionados ao processo biológico (Figura 37) podemos perceber que a porcentagem de seqüências que tiveram termos GO atribuídos para processo metabólico (GO:0008152) foram relativamente semelhantes entre os dois estágios, no entanto, ao analisar seqüências provenientes de merozoítos, 18% estão relacionadas ao processo biossintético (GO:0009058), e dentre estas, 17,59% estão

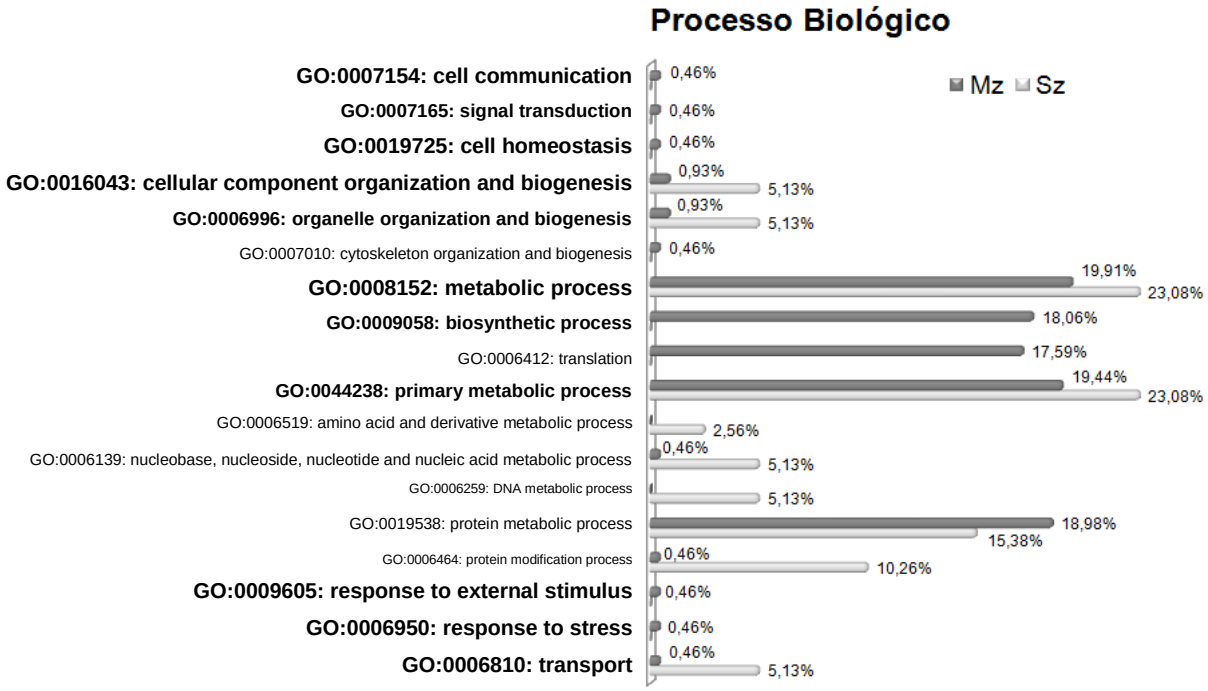


Figura 37 – Porcentagem de sequências (normalizadas) com termos GO atribuídos para processo biológico de acordo com os produtos protéicos reconstruídos a partir de *tags* diferencialmente expressas em merozoítos (Mz) ou esporozoítos (Sz).

relacionadas à tradução (GO:0006412), nenhuma seqüência de esporozoítos foi classificada dentro destas duas categorias. Para esporozoítos podemos observar que 5% das seqüências estão relacionadas à organização dos componentes celulares e biogênese (GO:0016043), 5% para processo metabólico de DNA (GO:0006259), e 5% na categoria transporte (GO:0006810), contra menos de 1% para cada uma destas em merozoítos. Podemos destacar que 10% das seqüências de esporozoítos estão relacionadas ao processo de modificação protéica (GO:0006464).

Após a análise dos termos GO relacionados ao componente celular (Figura 38), 20,39% das seqüências de merozoítos foram classificadas como célula (GO:0005623), das quais 18,45% pertencem aos termos citoplasma (GO:0005737) e ribossomo (GO:0005840). Para esporozoítos 33,33% foram classificadas como célula, sendo 11,11% de cromossomo (GO:0005694) e 11,11% de núcleo (GO:0005634), nenhuma das seqüências de esporozoítos foram classificadas para os termos citoplasma e ribossomo. Para o termo complexo protéico (GO:0043234) 16,67% das seqüências são de esporozoítos enquanto que apenas 2,91% são provenientes de merozoítos

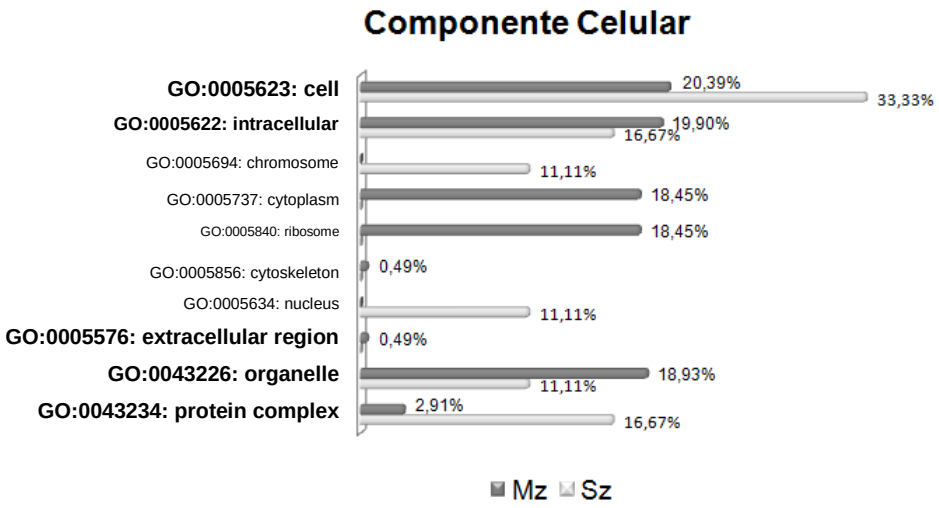


Figura 38 – Porcentagem de sequências (normalizadas) com termos GO atribuídos para componente celular de acordo com os produtos protéicos reconstruídos a partir de *tags* diferencialmente expressas em merozoítos (Mz) ou esporozoítos (Sz).

Na Figura 39 estão representadas as seqüências da ontologia função molecular. A categoria atividade de estrutura molecular (GO:0005198) somente foi observada em merozoítos (42,7% das seqüências). Seqüências classificadas como de ligação apresentaram porcentagens próximas entre esporozoítos e merozoítos, no entanto em esporozoítos 9,62% destas são de ligação a nucleotídeos (GO:0000166) e em merozoítos, 5,62% estão relacionadas ao termo ligação de proteínas (GO:0005515). Para a atividade catalítica (GO:0003824), 19,23% são de esporozoítos contra apenas 5,62% de merozoítos. Em relação à atividade de transporte (GO:0005215) 5,77% das seqüências são provenientes de esporozoítos.

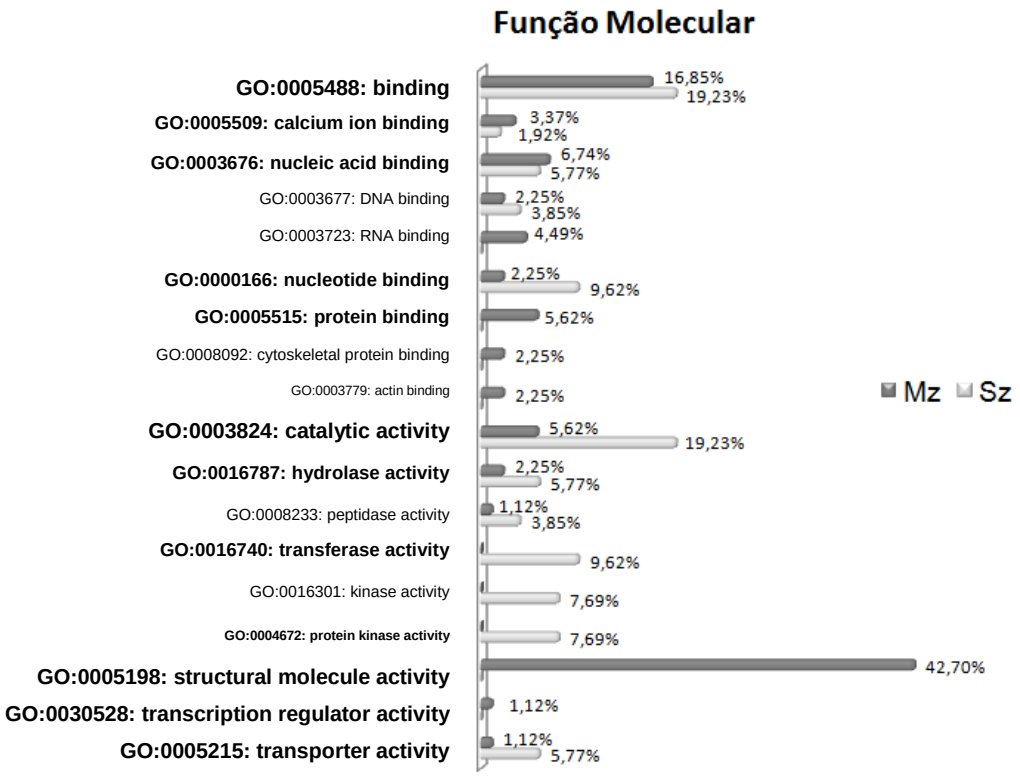


Figura 39 – Porcentagem de sequências (normalizadas) com termos GO atribuídos para função molecular de acordo com os produtos protéicos reconstruídos a partir de *tags* diferencialmente expressas em merozoítos (Mz) ou esporozoítos (Sz).

4.12.2 Anotação utilizando a base de dados KOG (*euKaryotic Orthologous Groups of proteins*)

As 197 seqüências reconstruídas e anotadas a partir de *tags* diferencialmente expressas entre merozoítos e esporozoítos foram submetidas a um processo de anotação utilizando a base de dados KOG (20 categorias). O total de 56 (28,42%) produtos protéicos foram classificados resultando em 9 categorias funcionais conforme mostrado na Tabela 14 e Figura 40, estes resultados também podem se visualizados no sítio (http://www.coccidia.icb.usp.br/sage_eimeria).

A maior parte das seqüências, 89,3% que apresentaram KOGs foi reconstruída a partir de *tags* diferencialmente expressas em merozoítos e apenas 10,7% a partir de seqüências provenientes de esporozoítos (Tabela 14).

Tabela 14 – Classificação dos 56 produtos protéicos reconstruídos a partir de *tags* diferencialmente expressas em merozoítos ou esporozoítos de *E. tenella* em categorias funcionais de KOG.

Categoria KOG	Categoria Funcional	# Total	%Total	# Mz	# Sz
B	Chromatin structure and dynamics	2	3,57	-	2
E	Amino acid transport and metabolism	1	1,79	-	1
I	Lipid transport and metabolism	2	3,57	-	2
J	Translation, ribosomal structure and biogenesis	39	69,64	39	-
K	Transcription	1	1,79	1	-
O	Posttranslational modification, protein turnover, chaperones	4	7,14	4	-
R	General function prediction only	2	3,57	1	1
T	Signal transduction mechanisms	3	5,36	3	-
Z	Cytoskeleton	2	3,57	2	-
TOTAL		**56**	**100**	**50**	**6**

Grande parte das seqüências (69,64%) foi classificada na categoria funcional **J**, que está relacionada à tradução, estrutura ribossômica e biogênese (Figura 37). As demais seqüências foram distribuídas de uma forma mais equilibrada entre as outras oito categorias, cujo porcentual variou de 1,79 a 7,14%. As categorias **K** (transcrição) e **E** (transporte e metabolismo de aminoácidos) apresentaram o menor número de seqüências, cada uma composta por 1,79%. Com 3,57% das seqüências estão representadas as categorias **B** (estrutura e dinâmica da cromatina), **I**

(transporte e metabolismo de proteínas), **R** (predição de função geral) e **Z** (citoesqueleto). A categoria **T** (mecanismos de transdução de sinal) é composta por 5,36% das seqüências, enquanto que a categoria **O** (modificação pós-traducional, modificação de proteínas e chaperonas) apresentou um percentual de 7,14%. Na Tabela 15 estão representadas todas as seqüências que tiveram códigos de KOG atribuídos e suas funções específicas.

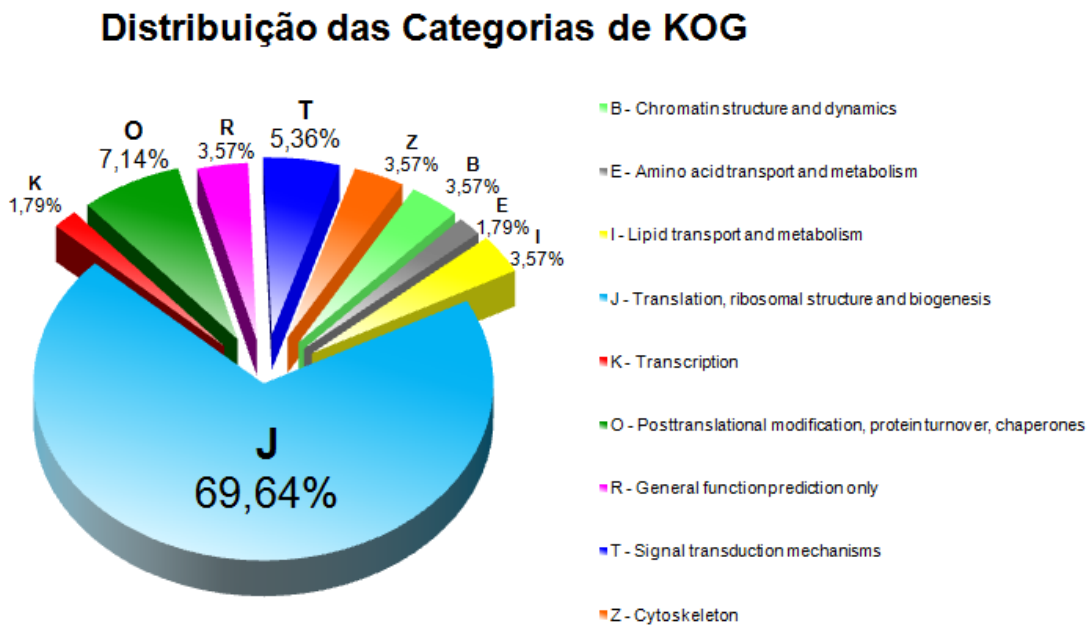


Figura 40 – Distribuição em porcentagem dos 56 produtos protéicos reconstruídos a partir de *tags* diferencialmente expressas em merozoítos ou esporozoítos de *E tenella* classificados de acordo com as categorias de KOG.

Tabela 15 – Classificação dos produtos protéicos reconstruídos a partir de *tags* diferencialmente expressas de *E. tenella* em categorias funcionais de KOG, função específica e número de KOG.

Seqüência Id	Categoria de KOG	Função específica	# KOG
Et_SAGE1236_cDNA	B	Histones H3 and H4	KOG1745
Et_SAGE1028_cDNA	B	Histones H3 and H4	KOG1745
Et_SAGE5992_cDNA	E	Glutamate/leucine/phenylalanine/valine dehydrogenases	KOG2250
Et_SAGE6544_cDNA	I	Lipid phosphate phosphatase and related enzymes of the PAP2 family	KOG3030
Et_SAGE8608_cDNA	I	Phosphatidylinositol transfer protein SEC14 and related proteins	KOG1471
Et_SAGE2994_cDNA	J	60S ribosomal protein L38	KOG3499
Et_SAGE4143_cDNA	J	60S ribosomal protein L38	KOG3499

continua

Tabela 15 – (continuação)

Seqüência Id	Categoria de KOG	Função específica	# KOG
Et_SAGE0675_cDNA	J	40S ribosomal protein S13	KOG0400
Et_SAGE0689_cDNA	J	60S ribosomal protein L37	KOG0402
Et_SAGE5846_cDNA	J	60s ribosomal protein L24	KOG1722
Et_SAGE3957_cDNA	J	60s ribosomal protein L24	KOG1722
Et_SAGE4101_cDNA	J	40S ribosomal protein S23	KOG1749
Et_SAGE0263_cDNA	J	40S ribosomal protein S8	KOG3283
Et_SAGE1740_cDNA	J	60S ribosomal protein L11	KOG0397
Et_SAGE4592_cDNA	J	60S ribosomal protein L30	KOG2988
Et_SAGE8495_cDNA	J	40S ribosomal protein S3	KOG3181
Et_SAGE7689_cDNA	J	60S ribosomal protein L35A/L37	KOG0887
Et_SAGE3646_cDNA	J	60s ribosomal protein L15	KOG1678
Et_SAGE6890_cDNA	J	60s ribosomal protein L2/L8	KOG2309
Et_SAGE9037_cDNA	J	60s ribosomal protein L2/L8	KOG2309
Et_SAGE9512_cDNA	J	60S ribosomal protein L36	KOG3452
Et_SAGE7808_cDNA	J	60S ribosomal protein L18A	KOG0829
Et_SAGE0478_cDNA	J	40S ribosomal protein S12	KOG3406
Et_SAGE4275_cDNA	J	40S ribosomal protein S17	KOG0187
Et_SAGE4673_cDNA	J	40S ribosomal protein S7	KOG3320
Et_SAGE7414_cDNA	J	60S acidic ribosomal protein P2	KOG3449
Et_SAGE1929_cDNA	J	Ubiquitin/40S ribosomal protein S27a fusion	KOG0004
Et_SAGE6107_cDNA	J	40S ribosomal protein S15	KOG0898
Et_SAGE6008_cDNA	J	40s ribosomal protein s10	KOG3344
Et_SAGE5444_cDNA	J	60S ribosomal protein L32	KOG0878
Et_SAGE7411_cDNA	J	60s acidic ribosomal protein P1	KOG1762
Et_SAGE2865_cDNA	J	60s ribosomal protein L19	KOG1696
Et_SAGE3730_cDNA	J	Ribosomal protein S7	KOG3291
Et_SAGE8099_cDNA	J	40s ribosomal protein S26	KOG1768
Et_SAGE9363_cDNA	J	60S ribosomal protein L9	KOG3255
Et_SAGE4699_cDNA	J	40S ribosomal protein S15/S22	KOG1754
Et_SAGE7722_cDNA	J	40S ribosomal protein S2	KOG0886
Et_SAGE7507_cDNA	J	40S ribosomal protein S4	KOG0378
Et_SAGE5210_cDNA	J	60S ribosomal protein L5	KOG0875
Et_SAGE2112_cDNA	J	60S ribosomal protein L31	KOG0893
Et_SAGE0093_cDNA	J	40S ribosomal protein S2/30S ribosomal protein S5	KOG0877
Et_SAGE0125_cDNA	J	60S ribosomal protein L7A	KOG3166
Et_SAGE5007_cDNA	J	40S ribosomal protein SA (P40)/Laminin receptor 1	KOG0830
Et_SAGE0135_cDNA	J	60S ribosomal protein L7	KOG3184
Et_SAGE2948_cDNA	K	Transcription factor containing NAC and TS-N domains	KOG2239
Et_SAGE2794_cDNA	O	Molecular chaperone (HSP90 family)	KOG0019
Et_SAGE8907_cDNA	O	Multifunctional chaperone (14-3-3 family)	KOG0841
Et_SAGE7668_cDNA	O	Thioredoxin	KOG0907
Et_SAGE1072_cDNA	O	Cyclophilin type peptidyl-prolyl cis-trans isomerase	KOG0865
Et_SAGE1996_cDNA	R	Predicted transporter (major facilitator superfamily)	KOG0254
Et_SAGE3991_cDNA	R	FOG: Transposon-encoded proteins with TYA, reverse transcriptase, integrase domains in various combinations	KOG0017
Et_SAGE5438_cDNA	T	G protein beta subunit-like protein	KOG0279
Et_SAGE7853_cDNA	T	Calmodulin and related proteins (EF-Hand superfamily)	KOG0027
Et_SAGE6140_cDNA	T	Calmodulin and related proteins (EF-Hand superfamily)	KOG0027
Et_SAGE1591_cDNA	Z	Actin depolymerizing factor	KOG1735
Et_SAGE5642_cDNA	Z	Actin and related proteins	KOG0676

conclusão

Em seguida, os dados de contagem das seqüências presentes em cada estágio (merozoíto e esporozoíto) foram normalizados e comparados em relação à classificação das categorias KOG (Figura 41). A única categoria compartilhada entre os dois estágios evolutivos estudados é a **R**, que classifica estas proteínas como de funções gerais. As categorias **B**, **E** e **I** estão apenas representadas por sequências de esporozoítos, enquanto que as categorias **J, K**, **O**, **T** e **Z** estão presentes apenas nos estágios merozoítas. Dentre estas, a categoria **J** possui o maior número de seqüências, correspondente a 78%.

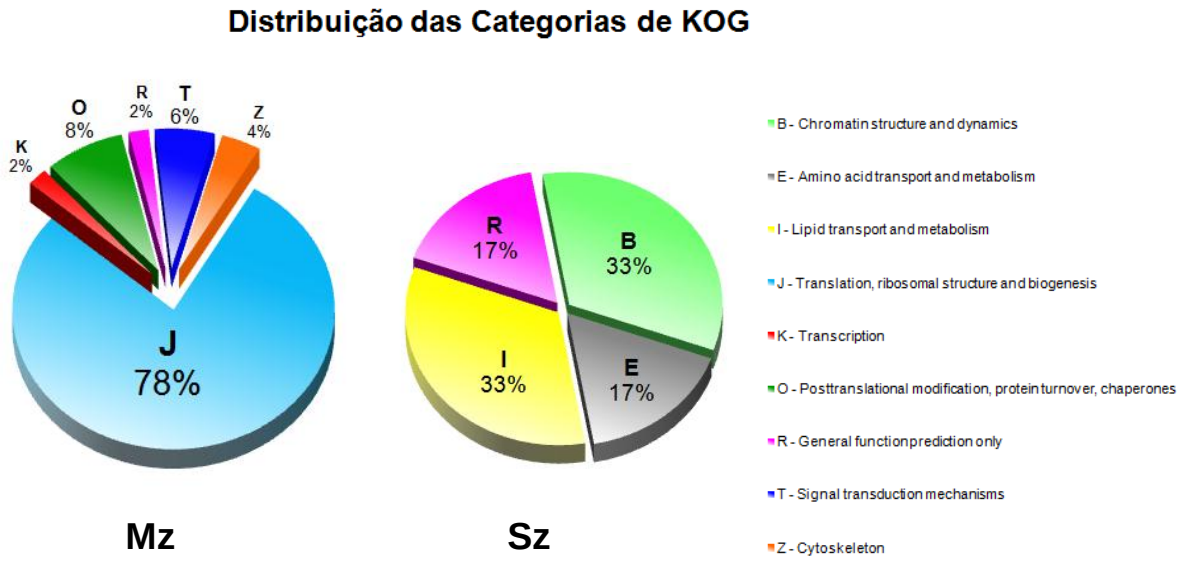


Figura 41 – Porcentagem das sequências classificadas de acordo com as categorias de KOG dividido de acordo com os produtos protéicos reconstruídos a partir de *tags* mais expressas em merozoítos (Mz) ou esporozoítos (Sz).

4.13 PCR em tempo real

Experimentos de PCR em tempo real (qPCR – PCR quantitativo) foram realizados com o objetivo de se realizar uma validação preliminar dos resultados obtidos através da técnica de LongSAGE. A partir das seqüências reconstruídas (item 2.9), *primers* (Tabela 2 e 3) foram desenhados para um pequeno conjunto de genes.

Como optamos pela quantificação relativa da expressão gênica, para os experimentos de RT-qPCR o primeiro procedimento realizado foi a busca de genes expressos de forma constitutiva em merozoítos e esporozoítos. Adicionalmente, a expressão destes genes também foi verificada em 3 estágios de oocistos (não esporulados, esporoblastos e esporulados). Infelizmente, ao se comparar estes 5 estágios, os valores de C_t apresentaram grande variação (>2) (dados não mostrados). Para merozoítos e esporozoítos, o gene escolhido como de referência foi a histona H2A por apresentar a menor variação de C_t nos diferentes experimentos (Figura 42).

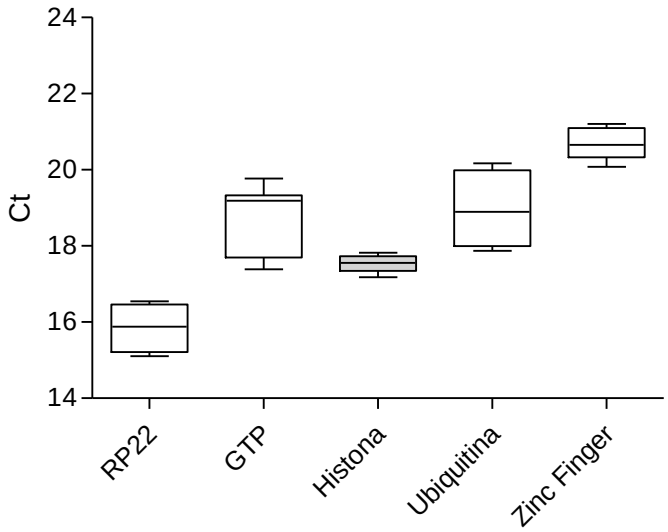


Figura 42 – Variação dos valores de Ct (PCR *cycle threshold*) para os genes candidatos à *housekeeping* genes constitutivamente expressos. As barras correspondem à média dos valores obtidos em diferentes experimentos utilizando amostras de cDNA sintetizadas a partir de RNAm de merozoítos e/ou esporozoítos, as caixas representam o 25º e 75º percentil, e os traços correspondem aos valores mínimos e máximos obtidos.

Em seguida, a expressão dos genes diferencialmente expressos foi avaliada. Um total de 6 genes foi testado e os resultados estão apresentados na Figura 43. De um modo geral, para os genes testados, a expressão diferencial estágio-específica tanto em merozoítos quanto em esporozoítos foi congruente em ambas as técnicas (LongSAGE e qRT-PCR). Em relação aos dados quantitativos, a expressão dos genes diferencialmente expressos em esporozoítos (*up-regulated*) é altamente

concordante para ambas as técnicas, apresentando um coeficiente de correlação de Pearson (P= 0,999) e um R^2= 0,999, entretanto, para merozoítos foi observada uma correlação moderada entre as técnicas (P= 0,67 e R^2=0,45). Analisando os dados em conjunto na Figura 43 (merozoítos e esporozoítos), encontramos uma boa correlação inter-plataformas com P= 0,79 e R^2= 0,63. Os genes cuja expressão diferencial apresentou maior discordância entre as plataformas foram a actina e o fator depolimerizante de actina (Figura 43).

Nos dados de anotação foi observado que os produtos protéicos referentes aos transcritos reconstruídos a partir das *tags* 1.360 (GTP), 2.126 (PUP) e 5.173 (eIF5) bem como as respectivas *tags*, apresentavam um direcionamento *antisense*. Como a síntese do cDNA foi realizada com sucesso ao se empregar *primers forward,* há evidência experimental de que o direcionamento destes transcritos seja *antisense.*

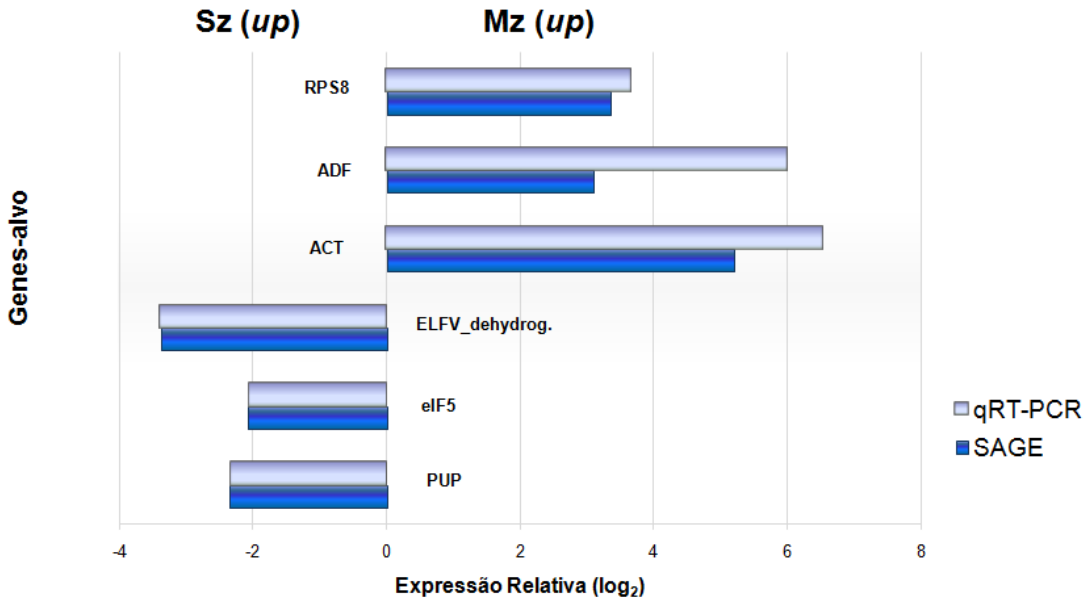


Figura 43 – Comparação entre a expressão relativa (log2) obtida a partir das técnicas de qRT-PCR e LongSAGE para os genes: RPS8 (*Ribosomal Protein S8*), ADF (*Actin Depolymerizing factor*) e ACT (*Actin*), diferencialmente expressos em merozoítos (Mz-*up*)*;* ELFV_dehydrog (*Glut/Leu/Phe/Val dehydrogenase family protein*)*,* eIF5 (Eukaryotic translation initiation factor 5) *e* PUP *(Putative Uncharacterized Protein*), diferencialmente expressos em esporozoítos (Sz-*up*).

4.14 Análise comparativa da expressão gênica em merozoítos e esporozoítos de *E. tenella* obtida por LongSAGE e *Northern* Digital (ESTs/ORESTES)

4.14.1 *Northern* Digital de ESTs/ORESTES de merozoítos de segunda geração e esporozoítos de *E. tenella*

As 48.361 seqüências provenientes de diversos estágios evolutivos (ver item 3.13.1) de *Eimeria tenella* foram agrupadas e montadas (*clustering*) resultando em 8.700 eventos, das quais, 3.724 seqüências reconstruídas (*contigs*) e 4.976 seqüências únicas (*singlets*).

Para análise estatística, apenas os *contigs* (3.008 eventos) contendo seqüências provenientes de merozoítos de segunda geração e esporozoítos foram utilizados. Destes, 2.101 tem leituras provenientes de esporozoítos e 1.787 de merozoítos, apenas 880 eventos possuem leituras de ambos os estágios (29,25%) (Figura 44).

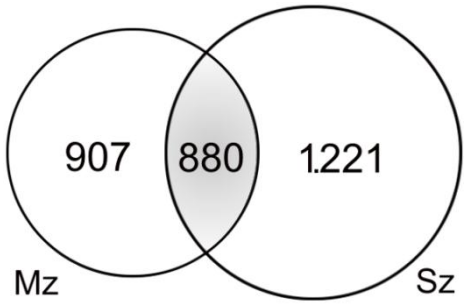


Figura 44 – Diagrama de Venn mostrando os 3.008 eventos de ESTs/ORESTES distribuídos de acordo com estágio evolutivo: Mz (merozoítos de segunda geração) e Sz (esporozoítos).

Dos 3.008 eventos, 660 *contigs* foram classificados como diferencialmente expressos. Após anotação automática, 274 (41,5%) apresentaram resultados positivos de BLAST em buscas de similaridade por BLASTP contra a base de dados nr e 386 (58,5%) produtos protéicos não apresentaram resultados positivos, sendo

classificados como proteínas hipotéticas (http://www.coccidia.icb.usp.br/sage_eimeria). Dos 660 *contigs*, 363 são diferencialmente expressos em merozoítos (*up-regulated*) dos quais 205 (56,5%) apresentaram resultados positivos de BLAST. Para esporozoítos, 297 *contigs* foram classificados como diferencialmente expressos, dos quais apenas 68 (22,9%) apresentaram resultados positivos de BLAST.

Dos 205 produtos protéicos provenientes de merozoítos, 42 são similares a antígenos (41 antígenos de superfície), 53 a proteínas de ribossomo, 4 a produtos similares a proteínas de micronema (4,5 e 8), 3 a HSP (*Heat Shock Protein* – 90, 10 e organelar) e, outros similares a calmodulina, miosinas e fator depolimerizante de actina. Para os 68 produtos protéicos de esporozoítos, encontramos 5 antígenos de superfície dos quais 3 também foram observados em merozoítos, 3 proteínas de micronema (1, 7 e 13), 2 HSP (70 e organelar) e outros produtos de funções diversas como, por exemplo, facilitador do transporte de glicose, subsitilinas, fator de elongação 5, proteína quinase e histonas.

4.14.2 Comparação dos dados de expressão diferencial entre LongSAGE e *Northern* digital

Ao se comparar as 199 seqüências reconstruídas a partir de *tags* (LongSAGE) diferencialmente expressas de merozoítos de segunda geração e esporozoítos contra as 660 seqüências clusterizadas de ESTs/ORESTES destes mesmos estágios observamos que 144 (72,36%) sequências de LongSAGE estão presentes dentro do conjunto de dados diferencialmente expressos de ESTs (Figura 45). Das 144 seqüências, 45 (66,17%) foram reconstruídas a partir de *tags* diferencialmente expressas em esporozoítos, enquanto que 99 (75,57%) a partir de merozoítos. Somente um transcrito apresentou discordância em relação à sua freqüência estágio-específica, nos ESTs foi mais observado (1,5 vezes) em merozoítos, enquanto que nos dados de LongSAGE apresentou-se 1,5 vezes mais expresso em esporozoítos. Ainda analisando estes 144 reconstruídos, 12 produtos protéicos apresentaram diferenças em relação à anotação automática.

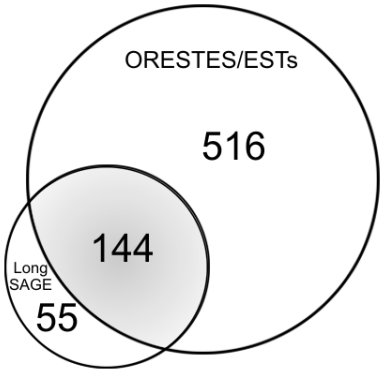


Figura 45 – Diagrama de Venn mostrando a distribuição das sequências diferencialmente expressas de ESTs/ORESTES e de LongSAGE.

Em seguida, os 55 reconstruídos a partir de *tags* de LongSAGE não compartilhadas com os dados de ESTs diferencialmente expressos (Figura 45) foram comparadas aos *contigs* de ESTs classificados como não diferenciais (2.348). Destes reconstruídos, 51 são compartilhados, dentre os quais 37 apresentam concordância em relação à expressão estágio específica e apenas 4 não foram encontrados (dados não mostrados).

4.14.3 Análise comparativa entre os transcritos mais expressos em LongSAGE e *Northern* digital (ESTs/ORESTES)

Os produtos protéicos numericamente mais expressos em cada estágio foram comparados entre as duas plataformas. Das 100 *tags* mais expressas em merozoítos, 68 foram reconstruídas e puderam ser usadas para a comparação inter-plataformas Destas, 42 foram classificadas como diferencialmente expressos (29 apresentaram resultados positivos de BLAST) e 26 como não diferencialmente expressas dos quais somente 12 apresentaram resultados positivos de BLAST (Figura 46). Para esporozoítos 71 *tags* foram reconstruídas dentre estas 35 foram classificadas como diferencialmente expressas (10 com resultados positivos de BLAST) e 36 como não diferenciais (11 com resultados positivos de BLAST) (Figura 47).

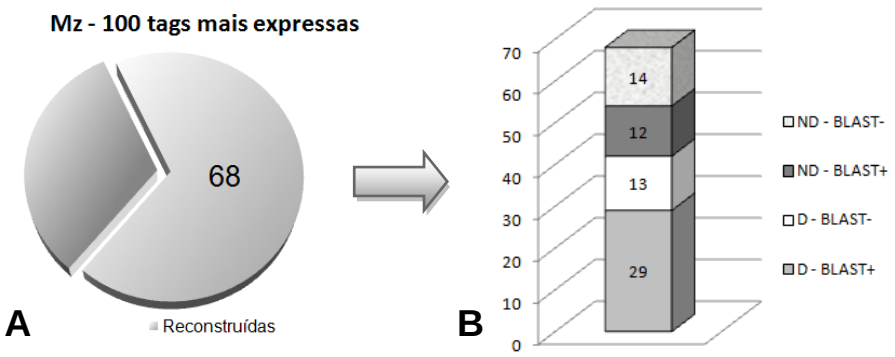


Figura 46 – Distribuição das 100 *tags* mais expressas numericamente em merozoítos. (A) número de *tags* reconstruídas, (B) número de *tags* reconstruídas de acordo com a classificação estatística e anotação automática: ND = não diferencial, D= diferencialmente expressa, BLAST- = não possui resultados de BLAST, BLAST+= possui resultados positivos de BLAST.

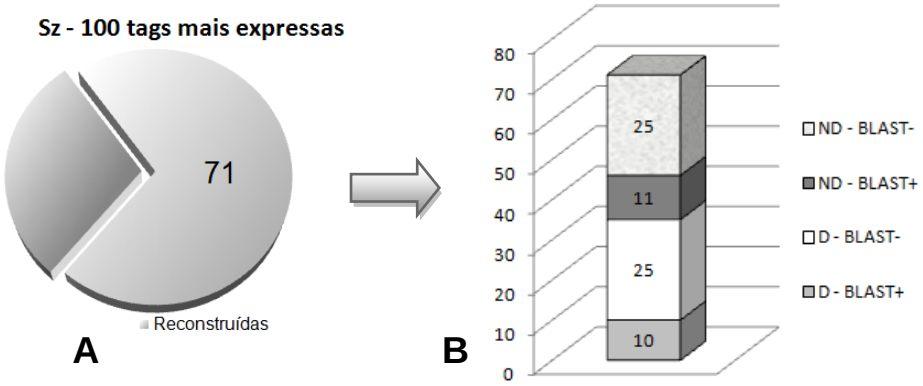


Figura 47 – Distribuição das 100 *tags* mais expressas numericamente em esporozoítos. (A) número de *tags* reconstruídas, (B) número de *tags* reconstruídas de acordo com a classificação estatística e anotação automática: ND = não diferencial, D= diferencialmente expressa, BLAST- = não possui resultados de BLAST, BLAST+= possui resultados positivos de BLAST.

Para as seqüências ESTs/ORESTES, das 100 mais expressas em merozoítos (Figura 48), 93 foram classificadas como diferencialmente expressas (73 com resultados positivos de BLAST) e 7 como não diferenciais (3 com resultados de

BLAST) e das 100 mais expressas em esporozoítos (Figura 48), 89 foram classificadas como diferencialmente expressas (28 com resultados positivos de BLAST) e 11 como não diferenciais (9 com resultados positivos de BLAST).

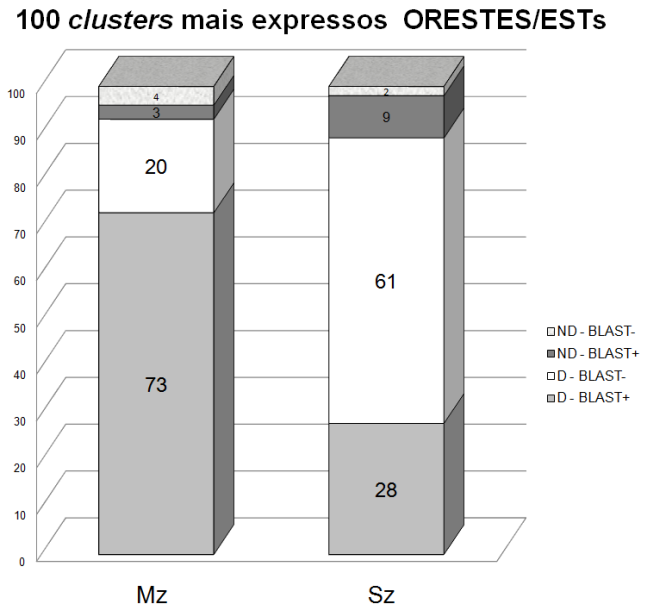


Figura 48 – Distribuição dos 100 *clusters* numericamente mais expressos em merozoítos (Mz) e esporozoítos (Sz) de acordo com a classificação estatística e anotação automática: ND = não diferencial, D= diferencialmente expresso, BLAST- = não possui resultados de BLAST, BLAST+= possui resultados positivos de BLAST.

Quando os produtos protéicos provenientes de merozoítos gerados pelas duas técnicas foram comparados entre si, foi observado que apenas 26 (38,2%) produtos são compartilhados entre estas plataformas (Figura 49), dos quais 4 são hipotéticos (Tabela 16). A mesma análise para esporozoítos resultou em um compartilhamento de 40 (56,3%) produtos protéicos (Figura 49), dentre estes, 24 hipotéticos (Tabela 17).

Figura 49 – Diagrama de Venn mostrando a distribuição das sequências numericamente mais expressas compartilhadas entre os dados provenientes de ESTs/ORESTES e LongSAGE (Mz – Merozoítos; Sz- Esporozoítos).

Tabela 16 – Seqüências de cDNAs reconstruídas a partir de *tags* numericamente mais expressas em merozoítos e compartilhadas entre LongSAGE e *Northern* Digital de ESTs /ORESTES de *E. tenella*. Estão apresentados na tabela a categoria (Cat.), o poder de expressão normalizado (PE(N)), os resultados da busca de similaridade por BLAST contra a base de dados nr, e o organismo.

Id do cDNA	Cat.	PE(N)	Resultado de BLAST	Organismo
Et_SAGE0173	U	27,85	Similar to 40S ribosomal subunit protein S6, putative	Theileria annulata
Et_SAGE0247	U	24,87	Similar to 60S ribosomal protein L22	Theileria parva
Et_SAGE1328	D	22,64	Similar to surface antigen 7	Eimeria tenella
Et_SAGE1431	D	33,96	Similar to 19 kDa sporozoite antigen	Eimeria tenella
Et_SAGE1463	D	172,28	Similar to microneme protein 4	Eimeria tenella
Et_SAGE1591	D	17,17	Similar to actin depolymerizing factor	Eimeria tenella
Et_SAGE1740	D	70,55	Similar to ribosomal protein L11	Solea senegalensis
Et_SAGE1938	D	17,39	Similar to surface antigen 22	Eimeria tenella
Et_SAGE2865	D	21,64	Similar to 60S ribosomal protein L19, putative	Babesia bovis
Et_SAGE3072	D	109,25	hypothetical protein	
Et_SAGE3839	D	27,72	Similar to microneme protein 5	Eimeria tenella
Et_SAGE4219	U	18,81	Similar to 60S ribosomal protein L10, putative	Cryptosporidium muris
Et_SAGE4421	D	102,51	Similar to surface antigen	Eimeria tenella
Et_SAGE5210	D	17,78	Similar to RL5_EIMTE 60S ribosomal protein L5	Eimeria tenella
Et_SAGE5438	D	17,76	Similar to receptor for activated C kinase 1	Toxoplasma gondii
Et_SAGE5642	D	21,88	Similar to ACT_TOXGO Actin	Toxoplasma gondii
Et_SAGE5846	D	78,22	Similar to 60S ribosomal protein L24	Plasmodium chabaudi
Et_SAGE6107	D	91,71	Similar to 40s ribosomal protein s15, putative	Theileria annulata
Et_SAGE6121	D	167,21	Similar to antigen	Eimeria acervulina
Et_SAGE6150	D	18,70	hypothetical protein	
Et_SAGE7324	U	20,91	Similar to microneme protein 4	Eimeria tenella
Et_SAGE7450	D	38,33	Similar to surface antigen 18	Eimeria tenella
Et_SAGE8099	D	17,17	Similar to 40S ribosomal protein S26e	Theileria parva
Et_SAGE8495	D	21,20	Similar to 40S ribosomal protein S3, putative	Cryptosporidium muris
Et_SAGE8631	D	35,19	hypothetical protein	
Et_SAGE_8276	D	20,08	hypothetical protein	

Dos 26 produtos numericamente mais expressos em merozoítos, encontrados tanto por Northern Digital de ESTs/ORESTES, quanto nos dados de LongSAGE grande parte dos genes é representada por proteínas de ribossomo (38,46%), por antígenos (23%), e proteínas de micronema (11,54%). Podemos verificar também a presença do transcrito similar a actina e fator depolimerizante de actina, receptor para ativação de C quinase 1 e 4 proteínas hipotéticas. Grande parte destes transcritos foi classificada como diferencialmente expresso (84,6%) e apenas 4 produtos foram considerados como não diferenciais, os quais incluem 3 proteínas de ribossomo e uma proteína de micronema 4.

Tabela 17 – Seqüências de cDNAs reconstruídas a partir de *tags* numericamente mais expressas em esporozoítos e compartilhadas entre LongSAGE e *Northern* Digital de ESTs /ORESTES de *E. tenella*. Estão apresentados na tabela a categoria (Cat.), o poder de expressão normalizado (PE(N)), os resultados da busca de similaridade por BLAST contra a base de dados nr, e o organismo.

Id do cDNA	Cat.	PE(N)	Resultado de BLAST	Organismo
Et_SAGE0208	U	1,06	hypothetical protein	
Et_SAGE0251	U	1,31	Similar to surface antigen 10	Eimeria tenella
Et_SAGE0488	D	6,90	Similar to unknown	Eimeria tenella
Et_SAGE1236	D	5,59	Similar to AF361949_1 histone 3	Eimeria tenella
Et_SAGE1476	U	1,12	hypothetical protein	
Et_SAGE1504	D	6,11	hypothetical protein	
Et_SAGE1717	D	2,30	hypothetical protein	
Et_SAGE1764	D	2,39	hypothetical protein	
Et_SAGE1873	U	1,35	Similar to hypothetical protein SH1975	Staphylococcus haemolyticus
Et_SAGE1893	D	2,62	hypothetical protein	
Et_SAGE1917	D	3,41	hypothetical protein	
Et_SAGE1996	D	1,28	Similar to AF518411_1 facilitative glucose transporter; GT1	Toxoplasma gondii
Et_SAGE2223	U	1,25	hypothetical protein	
Et_SAGE2505	D	4,59	hypothetical protein	
Et_SAGE3237	D	2,19	hypothetical protein	
Et_SAGE3549	U	1,02	hypothetical protein	
Et_SAGE3612	U	1,19	hypothetical protein	
Et_SAGE3813	D	3,76	hypothetical protein	
Et_SAGE4544	D	4,04	hypothetical protein	
Et_SAGE4769	D	2,71	Similar to unknown	Eimeria tenella
Et_SAGE5236	D	1,61	hypothetical protein	
Et_SAGE5422	U	1,28	Similar to AF420596_1 subtilisin-like protease	Toxoplasma gondii
Et_SAGE5695	U	1,37	Similar to SERPIN1 protein precursor	Eimeria tenella
Et_SAGE5709	U	1,44	Similar to surface antigen 9	Eimeria tenella
Et_SAGE5992	D	10,40	Similar to Glutamate/Leucine/Phenylalanine/Valine dehydrogenase family protein	Babesia bovis

continua

Tabela 17 – (continuação)

Id do cDNA	Cat.	PE(N)	Resultado de BLAST	Organismo
Et_SAGE6177	U	1,25	hypothetical protein	
Et_SAGE6188	U	2,31	hypothetical protein	
Et_SAGE6258	U	1,59	Similar to cystathionine-beta-synthase isoform 8	Pan troglodytes
Et_SAGE6451	U	1,50	Similar to RNA recognition motif. family protein	Cryptosporidium muris
Et_SAGE6749	U	1,45	Similar to dense granule protein 12	Toxoplasma gondii
Et_SAGE7224	U	1,52	hypothetical protein	
Et_SAGE7435	D	2,27	Similar to unknown	Eimeria tenella
Et_SAGE7559	D	1,79	hypothetical protein	
Et_SAGE7620	D	2,32	hypothetical protein	
Et_SAGE7752	D	4,23	Similar to unknown	Eimeria tenella
Et_SAGE7755	D	10,88	Similar to unknown	Eimeria tenella
Et_SAGE8175	D	8,19	hypothetical protein	
Et_SAGE8194	U	1,05	hypothetical protein	
Et_SAGE8801	D	7,66	hypothetical protein	
Et_SAGE9239	U	1,17	hypothetical protein	

conclusão

Para esporozoítos, 40 produtos mais expressos foram encontrados em ambas as técnicas (LongSAGE e Northern Digital de ESTs/ORESTES) e destes, grande parte foi classificada como proteínas hipotéticas (60%). Em relação à expressão gênica, 17 (42,5%) transcritos não foram classificados como diferencialmente expressos sendo que 8 apresentaram resultados positivos de BLAST, incluindo 2 antígenos de superfície, uma proteína de granulo denso, uma proteína que possui motivos de reconhecimento de RNA, um precursor de serpina, entre outros. Dentro dos transcritos classificados como diferencialmente expressos podemos destacar a histona 3, um facilitador do transporte de glicose e uma proteína similar à Glutamato/Leucina/Fenilalanina/Valina desidrogenase.

5 DISCUSSÃO

Dentre as espécies de *Eimeria* de galinha, *E. tenella* apresenta a maior prevalência e virulência no campo, sendo a espécie mais estudada e empregada como modelo para o entendimento da coccidiose aviária (Chapman e Shirley, 2003). Até o momento foram depositadas no *GenBank* cerca de 35.000 sequências ESTs (*Expressed Sequence Tags*) de *E. tenella*, sendo que, 80% destas sequências foram obtidas a partir de estágios esporozoítos e merozoítos. Além destas ESTs, o nosso grupo foi o responsável pelo sequenciamento de mais de 15.000 ESTs do tipo ORESTES para esta mesma espécie. Desta forma um total de 48.361 leituras foi utilizado para agrupamento e reconstrução (*clustering*), resultando em 8.700 eventos, 3.724 *clusters* e 4.976 *singlets*. Este resultado é próximo ao número de genes estimados para *E. tenella*, cerca de 9.000.

Na ausência de dados funcionais para a grande parte dos genes recém isolados as evidências de expressão diferencial podem ser utilizadas para explorar os dados obtidos, tanto na pesquisa básica, quanto na aplicada (Audic e Claverie, 1997).

Como vários estágios evolutivos estão presentes nas leituras de ESTs/ORESTES, foi possível realizar uma análise de *Northern* Digital, utilizando a freqüência de leituras relativa a cada estágio evolutivo presente em cada transcrito reconstruído. No entanto esta abordagem não é exatamente quantitativa, já que este conjunto de dados é formado por bibliotecas de diversas origens e metodologias, como por exemplo, as bibliotecas normalizadas de ORESTES, bibliotecas de subtração e bibliotecas de ESTs convencionais 5’ e 3’. Estes eventos foram submetidos a um protocolo de anotação automática (Ferro, 2008), cerca de 20% dos cDNAs reconstruídos apresentaram resultados positivos de BLAST contra a base de dados *nr*, a maior parte dos cDNAs reconstruídos ainda estão por serem caracterizados. As análises de *Northern* digital mostraram que apenas 20% dos genes são compartilhados entre cada estágio, mostrando um perfil de expressão diferenciado, com conjuntos de genes altamente expressos em cada fase do parasita.

Decidimos realizar uma análise quantitativa em larga escala empregando a técnica de LongSAGE, utilizando as fases de esporozoítos e merozoítos de segunda geração, ambos estágios invasivos e responsáveis pelas fases de proliferação,

críticas para a patogênese no hospedeiro (Kinnaird *et al.*, 2004). A escolha foi baseada na grande disponibilidade de seqüências destas fases, além da sua importância dentro do ciclo de desenvolvimento. Os esporozoítos estão intimamente relacionados aos mecanismos de resposta imune do hospedeiro, além de serem considerados como o estágio ideal para o desenvolvimento de vacinas recombinantes a partir de antígenos específicos. Em teoria, a interrupção da atividade dos esporozoítos seria capaz de prevenir a infecção (Min *et al.*, 2004; Shirley *et al.*, 2005). Os merozoítos de segunda geração por sua vez, são responsáveis pelos maiores danos causados ao intestino (Schmatz, 1997).

LongSAGE é uma variação da técnica de SAGE que gera *tags* mais longas, de 21pb, o que facilita a identificação inequívoca de uma *tag* experimental em um banco de dados como também aumenta a sua taxa de mapeamento (Malig *et al.*, 2006; Pleasance *et al.*, 2003).

Considerada como uma plataforma aberta de larga escala, SAGE possibilita a descoberta de novos genes, incluindo genes de baixa abundância e pode auxiliar a validação das predições in *silico* e de regiões gênicas não anotadas (Saha *et al.*, 2002). Por ser capaz de gerar um perfil digital, esta metodologia também permite catalogar tanto a identidade quanto a freqüência dos transcritos de uma determinada amostra experimental (Tuteja e Tuteja, 2004).

Apesar da estratégia de LongSAGE ser conceitualmente similar ao seqüenciamento de ESTs, é pelo menos uma ordem de magnitude mais eficiente, permitindo a descoberta de genes raros (Saha *et al.*, 2002). De fato, Sun *et al.* (2004) observaram uma sensibilidade 26 vezes maior de detecção de transcritos raros empregando SAGE quando comparada a ESTs. Provavelmente esta diferença se deva ao fato de que um número maior de seqüências é coletado pela técnica de SAGE e, como o tamanho das *tags* é muito semelhante, não há predileção na detecção destes transcritos com relação ao seu tamanho (Keime *et al.*, 2007). Além disso, a quantidade de dados de expressão gênica chega a ser 20 vezes maior em SAGE quando comparada à mesma quantidade de ESTs seqüenciadas (Moody, 2001), mesmo considerando que 1% dos transcritos possivelmente não são detectados, devido a ausência do sítio CATG (Unneberg *et al.*, 2003).

Quando comparamos LongSAGE a outras técnicas de plataforma aberta em larga escala, como o SAGE convencional e o MPSS, o LongSAGE apresenta vantagens em relação ao mapeamento, detecção de transcritos de baixa abundância

e descoberta de novos genes. Analisando o viés composicional GC, LongSAGE apresenta o menor viés quando comparado a estas mesmas plataformas (Siddiqui *et al.*, 2006).

Li *et al.* (2006) ao comparar bibliotecas de SAGE convencional *versus* LongSAGE encontrou uma maior especificidade no mapeamento, maior descoberta de genes e uma menor chance de falsos positivos e ainda foi capaz de detectar um número maior de genes diferencialmente expressos.

O MPSS é uma técnica muito semelhante ao SAGE. É baseada no isolamento de pequenas seqüências próximas à cauda PoliA, combinando clonagem *in vitro* de cDNAs na superfície de *microbeads* à uma alta capacidade de seqüenciamento não baseado em gel, gerando desta forma, milhões de seqüências por experimento. Apesar de ser protegida por direitos autorais e de ter sido descontinuada, diversas bibliotecas empregando-se MPSS foram construídas. Estudos relatam que o aumento do número de bibliotecas de MPSS não necessariamente aumenta o número de seqüências novas, o que provavelmente deve estar relacionado à algum viés durante a construção da biblioteca (Hene *et al.*, 2007). De fato, o sítio para a enzima *Dpn*II mesmo apresentando 4 bases é cerca de 50% mais raro do que a utilizada em LongSAGE. Apesar do MPSS ser capaz de gerar um volume muito maior de dados, a técnica de LongSAGE é capaz de identificar mais genes com um mapeamento 6 vezes maior (Siddiqui *et al.*, 2006).

Ao contrário das plataformas fechadas, LongSAGE não depende de conhecimento a priori de seqüências genômicas (Calsa e Figueira, 2007), o que é vantajoso em relação aos métodos de hibridização (*microarrays*) que possuem alta capacidade de análise de dados em larga escala, porém dependem do conhecimento prévio das seqüências a serem estudadas (Alba *et al.*, 2004; Skuce *et al.*, 2005; Yamamoto *et al.*, 2001). Normalmente os sinais de hibridização não são facilmente reprodutíveis, necessitando de replicatas experimentais e, além disso, podem ser afetados por inúmeras propriedades desconhecidas, como por exemplo, complexidade da biblioteca de cDNA (Audic e Claverie, 1997). Genes pertencentes a famílias gênicas altamente conservadas podem também gerar resultados equivocados devido a hibridização cruzada (Cleary *et al.*, 2002). Em particular, *microarrays Affymetrix* possuem um viés para seqüências ricas em AT (Lu *et al.*, 2004; van Ruissen *et al.*, 2005) podendo alterar os dados de expressão gênica. Outra plataforma fechada muito utilizada é o real-time PCR (RT-qPCR) que apesar

de possuir alta sensibilidade, gera dados em baixa escala limitados à transcritos conhecidos (Wang, 2007).

Considerando que em 2003, época em que projeto de pesquisa foi concebido, os *microarrays* eram extremamente caros e associado ao fato de que a caracterização das seqüências de *E. tenella* era bastante limitada, optamos por utilizar a técnica de LongSAGE para estudos de expressão gênica em *E. tenella.* A análise de transcriptomas de apicomplexa empregando-se SAGE tem evidenciado que os mecanismos transcricionais possuem um papel fundamental no desenvolvimento destes parasitas (Meissner *et al.*, 2007).

Nesta época, os estudos estatísticos e análises de SAGE não eram tão bem conhecidos e explorados e decidimos construir 4 bibliotecas, 2 de cada estágio infectante. Atualmente, recomenda-se a construção de apenas uma biblioteca para cada estágio ou célula a ser estudada, não havendo a necessidade de replicatas experimentais (Dinel *et al.*, 2005) e, para ser ter uma maior sensibilidade, confiabilidade e reprodutibilidade deve-se aumentar o número de *tags* seqüenciadas (Wang, 2007). Desta forma, no presente estudo, as duas bibliotecas de cada estágio foram analisadas em conjunto, formando uma biblioteca única para cada estágio. A geração de cerca de 35.000 *tags* é um número pequeno para a cobertura de todos os genes de *Eimeria tenella*, embora o número de *tags* únicas encontradas 9.516 seja bem próximo ao número de genes estimados para esta espécie. De fato, como este projeto era vinculado a um projeto de pesquisa individual financiado pela FAPESP (Fundação de Amparo à pesquisa do Estado de São Paulo) foi solicitada verba adicional para o sequenciamento de um número maior de leituras, mas infelizmente, a mesma foi negada. Independentemente de nossas escolhas, um projeto em larga escala para o seqüenciamento de *tags* de LongSAGE é inédito para o gênero *Eimeria*.

Para a construção das bibliotecas de *E. tenella*, uma das maiores dificuldades enfrentadas foi a obtenção de uma grande quantidade de esporozoítos para posterior extração de RNAm. Apesar de cada oocisto conter 8 esporozoítos, a metodologia de purificação destes parasitas envolve múltiplas etapas e é extremamente laboriosa. Os oocistos devem ser parcialmente quebrados (assim como acontece na moela da ave) de forma a liberar os esporocistos, entretanto, qualquer alteração da amostra, como por exemplo, a presença de debris, pode alterar o rendimento. A etapa seguinte consiste em uma mimetização da ação dos

sais biliares e tripsina que ocorre no intestino delgado da ave, desta forma, os esporocistos são submetidos a uma etapa de excistação *in vitro* para a digestão do corpúsculo de Stieda e saída ativa dos esporozoítos, este processo pode demorar cerca de 3 horas.

Após a excistação, os esporozoítos são purificados (separados dos esporocistos que não liberaram seus esporozoítos, dos oocistos não rompidos e dos *debris*). Para tanto, utilizamos um método híbrido de purificação baseado em uma coluna de troca iônica e um método de filtração. Em cada passo deste processo existem perdas, e no final cerca de 5 a 10% do material inicial é purificado, ou seja, para cada oocisto obtem-se menos de 1 esporozoíto purificado. Levando em consideração que estes parasitas somente são obtidos através de passagens *in vivo* e que cada ave infectada é capaz de produzir cerca de 2 a 5.10^7 oocistos, como foi purificado um número elevado de esporozoítos, grande quantidade de animais foi utilizado.

A implementação de uma nova técnica, normalmente vem acompanhada de algumas dificuldades iniciais. Em nosso caso, a primeira biblioteca construída não foi concluída com sucesso, já que não foi obtida quantidade suficiente de concatâmeros para as etapas posteriores de clonagem e seqüenciamento das *tags*. A partir deste resultado, tivemos que adaptar o protocolo original, utilizando uma quantidade de RNAm inicial muito maior do que a recomendada (cerca de 20 vezes). É importante ressaltar que, para garantir a qualidade de nossos dados, além da alta qualidade de RNAm empregado, todos os testes de controle sugeridos pelo kit foram realizados.

Diferentemente das outras técnicas, SAGE é altamente dependente da qualidade do seqüenciamento usado para gerar os dados onde uma única base errada pode alterar os dados de expressão (Green *et al.*, 2001). Baseando-se nesta premissa decidimos utilizar um critério de qualidade elevado para a seleção das seqüências. Grande parte dos grupos que trabalham com seqüenciamento de DNA, e mesmo com SAGE, adota a qualidade Phred 20 como aceitável, o que corresponde a um erro esperado a cada 100 pb. Assim, utilizando-se *tags* de 21 pb, cerca de uma em cada cinco *tags* (20%) poderia estar comprometida. Em função dessa taxa de erro ser muito alta, foi utilizado o valor de Phred 32, o qual corresponde a 1 erro esperado para cada 1.585 pb. Assim para um tamanho de *tag* de 21 pb, somente 1 em cada 75 *tags* (1,33%) teria um erro de seqüenciamento.

Ainda, como as *tags* que apresentam erros derivados de seqüenciamento possuem normalmente uma ou duas bases trocadas (Chen *et al.*, 2002; Kim *et al.*, 2006), o critério de qualidade Phred 32 foi aplicado para 100% das bases das *tags*.

Considerando que na etapa de replicação dos clones bacterianos anterior ao plaqueamento pode ocorrer duplicação de concatâmeros (Dinel *et al.*, 2005), para o pré-processamento das seqüências, um novo componente foi implementado ao Egene (Durham *et al.*, 2005). Este componente compara todas as seqüências entre si e os concatâmeros comuns do tamanho de 4 *tags* são mascarados.

O kit de LongSAGE vem acompanhado de um *software* para a extração das *tags*, no entanto, este programa possui uma documentação escassa, o método de extração não é explícito, além do código não ser aberto, o que nos fez procurar por outras opções. Além disso, poucos programas para extração de *tags* de SAGE estão disponíveis, nem sempre são facilmente utilizáveis e grande parte destes é destinada apenas à extração de *tags* de SAGE convencional. Desta forma, com a experiência e tradição do nosso grupo no desenvolvimento de ferramentas de bioinformática versáteis e aplicáveis a qualquer organismo, foi criado um pacote de programas denominado de SAGE Analysis destinado à extração e contagem de *tags* provenientes de qualquer plataforma de expressão digital, como SAGE, LongSAGE, MPSS entre outras, este pacote apresenta alta flexibilidade e é compatível com os sistemas operacionais Windows, Linux e Macintosh. Para validação deste software, foi desenvolvido um pacote de programas, o SAGE Suíte. Este pacote é capaz de simular a síntese de uma biblioteca de SAGE, gerar listas de *tags* randômicas de diferentes tamanhos com contagens facilmente modeláveis, arquivos de *ditags* e de concatâmeros abordando diversos parâmetros além de permitir a introdução de erros, possibilitando testar e validar a recuperação de *tags* pelo pacote SAGE Analysis.

O pacote SAGE Analysis também foi importante para atender as peculiaridades da técnica de LongSAGE. A enzima *Mme*I cliva o DNA gerando extremidades coesivas de duas bases (Saha *et al.*, 2002), ao contrário do SAGE convencional que emprega outra enzima que forma extremidades cegas. Devido a presença de extremidades coesivas, a ligação das *tags* para formação de *ditags* em LongSAGE não é randômica, diferentemente do observado para SAGE convencional. Segundo Velculescu *et al.* (1995) para SAGE convencional todas as *ditags* idênticas deveriam ser excluídas durante o processo de extração das *tags*,

pois podem representar artefatos da etapa de amplificação, entretanto, em organismos onde um pequeno número de genes é altamente expresso em cada fase de vida, a exclusão das *ditags* idênticas, pode acarretar numa quantificação subestimada destes genes. Trabalhos mais recentes, ao invés de simplesmente excluir todas as *ditags* idênticas, empregam um modelo estatístico capaz de estimar a proporção de *ditags* idênticas que representariam artefatos (Dinel *et al.*, 2005). Para LongSAGE, como a ligação das *tags* não ocorre de forma randômica, a retirada de *ditags* repetidas leva a uma alteração nos dados de expressão gênica podendo introduzir erros (Emmersen *et al.*, 2007; Khattra *et al.*, 2007).

Outra peculiaridade, refere-se ao tamanho das *ditags*, a enzima *Mme*I cliva em sítios de corte localizados a 21 e 19 bases *downstream* ao sítio de reconhecimento, e alternativamente também a 20 e 18 bases (Nakonieczna *et al.*, 2009). Como é comum observar *ditags* de 32 pb (duas *tags* de 17 pb subtraindo-se a sobreposição de duas bases das extremidades coesivas), 33 pb (uma *tag* de 17 pb e outra de 18 pb, menos a sobreposição) e, finalmente, *ditags* de 34 pb (duas *tags* de 18 pb), o pacote SAGE Analysis foi concebido de forma a permitir a escolha do tamanho mínimo e máximo das *ditags.* Este fenômeno de variação do tamanho das *tags* já havia sido relatado por Akmaev e Wang (2004) que observaram que mais de 50% das *ditags* apresentavam 33 pb, fato este também observado em nossos dados.

Para aumentar a recuperação das *tags* a partir das seqüências dos concatâmeros, uma estratégia de recuperação de *tags* das pontas (*monotags*) também foi implementada aumentando o rendimento total de *tags* em cerca de 5%.

Analisando os nossos dados, foram obtidas 35.248 *tags* totais, das quais 9.516 eram únicas. Destas apenas 1.975 *tags* únicas foram encontradas em ambos os estágios evolutivos estudados, indicando novamente a presença de genes regulados de acordo com estágio evolutivo, semelhante ao encontrado em *Toxoplasma gondii*, da qual a análise global de RNAm (utilizando mais de 8000 genes) indicou uma expressão gênica altamente dinâmica e estágio específica (Xia *et al.*, 2008).

Com relação à distribuição das *tags*, mais de 88% das *tags* únicas apresentaram uma contagem inferior a 5, *tags* com con*tag*ens entre 5 e 50 representaram 11,1%, e *tags* altamente expressas, menos de 1% corresponderam a 30,2% de todas as *tags*. Nas análises de SAGE de cana de açúcar, 70% dos

transcritos apresentaram contagem igual ou menor a 5 (Calsa e Figueira, 2007), assim como 80% das *tags* provenientes de SAGE de arroz tiveram contagem inferior a 5 (Bao *et al.*, 2005). Em *Toxoplasma gondii,* 70% das *tags* apresentaram contagem entre 2 a 10 vezes (Radke *et al.*, 2005). Portanto, esta alta taxa de *tags* com baixa expressão não é uma peculiaridade de *Eimeria*. Um perfil semelhante, com poucos genes representando um grupo de transcritos altamente expressos também foi encontrado em *Plasmodium falciparum*, do qual 0,3% das *tags* únicas representavam os genes de alta expressão (Patankar *et al.*, 2001). Considerando que a *E. tenella* apresenta um perfil transcricional com grupos de genes estágio-específicos, o padrão de distribuição da freqüência das *tags* parece ser coerente. Em uma célula somática típica os RNAm são divididos em três classes, altamente expressos, que normalmente são 10-15 genes que representam de 10-20% da quantidade de RNAm total, genes intermediários (1.000-2.000) que representam 40-45% do total e finalmente os transcritos raros (15.000-20.000) representando o restante (Sterky e Lundeberg, 2000).

Assim como esperado, grande parte das *tags* apresentou contagem igual a um, correspondente a 66,5% das *tags* únicas e 18% das *tags* totais, este achado é extremamente comum em bibliotecas de SAGE, devido à alta sensibilidade da técnica. A alta freqüência de *tags* com contagem igual a 1 não é uma característica única de nossas bibliotecas, vários trabalhos relatam bibliotecas de SAGE com distribuição similar. Keime *et al.* (2007) mapeando *tags* de LongSAGE no genoma humano observou que 13% das *tags* totais possuíam apenas uma contagem e que estas representavam 75% do conjunto total de *tags*. Dados semelhantes foram relatados por Gibbings *et al.* (2003) e Nielsen *et al.* (2005) que encontraram 70% de *singletons* em seus dados. Nas bibliotecas de *Leishmania donovani* cerca de 70% das *tags* também apresentavam contagem igual a um (Li *et al.*, 2008).

Segundo Wang (2007), em geral 70 a 80% das *tags* apresentam con*tag*ens únicas, e a única maneira de diminuir esta alta porcentagem está em aumentar o número de *tags* seqüenciadas. Entretanto, mesmo com um número elevado de *tags* a taxa de *singletons* não reduz tão drasticamente. Chen *et al.* (2002), analisaram 101 bibliotecas de SAGE de humanos, com o objetivo de verificar se a alta ocorrência de *tags* de contagem igual a um seria reduzida a medida que um maior volume de dados fosse obtido. Ao averiguar apenas uma biblioteca, 75% de *tags* de cópia única foram encontradas, aumentando-se este número para dez bibliotecas a

taxa de *singletons* cai para 50%, e a análise de 101 bibliotecas resultou ainda em 48% de *singletons,* demonstrando a alta sensibilidade da técnica em detectar transcritos de baixíssima abundância. Muitos destes transcritos podem estar representados por menos de uma cópia por célula podendo ser expressos em pontos muito específicos do ciclo celular ou em resposta a um nível particular de stress celular em subconjuntos populacionais (Hene *et al.*, 2007).

Tags com contagem igual a um ainda são consideradas por muitos autores como artefatos (Keime *et al.*, 2007; Khattra *et al.*, 2007; Pleasance *et al.*, 2003) e vários trabalhos eliminam estas *tags* das análises (mesmo que estas representem cerca de 70% das *tags* únicas obtidas), como por exemplo, em *Arabidopsis* (Fizames *et al.*, 2004), *Schistossoma mansoni* (Ojopi *et al.*, 2007), arroz (Gibbings *et al.*, 2003), entre outros. Entretanto, segundo Keime *et al.* (2007), a eliminação de *tags* com contagem igual a um não elimina todas as *tags* que apresentam erros de seqüenciamento. Poroyko *et al.* (2005) utilizando a técnica de GLGI (*Generation of Longer cDNA fragments from serial analysis of gene expression tags for Gene Identification*) (Chen *et al.*, 2002), testaram 55 *tags* de contagem igual a um e provaram que 63% destas *tags* correspondiam a transcritos verdadeiros.

Em nosso trabalho optamos pela não eliminação das *tags* de cópia única, já que utilizamos critérios altos de qualidade na seleção das seqüências. Além disso, se estas *tags* representam transcritos reais, sua eliminação acarretaria numa análise distorcida dos dados de expressão gênica, diminuindo assim uma das vantagens desta técnica, que é exatamente a detecção de transcritos de baixa abundancia, normalmente não detectados através de outras metodologias (Wang, 2007).

Para o mapeamento das *tags* no genoma, utilizamos uma abordagem bem simples. O conjunto de *tags* virtuais extraídas levou em consideração apenas a presença do sítio de restrição para enzima *Nla*III, não avaliando a presença de íntrons, éxons e outros critérios mais acurados. Esta abordagem, no entanto, teve que ser empregada já que o genoma de *E. tenella* está montado em 4.707 *contigs* e não está completamente anotado. Metodologia semelhante foi utilizada para análises de SAGE em *Plasmodium falciparum* (Patankar *et al.*, 2001). No nosso trabalho obtivemos uma taxa de mapeamento de 59%, embora o grau de mapeamento não seja alto, esse é compatível ao relatado por outros autores para vários organismos (Fizames *et al.*, 2004; Funaguma *et al.*, 2007; Ojopi *et al.*, 2007; Poroyko *et al.*, 2005; Wahl *et al.*, 2004). Mesmo para humanos e ratos que são

organismos bem estudados, cerca de 38% das *tags* não foram encontradas no genoma (Keime *et al.*, 2007; Wahl *et al.*, 2005a). Se considerarmos que as bases de um genoma fossem distribuídas de forma equilibrada, em teoria 99,8% das *tags* de 21 pb deveriam ser mapeadas em apenas um único sítio genômico da ordem de $3x10^9$ nucleotídeos (genoma humano). Apesar dos cálculos apontarem para uma alta taxa de mapeamento inequívoco, a análise de 16.000 genes humanos apontou que apenas 75% das *tags* foram encontradas em um único sítio neste conjunto (Saha *et al.*, 2002).

Analisando o nosso conjunto de *tags* mapeadas no genoma (59%), a alta taxa de *tags* mapeadas em um único sítio (95%) é considerada como um fator positivo, já que uma das maiores preocupações em relação à técnica de SAGE é o mapeamento inequívoco destas pequenas seqüências (Malig *et al.*, 2006).

Assim como as *tags* com contagem igual a 1, as *tags* não mapeadas também são consideradas por muitos como artefatos. Estas podem ser o resultado de digestão incompleta com a enzima de restrição *Nla*III (o que não parece ser o nosso caso, já que efetuamos os controles e nenhuma das *tags* apresenta em sua seqüência sitio CATG para reconhecimento da enzima), poliadenilação alternativa, splicing na extremidade 3' UTR, junção entre dois éxons, regiões polimórficas dos transcritos, SNPs e presença de bases da cauda PoliA na seqüência das *tags* (Harbers e Carninci, 2005; Keime *et al.*, 2007). Kuo *et al.* (2006) analisando *tags* preditas de transcritos conhecidos observaram que de 2 a 6% destas *tags* recaem sobre junções de *splicing.* Chen *et al.* (2002) comprovaram que a maioria das *tags* não encontradas no genoma representa *tags* verdadeiras. Foram testadas mais de 1000 *tags* e destas aproximadamente 400 correspondiam a *tags* não mapeadas, estes autores observaram que 70% das *tags* não mapeadas puderam ser caracterizadas como transcritos reais. Rivals *et al.* (2007) também foram capazes de detectar novos transcritos a partir de *tags* não mapeadas.

Uma das mais importantes aplicações de SAGE está na identificação de genes diferencialmente expressos (Tuteja e Tuteja, 2004). Assim, além do desenvolvimento de um software para extração e contagem das *tags*, decidimos utilizar uma ferramenta estatística para análise dos dados e classificação dos genes diferencialmente expressos. O Kemp é um programa baseado em um teste freqüentista desenvolvido para a comparação de perfis de expressão digital e já foi empregado com sucesso para análise de bibliotecas de SAGE de carcinoma de

células escamosas da laringe (Silveira *et al.*, 2008). Outra vantagem da aplicação deste teste está na sua maior eficiência para a avaliação das *tags* de baixa freqüência, quando comparado ao teste χ^2 (Varuzza, 2008). Este programa estatístico foi acoplado ao Sage Analysis, o que permitiu a extração, contagem e análise estatística dos dados em um único pacote.

Este teste de significância calcula os valores de p (*p-value*) que representam uma medida numérica da coerência dos dados observados em relação à hipótese nula (não diferenciação da expressão das *tags*). Para os cálculos dos valores de p, este teste utiliza uma técnica exata independente da quantidade de bibliotecas e das contagens totais das *tags* em todas as bibliotecas, mas altamente dependente das verossimilhanças (função de probabilidade condicional). Como a chance de se encontrar uma *tag* específica em uma biblioteca é um evento raro, para o modelamento estatístico utilizou-se a distribuição de Poisson.

Normalmente para a seleção de genes diferencialmente expressos, um único valor de corte canônico é utilizado para todas as *tags*, geralmente baseado no *p-value,* utilizando-se os valores 0,1; 0,05 ou 0,01. Estes valores de corte são utilizados para definir se uma *tag* é diferencialmente expressa ou não, ou seja, aceitar ou rejeitar a hipótese nula. Em qualquer perfil de expressão gênica, a abundância relativa de cada *tag* pode variar de forma significante e a utilização de um nível de corte fixo para todas as *tags* pode alterar os dados de expressão, principalmente para as *tags* de baixa abundância, já que normalmente as *tags* de alta abundância apresentam valores de p menores (Varuzza, 2008). O Kemp possui a grande vantagem de calcular um valor de nível crítico para cada *tag* em particular (α), o que minimiza a combinação linear dos erros do tipo I (falso positivo) e II (falso negativo). Como nosso objetivo era analisar as *tags* diferencialmente expressas, foi utilizado um peso maior (4) para a hipótese nula (erros do tipo I) selecionando assim *tags* realmente diferencialmente expressas, minimizando os erros do tipo II (falsos negativo).

A classificação dos genes diferencialmente expressos foi realizada de forma automática. As *tags* com valores de p menor do que α (nível crítico) foram classificadas como diferencialmente expressa.

Considerando que um dos principais objetivos deste trabalho é a identificação de genes diferencialmente expressos entre os estágios estudados, apenas as *tags* diferencialmente expressas foram utilizadas para o mapeamento nos transcritos. O

custo computacional para o mapeamento de 9.516 *tags* seria extremamente alto e o tempo de processamento extremamente longo, o que não descarta a possibilidade da utilização destes dados em novos estudos.

Para se construir uma base de dados de referência para SAGE múltiplos transcritos expressos a partir do mesmo gene necessitam ser agrupados, já que leituras presentes nas bases de dados de ESTs são altamente redundantes. Este processo não é fácil, diferentes genes podem ser agrupados como um único, devido à alta taxa de similaridade, enquanto que transcritos do mesmo gene podem ser separados em grupos distintos (Wang, 2007).

Para o mapeamento das *tags* nos transcritos de ESTs/ORESTES de *E. tenella* resolvemos utilizar uma abordagem inédita, utilizando o programa GenSeed (Sobreira e Gruber, 2008) que utiliza as *tags* como sementes para realizar uma reconstrução progressiva através de múltiplos ciclos de busca de similaridade, recuperação de seqüências e montagem. A grande vantagem deste método está na utilização das leituras dos cDNAs e não das montagens, pois neste último, dependendo dos parâmetros utilizados as *tags* podem ser cortadas de tal forma que parte delas possam estar representadas em diferentes *contigs*, principalmente em casos onde um número alto de eventos é encontrado no processo de *clusterização.* A utilização das *tags* como semente aumenta a chance de mapeamento destas e de uma posterior caracterização das seqüências reconstruídas. Das 270 *tags* classificadas como diferencialmente expressas, 199 (73,7%) foram reconstruídas utilizando esta abordagem, enquanto que se optássemos pela utilização de BLASTN contra o banco de ESTs/ORESTES *clusterizado* apenas 169 (62,6%) das seqüências seriam utilizadas para as análises posteriores (dados não mostrados).

Mesmo que nem todas as *tags* possam ser mapeadas, pois os dados de transcriptoma nem sempre estão completos para a maioria dos organismos (Malig *et al.*, 2006), basear o mapeamento das *tags* em genes preditos de bases de dados genômicas é problemático. Raramente as anotações gênicas incluem as UTRs dos transcritos expressos (Pleasance *et al.*, 2003), principalmente em genomas *drafts* não curados como o de *E. tenella*. Além disso, a arquitetura gênica deste parasita têm se mostrado extremamente complexa, possuindo um grande número de íntrons que freqüentemente apresentam *splicing* alternativo (Lal *et al.*, 2009).

A aplicação dos dados de expressão gerados em larga escala depende de sua correta interpretação, o que requer o uso de um arsenal de técnicas e

programas de bioinformática. Dentro deste contexto, para anotação das sequências reconstruídas a partir das *tags* diferencialmente expressas, empregamos um procedimento de anotação automática utilizando-se componentes recém desenvolvidos da plataforma Egene (Durham *et al.*, 2005). Esta metodologia foi utilizada com sucesso na anotação dos dados de ORESTES de três espécies de *Eimeria* de galinha doméstica (Ferro, 2008).

A anotação consiste em um processo múltiplo pelo qual seqüências brutas de DNA ou aminoácidos são analisados com a finalidade de se atribuir características biológicas e funções a estas seqüências (Stein, 2001). A identificação de genes em genomas eucarióticos, que apresentam éxons e íntrons é classicamente feita por uma combinação entre o uso de diferentes preditores de genes e de mapeamento de cDNAs. Os preditores de genes utilizam um arsenal de diferentes técnicas de predição, utilizando-se um conjunto representativo de genes (*golden standard*) para o treinamento. Esses programas baseiam-se principalmente na existência de regiões contendo fases de leitura aberta, desvios composicionais (ex. maior conteúdo G+C nos éxons), vieses (*biases*) na utilização de códons, etc. Para *Eimeria tenella*, o conteúdo G+C do genoma é muito próximo de 50% e os desvios composicionais nas regiões codificadores são praticamente indetectáveis. Este parasita apresenta uma utilização de códons bastante equilibrada, o que também torna mais difícil a distinção entre as regiões codificadoras e não codificadoras. Finalmente, o parasita apresenta uma estrutura genômica peculiar com grande quantidade de seqüências repetitivas seriadas (microssatélites) distribuídas de forma não homogênea ao longo de todo o genoma, sendo encontradas mais freqüentemente em regiões codificadoras, determinando assim um padrão genômico segmentado, conforme descrito para o cromossomo 1 de *E. tenella* (Ling *et al.*, 2007), tornando o processo de predição gênica ainda mais complicado. Não obstante, conforme destacado anteriormente, os genes destes organismos possuem uma grande quantidade de íntrons (Pinney *et al.*, 2005), o que prejudica a performance dos preditores de genes.

A anotação de genes também pode ser obtida a partir de dados de similaridade com seqüências biológicas de função conhecida, encontro de motivos protéicos, regiões transmembranares, peptídeo sinal, entre outros (Ferro, 2008). Dentre as evidências utilizadas para a anotação automática, a busca de similaridade utilizando BLAST (Altschul *et al.*, 1997) constitui uma das principais etapas de caracterização das seqüências especialmente quanto ao seu grau de conservação e

sua possível função, e por esse conjunto de características, foi a principal abordagem escolhida para análise dos transcritos no presente projeto.

A anotação automática apresentou resultados extremamente positivos, das 199 seqüências submetidas a este protocolo, 197 (99%) foram aceitas e apresentaram pelo menos uma ORF contendo no mínimo 50 resíduos de aminoácidos. Mesmo após a curagem manual, apenas 5 transcritos reconstruídos foram alterados com relação à denominação de seus produtos protéicos. Além do mais, dentro das seqüências que apresentaram resultados positivos de BLAST, a determinação da ORF pelo programa (obedecendo aos critérios escolhidos pelo usuário) foi sempre a correta. Para reconstruídos com resultados negativos de BLAST o mesmo parece não ter ocorrido, entretanto, não possuímos evidências suficientes para afirmar com certeza se a escolha da ORF foi realmente adequada.

Das 197 seqüências anotadas, 102 (51,7%) produtos protéicos não apresentaram resultados positivos em buscas de similaridade por BLASTP, apesar de não satisfatório, este resultado já era esperado, já que apenas 20% das seqüências ESTs/ORESTES de *E. tenella* apresentaram resultados positivos de BLAST. Este perfil é relativamente comum em outros organismos relacionados, das 5268 proteínas preditas no genoma de *Plasmodium falciparum* cerca de 60% não apresentou resultados de similaridade com proteínas de outros organismos (Gardner *et al.*, 2002), em *Toxoplasma gondii* metade dos genes preditos estão anotados como proteínas hipotéticas (Xia *et al.*, 2008).

Apesar de muitos destes produtos reconstruídos estarem anotados como proteínas hipotéticas, a freqüência das *tags* utilizadas como semente, em sua grande maioria era de média e alta expressão, das quais menos de 8% foram caracterizadas como de baixa expressão (contagem inferior a 5), o que demonstra que muito provavelmente estes reconstruídos correspondam a transcritos reais ainda não caracterizados. Para minimizar este problema novas abordagens poderão ser utilizadas, como a criação de um banco de dados de ESTs de organismos do filo Apicomplexa para realizar novas buscas de similaridade, o que poderia permitir a caracterização de genes conservados. Buscas em organismos Alveolata também possibilitariam a identificação de genes Filo específicos.

Dos 95 produtos protéicos que apresentaram resultados de similaridade, grande parte destes (87,4%) apresentou resultados positivos contra organismos do Filo Apicomplexa, o que é um bom indício tanto para a caracterização destes

reconstruídos, quanto para a credibilidade do protocolo de anotação utilizado. É importante ainda ressaltar que os resultados obtidos contra fungos, bactérias, dinoflagelados, correspondem em sua maioria a proteínas ribossômicas, geralmente conservadas entre organismos filogeneticamente mais distantes.

Foram identificados 130 transcritos reconstruídos a partir de *tags* diferencialmente expressa em merozoítos e 67 em esporozoítos. Analisando-se as seqüências com resultados positivos de BLAST mais expressas em merozoítos quando comparadas a esporozoítos, podemos observar que na distribuição dos produtos protéicos, 48,1% correspondem a proteínas de ribossomo, 25,3% a antígenos (em sua grande parte antígenos de superfície) e 3,8% a proteínas de micronema, o restante inclui produtos de funções diversas. Em esporozoítos apenas 16 reconstruídos apresentaram resultados positivos de BLAST, resultados semelhantes foram também observados por Ng *et al.* (2002). Além disso, em esporozoítos, a distribuição das proteínas não apresenta uma divisão tão bem estabelecida como acontece para merozoítos, das quais 5 correspondem a proteínas de *E. tenella* ainda não caracterizadas e os produtos restantes são bem variados incluindo histonas, proteínas de membrana associadas a ribossomos, facilitador do transporte de glicose, proteína de ligação à ATP, catepsina, entre outras. Apesar de ambos os estágios estarem relacionados ao processo de invasão, o perfil transcricional obtido foi relativamente diferente.

Merozoítos apresentam um grande número de proteínas de ribossomo, de fato, Schaap *et al.* (2005) ao analisar mais de 26.000 ESTs de *E. tenella* observaram que cerca de 10% dos ESTs de merozoítos codificam proteínas de ribossomo enquanto apenas 0,2% destas são codificadas nos esporozoítos. A identificação de um grande número de genes relacionados à expressão gênica e síntese de proteínas em merozoítos também já havia sido relatada (Ng *et al.*, 2002). Um estudo de proteoma de *E. tenella* revelou uma maior quantidade de proteínas envolvidas na transcrição, síntese protéica e ciclo celular em merozoítos quando comparados aos esporozoítos (Lal *et al.*, 2009). Para a espécie *E. acervulina*, uma análise de 1.847 ESTs de merozoítos de segunda geração e esquizontes revelou que 4 dos *contigs* mais abundantes codificavam proteínas de ribossomo (Miska *et al.*, 2008). Em *Plasmodium falciparum*, a expressão de cada unidade de RNAr é regulada de acordo com o estágio de desenvolvimento do parasita resultando na expressão de diferentes conjuntos de RNAs ribossômicos. Aparentemente estas

mudanças são capazes de alterar a taxa de tradução (globalmente ou para determinados RNAm), o que acarreta em mudanças na taxa de crescimento celular e nos padrões de desenvolvimento (Gardner *et al.*, 2002), podendo estar altamente correlacionada à expressão estágio específica encontrada nestes organismos.

Em taquizoítos de *Toxoplasma gondii* 12% das proteínas correspondem a proteínas de ribossomo e o segundo grupo mais abundante é de antígenos de superfície (Kappe *et al.*, 2001), resultado este semelhante aos nossos dados para merozoítos. Grandes famílias de antígenos de superfície são normalmente encontradas nos genomas de patógenos e exercem um papel fundamental na variação antigênica e evasão da imunidade do hospedeiro (Roos, 2005). Em *P. falciparum* cerca de 40% dos genes específicos de merozoítos são representados por antígenos de superfície (Llinas e DeRisi, 2004). Uma análise recente do proteoma de *E. tenella* foi capaz de detectar 47 antígenos de superfície em merozoítos contra apenas 4 em esporozoítos (Lal *et al.*, 2009).

Analisando-se ainda apenas os produtos protéicos reconstruídos a partir de *tags* diferencialmente expressas, as proteínas de micronema (EtMIC) foram identificadas somente em merozoítos, embora nos dados de ORESTES (dados não mostrados) várias proteínas de micronema também foram observadas em esporozoítos. As micronemas, organelas localizadas na porção apical das fases invasivas de todos os parasitas do Filo apicomplexa possuem um papel fundamental na invasão das células hospedeiras e contêm proteínas críticas para o processo de adesão (Ryan *et al.*, 2000). As micronemas exercem um papel importante na motilidade, adesão e reconhecimento da célula hospedeira, sendo a primeira organela a liberar seu conteúdo durante o processo de invasão (Bromley *et al.*, 2003). Também estão envolvidas no tráfico e seqüestro de ligantes para os receptores da célula hospedeira, liberando altas concentrações destas adesinas na extremidade apical do parasita quando há contato com a célula alvo (Tomley e Soldati, 2001).

Ainda em merozoítos podemos destacar uma proteína com atividade peptidil-prolil cis/trans isomerase que acelera o dobramento (*folding)* de proteínas. Provavelmente trata-se uma proteína da família das ciclofilinas, conhecidas por interagirem com as fosfatases calcineurinas mediadoras cruciais da via de sinalização de cálcio (Hosse *et al.*, 2008). Relacionada ao processamento de RNA e

transdução de sinal esta proteína também pode ser utilizada como alvo terapêutico contra organismos do Filo apicomplexa (Krücken *et al.*, 2009).

Heat shock protein 90 (Hsp90) é uma chaperona molecular, considerada como uma das proteínas mais abundantes e conservadas evolutivamente. Possuem a função de manter a conformação adequada das proteínas auxiliando no dobramento das cadeias polipeptídicas nascentes além de estarem envolvidas no re-dobramento das proteínas desnaturadas após estresse proteotóxico. Normalmente observa-se uma maior concentração destas proteínas com o aumento da temperatura sob situações de estresse celular e oxidativo. Durante a invasão, estabelecimento e desenvolvimento dos parasitas na célula hospedeira, estes organismos são expostos a estresse ambiental e a Hsp90 possui um papel fundamental na sobrevivência destes protozoários. Em taquizoítos de *Toxoplasma gondii* os níveis de Hsp90 assim como a sua transcrição é aumentada durante o estresse, além disso, este estágio de desenvolvimento também é capaz de secretar esta proteína (Echeverria *et al.*, 2005). Cerca de 2% dos genes de *P. falciparum* codificam chaperonas e durante exposição à drogas foi relatado um aumento de transcritos de Hsp90 (Natalang *et al.*, 2008). Em *E. tenella* estas proteínas são consideradas praticamente como constitutivas sendo detectadas em oocistos esporulados e não esporulados, esporozoítos e merozoítos, no entanto, níveis diferentes destes transcritos são observados em cada fase evolutiva. Quando analisados por PCR em tempo real, estas chaperonas mostram-se mais presentes nos estágios de merozoítos em relação aos esporozoítos (Miska *et al.*, 2005), corroborando com os nossos dados. A localização celular desta proteína também é alterada quando esporozoítos mantidos em PBS são transferidos para meio contendo soro fetal bovino, estas proteínas que antes estavam no citoplasma dirigem-se para a porção apical do parasita, mesmo padrão de migração é observado em esporozoítos intracelulares. Em esquizontes a distribuição é uniforme, em merozoítos a HSP tem localização próxima do núcleo e também podem ser encontradas no vacúolo parasitóforo. Ainda neste estudo foi relatado que o uso de anticorpos contra Hsp90 pode impedir a invasão dos parasitas nas células hospedeiras (Péroval *et al.*, 2006).

Proteínas associadas ao complexo da cadeia polipeptídica nascente subunidade α (NAC) também foram encontradas em merozoítos e são provavelmente as primeiras proteínas citosólicas a entrar em contato com a cadeia

polipeptídica nascente, prevenindo assim interações inapropriadas com outros fatores (Beatrix *et al.*, 2000).

A superóxido dismutase (SOD) também foi observada em merozoítos está envolvida no controle do estresse oxidativo, causado pela produção de compostos tóxicos e espécies reativas de oxigênio (ROS). A SOD é responsável pela reação de dismutação de O_2^- para H_2O_2 em *Plasmodium falciparum* (Bozdech e Ginsburg, 2004) e também está implicada em neutralizar e eliminar os ROS em *E. tenella* (Georgieva *et al.*, 2006).

A miosina faz parte do sistema actina miosina, necessário para o movimento de *gliding* e penetração dos “zoítos” na célula hospedeira. Em *Toxoplasma gondii* a miosina A é ancorada na membrana plasmática e se liga a miosina de cadeia leve (TgMCL1), similar à encontrada em nossos dados de merozoítos e, este complexo encontra-se distribuído de forma homogênea no citoplasma das células do parasita. Aparentemente o citoesqueleto de actina destes organismos do Filo apicomplexa é muito bem utilizado, usando uma combinação de movimento aumentado pela miosina juntamente com a força gerada pela polimerização da actina (Herm-Götz *et al.*, 2002).

As fosfatases ácidas estão envolvidas na transdução de sinal (Klotz *et al.*, 2005). A fosforilação de proteínas tem um papel fundamental na regulação de funções celulares, causando a inibição ou ativação de muitas enzimas envolvidas em diversas vias bioquímicas.

Uma ATP sintase sub-unidade G1 (V-ATPase) vacuolar também foi observada em merozítos e está relacionada a acidificação do ambiente intracelular. São encontradas como proteínas de ligação às membranas em várias organelas, assim como na membrana plasmática de células especializadas podendo estar envolvida no acoplamento da degradação de ATP com a translocação de H^+. Esta V-ATPase em *P. falciparum* está implicada na regulação do armazenamento do cálcio intracelular nos compartimentos ácidos e a sua alteração modifica negativamente a homeostase intracelular de cálcio, podendo levar à morte celular, necrose e apoptose (Natalang *et al.*, 2008).

As tiorredoxinas são proteínas doadoras de hidrogênio que participam das reações de redução, sendo encontradas em praticamente todos os organismos vivos. Em geral agem como uma dissulfeto oxirredutase. O sistema de tiorredoxina foi caracterizado para *P. falciparum,* a ação desta proteína é capaz de fornecer

equivalentes reduzidos para uma ampla variedade de aceptores, além de ser muito eficiente na redução não específica das pontes de dissulfeto das proteínas (Bozdech e Ginsburg, 2004). Algumas moléculas que requerem a ativação via tiorredoxina ou são associadas a esta estão localizadas no apicoplasto de *T. gondii* (DeRocher *et al.*, 2008).

A calmodulina também é uma proteína que se liga a moléculas de cálcio, em *T. gondii*, foi encontrada no complexo apical concentrada principalmente no conóide. Proteínas do complexo apical que se ligam a moléculas de cálcio provavelmente estão envolvidas na regulação motora, já que o movimento do conóide e de suas fibras é aumentado com a elevação da concentração de cálcio intracelular (Hu *et al.*, 2006), apresentando desta forma, um papel importante no processamento da sinalização de cálcio. O complexo calmodulina-cálcio é capaz de se ligar a outros componentes e regular diversas funções celulares, incluindo a secreção, organização do citoesqueleto, ativação enzimática (Pezzella-D'Alessandro *et al.*, 2001) e evasão celular (Santos *et al.*, 2009).

A guanilil ciclase, presente também nos dados de merozoítos, catalisa a formação do GMP cíclico (cGMP) a partir de GTP. Esta proteína age como um mensageiro intracelular ativando a cGMP quinases dependentes e regulando os canais de íon sensíveis a cGMP. A inibição de proteínas quinases dependentes de cGMP (PKG) em *Toxoplasma* e *Eimeria* inibe a secreção de adesinas, a motilidade (*glinding*), adesão celular e invasão. O mecanismo pelo qual estas PKGs regulam a secreção das proteínas de micronema ainda é desconhecido e deve estar envolvido na comunicação com a via dependente de cálcio (Wiersma *et al.*, 2004). Além disso, em *Plasmodium berguei* o controle da hidrólise de cGMP é essencial para prevenir uma super estimulação de PKGs que são proteínas reguladas de forma estágio específica (Moon *et al.*, 2009).

A proteína 14-3-3 de *Eimeria* também foi identificada. Sabe-se que esta proteína estimula interações proteína-proteína, funcionando tanto como âncoras ou chaperonas, além de estarem envolvidas na regulação da localização sub-celular de proteínas e inibição de algumas enzimas. Esta proteína foi estudada em *T. gondii* e localiza-se principalmente no citosol dos taquizoítos. Uma pequena porção destas está associada à membranas evidenciada pela sua co-localização com SAG1 (*surface antigen* 1), entretanto, sua função nos protozoários ainda não foi totalmente estabelecida (Assossou *et al.*, 2003).

Analisando em conjunto estes produtos protéicos diferencialmente expressos em merozoítos, podemos observar que estão relacionados principalmente à manutenção e dobramento das proteínas, como a ciclofilina hipotética, Hsp90, NAC e proteínas 14-3-3; resposta a estresse, muito comum em “zoítos” como SOD, tiorredoxina e Hsp90; transdução de sinal, incluindo novamente a ciclofilina, a fosfatase ácida e as proteínas associadas ao cálcio intracelular intimamente relacionadas ao processo de invasão, como a miosina, a ATP sintase vacuolar, a calmodulina e, indiretamente a guanilil ciclase e a ciclofilina.

Em esporozoítos podemos destacar a catepsina C e uma proteína transportadora de *cassete* de ligação de ATP subfamília A. As catepsinas pertencem à família das cisteíno-proteases normalmente ativadas sob o pH ácido dos lisossomos. Estas proteínas exercem atividades fundamentais para a invasão, biogênese organelar e sobrevivência intracelular dos parasitas do Filo apicomplexa. As catepsinas C são estruturalmente e cataliticamente únicas dentro da sub-família das papaínas, pois possuem atividade de exopeptidase. Estas proteínas são capazes de remover de forma seqüencial os dipeptídeos da porção N-terminal das cadeias polipeptídicas. Estudos demonstram que a catepsina C pode ter importância no crescimento celular, diferenciação e ativação das proteases, além disso, pode ser secretada para o compartimento extracellular podendo estar relacionada à degradação da matriz extracelular (Que *et al.*, 2007).

Em *T. gondii*, três catepsinas C foram encontradas, a C1 é a mais expressa em taquizoítos, sendo observada nos grânulos densos e não é encontrada em bradizoítos. A liberação de proteínas dos grânulos densos é um dos processos mais importantes destes parasitas. Estas proteínas são transmembranares em sua grande parte, são secretadas para o espaço vacuolar na forma de proteínas solúveis e são responsáveis por remodelar o vacúolo parasitóforo em um compartimento ativo. Podem estar implicadas nas trocas entre o parasita e a célula hospedeira como a aquisição de nutrientes, transdução de sinal ou manutenção do vacúolo parasitóforo (Que *et al.*, 2007). Catepsina C também foi detectada no vacúolo parasitóforo de *Plasmodium falciparum* onde participa do processo de quebra da hemoglobina (Klemba *et al.*, 2004). Aparentemente os dipeptídeos gerados pela ação da catepsina C podem ser transportados para o citosol do parasita de forma ativa onde são hidrolisados em aminoácidos por aminopeptidases neutras citosólicas (Klemba *et al.*, 2004).

A proteína transportadora de *cassete* de ligação de ATP subfamília A, pertence à super-família do cassete de ligação a ATP (ABC) e é uma das proteínas mais conservadas evolutivamente. Os transportadores ABC utilizam a hidrólise de ATP para fornecer energia à diversos sistemas biológicos relacionados principalmente manutenção da homeostase celular. Possuem também um papel fundamental na desintoxicação celular de endobióticos e xenobióticos, sendo que a super expressão de algumas proteínas desta família contribui para a resistência às drogas em vários organismos, desde células neoplásicas humanas, bactérias, fungos, nematóides até os protozoários parasitas, como *Plasmodium berguei* (González-Pons *et al.*, 2009). Em *Cryptosporidium parvum* estas proteínas estão localizadas na fronteira entre o parasita e a célula hospedeira e são capazes de transportar grandes ânions orgânicos. Aparentemente possuem maior importância nos estágios intracelulares, entretanto, também são encontradas em oocistos e esporozoítos. Sua distribuição em esporozoítos parece estar relacionada aos grânulos densos que sofrem exocitose durante a invasão das células hospedeiras (Perkins *et al.*, 1999).

Ambos os produtos protéicos descritos acima são importantes para o desenvolvimento de esporozoítos, a catepsina C serve como fonte de aminoácidos e a proteína transportadora de *cassete* de ligação de ATP intimamente relacionada com transporte celular e resistência às drogas.

Ainda em relação aos dados de anotação com resultados positivos de BLAST, pudemos identificar em ambos os estágios *tags* distintas reconstruídas que apresentaram o mesmo número de acesso no *Genbank* e, que desta maneira, podem representar partes do mesmo gene, isoformas, ou mesmo transcritos diferenciais resultantes de *splicing* alternativo na extremidade 3’ UTR, poliadenilação heterogênea, RNAs antisense, SNPs, entre outros. Estudos de SAGE em *Arabidopsis thaliana* revelaram que 15% dos transcritos diferencialmente expressos correspondiam a transcritos alternativos (Robinson e Parkin, 2008).

Dentre estes transcritos alternativos foi observado casos sugestivos de poliadenilação heterogênea (Stern *et al.*, 2003) (apesar de não termos encontrado os sítios de poliadenilação) e *splicing* alternativo da extremidade 3’ UTR (Keime *et al.*, 2007; Kuo *et al.*, 2006; Unneberg *et al.*, 2003). Em ambos os exemplos, sempre as *tags* mais próximas à cauda poliA apresentavam uma contagem absoluta de no mínimo 10 vezes maior do que a segunda *tag* encontrada no transcrito. Segundo

Malig *et al.* (2006) 20% das *tags* experimentais mapeadas em um transcrito estão localizadas no segundo sítio CATG mais próximo à cauda poliA.

Poliadenilação heterogênea já foi relatada em estudos de SAGE. Em *Schistossoma mansoni* 13 genes apresentaram poliadenilação alternativa acarretando em uma diminuição da porção 3' terminal dos RNAs mensageiros, o que poderia afetar a estabilidade e regulação destes transcritos (Ojopi *et al.*, 2007). Em *T. gondii* a presença de poliadenilação heterogênea em um gene diferencialmente expresso referente à proteína de grânulo denso 7 (GRA7), foi comprovada experimentalmente através de RACE (*Rapid Amplification of cDNA Ends)* (Radke *et al.*, 2005). Poliadenilação heterogênea também foi relatada em transcritos de SAGE proveniente de bibliotecas de trigo (Poole *et al.*, 2008)

O *splicing* alternativo é uma das maiores fontes de diversidade transcricional em células humanas (Kirschbaum-Slager *et al.*, 2005). Formas alternativas de *splicing* foram encontradas em mais de 59% dos transcritos de humanos e 41% em ratos (Wahl *et al.*, 2005a). Em organismos do Filo apicomplexa variantes de splicing tem sido mais exploradas e caracterizadas em espécies do gênero *Plasmodium* spp., sendo encontrados nos genes *maebl, stevor, 41-3* e *b7* e possivelmente nos genes *yir* (*Plasmodium* interspersed repeats- *pir* de *P. yoelli*) (Fonager *et al.*, 2007). Em *T. gondii* a miosina B e C são variantes de *splicing* que diferem apenas na extremidade 3' (carboxi-terminal). Devido esta diferença, estas isoformas apresentam solubilidades distintas e diferentes localizações subcelulares. Estes 2 transcritos estão presentes em taquizoítos, no entanto, em bradizoítos são super-expressos. Aparentemente a produção de miosina B é menos freqüente quando comparada a miosina C, já que a super expressão de miosina B leva a um defeito na divisão celular, redução da virulência e atraso na replicação (Delbac *et al.*, 2001). Outro trabalho também em *T. gondii* relatou a presença de duas variantes de *splicing* do gene que codifica uma hipoxantina-xantina-guanina fosforibosiltransferase (HXGPRT), estas variantes diferem apenas na presença ou ausência de 49 aminoácidos, mas suas características quanto à atividade, substrato e padrão de expressão são similares, diferindo apenas quanto a localização celular (Chaudhary *et al.*, 2005). Em *Theileria lestoquardi* foi identificado um gene denominado de clone 5 que possui duas isoformas do mesmo gene codificador. A única diferença entre as variantes está na retenção de um íntron de 34 pb o qual pode gerar isoformas com domínios transmembranares (Bakheit *et al.*, 2006).

Quando *tags* distintas provenientes de SAGE recebem uma mesma anotação pode ser um indício da presença de variantes de *splicing* na extremidade 3' UTR deste gene. A porção 3'UTR tem um importante papel na regulação da tradução, estabilidade dos RNAm e sublocalização celular (Kuersten e Goodwin, 2003; Wahl *et al.*, 2004; Wahl *et al.*, 2005a). Variantes de *splicing* foram detectadas em bibliotecas de SAGE de humanos (Keime *et al.*, 2007); (Kim *et al.*, 2006), de ratos (Wahl *et al.*, 2005a) em raiz de milho (Poroyko *et al.*, 2005), cana de açúcar (Calsa e Figueira, 2007) e em tecido de folhas de *Arabidopisis thaliana* mantidas sob condições de baixa temperatura (Robinson e Parkin, 2008). Estes estudos mostram a relevância dos mecanismos transcricionais para o controle da expressão gênica. Além disso, caracterização destes variantes pode auxiliar na elucidação dos mecanismos moleculares utilizados por estes organismos.

Ainda analisando os transcritos diferenciais encontrados em nossos dados, *tags* potencialmente *antisense* também foram encontradas. Dois casos distintos foram observados: a presença de *sense* e *antisense* no mesmo sítio de pontuação (CATG) e a identificação de *tags* apenas no sentido *antisense*. Foram encontrados 9 pares de *tags sense-antisense* e 18 *tags* potencialmente *antisense* (baseada na comparação das seqüências reconstruídas contra bases de dados de ESTs direcionais de *E. tenella*), totalizando 13,7% das *tags* diferencialmente expressas. Ensaios de RT-PCR evidenciaram experimentalmente que 3 *tags* apresentavam direcionamento *antisense.*

Uma característica elegante da técnica de SAGE é que a posição e orientação de uma determinada *tag* em um transcrito pode ser determinada de modo acurado, permitindo a detecção de transcritos *antisense* (Gibbings *et al.*, 2003). O controle de expressão gênica por transcritos *antisense* naturais (NAT) foi descoberta em procariotos há mais de 20 anos. Como o RNAm possui apenas uma única fita, a presença de uma fita *antisense* complementar pode alterar a transcrição, elongação, processamento, estabilidade e tradução destes RNAs. Os transcritos *antisense* podem ser codificadores ou não, complementares a regiões codificadoras de RNAm processadas e maduras ou podem ser complementares apenas ao transcrito primário não processado (Kiyosawa *et al.*, 2003). Existem dois tipos de RNA antisense endógeno ou NAT, os RNAs *antisense* em *cis* e os em *trans*. Os RNAs *antisense* em *cis* são transcritos a partir de uma fita oposta do RNA *sense* na mesma localização genômica e, são caracterizados por possuírem uma sobreposição

geralmente longa e perfeita com o RNA *sense*. Os RNAs *antisense* em *trans* são transcritos a partir de outras regiões genômicas e a sobreposição normalmente é pequena e imperfeita (Chen *et al.*, 2004).

Kiyosawa *et al.*, 2003 estudando genes de ratos encontraram 2.481 pares de transcritos *sense-antisense* demonstrando que talvez a regulação da expressão por transcritos *antisense* seja muito mais comum do que o previsto. Chen *et al.* (2004) relataram que 22% dos genes humanos formam pares *sense-antisense*. Aparentemente estes pares são conservados entre os vertebrados, provavelmente já estavam presentes nos ancestrais e podem ter influenciado a evolução destes organismos. A ocorrência destes pares parece ocorrer com maior freqüência na região UTR dos transcritos (Dahary *et al.*, 2005). Geralmente a expressão dos transcritos *sense* é maior do que os seus correspondentes *antisense.* (Wahl *et al.*, 2005b), como observado no estudo de mais de 18.000 transcritos de ratos, dos quais 1.468 transcritos eram *antisense* e destes, 78% dos transcritos *antisense* apresentavam um par *sense.* Uma análise focada em *tags* de humanos que não mapearam em transcritos conhecidos, identificou que na média 61% dos transcritos de cada biblioteca apresentam-se na posição *sense*, 10% apenas na posição *antisense* e 29% eram pares *sense-antisense* (Keime *et al.*, 2007). A obtenção de *tags antisense* também foi relatada em outros organismos como *Schistossoma mansoni* (Taft *et al.*, 2009), *Brassica napus* (Obermeier *et al.*, 2009) e *Oriza sativa* (Gibbings *et al.*, 2003). A partir dos dados de SAGE de *Giardia* spp foi observado que os transcritos *antisense* são gerados a partir de promotores que realizam uma transcrição bidirecional. Estes transcritos *antisense* são detectados em níveis baixos e estão distribuídos de forma homogênea por todo o genoma (Teodorovic *et al.*, 2007). Em *P. falciparum,* a síntese de transcritos *antisense* parece ser um fenômeno transcricional comum catalisado pela RNA Polimerase II que aparentemente não possui a capacidade de distinguir a fita *sense* da *antisense*. Estes transcritos *antisense* normalmente não apresentam ORFs e podem representar novos reguladores da expressão estágio específica competindo com os fatores de transcrição *sense* ou modelando a estrutura da cromatina e podem também formar uma estrutura duplex para regulação da transcrição e estabilidade do RNAm (Militello *et al.*, 2005). Para o gene MSP3 (merozoite surface protein 3) de *P. falciparum* o transcrito *sense* aparece cerca de 60 vezes mais do que o seu par antisense. Em estudos de SAGE de *P. falciparum*, 12% das *tags* foram

caracterizadas como somente *antisense*, das quais 70% eram de baixa abundância e 5% constituíam pares *sense-antisense* (Gunasekera *et al.*, 2004). Patankar *et al.* (2001) relatou que 17% das *tags* de alta expressão estavam na posição *antisense*. Em *T. gondii*, 21,5% das *tags* mapeadas em um único sítio genômico também correspondiam a transcritos *antisense* (Radke *et al.*, 2005). Nos nossos dados também foi observado um maior número de *tags sense* do que *antisense* nos pares *sense-antisense*.

Transcrito alternativo, possivelmente resultante de SNP também foi encontrado em nossas análises. Estima-se que *tags* alternativas derivadas de SNP ocorram em 8,6% dos genes humanos. Dentro deste conjunto putativo, 61,9% dos genes apresentaram *tags* obtidas experimentalmente contendo estes SNPs (Silva *et al.*, 2004). Em *T. gondii* analisando um conjunto de cerca de 200.000 *tags* mais de 400 *tags* apresentaram SNPs (Radke *et al.*, 2005).

Mesmo analisando um pequeno conjunto de *tags*, 197 transcritos reconstruídos a partir de *tags* diferencialmente expressas, observamos alguns transcritos alternativos demonstrando que a técnica de LongSAGE foi capaz de detectar estes transcritos. Estudos futuros utilizando o conjunto total de *tags* poderão ser realizados com o objetivo de se caracterizar estes transcritos alternativos mais profundamente.

Como *tags* mitocondriais também já haviam sido encontradas em SAGE proveniente de *Saccharomyces cerevisiae* (Velculescu *et al.*, 1997), *C. elegans* (principalmente *antisense*) (Jones *et al.*, 2001), *Oryza sativa* (Song *et al.*, 2007) e *P. falciparum* (Patankar *et al.*, 2001), decidimos mapear as 9.516 *tags* únicas no genoma anotado de mitocôndria de *E. tenella* (Romano, 2004).

Plasmodium, *Toxoplasma* e outros organismos do Filo apicomplexa, assim como maior parte dos tripanossomatídeos apresenta uma mitocôndria com capacidade de realizar fosforilação oxidativa (Ginger, 2006). Em apicomplexa, durante o processo evolutivo, a maior parte dos genes mitocondriais foram transferidos para o núcleo (Gardner *et al.*, 2002; Ossorio *et al.*, 1991) permaneceram na mitocôndria apenas os genes cujos produtos são altamente tóxicos para serem liberados no citoplasma. O mesmo parece ter ocorrido para *Eimeria* spp., pois somente genes de citocromo são encontrados na mitocôndria destes organismos (Romano, 2004). Analisando os nossos dados observamos que apenas 13 *tags* foram encontradas no genoma mitocondrial, das quais 8 apresentam-se na direção

antisense. Destas 13 *tags,* somente 4 apresentaram resultados positivos de anotação. É importante ressaltar que não temos como afirmar se estas *tags* realmente são verdadeiras e não artefatos. Como as *tags* localizadas na mitocôndria também foram encontradas no genoma nuclear e, considerando que os genes de citocromo *b* e *c* de *E. tenella* não possuem cópias nucleares (Romano, 2004), é um indicativo de que aparentemente as seqüências disponíveis do genoma nuclear não foram filtradas contra esta organela. Desta forma, as *tags* foram reconstruídas a partir das seqüências do genoma nuclear e então submetidas ao processo de anotação.

Apesar de não termos como afirmar que as *tags* encontradas no genoma mitocondrial de *E. tenella* sejam realmente verdadeiras, indícios da presença de cauda poliA nos RNAs ribossômicos, assim como a presença de transcritos *antisense* provenientes desta organela em outros organismos nos leva a acreditar que estas possam ser de fato reais. Em *P. falciparum* 5,3% das *tags* mais abundantes são provenientes de mitocôndria (Patankar *et al.*, 2001).

Em humanos, RNAs *antisense* de citocromo oxidase C também foram observados e podem induzir mudanças morfológicas e morte celular em linhagens celulares hematopoiéticas, provavelmente pela indução de respiração mitocondrial insuficiente (Shirafuji *et al.*, 1997). A detecção de transcritos *antisense* relacionados aos RNAr foi encontrada em duas regiões reguladoras de mitocôndrias de humanas apresentando-se de forma estável e providos de cauda poliA (Tullo *et al.*, 1994). Pequenos fragmentos de RNAr em ambas as fitas também são encontrados no genoma mitocondrial de *P. falciparum* e muitos destes apresentam poliadenilação, provavelmente para conferir proteção contra eventuais enzimas com atividade exonucleotídica 3’ (Gillespie *et al.*, 1999). Esta mesma característica também foi descrita para *Theileria parva* (Nene *et al.*, 1998).

Além do genoma mitocondrial, os organismos do Filo apicomplexa possuem uma organela denominada de apicoplasto provida de genoma próprio. A função exata desta organela ainda não foi elucidada, mas está intimamente associada a síntese de lipídeos, heme e isopropenóides e vários destes processos metabólicos podem também constituir alvos terapêuticos interessantes (Waller e McFadden, 2005). Nenhuma *tag* foi encontrada no genoma de apicoplasto de *E. tenella*, caracterizado por Cai *et al.* (2003), apesar dos transcritos derivados deste genoma possuírem cauda poliA. No entanto, a poliadenilação destes transcritos ao invés de

conferir estabilidade, como acontece nos transcritos citoplasmáticos, deve marcar estes RNAs para serem degradados, assim como acontece nos cloroplastos das plantas (Hayes *et al.*, 1999). Talvez este fato possa explicar a ausência de *tags* derivadas do genoma desta organela.

Nossos dados também foram submetidos ao mapeamento e quantificação das seqüências por termos de ontologia gênica (GO) e busca de ortólogos na base de dados KOG. Por serem curadas, estas bases são mais confiáveis, entretanto, o tamanho destes bancos de dados é muito menor. Três ontologias de termos são utilizadas na classificação de GO: função molecular, processo biológico e componente celular. Das 197 sequências analisadas apenas 72 (36,5%) apresentaram resultados positivos de Interpro, sendo 55 seqüências provenientes de reconstruídos mais expressos em merozoítos e 17 em esporozoítos.

Da mesma forma que observado anteriormente em relação aos resultados de anotação, um maior número de seqüências provenientes de merozoítos tiveram termos GO atribuídos levando-se em consideração apenas os resultados positivos de Interpro. Para comparação entre os dados obtidos de seqüências diferencialmente expressas em merozoítos e esporozoítos apenas os resultados normalizados foram utilizados, já que um número diferente de termos foi atribuído para cada conjunto de seqüências.

Analisando-se as três ontologias principais, as seqüências de merozoítos tiveram mais termos de GO relacionados à ontologia de componente celular, apresentando uma diferença superior a 20% (dados normalizados), enquanto que as seqüências de esporozoítos tiveram mais termos atribuídos dentro da ontologia função molecular, com uma diferença também superior a 20%. Analisando separadamente cada uma das três ontologias gênicas principais, observamos que para processo biológico, grande parte dos produtos protéicos de merozoítos (18%) faz parte do processo biológico biossintético e dentro deste conjunto a maior parte está envolvida na tradução e metabolismo de proteínas. As seqüências de esporozoítos foram encontradas em maior proporção em relação à organização dos componentes celulares e biogênese e transporte. Analisando a ontologia gênica de componente celular, observamos uma maior porcentagem de seqüências relacionadas à célula, núcleo e complexo protéico em esporozoítos e em merozoítos relacionadas à organela ribossomo. Para função molecular ambas tiveram representações semelhantes em *binding*, porém, o termo de ligação a nucleotídeos

foi mais encontrado em esporozoítos, assim como atividade catalítica, incluindo atividade de quinase e também atividade de transporte. Somente foram encontrados termos de ligação às proteínas nas seqüências diferencialmente expressas de merozoítos, assim como seqüências pertencentes à categoria de estrutura molecular.

Os termos de GO apresentam uma boa correlação com os dados de anotação, em merozoítos encontramos diversos termos relacionados à tradução e processo metabólico de proteínas, associados à organela ribossomo e atividade estrutural de moléculas termos estes relacionados provavelmente ao grande conjunto de proteínas de ribossomo encontradas neste estágio. Já as proteínas de micronema e os antígenos de superfície possuem intima relação ao processo de *binding*. Em esporozoítos, as histonas estão relacionadas aos termos encontrados para núcleo, a atividade catalítica deve referir-se às proteínas de cassete de transportador da família ABC, catepsina C, fosfatase ácida, desidrogenases, além das proteínas cujos domínios são relacionados à atividade de quinases e foram caracterizadas como genes desconhecidos em *Eimeria.* Com relação à atividade de transporte podemos identificar um facilitador do transporte de glicose neste estágio.

Nas buscas de ortólogos na base de dados de KOG apenas 56 produtos protéicos foram classificados e novamente a maior parte (50) pertenciam ao grupo de seqüências diferencialmente expressas em merozoítos. Apesar de poucas seqüências serem encontradas nas categorias funcionais de KOG atribuídos, este resultado não é inesperado, já que apenas 7 organismos eucarióticos foram utilizados para montar o banco de KOG, dentre eles o mais próximo filogeneticamente de *Eimeria tenella* é o fungo do microsporídia *Encephalitozoon cuniculi*. Além disso, apesar de ser considerado como um banco de ortólogos padrão, seus dados não são atualizados desde 2003. Em teoria, para identificação de ortólogos, análises filogenéticas de famílias inteiras de proteínas homólogas devem ser analisadas, porém estas análises são laboriosas e lentas, principalmente em alta escala (Dessimoz *et al.*, 2006) e portanto, a utilização de um banco de dados de ortólogos nos pareceu mais conveniente.

Das 20 categorias funcionais de KOG, nossos produtos foram encontrados dentro de 9, sendo a **J**, a mais representativa que novamente está relacionada à tradução, estrutura do ribossomo e biogênese para ambos os estágios. Analisando-se separadamente a proporção de seqüências de acordo com o estágio, pudemos

observar que das 9 categorias funcionais, apenas a categoria **R**, relacionada a produtos de função geral estava presente nos dois estágios (merozoítos e esporozoítos). Esporozoítos apresentaram seqüências do grupo relacionado à estrutura da cromatina (histonas), transporte de lipídeos e metabolismo (catepsina C) e transporte de aminoácidos (desidrogenases). Em merozoítos o maior grupo estava relacionado à categoria **J**, as demais proteínas foram distribuídas em 5 categorias relacionadas à transcrição, modificação pós-traducional, transdução de sinal e citoesqueleto, com um número bem menor de seqüências em cada. Assim como observado nos dados de anotação, *tags* distintas, que podem representar o mesmo transcrito, foram também observadas dentro das funções específicas de KOG, como as seqüências de histona e as proteínas de ribossomo RPL24, RPL38 e RPL2/L8.

Em suma, para merozoítos, podemos observar que os dados de anotação baseados em BLAST, dados de GO e dados de KOG estão relacionados aos produtos protéicos envolvidos na tradução, modificação e manutenção da conformação das proteínas. Enquanto que em esporozoítos os resultados positivos são menos freqüentes e estão relacionadas à estrutura da cromatina, transporte e atividade catalítica. Assim, esses resultados sugerem que, diferentemente de merozoítos, a maioria dos genes expressos preferencialmente em esporozoítos codifica proteínas de função desconhecida. Apesar dos dois estágios serem invasivos, podemos notar que a classificação de seus produtos protéicos é distinta, além disso, a quantidade de seqüências classificadas (resultados de BLAST positivos, termos GO e categorias de KOG) provenientes de esporozoítos também é muito menor do que as de merozoítos. A fração analisada de seqüências pode nos dar uma idéia das diferenças entre ambos os estágios, mas não exclui a possibilidade de que uma melhor caracterização das proteínas de esporozoítos possa alterar este perfil obtido.

A técnica de LongSAGE apresenta diversas vantagens em relação a outras metodologias, no entanto, por se tratar de uma plataforma de larga escala nem sempre o perfil quantitativo gerado é tão acurado. Com a disponibilidade de inúmeras tecnologias para medição da expressão gênica, é necessário realizar estudos em plataformas distintas para aumentar a significância e confiabilidade dos resultados. Diversos trabalhos que empregaram SAGE utilizaram o RT-qPCR para validação experimental estudando um pequeno conjuntos de genes (Funaguma *et*

al., 2007; Long *et al.*, 2003; Poroyko *et al.*, 2005). O PCR em tempo real (Heid *et al.*, 1996) é considerado como o método padrão para mensurar a expressão gênica diferencial devido a sua alta especificidade e sensibilidade (Deepak *et al.*, 2007).

Assim, decidimos empregar o real time PCR como forma de validar nossos dados quantitativos. Devido ao baixo custo e relativa facilidade de se delinear os experimentos, optamos pela utilização do SYBR® Green, método utilizado por metade dos pesquisadores que trabalham com esta técnica. Cientes de que este corante não é específico e pode gerar sinais falso-positivos ao detectar dímeros de *primers* ou produtos inespecíficos (Ponchel, 2006), empregamos *primers* cuidadosamente desenhados e analisamos as curvas de *melting.* Além disso, o kit utilizado possui uma DNA polimerase *hot-start*, diminuindo assim a chance de amplificação de produtos indesejáveis. Interessantemente, um trabalho recente mostrou que a correlação entre os dados de PCR em tempo real empregando-se SYBR® Green ou TaqMan (normalmente conhecido pela sua alta especificidade (Valasek e Repa, 2005)) foi altíssima, apresentando um coeficiente de correlação igual a 0,97 (Arikawa *et al.*, 2008).

A quantificação absoluta tem por objetivo mensurar o número de cópias do ácido nucléico em uma determinada amostra, é bastante laboriosa, pois em cada experimento, a amostra a ser estudada deve ser comparada à uma curva padrão gerada por diluições seriadas a partir de uma amostra de concentração conhecida (Valasek e Repa, 2005), aumentando o custo e o tempo de execução do experimento. Além disso, a eficiência da reação de amplificação entre o gene de interesse e o material utilizado na curva padrão devem ser semelhantes (Dussault e Pouliot, 2006; Pfaffl, 2006). Como nosso interesse era comparar os dados quantitativos obtidos por LongSAGE através de outra técnica, optamos pela quantificação relativa, que em teoria é mais simples e menos suscetível a erros de metodológicos (Pfaffl, 2006). Uma boa qualidade do RNAm também é essencial para execução de qualquer ensaio independente da escolha do método de quantificação (Fleige e Pfaffl, 2006), para tanto utilizamos amostras de RNAm de alta qualidade, tratadas com DNAse e quantificadas de forma acurada com um kit específico que somente reporta a quantidade de moléculas de ácido nucleico de interesse, no caso RNA. A utilização de RNAm apesar de mais cara é recomendada, pois já foi relatado que a proporção entre a quantidade de RNAm e RNAr em amostras de RNA total pode variar em diferentes condições e/ou tecidos, podendo alterar a mensuração da

expressão gênica (Morse *et al.*, 2005). Utilizamos também *primers* específicos para a síntese de cDNA, o que encarece o processo, mas é a estratégia mais adequada quando um pequeno conjunto de genes é analisado (Kubista *et al.*, 2006), pois garante que a síntese de cDNA ocorre de fato para cada gene a ser estudado (Shipley, 2006).

O primeiro desafio foi a busca e a validação de genes constitutivamente expressos (Dheda *et al.*, 2004) em *Eimeria tenella*. Sabe-se que os organismos do Filo apicomplexa possuem uma expressão gênica altamente estágio específica, ao contrário dos organismos da Ordem kinetoplastida, como a *Leishmania*, da qual apenas 0,2 a 0,5% dos genes apresentam expressão diferencial entre as formas amastigotas e promastigotas (Cohen-Freue *et al.*, 2007). Para organismos do Filo apicomplexa, por exemplo, *Plasmodium falciparum*, o perfil transcricional assemelha-se a uma cascata de expressão gênica que resulta no processo "*just in time*" onde os produtos são fabricados de acordo com a necessidade do parasita (Gissot *et al.*, 2005), o mesmo fenômeno acontece em *Toxoplasma gondii* (Mullapudi *et al.*, 2009). Assim, estima-se que nestes parasitas, a quantidade de genes *housekeeping* expressos de maneira constitutiva seja baixa (Llinas e DeRisi, 2004).

Para seleção dos candidatos, baseamos nossa busca nos resultados de LongSAGE utilizando o critério de *p-value* igual a 1 e resultados positivos de BLAST. Cinco candidatos foram selecionados: ubiquitina, histona H2A, proteína de ribossomo 22, uma seqüência possuindo o domínio z*inc finger* e uma proteína de ligação a GTP nuclear ran/tc4. Além dos critérios acima mencionados, estes genes foram comparados aos ortólogos quanto aos dados de expressão nas bases de dados de *Plasmodium falciparum* e *Toxoplasma gondii* e como foi observado um perfil transcricional constitutivo, esperávamos que este mesmo perfil pudesse se repetir em *Eimeria tenella*.

A ubiquitina é uma proteína de 76 aminoácidos extremamente conservada entre as seqüências de organismos eucarióticos, apresentando taxas de similaridade superiores a 98% entre humanos, leveduras e parasitas do Filo apicomplexa. Em organismos do Filo apicomplexa o gene que codifica para esta proteína possui uma única cópia. O complexo ubiquitina/proteassoma pertencente a uma via regulatória pós-transcricional essencial, que regula uma ampla gama de atividades celulares incluindo reparo de DNA, transcrição, divisão celular, endocitose, tráfico intracelular entre outros (Ponts *et al.*, 2008). No entanto ao estudar a expressão deste gene em

oocistos esporulados, não esporulados, parcialmente esporulados, merozoítos de segunda geração e esporozoítos, a variação de C_t encontrada foi muito grande, mais de 2 logs, não permitindo a qualificação deste gene como controle. De fato, as proteínas envolvidas na via da ubiquitina provenientes de organismos do Filo apicomplexa possuem uma maior diversidade (Ponts *et al.*, 2008), o que poderia explicar este comportamento variável em nossos experimentos.

A proteína de ligação a GTP nuclear Ran/TC4 é um membro conservado da superfamília Ras de pequenas GTPases que regulam todo o transporte mediado por receptores entre o núcleo e o citoplasma. A Ran pertence ao grupo de pequenas proteínas de ligação a GTP, que possuem um papel importante tanto nas vias de importação quanto de exportação e, geralmente apresentam altas concentrações nucleares. Esta proteína possui um alto nível de similaridade e função conservada entre organismos eucarióticos incluindo os protozoários (Frankel e Knoll, 2009). Apresenta-se de forma abundante em *Toxoplasma gondii* onde da mesma forma é responsável pela maior parte do transporte entre o núcleo e o citoplasma (Frankel e Knoll, 2008) e também já foi descrita em *P. falciparum* (Ramachandran *et al.*, 2002) e *Babesia divergens* (Delbecq *et al.*, 2003). Apesar de ser de grande importância para o metabolismo destes parasitas, uma grande oscilação dos valores de Ct coletados foi observada.

A proteína de ribossomo 22 apresentou uma variação menos discrepante, mas ainda muito próxima de 2 log.

Proteínas que apresentam o motivo *zinc finger* em sua seqüência pode ligar-se a DNA, RNA, proteínas e substratos lipídicos, servindo como âncoras estáveis. Estão associadas à transcrição, tradução, tráfico de RNAm, organização do citoesqueleto e dobramento de proteínas. Esta proteína apresentou uma variação de C_t menor que 2 log, mas ainda acima de 1,5 log não sendo ideal para ser usada como gene de referência, principalmente quando os dados de variação de expressão não são muito grandes, às vezes, menores do que 10. Este resultado só foi encontrado ao se comparar os estágios de merozoítos e esporozoítos, pois utilizando os cinco estágios a flutuação do C_t foi muito superior, o que mostra a dificuldade de se encontrar genes expressos de forma constitutiva em todos os estágios de vida destes organismos. Em *Plasmodium falciparum* a super representação de proteínas com motivos de *zinc finger* está potencialmente

envolvida na regulação da estabilidade do RNAm (Natalang *et al.*, 2008) e assim como encontrado em nossos dados, apresenta-se diferencialmente expressa.

As histonas H2A, H2B, H3 e H4 são proteínas nucleares carregadas positivamente dispostas na forma de octâmeros que ficam envoltas ao DNA na sua forma *coiled* carregado negativamente, formando os nucleossomos. As histonas passam por grandes modificações pós-traducionais que podem influenciar na organização e estabilidade da estrutura da cromatina sendo importante para o controle da transcrição e do ciclo celular (Longhurst e Holder, 1997). Importante ressaltar que as histonas canônicas de eucariotos inferiores e plantas são poliadeniladas (Marzluff *et al.*, 2008). As histonas podem estar envolvidas na diferenciação dos estágios de vida de apicomplexa, sendo acompanhada por significantes alterações de expressão gênica, por processos de acetilação, metilação e presença de histonas variantes (Sullivan e Hakimi, 2006). A histona H2A é pequena e altamente conservada e foi escolhida como gene de referência para nossas análises, pois para merozoítos e esporozoítos (alvos de estudo deste trabalho) apresentou uma variação dos dados de C_t menor do que 1 log. Infelizmente este mesmo padrão não foi observado nos estágios de oocistos (dados não mostrados) o que provavelmente possa estar relacionado à presença de variantes. Recentemente, em *Toxoplasma gondii* foram caracterizados três variantes desta família cuja expressão é diferencial entre taquizoítos e bradizoítos (Dalmasso *et al.*, 2009).

A partir da escolha do gene de referência, os ensaios experimentais para verificar a expressão diferencial foram realizados. Foram testados 6 genes, três genes diferencialmente expressos (*up-regulated*) para cada estágio. Para a escolha dos genes, a seqüência da *tag* respectiva estava presente em todos os fragmentos amplificados.

Para esporozoítos foram avaliadas uma proteína hipotética, o fator de início da tradução 5 (eIF5) e uma proteína caracterizada como uma Glutamato/Leucina/Fenilalanina/Valina desidrogenase (ELFV_dehydrog.). O início da tradução é um processo celular sofisticado, bem regulado e altamente coordenado em eucariotos em que vários fatores de iniciação são utilizados, dentre os quais, o eIF5 identificado também em *Toxoplasma* (Sauvage *et al.*, 2006) e que representa um dos 6 fatores de inicio da transcrição caracterizados neste organismo.

As proteínas ELFV desidrogenase são estrutural e funcionalmente relacionadas por conterem uma região rica em glicina e um resíduo de lisina relacionado à atividade catalítica. Este motivo protéico também foi encontrado nos dados de proteômica de esporozoítos de *P. falciparum* (Lasonder *et al.*, 2008).

Quando analisamos nossos resultados, observamos que os dados quantitativos de expressão destes três genes foram muito próximos entre LongSAGE e RT-qPCR, apresentando um coeficiente de correlação de Pearson e R^2 de 0, 999.

Para os merozoítos foram selecionados os genes de actina, fator depolimerizante de actina e proteína de ribossomo S8. A actina e o fator depolimerizante de actina (ADF) estão relacionados ao processo de invasão dependente do sistema contráctil de actina-miosina. O ADF possui um papel importante no remodelamento do citoesqueleto de actina, sendo capaz de ligar-se a actina na forma polimerizada ou monomérica rompendo os filamentos de actina em fragmentos. Em *Toxoplasma gondii*, estudos de localização de actina foram particularmente difíceis por causa do rápido *turnover* destes filamentos promovido pelos ADFs, portanto, para realização destes estudos, estes fatores de depolimerização devem ser reprimidos (Schatten e Ris, 2002).

A análise de bancos de dados de seqüências revelou duas isoformas de ADF em *Eimeria* e *Toxoplasma*, no entanto apenas uma foi caracterizada em *Toxoplasma gondii* (Maciver e Hussey, 2002). Em *Eimeria tenella* o ADF foi observado em merozoítos, esporozoítos e oocistos (Xu *et al.*, 2008).

Em relação aos nossos dados, ambos os genes encontraram-se mais expressos nos experimentos de RT-qPCR do que em LongSAGE, apresentando uma diferença de cerca de 2 a 3 vezes maior nos dados de real-time. Este resultado pode significar que a concordância entre as duas plataformas não é tão boa devido a diferença de sensibilidade das duas técnicas e a alta especificidade do real-time. Alternativamente, o fato de considerarmos que todas as reações de RT-qPCR possuíam eficiências iguais, pode também ter influenciado os valores de C_t em um ou dois ciclos, o que poderia ter prejudicado a concordância inter-plataformas. Importante salientar que como no genoma de *Eimeria tenella* o fragmento amplificado está na fronteira íntron-éxon, variantes de *splicing* poderiam ser também amplificadas alterando os dados de expressão. Além disso, em buscas de similaridade nos ESTs montados de *Eimeria tenella* este fragmento também

apresenta dois alinhamentos perfeitos em *contigs* distintos. No entanto, quando os *primers* foram desenhados, a base de dados de ESTs não contava com o número atual de seqüências e portanto, estes dois alinhamentos não haviam sido detectados. Estes genes não foram escolhidos ao acaso, sabe-se que apesar do amplo uso da actina como gene de referência, a sua utilização não é aplicável a qualquer organismo sem uma padronização prévia. Um estudo recente sobre ADF de *Eimeria tenella* utilizou nos ensaios experimentais de RT-qPCR a actina como gene de referência (Xu *et al.*, 2008). Muito provavelmente a expressão de ADF detectada neste trabalho não reflete a realidade, já que pelos nossos dados, o gene de referência empregado apresenta uma expressão estágio específica significativa, sendo 3 vezes superior em merozoítos. Isto explica o fato dos autores não terem detectado grandes diferenças de expressão de ADF nos estágios *zoítas*, ao contrário do encontrado em nossos experimentos.

Já em relação à proteína de ribossomo S8, os dados quantitativos entre as plataformas foram mais equivalentes.

Analisando a expressão dos três genes em conjunto, observamos que em merozoítos a correlação entre real-time e LongSAGE foi moderada, com um coeficiente de Pearson de 0,67.

No entanto, analisando-se em conjunto os 6 genes, a correlação entre as duas técnicas foi boa com um coeficiente de correlação de Pearson igual a 0,79. Assim, estes dados serviram para fazer uma validação preliminar dos resultados obtidos em LongSAGE tanto em relação ao estágio de maior expressão quanto aos dados quantitativos.

Para realizar estudos comparativos entre dados de expressão, utilizamos os resultados de *Northern Digital* de ESTs/ORESTES de *Eimeria tenella*. Dos 3.008 eventos menos de 30% são compartilhados entre merozoítos de segunda geração e esporozoítos, resultados estes muito semelhantes ao observado nos dados de LongSAGE para estes mesmos estágios.

Nos dados de *Northern* Digital de ESTs/ORESTES, 660 transcritos foram classificados como diferencialmente expressos e, com relação aos dados de anotação com resultados positivos de BLAST, o perfil transcricional mostrou-se muito semelhante ao obtido em LongSAGE. Dentro do conjunto de genes expressos preferencialmente em merozoítos, uma grande quantidade de proteínas de ribossomo e antígenos de superfície foram encontradas além de algumas proteínas

de micronema, calmodulina, miosina e fator depolimerizante de actina. Em esporozoítos poucos resultados de BLAST positivos foram observados e os genes apresentaram funções diversas.

Comparando-se os dados de expressão diferencial entre LongSAGE (199 reconstruídos) e ESTs/ORESTES (660 seqüências), observamos que os 144 reconstruídos a partir de *tags* de LongSAGE encontram-se entre os genes diferencialmente expressos de ESTs/ORESTES. Somente um transcrito apresentou discordância em relação ao estágio mais expresso após a comparação inter-plataformas, ou seja, em LongSAGE mostrou-se diferencialmente expresso para esporozoítos (1,5 vezes mais expresso) e nos dados de ESTs foi mais freqüente em merozoítos (1,5 vezes mais expresso). Por se tratar de uma diferença de expressão ligeiramente pequena, este resultado não parece ser tão relevante.

As diferenças quanto à anotação automática foram pequenas e, devem-se principalmente aos critérios distintos utilizados em cada análise. Em LongSAGE foram considerados resultados positivos os alinhamentos com *e-values* inferiores a 1e-04 e porcentagem mínima de identidade igual a 45%, enquanto que em ESTs/ORESTES o critério foi mais estringente utilizando e-value inferior a 1e-06.

Apesar deste trabalho ter como objetivo principal a análise de transcritos diferencialmente expressos, resolvemos também comparar os dados inter-plataformas dos genes numericamente mais expressos. A concordância inter-plataforma desta vez não foi tão boa, menos de 40% dos transcritos foram compartilhados em merozoítos e menos de 60% em esporozoítos, no entanto, se considerarmos que são plataformas completamente distintas, com sensibilidade e vieses diferentes, este resultado parece ser aceitável.

Ao verificar os produtos protéicos diferencialmente expressos numericamente mais representativos nas duas plataformas, podemos destacar genes de grande interesse e importância, alguns destes já descritos anteriormente.

Em merozoítos encontramos o ADF e actina, o receptor 1 para quinase C ativada (RACK1) e as proteínas de micronema 4 e 5. Em esporozoítos podemos destacar a histona 3, que inclusive parece ter um controle pós transcricional a partir de RNA *antisense* e o facilitador de transporte de glicose. Estudos empregando oócitos de *Xenopus* demonstraram que RNAs complementares que codificam para transportadores de glicose em *P. kowlesi, P. yoelli* e *T. gondii* induzem um grande aumento de obtenção de glicose, cerca de 20 vezes (Joët *et al.*, 2002). Nas espécies

de *Plasmodium* estes transportadores já foram validados como alvos de drogas (Joët *et al.*, 2003). No entanto, um estudo recente em *Toxoplasma gondii* mostrou que este transportador não é essencial para a sobrevivência *in vitro* e para virulência *in vivo* dos taquizoítos (Blume *et al.*, 2009).

A proteína de micronema 4 foi encontrada em esporozoítos e merozoítos, e é diferencialmente expressa durante a esquizogonia (Tomley *et al.*, 2001). Possui 31 módulos EGF (epidermal growth factor) dos quais provavelmente 22 estão envolvidos na ligação a moléculas de cálcio, sendo secretada e expressa na superfície do parasita (Tomley *et al.*, 2001). As proteínas de micronema 4 e 5 interagem entre si formando um complexo protéico oligomérico de alto peso molecular. A formação de complexos entre estas proteínas de micronema solúveis e transmembranares provê um mecanismo onde estas proteínas atingem os micronemas através de um mecanismo de cooperação. Para isto, vários sinais presentes na cauda citosólica das proteínas transmembranares são utilizados com a participação ativa da carga contida nas proteínas solúveis, permitindo um correto dobramento e tráfico destas proteínas (Periz *et al.*, 2007).

O receptor 1 para quinase C ativada (RACK1) está implicado em múltiplas interações entre proteínas, incluindo sua manutenção estrutural para sinalização de moléculas. Devido às repetições WD, este receptor apresenta uma estrutura complexa que funciona como uma âncora. Esta proteína estrutural, relacionada à sinalização celular, foi estudada em *Toxoplasma gondii* e foi verificado que este receptor serve como ponto de integração para diversas vias de transdução de sinal. São capazes de se ligar a proteínas quinase incluindo as ativadas por cálcio, integrinas e fosfodiesterases, possuindo importância nos mecanismos de controle celular e transcrição. Devido à sua relação com a via dependente de cálcio, o RACK1 pode ser utilizado como importante alvo terapêutico interferindo nos mecanismos de invasão e evasão parasitária (Moran *et al.*, 2007).

As novas tecnologias de sequenciamento em larga escala capazes de gerar centena de milhares de sequências como o 454, o Solexa e o SOLiD (Hutchison, 2007) muito provavelmente irão permitir uma análise mais profunda do transcriptomas devido á sua alta cobertura e capacidade de gerar muitos dados em pouco tempo, entretanto, esta alta capacidade está associada a uma perda substancial de tamanho e precisão das seqüências geradas (Wold e Myers, 2008). E apesar de promissoras, ainda sofrem problemas relativos à vieses, por exemplo nas

seqüências provenientes de Solexa, o viés GC necessita de uma cobertura de 20 vezes para total diluição dos erros, e não é recomendada para estudos de SNPs (Dohm *et al.*, 2008). Um bilhão de seqüências de transcritos de *S.cerevisae* foi capaz de identificar somente 91% dos genes (Graveley, 2008). Cheung *et al.* (2008) comparam dados obtidos a partir de sequenciamento pelo método de Sanger e pirosequenciamento. Estes autores mostraram que estas técnicas são complementares e que muitas destas sequencias são detectadas por apenas uma das metodologias empregadas. Shin *et al.* (2008) analisaram o primeiro estágio larval de *C. elegans* gerando 30 milhões de bases utilizando equipamento 454 e encontraram uma sobreposição de apenas 50% com os dados de SAGE e pirosequenciamento.

A utilização da LongSAGE em organismos do gênero *Eimeria* é inédita e permitiu a análise quantitativa dos transcritos provenientes de dois estágios invasivos de *Eimeria tenella.* Estes dados poderão ser utilizados para diversos estudos envolvendo a caracterização de transcritos alternativos, como os derivados de *splicing* e também seqüências *antisense*, além de viabilizar a escolha de candidatos interessantes para estudos mais profundos relacionados à expressão gênica estágio-específica, mecanismo fundamental para o desenvolvimento do ciclo de vida destes parasitas.

6 CONCLUSÕES

- ✓ Foram construídas com sucesso bibliotecas de LongSAGE a partir dos estágios invasivos de merozoítos de segunda geração e esporozoítos de *E. tenella*, o que resultou em mais de 35.000 *tags,* das quais 9.516 são únicas;

- ✓ Para extração, contagem e análise estatística das *tags* um pacote de programas intitulado SAGE Analysis foi desenvolvido pela nossa equipe de bioinformática;

- ✓ Menos de 1% das *tags* encontradas é de alta expressão (>50x), o que significa que um pequeno conjunto de genes é altamente expresso em cada fase do parasita;

- ✓ Aproximadamente 60% das *tags* foram mapeadas no genoma *E. tenella*, das quais 95% foram localizadas em apenas um sítio, o que sugere um mapeamento inequívoco da maior parte das *tags*;

- ✓ A técnica de LongSAGE permitiu a identificação de transcritos alternativos, como os provenientes de poliadenilação heterogênea, *splicing* alternativo, transcritos *antisense*, além de *tags* mitocondriais;

- ✓ Após análise estatística foram identificados 270 *tags* diferencialmente expressas entre estes estágios invasivos, das quais 199 foram mapeadas e reconstruídas a partir dos dados do transcriptoma de *E. tenella* (ESTs/ORESTES), destas, 197 foram anotadas;

- ✓ Menos da metade dos transcritos anotados apresentou resultados positivos de BLAST, o que pode indicar que um grande número de genes de *E. tenella* ainda está por ser caracterizado, principalmente para esporozoítos;

- ✓ O protocolo de anotação automática gerou dados consistentes, 87,4% dos transcritos com resultados de BLAST positivos apresentaram similaridade com seqüências de organismos do Filo Apicomplexa, e poucas alterações foram efetuadas após a curagem manual;

- ✓ Foi observado um perfil semelhante entre os dados de anotação, GO e KOG, onde um número maior de produtos protéicos provenientes de merozoítos apresentou resultados positivos relacionados à tradução, manutenção e dobramento de proteínas, resposta à estresse, transdução de sinal e invasão. Em esporozoítos, os poucos resultados encontrados estão relacionados ao metabolismo de aminoácidos, transporte e atividade catalítica;

- ✓ O pequeno conjunto de genes submetidos a experimentos de RT-qPCR apresentou uma boa correlação com os dados obtidos por LongSAGE, tanto com relação à expressão estágio-específica, quanto aos dados quantitativos.

- ✓ A comparação dos dados de LongSAGE com os de *Northern* Digital de ESTs/ORESTES de *E. tenella* sugere que há uma boa correlação entre os dados diferencialmente expressos, tanto na anotação quanto na expressão estágio específica;

REFERÊNCIAS BIBLIOGRÁFICAS

Adams MD, Kelley JM, Cocaine JD, Dubnick M, Polymeropoulos MH, Xiao H, *et al*. Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 1991 Jun 21;252(5013):1651-6.

Akmaev VR, Wang CJ. Correction of sequence-based artifacts in serial analysis of gene expression. Bioinformatics. 2004 May 22;20(8):1254-63.

Alba R, Fei Z, Payton P, Liu Y, Moore SL, Debbie P, Cohn J, D'Ascenzo M, Gordon JS, Rose JK, Martin G, Tanksley SD, Bouzayen M, Jahn MM, Giovannoni J. ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development. Plant J. 2004 Sep;39(5):697-714.

Allen PC, Fetterer RH. Recent advances in biology and immunobiology of *Eimeria* species and in diagnosis and control of infection with these coccidian parasites of poultry. Clin Microbiol Rev. 2002 Jan;15(1):58-65.

Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol. 2009 Jan;5(1):e1000262.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.

Arikawa E, Sun Y, Wang J, Zhou Q, Ning B, Dial SL, Guo L, Yang J. Cross-platform comparison of SYBR Green real-time PCR with TaqMan PCR, microarrays and other gene expression measurement technologies evaluated in the MicroArray Quality Control (MAQC) study. BMC Genomics. 2008;9:328.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.

Assossou O, Besson F, Rouault JP, Persat F, Brisson C, Duret L, Ferrandiz J, Mayencon M, Peyron F, Picot S. Subcellular localization of 14-3-3 proteins in *Toxoplasma gondii* tachyzoites and evidence for a lipid raft-associated form. FEMS Microbiol Lett. 2003 Jul 29;224(2):161-8.

Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res. 1997 Oct;7(10):986-95.

De acordo com:
International Committee of Medical Journal Editors (ICMJE) – Vancouver Style.
http://www.nlm.nih.gov/bsd/uniform_requirements.html

Avison MB. Differential display, subtractive hybridization,amplification suppression and SAGE techniques for measuring gene expression. In: Avison MB, editor. Measuring Gene Expression. New York (USA): Taylor & Francis Group; 2008. p. 217-44.

Bakheit MA, Scholzen T, Ahmed JS, Seitzer U. Molecular characterization of a *Theileria lestoquardi* gene encoding for immunogenic protein splice variants. Parasitol Res. 2006 Dec;100(1):161-70.

Ball SJ, Pittilo RM, Long PL. Intestinal and extraintestinal life cycles of eimeriid coccidia. Adv Parasitol. 1989;28:1-54.

Bao J, Lee S, Chen C, Zhang X, Zhang Y, Liu S, Clark T, Wang J, Cao M, Yang H, Wang SM, Yu J. Serial analysis of gene expression study of a hybrid rice strain (LYP9) and its parental cultivars. Plant Physiol. 2005 Jul;138(3):1216-31.

Beatrix B, Sakai H, Wiedmann M. The alpha and beta subunit of the nascent polypeptide-associated complex have distinct functions. J Biol Chem. 2000 Dec 1;275(48):37838-45.

Bell AS, Ranford-Cartwright LC. Real-time quantitative PCR in parasitology. Trends Parasitol. 2002 Aug;18(8):337-42.

Belli SI, Walker RA, Flowers SA. Global protein expression analysis in apicomplexan parasites: current status. Proteomics. 2005 Mar;5(4):918-24.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol. 2004 Jul 16;340(4):783-95.

Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999 Jan 15;27(2):573-80.

Bianchetti L, Wu Y, Guerin E, Plewniak F, Poch O. SAGETTARIUS: a program to reduce the number of tags mapped to multiple transcripts and to plan SAGE sequencing stages. Nucleic Acids Res. 2007;35(18):e122.

Blake DP, Qin Z, Cai J, Smith AL. Development and validation of real-time polymerase chain reaction assays specific to four species of Eimeria. Avian Pathol. 2008 Feb;37(1):89-94.

Blume M, Rodriguez-Contreras D, Landfear S, Fleige T, Soldati-Favre D, Lucius R, Gupta N. Host-derived glucose and its transporter in the obligate intracellular pathogen *Toxoplasma gondii* are dispensable by glutaminolysis. Proc Natl Acad Sci U S A. 2009 Aug 4;106(31):12998-3003.

Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, Morin PJ, Buetow KH, Strausberg RL, De Souza SJ, Riggins GJ. An anatomy of normal and malignant gene expression. Proc Natl Acad Sci U S A. 2002 Aug 20;99(17):11287-92.

Bozdech Z, Ginsburg H. Antioxidant defense in *Plasmodium falciparum* - data mining of the transcriptome. Malar J. 2004 Jul 9;3:23.

Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL. Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. Genome Biol. 2003;4(2):R9.

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol. 2000 Jun;18(6):630-4.

Bromley E, Leeds N, Clark J, McGregor E, Ward M, Dunn MJ, Tomley F. Defining the protein repertoire of microneme secretory organelles in the apicomplexan parasite *Eimeria tenella*. Proteomics. 2003 Aug;3(8):1553-61.

Bumstead J, Tomley F. Induction of secretion and surface capping of microneme proteins in *Eimeria tenella*. Mol Biochem Parasitol. 2000 Oct;110(2):311-21.

Cai X, Fuller AL, McDougald LR, Zhu G. Apicoplast genome of the coccidian *Eimeria tenella*. Gene. 2003 Dec 4;321:39-46.

Calsa T, Jr., Figueira A. Serial analysis of gene expression in sugarcane (*Saccharum* spp.) leaves revealed alternative C4 metabolism and putative antisense transcripts. Plant Mol Biol. 2007 Apr;63(6):745-62.

Chapman HD, Shirley MW. The Houghton strain of *Eimeria tenella*: a review of the type strain selected for genome sequencing. Avian Pathol. 2003 Apr;32(2):115-27.

Chapman HD, Cherry TE, Danforth HD, Richards G, Shirley MW, Williams RB. Sustainable coccidiosis control in poultry production: the role of live vaccines. Int J Parasitol. 2002 May;32(5):617-29.

Chaudhary K, Donald RG, Nishi M, Carter D, Ullman B, Roos DS. Differential localization of alternatively spliced hypoxanthine-xanthine-guanine phosphoribosyltransferase isoforms in *Toxoplasma gondii*. J Biol Chem. 2005 Jun 10;280(23):22053-9.

Chen J, Sun M, Lee S, Zhou G, Rowley JD, Wang SM. Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. Proc Natl Acad Sci U S A. 2002 Sep 17;99(19):12257-62.

Chen J, Sun M, Kent WJ, Huang X, Xie H, Wang W, Zhou G, Shi RZ, Rowley JD. Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Res. 2004;32(16):4812-20.

Cheung F, Win J, Lang JM, Hamilton J, Vuong H, Leach JE, Kamoun S, Andre Levesque C, Tisserat N, Buell CR. Analysis of the *Pythium ultimum* transcriptome using Sanger and Pyrosequencing approaches. BMC Genomics. 2008;9:542.

Clark JD, Billington K, Bumstead JM, Oakes RD, Soon PE, Sopp P, Tomley FM, Blake DP. A toolbox facilitating stable transfection of *Eimeria* species. Mol Biochem Parasitol. 2008 Nov;162(1):77-86.

Cleary MD, Singh U, Blader IJ, Brewer JL, Boothroyd JC. *Toxoplasma gondii* asexual development: identification of developmentally regulated genes and distinct patterns of gene expression. Eukaryot Cell. 2002 Jun;1(3):329-40.

Cohen-Freue G, Holzer TR, Forney JD, McMaster WR. Global gene expression in *Leishmania*. Int J Parasitol. 2007 Aug;37(10):1077-86.

Cox D. Partial likelihood. Biometrika. 1975;62(2):269-76.

Current WL, Upton SJ, Long PL. Taxonomy and Life Cycles. In: Long PL, editor. Coccidiosis of Man and Domestical Animals. Boston: CRC Press Inc; 1990. p. 1-17.

Dahary D, Elroy-Stein O, Sorek R. Naturally occurring antisense: transcriptional leakage or real overlap? Genome Res. 2005 Mar;15(3):364-8.

Dalloul RA, Lillehoj HS. Recent advances in immunomodulation and vaccination strategies against coccidiosis. Avian Dis. 2005 Mar;49(1):1-8.

Dalmasso MC, Onyango DO, Naguleswaran A, Sullivan WJ Jr, Angel SO. *Toxoplasma* H2A variants reveal novel insights into nucleosome composition and functions for this histone family. J Mol Biol. 2009 Sep 11;392(1):33-47.

de Venevelles P, Chich JF, Faigle W, Loew D, Labbe M, Girard-Misguich F, Pery P. Towards a reference map of *Eimeria tenella* sporozoite proteins by two-dimensional electrophoresis and mass spectrometry. Int J Parasitol. 2004 Nov;34(12):1321-31.

Deepak S, Kottapalli K, Rakwal R, Oros G, Rangappa K, Iwahashi H, Masuo Y, Agrawal G. Real-Time PCR: Revolutionizing Detection and Expression Analysis of Genes. Curr Genomics. 2007 Jun;8(4):234-51.

Delbac F, Sanger A, Neuhaus EM, Stratmann R, Ajioka JW, Toursel C, Herm-Gotz A, Tomavo S, Soldati T, Soldati D. *Toxoplasma gondii* myosins B/C: one gene, two tails, two localizations, and a role in parasite division. J Cell Biol. 2001 Nov 12;155(4):613-23.

Delbecq S, Precigout E, Schetters T, Gorenflot A. *Babesia divergens*: cloning of a Ran binding protein 1 homologue. Vet Parasitol. 2003 Jul 29;115(3):205-11.

DeRocher AE, Coppens I, Karnataki A, Gilbert LA, Rome ME, Feagin JE, Bradley PJ, Parsons M. A thioredoxin family protein of the apicoplast periphery identifies

abundant candidate transport vesicles in *Toxoplasma gondii*. Eukaryot Cell. 2008 Sep;7(9):1518-29.

Dessimoz C, Boeckmann B, Roth AC, Gonnet GH. Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. Nucleic Acids Res. 2006;34(11):3309-16.

Dheda K, Huggett JF, Bustin SA, Johnson MA, Rook G, Zumla A. Validation of housekeeping genes for normalizing RNA expression in real-time PCR. Biotechniques. 2004 Jul;37(1):112-4, 6, 8-9.

Dias Neto E, Correa RG, Verjovski-Almeida S, Briones MR, Nagai MA, da Silva W, Jr., Zago MA, Bordin S, Costa FF, Goldman GH, Carvalho AF, Matsukuma A, Baia GS, Simpson DH, Brunstein A, de Oliveira PS, Bucher P, Jongeneel CV, O'Hare MJ, Soares F, Brentani RR, Reis LF, de Souza SJ, Simpson AJ. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proc Natl Acad Sci U S A. 2000 Mar 28;97(7):3491-6.

Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED, Siebert PD. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc Natl Acad Sci U S A. 1996 Jun 11;93(12):6025-30.

Dimmer EC, Huntley RP, Barrell DG, Binns D, Draghici S, Camon EB, Hubank M, Talmud PJ, Apweiler R, Lovering RC. The Gene Ontology - Providing a Functional Role in Proteomic Studies. Proteomics. 2008 Jul 17.

Dinel S, Bolduc C, Belleau P, Boivin A, Yoshioka M, Calvo E, Piedboeuf B, Snyder EE, Labrie F, St-Amand J. Reproducibility, bioinformatic analysis and power of the SAGE method to evaluate changes in transcriptome. Nucleic Acids Res. 2005;33(3):e26.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 2008 Sep;36(16):e105.

Dubremetz JF, Garcia-Reguet N, Conseil V, Fourmaux MN. Apical organelles and host-cell invasion by Apicomplexa. Int J Parasitol. 1998 Jul;28(7):1007-13.

Dufva M. Introduction to Microarray Technology. In: Dufva M, editor. DNA Microarrays for Biomedical Research. New York (USA): Humana Press; 2009. p. 1-22.

Duncan R. DNA microarray analysis of protozoan parasite gene expression: outcomes correlate with mechanisms of regulation. Trends Parasitol. 2004 May;20(5):211-5.

Durham AM, Kashiwabara AY, Matsunaga FT, Ahagon PH, Rainone F, Varuzza L, Gruber A. EGene: a configurable pipeline generation system for automated sequence analysis. Bioinformatics. 2005 Jun 15;21(12):2812-3.

Dussault AA, Pouliot M. Rapid and simple comparison of messenger RNA levels using real-time PCR. Biol Proced Online. 2006;8:1-10.

Echeverria PC, Matrajt M, Harb OS, Zappia MP, Costas MA, Roos DS, Dubremetz JF, Angel SO. *Toxoplasma gondii* Hsp90 is a potential drug target whose expression and subcellular localization are developmentally regulated. J Mol Biol. 2005 Jul 22;350(4):723-34.

Emmersen J, Heidenblut AM, Hogh AL, Hahn SA, Welinder KG, Nielsen KL. Discarding duplicate ditags in LongSAGE analysis may introduce significant error. BMC Bioinformatics. 2007;8:92.

Entzeroth R, Mattig FR, Werner-Meier R. Structure and function of the parasitophorous vacuole in *Eimeria* species. Int J Parasitol. 1998 Jul;28(7):1015-8.

Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998 Mar;8(3):186-94.

Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 1998 Mar;8(3):175-85.

Fernandez S, Pagotto AH, Furtado MM, Katsuyama AM, Madeira AM, Gruber A. A multiplex PCR assay for the simultaneous detection and discrimination of the seven *Eimeria* species that infect domestic fowl. Parasitology. 2003 Oct;127(Pt 4):317-25.

Fernando MA. *Eimeria*: Infections of the Intestine. In: Long PL, editor. Coccidiosis of Man and Domestical Animals. Boston: CRC Press Inc; 1990. p. 63-75.

Ferro M. Desenvolvimento e validação de protocolos para a anotação automática de seqüências ORESTES de *Eimeria* spp. de galinha doméstica [Dissertação (Mestrado em Ciências - Biologia da Relação Patógeno Hospedeiro)]. São Paulo: Instituto de Ciências Biomédicas da Universidade de São Paulo; 2008.

Fizames C, Munos S, Cazettes C, Nacry P, Boucherez J, Gaymard F, Piquemal D, Delorme V, Commes T, Doumas P, Cooke R, Marti J, Sentenac H, Gojon A. The *Arabidopsis* root transcriptome by serial analysis of gene expression. Gene identification using the genome sequence. Plant Physiol. 2004 Jan;134(1):67-80.

Fleige S, Pfaffl MW. RNA integrity and the effect on the real-time qRT-PCR performance. Mol Aspects Med. 2006 Apr-Jun;27(2-3):126-39.

Fonager J, Cunningham D, Jarra W, Koernig S, Henneman AA, Langhorne J, Preiser P. Transcription and alternative splicing in the yir multigene family of the malaria parasite *Plasmodium y. yoelii*: identification of motifs suggesting epigenetic and post-transcriptional control of RNA expression. Mol Biochem Parasitol. 2007 Nov;156(1):1-11.

Frankel MB, Knoll LJ. Functional analysis of key nuclear trafficking components reveals an atypical Ran network required for parasite pathogenesis. Mol Microbiol. 2008 Oct;70(2):410-20.

Frankel MB, Knoll LJ. The ins and outs of nuclear trafficking: unusual aspects in apicomplexan parasites. DNA Cell Biol. 2009 Jun;28(6):277-84.

Funaguma S, Hashimoto S, Suzuki Y, Omuro N, Sugano S, Mita K, Katsuma S, Shimada T. SAGE analysis of early oogenesis in the silkworm, *Bombyx mori*. Insect Biochem Mol Biol. 2007 Feb;37(2):147-54.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 2002 Oct 3;419(6906):498-511.

Ge X, Wu Q, Wang SM. SAGE detects microRNA precursors. BMC Genomics. 2006;7:285.

Georgieva NV, Koinarski V, Gadjeva V. Antioxidant status during the course of *Eimeria tenella* infection in broiler chickens. Vet J. 2006 Nov;172(3):488-92.

Gibbings JG, Cook BP, Dufault MR, Madden SL, Khuri S, Turnbull CJ, Dunwell JM. Global transcript analysis of rice leaf and seed using SAGE technology. Plant Biotechnol J. 2003 Jul;1(4):271-85.

Gillespie DE, Salazar NA, Rehkopf DH, Feagin JE. The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum* have short A tails. Nucleic Acids Res. 1999 Jun 1;27(11):2416-22.

Ginger ML. Niche metabolism in parasitic protozoa. Philos Trans R Soc Lond B Biol Sci. 2006 Jan 29;361(1465):101-18.

Gissot M, Briquet S, Refour P, Boschet C, Vaquero C. PfMyb1, a *Plasmodium falciparum* transcription factor, is required for intra-erythrocytic growth and controls key genes for cell cycle regulation. J Mol Biol. 2005 Feb 11;346(1):29-42.

González-Pons M, Szeto AC, González-Mendez R, Serrano AE. Identification and bioinformatic characterization of a multidrug resistance associated protein (ABCC) gene in Plasmodium berghei. Malar J. 2009;8:1.

Graveley BR. Molecular biology: power sequencing. Nature. 2008 Jun 26;453(7199):1197-8.

Green CD, Simons JF, Taillon BE, Lewin DA. Open systems: panoramic views of gene expression. J Immunol Methods. 2001 Apr;250(1-2):67-79.

Gruber A. Expressed sequence tags. In: Dear PH, editor. Bioinformatics. Bloxham Mill, Oxfordshire (UK): Scion Publishing Limited; 2007. p. 141-67.

Guerfali FZ, Laouini D, Guizani-Tabbane L, Ottones F, Ben-Aissa K, Benkahla A, Manchon L, Piquemal D, Smandi S, Mghirbi O, Commes T, Marti J, Dellagi K.

Simultaneous gene expression profiling in human macrophages infected with *Leishmania major* parasites using SAGE. BMC Genomics. 2008;9:238.

Gunasekera AM, Patankar S, Schug J, Eisen G, Kissinger J, Roos D, Wirth DF. Widespread distribution of antisense transcripts in the *Plasmodium falciparum* genome. Mol Biochem Parasitol. 2004 Jul;136(1):35-42.

Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, Berriman M, Florens L, Janssen CS, Pain A, Christophides GK, James K, Rutherford K, Harris B, Harris D, Churcher C, Quail MA, Ormond D, Doggett J, Trueman HE, Mendoza J, Bidwell SL, Rajandream MA, Carucci DJ, Yates JR, 3rd, Kafatos FC, Janse CJ, Barrell B, Turner CM, Waters AP, Sinden RE. A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. Science. 2005 Jan 7;307(5706):82-6.

Hammod DM. Life Cycles and Development of Coccidia. In: Hammod DM, Long PL, editors. The Coccidia. Baltimore: University Park Press e London- Butterworths; 1973. p. 45-80.

Harbers M, Carninci P. Tag-based approaches for transcriptome research and genome annotation. Nat Methods. 2005 Jul;2(7):495-502.

Hayes R, Kudla J, Gruissem W. Degrading chloroplast mRNA: the role of polyadenylation. Trends Biochem Sci. 1999 May;24(5):199-202.

Heid CA, Stevens J, Livak KJ, Williams PM. Real time quantitative PCR. Genome Res. 1996 Oct;6(10):986-94.

Hene L, Sreenu VB, Vuong MT, Abidi SH, Sutton JK, Rowland-Jones SL, Davis SJ, Evans EJ. Deep analysis of cellular transcriptomes - LongSAGE versus classic MPSS. BMC Genomics. 2007;8:333.

Herm-Götz A, Weiss S, Stratmann R, Fujita-Becker S, Ruff C, Meyhöfer E, Soldati T, Manstein DJ, Geeves MA, Soldati D. *Toxoplasma gondii* myosin A and its light chain: a fast, single-headed, plus-end-directed motor. EMBO J. 2002 May 1;21(9):2149-58.

Hosse RJ, Krücken J, Bierbaum S, Greif G, Wunderlich F. *Eimeria tenella*: genomic organization and expression of an 89kDa cyclophilin. Exp Parasitol. 2008 Feb;118(2):275-9.

Hu K, Johnson J, Florens L, Fraunholz M, Suravajjala S, DiLullo C, Yates J, Roos DS, Murray JM. Cytoskeletal components of an invasion machine--the apical complex of *Toxoplasma gondii*. PLoS Pathog. 2006 Feb;2(2):e13.

Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999 Sep;9(9):868-77.

Hutchison CA, 3rd. DNA sequencing: bench to bedside and beyond. Nucleic Acids Res. 2007;35(18):6227-37.

Jaluria P, Konstantopoulos K, Betenbaugh M, Shiloach J. A perspective on microarrays: current applications, pitfalls, and potential uses. Microb Cell Fact. 2007;6:4.

Joët T, Eckstein-Ludwig U, Morin C, Krishna S. Validation of the hexose transporter of *Plasmodium falciparum* as a novel drug target. Proc Natl Acad Sci U S A. 2003 Jun 24;100(13):7476-9.

Joët T, Holterman L, Stedman TT, Kocken CH, Van Der Wel A, Thomas AW, Krishna S. Comparative characterization of hexose transporters of *Plasmodium knowlesi*, *Plasmodium yoelii* and *Toxoplasma gondii* highlights functional differences within the apicomplexan family. Biochem J. 2002 Dec 15;368(Pt 3):923-9.

Johnston DA, Fernando MA. *Eimeria* spp. of the domestic fowl: analysis of genetic variability between species and strains using DNA polymorphisms amplified by arbitrary primers and denaturing gradient-gel electrophoresis. Parasitol Res. 1995;81(2):91-7.

Jones SJ, Riddle DL, Pouzyrev AT, Velculescu VE, Hillier L, Eddy SR, Stricklin SL, Baillie DL, Waterston R, Marra MA. Changes in gene expression associated with developmental arrest and longevity in *Caenorhabditis elegans*. Genome Res. 2001 Aug;11(8):1346-52.

Kall L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 2004 May 14;338(5):1027-36.

Kappe SH, Gardner MJ, Brown SM, Ross J, Matuschewski K, Ribeiro JM, Adams JH, Quackenbush J, Cho J, Carucci DJ, Hoffman SL, Nussenzweig V. Exploring the transcriptome of the malaria sporozoite stage. Proc Natl Acad Sci U S A. 2001 Aug 14;98(17):9895-900.

Kawahara F, Taira K, Nagai S, Onaga H, Onuma M, Nunoya T. Detection of five avian *Eimeria* species by species-specific real-time polymerase chain reaction assay. Avian Dis. 2008 Dec;52(4):652-6.

Keime C, Semon M, Mouchiroud D, Duret L, Gandrillon O. Unexpected observations after mapping LongSAGE tags to the human genome. BMC Bioinformatics. 2007;8:154.

Kempthorn O. Of what use are tests of significance and tests of hypothesis. Commun Stat-Theor M. 1976;5(8):763-77.

Khattra J, Delaney AD, Zhao Y, Siddiqui A, Asano J, McDonald H, Pandoh P, Dhalla N, Prabhu AL, Ma K, Lee S, Ally A, Tam A, Sa D, Rogers S, Charest D, Stott J, Zuyderduyn S, Varhol R, Eaves C, Jones S, Holt R, Hirst M, Hoodless PA, Marra MA. Large-scale production of SAGE libraries from microdissected tissues, flow-sorted cells, and cell lines. Genome Res. 2007 Jan;17(1):108-16.

Kim YC, Jung YC, Xuan Z, Dong H, Zhang MQ, Wang SM. Pan-genome isolation of low abundance transcripts using SAGE tag. FEBS Lett. 2006 Dec 11;580(28-29):6721-9.

Kinnaird JH, Bumstead JM, Mann DJ, Ryan R, Shirley MW, Shiels BR, Tomley FM. EtCRK2, a cyclin-dependent kinase gene expressed during the sexual and asexual phases of the *Eimeria tenella* life cycle. Int J Parasitol. 2004 May;34(6):683-92.

Kirschbaum-Slager N, Parmigiani RB, Camargo AA, de Souza SJ. Identification of human exons overexpressed in tumors through the use of genome and expressed sequence data. Physiol Genomics. 2005 May 11;21(3):423-32.

Kiyosawa H, Yamanaka I, Osato N, Kondo S, Hayashizaki Y. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. Genome Res. 2003 Jun;13(6B):1324-34.

Klemba M, Gluzman I, Goldberg DE. A *Plasmodium falciparum* dipeptidyl aminopeptidase I participates in vacuolar hemoglobin degradation. J Biol Chem. 2004 Oct 8;279(41):43000-7.

Klotz C, Marhofer RJ, Selzer PM, Lucius R, Pogonka T. *Eimeria tenella*: identification of secretory and surface proteins from expressed sequence tags. Exp Parasitol. 2005 Sep;111(1):14-23.

Knoll-Gellida A, Andre M, Gattegno T, Forgue J, Admon A, Babin PJ. Molecular phenotype of zebrafish ovarian follicle by serial analysis of gene expression and proteomic profiling, and comparison with the transcriptomes of other animals. BMC Genomics. 2006;7:46.

Kogut MH. Host Specificity of the Coccidia. In: Long PL, editor. Coccidiosis of Man and Domestical Animals. Boston: CRC Press Inc; 1990. p. 43-62.

Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39:309-38.

Kozian DH, Kirschbaum BJ. Comparative gene-expression analysis. Trends Biotechnol. 1999 Feb;17(2):73-8.

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 Jan 19;305(3):567-80.

Krücken J, Greif G, von Samson-Himmelstjerna G. In silico analysis of the cyclophilin repertoire of apicomplexan parasites. Parasit Vectors. 2009;2(1):27.

Kubista M, Andrade JM, Bengtsson M, Forootan A, Jonak J, Lind K, Sindelka R, Sjoback R, Sjogreen B, Strombom L, Stahlberg A, Zoric N. The real-time polymerase chain reaction. Mol Aspects Med. 2006 Apr-Jun;27(2-3):95-125.

Kuersten S, Goodwin EB. The power of the 3' UTR: translational control and development. Nat Rev Genet. 2003 Aug;4(8):626-37.

Kuo BY, Chen Y, Bohacec S, Johansson O, Wasserman WW, Simpson EM. SAGE2Splice: unmapped SAGE tags reveal novel splice junctions. PLoS Comput Biol. 2006 Apr;2(4):e34.

Lal K, Bromley E, Oakes R, Prieto JH, Sanderson SJ, Kurian D, Hunt L, Yates JR, 3rd, Wastling JM, Sinden RE, Tomley FM. Proteomic comparison of four *Eimeria tenella* life-cycle stages: unsporulated oocyst, sporulated oocyst, sporozoite and second-generation merozoite. Proteomics. 2009 Oct;9(19):4566-76.

Lasonder E, Janse CJ, van Gemert GJ, Mair GR, Vermunt AM, Douradinha BG, van Noort V, Huynen MA, Luty AJ, Kroeze H, Khan SM, Sauerwein RW, Waters AP, Mann M, Stunnenberg HG. Proteomic profiling of *Plasmodium* sporozoite maturation identifies new proteins essential for parasite development and infectivity. PLoS Pathog. 2008 Oct;4(10):e1000195.

Levine ND. Introduction, History, and Taxonomy. In: Hammod DM, Long PL, editors. The Coccidia. Baltimore: University Park Press e London- Butterworths; 1973. p. 1-22.

Levine ND. Progress in taxonomy of the Apicomplexan protozoa. J Protozool. 1988 Nov;35(4):518-20.

Li L, Brunk BP, Kissinger JC, Pape D, Tang K, Cole RH, Martin J, Wylie T, Dante M, Fogarty SJ, Howe DK, Liberator P, Diaz C, Anderson J, White M, Jerome ME, Johnson EA, Radke JA, Stoeckert CJ, Jr., Waterston RH, Clifton SW, Roos DS, Sibley LD. Gene discovery in the apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. Genome Res. 2003 Mar;13(3):443-54.

Li Q, Zhao Y, Ni B, Yao C, Zhou Y, Xu W, Wang Z, Qiao Z. Comparison of the expression profiles of promastigotes and axenic amastigotes in *Leishmania donovani* using serial analysis of gene expression. Parasitol Res. 2008 Sep;103(4):821-8.

Li YJ, Xu P, Qin X, Schmechel DE, Hulette CM, Haines JL, Pericak-Vance MA, Gilbert JR. A comparative analysis of the information content in long and short SAGE libraries. BMC Bioinformatics. 2006;7:504.

Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science. 1992 Aug 14;257(5072):967-71.

Lillehoj HS, Lillehoj EP. Avian coccidiosis. A review of acquired intestinal immunity and vaccination strategies. Avian Dis. 2000 Apr-Jun;44(2):408-25.

Ling KH, Rajandream MA, Rivailler P, Ivens A, Yap SJ, Madeira AM, Mungall K, Billington K, Yee WY, Bankier AT, Carroll F, Durham AM, Peters N, Loo SS, Isa MN, Novaes J, Quail M, Rosli R, Nor Shamsudin M, Sobreira TJ, Tivey AR, Wai SF, White S, Wu X, Kerhornou A, Blake D, Mohamed R, Shirley M, Gruber A, Berriman M, Tomley F, Dear PH, Wan KL. Sequencing and analysis of chromosome 1 of *Eimeria tenella* reveals a unique segmental organization. Genome Res. 2007 Mar;17(3):311-9.

Llinas M, DeRisi JL. Pernicious plans revealed: *Plasmodium falciparum* genome wide expression analysis. Curr Opin Microbiol. 2004 Aug;7(4):382-7.

Long EL, Sonstegard TS, Long JA, Van Tassell CP, Zuelke KA. Serial analysis of gene expression in turkey sperm storage tubules in the presence and absence of resident sperm. Biol Reprod. 2003 Aug;69(2):469-74.

Long PL, Joyner LP. Problems in the identification of species of *Eimeria*. J Protozool. 1984 Nov;31(4):535-41.

Long PL, Millard BJ, Joyner LP, Norton CC. A guide to laboratory techniques used in the study and diagnosis of avian coccidiosis. Folia Vet Lat. 1976 Jul-Sep;6(3):201-17.

Longhurst HJ, Holder AA. The histones of *Plasmodium falciparum*: identification, purification and a possible role in the pathology of malaria. Parasitology. 1997 May;114 (Pt 5):413-9.

Lu J, Lal A, Merriman B, Nelson S, Riggins G. A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. Genomics. 2004 Oct;84(4):631-6.

Maciver SK, Hussey PJ. The ADF/cofilin family: actin-remodeling proteins. Genome Biol. 2002;3(5):reviews3007.

MacPherson JM, Gajadhar AA. Differentiation of seven *Eimeria* species by random amplified polymorphic DNA. Vet Parasitol. 1993 Jan;45(3-4):257-66.

Malig R, Varela C, Agosin E, Melo F. Accurate and unambiguous tag-to-gene mapping in serial analysis of gene expression. BMC Bioinformatics. 2006;7:487.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res. 2002 Jan 1;30(1):281-3.

Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 2007 Jan;35(Database issue):D237-40.

Margulies EH, Kardia SL, Innis JW. Identification and prevention of a GC content bias in SAGE libraries. Nucleic Acids Res. 2001 Jun 15;29(12):E60-0.

Martin DM, Berriman M, Barton GJ. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics. 2004 Nov 18;5:178.

Marzluff WF, Wagner EJ, Duronio RJ. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. Nat Rev Genet. 2008 Nov;9(11):843-54.

Meissner M, Agop-Nersesian C, Sullivan WJ, Jr. Molecular tools for analysis of gene function in parasitic microorganisms. Appl Microbiol Biotechnol. 2007 Jul;75(5):963-75.

Militello KT, Patel V, Chessler AD, Fisher JK, Kasper JM, Gunasekera A, Wirth DF. RNA polymerase II synthesizes antisense RNA in *Plasmodium falciparum*. RNA. 2005 Apr;11(4):365-70.

Min W, Dalloul RA, Lillehoj HS. Application of biotechnological tools for coccidia vaccine development. J Vet Sci. 2004 Dec;5(4):279-88.

Miska KB, Fetterer RH, Barfield RC. Analysis of transcripts expressed by *Eimeria tenella* oocysts using subtractive hybridization methods. J Parasitol. 2004 Dec;90(6):1245-52.

Miska KB, Fetterer RH, Rosenberg GH. Analysis of transcripts from intracellular stages of *Eimeria acervulina* using expressed sequence tags. J Parasitol. 2008 Apr;94(2):462-6.

Miska KB, Fetterer RH, Min W, Lillehoj HS. Heat shock protein 90 genes of two species of poultry *Eimeria*: expression and evolutionary analysis. J Parasitol. 2005 Apr;91(2):300-6.

Moody DE. Genomics Techniques: An overview of methods for the study of gene expression. J Anim Sci. 2001;79:E128-E35.

Moon RW, Taylor CJ, Bex C, Schepers R, Goulding D, Janse CJ, Waters AP, Baker DA, Billker O. A cyclic GMP signalling module that regulates gliding motility in a malaria parasite. PLoS Pathog. 2009 Sep;5(9):e1000599.

Moran JM, Smith SS, Hager KM. *Toxoplasma gondii* possesses a receptor for activated C kinase ortholog. Biochem Biophys Res Commun. 2007 Nov 23;363(3):680-6.

Morgan JA, Morris GM, Wlodek BM, Byrnes R, Jenner M, Constantinoiu CC, Anderson GR, Lew-Tabor AE, Molloy JB, Gasser RB, Jorgensen WK. Real-time polymerase chain reaction (PCR) assays for the specific detection and quantification of seven *Eimeria* species that cause coccidiosis in chickens. Mol Cell Probes. 2009 Apr;23(2):83-9.

Morrissette NS, Sibley LD. Cytoskeleton of apicomplexan parasites. Microbiol Mol Biol Rev. 2002 Mar;66(1):21-38; table of contents.

Morse DL, Carroll D, Weberg L, Borgstrom MC, Ranger-Moore J, Gillies RJ. Determining suitable internal standards for mRNA quantification of increasing cancer progression in human breast cells by real-time reverse transcriptase polymerase chain reaction. Anal Biochem. 2005 Jul 1;342(1):69-77.

Mulder N, Apweiler R. InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol. 2007;396:59-70.

Mullapudi N, Joseph SJ, Kissinger JC. Identification and functional characterization of cis-regulatory elements in the apicomplexan parasite *Toxoplasma gondii*. Genome Biol. 2009;10(4):R34.

Muller PY, Janovjak H, Miserez AR, Dobbie Z. Processing of gene expression data generated by quantitative real-time RT-PCR. Biotechniques. 2002 Jun;32(6):1372-4, 6, 8-9.

Munasinghe A, Patankar S, Cook BP, Madden SL, Martin RK, Kyle DE, Shoaibi A, Cummings LM, Wirth DF. Serial analysis of gene expression (SAGE) in *Plasmodium falciparum*: application of the technique to A-T rich genomes. Mol Biochem Parasitol. 2001 Mar;113(1):23-34.

Nagaraj SH, Gasser RB, Ranganathan S. A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform. 2007 Jan;8(1):6-21.

Nakonieczna J, Kaczorowski T, Obarska-Kosinska A, Bujnicki JM. Functional analysis of MmeI from methanol utilizer Methylophilus methylotrophus, a subtype IIC restriction-modification enzyme related to type I enzymes. Appl Environ Microbiol. 2009 Jan;75(1):212-23.

Natalang O, Bischoff E, Deplaine G, Proux C, Dillies MA, Sismeiro O, Guigon G, Bonnefoy S, Patarapotikul J, Mercereau-Puijalon O, Coppee JY, David PH. Dynamic RNA profiling in *Plasmodium falciparum* synchronized blood stages exposed to lethal doses of artesunate. BMC Genomics. 2008;9:388.

Nene V, Morzaria S, Bishop R. Organisation and informational content of the *Theileria parva* genome. Mol Biochem Parasitol. 1998 Sep 1;95(1):1-8.

Ng ST, Sanusi Jangi M, Shirley MW, Tomley FM, Wan KL. Comparative EST analyses provide insights into gene expression in two asexual developmental stages of Eimeria tenella. Exp Parasitol. 2002 Jun-Jul;101(2-3):168-73.

Novaes J, Kashiwabara AY, Varuzza L, Nagao LT, Manha APS, Fernandez S, Durham AM, Gruber A, Madeira AMBN. Survey of *Eimeria* spp. transcripts using open reading frame ESTs (ORESTES). In: The IXth International Coccidiosis Conference; 2005; Foz do Iguassu, Parana, Brazil. 2005. p. 150.

Obermeier C, Hosseini B, Friedt W, Snowdon R. Gene expression profiling via LongSAGE in a non-model plant species: a case study in seeds of *Brassica napus*. BMC Genomics. 2009;10:295.

Ojopi EP, Oliveira PS, Nunes DN, Paquola A, DeMarco R, Gregorio SP, Aires KA, Menck CF, Leite LC, Verjovski-Almeida S, Dias-Neto E. A quantitative view of the transcriptome of *Schistosoma mansoni* adult-worms using SAGE. BMC Genomics. 2007;8:186.

Ossorio PN, Sibley LD, Boothroyd JC. Mitochondrial-like DNA sequences flanked by direct and inverted repeats in the nuclear genome of *Toxoplasma gondii*. J Mol Biol. 1991 Dec 5;222(3):525-36.

Palm D, Weiland M, McArthur AG, Winiecka-Krusnell J, Cipriano MJ, Birkeland SR, Pacocha SE, Davids B, Gillin F, Linder E, Svard S. Developmental changes in the adhesive disk during *Giardia* differentiation. Mol Biochem Parasitol. 2005 Jun;141(2):199-207.

Patankar S, Munasinghe A, Shoaibi A, Cummings LM, Wirth DF. Serial analysis of gene expression in *Plasmodium falciparum* reveals the global expression profile of erythrocytic stages and the presence of anti-sense transcripts in the malarial parasite. Mol Biol Cell. 2001 Oct;12(10):3114-25.

Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci U S A. 1994 May 24;91(11):5022-6.

Peirson SN, Butler JN, Foster RG. Experimental validation of novel and conventional approaches to quantitative real-time PCR data analysis. Nucleic Acids Res. 2003 Jul 15;31(14):e73.

Periz J, Gill AC, Hunt L, Brown P, Tomley FM. The microneme proteins EtMIC4 and EtMIC5 of *Eimeria tenella* form a novel, ultra-high molecular mass protein complex that binds target host cells. J Biol Chem. 2007 Jun 8;282(23):16891-8.

Perkins ME, Riojas YA, Wu TW, Le Blancq SM. CpABC, a *Cryptosporidium parvum* ATP-binding cassette protein at the host-parasite boundary in intracellular stages. Proc Natl Acad Sci U S A. 1999 May 11;96(10):5734-9.

Péroval M, Péry P, Labbé M. The heat shock protein 90 of *Eimeria tenella* is essential for invasion of host cell and schizont growth. Int J Parasitol. 2006 Sep;36(10-11):1205-15.

Pezzella-D'Alessandro N, Le Moal H, Bonhomme A, Valere A, Klein C, Gomez-Marin J, Pinon JM. Calmodulin distribution and the actomyosin cytoskeleton in *Toxoplasma gondii*. J Histochem Cytochem. 2001 Apr;49(4):445-54.

Pfaffl MW. Relative quantification. In: Dorak MT, editor. Real-time PCR. New York (USA): Taylor & Francis Group; 2006. p. 63-82.

Pinney JW, Shirley MW, McConkey GA, Westhead DR. metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. Nucleic Acids Res. 2005;33(4):1399-409.

Pleasance ED, Marra MA, Jones SJ. Assessment of SAGE in transcript identification. Genome Res. 2003 Jun;13(6A):1203-15.

Ponchel F. Real-time PCR using SYBR® Green. In: Dorak MT, editor. Real-time PCR. New York (USA): Taylor & Francis Group; 2006. p. 139-54.

Ponts N, Yang J, Chung DW, Prudhomme J, Girke T, Horrocks P, Le Roch KG. Deciphering the ubiquitin-mediated pathway in apicomplexan parasites: a potential strategy to interfere with parasite virulence. PLoS ONE. 2008;3(6):e2386.

Poole RL, Barker GL, Werner K, Biggi GF, Coghill J, Gibbings JG, Berry S, Dunwell JM, Edwards KJ. Analysis of wheat SAGE tags reveals evidence for widespread antisense transcription. BMC Genomics. 2008;9:475.

Poroyko V, Hejlek LG, Spollen WG, Springer GK, Nguyen HT, Sharp RE, Bohnert HJ. The maize root transcriptome by serial analysis of gene expression. Plant Physiol. 2005 Jul;138(3):1700-10.

Que X, Engel JC, Ferguson D, Wunderlich A, Tomavo S, Reed SL. Cathepsin Cs are key for the intracellular survival of the protozoan parasite, *Toxoplasma gondii*. J Biol Chem. 2007 Feb 16;282(7):4994-5003.

Quére R, Manchon L, Lejeune M, Clément O, Pierrat F, Bonafoux B, Commes T, Piquemal D, Marti J. Mining SAGE data allows large-scale, sensitive screening of antisense transcript expression. Nucleic Acids Res. 2004;32(20):e163.

Radke JR, Behnke MS, Mackey AJ, Radke JB, Roos DS, White MW. The transcriptome of *Toxoplasma gondii*. BMC Biol. 2005;3:26.

Ramachandran V, Dorin D, Khiong CW, Kara UA, Doerig C. A *Plasmodium falciparum* homologue of the Ran binding protein 1, a protein involved in nucleocytoplasmic transport. Mol Biochem Parasitol. 2002 Aug 7;123(1):67-71.

Reinartz J, Bruyns E, Lin JZ, Burcham T, Brenner S, Bowen B, Kramer M, Woychik R. Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. Brief Funct Genomic Proteomic. 2002 Feb;1(1):95-104.

Rivals E, Boureux A, Lejeune M, Ottones F, Perez OP, Tarhio J, Pierrat F, Ruffle F, Commes T, Marti J. Transcriptome annotation using tandem SAGE tags. Nucleic Acids Res. 2007;35(17):e108.

Robinson SJ, Parkin IA. Differential SAGE analysis in *Arabidopsis* uncovers increased transcriptome complexity in response to low temperature. BMC Genomics. 2008;9:434.

Romano CM. Caracterização molecular e análise comparativa de genomas mitocondriais de *Eimeria* spp. de galinha doméstica [Dissertação (Mestrado em Ciências - Biologia da Relação Patógeno-Hospedeiro)]. São Paulo: Instituto de Ciências Biomédicas da Universidade de São Paulo; 2004.

Roos DS. Genetics. Themes and variations in apicomplexan parasite biology. Science. 2005 Jul 1;309(5731):72-3.

Rosinski-Chupin I, Chertemps T, Boisson B, Perrot S, Bischoff E, Briolay J, Couble P, Menard R, Brey P, Baldacci P. Serial Analysis of Gene Expression in *Plasmodium berghei* salivary gland sporozoites. BMC Genomics. 2007;8:466.

Ryan R, Shirley M, Tomley F. Mapping and expression of microneme genes in *Eimeria tenella*. Int J Parasitol. 2000 Dec;30(14):1493-9.

Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE. Using the transcriptome to annotate the genome. Nat Biotechnol. 2002 May;20(5):508-12.

Sambrook J, Russell DW. Molecular Cloning - A Laboratory Manual. 3rd ed. New York: Cold Spring Harbor Laboratory Press; 2001.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463-7.

Santos JM, Lebrun M, Daher W, Soldati D, Dubremetz JF. Apicomplexan cytoskeleton and motors: key regulators in morphogenesis, cell division, transport and motility. Int J Parasitol. 2009 Jan;39(2):153-62.

Sauvage V, Millot JM, Aubert D, Visneux V, Marle-Plistat M, Pinon JM, Villena I. Identification and expression analysis of ABC protein-encoding genes in *Toxoplasma gondii*. *Toxoplasma gondii* ATP-binding cassette superfamily. Mol Biochem Parasitol. 2006 Jun;147(2):177-92.

Schaap D, Arts G, van Poppel NF, Vermeulen AN. De novo ribosome biosynthesis is transcriptionally regulated in *Eimeria tenella*, dependent on its life cycle stage. Mol Biochem Parasitol. 2005 Feb;139(2):239-48.

Schatten H, Ris H. Unconventional specimen preparation techniques using high resolution low voltage field emission scanning electron microscopy to study cell motility, host cell invasion, and internal cell structures in *Toxoplasma gondii*. Microsc Microanal. 2002 Apr;8(2):94-103.

Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995 Oct 20;270(5235):467-70.

Schmatz DM. The mannitol cycle in *Eimeria*. Parasitology. 1997;114 Suppl:S81-9.

Schnitzler BE, Thebo PL, Mattsson JG, Tomley FM, Shirley MW. Development of a diagnostic PCR assay for the detection and discrimination of four pathogenic .*Eimeria* species of the chicken. Avian Pathol. 1998;27(5):490-7.

Schnitzler BE, Thebo PL, Tomley FM, Uggla A, Shirley MW. PCR identification of chicken *Eimeria*: a simplified read-out. Avian Pathol. 1999 Feb;28(1):89-93.

Shi T, Yan W, Ren H, Liu X, Suo X. Dynamic development of parasitophorous vacuole of *Eimeria tenella* transfected with the yellow fluorescent protein gene fused to different signal sequences from apicomplexan parasites. Parasitol Res. 2009 Jan;104(2):315-20.

Shin H, Hirst M, Bainbridge MN, Magrini V, Mardis E, Moerman DG, Marra MA, Baillie DL, Jones SJ. Transcriptome analysis for *Caenorhabditis elegans* based on novel expressed sequence tags. BMC Biol. 2008;6:30.

Shipley GL. An introduction to real-time PCR. In: Dorak MT, editor. Real-time PCR. New York (USA): Taylor & Francis Group; 2006. p. 63-82.

Shirafuji N, Takahashi S, Matsuda S, Asano S. Mitochondrial antisense RNA for cytochrome C oxidase (MARCO) can induce morphologic changes and cell death in human hematopoietic cell lines. Blood. 1997 Dec 1;90(11):4567-77.

Shirley MW. The genome of *Eimeria tenella*: further studies on its molecular organisation. Parasitol Res. 1994;80(5):366-73.

Shirley MW. *Eimeria* species and strains of chicken. In: Eckert J, Braun R, Shirley MW, Coudert P, editors. COST 89/820 Biotechnology: Guidelines on techniques in coccidiosis research. Brussels, Luxembourg European Comission; 1995. p. 1-24.

Shirley MW. The genome of *Eimeria* spp., with special reference to *Eimeria tenella*--a coccidium from the chicken. Int J Parasitol. 2000 Apr 10;30(4):485-93.

Shirley MW, Bumstead N. Intra-specific variation within *Eimeria tenella* detected by the random amplification of polymorphic DNA. Parasitol Res. 1994;80(4):346-51.

Shirley MW, Harvey DA. *Eimeria tenella*: genetic recombination of markers for precocious development and arprinocid resistance. Appl Parasitol. 1996 Dec;37(4):293-9.

Shirley MW, Harvey DA. A genetic linkage map of the apicomplexan protozoan parasite *Eimeria tenella*. Genome Res. 2000 Oct;10(10):1587-93.

Shirley MW, Smith AL, Tomley FM. The biology of avian *Eimeria* with an emphasis on their control by vaccination. Adv Parasitol. 2005;60:285-330.

Shirley MW, Smith AL, Blake DP. Challenges in the successful control of the avian coccidia. Vaccine. 2007 Jul 26;25(30):5540-7.

Shirley MW, Blake D, White SE, Sheriff R, Smith AL. Integrating genetics and genomics to identify new leads for the control of *Eimeria* spp. Parasitology. 2004a;128 Suppl 1:S33-42.

Shirley MW, Ivens A, Gruber A, Madeira AM, Wan KL, Dear PH, Tomley FM. The *Eimeria* genome projects: a sequence of events. Trends Parasitol. 2004b May;20(5):199-201.

Sibley LD. Intracellular parasite invasion strategies. Science. 2004 Apr 9;304(5668):248-53.

Siddiqui AS, Delaney AD, Schnerch A, Griffith OL, Jones SJ, Marra MA. Sequence biases in large scale gene expression profiling data. Nucleic Acids Res. 2006;34(12):e83.

Silva AP, De Souza JE, Galante PA, Riggins GJ, De Souza SJ, Camargo AA. The impact of SNPs on the interpretation of SAGE and MPSS experimental data. Nucleic Acids Res. 2004;32(20):6104-10.

Silveira NJ, Varuzza L, Machado-Lima A, Lauretto MS, Pinheiro DG, Rodrigues RV, Severino P, Nobrega FG, Silva WA Jr, de BPCA, Tajara EH. Searching for molecular markers in head and neck squamous cell carcinomas (HNSCC) by statistical and bioinformatic analysis of larynx-derived SAGE libraries. BMC Med Genomics. 2008;1:56.

Skuce PJ, Yaga R, Lainson FA, Knox DP. An evaluation of serial analysis of gene expression (SAGE) in the parasitic nematode, *Haemonchus contortus*. Parasitology. 2005 May;130(Pt 5):553-9.

Smith AL, Hesketh P, Archer A, Shirley MW. Antigenic diversity in *Eimeria maxima* and the influence of host genetics and immunization schedule on cross-protective immunity. Infect Immun. 2002 May;70(5):2472-9.

Sobreira TJ, Gruber A. Sequence-specific reconstruction from fragmentary databases using seed sequences: implementation and validation on SAGE, proteome and generic sequencing data. Bioinformatics. 2008 Jun 9.

Song S, Qu H, Chen C, Hu S, Yu J. Differential gene expression in an elite hybrid rice cultivar (*Oryza sativa*, L) and its parental lines based on SAGE data. BMC Plant Biol. 2007;7:49.

Stein L. Genome annotation: from sequence to biology. Nat Rev Genet. 2001 Jul;2(7):493-503.

Sterky F, Lundeberg J. Sequence analysis of genes and genomes. J Biotechnol. 2000 Jan 7;76(1):1-31.

Stern MD, Anisimov SV, Boheler KR. Can transcriptome size be estimated from SAGE catalogs? Bioinformatics. 2003 Mar 1;19(4):443-8.

Striepen B, White MW, Li C, Guerini MN, Malik SB, Logsdon JM, Jr., Liu C, Abrahamsen MS. Genetic complementation in apicomplexan parasites. Proc Natl Acad Sci U S A. 2002 Apr 30;99(9):6304-9.

Sullivan WJ Jr, Hakimi MA. Histone mediated gene activation in *Toxoplasma gondii*. Mol Biochem Parasitol. 2006 Aug;148(2):109-16.

Sun M, Zhou G, Lee S, Chen J, Shi RZ, Wang SM. SAGE is far more sensitive than EST for detecting low-abundance transcripts. BMC Genomics. 2004 Jan 5;5(1):1.

Taft AS, Vermeire JJ, Bernier J, Birkeland SR, Cipriano MJ, Papa AR, McArthur AG, Yoshino TP. Transcriptome analysis of *Schistosoma mansoni* larval development using serial analysis of gene expression (SAGE). Parasitology. 2009 Apr;136(5):469-85.

Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997 Oct 24;278(5338):631-7.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000 Jan 1;28(1):33-6.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003 Sep 11;4:41.

Tenter AM, Barta JR, Beveridge I, Duszynski DW, Mehlhorn H, Morrison DA, Thompson RC, Conrad PA. The conceptual basis for a new classification of the coccidia. Int J Parasitol. 2002 May;32(5):595-616.

Teodorovic S, Walls CD, Elmendorf HG. Bidirectional transcription is an inherent feature of *Giardia lamblia* promoters and contributes to an abundance of sterile antisense transcripts throughout the genome. Nucleic Acids Res. 2007;35(8):2544-53.

Tomley FM, Soldati DS. Mix and match modules: structure and function of microneme proteins in apicomplexan parasites. Trends Parasitol. 2001 Feb;17(2):81-8.

Tomley FM, Billington KJ, Bumstead JM, Clark JD, Monaghan P. EtMIC4: a microneme protein from *Eimeria tenella* that contains tandem arrays of epidermal growth factor-like repeats and thrombospondin type-I repeats. Int J Parasitol. 2001 Oct;31(12):1303-10.

Tomley F. Apicomplexan biology in the post-genomic era: perspectives from the European COST Action 857. Int J Parasitol. 2009 Jan;39(2):133-4.

Tullo A, Tanzariello F, D'Erchia AM, Nardelli M, Papeo PA, Sbisa E, Saccone C. Transcription of rat mitochondrial NADH-dehydrogenase subunits. Presence of antisense and precursor RNA species. FEBS Lett. 1994 Oct 31;354(1):30-6.

Tuteja R, Tuteja N. Serial Analysis of Gene Expression: Applications in Human Studies. J Biomed Biotechnol. 2004;2004(2):113-20.

Tyzzer EE. Coccidiosis in gallinaceous birds. Am J Hyg. 1929(10):269-83.

Unneberg P, Wennborg A, Larsson M. Transcript identification by analysis of short sequence tags--influence of tag length, restriction site and transcript database. Nucleic Acids Res. 2003 Apr 15;31(8):2217-26.

Valasek MA, Repa JJ. The power of real-time PCR. Adv Physiol Educ. 2005 Sep;29(3):151-9.

van Ruissen F, Schaaf GJ, Kool M, Bass F, Ruijter JM. Scaling of Gene Expression Data Allowing the Comparison of Different Gene Expression Platforms. In: Nielsen KL, editor. Serial Analysis of Gene Expression (SAGE). Totowa, New Jersey (USA): Humana Press; 2008.

van Ruissen F, Ruijter JM, Schaaf GJ, Asgharnegad L, Zwijnenburg DA, Kool M, Baas F. Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. BMC Genomics. 2005;6:91.

Varuzza L. Métodos Estatísticos para a análise de bibliotecas digitais de expressão gênica [Doutorado (Programa Interunidades de Bioinformática)]. São Paulo: Universidade de São Paulo; 2008.

Velculescu VE. Essay: Amersham Pharmacia Biotech & Science prize. Tantalizing transcriptomes--SAGE and its use in global gene expression analysis. Science. 1999 Nov 19;286(5444):1491-2.

Velculescu VE, Vogelstein B, Kinzler KW. Analysing uncharted transcriptomes with SAGE. Trends Genet. 2000 Oct;16(10):423-5.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science. 1995 Oct 20;270(5235):484-7.

Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Jr., Hieter P, Vogelstein B, Kinzler KW. Characterization of the yeast transcriptome. Cell. 1997 Jan 24;88(2):243-51.

Wahl MB, Heinzmann U, Imai K. LongSAGE analysis significantly improves genome annotation: identifications of novel genes and alternative transcripts in the mouse. Bioinformatics. 2005a Apr 15;21(8):1393-400.

Wahl MB, Heinzmann U, Imai K. LongSAGE analysis revealed the presence of a large number of novel antisense genes in the mouse genome. Bioinformatics. 2005b Apr 15;21(8):1389-92.

Wahl MB, Caldwell RB, Kierzek AM, Arakawa H, Eyras E, Hubner N, Jung C, Soeldenwagner M, Cervelli M, Wang YD, Liebscher V, Buerstedde JM. Evaluation of the chicken transcriptome by SAGE of B cells and the DT40 cell line. BMC Genomics. 2004 Dec 21;5(1):98.

Waller RF, McFadden GI. The apicoplast: a review of the derived plastid of apicomplexan parasites. Curr Issues Mol Biol. 2005 Jan;7(1):57-79.

Wan KL, Chong SP, Ng ST, Shirley MW, Tomley FM, Jangi MS. A survey of genes in *Eimeria tenella* merozoites by EST sequencing. Int J Parasitol. 1999 Dec;29(12):1885-92.

Wang SM. Understanding SAGE data. Trends Genet. 2007 Jan;23(1):42-50.

Wang X, Zhao Y, Wong K, Ehlers P, Kohara Y, Jones SJ, Marra MA, Holt RA, Moerman DG, Hansen D. Identification of genes expressed in the hermaphrodite germ line of C. elegans using SAGE. BMC Genomics. 2009;10:213.

Welsh J, Chada K, Dalal SS, Cheng R, Ralph D, McClelland M. Arbitrarily primed PCR fingerprinting of RNA. Nucleic Acids Res. 1992 Oct 11;20(19):4965-70.

Wiersma HI, Galuska SE, Tomley FM, Sibley LD, Liberator PA, Donald RG. A role for coccidian cGMP-dependent protein kinase in motility and invasion. Int J Parasitol. 2004 Mar 9;34(3):369-80.

Williams DL, Sayed AA, Bernier J, Birkeland SR, Cipriano MJ, Papa AR, McArthur AG, Taft A, Vermeire JJ, Yoshino TP. Profiling *Schistosoma mansoni* development using serial analysis of gene expression (SAGE). Exp Parasitol. 2007 Nov;117(3):246-58.

Williams RB. Epidemiological aspects of the use of live anticoccidial vaccines for chickens. Int J Parasitol. 1998 Jul;28(7):1089-98.

Wold B, Myers RM. Sequence census methods for functional genomics. Nat Methods. 2008 Jan;5(1):19-21.

Wu X. Workflow-based systematic design of high throughput genome annotation [Ph. D. Thesis]. London: Imperial College London Department of Computing; 2008.

Xia D, Sanderson SJ, Jones AR, Prieto JH, Yates JR, Bromley E, Tomley FM, Lal K, Sinden RE, Brunk BP, Roos DS, Wastling JM. The proteome of *Toxoplasma gondii*: integration with the genome provides novel insights into gene expression and annotation. Genome Biol. 2008;9(7):R116.

Xu JH, Qin ZH, Liao YS, Xie MQ, Li AX, Cai JP. Characterization and expression of an actin-depolymerizing factor from *Eimeria tenella*. Parasitol Res. 2008 Jul;103(2):263-70.

Yamamoto M, Wakatsuki T, Hada A, Ryo A. Use of serial analysis of gene expression (SAGE) technology. J Immunol Methods. 2001 Apr;250(1-2):45-66.

Yan W, Liu X, Shi T, Hao L, Tomley FM, Suo X. Stable transfection of *Eimeria tenella*: constitutive expression of the YFP-YFP molecule throughout the life cycle. Int J Parasitol. 2009 Jan;39(1):109-17.

Zou J, Liu X, Shi T, Huang X, Wang H, Hao L, Yin G, Suo X. Transfection of *Eimeria* and *Toxoplasma* using heterologous regulatory sequences. Int J Parasitol. 2009 Sep;39(11):1189-93.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo