Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Pós-Graduação em Biologia Computacional e Sistemas

DIOGO ANTONIO TSCHOEKE

Genômica comparativa de protozoários

Tese apresentada ao Instituto Oswaldo Cruz
como parte dos requisitos para obtenção do título
de Doutor em Biologia Computacional e Sistemas

Orientador: Dr. Alberto Mártin Rivera Dávila

RIO DE JANEIRO

2013

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Pós-Graduação em Biologia Computacional e Sistemas

DIOGO ANTONIO TSCHOEKE

Genômica comparativa de protozoários

Orientador: **Dr. Alberto Mártin Rivera Dávila**

Aprovada em: 17/12/2013

EXAMINADORES:
Dr^{a}.– Claudia M. D'Ávila Levy
Dr^{a}. – Maria Luiza Machado Campos
Dr. – Christian M. Probst
Suplente: Dr^{a}. Maria Cláudia Cavalcanti
Suplente: Dr. Leonardo B. Koerich

Rio de Janeiro, 17 de Dezembro de 2013

Dedicatória

Dedico este trabalho a minha família,
que sempre me apoiou em todos os momentos
e a todos aqueles que me ajudaram
no decorrer do doutorado

Agradecimentos

Gostaria de agradecer a minha “namorida” Amanda pela ajuda, pela força e por me dar o suporte que sempre precisei durante estes quase quatro anos de doutorado. Além de ter tido o trabalho de revisar esta tese toda! Amo-te! ATF! ☺

A minha família Nivio, Susan, Jeferson e Larissa pelo apoio. Mesmo longe sei o quanto vocês foram presentes nesta jornada! Amo vocês!

Quero agradecer a minha “família carioca” que me adotou nestes quase sete anos de Rio! Obrigado pelo carinho e pelos ótimos finais de semana!

Meu muito obrigado ao Dr. Alberto Dávila por me orientar nesta jornada, que eu sei o quanto foi estressante! Mas conseguimos chegar ao seu fim!

Ao pessoal do LBCS (e para aquelas que não são mais do lab) por toda a amizade e carinho, já são 6 anos de lab! Rodrigo, Joana, Luisa, Nelson, Gisele, Aline, Dri, Bernardo, Monete, Leandro, Rachel, Elisa, Antônio, Renata, André, Rafael, Fabio Mota, Lana, Mayla, Lennon, Darueck, Priscila, Janaína e Fillipe. Muito obrigado pessoal, as horas de conversa e descontração no lab sempre foram muito boas!

Aos colaboradores que ajudaram com o andamento desta tese: Serra, Marta, Maria Luiza, Christian, Antônio, Monete, Leandro, Floriano, André. Obrigado, sem a ajuda de vocês a realização deste trabalho ficaria comprometida.

Dra. Cláudia Levy, obrigado por aceitar o convite para revisar esta tese e para participar da banca. Obrigado também à Dra Maria Luiza e a Dra Maria Claudia (Yoko) por aceitarem prontamente o convite para integrarem a banca avaliadora. Quero agradecer também ao Dr. Christian e Dr. Leonardo por aceitaram o convite e por avaliarem esta tese! Obrigado!

Um muito obrigado a todos os professores do PVS por compartilharem muitos sábados do ano trabalhando! Andreza, Renato, Jean,Luiz, Rafael, Rodrigo, Cadu, Lucas e Osmar! Obrigado pessoal!

Obrigado a todo o pessoal da secretaria da Fiocruz, principalmente a nossa secretária Alessandra, pela ajuda durante o doutorado!

Obrigado a todos que me ajudaram nesta longa caminhada e peço desculpas se me esqueci de algum nome!

Epígrafe

“Tenho a impressão de ter sido uma
criança brincando à beira-mar,
divertindo-me em descobrir uma
pedrinha mais lisa ou uma
concha mais bonita que as outras,
enquanto o imenso oceano da
verdade continua misterioso
diante de meus olhos”
Isaac Newton

“É preciso correr muito para
ficar no mesmo lugar.
Se você quer chegar a outro
lugar, corra duas vezes mais”
Lewis Carrol

“A ciência nunca resolve um problema
sem criar pelo menos outros dez”
George Bernard Shaw

Resumo

Os protozoários são definidos como organismos eucariotos unicelulares, e apresentam grande diversidade e variedade. Cerca de 200 mil espécies são descritas e quase 10.000 são parasitas. As espécies patogênicas causam doenças como a malária, doença do sono, doença de Chagas, leishmaniose, amebíase e giardíase. Portanto, estudos comparativos entre os protozoários são importantes porque estes podem mostrar semelhanças e diferenças entre essas espécies. A identificação de ortólogos é importante para a categorização funcional de genomas, porque ortólogos tipicamente ocupam o mesmo nicho funcional nos diferentes organismos, enquanto a identificação de parálogos é importante porque eles são submetidos a uma diversificação funcional via duplicação, através dos processos de neofuncionalização e subfuncionalização. A fim de realizar uma análise comparativa de 22 protozoários, 204.624 proteínas não redundantes de *Plasmodium*, *Entamoeba*, *Trypanosoma*, *Leishmania*, *Giardia*, *Theileria*, *Toxoplasma*, *Trichomonas* e *Cryptosporidium*, foram submetidos ao programa OrthoMCL, resultando em 26.101 grupos homólogos. Entre eles, 21.119 grupos são ortólogos, incluindo 7.679 co-ortólogos (grupos que contêm parálogos recentes), e 4982 são parálogos internos. Entre os ortólogos, 348 são compartilhados por todas as 22 espécies e representam o núcleo proteômico de Protozoa. Com este núcleo realizamos uma análise filogenômica, usando os 348 ortólogos concatenados, resultando em uma supermatriz de 328.228 posições, que geraram uma árvore de espécies para os 22 protozoários. Quando inferimos os diferentes Núcleos Proteômicos, Kinetoplastida tem 5.000 grupos ortólogos e 67,92 % (3396/5000) são Kinetoplastida específicos, além disso, 46,29% (1592/3396) destes ortólogos são anotados como "hipotéticos". O núcleo proteômico de Apicomplexa tem 986 grupos ortólogos e 27,82% (224/986) são específicos, enquanto que 40,63% (92 /224) destes são classificados como hipotéticos. O núcleo proteômico de *Entamoeba* tem 5.915 grupos ortólogos e 75,08% (4441/5915) destes grupos são específicos, sendo que 65,41% (2905/4441) são anotados como hipotéticos. Analisando os parálogos, *Trichomonas vaginalis* foi a espécie que apresentou o maior número de grupos parálogos internos, 2933, e também mostrou 948 co-ortólogos totalizando 3.881 parálogos. Um aprofundamento da análise na ordem Kinetoplastida mostrou que *Trypanosoma cruzi* apresenta o número mais elevado de duplicações, totalizando 5.777 parálogos, sendo 4963 co-ortólogos e 814 parálogos internos. Os resultados da montagem e análise de *L. amazonensis* resultaram 29.670.588 bases e 8802 CDS identificadas. A análise comparativa do gênero *Leishmania* mostrou que as seis espécies estudadas compartilham 7016 ortólogos, enquanto *L. amazonensis* e *L. mexicana* têm o maior número de ortólogos-específicos e *L. braziliensis* o maior número de paralogos internos. A análise filogenômica mostrou a posição taxonômica esperada de *L. amazonensis* e juntamente com *L. mexicana* formando o "complexo Mexicana", além da separação esperada do subgênero *Leishmania*. Encontramos potenciais proteínas análogas entre *L. amazonensis* e *Homo sapiens* e dentro do genoma de *L. amazonensis,* denominados análogos intragenômicos. Finalmente, a mineração por genes de RNAi mostrou que *L. amazonensis* , provavelmente não apresenta esta via funcional.

Abstract

Protozoa are defined as single celled eukaryotic organisms showing an extremely diversity and variety. Approximately 200,000 species are described and nearly 10,000 are parasitic. The pathogenic species cause diseases such as malaria, sleeping sickness, Chagas disease, leishmaniasis, amoebiasis and giardiasis. Therefore, comparative studies among Protozoa are important because they may identify similarities and differences in these species. Orthologs identification is central to functional characterization of genomes because orthologs typically occupy the same functional niche in different organisms, while paralogs identification is important because they undergo a functional diversification by duplication, via the processes of neofunctionalization and subfunctionalization. In order to perform comparative protozoa analysis, 204,624 non-redundant proteins from *Plasmodium*, *Entamoeba*, *Trypanosoma*, *Leishmania*, *Giardia*, *Theileria*, *Toxoplasma*, *Trichomonas* and *Cryptosporidium*, totalizing 22 species, were submitted to OrthoMCL resulting in 26,101 homologs groups. Among them, 21,119 groups are orthologs including 7,679 co-orthologs (groups that contain recent paralogs) and 4,982 are inparalogs. Among the orthologs, 348 are shared by all 22 species, representing the Protozoa core proteome, with this core we performed a phylogenomic analysis with the 348 concatenated orthologs, resulting in a global supermatrix of 328,228 positions that generate a species tree for the 22 protozoa. When we inferred Core Proteome, the Kinetoplastida core has 5,000 orthologous groups and 67.92% (3,396/5000) are Kinetoplastida Specific, besides 46.29% (1,592/3,396) of these orthologs are annotated as “hypothetical”. Apicomplexa Core Proteome has 986 orthologous groups and 27.82% (224/986) are Apicomplexa Specific whereas 40.63% (92/224) were classified as hypothetical proteins. Entamoeba Core Proteome has 5,915 orthologous groups and 75.08% (4,441/5,915) of these groups are Entamoeba Specific and 65.41% (2,905/4441) were annotated as hypothetical. Analyzing the paralogs, *Trichomonas vaginalis* was the species that presented the highest number of inparalogs groups, 2,933 and also showed 948 co-orthologs totalizing 3881 paralogs. A deep look into the Kinetoplastida order showed that *Trypanosoma cruzi* has the highest duplication number, totalizing 5777 paralogs, 4963 co-orthologs and 814 inparalogs groups. The *L. amazonensis* analysis resulted in 29,670,588 bases assembled, and 8802 CDS identified. Comparative analysis into the *Leishmania* genus showed that these 6 species share 7016 ortologs, whilst *L. amazonensis* and *L. mexicana* has the biggest number of specific orthologs and *L. braziliensis* biggest number of inparalogs. Phylogenomic analysis showed the expected *L. amazonensis* taxonomic position together with *L. mexicana* forming the “Mexicana complex” and the New and Old *Leishmania* (L.). spp. separation. Potential analagous proteins were found between *L. amazonensis* and *Homo sapiens*, and also into the *L. amazonensis* proteome. Finally, RNAi analysis showed that *L. amazonensis*, probably, do not have functional RNAi pathway

Abreviaturas

A+T - porcentagem de bases Adenina e timina em uma porção de DNA

AIDS – do inglês *Acquired Immunodeficiency Syndrome*

BLAST – do Inglês *Basic Local Alignment Search Tool*

CDD – do inglês *Conserved Domain Database*

CDS – do inglês *Codifying Sequence*

COG – do Inglês *Cluster of Orthologous Groups*

DNA - Ácido desoxirribonucléico

DTN – Doença Tropical Negligenciada

EST – do inglês *Espressed Sequence Tag*

E-value – valor de probabilidade de um resultado do BLAST ter sido obtido ao acaso, do inglês *Expectation value*

GC – porcentagem de bases Guanina e Citosina em uma porção de DNA

KEGG – do inglês *Kyoto Encyclopedia of Genes and Genomes*

KOG – do inglês *Eukaryotic Orthologous Groups*

LC – Leishmaniose Cutânea

LCD - Leishmaniose Cutânea Difusa

LGT – do inglês *Lateral Gene Transfer*

LTA – Leishmaniose Tegumentar Americana

LV – Leishmaniose Visceral

MASP – do inglês *Mucin associated surface protein*

MCL – do inglês *Markov Cluster algorithm*

NGS – do inglês *Next Generation Sequence*

pb – pares de bases

PDB – do inglês *Protein Data Bank*

Perl – do Inglês *Practical Extraction and Report Language*

RNA – Ácido ribonucléico

SNP – do inglês *Single Nucleotide Polimorphism*

STINGRAY - do inglês *System for Integrated Genomic Resources and Analysis*

TDR – do inglês *Special Programme for Research and Training in Tropical Diseases*

tRNA – RNA transportador

VSG – do inglês *Variant Surface Glicoprotein*

WHO - do inglês *World Health Organization*

Lista de ilustrações

Lista de Figuras

Figura 1.1: Ilustração composta por várias fotos/figuras de protozoários	3
Figura 1.2: Exemplos de estruturas homólogas e análogas.	9
Figura 1.3: Exemplo das relações de homologia	11
Figura 1.4: Triângulo de ortólogos formados por melhores resultados recíprocos	18
Figura 1.5: Fluxograma do algoritmo OrthoMCL	19

1 - INTRODUÇÃO 1
1.1 OS PROTOZOÁRIOS E SEUS IMPACTOS 1
1.2 CARACTERIZAÇÃO DAS ESPÉCIES 4
1.3 GENÔMICA COMPARATIVA, HOMOLOGIA, GENES HOMÓLOGOS E GENES ÓRFÃOS. 8
1.4 GENÔMICA COMPARATIVA DE APICOMPLEXA 12
1.5 A GENÔMICA COMPARATIVA DE TRIPANOSSOMATÍDEOS 14
1.6 A FILOGENÔMICA 16
1.7 GRUPOS DE ORTÓLOGOS E OS BANCOS COG E KOG 16
1.8 O PROGRAMA ORTHOMCL 18
1.9 O PARASITA *LEISHMANIA* (*LEISHMANIA*) *AMAZONENSIS* 20
1.10 ORGANIZAÇÃO DA TESE 22
2 - OBJETIVOS 23
2.1 - OBJETIVO GERAL 23
2.2 - OBJETIVOS ESPECÍFICOS 23
3 - ARTIGO 1: GENÔMICA COMPARATIVA DE PROTOZOÁRIOS: IDENTIFICAÇÃO DE GENES GRUPOS E ESPÉCIE-ESPECÍFICOS E SUA CATEGORIZAÇÃO FUNCIONAL. 24
4 – ARTIGO 2: SOBRE AS PARTICULARIDADES DOS PROTOZOÁRIOS: INFERINDO AS EXPANSÕES DE FAMÍLIAS E PROTEÍNAS ÓRFÃS 83
5 - ARTIGO 3: “GENÔMICA COMPARATIVA DO PARASITA *LEISHMANIA AMAZONENSIS*” 111
6- DISCUSSÃO 177
7 - CONCLUSÕES 187
8 - REFERÊNCIAS 189
9: ANEXOS 197
9.1- ARTIGO 4: PROTOZOADB 2.0: UMA FERRAMENTA PARA A EXTRAÇÃO DE INFORMAÇÕES A PARTIR DO GENOMA DE 22 ESPÉCIES DE PROTOZOÁRIOS PATOGÊNICOS. 197
9.2- ARTIGO 5: “STINGRAY: SISTEMA INTEGRADO PARA ANALISES DE RECURSOS GENÔMICOS” 216

1 - Introdução

1.1 Os protozoários e seus impactos

A palavra Protozoa foi criada por Goldfuss em 1818 e significa "Primeiros Animais" (Imam, 2009). Siebold, assim como Goldfuss, considerou os protozoários como invertebrados primitivos e os dividiu em duas classes: Infusórios e Rhizopoda (atualmente termos equivalentes a ciliados e amebas). Em 1858, Owen estabeleceu o reino Protozoa para os organismos unicelulares mais primitivos (Cavalier-Smith, 2009), colocando organismos unicelulares como bactérias, amebas e diatomáceas em um reino separado. Hogg, em 1860, introduziu o termo “Protoctista” para incluir estas formas que não possuíam afinidade com plantas animais ou ainda não demonstravam afinidades claras entre si. Em 1876, Haeckel utiliza o termo "Protista" para tentar traduzir em um sistema de classificação a divisão do reino animal em: organismos unicelulares (protozoários) e organismos multicelulares (Metazoa), dividindo assim a árvore da vida em três reinos: (i) Animalia; (ii) Plantae; e (iii) Protista, onde cada um deles surgiu monofileticamente do lodo primordial (Cavalier-Smith, 2010).

Nos dias de hoje, quando falamos sobre os protozoários, existe uma gama de definições e classificações. Cavalier-Smith em 1993, por exemplo, utilizou a definição: "eucariotos unicelulares fagotróficos com mitocôndrias", que pode ser uma definição simples incluindo a grande maioria dos protozoários, inclusive os organismos do filo Chromista e excluindo poucas espécies. Entretanto, tal definição não seria suficientemente precisa para definir os limites exatos deste reino. Quando adotamos um sistema de cinco reinos para os eucariotos, por exemplo, os protozoários são considerados os mais basais enquanto os outros quatro reinos são mais derivados: Animalia, Fungi (com ancestralidade heterotrófica), Plantae e Chromista (com ancestralidade fototrófica) (Cavalier-Smith, 2010). Outra abordagem mais conservadora é tratar Protozoa como um sub-reino, porém não especificando se pertence à Animalia ou Protista. Já a abordagem mais radical seria abandonar completamente Protozoa como um táxon e dividir seus filos em um amplo reino como Protista, Protoctista ou mesmo Phytobiota (=Plantae), ou alternativamente, dividí-lo em vários reinos menores (Cavalier-Smith, 1993). Todavia, de maneira genérica, os protozoários, atualmente, são definidos como: organismos unicelulares eucariontes. Entretanto, o fato deles compartilharem características dos procariotos e eucariotos multicelulares faz com que sua classificação seja um pouco complexa, como já mencionado. Uma definição filogenética do reino Protozoa poderia ser a seguinte: “eucariotos, exceto aqueles que não possuem primitivamente mitocôndrias, peroxissomas (Archezoa) e os caracteres derivados compartilhados que definem os quatro maiores reinos derivados: Animalia, Fungi, Plantae e Chromista” (Cavalier-Smith, 1993). Nos dias atuais, existem descritas mais de 200 mil espécies de protozoários, das quais cerca de 10 mil são parasitas de invertebrados e de quase todas as espécies de vertebrados (Imam, 2009). Esta diversidade de formas (Fig. 1.1) varia desde as não patogênicas até aquelas causadoras de importantes doenças em países tropicais, tais como: malária, doença do sono, doença de Chagas, leishmaniose, amebíase e giardíase, que juntas ameaçam mais de um quarto da população mundial (Imam, 2009).

Várias enfermidades, também associadas a protozoários, apresentam impacto relevante na saúde humana e na agroindústria, como por exemplo, no caso da Giardíase, Toxoplasmose, Babesiose, Teileriose e Criptosporidiose (Pain et al., 2005; Brayton et al., 2007; Carlton et al., 2008; Thompson, 2009; Widmer et al., 2009; Elmore et al., 2010). Um conjunto de doenças importantes para a agroindústria, causadas por protozoários, são os Apicomplexa do agrupamento dos Piroplasmídeos como, por exemplo, a Babesiose e Teileriose. A Babesiose é uma doença hemoprotozoaria causada por *Babesia bovis*, transmitida através de carrapatos vetores, enzoótica em ruminantes na maioria das áreas subtropicais e tropicais do mundo. Também reconhecida como uma doença zoonótica emergente em seres humanos, particularmente em indivíduos imunocomprometidos (Brayton et al., 2007).

Dentre as dez doenças especificadas como prioridades de investigação pelo Programa Especial da Organização Mundial da Saúde para Pesquisa e Treinamento em Doenças Tropicais (WHO, 2012), quatro são causadas por protozoários parasitas (malária, leishmaniose, doença de Chagas, doença do sono). Estas doenças, e outras como a amebíase e tricomoníase, apresentam um aumento alarmante de casos de refratividade ao tratamento principal. O fracasso do tratamento tem, potencialmente, uma origem multifatorial e como uma das principais preocupações a resistência a drogas (Sobel et al., 1999; Burri & Keiser, 2001; Arevalo et al., 2007; Arango et al., 2008). Além disso, das várias doenças que afetam a saúde humana, existem aquelas denominadas Doenças Tropicais Negligenciadas (DTN), sendo que cerca de um bilhão de pessoas são afetadas por uma ou mais DTN e são assim denominadas porque persistem em comunidades mais pobres (Yamey & Torreele, 2002; Winters, 2006). A Organização Mundial de Saúde reconhece cerca de 17 DTN e, entre elas, três são causadas por protozoários: doença de Chagas (*Trypanosoma cruzi*), tripanossomíase africana humana (doença do sono) (*T. brucei*) e leishmaniose (*Leishmania* spp.). Lembrando que além dessas, outras parasitoses também afetam uma grande parcela da população mundial, especialmente em países em desenvolvimento, causando mortes ou limitando a qualidade de vida.

No Brasil, por exemplo, boa parte das doenças parasitárias que afetam os seres humanos é causada por protozoários, como: Malária (*Plasmodium* spp.), Giardíase (*Giardia lamblia*), Doença de Chagas (*T. cruzi*), Toxoplasmose (*Toxoplasma gondii*), Criptosporidiose (*Cryptosporidium spp*.), Tricomoníase (*Trichomonas vaginalis*) e Amebíase (*Entamoeba histolytica*). Estas DTN e outras parasitoses têm recebido baixa prioridade quando comparadas às doenças que causam alta mortalidade, entretanto, a grande maioria destas DTN pode ser prevenida ou controlada. Contudo, os grupos de pessoas mais afetadas (entre elas majoritariamente crianças) são pobres, não representam um mercado lucrativo para a indústria de medicamentos e possuem baixo acesso aos métodos preventivos destas doenças (dos Santos et al., 2012).

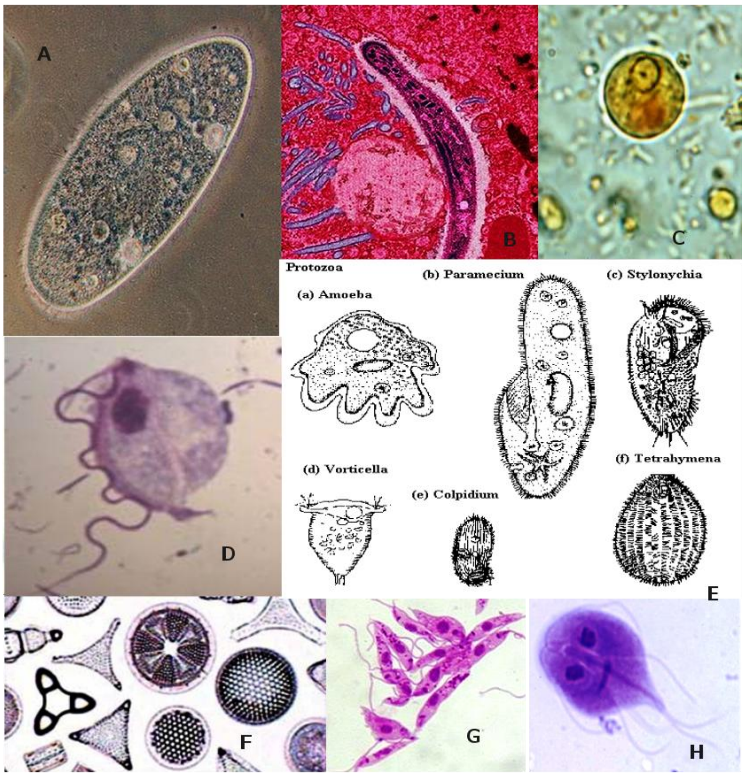


Figura 1.1: Ilustração composta por fotos/figuras de protozoários, patogênicos ou não, para que seja possível visualizar a grande diversidade de formas encontradas entre estes organismos. A: *Paramecium* sp.; B: *Plasmodium* SP.; C: *Entamoeba* sp. D: *Trichomonas* sp. E: prancha de diversidade de protozoa (ameba, paramécio, stylonychia, vorticela, colpidium e tetrahymena); F: diversidade de carapaças de diatomáceas; G: *Leishmania* sp.; H: *Giardia* sp. Figura montada com imagens disponíveis na Internet.

Os estudos genômicos e proteômicos desses protozoários têm ajudado na identificação de vias estágio-especificas não documentadas (Atwood et al., 2005). Em função disto, a comparação entre genomas de diferentes patógenos, por exemplo, tem ajudado a encontrar diferenças relevantes e linhagens específicas entre eles. Sendo que estas diferenças ajudam a reconhecer melhor genes importantes relacionados à patogenicidade de cada protozoário (El-Sayed et al., 2005b).

1.2 Caracterização das espécies

O filo Apicomplexa é um filo eucarioto cujos membros compartilham um aparelho secretor apical comum mediando a locomoção, invasão celular ou tissular (Abrahamsen et al., 2004). Estes organismos formam o superfilo Alveolata que se originou há aproximadamente 930 milhões de anos atrás (Gardner et al., 2005). Atualmente, existem em torno de seis mil espécies descritas, sendo as mais conhecidas parasitas obrigatórias e algumas destas espécies causam importantes doenças em humanos ou animais como: a malária causada por espécies de *Plasmodium*, Toxoplasmose que têm o *Toxoplasma gondii* como agente etiológico, a coccidiose (*Eimeria* spp.), Babesiose (*Babesia* spp.), Teilerioses (*Theileria* spp.), Criptosporidiose (*Cryptosporidium* spp.), *Neospora*, *Sarcocystis* e *Cyclospora* (Abrahamsen et al., 2004; Sato, 2011).

A Babesiose é uma doença hemoprotozoaria causada pela espécie *Babesia bovis*, transmitida através de carrapatos vetores, enzoótica[1] em ruminantes na maioria das áreas tropicais e subtropicais do mundo. Reconhecida como uma doença zoonótica emergente de seres humanos, particularmente em indivíduos imunocomprometidos (Brayton et al., 2007).

O gênero *Theileria* é um patógeno eucarioto intracelular capaz de transformar reversivelmente suas células hospedeiras. *T. annulata* e *T. parva* são transmitidas, assim como *B. bovis,* por carrapatos hemoparasitas e originam doenças linfo-proliferativas no gado, conhecidas como: Teileriose Tropical (*T. annulata*); e Febre da Costa do Leste (*T. parva*). A mortalidade e morbidade da Teileriose são atribuídas à capacidade do esquizonte transformar, malignamente, o linfócito bovino, sua célula hospedeira (Pain et al., 2005).

[1] Doença enzoótica**:** Doença de ocorrência estável em período sucessivos, em um rebanho ou determinada região.

O gênero *Cryptosporidium* compreende dois grupos de parasitas que se adaptaram a diferentes ambientes no trato gastrointestinal: (i) intestino delgado/cólon, onde a maioria das espécies como *C. parvum* e *C. hominis* se multiplica e; (ii) estômago, local de infecção de algumas espécies como *C. muris* que divergiu das espécies intestinais provavelmente como resultado da adaptação ao ambiente diferente do hospedeiro (Widmer et., 2009). As espécies de *Cryptosporidium* causam gastroenterite aguda e diarréia em todo o mundo, sendo transmitidos através da ingestão de oocistos completando seu ciclo de vida em um único hospedeiro. Um grau substancial de morbidade e mortalidade está associado com as infecções em pacientes com AIDS (Abrahamsen et al., 2004). Apesar dos esforços intensivos nos últimos 20 anos, não existe atualmente terapia eficaz para tratar ou prevenir infecções de *C. parvum* e *C. hominis* em humanos. Com isto, seu controle concentra-se na eliminação de oocistos nos suprimentos de água (Xu et al., 2004; Widmer et al., 2009).

O *Toxoplasma gondii* é um protozoário do filo Apicomplexa, sendo que por volta de 1970 foi reconhecido como um coccídeo (assim como *Cryptosporidium*). Este parasita é amplamente distribuído no mundo e estima-se que infecte 1/3 da população humana mundial. Apresenta a capacidade de infectar várias espécies de animais de sangue quente, sendo um significativo patógeno zoonótico e veterinário (Weiss & Dubey, 2009). A maioria das pessoas infectadas após o nascimento é assintomática, no entanto, alguns podem desenvolver mal-estar, febre, linfo-adenopatia e várias outras síndromes clínicas, especialmente nos indivíduos com AIDS, onde pode causar encefalite, infecções sistêmicas, infecção congênita e mortalidade neonatal (Elmore et al., 2010).

Estima-se que a linhagem de *Plasmodium* surgiu entre 100-180 milhões de anos atrás, sendo que as espécies deste gênero são conhecidas por infectar aves, mamíferos e répteis (Carlton et al., 2002). A malária humana é causada por espécies deste gênero, *Plasmodium*, que são parasitas intracelulares transmitidas de um hospedeiro para outro por mosquitos vetores do gênero *Anopheles* spp. (Gardner et al., 2002). Quatro espécies que infectam o Homem são descritas: *P. falciparum*, *P. vivax*, *P. ovale* e *P. malariae* (Hall & Carlton, 2005). *P. falciparum* é a espécie mais letal e virulenta, sendo conhecida como "terçã maligna". *P. vivax* (que provoca "terçã benigna") é a mais prevalente e possui ampla distribuição mundial, sendo responsável por 50% dos casos de malária na América Central e do Sul, Ásia e subcontinente Indiano (Carlton, 2003; Carlton et al., 2008). Entretanto, *P. knowlesi* que é um parasita cujo hospedeiro natural é *Macaca fascicularis*, cada vez mais vem sendo reconhecido como uma causa significativa da malária humana, particularmente no Sudeste Asiático, podendo ser a quinta espécie a causar malária em seres humanos (Hall et al., 2005). Apesar de mais de um século de esforços para sua erradicação ou controle, a doença continua sendo uma ameaça crescente para a saúde pública e para o desenvolvimento econômico dos países tropicais e subtropicais do mundo. Aproximadamente 40% da população mundial vive em áreas onde a malária é transmitida, estimando-se que existam de 300 a 500 milhões de casos e até 2,7 milhões de mortes/ano, o que a torna uma das doenças mais importantes que afeta a humanidade. Em países com malária endêmica, as taxas anuais de crescimento econômico ao longo de um período de 25 anos foram 1,5% menor que em outros países. A resistência aos medicamentos anti-maláricos, inseticidas, a decadência da infraestrutura de saúde pública, os movimentos da população, agitação política e as mudanças ambientais contribuem para a propagação da malária (Gardner et al., 2002). Para tentar entender e combater o parasita, boa parte dos procedimentos experimentais utilizados em *P. falciparum* foi inicialmente desenvolvidos nas espécies de malária de roedores *P. chabaudi*, *P. yoelii* e *P. berghei*, que são estreitamente relacionadas entre si (Carlton et al., 2002; Pain et al., 2008). Uma vez que estas três espécies causadoras de malária, de uso comum em laboratórios, fornecem modelos de sistemas que permitem abordar questões que são difíceis de responder com as espécies infecciosas em humanos, pois reproduzem muitas das características biológicas dos parasitas da malária humana, que são de difícil cultivo em laboratório (Hall et al., 2005).

A *Entamoeba histolytica* é um parasita do intestino humano, amitocondriado, normalmente contraído pela ingestão de água ou alimento contaminado. A infecção com o parasita é endêmica em muitas partes do mundo onde a infraestrutura de saneamento é pobre. Estima-se que afeta pelo menos 50 milhões de pessoas a cada ano, causando mais 100 mil mortes. A amebíase pode resultar na invasão da parede intestinal, levando à diarréia e disenteria (fezes com sangue) (Lorenzi et al., 2010). Recentemente, Diamond e Clark em 1993, separaram esta espécie em duas: a potencialmente virulenta *E. histolytica* e a forma avirulenta *E. dispar*, parente mais próximo descrito de *E. histolytica*, sendo morfologicamente idêntica, conhecida por viver como um comensal no intestino e por não ser virulenta (Loftus et al., 2005; Weedall & Hall, 2011).

O parasito Diplomonadida *Giardia lamblia* (sinonímia: *G. intestinalis* ou *G. duodenalis*) é um organismo unicelular, amitoncondriado, que infecta o intestino delgado de humanos e uma variedade de outros mamíferos (Adam, 2000). É um Protozoário enigmático, membro da ordem Diplomonadida, que inclui tanto espécies de vida livre e espécies parasitárias (Morrison et al., 2007; Thompson, 2009). A giardíase se desenvolve quando cistos são ingeridos (Adam, 2000), sendo comum entre pessoas com higiene fecal-oral pobre. Os principais modos de transmissão incluem o abastecimento de água contaminado ou atividade sexual. Seus sintomas são muito variáveis, entretanto compreendem geralmente diarréia, dor epigástrica, náuseas, vômitos e perda de peso (Morrison et al., 2007; Thompson, 2009).

Trichomonas vaginalis é um protista flagelado, membro da linhagem Parabasilia, eucariotos microaerofílicos, que não possuem mitocôndrias e peroxissomos, mas apresentam uma organela diferenciada denominada hidrogenossomo. Esta espécie é causadora da tricomoníase, uma infecção humana comum, sexualmente transmissível, com aproximadamente 170 milhões de casos anuais em todo o mundo (Carlton et al., 2007).

Os tripanosomatídeos são um grupo notável de protistas (Simpson et al., 2006) e formam um clado rico em espécies (Kinetoplastea) dentro do filo Euglenozoa (Adl et al., 2005). O seu nome reflete a presença de uma estrutura característica, o cinetoplasto, que consiste de uma massa densa de DNA extranuclear contido dentro da sua mitocôndria única. Os Kinetoplastidas são tradicionalmente considerados como composto por dois subgrupos: (i) os tripanosomatídeos uniflagelados que incluiu importantes gêneros parasitas, como *Trypanosoma* e *Leishmania*; (ii) os bodonideos biflagelados, com uma variedade de gêneros de vida livre e parasitária (Deschamps et al., 2011). As espécies do gênero *Trypanosoma* podem causar a doença do sono e a doença de Chagas, enquanto que *Leishmania* pode matar e debilitar centenas de milhares de pessoas em todo o mundo a cada ano (Grimaldi & Tesh, 1993; Simpson et al., 2006). Coletivamente causam doenças e morte de milhares de seres humanos e muitas infecções em outros mamíferos, principalmente nos países em desenvolvimento das regiões tropicais e subtropicais. Não existem vacinas para essas doenças e apenas algumas drogas estão disponíveis, porém inadequadas devido à toxicidade aos hospedeiros e resistência dos parasitas. Cada doença é transmitida por um inseto diferente e possui seu próprio ciclo de vida, tecido-alvos e patogênese distinta em seus hospedeiros mamíferos. Além disto, estes parasitas empregam diferentes estratégias de evasão do sistema imunológico

(Mauricio et al., 1999; Barrett et al., 2003; Pays et al., 2004; Berriman et al., 2005; El-Sayed et al., 2005a; Murray et al., 2005). A infecção por *Leishmania*spp. resulta em um amplo espectro de doenças humanas, denominado leishmanioses, apresentando uma incidência anual de 2 milhões de casos em 88 países e transmitidas por flebotomíneos do gênero *Lutzomyia* e *Phebotomus*. A espécie parasita do Novo Mundo, *L. braziliensis*, é o agente causador da leishmaniose tegumentar americana (LTA), enquanto no Velho Mundo as espécies *L. major* e *L. infantum*, presentes na África, Europa e Ásia, são parasitas causadoras da leishmaniose cutânea (LC) e leishmaniose visceral (LV), respectivamente (Grimaldi & Tesh, 1993; Mauricio et al., 1999; Ivens et al., 2005; Murray et al., 2005; Peacock et al., 2007). O *T. cruzi*, agente etiológico da doença de Chagas (Vallejo et al., 2002; WHO, 2010) é um grave problema de saúde na América Latina, sendo uma das causas de doença cardíaca. Estima-se que 16-18 milhões de pessoas estejam infectadas levando a óbito cerca de 13-21 mil pessoas/ano (Miles et al., 2009; Subileau et al., 2009; WHO, 2010). *T. brucei*, agente etiológico da "doença do sono", é um parasita transmitido pela mordida de moscas hematófagas tsé-tsé (*Glossina* spp.). As estimativas do nível de transmissão alcançam cerca de 300 mil novos casos anualmente, causando a morte de 50 mil pessoas em 36 países na África Sub-saariana, especialmente na população pobre dos campos de alguns dos países menos desenvolvidos da África Central (Berriman et al., 2005; WHO, 2010).

1.3 Genômica Comparativa, Homologia, genes homólogos e genes órfãos.

Com a disponibilidade crescente de genomas completos de diversas espécies, estudos funcionais destes organismos permitem a observação de características relacionadas à epidemiologia, traços e fenótipos de cada uma destas espécies, assim como analisar a dinâmica evolutiva entre espécies distintas, frente à conservação de genes entre as mesmas, por exemplo. Nesse contexto, identifica-se o conceito de Genômica Comparativa que se baseia principalmente no estudo da homologia e da dinâmica evolutiva dos organismos, seus respectivos genes e proteínas, podendo apresentar grande utilidade no entendimento da evolução das espécies pela comparação de seus genomas completos ou de genes-específicos de cada espécie (Hardison, 2003).

O conceito de homologia foi herdado da Botânica e da Zoologia e introduzido por Richard Owen em 1843. Neste contexto, a homologia era utilizada para definir órgãos com mesma origem (porém não necessariamente a mesma função), enquanto que órgãos com origens diferentes, porém com funções semelhantes (como a asa do morcego e da borboleta) eram denominados análogos, vide figura 1.2. Em 1970 Walter Fitch introduz este conceito na Biologia Molecular, onde genes homólogos possuem mesma função e origem comum. Dentro dessa idéia desdobram-se duas subcategorias principais: os genes ortólogos e parálogos. Cabe ressaltar que ao confrontar as sequências gênicas e seus produtos contra diversos bancos de dados disponíveis na Internet, e usando os programas apropriados podemos inferir estrutura e função hipotéticas aos genes ortólogos e parálogos (Koonin, 2005) e a existência de uma origem comum para a sequência. Portanto, homologia é um termo qualitativo e não quantitativo que fornece uma noção de ancestralidade (Koonin & Galperin, 2002).

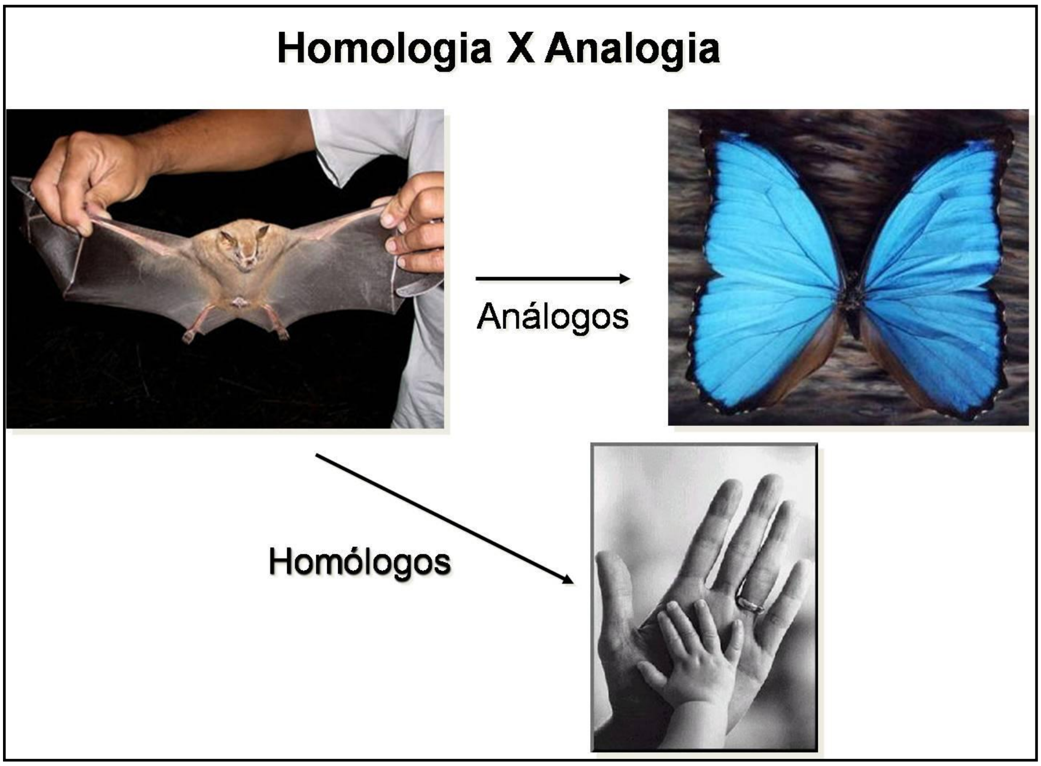


Figura 1.2: Exemplos de estruturas homólogas e análogas. A mão humana e a asa do morcego são exemplos de estruturas homólogos, pois compartilham a mesma origem, entretanto a asa do morcego e a asa da borboleta não compartilham uma origem comum, mas possuem funções idênticas, sendo denominados de estruturas análogas.

Os genes ortólogos descendem de um gene presente no último ancestral em comum às duas espécies comparadas. Tipicamente, ocupam o mesmo nicho funcional em diferentes organismos (Koonin et al., 2004). Sua definição inclui duas declarações distintas, importantes de serem levadas em consideração: (i) exigência de um único gene ancestral é central ao conceito de ortologia; (ii) presença de um gene ancestral no último ancestral comum das espécies comparadas, ao invés de um gene ancestral mais antigo. Esta definição assume a existência de um ancestral comum distinto das espécies comparadas, uma proposição, por vezes contestada para procariontes devido à alta incidência de transferência lateral de genes (Koonin, 2005; Makarova et al., 2005). Uma propriedade dos genes ortólogos é que estes normalmente desempenham funções equivalentes nos respectivos organismos, evitando-se a expressão "funções idênticas", uma vez que em contextos biológicos diferentes, as funções podem não ser necessariamente as mesmas. Enfatizado a assimetria das relações entre ortologia e função: genes ortólogos na maioria das vezes apresentam funções equivalentes, entretanto, o inverso pode não ser verdadeiro, uma vez que é comum situações onde funções equivalentes são realizadas por proteínas não-ortólogas. Em função disto, não é correto referir-se a genes ortólogos funcionais, porém é correto afirmar que genes ortólogos têm função semelhante (Koonin, 2005).

Os genes parálogos são resultantes de um processo de duplicação gênica e tendem a evoluir em novas funções (Koonin et al., 2004). O estudo de genes parálogos traz a compreensão da evolução dos organismos. Em 1970, Ohno propõe em seu livro, *Evolution by Gene Duplication*, que tão logo se inicia o processo de duplicação dos genes: (i) a pressão de seleção sobre um destes genes diminui, de modo a possibilitar a evolução do mesmo para incorporar novas funcionalidades ao genoma do organismo. Ou seja, após um processo de duplicação gênica a nova cópia fica livre para acumular mutações e desenvolver novas funções contribuindo para a diversificação funcional do gene. (ii) Inativação, acumulação de mutações deletérias levando o gene a transformar- se em pseudogene, processo denominado não-funcionalização. Este fenômeno ocorre em 99% dos casos de duplicação. No entanto, outras teorias propõem o relaxamento da pressão sobre as duas cópias com a divisão da função ancestral entre as duas cópias ou incorporação de novas funções por qualquer uma dessas cópias – fenômenos denominados subfuncionalização e sub-neofuncionalização, respectivamente (Koonin, 2005; Makarova et al., 2005). Quando esta duplicação gênica ocorre antes de uma especiação os genes são chamados Parálogos Externos e não são restritos a um mesmo genoma. Quando a duplicação gênica é restrita a um mesmo genoma podemos denominá-los Parálogos Internos. Estes, geralmente retêm a função ancestral, pois divergiram há menos tempo que os genes parálogos externos, ou seja, sua duplicação ocorreu após um processo de especiação (Sonnhammer & Koonin, 2002; Koonin, 2005). A ortologia e paralogia são termos intimamente ligados, porque se uma duplicação (ou uma série de duplicações) ocorreu após o evento de especiação que separava a comparação das espécies, a ortologia torna-se uma relação entre conjuntos de parálogos, ao invés de genes individuais e, neste caso, tais genes são chamados co-ortólogos (Sonnhammer & Koonin, 2002; Koonin, 2005; Makarova et al., 2007). A figura 1.3 explica a definição de genes homólogos ortólogos e parálogos utilizando o gene da hemoglobina.

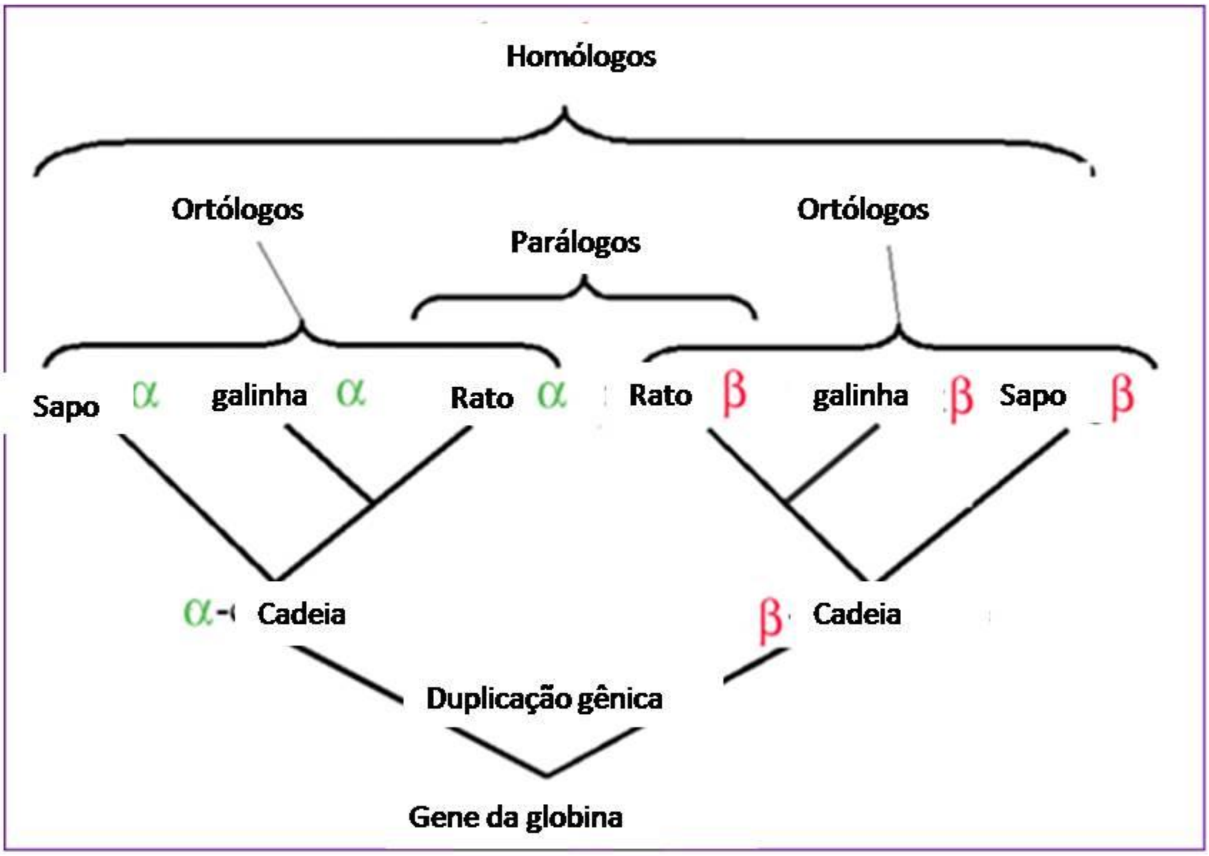


Figura 1.3: Exemplo das relações de homologia, ortologia e paralogia utilizando o gene da hemoglobina. Genes ortólogos e parálogos são dois tipos de sequências homólogas. Ortologia descreve genes em diferentes espécies que derivam de um ancestral comum. Paralogia descreve genes homólogos dentro de uma espécie que divergiram por duplicação gênica
(Fonte: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html).

O termo "órfão" originalmente possuía duplo significado: (i) regiões codificantes sem função conhecida ou (ii) regiões codificantes sem similaridade com outros genes no banco de dados. Atualmente, esta última definição (ii) é geralmente a mais utilizada. Todos os projetos genoma identificaram uma fração substancial de janelas abertas de leitura que não possuem nenhuma semelhança com outros genes no banco de dados. A origem evolutiva dos genes órfãos ainda é enigmática, entretanto, uma possível explicação para sua origem é de que estes evoluem tão rápido que a similaridade da sequência se perde até mesmo dentro de períodos evolutivos relativamente curtos (Domazet-Loso & Tautz, 2003).

1.4 Genômica comparativa de Apicomplexa

Em termos de afinidade evolutiva, estudos envolvendo a comparação dos genes homólogos de diferentes espécies de *Plasmodium* têm mostrado que este gênero é composto por vários ramos. As quatro espécies de malária humana formam um clado separado, onde *P. vivax* está agrupado mais próximo a *P. knowlesi* e outros parasitas de macacos, enquanto *P. falciparum* está mais estreitamente associados a espécies da malária aviária. As espécies de malária de roedores também formam um clado distinto. A composição do genoma varia de espécie para espécie e não é linhagem-específica. *P. falciparum*, por exemplo, possui 81% de A+T, enquanto *P. vivax* apresenta 62% de A+T. As espécies de malária de roedores possuem genomas igualmente ricos em A+T, assim como *P. falciparum*, enquanto *P. knowlesi* e *P. vivax* apresentam valores mais baixos de A+T. Cada espécie de *Plasmodium* parece ter entre 5-6 mil genes preditos por genoma, destes, 60% representam genes ortólogos entre as espécies. Esta diferença no número de genes entre espécies deve-se à expansão/contração gênica em diferentes linhagens (Carlton et al., 2005, 2008; Hall & Carlton, 2005). Sendo que as principais diferenças no conteúdo gênico entre as espécies de *Plasmodium* ocorrem entre genes envolvidos na interação com o sistema imune do hospedeiro, e muitos destes genes estão localizados nas regiões subteloméricas do genoma do parasita. O sequenciamento destas regiões em várias espécies identificou famílias de antígenos, alguns dos quais conservadas entre as espécies que causam malária em humanos e roedores (Carlton et al., 2002).

Comparações das taxas de evolução de genes duplicados contra não-duplicados em *P. falciparum* e *P. yoelii* demonstraram que genes duplicados evoluem mais rapidamente ao nível de nucleotídeos e possuem taxas aceleradas de ganho e perda de íntrons. Isso sustenta a teoria de que a expansão e diversificação de famílias de genes parálogos estão desempenhando um papel importante na evolução do parasita da malária (Hall & Carlton, 2005). Além disto, linhagens de *Plasmodium* exibem uma expansão diferencial de famílias gênicas que moldam a biologia específica de cada espécie (Carlton et al., 2008). Os genomas de *Plasmodium* que parasitam mamíferos (*P. falciparum*, *P. knowlesi*, *P. vivax* e *P. yoelii*) são notavelmente uniformes, 77% dos genes são ortólogos entre as quatro espécies e quase metade codifica proteínas hipotéticas conservadas de função desconhecida. *P. berghei* e *P. chabaudi* também são parasitas de mamíferos e quando analisadas estas seis espécies em conjunto, 3336 ortólogos foram encontrados entre elas, dos quais 3305 (99%) apresentaram inclusive a posição conservada (Carlton et al., 2008).

Em outra análise envolvendo *P. falciparum* e *P. yoelii*, por exemplo, encontrou-se um núcleo conservado de 4500 genes nas regiões centrais dos 14 cromossomos (Hall et al., 2005). A sintenia[2] entre estas duas espécies é elevada nas regiões centrais, onde se encontram os genes *housekeeping*, mas não nas regiões onde estão os genes envolvidos na variação antigênica/evasão do sistema imune do hospedeiro, que possuem mais de 800 cópias localizadas em regiões subteloméricas (Carlton et al., 2002). *P. knowlesi*, como dito anteriormente, é filogeneticamente próximo à *P. vivax* e cerca de 80% (4156/5185) dos genes ortólogos preditos em *P. knowlesi* podem ser identificados em *P. falciparum* e *P. vivax*. As famílias antigênicas variáveis específicas de *P. knowlesi*, como os genes SICAvar e KIR, formam as maiores expansões nesta espécie (Pain et al., 2008).

Agrupamentos de similaridade utilizando proteômas preditos de *B. bovis*, *T. parva* e *P. falciparum* resultaram na criação de 1945 agrupamentos ortólogos entre os três organismos (COG). Entretanto, o proteôma de *B. bovis* é mais estreitamente relacionado com o de *T. parva*, uma vez que cerca de metade das proteínas restantes de *B. bovis*, não incluídas no agrupamento entre os três organismos, ficaram nos agrupamentos ortólogos entre *T. parva* e *B. bovis*; enquanto *B. bovis* e *P. falciparum* compartilham apenas 111 grupos bidirecionais. Destas análises de agrupamento, 706 genes foram únicos para *B. bovis*, 1.107 para *T. parva* e 3309 para *P. falciparum* (Brayton et al., 2007). Análises de sintenia com *P. falciparum* e *C. parvum* ou entre *T. parva* e *C. parvum* mostraram um total de 435 regiões microsintenicas contendo 1279 ortólogos (Gardner et al., 2005). Entre os genomas de *T. annulata* e *T. parva* existem 3265 genes ortólogos e as regiões não-subteloméricas destes genomas mostram grande sintenia com poucas inversões de pequenos blocos de sequência, sem rearranjos inter-cromossômicos. A pequena divergência de similaridade de sequências entre *T. parva* e *T. annulata* sugere que elas expandiram-se após a especiação (Pain et al., 2005).

Já o genoma de *C. parvum* é semelhante a outros Apicomplexas formadores de cistos (por exemplo, *Eimeria* e *Toxoplasma*), possuindo aproximadamente 3807 genes codificadores de proteínas, valor abaixo dos cerca de 5-6mil genes preditos para *Plasmodium*. Esta diferença deve-se principalmente à ausência do genoma do

[2] Sintenia: Propriedade de dois ou mais genes estarem localizados no mesmo cromossomo. Termo utilizado para descrever a conservação na ordem de genes entre espécies relacionadas.

apicoplasto e da mitocôndria, bem como a presença de poucos genes que codificam funções metabólicas e proteínas variantes de superfície (Abrahamsen et al., 2004). Os genomas de *C. hominis* e *C. parvum* são muito semelhantes, exibindo apenas 3-5% de divergência de sequências, sem inserções, supressões ou grandes rearranjos evidentes. De fato, os genes das duas espécies são essencialmente idênticos, os poucos genes de *C. parvum* não encontradas em *C. hominis* são um mosaico de sequências provenientes de diversos progenitores, incluindo algas endossimbiontes hipotéticas formadoras do apicoplasto, mitocôndrias ou genes adquiridos a partir de procariontes por transferência lateral (Xu et al., 2004).

1.5 A genômica comparativa de Tripanossomatídeos

Os genomas das espécies de Tripanossomatídeos possuem grandes áreas sintênicas, de todos os genes em *T. brucei* e *L. major*, 68 e 75% permanecem no mesmo contexto genômico, respectivamente. O núcleo proteômico dos *Tritryp*[3] é formado por 6158 genes e um alinhamento da sequência de aminoácidos de uma grande amostra de seus COG (em 3-vias) revelou uma média de 57% de identidade entre *T. brucei* e *T. cruzi*; e 44% entre *L. major* e os outros dois *Trypanosoma* (El-Sayed et al., 2005b). As análises dos genomas dos *Tritryp* revelaram diferenças para com outros eucariotos no reparo de DNA, iniciação da replicação que refletem seu DNA mitocondrial “atípico”. Eles contêm, por exemplo, várias cópias dos quatro genes núcleos (H2A, H2B, H3 e H4) e ligadores (H1) de histona (Ivens et al., 2005).

A família de metaloproteases de superfície, GP63 é encontrada nos três Tripanossomatídeos e tem sido relacionada com a virulência, infecção de células hospedeiras e liberação de proteínas de superfície do parasita. *L. major* apresenta quatro genes GP63 e dois GP63-like, *T. brucei* possui 13 cópias e *T. cruzi* contem mais de 420 genes e pseudogenes de GP63 (El-Sayed et al., 2005b). Esta expansão de famílias gênicas, pela duplicação em *tandem,* é um mecanismo pelo qual os parasitas podem aumentar os níveis de expressão para compensar uma falta generalizada no controle transcricional (Berriman et al., 2005; Ivens et al., 2005).

A duplicação gênica seguida de divergência também é explorada para a geração de diversidade antigênica, particularmente em *T. brucei* e *T. cruzi* que exibem extensa modificação pós-traducional, especialmente para as proteínas de superfície e as secretadas (Ivens et al., 2005). *T. brucei* tem 900 pseudogenes e 1700 genes específicos, além de apresentar um repertório de 806 genes de

[3] *Tritryps*: Nome dado a junção dos três organismos: *Leishmania major*, *Trypanosoma brucei* e *Trypanosoma cruzi.*

glicoproteínas variantes de superfície (VSG), em grandes arranjos subteloméricos, usados pelo parasita para evadir do sistema imunológico dos mamíferos. A maioria dos genes de VSG são pseudogenes e podem ser utilizados para gerar um mosaico de genes expressos por recombinação ectópica. As comparações do citoesqueleto e sistemas de tráfico endocítico com os dos seres humanos e outros organismos eucarióticos revelaram mais diferenças entres estes organismos e os *Tritryps* (Berriman et al., 2005).

Quanto ao genoma de *T. cruzi,* mais de 50% é constituído por sequências repetidas, tais como retrotransposons e genes de grandes famílias de moléculas de superfície incluindo as trans-sialidases, mucinas, gp63 e uma grande família de genes de mucinas associadas a proteínas de superfície (MASP). Análises no genoma de *T. cruzi* revelaram 1052 agrupamentos parálogos (de mais de dois genes) contemplando 8419 cópias, dos quais 46 clusters (somando 3836 cópias) continham 20 ou mais parálogos (El-Sayed et al., 2005a). Já o genoma de *L. major* possui 911 genes de RNA preditos, 39 pseudogenes e 8272 genes codificadores de proteínas dos quais 36% possui uma função hipotética atribuída. Estes 8272 genes incluem aqueles, possivelmente, envolvidos na interação parasita-hospedeiro, tais como enzimas proteolíticas e uma extensa maquinaria para a síntese de complexos glicoconjugados de superfície.

A maioria dos genes de *L. major* é compartilhada pelos genomas de *T. brucei* e *T. cruzi*. No entanto, 74 grupos de ortólogos contêm genes apenas de *L. major* e *T. brucei*, enquanto 482 grupos de ortólogos apresentam genes apenas de *L. major* e *T. cruzi*, entretanto 910 genes de *L. major* não são ortólogos nos outros dois genomas (Ivens et al., 2005). Uma comparação entre os genomas de *L. braziliensis* e *L. infantum* com o genoma de *L. major* revelou uma conservação de sintenia de mais de 99% dos genes entre os três genomas e identificou apenas 200 genes com uma distribuição diferencial entre as espécies. *L. braziliensis*, por exemplo, possui componentes de uma via hipotética de interferência medida por RNA (RNAi), elementos transponíveis associados a telômeros e retrotransposons associados a sequências *spliced leader*. Mais de 1000 genes *Leishmania*-específicos foram encontrados, muitos dos quais permanecem sem caracterização. Dentre os genes espécies específicos, foram encontrados 5 genes *L. major*-específicos, 26 genes *L. infantum*-específicos e 47 *L. braziliensis*-específicos distribuídos por todo o genoma, sendo que a duplicação em *tandem* seguida pela diversificação pode ser a responsável por estas diferenças espécie-específicas (Peacock et al., 2007).

1.6 A filogenômica

A filogenômica é definida essencialmente como a interseção entre a evolução e a genômica (Graham et al., 2004; Delsuc et al., 2005). O termo tem sido usado para se referir às análises relacionadas a dados genômicos e a reconstruções evolutivas, especialmente filogenéticas. Outra definição é a reconstrução da vida evolutiva dos organismos com base nas análises dos seus genomas (Brown & Sjölander, 2006; Conte et al., 2008). Dentre as aplicações da filogenômica estão: (i) a predição da função de um gene baseada na história evolutiva, representada em uma árvore filogenética, idéia original de Jonathan Eisen (Eisen, 1998; Eisen & Wu, 2002; Eisen & Fraser, 2003); (ii) a reconstrução de uma árvore de espécies combinando informações de vários genes ou genomas inteiros; (iii) a integração de análises genômicas para a reconstrução evolutiva (Wu & Eisen, 2008). A abordagem filogenômica faz uso de um grande número de genes descobertos por projetos genoma e/ou pelo sequenciamento dos *EST (Expressed Sequence Tag)*. Os maiores alinhamentos pré-filogenômicos consistiram inicialmente de cerca de 10 genes. Nos dias de hoje, um alinhamento filogenômico compreende entre 50 e várias centenas de genes (Telford, 2007). A filogenômica pode utilizar genomas inteiros para inferir uma árvore de espécie e tornou-se o padrão para a reconstrução de filogenias de espécies confiáveis (Daubin et al., 2002; Ciccarelli et al., 2006). Atualmente, existe a opção de se concatenar sequências de genes múltiplos, na tentativa de se obter mais sinais filogenéticos para construir árvores no nível genômico, como as “árvores do genoma” - também chamadas “árvores de supermatriz”. A vantagem desta abordagem é que estas árvores são menos suscetíveis a erros estocásticos do que as árvores construídas a partir de um único gene (Doolittle, 1999; Dutilh et al., 2004; Jeffroy et al., 2006).

1.7 Grupos de Ortólogos e os Bancos COG e KOG

O banco de dados COG (*Clusters of Orthologous Groups*) foi construído com a idéia de classificar proteínas de genomas completamente sequenciados, utilizando como base o conceito da ortologia (Tatusov et al., 1997, 2000). As aplicações mais importantes do COG são a anotação funcional dos genomas recém sequenciados e análises evolutivas em larga escala, uma vez que inferir relações de ortologia e paralogia entre os genes/proteínas é importante para os aspectos funcionais e evolutivos da genômica (Tatusov et al., 2003).

Levando-se em consideração a existência de relacionamentos ortólogos de “um para muitos” e “muitos para muitos”, a tarefa de identificar ortólogos foi redefinida como uma delimitação de grupos de genes ortólogos, onde cada COG consiste de genes ortólogos ou de grupos parálogos de três ou mais linhagens filogenéticas que apresentam melhor resultado recíproco. Em outras palavras, duas proteínas de organismos diferentes que pertencem ao mesmo COG são ortólogos, além disso, assume-se que cada COG pode ter evoluído a partir de um gene ancestral, através de uma série de eventos de especiação e duplicação.

A abordagem para a identificação de conjuntos de proteínas ortólogas com base no agrupamento dos melhores resultados recíprocos do BLAST foi implementada na coleção do COG. Este protocolo de construção do COG incluiu: (i) procedimento automático para a detecção de possíveis ortólogos; (ii) separação manual dos vários domínios de proteínas em domínios de componentes; e (iii) verificação e anotação destes resultados. Essa identificação do COG baseia-se neste padrão de melhor resultado recíproco, sendo que o mais simples e o mais importante destes padrões é um triângulo, formado por genes ortólogos, vide figura 1.4. Ou seja, cada COG começa como um conjunto de proteínas que surgem como melhores resultados recíprocos de pelo menos três genomas divergentes após uma comparação de sequências de “todos-contra-todos” (Tatusov et al., 1997, 2003). Para permitir esta anotação automática de novas proteínas foi criado o sistema COGnitor, que justamente possibilita a anotação automática funcional e filogenética de genes e conjuntos de genes.

No processo que foi utilizado durante a construção do banco de dados COG, o critério para acrescentar prováveis novos ortólogos de outros genomas do COG foi baseado na coerência entre os relacionamentos observados. Para tal verificação, uma proteína é comparada ao banco de dados de sequências (COG) e acaba sendo incluída em um COG quando apresentar pelo menos dois melhor resultados recíprocos, conforme a figura 1.4. Durante a construção deste banco inicial (COG) utilizou-se as proteínas codificadas a partir de genomas completos, entretanto não é exigência de que proteínas recém-incluídas ao banco COG também sejam provenientes de um genoma completo (Tatusov et al., 1997).

A classificação de proteínas codificadas em um genoma sequenciado é fundamental para fazer com que as informações do genoma sejam úteis para estudos funcionais e evolutivos. Para facilitar tais estudos, o banco do COG foi dividido em 25 grandes categorias funcionais que abrangem quatro grandes áreas:

(i) armazenamento e processamento de informações; (ii) sinalização e processamento celular; (iii) metabolismo; além da inclusão de uma classe para a qual existe apenas uma (iv) previsão funcional geral (Tatusov et al., 2001, 2003). Atualmente, a Coleção do COG é constituída por 138.458 proteínas, que formam 4.873 COG e abrange 75% das 185.505 proteínas codificadas em 66 genomas de organismos unicelulares. O conjunto KOG (*Eukaryotic Orthologous Groups*) ou grupos de ortólogos de eucarióticos é formado por proteínas de sete genomas eucarióticos que consistem em 4852 grupos de ortólogos e incluem 59.838 proteínas, ou aproximadamente 54% dos 110.655 produtos gênicos de eucariotos analisados (Tatusov et al., 2003).

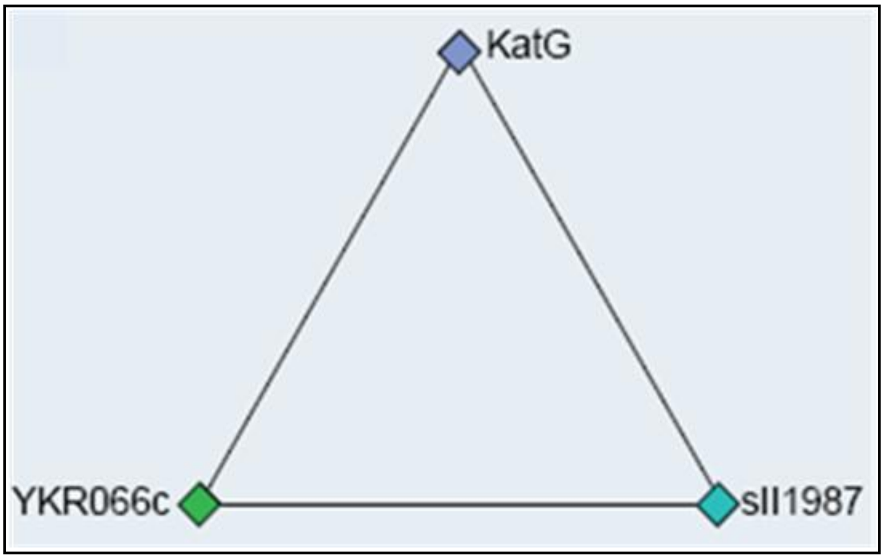


Figura 1.4: Triângulo de ortólogos formados por melhores resultados recíprocos, ou seja, COG mínimo. Origem das proteínas: KatG, *Escherichia coli*; sll1987, *Synechocystis* sp., YKR066c e *Saccharomyces cerevisiae*. Note que todos os melhores resultados recíprocos são simétricos (Tatusov et al., 1997).

1.8 O programa OrthoMCL

O procedimento de funcionamento do OrthoMCL (fig 1.5) inicia-se com comparações de “todo-contra-todos” utilizando o BLASTP, em um conjunto de sequências de proteínas dos genomas de interesse. As possíveis relações ortólogas são identificadas entre os genomas via melhores pares de similaridade recíprocos. Para cada possível ortólogo, prováveis parálogos "recentes" são identificados dentro do mesmo genoma como sequências que se encontram mutuamente mais semelhantes entre si, do que em relação a qualquer sequência de outro genoma. Em seguida, estas possíveis relações ortólogas e parálogas são convertidas em um grafo no qual os nós representam as sequências de proteínas e as arestas representam seus relacionamentos ponderados (normalizados), ou seja, estes relacionamentos recebem pesos diferentes. Os pesos são inicialmente calculados como a média de -log10 do *e-value* dos resultados do BLAST para cada par de sequências. Entretanto, levando-se em consideração que a grande similaridade dos parálogos "recentes" em relação ao ortólogos pode influenciar o processo de agrupamento, os pesos das arestas são então normalizados para refletir o peso médio para todos os pares de ortólogos nestas duas espécies (ou parálogos "recentes" ao se comparar dentro das espécies). Ajustando este viés sistemático entre as arestas que ligam as sequências dentro do mesmo genoma e arestas que ligam as sequências de genomas diferentes. O grafo resultante é representado por uma matriz de similaridade simétrica onde o algoritmo do MCL é aplicado para resolver os relacionamentos ortólogos de “muitos-para-muitos” inerentes a comparações entre vários genomas. Como resultado final deste processo, o OrthoMCL gera grupos contendo seqüências de pelo menos duas espécies: agrupamentos de grupos de ortólogos e parálogos "recentes".

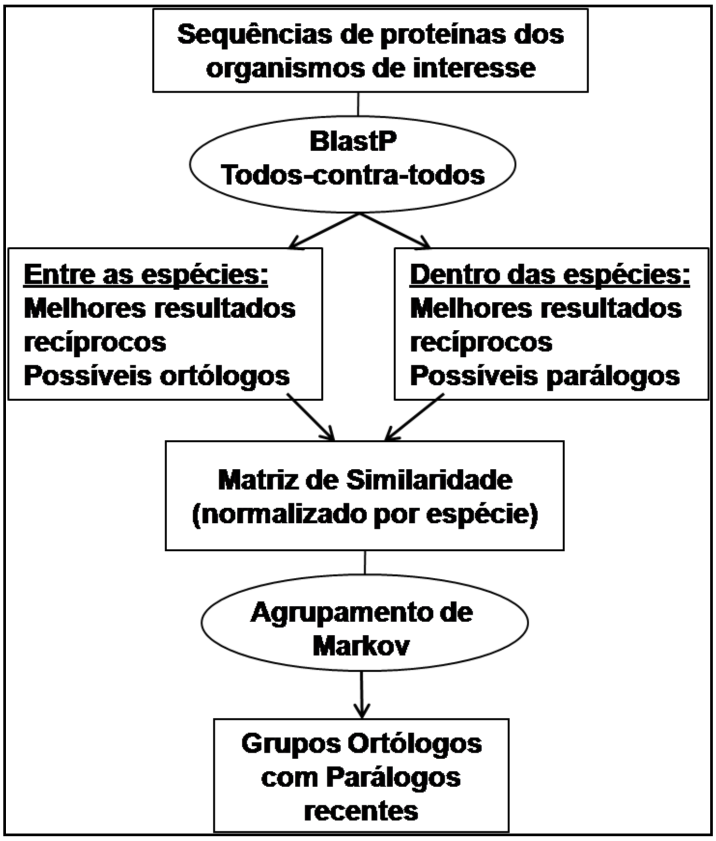

Sequências de proteínas dos organismos de interesse
BlastP Todos-contra-todos
Entre as espécies: Melhores resultados recíprocos Possíveis ortólogos
Dentro das espécies: Melhores resultados recíprocos Possíveis parálogos
Matriz de Similaridade (normalizado por espécie)
Agrupamento de Markov
Grupos Ortólogos com Parálogos recentes

Figura 1.5: Fluxograma do algoritmo OrthoMCL para agrupamento de proteínas ortólogas. O programa realiza, inicialmente, um BlastP de todos-contra-todos, passando pela geração de uma matriz de similaridade normalizado por espécie e finaliza gerando os grupos homólogos para o conjunto de proteínas submetidas dos diferentes genomas (Adaptado de Li et al., 2003).

1.9 O parasita *Leishmania* (*Leishmania*) *amazonensis*

A leishmaniose é uma doença infecciosa causada por parasitas do gênero *Leishmania*, distribuída em 88 países, como mencionado anteriormente. Estima-se que mais de 12 milhões de pessoas estão, atualmente, infectadas com *Leishmania* spp. Além disto, cerca de 350 milhões de pessoas que vivem nessas áreas de risco (zonas rurais e suburbanas pobres) são vulneráveis à infecção, sendo que cerca de 90% dos casos mundiais ocorrem no Afeganistão, Argélia, Brasil, Índia, Irã, Nepal, Peru, Arábia Saudita, Sudão e Síria (Desjeux, 1996; WHO, 2010). Dentre as espécies conhecidas de *Leishmania*, encontram-se pelo menos 20 espécies descritas pelo mundo que podem causar uma grande variedade e um complexo grupo de doenças humanas. Este espectro de doenças é caracterizado por sintomas que vão desde lesões cutâneas até a leishmaniose visceral (LV), que pode ser fatal dependendo das espécies e da resposta imune do hospedeiro (Barral et al., 1991; Grimaldi & McMahan-Pratt, 1991; Weigle & Saravia, 1996). Este gênero apresenta grande impacto em todo o mundo, com taxas de morbidade e mortalidade consideráveis, principalmente em países em desenvolvimento. Devido, em grande parte, à falta de um tratamento e de vacinas eficazes é um problema grave, uma vez que a maioria das drogas disponíveis, como por exemplo, o antimônio pentavalente, é tóxico ou pode causar grandes efeitos colaterais (Choi & El-Sayed, 2012; dos Santos et al., 2012).

De todas as formas conhecidas, a forma mais severa desta doença é a LV, causada principalmente por *L. donovani* e *L. infantum*. Nestes casos os parasitas afetam principalmente o fígado e o baço podendo gerar graves sintomas no hospedeiro, como: (i) imunossupressão; (ii) febre progressiva; (iii) perda de peso; (iv) anemia. Sendo que a ausência de um tratamento eficaz desta doença também pode ser fatal (Desjeux, 1996; Chappuis et al., 2007). Em outra variedade da doença, a Leishmaniose cutânea (LC), os parasitas *L. amazonensis*, *L. major, L. tropica e L. aethiopica* podem causar ulceração localizada a longo prazo, levando a cronicidade, latência e tendência a metástase no corpo humano (Akilov et al., 2007). A leishmaniose tegumentar americana (LTA) ou leishmaniose mucocutânea, causada principalmente pela espécie *L. braziliensis*, provoca a destruição do tecido da nasofaringe com lesões desfigurantes. Existe mais um tipo de manifestação clínica classificada de leishmaniose, a leishmaniose cutânea difusa (LCD), causada pelos parasitas *L. amazonensis*, *L. guyanensis* e *L. aethiopica*. Esta é uma doença de longa duração, devida a uma resposta imune celular deficiente, apresentando uma lesão primária progressiva e posteriormente a presença de múltiplas lesões metastáticas (Marsden, 1986; Cupolilo et al., 2003; Desjeux, 2004).

Quanto à leishmaniose no Brasil, a LC é uma doença endêmica causada por pelo menos seis espécies de *Leishmania* do subgênero *Viannia* e *Leishmania*. Os principais agentes de LC no sul da bacia amazônica são *L.* (*V.*) *braziliensis* e *L*. (*L.*) *amazonensis*, sendo que elas não demonstram diferenças nas manifestações clínicas, porém os efeitos das doenças são diferentes (Passos et al., 1999; TDR, 2013). Levando-se em conta a distribuição da *L. amazonensis*, ela é considerada uma espécie do Novo Mundo, pertencendo ao mesmo complexo da espécie *L. mexicana*, sendo que ambas contém apenas 34 cromossomos, devido à fusão do cromossomo 8 com 29 e do cromossomo 20 com 36 (Stiles et al., 1999). Enquanto as espécies *L.*(*L.*) *major* e *L.*(*L.*) *infantum*, que pertencem ao mesmo subgênero de *L.*(*L.*) *amazonensis*, são classificas como *Leishmania* do Velho Mundo e apresentam 36 cromossomos. Já as outras espécies de *Leishmania* do Novo Mundo, que pertencem ao subgênero *Viannia*, como *L.* (*V.*) *braziliensis* e *L.* (*V.*) *guyanensis*, possuem 35 cromossomos, uma vez que os cromossomos 20 e 36 estão fusionados (Wincker et al., 1996).

Como já mencionado, a espécie *L. amazonensis* está relacionada com a LC e LCD, no entanto, esta espécie já foi isolada de pacientes com LTA e VL. Por isto, infecções por *L. amazonensis* podem causar todos os tipos de manifestações clínicas conhecidas da leishmaniose (Barral et al., 1991). Embora este parasita possa levar a manifestação clínica dos quatro tipos diferentes de leishmaniose, o tipo mais comum que ela causa é leishmaniose cutânea com uma única lesão cutânea. Esta diversidade de doenças causadas por *L. amazonensis* aumenta a importância do sequênciamento do seu genoma e do seu estudo.

1.10 Organização da Tese

Esta tese, que possui como tema central a genômica comparativa de protozoários, é composta por três artigos que originaram um capítulo cada. O capítulo 3: “Genômica Comparativa de Protozoários: identificação de genes grupos e espécie-específicos e sua categorização funcional”, trata sobre a genômica comparativa de 22 protozoários focando os genes ortólogos e grupo específicos. O capítulo 4: “Sobre as particularidades dos Protozoários: inferindo as expansões de famílias e proteínas órfãs”, retrata o estudo de genômica comparativa dos mesmos 22 protozoários, mas focando os genes parálogos e suas expansões. A anotação e análise do genôma de *Leishmania amazonensis* originaram o capítulo 5: “Genômica comparativa e filogenômica do parasita *Leishmania amazonensis*”, que trata da anotação do genoma de *L. amazonensis* e da análise comparativa de seis espécies do gênero *Leishmania*, dentre elas *L. amazonensis*.

Além disto, anexamos dois artigos que possuem relação indireta com esta tese, uma vez que apresentam ferramentas e bancos de dados desenvolvidos/utilizados durante o andamento desta tese. O primeiro artigo em anexo: “ProtozoaDB 2.0: uma ferramenta para a extração de informações a partir do genoma de 22 espécies de protozoários patogênicos”, trata da origem e armazenamento dos dados utilizados nas análises de homologia. Além de permitir a visualização das informações de homologia geradas para os 22 protozoários estudados neste trabalho (cap. 3 e 4). O segundo artigo anexado: “STINGRAY: Sistema Integrado para Analises de Recursos Genômicos” discorre sobre o sistema de anotação, STINGRAY, desenvolvido pelo Laboratório de Biologia Computacional e Sistemas, utilizado para a anotação do genoma de *Leishmania amazonensis* (cap. 5).

2 - Objetivos

2.1 - Objetivo geral

Inferir as similaridades e diferenças de protozoários parasitas por meio da identificação de grupos homólogos, grupos específicos e proteôma núcleo de protozoários, através da genômica comparativa.

2.2 - Objetivos específicos

2.2.1 - Identificar os ortólogos entre as 22 espécies de protozoários

2.2.2 - Categorização funcional "*in silico*" do proteôma núcleo de 22 protozoa

2.2.3 - Identificar as proteínas grupo-específicos

2.2.4 - Categorização funcional "*in silico*" dos ortólogos grupos-específicos

2.2.5 - Identificar os parálagos para cada uma das espécies estudadas

2.2.6 - Categorizar funcionalmente "*in silico*" os parálogos específicos

2.2.7 - Identificar as proteínas órfãs dos organismos estudados.

2.2.8 - Anotar o genoma de *Leishmania amazonensis*

2.2.9 - Categorização funcional do genoma de *L. amazonensis*

2.2.10 - Analisar comparativamente o genoma de *L. amazonensis* e de outras cinco espécies de *Leishmania* spp.

3 - Artigo 1: Genômica Comparativa de Protozoários: identificação de genes grupos e espécie-específicos e sua categorização funcional.

Tschoeke DA, Jardim R, Lima JA, da Cruz SMR, Cuadrat RR, Mattoso M, Campos MLM, Dávila AMR. **The comparative genomics of Protozoa: identification of group and species-specific genes and their functional categorization.**

Artigo submetido para a revista *BMC Genomics*

Article title: The comparative genomics of Protozoa: identification of group and species-specific genes and their functional categorization
MS ID : 1440858931148857
Authors : Diogo A Tschoeke, Rodrigo Jardim, Joana Lima, Sergio MS Cruz, Rafael RC Cuadrat, Marta Mattoso, Maria LM Campos and Alberto MR Davila
Journal : BMC Genomics
A pdf file has been generated from your submitted manuscript and figures. We would be most grateful if you could check this file and let us know if any aspect is missing or incorrect. Any additional files you uploaded will also be sent in their original format for review.
http://www.biomedcentral.com/imedia/1440858931148857_article.pdf (4304K)

Este artigo corresponde aos objetivos específicos 2.2.1, 2.2.2, 2.2.3 e 2.2.4: Identificar os ortólogos entre as 22 espécies de protozoários; Categorização funcional "*in silico*" do proteôma núcleo de 22 protozoa; Identificar as proteínas grupo-específicos; e Categorização funcional "*in silico*" dos ortólogos grupos-específicos

O presente artigo faz uma análise dos genes ortólogos, utilizando os proteomas armazenados no ProtozoaDB 2.0. Aprofundando a análise destes genes ortólogos, encontramos que o proteoma núcleo de protozoa é constituído por 348 ortólogos, o proteoma núcleo de Kinetoplastida possui 5 mil ortólogos, enquanto que Apicomplexa e *Entamoeba* possuem 986 e 5915 ortólogos, respectivamente. Os resultados gerados por esta análise foram armazenados no ProtozoaDB 2.0 para sua melhor organização e visualização.

The comparative genomics of Protozoa: identification of group and species-specific genes and their functional categorization

Diogo A. Tschoeke[1,2], Rodrigo Jardim[1,2], Joana Lima[1,2], Sérgio Manuel Serra da Cruz[3], Rafael Cuadrat[1,2], Marta Mattoso[4], Maria Luiza Machado Campos[5], Alberto M. R. Dávila*[1,2]

1 Computational and Systems Biology Pole, FIOCRUZ, Rio de Janeiro, Brazil

2 Computational and Systems Biology Lab, Oswaldo Cruz Institute, FIOCRUZ, Rio de Janeiro, Brazil

3 PPGMMC - Postgraduate Program in Mathematical and Computational Modeling - Federal Rural University of Rio de Janeiro, Rio de Janeiro, Brazil

4 COPPE/PESC – Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

5 PPGI/UFRJ – Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

*****Corresponding: davila@fiocruz.br

Abstract:

Background: Pathogenic Protozoa species cause diseases such as malaria, sleeping sickness, Chagas disease, leishmaniasis, amoebiasis and giardiasis. Comparative studies among Protozoa are important because they may identify similarities and differences in these species. Furthermore, orthologous identification is central to functional characterization of genomes because they typically occupy the same functional niche in different organisms.

Results: 204,624 non-redundant proteins from *Plasmodium*, *Entamoeba*, *Trypanosoma*, *Leishmania*, *Giardia*, *Theileria*, *Toxoplasma*, *Trichomonas* and *Cryptosporidium*, totalizing 22 species, were submitted to OrthoMCL resulting in 26,101 homologs groups. Among them, 21,119 groups are orthologs and among these, 348 orthologs are shared by all 22 species, representing the Protozoa core proteome. We performed a phylogenomic analysis with the 348 concatenated PCP orthologs, resulting in a global supermatrix of 328,228 positions that generated a species tree for the 22 Protozoa. The inferred Kinetoplastida Core Proteome has 5,000 orthologous groups. From these, 67.92% (3,396/5000) are Kinetoplastida Specific and 46.29% (1,592/3,396) of these orthologs are annotated as “hypothetical”. Apicomplexa Core Proteome has 986 orthologous groups, from which 27.82% (224/986) are Apicomplexa Specific whereas 40.63% (92/224) were classified as hypothetical proteins. *Entamoeba* Core Proteome has 5,915 orthologous groups. On these groups, 75.08% (4,441/5,915) are *Entamoeba* Specific and 65.41% 2,905 (2,905/4441) were annotated as hypothetical.

Conclusion: The PCP genes are generally related to the maintenance of the cell and information processing, because there were found ribosomal proteins, histones, cytoskeletal proteins and tRNA synthetase. Among the specific orthologous groups, most of the orthologs have hypothetical functions, however functions like ABC-transporter, RNA-helicase and Rab-GTPase could also be found.

Background

The Protozoa

Protozoa are defined as single celled Eukaryotic organisms. However, this definition is very simplistic, mainly because of the small size of their cells, evolutionary history and the lack of uniform morphological features. Protozoa share characteristics of prokaryotic and derived Eukaryotic organisms, and their classification is a complex task [1]. Still, a wide range of definitions are available, Cavalier-Smith [2] defined them as 'unicellular phagotrophic Eukaryotes with mitochondria". This was also a very simple definition that includes the vast majority of Protozoa and excludes some organisms, but it would also include a few Chromista. Such description would not be sufficiently precise to define the kingdom's limits. One simple, phylogenetic definition of the Protozoa kingdom is: Eukaryotes, other than those that primitively lack mitochondria and peroxisomes (Archezoa), which lack the shared derived characters that define the four higher kingdoms: Animalia, Fungi, Plantae and Chromista [2, 3]. Electron microscopic discoveries eventually led to Bacteria being separated as a distinct kingdom and a five-kingdom system for Eukaryotes: (i) basal Protozoa and four derived kingdoms: two ancestrally heterotrophic (ii) Animalia; (iii) Fungi; and two ancestrally phototrophic (iv) Plantae; (v) Chromista [3, 4]. The most conservative approach is to treat Protozoa as a subkingdom, although it does not specify whether it belongs to Animalia or Protista.

Protozoa and their impact on health

Apicomplexa is an Eukaryotic phylum, classified as Protozoa, whose members share a common apical secretory apparatus mediating locomotion and tissue or cellular invasion [5], that forms the superphylum Alveolata, which is predicted to has been originated about 930 million years ago [6]. Currently, near 6,000 species are known. Most of them, named apicomplexans are obligated parasites, and some of them cause important human or animal diseases such as malaria (*Plasmodium* spp.), toxoplasmosis (*Toxoplasma gondii*), coccidiosis (*Eimeria* spp.), babesiosis (*Babesia* spp.), theileriosis (*Theileria* spp.), cryptosporidiosis (*Cryptosporidium* spp.), neosporosis (*Neospora)*, sarcosporidiosis (*Sarcocystis)*, and Cyclosporiasis (*Cyclospora)* [5, 7].

Babesia bovis is a tick-borne hemoprotozoan that causes Babesiosis transmitted by tick vectors, enzootic in ruminants in most sub-temperate and tropical areas of the World, recognized as an emerging zoonotic disease of humans, particularly in immunocompromised individuals [8]. *Theileria* genus contains the only intracellular

Eukaryotic pathogens capable of reversibly transforming its host cells. *Theileria annulata* and *Theileria parva* are tick-borne hemoparasites (hemoprotozoan) that give rise to lymphoproliferative diseases of cattle, known as tropical theileriosis (*T. annulata*) and East Coast fever (*T. parva*). Morbidity and mortality due to theileriosis are attributed to the ability of the schizont stage to malignantly transform its host cell, the bovine lymphocyte [9].

The *Cryptosporidium* genus comprises two groups of parasites which have adapted to different environments in the gastrointestinal tract; the small intestine/colon is where the majority of species multiply, and the stomach is the site of infection by a few species. *C. muris* has diverged from that of the intestinal species *C. parvum* and *C. hominis*, perhaps due the adjustment to the different host environment [10]. *Cryptosporidium* species causes acute gastroenteritis and diarrhea, and transmission occurs by ingestion of oocysts, completing its life cycle in a single host. A substantial degree of morbidity and mortality is associated with infections in AIDS patients [5]. Despite intensive efforts over the past 20 years, there is no effective therapy for treating or preventing *C. parvum*, *C. hominis* infection in humans, and control focuses on eliminating oocysts in water supplies [10, 11].

Toxoplasma gondii is a ubiquitous Apicomplexan Protozoa parasite of humans that is estimated to infect 1/3 of the World's human population and Around the 1970s, it was identified as a Coccidian (like C*ryptosporidium*). *T. gondii* can infect many species of warm-blooded animals and it is a significant zoonotic and veterinary pathogen [12]. Most people infected after birth are asymptomatic, however, some may develop fever, malaise, and lymphadenopathy, and other several clinical syndromes. Occasionally, individuals with HIV/AIDS can develop encephalitis or systemic infections, congenital infection and neonatal mortality [13].

The *Plasmodium* lineage is estimated to have arisen some 100-180 million years ago, and species of the parasite are known to infect birds, mammals and reptiles [14]. The malaria disease in humans is caused by a species of the genus *Plasmodium*, intracellular Protozoan parasites that are transmitted from host to host by the mosquito vectors *Anopheles*. Despite more than a century of efforts to control malaria, the disease remains a major and growing threat to the public health and economic development of countries in the tropical and subtropical regions of the World. Approximately 40% of the World's population live in areas where malaria is transmitted. There are 300-500 million estimated cases and up to 2.7 million deaths annually, making it one of the most important diseases affecting mankind. Resistance to anti-malarial drugs and insecticides, the decay of public health infrastructure, population movements, political unrest, and environmental changes are contributing to the spread of malaria. In countries with endemic malaria, the annual economic growth rates over a 25-year period were 1.5% lower than in other countries [15]. There are four species of *Plasmodium* that infect humans: *P. falciparum*, *P. vivax*, *P. ovale* and *P. malariae* [16]. *P. falciparum* is the most lethal and virulent species, and it is known as “malign tertian”. *P. vivax* (that causes “benign tertian”) is the most prevalent and it has achieved the widest global distribution, being responsible for 50% of malaria cases in Central and South America, Asia and the Indian subcontinent [17, 18]. *P. knowlesi* is an intracellular parasite whose natural vertebrate host is *Macaca fascicularis*; however, it is now increasingly recognized as a significant cause of human malaria, being the fifth species that can cause malaria in humans, particularly in southeast Asia. Three closely related species, *P. chabaudi*, *P. yoelii*, and *P. berghei*, are rodent malaria species, and they reproduce many of the biological characteristics of the human malaria parasite [16]. For this reason, the experimental procedures are refined for use with *P. falciparum,* providing model systems that allow issues to be addressed which are impracticable with the human-infectious species *P. falciparum* and *P. vivax* [14, 19].

Entamoeba histolytica is a parasite of the human large intestine, commonly acquired by ingesting contaminated water or food. Infection with the parasite is endemic in many parts of the World where sanitation infrastructure is deficient. Amoebiasis results from invasion of the gut wall, leading to diarrhea and dysentery (bloody stools).It affects at least 50 million people every year, causing over 100,000 deaths [20]. In 1993, Diamond and Clark separated this species into two: the potentially virulent *E. histolytica* and the avirulent *Entamoeba dispar*, which is the closest described relative of *E. histolytica* and is morphologically identical.It is not known to be virulent, but rather to live as a commensal in the gut [21, 22].

Giardia lamblia (synonym *G. intestinalis*, *G. duodenalis*) is an unicellular organism member of the Diplomonadida, which includes both free-living and parasitic species [23]. It infects the small intestines of humans and of a variety of other mammalian hosts worldwide. Giardiasis results when the environmentally resistant cyst is ingested [24]. It is common among people with poor fecal-oral hygiene, and major modes of transmission include contaminated water supplies or sexual activity. Symptoms are highly variable but include persistent, usually short-term, diarrhea, epigastric pain, nausea, vomiting and weight loss [23, 25].

Trichomonas vaginalis is a flagellated Protozoa, member of the parabasilid lineage of microaerophilic Eukaryotes that lacks mitochondria and peroxisomes. It causes trichomoniasis, a common but overlooked sexually transmitted human infection, with approximately 170 million cases occurring annually worldwide [26].

Kinetoplastids is a remarkable group of Protozoa [27] forming a species-rich clade (the Kinetoplastea) within the phylum Euglenozoa [28]. Among this group, some *Trypanosoma* species cause sleeping sickness and Chagas disease, whereas some *Leishmania* species cause different leishmaniases, killing and debilitating hundreds of thousands of people worldwide each year [27, 29, 30]. Collectively, they cause diseases and death in millions of humans and countless infections in other mammals, primarily in tropical and subtropical regions and developing countries.. There are no vaccines for these diseases and only few drugs are available, which are mostly inadequate because of toxicity and resistance. Each parasite is transmitted by a different insect and has its own life-cycle features, different target tissues, and distinct disease pathogenesis in its mammalian host. These parasites also use different immune evasion strategies[31].

The comparative genomics of Apicomplexa

In terms of evolutionary relatedness, studies involving the comparison of homologous genes from different *Plasmodium* species have shown that the genus is comprised of several deep branches. The four species responsible for malaria in humans - *P. falciparum*, *P. vivax*, *P. malariae* and *P. ovale* - form one clade, with *P. vivax* grouping apart in another clade close to monkey malaria parasites, such as *P. knowlesi*, and *P. falciparum*is, more related to avian malaria species. The rodent malaria species also form a different clade. Genome composition varies between species, and is not host lineage-specific; for example, the (A+T) genome composition of *P. falciparum* is 81% compared to 62% in *P. vivax*. The rodent malaria species have similarly high (A+T)-rich genomes, whereas *P. knowlesi* and *P. vivax* are less biased. Each *Plasmodium* species appears to have 5,000-6,000 predicted genes per genome. Of these, 60% represent orthologous genes between the species. The difference in gene number between species is due to gene expansion/contraction in distinct lineages [16]. The major differences in gene content among *Plasmodium* species occur between genes involved in interaction with the host immune system. Many of these genes are located in the subtelomeric regions of the parasite genome, and sequencing of these regions in several species has identified such antigen families, some of which are conserved between all species studied [14].

The comparative genomics of *Trypanosoma*tids

The genomes of the trypanosomatid species are highly syntenic: of all the genes in *T. brucei* and *L. major*, 68% and 75% remain in the same genomic context, respectively. The *T. cruzi*, *T. brucei*, and *Leishmania major* (Tritryp) core proteome is formed by 6158 genes. An alignment of the aminoacid sequence of a large sample of three-way COGs reveals an average of 57% identity between *T. brucei* and *T. cruzi*, and 44% identity between *L. major* and the two other trypanosomes [32, 33]. Analyses of the Tritryp genomes confirmed differences from other Eukaryotes in DNA repair and initiation of replication and reflect their unusual mitochondrial DNA. They contain multiple copies of the four core (H2A, H2B, H3, and H4) and linker (H1) histone genes [34]. The gp63 family of surface metalloproteases is found in the three trypanosomatids and has been implicated in virulence, host cell infection, and release of parasite surface proteins. *L.major* has only four copies of the gp63 gene and two copies of the gp63-like gene, *T.brucei* has only 13, and *T.cruzi* contains more than 420 genes and pseudogenes [32, 35]. Expansion of gene families by tandem duplication is a mechanism by which the parasites can increase expression levels to compensate for a general lack of transcriptional control [34, 36].

Methods

Dataset

Proteins sequences from 22 Protozoa (*Babesia bovis, Crypstosporidium parvum, C. hominis, C. muris, Entamoeba dispar, E. histolytica, Giardia lamblia, Leishmania braziliensis, L. infantum, L. major, Plasmodium berghei, P. chabaudi, P. falciparum, P. knowlesi, P. vivax, P. yoelii, Theileria annulata, T. parva, Toxoplasma gondii, Trichomonas vaginalis, Trypanosoma brucei, T. cruzi*) were downloaded from GenBank (release 181.0) (http://www.ncbi.nlm.nih.gov/genbank/) and RefSeq (release 24.0) (http://www.ncbi.nlm.nih.gov/RefSeq/) in fasta format.

Redundancy removal

Identical protein sequences were removed using the software CD-HIT [37] with the options sequence identity threshold and length difference cutoff equal to 1.00.

Homology identification

Sequences of 22 Protozoa were used for homologs identification by OrthoMCL with the cutoff of 1e-5 for P-value [38]. Considering the huge amount of data and the substantial computational time to infer the relationship between proteins, our analysis was performed in a cluster machine at COPPE/UFRJ, with 640 cores of processing, according to the protocol described by Coutinho [39]. The OrthoMCL output was used to infer proteins groups of orthologs and paralogs among Protozoa, as well as group-specific proteins and orphans proteins for each species. We loaded all the results into ProtozoaDB [40] using in house Perl/Ruby scripts specifically developed for this purpose.

Homolog statistics

Statistics of homologs were obtained from OrthoMCL results, including: (i) number of groups created, (ii) number of orthologous and paralogous groups, (iii) number of proteins used for each species. They were obtained using scripts written in Perl, Ruby and R, as well as UNIX commands.

Orthologous proteins identification

The identification of orthologous proteins was performed accessing the output file generated by the OrthoMCL software. We could then, identify orthologous proteins common to 22 studied Protozoa, until orthologous proteins shared by two species. The protein function of these orthologs was assigned considering the annotation of the given sequence. The orthologous proteins shared between the evaluated species, generated by OrthoMCL, were represented by Venn diagrams using R software [41].

Protozoa core proteome (PCP) identification and functional categorization

The PCP was identified and analyzed among the orthologous groups, defined as proteins shared by the 22 species of Protozoa studied. The functional categorization was conducted through an analysis with Blast and RpsBlast [42] against the database of orthologous genes in Prokaryotes (COG/NCBI) and Eukaryotic orthologous genes (KOG/NCBI). It used the value of 1e-5 for e-value, in both programs, to find out to which functional category each one of the orthologous groups belongs, according to a classification proposed by Tatusov in 1997 [43] and 2003 [44]. However, we removed the sequences from unicellular Eukaryotic species that were present in COG database, using an in-house script. Finally, the graphics of the functional categories were created with R software.

Finding putative specific groups

Specific proteins analyses were performed by a script using the OrthoMCL result. Orthologous proteins in Kinetoplastida, Apicomplexa and *Entamoeba* groups were detected by in-house Perl scripts developed for OrthoMCL orthologous groups, which check if the orthologous groups have only Kinetoplastida proteins. Then these orthologous groups were classified as specific to Kinetoplastida, for example.

In order to detect orthologous groups that are specific in RefSeq universe, ee performed a second check in these putative specific groups, running a Blast analysis against the RefSeq database (release 56; 18,132,578 sequences), with "-gi_negative_list" option (given all gi of Kinetoplastida proteins found in GenBank, as list), to identify proteins that are specific to *Entamoeba* / Apicomplexa / *Trypanosoma* and exclusive to them in accordance with RefSeq. In other words, proteins that are putative specifics and did not show similarity to other species than those defined as taxonomic groups were defined as specie-specific.

Phylogenomics to infer species tree

Our methodology was based on the study by Ocaña & Dávila [45] for the selection and construction of the species tree. The 348 PCP sequences were retrieved in fasta format from the OrthoMCL orthologous groups and then aligned using Mafft v5.861 [46] with default parameters. The individual alignments were concatenated using an in-house Perl script. A supermatrix tree was obtained using concatenated multiple alignments from entire protein sequences, resulting in a global supermatrix of 328,228 positions in a total of 22 species. This supermatrix was used to generate the tree with MEGA 5 [47], inferred by Neighbor Joining using 100 bootstrap replicates. However, because it is not simple to use multiple models in a single (concatenated) alignment, we decided to adopt JTT that was also the model adopted in the phylogenomics studies of Ciccarelli [48]. JTT assumed that there were two classes of sites, one class being invariable and the other class being free to change.

Results

Dataset and redundancy removal

A total of 346,468 proteins were obtained from NCBI (Genbank and RefSeq) databases for the 22 organisms. After redundancy removal with cd-hit program, a non-redundant dataset (called nr-Protozoa) containing a total of 204,624 entries was obtained. Table 1 shows the total number of obtained proteins from NCBI, the number of proteins after the redundancy removal and the number of predicted proteins reported in the literature, separated by species.

Homology identification

According to our calculations and based on data from literature, after redundancy removal, the minimum percentage proteome size used for orthology analysis correspond to *T. vaginalis* (84.1% of the predicted proteome size reported in the literature) (table 1). Only 83.61% (171,096/204,624) of the nr-Protozoa were grouped in some of those 26,101 homologs groups obtained by OrthoMCL. Of these groups, 19.09% (4,982/26,101) are paralogs and 80.91% (21,119/26,101) are orthologous groups shared from 2 till 22 species. A large amount of orthologous groups (figure 1) was observed as being shared by two species (7628 or 36.12%) and three species (3795 or 17.97%). Among those orthologous, we can highlight some in *Entamoeba*: Rab GTPase family; *Trypanosoma*: histone deacetylase; *Theileria*: Tpr-related protein family member; *Cryptosporidium*: peptidase calcium-dependent protein kinase; *Plasmodium*: CIR protein and BIR protein; *Leishmania*: amastin-like protein. When we analyze the genes shared only by two species, *Entamoeba* genus has 4441 orthologs (58.22%), and *Trypanosoma* genus has 900 orthologs (11.8%) (figure 2). Furthermore, the total number of orthologous groups shared by those 22 organisms is 348, representing the Protozoa Core Proteome (PCP). Moreover, this PCP represents 9.43% (348/3961) of the *B. bovis* proteome, 4.35% of the *L. major* proteomes and represents 0.69% (348/50189) of the *T. vaginalis* proteome (table 1).

T. vaginalis was the species that showed the largest number of proteins clustered by OrthoMCL (which had similarity with another protein to be clustered) with 44,241 (88.15% of proteome) proteins grouped (table 2), but only 4,708 homologs groups (table 3 and figure 3) were found, and it was the second species with the smallest number of orthologous groups (Table 3 and figure 4): 37.7% (1,775/4708 groups). *G. lamblia* was the species that showed the fewest number of proteins clustered into homologs: 39.89% (2,857/7,163) and 90% (3903/4306) of the non-clustered proteins have “hypothetical” function (figure 4). Furthermore, this latter species has the smallest number of orthologous groups compared with other analyzed species (figure 3 and table 2): 82.2% of the total homologs groups (1,181/1,436).

These results were loaded and can be viewed at ProtozoaDB [40] website (http://Protozoadb.biowebdb.org).

Protozoa Core Proteome (PCP) identification and functional categorization

PCP functional categorization showed that the most abundant category found by KOG and COG was “J” with 20,7% (72/348) and 29.6% (103/348), respectively (figure 5A). Along the PCP most of these orthologs are involved in translation or ribosomal structure, distributed in functional categories "J", "A", “L” and "O". Furthermore, in table 5,we can observe the function of some PCP proteins. We also conducted analyses based on score and e-value of their best Blast hits results in 348 PCP, in order to identify if they were closer to Eukaryotes or Prokaryotes. Our results show that 78.43% (273/348) groups were closer to Eukaryotes than Prokaryotes, 21% (73/348) groups showed similarity only to Eukaryotes and 0.57% (2/348) was closer to Prokaryotes: (i) methiony-tRNA synthetase was more similar to Alfa Bacteria and (ii) histone acethyl transferase was more similar to Archaea. Moreover, of the 348 PCP, 79.02% (275/348) groups also showed similarity against Prokaryotes organisms in COG and some are closer to Bacteria or Archaea. The categories “J” (79.46%) were more related to Archaea than Bacteria and were involved in information process, while the categories “G” (72.09%) were more similar to Bacteria and encode metabolic enzymes. Table 5 and figure 5B show the function and COG categories that were closer to Bacteria or Archaea. The figure 6 shows a phylogenetic tree Tyrosyl-tRNA synthetase (TyrRS) (COG0162 - information process protein), which indicates that this ortholog was closer to Archaea than Bacteria. The figure 7 shows that the metabolic gene Glucose-6-phosphato isomerase (GPI) (COG0164) was closer to Bacteria than Archaea.

Venn diagram analysis

Venn diagrams about (i) Kinetoplastida; (ii) *Plasmodium;* (iii) Piroplasmids; (iv) Coccids and (v) “other Protozoa” are shown in figures 8, 9, 10, 11 and 12, respectively. In these figures we can observe the core for each taxonomic group, as well as orthologous groups shared between these species.

a) Kinetoplastida

The five species of Kinetoplastida share 5000 orthologous groups. This core comprises 63.9% (5000/7825) of the *L. braziliensis* proteome, 63.52% (5000/7872) of the *L. infantum* proteome, 62.48% (5000/8003) of the *L. major* proteome, 58.59% (5000/8540) of the *T. brucei* proteome and 26% (5000/19247) of the *T. cruzi* proteome. Among those orthologs we found functions like: surface protease GP63, calpain-like cysteine peptidase. We also observed that nearly 17% (1368/7900) of the *Leishmania* spp. proteomes are shared only among them, not shared by *Trypanosoma* spp. While *T. brucei* and *T. cruzi* share 955 orthologs, comprising 10.5% (900/8540) and 4.7% of their proteomes, respectively (figure. 8).

b) *Plasmodium*

The six species of the genus *Plasmodium* showed 3,327 orthologous groups shared between them (figure. 9), with functions like: histone deacetylase. These orthologs comprise 34.2% (3327/9730) of the *P. berghei* proteome, 22.73% (3327/14639) of the *P. chabaudi* proteome, 45.55% (3327/7303) of the *P. yoelii* proteome, 56.62% (3327/5876) of the *P. falciparum* proteome, 65.21% (3327/5102) of the *P. knowlesi* proteome and 61.65% (3327/5397) of the *P. vivax* proteome. We also observed 290 orthologs shared by only rodent parasites (*P. berguei*, *P. chabaudi* and *P. yoelli*), and 165 orthologs are shared by *Plasmodium* parasites from humans. Nevertheless, the proteomic core of these two groups (parasites of rodents and humans) has 3763 and 4334, respectively.

c) Piroplasmids

The Piroplasmids proteomic core of these three species has 2561 orthologs, comprising 69.38% (2561/3691) of the *B. bovis* proteome, 67.57% (2561/3790) of the *T. annulata* proteome and 63.23% (2561/4050) of the *T. parva* proteome. Some of these orthologs have functions like: cytidine triphosphate synthetase and erythrocyte membrane-associated antigen. The core of the genus *Theileria* has 3176 orthologs like choline kinase and comprises 83.4% (3176/3790) and 78.42% (3176/4050) of the *T. annulata* and *T. parva* proteome, respectively (figure 10).

d) Coccideans

The Coccidia core (*T. gondii* and *Cryptosporidium* spp.) is comprised by 1711 orthologs which are shared by the four species and represent 44.69% (1711/3829), 44.04% (1711/3885), 43.54% (1711/3930) and 21.43% (1711/7984) of the *C. parvum*, *C. hominis, C.muris* and *T. gondii* proteomes, respectively, whereas *C. parvum* and *C. hominis* shares 399 orthologs, ATP-binding cassette for example, that comprises 10% (399/3860) of their proteomes, is not shared with *C. muris.* In figure 11, we see the Venn diagram for this group.

e) “Other Protozoa”

The species *E. dispar*, *E. histolytica*, *G. lamblia* and *T. vaginalis* have 667 orthologs as core, which comprise 7.75% (667/8606), 8.37% (667/7973), 9.31% (667/7163) and 1.33% (667/50189) of their proteomes, respectively (figure 12). Most of these orthologs are related to metabolism, like kinases and 1,4-alpha-glucan branching. Moreover, *T. vaginalis* and *G. lamblia* share 254 orthologous with each other, comprising 3.55% (254/7163) and 0.5% (254/50189) of the respective proteomes, like Carbamate kinase which is not shared with *Entamoeba* spp.

Analysis of putative specific proteins

At tables 6 and 7, the number of specific groups (orthologous) is showed. Apicomplexa proteome core is comprised of 986 orthologous groups, however 22.72% (224/986) are specific only to this group (Apicomplexa Specific Proteins - ASP). Some of the ASP are helicase, histone and RNA methyltransferase. For the five species of Kinetoplastida, 67.92% (3396/5000) are specific to this taxonomic group (Kinetoplastida Specific Proteins – KSP) and some of the specific proteins functions are hexose transporter, 40S ribosomal protein L14, 60S ribosomal protein L28, cytochrome P450 reductase. The *Entamoeba* Core Proteome (ECP) is the largest of the four observed core, comprising 5915 groups. Moreover, it has the highest percentage of specific groups (ESP) 75.08% (4441/5915) and the largest number of orthologous with hypothetical function 65.41% (2905/4441),while, Kinetoplastida has 46.29% (1592/3396) orthologs with hypothetical function and Apicomplexa 40.63% (92/224). If we analyze the groups number shared only by three species: rodent *Plasmodium* and human *Plasmodium*: 7.39% (278/3763) and 3% (130/4334) of their core are specific, respectively. In rodent's *Plasmodium* we find CIR/BIR protein and chloroquine resistance marker, while in Human's *Plasmodium* we find early transcribed membrane protein and rifin-like protein. Table 8 shows the function of 25 KSP, ASP and ESP, and table 7 shows the core size for all Protozoa (PCP), Apicomplexa, Kinetoplastida and *Entamoeba,* with their respective percentage of specific proteins, hypothetical groups number and representativity of KOG category "R".

By functional categorization using COG in specific orthologous groups (figure 13), we noted that the largest category for Kinetoplastida Specific Proteins (KSP), *Entamoeba* Specific Proteins (ESP) and Apicomplexa Specific Proteins (ASP) were "R" with 22.95% (201/876), 34.4% (365/1061) and 29.47% (28/95) of the characterized groups, respectively. When we functionally characterized the KSP, ASP and ESP using KOG categories (figure 14), we observed that in KSP the largest category was "R" with 15.17% (265/1746) of the characterized groups. In ESP and

ASP, the “T” category was the most abundant, with 16.41% (320/1950) and 11.97% (34/284), respectively.

Specie-Specifics proteins after Blast analysis against Refseq database

After Blast analysis against RefSeq database, 1.38% (47/3396) of the KSP showed no similarity to other species than Kinetoplastida, of which 45 have hypothetical function, one kinetoplast DNA-associated protein and the other elks delta-like protein. Among the 986 ASP, all showed similarity against other species and 6.46% (287/4441) of the ESP showed no similarity outside genus, whereas 227 have hypothetical function and 60 have some function defined as: caldesmon, DNA repair protein Rad-50. The complete list with the function of these proteins can be observed in table 9.

Phylogenomic analysis

For the phylogenomic analysis, we get the 348 PCP and construct a supermatrix of the 22, resulting in global supermatrix of 328,228 positions. The tree was generated with MEGA 5, by Neighbor Joining using 100 bootstrap replicates, with JTT model. In the phylogenomic tree (figure 15), we observed that the genus *Plasmodium* is a monophyletic group with bootstrap support of 100 to these branches. Piroplasmid is next to the *Plasmodium* genus, and we observe within Piroplasmid the grouping of *Theileria* genus and *Babesia*. The genus *Cryptosporidium* and *T. gondii* was the most basal branch with 100 of bootstrap value. The Kinetoplastida forms a monophyletic group with the separation of *Trypanosoma* and *Leishmania. Entamoeba* appears between Kinetoplastida and Apicomplexa forming a monophyletic group, and *Giardia lamblia* and *Trichomonas vaginalis* as outermost group.

Discussion

The redundancy caused by simultaneous usage of GenBank and Refseq entries was successfully removed as shown in Table 1. This allowed us to perform all the comparative analysis with the most complete possible dataset available at NCBI.

Shared orthologs number vary according to species relationships

The identification of the 348 orthologous groups (figure 5A and table 2 and 3) shared by the 22 Protozoa and reported in the present study (PCP) is novel because it has not been previously reported. When analyzing the KCP (figure 8) it is observed that Kinetoplastida shares 5000 orthologous groups. A similar number was reported by El-Sayed [32] that found 6158 orthologs shared by TriTryps. The smaller number of shared proteins found in our study is directly related to the fact that the KCP inferred by us involved five species and not only three (TriTryps). The three Piroplasmids species (figure 10) share 66.73% (2561/3843) orthologous groups, this value might be considered high since the average size of the genomes is about 3800 genes, and Brayton and colleagues (2007) proposed a core of approximately 2600 orthologs. On the other hand, the four species *E. dispar*, *E. histolytica*, *G. lamblia* and *T. vaginalis* share only 667 orthologous groups (figure 12), a low number compared with 71% (5915/8289) of orthologs shared between *E. histolytica* and *E. dispar*. This result is expected considering both species are phylogenetically close and that until 1993 they were considered to be only one species [22]. This group formed by *E. dispar*, *E. histolytica*, *G. lamblia* and *T. vaginalis* shares some particularities like oxidative and/or nitrosative stress resistance genes. These results corroborate previous studies [21], including the rubrerythrin absence in *G. lamblia*. While our study is the first attempt to infer the protein core among the 22 Protozoa species, the low number of proteins shared by all of them was something expected because the more phylogenetically distant the species, the smaller the number of genes shared by them. This is corroborated when species belonging to the same genus are analyzed, the amount of proteins shared by those species is higher than the number of proteins shared by phylogenetically distant species: e.g. *Entamoeba* spp., *G. lamblia* and *T. vaginalis* share only 667 orthologs, whereas the phylogenetically related *Leishmania* species [30, 33], shares 6368 orthologs among them (figure 4 and table 3). In another study, Brayton [8] analyzed species of the same phylum comparing the proteomes of *B. bovis*, *T. parva* and *P. falciparum* and inferred 1945 shared orthologous in triplicate genes, and approximately 900 of the remaining proteins of *B. bovis,* were not included in orthologous groups shared by the three organisms, but were grouped into genes shared between *B. bovis* and *T. parva* only resulting in approximately 2650 orthologs. Our results show that the number of proteins shared between these two species (*B. bovis* and *T. parva)* is 2667 groups, a similar number of orthologs was found by Brayton et al [8]). Although the three species analyzed by Brayton [8] belong to the phylum Apicomplexa, they are somewhat evolutionarily distant: *B. bovis* and *T. parva* are classified as Piroplasmids [7], while *P. falciparum* is classified as Haemosporida. Again, when organisms that belong to the same genus within the phylum Apicomplexa (for example, *P.*

falciparum, *P. berguei*, *P. chabaudi* and *P. yoelii)* are compared, the number of orthologous shared by those four species is 4391 [49]. In our study, we found a different result, consisting of 3437 orthologous groups shared by the same four species. However, Hall [49] inferred these orthologs through bidirectional BLAST searches, while we used OrthoMCL. These different approaches can explain this difference in the results. When the six species of the *Plasmodium* were analyzed by us, it was inferred that 3327 clusters of orthologous proteins are shared by them. This result is similar to the findings of Carlton [17] that found 3336 orthologs shared by the same six species. While most of these groups, shared by the six *Plasmodium* species, have hypothetical functions, we found some of those genes to be related to surface proteins like merozoite surface protein 7 (1 ortholog), erythrocyte membrane-associated antigen (49 orthologs) and Merozoite Surface Protein 8 (1 ortholog), which are *Plasmodium*-specific, corroborating Carlton's findings [17]. We also found an ABC transporter to be a specific ortholog for *Plasmodium* spp. This result was initially unexpected because ABC transporter is spread along many organisms. However, considering those genes are known to be under positive selection [50] and show high mutation rates, that could explain why the ABC transporter found in *Plasmodium* spp does not group together with the ABC transporter orthologs of the remaining Protozoa.

PCP analysis and functional categorization

The results indicating only 0.72% (2/275) (figure 5A) of the PCP are closer to Prokaryotes than to Eukaryotes (methionyl-tRNA synthetase and histone acetyltransferase) have not been previously reported. It indicates that while these proteins are common to Prokaryotes and Eukaryotes, they are phylogenetically closer to Prokaryotes. This is consistent with data from literature, because methionyl-tRNA synthetase was possibly transferred from alfa-Proteobacteria to Protozoa. But this transfer was only reported for *G. lamblia*, *P. falciparum*, *P. yoeli* and *L. major* [51] while, in our study, this protein is shared by the 22 Protozoa species, providing further support for the ancient transfer theory arguing that aminoacyl-tRNA synthetases enzymes existed prior to the divergence of Bacteria and Eukaryotes (Brown & Doolittle, 1999). The other possibly transferred gene, histone acetyl transferase-like, has some homologs in Bacteria that have been proposed to be the ancestors of their Eukaryotic homologs. Additionally, the discovery of histones in Archaea supports the argument that histones evolved before the divergence of

Archaea and Eukarya [53, 54]. Considering these results, the LGT (Lateral Gene Transfer) is evolutionarily important, especially among Bacteria and Protozoa [55]. Previous reports in the literature on *C. hominis* [11], *E. histolytica* [21], *L. major* [33, 34, 56] and *T. brucei* [36] showed that those genomes have a few genes of bacterial origin that contributed to some of the metabolic differences found in parasitic Protozoa. Such studies also reported that this possible LGT between Prokaryotes and Eukaryotes may be responsible for the mosaic of genes found in Protozoa. In a previous study, [56] found a number of LGT events: 68 in *L. major*, 63 in *E. histolytica*, 46 in *T. brucei*, 19 in *P. falciparum*, 21 in *G. lamblia*, 17 in *P. vivax*, 11 in *C. parvum*, 149 in *T. vaginalis*, 49 in *T. cruzi*, 16 in *T. gondii* and finally 16 in *P. yoelii*, most of them involved in aminoacid metabolism and enzyme reaction. In our study, we found Isocitrate Dehidrogenase (ICD) transferred in *L. major*, *L. infantum*, *L. braziliensis*, *T. vaginalis*, *E. histolytica* and *E. dispar* are phylogenetically closer to Bacteria than Eukaryotes (figure S1), corroborating and expanding previous findings [56, 57], since this ICD transfer wasdescribed only in *L. major* and *T. vaginalis.* Complementing Loftus [21] findings, we found ICD transferred in *E. histolytica*. Additionally, we suggest that this transfer might be previous to the separation of *Leishmania's* subgenus, because both *Viannia* and *Leishmania* subgenus possess this gene.

It should be highlighted that PCP entries are generally related to the maintenance of cell and information processing, like ribosomal proteins, histones, cytoskeletal proteins and tRNA synthetase, as shown in table 5. Furthermore, according to functional classification, the majority of genes are distributed in functional categories "J", "A" and "O", confirming that, in this case, conserved genes are expected to be housekeeping genes, then related to maintenance and processing. Additionally, they might be also categorized in two groups: "Storage and Information Processing" and "Process and Cellular Signaling," as observed in previous works [48, 58].

Despite Protozoa being Eukaryotic organisms, there are several reports in the literature showing they received genes from Prokaryotes (Bacteria and Archaea) and some studies argue that Eukaryotes are a chimeric construction [59–63]. There is a division between the “archaeal” and “bacterial” genes in Eukaryotes, considering informational genes and operational genes, respectively. The informational genes (involved in transcription, translation replication, and repair) are closely related to archaeal genes and they are named as “information processing genes”. The operational genes (involved in cellular metabolic processes such as amino acid biosynthesis, cell envelope and lipid synthesis, metabolic enzymes, components of membranes) are closely related to bacterial genes [60–62]. In other words, informational genes are involved in essential housekeeping functions and are closer to Archaea, while the metabolic genes are closer to Bacteria [64]. According to our results, orthologs classified as metabolic genes by COG are closer to Bacteria than Archaea (figure 5B and table 5), for example, glucose-6-phosphate isomerase (GPI) (figure 7). On the other hand, informational genes are closer to Archaea than Bacteria, for example Tyrosyl-tRNA synthetase (figure 8) (genes classified with COG categories “B”, “O”, “J”), corroborating previous work [62]. Nevertheless, the COG category K (Replication) was closer to Bacteria than Archaea, unlike what was argued on previous works [60–62]. But, when looking for orthologs that belong only to category “K” (Transcription), 94% (16/17) are closer to Archaea than Bacteria. The same scenario occurs for Replication (“L”): considering orthologs that belong only to “L” category, 64% (20/31) are closer to Archaea (64%) than Bacteria (36%). Contrary to what is described in literature, our findings suggest that genes related to Energy production and conversion (“C”) (as vacuolar proton translocating ATPase subunit A, vacuolar ATP synthase subunit B and vacuolar ATP synthase subunit) are closer to Archaea, considering that 70% (7/10) of these orthologs are closer to them. This unexpected results can be justified by the fact that Protozoa are organisms between prokaryotic and higher Eukaryotic organisms, sharing characteristics with these two groups, making their classification complex [1] or by the fact that Protozoa is basal in the Eukaryotic kingdom showing unresolved taxonomy position [65]. Our result showed that the Protozoa shares Eukaryotic and Prokaryotic characteristics, whereas 275 PCP are also similar to Eukaryotic and Prokaryotic organisms at the same time, corroborating previous work. As our results demonstrate, bacterial-like genes are found in the Eukaryotic genome and give origin to the Eukaryotic cell [61, 64, 66, 67]. In this scenario we highlight the Aminoacyl-tRNA synthetases (aaRSs) that probably was one of first to emerge from the RNA World [68], more specifically one type of the Aminoacyl-tRNA synthetase: the Tyrosyl tRNA-synthetase (TyrRS). This gene is closer to Archaea than Bacteria, as expected from informational genes. In our work this protein was also closer to Archaea. Brindefalk [67] reported that aaRS trees were consistent with the 3-domain life hypothesis. Cytoplasmic and archaeal aaRSs formed a group, excluding the bacterial and mitochondrial aaRSs. Bonnefond, when compared TyrRS sequences, also found Archae and Eukarya closer to each other than to their bacterial counterparts [68]. Our TyrRS tree of the 22 analyzed Protozoa (figure 6) was similar to the one found by Bonnefond [68] and Larson [69], where Protozoa are sister-clade of the Archaea. Furthermore, probably, we are working with citosolyc TyrRS and not with mitochondrial aaRS, because *G. lamblia* and *T. vaginalis* lack mitochondria. Our results show that these two organisms contain only a single TyrRS closer to Archaea, indicating that we are working the citosolyc TyrRS. The implication of this finding is that mitochondrial aaRS are closer to Bacteria, while citosolyc Protozoa aaRS are closer to Archaea than to aaRS from higher Eukaryotes or Bacteria [70], corroborating our results. Additionally, when we trace the phylogeny of the Glucose-6-phosphate isomerase (GPI), an important cytosolic enzyme involved in glycolysis and universally present in eukaryote [71, 72]. Our GPI-phylogenetic tree (figure 7) is closer to obtained by Grauvogel [72], suggesting that GPI is closer to Bacteria than Archaea. Since GPI from the amitocondriates *G. lamblia* and *T. vaginalis* are basal, closer to *Synechocystis* and *Nostoc*, two Cyanobacterias. Kinetoplastida are closer to gamma Bacteria as well as *Entamoeba.* However, Apicomplexa GPI are closer to Chlam-Spir (spirochaeta species), unlike to Grauvogel [72] who found them closer to dinophyta and glaucophyta GPI, but none of these proteins in Protozoa are closer to Archaea, as expected. The classical form of GPI is found in Gram-negative Bacteria and in the cytoplasm of the Eukaryotic cell and seems not to be encoded in the Archaeal genomes. However, some Archaea encodes a novel GPI or have different ways to metabolize glucose [73–75]. These results could explain the proximity of GPI Protozoa to Bacteria than Archaea. In other words, Protozoa received this protein from Bacteria, because Archaea does not have the protein, or encode a divergent copy that was retained in this domain. Our results corroborate this scenario, because all GPI Protozoa are closer to Bacteria.

Group-specific orthologs

By analyzing group-specific orthologs in Apicomplexa, *Entamoeba* and Kinetoplastida (table, 6, 7 and 8), our study is the first attempt to identify the core proteins for those taxonomic groups. KSP (Table 8) contains specific orthologs like drug resistance-related ABC transporter, this was unexpected considering ABC transporter is widespread among species [76]. However, it is known that this gene is under positive selection and show high mutation rates. That may explain the presence of KSP ABC transporter, not shared among other Protozoa species. The fact that hexose transporter, beta-fructofuranosidase, fructose-6-phosphate2-kinase, glycosyl transferase and glycosyl hydrolase are part of KSP, corroborate previous studies indicating that Kinetoplastida has a differentiated way to metabolize carbohydrates [27, 77]. The presence of RNA helicase, RNA methyltransferase and DEAD box helicase in ASP (table 8), could indicate a particular way of RNA processing by Apicomplexa, supporting previous studies that found proteins related to RNA processing [5, 11, 16]. Since then, DEAD box helicases are involved in various aspects of RNA metabolism and pre-mRNA splicing. Furthermore, the fact that Apicomplexa genes have introns [5], can explain the presence of this differentiated RNA processing machinery from others Protozoa. Kinetoplastida, for example, also have an unusual and specific RNA processing as: uninterrupted genes, most of them undergo trans-splicing and the presence of KSP-DEAD/DEAH box helicase, KSP-RNA helicase and KSP-mRNA capping methyltransferase [27, 34, 78] supporting our results. Among the ESP cyst wall-specific glycoprotein Jacob and several copies of Rab GTPases are part of these specifics orthologs. A type of Rab GTPases proteins is also present in *T. vaginalis*, however, these proteins do not have enough similarity to be grouped together in the same ortholog group, unlike the findings of Weedall & Hall [22] that found a large number of Rab GTPases encoded in common between *T. vaginalis* and *Entamoeba*. Finally, our findings show that: (i) 46.29% (1592/3396) of the KSP; (ii) 40.63% (92/224) of the ASP and (iii) 65.41% (2905/4441) of the ESP have hypothetical function (table 7). These results are new and we found this big number of Kinetoplastida specific orthologs, probably because they have very unusual characters, as: (i) specific mitochondrial RNA editing; (ii) mitochondrial DNA architecture unique; (iii) trans-splicing; (iv) genes into big polycistronic clusters; (v) uncommon nucleotides modifications; (vi) glycolysis compartmentalization [27, 78, 79]. Apicomplexa show specific structures, like rhoptries and micronemes forming the apical complex and they mediate locomotion, tissue, or cellular invasion, besides dense granules that are Apicomplexa specific [2, 5]. *Entamoeba* groups also have specific characteristics as: movement trough pseudopodia, lack of cilia, mitochondria, peroxisomes, and hydrogenosomes [2, 80]. Furthermore, *Entamoeba* shows another curious specificity: its cyst wall is made of chitin and we found chitin synthase only as ESP, no other studied Protozoa has the gene. Thus, the reasons enumerated above can explain the presence of these specific orthologs. However, this large number of hypothetical proteins was not totally unexpected, since these proteins are group-specific and there are no similar genes in other previously studied organisms. Therefore is not possible to transfer annotation by similarity to the KSP/ASP/ESP orthologs.

Phylogenomic analysis

As far as we know, our phylogenomic tree (fig. 15) is the first to use 348 concatenated proteins for Protozoa analysis, while Ocaña & Dávila (2011) used only 31 [45] and Deschamps used 64 Kinetoplastida concatenated genes [79]. *Plasmodium* genus showed to be a monophyletic group separated in two clades: rodent's parasites and human's parasites in both the 31-protein and 348-protein phylogenomic trees. As expected, *P. vivax* was closely related to *P. knowlesi* and *P. falciparum* [16], corroborating previous studies [45]. However the bootstrap support was higher in our phylogenomic tree. Piroplasmids formed a clade close to *Plasmodium* , also confirming Ocaña & Dávila [45], furthermore *Theileria* and *Babesia* formed together a clade corroborating the findings of Sato [7] and Brayton [8] who stated: "*Babesia* is Piroplasmid within the phylum Apicomplexa, similar to other members of this phylum, such as the phylogenetically closely positioned *Theileria".* This *Theileria-Babesia* clade was placed as outgroup of the *Plasmodium* genus in our tree, and the plastid-less *Cryptosporidium* genus and *T. gondii* (that form the coccids group) adopted a basal position related to the monophyletic Apicomplexa group. This result is not in agreement with Ocana & Dávila (2011) findings [45], that shown *Plasmodium* to be basal to the other Protozoa species. Since our tree using 348 proteins is more robust (even in terms of bootstrap support) than the one using only 31, then the scenario where *Cryptosporidium-Toxoplasma* is basal to Apicomplexa is probably more reliable. These same results further support the observations by Abrahamsen [5], about the life cycle of *C. parvum* being similar to that of other cyst-forming apicomplexans (e.g., *Eimeria* and *Toxoplasma*). Among the *Cryptosporidium species, C. muris* is considered a "gastric species", while *C. parvum/C. hominis* are named as "intestinal species" and are very close species, showing 95–98% identity at the DNA level [5, 10, 11]. As expected, our tree successfully reflected this separation, because *C. parvum* and *C. hominis* were placed in the same clade with high bootstrap value (100). As described in the literature, there are 13 species forming the Apicomplexa monophyletic group and our tree successfully reconstructed that scenario. This is not in agreement with Ocana and Dávila [45] tree that placed *Cryptosporidium* as a sister clade to *G. lamblia* violating the Apicomplexa monophyly. Analyzing the Kinetoplastida order in the tree, the expected separation between *Trypanosoma (T. bruc*ei*/T. cruzi) and Lei*shmani*a* *(L. braz*il*iensis / L. inf*antum */ L. maj*or) genus was properly inferred, corroborating Deschamps [79] and Ocaña and Dávila [45] trees. Inside the *Leishmania* genus, the correct separation between Old World (*L. major / L. infantum)* and New World Leishmani*a (L. braziliensis)* was properly inferred, corroborating Mauricio [81]. The inferred Kinetoplastida clade supports the expected monophyly of the order [45, 79]. It showed a 100 bootstrap support for all its internal branches, suggesting that a reliable tree was inferred. *Entamoeba spp.,* was inferred as a monophyletic group corroborating Ptáčková et al. [82], but its taxonomic position in the tree is unexpected since appears between Kinetoplastida and Apicomplexa, contrasting with Ocaña and Dávila [45] and Morrison findings [23]. Hopefully those two species should have been closer to two other amitochondrial Protozoa: *G. lamblia* and *T. va*gina*lis,* such as inferred by Ocaña and Dávila [45] and Morrison et al. [23], since *Entamoeba spp., G. lamblia* and *T. vaginalis* share a variety of metabolic adaptations [21]. This phylogenomic position of *Entamoeba*, is problematic. Our tree is similar to Bapeste et al [83] where *Giardia* is at the base of phylogenetic trees constructed of RNA or protein sequences and amoebae branch is in an intermediate position, not particularly early. Furthermore, if we root the tree closer to *Trichomonas* and *Giardia,* our tree is also similar to Baldauf [84], because when Excavata (*Trichomonas* and *Giardia*) are considered basal, Kinetoplastida stays closer to them, whereas *Entamoeba* branch appears in an intermediate position. Previous studies [2, 65] also show *Entamoeba* in an intermediate position at the tree, just like our philogenomic tree. However, as we can see, there is no consensus about *Entamoeba* position in the eukaryote tree. Cavalier Smith [2], for example, exclude Entamoebidae from the Archezoa (basal taxon where *Giardia* and *trichomonas* are classified) and place them in the Protozoa kingdom, in the subkingdom Dictyozoa, as a new phylum Entamoebia, whilst Ptáckova et al. [82] classify *Entamoeba* as archamoeba, a basal amoeba eukaryote that secondarily lost mitochondria. Finally, the last two species *G. lamblia* and *T. vaginalis* appear as outermost group compared to others, since vacuolar ATPase and elongation factor phylogenies identify *G. lamblia* as a basal eukaryote [23] and transcription machinery in amitochondriate *T. vaginalis* represents the earliest Eukaryotic lineages [85, 86] [85]. Furthermore, these orthologs were found in PCP, which might partially explain the basal position of this species. Finally these two species have a number of unusual characteristics, such as the mitochondria absence and fermentation metabolism enzymes [24, 26].

Dynamics of Superfamilies in Protozoa

Analyzing some gene families that are taxa specific or shared by all organisms, we found that VESA1 in *B. bovis* is not a homolog to PfEMP1 in *P. falciparum*, since VESA1 appears to be paralog only in *B. bovis* showing 132 copies. This is not in agreement with a previous study, between *B. bovis* and *P. falciparum,* [8] that found that PfEMP1, encoded by var gene in *P. falciparum*, is probably a homolog of VESA1 in *B. bovis*.

Dynamics of Cytoskeleton genes in the 22 Protozoa studied

Actin is an ubiquitous conserved protein considered essential in Eukaryotes, because it is involved in movement, morphology and trafficking [87]. It is a major cytoskeleton protein in Eukaryotic cells abundantly present in the nucleus and in the cytoplasm [88]. As expected, we found one actin in the PCP, which can be explained by the fact that actin is highly conserved (data now show) and playing an essential role in the structure and dynamics of most Eukaryotic cell [87, 89]. The presence of actin in Kinetoplastida has been corroborated in all species analyzed in this work. However, other studies have shown that *Leishmania* actin is a highly unconventional form of actin unlike other Eukaryotic actins and is associated with the kDNA [88, 90]. Furthermore, Cevallos et al. [89] found 12 loci annotated as actins, actin-like proteins or actin-related proteins. The presence of a diverse actins family in *T. cruzi* was unexpected. Nevertheless, our results corroborate and expand their findings because we found similar actin diversity not only in *T. cruzi*, but also in the five analyzed Kinetoplastida. In addition to PCP-actin, we found five KSP-actins, probably associated with kDNA [88, 90]. These five actins are probably KSP because they are associated with the kinetoplast. However, Kinetoplastida are not the only parasites that encode divergent actins, since the phylum Apicomplexa has some divergent actins too. In contrast to Skillman et al. [91] which found one actin in *T. gondii* and two in *P. falciparum*, we also found further than the PCP-actin, three ASP-actins orthologs. All Apicomplexa analyzed here have one actin into each ASP-actin ortholog, with the exception of *T. gondii* which shows two copies into each ASP-actin ortholog. Our study corroborates and expands Skillman´s findings [91], since we found divergent actins shared between all 13 apicomplexa species, not only between *T. gondii* and *P. falciparum*. We hypothesized that, by the fact of Apicomplexa moves itself through a unique form of gliding motility that is actin-dependent [91], this could explain the need of ASP-actin.

Other important genes that form the cytoskeleton are alpha-tubulin, beta-tubulin and gamma-tubulin [92, 93]. These three tubulins have been described as the basic components of microtubules, ubiquitous to all Eukaryotes and appear to be the minimal set of tubulins required by Eukaryotic cells, since they are crucial for the mitotic spindle apparatus, transport and motility [92–96]. Our results corroborate those previous studies, since we successfully found these three ubiquitous tubulins in the PCP. As expected these three genes are orthologs and shared among the 22 studied Protozoa. Despite these three Tubulins been conserved (data not shown), they are probably an ancient gene prior to the Protozoa diversification, since all analyzed species have this conserved ortholog. Moreover, other three tubulins were found: delta-tubulin, epsilon-tubulin and zeta-tubulin, comprising six known tubulin families. The identification of epsilon-tubulin and zeta-tubulin in *T. brucei* was curious, because Gull [97] found these two new tubulins after cloning the *T. brucei* delta-tubulin. These two new tubulins, epsilon and zeta, possess low similarity with other members of the family and appeared to be restricted to yeast, *Drosophila* and *Caenorhabditis elegans*. But Gull [97] has detected homologues in the genome databases of *Leishmania* and of *Plasmodium*. When we analyzed delta-tubulin and epsilon-tubulin, we achieved similar result than Gull, finding two delta-tubulin orthologs. The first delta-tubulin is an ortholog comprised by *Plasmodium* spp., *B. bovis*, *T. gondii* and *G. lamblia*, while the second is comprised by Kinetoplastida species and *T. vaginalis* sequences (confirmed by Interpro domain). We also found two epsilon-tubulin orthologs. One epsilon-tubulin ortholog is comprised by proteins from Kinetoplastida spp., *Plasmodium* spp., *T. vaginalis* and *G. lamblia*, confirmed by Interpro analyses, because the annotation of these sequences was incomplete, whereas the second epsilon-tubulin ortholog was found shared only by Piroplasmids and further function inference done using Interpro domain search too. In contrast to Dawson and Paredez [98] that state delta-tubulin and epsilon-tubulin are present only in excavates, our study shows that these tubulins are also present in Apicomplexa. Comparing our results with Wickstead and Gull [93], we also found delta-tubulin and epsilon-tubulin in *T. gondii*, *Plasmodium* spp., *T. vaginalis*, *G. lamblia* and Kinetoplastida spp. On the other hand, they did not find delta-tubulin and epsilon-tubulin in *B. bovis* and we were successful on finding these two tubulins in *B. bovis*. Furthermore, we found epsilon-tubulin as a specific-ortholog in *B. bovis*, *T. annulata* and *T. parva* unlike the Wickstead and Gull [93]. This is the first related work in the literature that described the presence of epsilon-tubulin in *Theileria* spp. The presence of delta-tubulin and epsilon-tubulin in these organisms are expected, since it is a characteristic of organisms with basal bodies and flagella. They are absent from organisms that have lost cilia/flagella, as in the case of *Entamoeba* spp. [36, 93], what can explain our findings. Finally, Zeta-tubulin is the last tubulin described and seems so divergent from the others five tubulins, that was found only in Kinetoplastida, confirming previous studies [36, 92, 94]. No study proposes a specific function for the zeta-tubulin, however Immunofluorescence and immunoelectron microscopy revealed that Zeta-tubulin is localized at the basal body region [99].

DNA processing

Searching for components of the major DNA repair processes in Eukaryotes, we found that the core of Eukaryotic nucleotide excision repair system is formed by four genes: Rad1, Rad2, Rad3 and Rad25, [6]. Nevertheless, in our study, unlike to Gardner [6], not all of the analyzed species has the core of Eukaryotic nucleotide excision repair system. We did not find RAD2 in *G. lamblia*, but found RAD2 as two orthologs: the first shared by *E. dispar*, *E. histolytica* and *T. vaginalis*; and the second shared by the other 18 species. Such result is not totally unexpected, because the amitochondrial protist: *G. lamblia* does not have some organelles and other features that are typical of Eukaryotes, like a complex DNA repair system, and depending on the gene analyzed is classified as basal Eukaryotes (Adam, 2000; Morrison et al., 2007). Furthermore, the fact that RAD2 is shared among *Entamoeba* spp. and *T. vaginalis* can be explained because they share a variety of adaptations and different evolutive pressure in their niches [18, 21]. The other three genes of the DNA repair processes: RAD25, RAD3 and RAD1 were found into PCP, as expected, since this is a crucial system for cellular maintenance.

Even sharing the core of Eukaryotic nucleotide excision repair system with all analyzed species, the Kinetoplastida DNA repair and initiation of replication are different from other Eukaryotes [35]. We found that DNA repair and initiation of replication in Kinetoplastida are also different from the other studied Protozoa, except in the core of Eukaryotic nucleotide excision repair system. Our results corroborate previous work [35], since we found some KSP related to initiation of replication: (i) DNA replication licensing factor; (ii) Eukaryotic translation initiation factor 4E; (iii) translation initiation factor 2 subunit; (iv) translation initiation factor IF-2; and (v) translation initiation factor eIF2B subunit delta. Furthermore, we found KSP related DNA repair: damage-specific DNA binding protein; SNF2 DNA repair protein; DNA repair protein BRCA2, mismatch repair protein MSH4; DNA repair and transcription factor protein; DNA repair and recombination protein, mitochondrial precursor and mitochondrial exoribonuclease DSS-1, which has no homologs with other Protozoa in this study. Ivens [34] suggest that the mechanisms regulating RNA polymerase II–directed transcription in the Tritryp are distinct from those operating in other Eukaryotes. In addition, we found that these mechanisms are distinct, not only in the Tritryp, but also in the five Kinetoplastida studied. Despite the presence of 3 orthologs in PCP related to RNA polymerase II–directed transcription: (i) RNA polymerase IIA largest subunit; (ii) DNA-directed RNA polymerases II; and (iii) DNA-directed RNA polymerase II subunit 3, we detected 2 KSP orthologs related DNA-directed RNA polymerase II. These KSP orthologs may reflect their unusual translational processes and unique mitochondrial DNA architecture [32, 79].

Metabolism Genes involved in energy production and apoproteins maturation

The metabolic gene pyruvate kinase, catalyze the last step of glycolysis forming pyruvate and ATP, and serve as a major regulator of glycolysis. There are some differences among pyruvate kinase from various species, for example: Kinetoplastida encode one pyruvate kinase, while *T. gondii* and *P. falciparum* encode a second enzyme localized in the apicoplast, in addition to the cytoplasmic form. The organisms acquire energy principally through the glycolysis, for example, in *C. parvum* the glycolytic pathway enzymes were found in the cytoplasm of the oocysts [100]. Additionally, Abrahamsen [5] found that pyruvate kinase, in *C. parvum,* is a "plant-like" enzyme that is either absent in or highly divergent from those typically found in mammals [100]. In our study, we found one pyruvate kinase shared by 22 Protozoa, including in this ortholog the possible *Cryptosporidium* "pyruvate kinase plant-like", which may suggest that this enzyme, in the 22 Protozoa studies, has "plant-like" origins. Furthermore, we found a second pyruvate kinase only in "apicomplexa having apicoplast". Cook [100] found this second enzyme only in *T. gondii* and *P. falciparum*, but we identify that this enzyme is also shared by *B. bovis*, *P. vivax*, *P. knowlesi*, *P. chabaudi*, *P. berghei*, *P. yoelii*, *T. gondii*, *T. annulata* and *T. parva*. In other words, this second enzyme is present in all "apicomplexa having apicoplast" studied and not in the *Cryptospodirium* plastid-less species, reinforcing the idea that cytoplasmic pyruvate kinase found in PCP is a pyruvate kinase with putative "plant-like" origins.

Metabolic genes involved in apoproteins maturation in mitochondria and plastids require Fe-S clusters. The assembly of Fe-S clusters formation in the *T. parva* mitochondrion appears to be similar to that in *Plasmodium*, but only sufS was identified in *T. parva* [6]. Additionally, we also identify, in *T. annulata,* sufS*,* which is a cysteine desulfurase that requires SufE for catalytic activity. Looking for Fe-S clusters proteins in all Protozoa, we identify one cysteine desulfurase, that belongs to Fe-S clusters, in the PCP; one cysteine desulfurase, common to *B. bovis*, *Plasmodium* spp., *T. gondii, T. parva* and *T. annulata,* but not found in *Cryptosporium* spp. since they are plastid-less and some of these Fe-S cluster genes are localized in apicoplast [7]. Moreover, *Plasmodium* spp*. and T. gondii* have sufD proteins, absent in *B. bovis* and *Theileria* spp; and sufC proteins were found in *P. falciparum, P. vivax, P. knowlesi, P. berghei, P. chabaudi* and *T. gondii*, also absent in *B. bovis, Theileria* spp and *P. yoelii*. These results corroborate previous work [7], which argues that like other plastid-bearing organisms, *Toxoplasma* and *Plasmodium* have the Suf system incorporating the SufBCD complex, while *Babesia* and *Theileria* lack genes specifying the components of the complex. Finally, we did not find Suf system in Kinetoplastida, probably because the mitochondrial essential FeS cluster synthesis, in those organisms occurs through the Isc pathway [101] .

The last metabolic gene described is the AdhE, a bifunctional dehydrogenase with acetaldehyde and alcohol dehydrogenase activities which first converts acetyl-CoA to acetaldehyde and then reduces the latter to ethanol. This AdhE predominantly occurs in Bacteria, but has been identified in *G. lamblia, E. dispar, E. histolytica* and *C. parvum* [5]. In addition, we also found this adhE in *C. hominis* and *C. muris*. Unexpectedly, our results show that this AdhE is an ortholog in *G. lamblia, Entamoeba* spp, *Cryptosporidium* spp., *Leishmania* spp., *Trypanosoma spp.* and *T. vaginalis.* Furthermore, we found that only alcohol dehydrogenase is an AdhE ortholog in *L. major*, *L. infantum*, *L. braziliensis*, *T. brucei, T. cruzi* and *T. vaginalis*. These organisms have alcohol dehydrogenase genes not fusioned to acetaldehyde dehydrogenase as AdhE. Possibly this AdhE was transferred independently in *G. lamblia, Entamoeba* spp and *Cryptosporidium* spp, since these organisms possess genes with alcohol dehydrogenase function only. Since we did not find acetaldehyde dehydrogenase in Kinetoplastida, and the alcohol dehydrogenase was ortholog to AdhE and not to other alcohol dehydrogenase, perhaps a gene fission followed by a gene loss could explain this result in these taxa.

Adaptation and resistance genes

An interesting adaptation gene is choline kinase (ChoK) whose activity is deregulated in transformed cell lines and its inhibition results in a reversible blockage of cell proliferation [6, 9]. In our results, only Apicomplexans species has ChoK proteins. This gene is ortholog in Apicomplexa, with exception of *B. bovis*, and is present in high copy number in *Theileria* genus when compared with other Apicomplexans. In this ortholog, we observed that *T. parva* has six ChoK copies, *T. annulata* seven copies and *Plasmodium* spp*., Cryptosporidium* spp*., T. gondii* only one copy each. In addition to this, one ChoK ortholog is specific to *Theileria* genus. The high ChoK copy number found in *Theileria spp.* may explain the fact that this pathogen is able to reversibly transform its host cells [6, 9], whilst other studied Protozoa do not posses this ability. Finally, is not clear how these parasite enzymes involved in lipid metabolism affect the host cell, however, none of the Theileria choline kinases have a recognizable signal peptide [102, 103].

Another interesting gene related to adaption found shared between *Entamoeba* and *T. vaginalis* was AIG1-like GTPases. These genes belongs to a large gene family, with unknown function, but their differential expression suggests that may be associated with virulence or adaptation [22]. We reported here an AIG1-like ortholog shared between *E. dispar, E. histolytica* and *T. vaginalis,* comprised by 18, 16 and 9 proteins, respectively. This result was unexpected, because: (i) a large number of Rab GTPases are shared between *T. vaginalis* and *Entamoba* spp., however we find AIG1-like GTPases shared among them; and (ii) we expected to find larger AIG1 expansions in *T. vaginalis*, not only 9 copies, since many gene families in this parasite have undergone a big expansion, and such gene family expansions are likely to improve an organism's adaptation to its environment [26]. Previous works report that AIG1 was similar to the AIG1 plant-like antibacterial protein, related to plant defense against Bacteria. Furthermore, AIG1 was also significantly up-regulated on *E. histolytica* during invasion of the mouse intestine, possibly acting as a defense mechanism against Bacteria present in the intestine [104] Therefore, these results suggest that *T. vaginalis* possibly has these AIG1 orthologous proteins to protect itself during invasion, acting as a defense mechanism against bacterial flora present in vagina.

Trehalose is a disaccharide whose major function is to protect against various stresses including: desiccation/dehydration, heat, cold, oxidation; and also may serve as an energy source in Coccids and Piroplasmids. Corroborating and expanding previous studies [6, 105], we detected that only Piroplasmids (*B. bovis*, *T. annulata* and *T. parva*) and Coccids (*C. muris*, *C. hominis*, *C. parvum* and *T. gondii*) encodes trehalose-6-phosphate synthase. None other Protozoa analyzed, in our study, possess the enzymes to manipulate trehalose. Probably, the trehalose protects the Piroplasmids parasites during its long sexual developmental cycle in the tick- vectors and act as a stress protective in Coccid´s oocyst stage, in the natural environment [6, 105]. Coccids and Piroplasmids using trehalose to protect against environmental adversity, while others Protozoa adopt different protective strategy.

In our study, we found one gene associated to chloroquine resistance (PfCRT). Unlike previous work [5] that found this PfCRT gene as ortholog between *P. falciparum* and *C. parvum,* we found that PfCRT is an ACP-ortholog shared among all the 13 analyzed Apicomplexa species. While it was a bit surprising to find this chloroquine resistance spread among Apicomplexa, previous work [106] show that chloroquine resistance in PfCRT gene is conferred through a mutation, K to T, at amino acid position 76 (K76T), reducing the choloroquine accumulation in the vacuolar food by accelerating efflux of choloroquine. Nevertheless, in our study, none PfCRT chloroquine resistance–linked gene analyzed carry the K76T mutation. Therefore the analyzed species are not chloroquine-resistant.

Comparing the orthologs shared between *Entamoeba* and *T. vaginalis,* we found, for example, BspA which are homologous to a bacterial fibronectin-binding protein (BspA of Bacteroides forsythus) and belong to a large gene family. It is a highly diverse family, including approximately 650 proteins characterized by the *Treponema pallidum* leucine-rich repeat. The unique other eukaryote known to encode BspA-like proteins, as *T. vaginalis*, is *E. histolytica* that contains 91 proteins. Furthermore, this BspA is described to occur in common as homolog among *E. histolytica*, *T. vaginalis* and the free-living *Tetrahymena thermophyla*, and was not described/found in *E. dispar* [22, 26]. However, in our study, unlike these previous works, we found two BspA-like orthologs shared among *E. histolytica*, *E. dispar* and *T. vaginalis*. The first ortholog shows one protein from each organism and the second BspA ortholog is formed by 10 copies from *E. histolytica*, 36 from *E. dispar* and 24 copies of the *T. vaginalis*. Additionally, four BspA-like orthologs were shared only between *E. dispar* and *E. hystolitica*, where each ortholog shows one BspA-like protein from each organism. One BspA was localized on the parasite surface [107] and its presence on surface suggests that these proteins are used to epithelial cell attachment and invasion by *Entamoeba* and *T. vaginalis*, as reported for *Tannerella forsythia* [108].

Conclusions

This is the first study that infers the orthologs, shared by the 22 Protozoa species, and proposes a proteomic core to them. Furthermore, this work intends to find the specific orthologs for Kinetoplastida species, Apicomplexa species and *Entamoeba* species.

The Protozoa Core Proteome is formed by 348 orthologous groups, shared by all 22 Protozoa studied species. We can observe that the Protozoa Core Proteome orthologs are generally related to the maintenance of the cell and information processing, because there were found ribosomal proteins, histones, cytoskeletal proteins and tRNA synthetase. Moreover, most of the genes are distributed in functional categories "J", "A" and "O", confirming the idea that these genes are more related to maintenance and processing, falling into two divisions: "Information Storage and Processing" and "Cellular Processes and Signaling". In PCP we did not find groups characterized as hypothetical and only 2.01% (7/348) of the groups belong to category “R” of KOG/NCBI. Furthermore, among 348 PCP, the informational genes (involved in transcription, translation replication, and repair) are closer to Archaea, and the operational genes (involved in cellular metabolic processes such as amino acid biosynthesis, cell envelope and lipid synthesis, metabolic enzymes, components of membranes) are closer to Bacteria.

Our phylogenomic analysis using 348 concatenated orthologous proteins from 22 Protozoa is the first report in literature. The generated tree has high confidence level, since the minimum bootstrap value among different clades was 95. Furthermore, the number of informative sites (328,228 positions) offers further confidence to this analysis, helping to understand the Protozoa evolution, and suggesting that *G. lamblia* and *T. vaginalis* are basal species.

The KPC has 5000 orthologs and 67.92% (3,396/5000) of those orthologs are Kinetoplastida Specific. Among them, it was noted proteins with specific functions, such as ABC transporter related to drug resistance and carbohydrate metabolism like hexose transporter; corroborating that Kinetoplastida has a different way of carbohydrate metabolism, although, 46.29% (1,592/3,396) of those orthologs are annotated as “hypothetical”.

The Apicomplexa Core proteome has 986 orthologous groups and 27.82% (224/986) are Apicomplexa Specific. Among these orthologs, it was observed that they have a particular way of DNA/RNA processing because there were found proteins related to them, such as: RNA helicase, RNA methyltransferase and histone binding proteins. Adittionally, 40.63% (92/224) of the Apicomplexa Specific orthologs were classified as hypothetical proteins.

The *Entamoeba* Core proteome has 5,915 orthologous groups and 75.08% (4,441/5,915) of these groups are *Entamoeba* Specific. This *Entamoeba* shows a particular way to control membrane traffics, due to the presence of a great number of Rab and Rab- GTPase proteins and its own way to defense, since we found AIG1 in these organisms. However, 65.41% (2,905/4441) of these specific orthologs were annotated as hypothetical or unknown function. We also found a considerable number of proteins annotated as hypothetical or with unknown function in the Kinetoplastida, Apicomplexa and *Entamoeba* Specific orthologs. These results are somewhat expected, due to the nature of this specific orthologs. Therefore, it is not possible to transfer annotation by similarity, since there are no similar proteins in other previously studied/sequenced organisms.

And finally, we hope contribute to the increase of information, knowledge, and understanding about Protozoa since inferring the similarities and differences among them facilitates a better understanding of their biology.

List of abbreviations

PCP: Protozoa Core Proteome
ACP: Apicomplexa Core Proteome
ECP Core Proteome
KCP Core Proteome
ASP Specific Proteome
ESP Specific Proteome
KSP Specific Proteome
COG Clusters of Orthologous Groups of Proteins
KOG: Eukaryotic Orthologous Genes
LGT: Lateral Gene Transfer
aaRSs: Aminoacyl-tRNA synthetases
TyrRS: Tyrosyl tRNA-synthetase
GPI: Glucose-6-phosphate isomerase
kDNA: Kinetoplast DNA
ChoK: choline kinase

Competing interests

Authors' contributions

DAT: analyzed data and write the manuscript. RJ: helps with analysis and revised the manuscript. SS, MM, MLMC and AMRD: revised the manuscript and helps with scientific questions JL and RC: revised the manuscript and help with venn diagram analysis.

Acknowledgements

To CAPES, FAPERJ and Fiocruz for the financial support.

References

1. Imam T: **The complexities in the classification of protozoa: a challenge to parasitologists**. *Bayero J Pure Appl Sci* 2009, **2**:159–164.

2. Cavalier-Smith T: *Kingdom Protozoa and Its 18 Phyla. Volume 57*; 1993:953–94.

3. Cavalier-Smith T: **Deep phylogeny, ancestral groups and the four ages of life.** *Philos Trans R Soc Lond B Biol Sci* 2010, **365**:111–32.

4. Cavalier-Smith T: **Predation and eukaryote cell origins: a coevolutionary perspective.** *Int J Biochem Cell Biol* 2009, **41**:307–22.

5. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto C a, Deng M, Liu C, Widmer G, Tzipori S, Buck G a, Xu P, Bankier AT, Dear PH, Konfortov B a, Spriggs HF, Iyer L, Anantharaman V, Aravind L, Kapur V: **Complete genome sequence of the apicomplexan, Cryptosporidium parvum.** *Science* 2004, **304**:441–5.

6. Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM, Hall N, Ren Q, Paulsen IT, Pain A, Berriman M, Wilson RJM, Sato S, Ralph S a, Mann DJ, Xiong Z, Shallom SJ, Weidman J, Jiang L, Lynn J, Weaver B, Shoaibi A, Domingo AR, Wasawo D, Crabtree J, Wortman JR, Haas B, Angiuoli S V, Creasy TH, Lu C, Suh B, et al.: **Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes.** *Science* 2005, **309**:134–7.

7. Sato S: **The apicomplexan plastid and its evolution.** *Cell Mol Life Sci* 2011, **68**:1285–96.

8. Brayton K a, Lau AOT, Herndon DR, Hannick L, Kappmeyer LS, Berens SJ, Bidwell SL, Brown WC, Crabtree J, Fadrosh D, Feldblum T, Forberger H a, Haas BJ, Howell JM, Khouri H, Koo H, Mann DJ, Norimine J, Paulsen IT, Radune D, Ren Q, Smith RK, Suarez CE, White O, Wortman JR, Knowles DP, McElwain TF, Nene VM: **Genome sequence of Babesia bovis and comparative analysis of apicomplexan hemoprotozoa.** *PLoS Pathog* 2007, **3**:1401–13.

9. Pain A, Renauld H, Berriman M, Murphy L, Yeats C a, Weir W, Kerhornou A, Aslett M, Bishop R, Bouchier C, Cochet M, Coulson RMR, Cronin A, de Villiers EP, Fraser A, Fosker N, Gardner M, Goble A, Griffiths-Jones S, Harris DE, Katzer F, Larke N,

Lord A, Maser P, McKellar S, Mooney P, Morton F, Nene V, O'Neil S, Price C, et al.: **Genome of the host-cell transforming parasite Theileria annulata compared with T. parva.** *Science* 2005, **309**:131–3.

10. G. Widmer, E. London, L. Zhang, G. Ge, S. Tzipori, J. Carlton J da S: *Preliminary Analysis of the Cryptosporidium Muris Genome*. CAB International; 2009:320–327.

11. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, Pearson WR, Dear PH, Bankier AT, Peterson DL, Abrahamsen MS, Kapur V, Tzipori S, Buck GA: **The genome of Cryptosporidium hominis**. *Science (80-)* 2004, **431**(October):1107–1112.

12. Weiss LM, Dubey JP: **Toxoplasmosis: A history of clinical observations.** *Int J Parasitol* 2009, **39**:895–901.

13. Elmore S a, Jones JL, Conrad P a, Patton S, Lindsay DS, Dubey JP: **Toxoplasma gondii: epidemiology, feline clinical aspects, and prevention.** *Trends Parasitol* 2010, **26**:190–6.

14. Carlton JM, Angiuoli S V, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, Peterson JD, Pop M, Kosack DS, Shumway MF, Bidwell SL, Shallom SJ, van Aken SE, Riedmuller SB, Feldblyum T V, Cho JK, Quackenbush J, Sedegah M, Shoaibi A, Cummings LM, Florens L, Yates JR, Raine JD, Sinden RE, Harris M a, Cunningham D a, et al.: **Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii.** *Nature* 2002, **419**:512–9.

15. 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 J a, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan M-S, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM a, et al.: **Genome sequence of the human malaria parasite Plasmodium falciparum.** *Nature* 2002, **419**:498–511.

16. Hall N, Carlton J: **Comparative genomics of malaria parasites.** *Curr Opin Genet Dev* 2005, **15**:609–13.

17. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, Crabtree J, Angiuoli S V, Merino EF, Amedeo P, Cheng Q, Coulson RMR, Crabb BS, Del Portillo HA, Essien K, Feldblyum T V, Fernandez-Becerra C, Gilson PR, Gueye AH, Guo X, Kang'a S, Kooij TWA, Korsinczky M, Meyer EV-S, Nene V, Paulsen I, White O, Ralph SA, Ren Q, Sargeant TJ, et al.: **Comparative genomics of the neglected human malaria parasite Plasmodium vivax.** *Nature* 2008, **455**:757–63.

18. Carlton J: **The Plasmodium vivax genome sequencing project.** *Trends Parasitol* 2003, **19**:227–31.

19. Pain a, Böhme U, Berry a E, Mungall K, Finn RD, Jackson a P, Mourier T, Mistry J, Pasini EM, Aslett M a, Balasubrammaniam S, Borgwardt K, Brooks K, Carret C, Carver TJ, Cherevach I, Chillingworth T, Clark TG, Galinski MR, Hall N, Harper D, Harris D, Hauser H, Ivens a, Janssen CS, Keane T, Larke N, Lapp S, Marti M, Moule

S, et al.: **The genome of the simian and human malaria parasite Plasmodium knowlesi.** *Nature* 2008, **455**:799–803.

20. Lorenzi H a, Puiu D, Miller JR, Brinkac LM, Amedeo P, Hall N, Caler E V: **New assembly, reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features and protein content information.** *PLoS Negl Trop Dis* 2010, **4**:e716.

21. Loftus B, Anderson I, Davies R, Alsmark UCM, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ, Nozaki T, Suh B, Pop M, Duchene M, Ackers J, Tannich E, Leippe M, Hofer M, Bruchhaus I, Willhoeft U, Bhattacharya A, Chillingworth T, Churcher C, Hance Z, Harris B, Harris D, Jagels K, Moule S, Mungall K, Ormond D, et al.: **The genome of the protist parasite Entamoeba histolytica.** *Nature* 2005, **433**:865–8.

22. Weedall GD, Hall N: **Evolutionary genomics of Entamoeba.** *Res Microbiol* 2011:1–9.

23. Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, Best A a, Cande WZ, Chen F, Cipriano MJ, Davids BJ, Dawson SC, Elmendorf HG, Hehl AB, Holder ME, Huse SM, Kim UU, Lasek-Nesselquist E, Manning G, Nigam A, Nixon JEJ, Palm D, Passamaneck NE, Prabhu A, Reich CI, Reiner DS, Samuelson J, Svard SG, Sogin ML: **Genomic minimalism in the early diverging intestinal parasite Giardia lamblia.** *Science* 2007, **317**:1921–6.

24. Adam RD: **The Giardia lamblia genome.** *Int J Parasitol* 2000, **30**:475–84.

25. Thompson RCA: **The Impact of Giardia on Science and Society**. In *GIARDIA CRYPTOSPORIDIUM From Mol to Dis*. Edited by Ortega-Pierres, Guadalupe; Cacciò, Simone M; Fayer, Ronald; Mank, Theo G; Smith, Huw V; Thompson R. CAB International, 2009; 2009:1–11.

26. Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao Q, Wortman JR, Bidwell SL, Alsmark UCM, Besteiro S, Sicheritz-Ponten T, Noel CJ, Dacks JB, Foster PG, Simillion C, Van de Peer Y, Miranda-Saavedra D, Barton GJ, Westrop GD, Müller S, Dessi D, Fiori PL, Ren Q, Paulsen I, Zhang H, Bastida-Corcuera FD, Simoes-Barbosa A, Brown MT, Hayes RD, Mukherjee M, et al.: **Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis.** *Science* 2007, **315**:207–12.

27. Simpson AGB, Stevens JR, Lukes J: **The evolution and diversity of kinetoplastid flagellates.** *Trends Parasitol* 2006, **22**:168–74.

28. Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S, James TY, Karpov S, Kugrens P, Krug J, Lane CE, Lewis LA, Lodge J, Lynn DH, Mann DG, McCourt RM, Mendoza L, Moestrup O, Mozley-Standridge SE, Nerad TA, Shearer CA, Smirnov A V, Spiegel FW, Taylor MFJR: **The new higher level classification of eukaryotes with emphasis on the taxonomy of protists.** *J Eukaryot Microbiol* 2005, **52**:399–451.

29. Barrett MP, Burchmore RJS, Stich A, Lazzari JO, Frasch AC, Cazzulo JJ, Krishna S: **The trypanosomiases.** *Lancet* 2003, **362**:1469–80.

30. Grimaldi G, Tesh RB: **Leishmaniases of the New World: current concepts and implications for future research.** *Clin Microbiol Rev* 1993, **6**:230–50.

31. Pays E, Vanhamme L, Pérez-Morga D: **Antigenic variation in Trypanosoma brucei: facts, challenges and mysteries.** *Curr Opin Microbiol* 2004, **7**:369–74.

32. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey E a, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu DC, Haas BJ, Tran A-N, Wortman JR, Alsmark UCM, Angiuoli S, Anupama A, Badger J, Bringaud F, Cadag E, Carlton JM, Cerqueira GC, Creasy T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC, et al.: **Comparative genomics of trypanosomatid parasitic protozoa.** *Science* 2005, **309**:404–9.

33. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, Kerhornou A, Ivens A, Fraser A, Rajandream M-A, Carver T, Norbertczak H, Chillingworth T, Hance Z, Jagels K, Moule S, Ormond D, Rutter S, Squares R, Whitehead S, Rabbinowitsch E, Arrowsmith C, White B, Thurston S, Bringaud F, Baldauf SL, et al.: **Comparative genomic analysis of three Leishmania species that cause diverse human disease.** *Nat Genet* 2007, **39**:839–47.

34. Ivens AC, Peacock CS, Worthey E a, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream M-A, Adlem E, Aert R, Anupama A, Apostolou Z, Attipoe P, Bason N, Bauser C, Beck A, Beverley SM, Bianchettin G, Borzym K, Bothe G, Bruschi C V, Collins M, Cadag E, Ciarloni L, Clayton C, Coulson RMR, Cronin A, Cruz AK, Davies RM, De Gaudenzi J, et al.: **The genome of the kinetoplastid parasite, Leishmania major.** *Science* 2005, **309**:436–42.

35. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran A-N, Ghedin E, Worthey E a, Delcher AL, Blandin G, Westenberger SJ, Caler E, Cerqueira GC, Branche C, Haas B, Anupama A, Arner E, Aslund L, Attipoe P, Bontempi E, Bringaud F, Burton P, Cadag E, Campbell D a, Carrington M, Crabtree J, Darban H, da Silveira JF, de Jong P, Edwards K, et al.: **The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease.** *Science* 2005, **309**:409–15.

36. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Böhme U, Hannick L, Aslett M a, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UCM, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth T-J, Churcher C, Clark LN, Corton CH, Cronin A, et al.: **The genome of the African trypanosome Trypanosoma brucei.** *Science* 2005, **309**:416–22.

37. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22**:1658–9.

38. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**:2178–89.

39. Coutinho F, Ogasawara E, Oliveira D, Braganholo V, Lima AAB, Dávila AMR, Mattoso M: **Many task computing for orthologous genes identification in protozoan genomes using Hydra**. *Concurr Comput Pract Exp* 2011:n/a–n/a.

40. Dávila AMR, Mendes PN, Wagner G, Tschoeke D a, Cuadrat RRC, Liberman F, Matos L, Satake T, Ocaña K a CS, Triana O, Cruz SMS, Jucá HCL, Cury JC, Silva FN, Geronimo G a, Ruiz M, Ruback E, Silva FP, Probst CM, Grisard EC, Krieger M a, Goldenberg S, Cavalcanti MCR, Moraes MO, Campos MLM, Mattoso M: **ProtozoaDB: dynamic visualization and exploration of protozoan genomes.** *Nucleic Acids Res* 2008, **36**(Database issue):D547–52.

41. R Core Team: **R: A Language and Environment for Statistical Computing**. 2013.

42. Altschul SF, Madden TL, Schäffer a a, 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, **25**:3389–402.

43. Tatusov RL: **A Genomic Perspective on Protein Families**. *Science (80-)* 1997, **278**:631–637.

44. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, others: **The COG database: an updated version includes eukaryotes**. *BMC Bioinformatics* 2003, **4**:41.

45. Ocaña K a CS, Dávila AMR: **Phylogenomics-based reconstruction of protozoan species tree.** *Evol Bioinform Online* 2011, **7**:107–21.

46. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Res* 2005, **33**:511–8.

47. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28**:2731–9.

48. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life.** *Science* 2006, **311**:1283–7.

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

50. Mu J, Myers R a, Jiang H, Liu S, Ricklefs S, Waisberg M, Chotivanich K, Wilairatana P, Krudsood S, White NJ, Udomsangpetch R, Cui L, Ho M, Ou F, Li H, Song J, Li G, Wang X, Seila S, Sokunthea S, Socheat D, Sturdevant DE, Porcella SF, Fairhurst RM, Wellems TE, Awadalla P, Su X: **Plasmodium falciparum genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs.** *Nat Genet* 2010, **42**:268–71.

51. Brown JR: **Ancient horizontal gene transfer.** *Nat Rev Genet* 2003, **4**:121–32.

52. Brown JR, Doolittle WF: **Gene descent, duplication, and horizontal transfer in the evolution of glutamyl- and glutaminyl-tRNA synthetases.** *J Mol Evol* 1999, **49**:485–95.

53. Hamon MA, Cossart P: **Histone modifications and chromatin remodeling during bacterial infections.** *Cell Host Microbe* 2008, **4**:100–9.

54. Sandman K, Hallam SJ, Delong EF, Reeve JN: **Histones in Crenarchaea**. 2005, **187**:5482–5485.

55. Cavalier-Smith T: **Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree.** *Biol Lett* 2010, **6**:342–5.

56. Alsmark C, Foster PG, Sicheritz-Ponten T, Nakjang S, Martin Embley T, Hirt RP: **Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes.** *Genome Biol* 2013, **14**:R19.

57. Opperdoes FR, Michels PAM: **Horizontal gene transfer in trypanosomatids**. *Trends Parasitol* 2007, **23**.

58. Liu W, Fang L, Li M, Li S, Guo S, Luo R, Feng Z, Li B, Zhou Z, Shao G, Chen H, Xiao S: **Comparative Genomics of Mycoplasma: Analysis of Conserved Essential Genes and Diversity of the Pan-Genome**. *PLoS One* 2012, **7**:e35698.

59. Leipe DD, Aravind L, Koonin E V: **Did DNA replication evolve twice independently?** *Nucleic Acids Res* 1999, **27**:3389–401.

60. Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, Henze K, Kretschmann E, Richly E, Leister D, Bryant D, Steel M a, Lockhart PJ, Penny D, Martin W: **A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes.** *Mol Biol Evol* 2004, **21**:1643–60.

61. Rivera MC, Lake J a: **The ring of life provides evidence for a genome fusion origin of eukaryotes.** *Nature* 2004, **431**:152–5.

62. Yutin N, Makarova KS, Mekhedov SL, Wolf YI, Koonin E V: **The deep archaeal roots of eukaryotes.** *Mol Biol Evol* 2008, **25**:1619–30.

63. Peretó J, López-García P, Moreira D: **Ancestral lipid biosynthesis and early membrane evolution.** *Trends Biochem Sci* 2004, **29**:469–77.

64. Lester L, Meade A, Pagel M: **The slow road to the eukaryotic genome.** *Bioessays* 2006, **28**:57–64.

65. Cavalier-Smith T: **The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa.** *Int J Syst Evol Microbiol* 2002, **52**(Pt 2):297–354.

66. Embley TM, Martin W: **Eukaryotic evolution, changes and challenges.** *Nature* 2006, **440**:623–30.

67. Brindefalk B, Viklund J, Larsson D, Thollesson M, Andersson SGE: **Origin and evolution of the mitochondrial aminoacyl-tRNA synthetases.** *Mol Biol Evol* 2007, **24**:743–56.

68. Bonnefond L, Giegé R, Rudinger-Thirion J: **Evolution of the tRNA(Tyr)/TyrRS aminoacylation systems.** *Biochimie* 2005, **87**:873–83.

69. Larson ET, Kim JE, Castaneda LJ, Napuli AJ, Zhang Z, Fan E, Zucker FH, Verlinde CLMJ, Buckner FS, Van Voorhis WC, Hol WGJ, Merritt E a: **The double-length tyrosyl-tRNA synthetase from the eukaryote Leishmania major forms an intrinsically asymmetric pseudo-dimer.** *J Mol Biol* 2011, **409**:159–76.

70. Arakaki TL, Carter M, Napuli AJ, Verlinde CLMJ, Fan E, Zucker F, Buckner FS, Van Voorhis WC, Hol WGJ, Merritt E a: **The structure of tryptophanyl-tRNA synthetase from Giardia lamblia reveals divergence from eukaryotic homologs.** *J Struct Biol* 2010, **171**:238–43.

71. Broutin H, Tarrieu F, Tibayrenc M, Oury B, Barnabé C: **Phylogenetic analysis of the glucose-6-phosphate isomerase gene in Trypanosoma cruzi.** *Exp Parasitol* 2006, **113**:1–7.

72. Grauvogel C, Brinkmann H, Petersen J: **Evolution of the glucose-6-phosphate isomerase: the plasticity of primary metabolism in photosynthetic eukaryotes.** *Mol Biol Evol* 2007, **24**:1611–21.

73. Koonin E V, Galperin MY: *Sequence - Evolution - Function : Computational Approaches in Comparative Genomics*. Boston, Dordrecht, London: Kluwer academic publishers; 2002.

74. Hansen T, Oehlmann M, Scho P, Kiel D-: **Novel Type of Glucose-6-Phosphate Isomerase in the Hyperthermophilic Archaeon Pyrococcus furiosus**. 2001, **183**:3428–3435.

75. Verhees CH, Huynen M a, Ward DE, Schiltz E, de Vos WM, van der Oost J: **The phosphoglucose isomerase from the hyperthermophilic archaeon Pyrococcus furiosus is a unique glycolytic enzyme that belongs to the cupin superfamily.** *J Biol Chem* 2001, **276**:40926–32.

76. Sauvage V, Aubert D, Escotte-Binet S, Villena I: **The role of ATP-binding cassette (ABC) proteins in protozoan parasites.** *Mol Biochem Parasitol* 2009, **167**:81–94.

77. Cáceres AJ, Quiñones W, Gualdrón M, Cordeiro A, Avilán L, Michels P a M, Concepción JL: **Molecular and biochemical characterization of novel glucokinases from Trypanosoma cruzi and Leishmania spp.** *Mol Biochem Parasitol* 2007, **156**:235–45.

78. Simpson AGB, Lukes J, Roger AJ: **The evolutionary history of kinetoplastids and their kinetoplasts.** *Mol Biol Evol* 2002, **19**:2071–83.

79. Deschamps P, Lara E, Marande W, López-García P, Ekelund F, Moreira D: **Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids.** *Mol Biol Evol* 2011, **28**:53–8.

80. Clark CG, Roger a J: **Direct evidence for secondary loss of mitochondria in Entamoeba histolytica.** *Proc Natl Acad Sci U S A* 1995, **92**:6518–21.

81. Mauricio IL, Gaunt MW, Stothard JR, Miles MA: **Glycoprotein 63 (gp63) genes show gene conversion and reveal the evolution of Old World Leishmania.** *Int J Parasitol* 2007, **37**:565–76.

82. Ptáčková E, Kostygov AY, Chistyakova L V, Falteisek L, Frolov AO, Patterson DJ, Walker G, Cepicka I: **Evolution of Archamoebae: morphological and molecular evidence for pelobionts including Rhizomastix, Entamoeba, Iodamoeba, and Endolimax.** *Protist* 2013, **164**:380–410.

83. Bapteste E, Brinkmann H, Lee J a, Moore D V, Sensen CW, Gordon P, Duruflé L, Gaasterland T, Lopez P, Müller M, Philippe H: **The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba.** *Proc Natl Acad Sci U S A* 2002, **99**:1414–9.

84. Baldauf SL: **The deep roots of eukaryotes.** *Science* 2003, **300**:1703–6.

85. Smith A, Johnson P: **Gene expression in the unicellular eukaryote Trichomonas vaginalis.** *Res Microbiol* 2011, **162**:646–54.

86. Gerbod D, Edgcomb VP, Noël C, Delgado-Viscogliosi P, Viscogliosi E: **Phylogenetic position of parabasalid symbionts from the termite Calotermes flavicollis based on small subunit rRNA sequences.** *Int Microbiol* 2000, **3**:165–72.

87. García-Salcedo J a, Pérez-Morga D, Gijón P, Dilbeck V, Pays E, Nolan DP: **A differential role for actin during the life cycle of Trypanosoma brucei.** *EMBO J* 2004, **23**:780–9.

88. Kapoor P, Kumar A, Naik R, Ganguli M, Siddiqi MI, Sahasrabuddhe A a, Gupta CM: **Leishmania actin binds and nicks kDNA as well as inhibits decatenation activity of type II topoisomerase.** *Nucleic Acids Res* 2010, **38**:3308–17.

89. Cevallos AM, Segura-Kato YX, Merchant-Larios H, Manning-Cela R, Alberto Hernández-Osorio L, Márquez-Dueñas C, Ambrosio JR, Reynoso-Ducoing O, Hernández R: **Trypanosoma cruzi: multiple actin isovariants are observed along different developmental stages.** *Exp Parasitol* 2011, **127**:249–59.

90. De Melo LDB, Sant'Anna C, Reis S a, Lourenço D, De Souza W, Lopes UG, Cunha-e-Silva NL: **Evolutionary conservation of actin-binding proteins in Trypanosoma cruzi and unusual subcellular localization of the actin homologue.** *Parasitology* 2008, **135**:955–65.

91. Skillman KM, Diraviyam K, Khan A, Tang K, Sept D, Sibley LD: **Evolutionarily divergent, unstable filamentous actin is essential for gliding motility in apicomplexan parasites.** *PLoS Pathog* 2011, **7**:e1002280.

92. Gull K: **The cell biology of parasitism in Trypanosoma brucei: insights and drug targets from genomic approaches?** *Curr Pharm Des* 2002, **8**:241–56.

93. Wickstead B, Gull K: **The evolution of the cytoskeleton.** *J Cell Biol* 2011, **194**:513–25.

94. Kumar V, Sharma R, Trivedi PC, Vyas GK, Khandelwal V: **Review article Traditional and novel references towards systematic normalization of qRT-PCR data in plants**. *Aust J Crop Sci* 2011, **5**:1455–1468.

95. Ma C, Tran J, Gu F, Ochoa R, Li C, Sept D, Werbovetz K, Morrissette N: **Dinitroaniline activity in Toxoplasma gondii expressing wild-type or mutant alpha-tubulin.** *Antimicrob Agents Chemother* 2010, **54**:1453–60.

96. Luis L, Serrano ML, Hidalgo M, Mendoza-León A: **Comparative Analyses of the β -Tubulin Gene and Molecular Modeling Reveal Molecular Insight into the Colchicine Resistance in Kinetoplastids Organisms.** *Biomed Res Int* 2013, **2013**:843748.

97. Gull K: **The biology of kinetoplastid parasites: insights and challenges from genomics and post-genomics.** *Int J Parasitol* 2001, **31**:443–52.

98. Dawson SC, Paredez AR: **Alternative cytoskeletal landscapes: cytoskeletal novelty and evolution in basal excavate protists.** *Curr Opin Cell Biol* 2013, **25**:134–41.

99. McKean PG, Vaughan S, Gull K: **The extended tubulin superfamily.** *J Cell Sci* 2001, **114**(Pt 15):2723–33.

100. Cook WJ, Senkovich O, Aleem K, Chattopadhyay D: **Crystal structure of Cryptosporidium parvum pyruvate kinase.** *PLoS One* 2012, **7**:e46875.

101. Gisselberg JE, Dellibovi-Ragheb T a, Matthews K a, Bosch G, Prigge ST: **The suf iron-sulfur cluster synthesis pathway is required for apicoplast maintenance in malaria parasites.** *PLoS Pathog* 2013, **9**:e1003655.

102. Wasmuth J, Daub J: **The origins of apicomplexan sequence innovation**. *Genome …* 2009:1202–1213.

103. Shiels B, Langsley G, Weir W, Pain A, McKellar S, Dobbelaere D: **Alteration of host cell phenotype by Theileria annulata and Theileria parva: mining for manipulators in the parasite genomes.** *Int J Parasitol* 2006, **36**:9–21.

104. Gilchrist C a, Houpt E, Trapaidze N, Fei Z, Crasta O, Asgharpour A, Evans C, Martino-Catt S, Baba DJ, Stroup S, Hamano S, Ehrenkaufer G, Okada M, Singh U, Nozaki T, Mann BJ, Petri W a: **Impact of intestinal colonization and invasion on the Entamoeba histolytica transcriptome.** *Mol Biochem Parasitol* 2006, **147**:163–76.

105. Yu Y, Zhang H, Zhu G: **Plant-type trehalose synthetic pathway in cryptosporidium and some other apicomplexans.** *PLoS One* 2010, **5**:e12593.

106. Takahashi N, Tanabe K, Tsukahara T, Dzodzomenyo M, Dysoley L, Khamlome B, Sattabongkot J, Nakamura M, Sakurai M, Kobayashi J, Kaneko A, Endo H, Hombhanje F, Tsuboi T, Mita T: **Large-scale survey for novel genotypes of Plasmodium falciparum chloroquine-resistance gene pfcrt.** *Malar J* 2012, **11**:92.

107. Davis PH, Zhang Z, Chen M, Zhang X, Chakraborty S, Stanley SL: **Identification of a family of BspA like surface proteins of Entamoeba histolytica with novel leucine rich repeats.** *Mol Biochem Parasitol* 2006, **145**:111–6.

108. Inagaki S, Onishi S, Kuramitsu HK, Sharma A: **Porphyromonas gingivalis vesicles enhance attachment, and the leucine-rich repeat BspA protein is required for invasion of epithelial cells by “Tannerella forsythia”.** *Infect Immun* 2006, **74**:5023–8.

109. Carlton J, Silva J, Hall N: **The genome of model malaria parasites, and comparative genomics.** *Curr Issues Mol Biol* 2005, **7**:23–37.

110. Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, Heiges M, Iodice J, Kissinger JC, Mackey AJ, Pinney DF, Roos DS, Stoeckert CJ, Wang H, Brunk BP: **ToxoDB: an integrated Toxoplasma gondii database resource.** *Nucleic Acids Res* 2008, **36**(Database issue):D553–6.

Figures legend

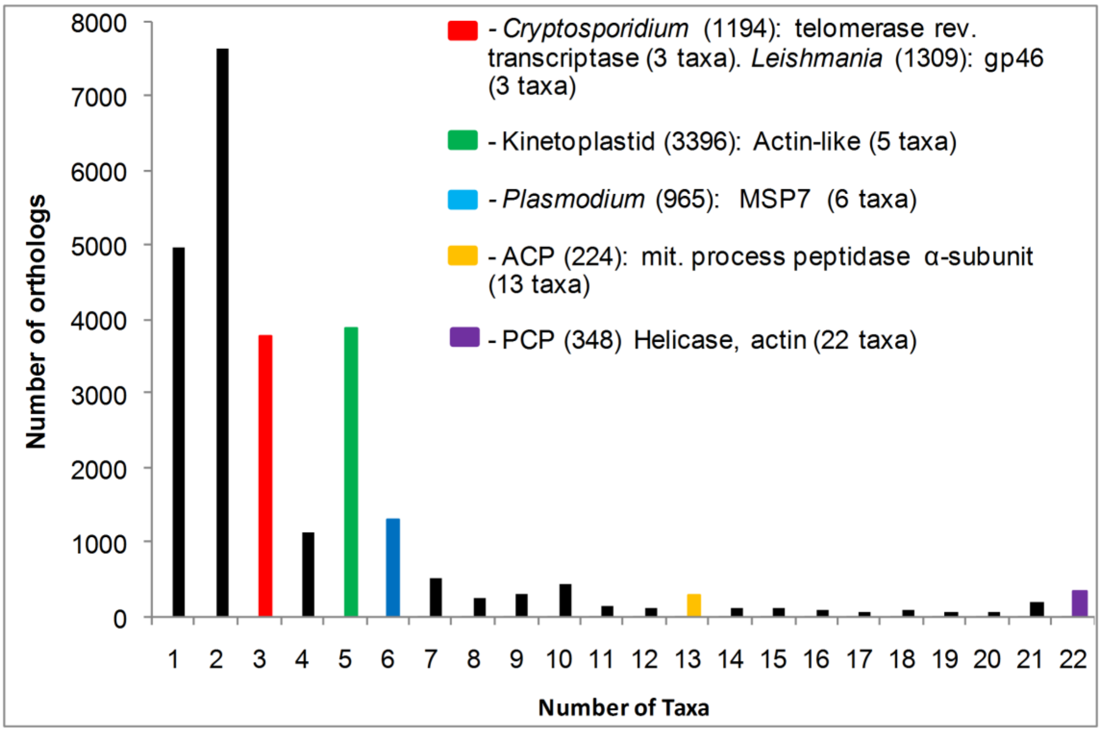


Figure 1: Homologs groups (N = 26,101 groups). Homologs groups number generated with OrthoMCL using 204,624 proteins within 22 organisms included in this study. The names, colors and numbers represents genus name, function and orthologs numbers shared between species.

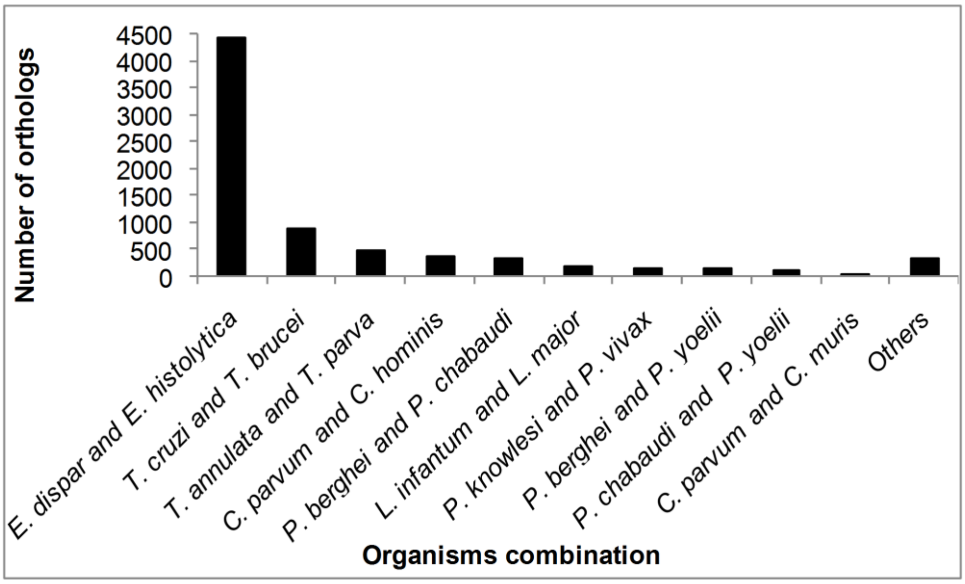


Figure 2: Orthologous groups shared by only two species (N = 7628 groups). Orthologous groups number created by OrthoMCL, for the 7628 orthologs groups shared only for 2 taxa. We observe the respective species combinations of the most abundant groups found, shared only by two taxa.

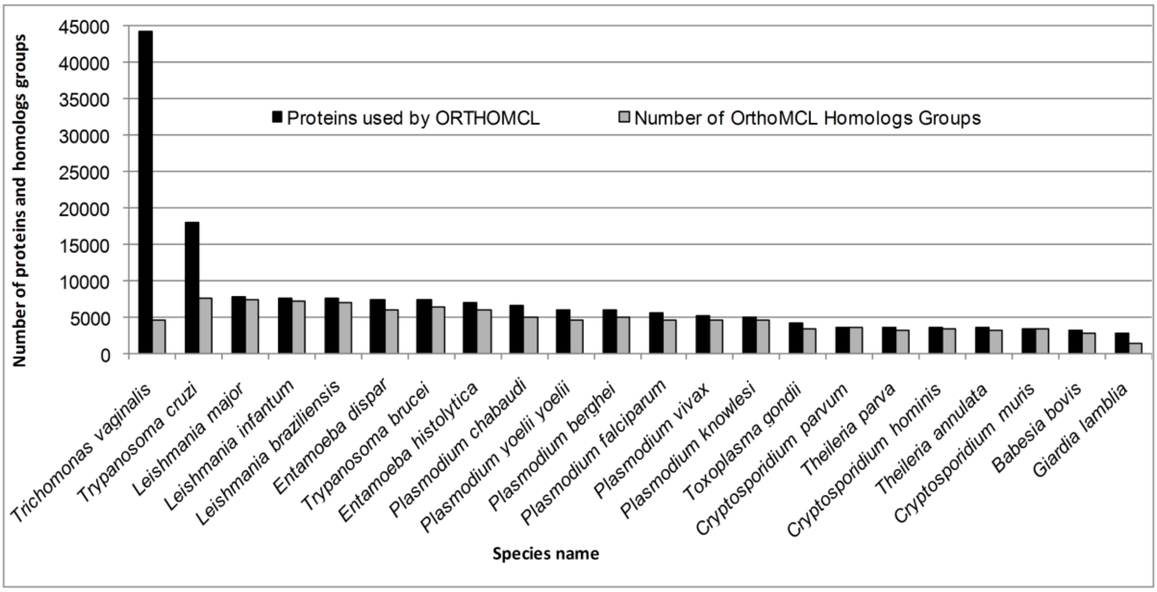


Figure 3. Number of proteins grouped (N = 171,096 proteins) and homologs groups created (N = 26,101) by OrthoMCL. Number of Homologs groups formed by OrthoMCL for each analyzed species. Number of proteins grouped (clusterized) by OrthoMCL for each organism into homologs groups.

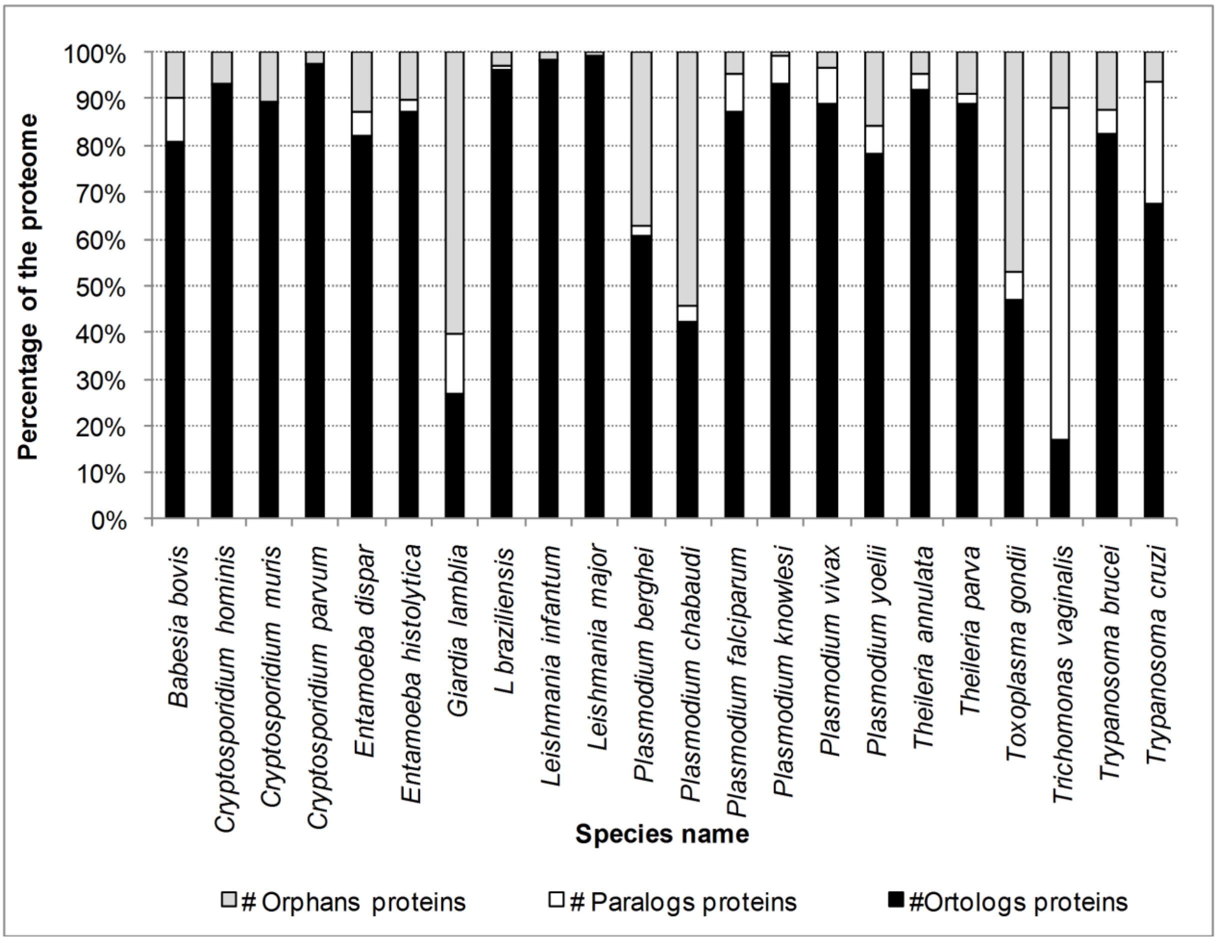


Figure 4: Relationship between homologs groups and proteins for each analyzed species. Percentage of proteins not grouped (considered orphans), percentage of paralogs proteins (inside paralogs groups) and percentage of orthologs proteins (inside orthologs groups) for each studied specie.

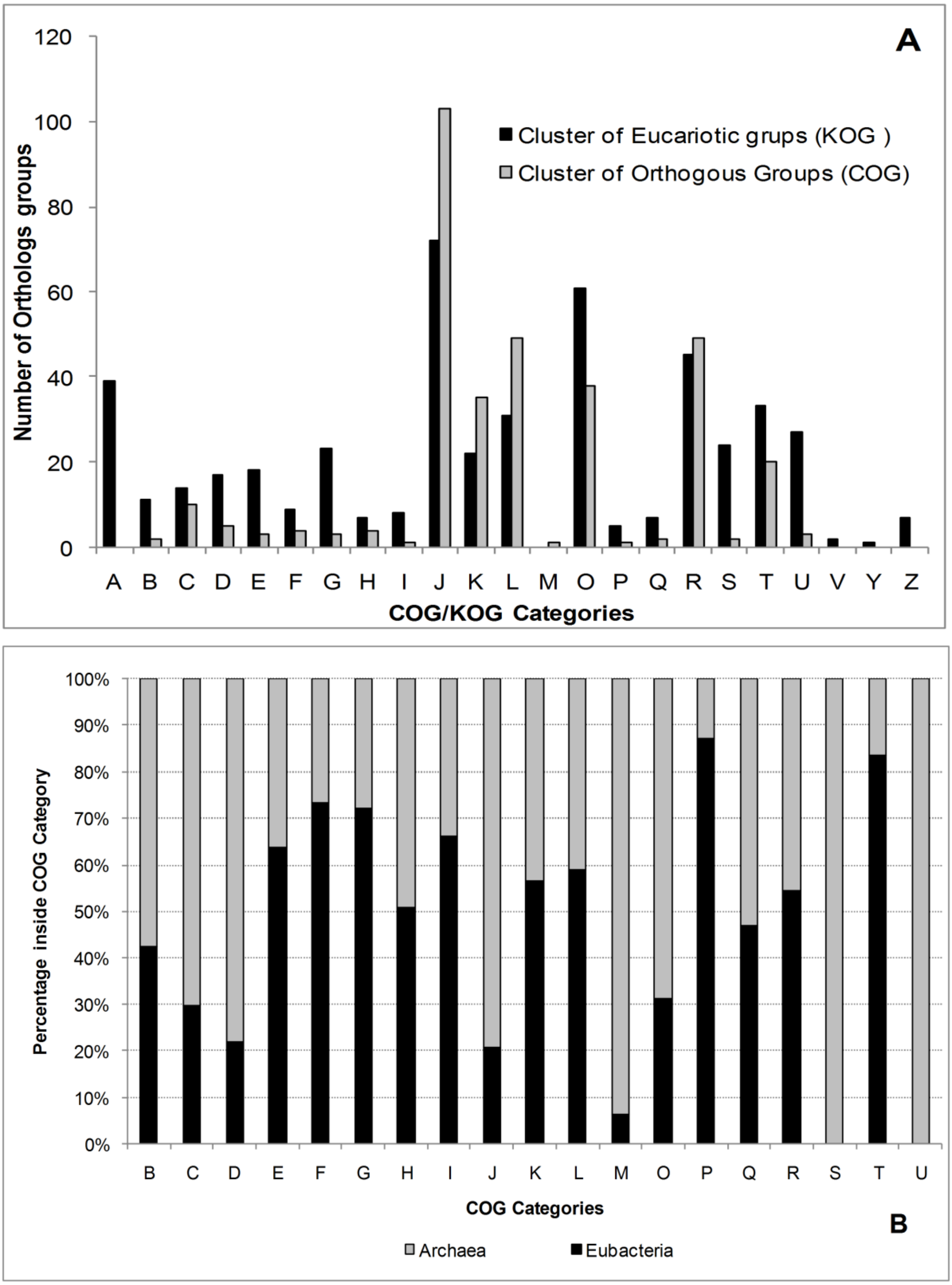


Figure 5: (A) Functional classification of the ortologous genes common to 22 protozoa, according COG/NCBI and KOG/NCBI for each groups (N = 348 groups). (B) Percentage of the 275 PCP orthologs closer to Archaea or Bacteria using the COG Functional classification (N = 275 groups). COG/KOG categories legend: Information Storage and Processing: [J] Translation, ribosomal structure and biogenesis; [A] RNA processing and modification; [K] Transcription; [L] Replication, recombination and repair; [B] Chromatin structure and dynamics. Cellular Processes and Signaling: [D] Cell cycle control, cell division, chromosome partitioning; [Y] Nuclear structure; [V] Defense mechanisms; [T] Signal transduction mechanisms; [M] Cell wall/membrane/envelope biogenesis; [Z] Cytoskeleton; [W] Extracellular structures; [U] Intracellular trafficking, secretion, and vesicular transport; [O] Posttranslational modification, protein

turnover, chaperones. Metabolism: [C] Energy production and conversion; [G] Carbohydrate transport and metabolism; [E] Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [H] Coenzyme transport and metabolism; [I] Lipid transport and metabolism; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport and catabolism. Poorly characterized: [R] General function prediction only; [S] Function unknown

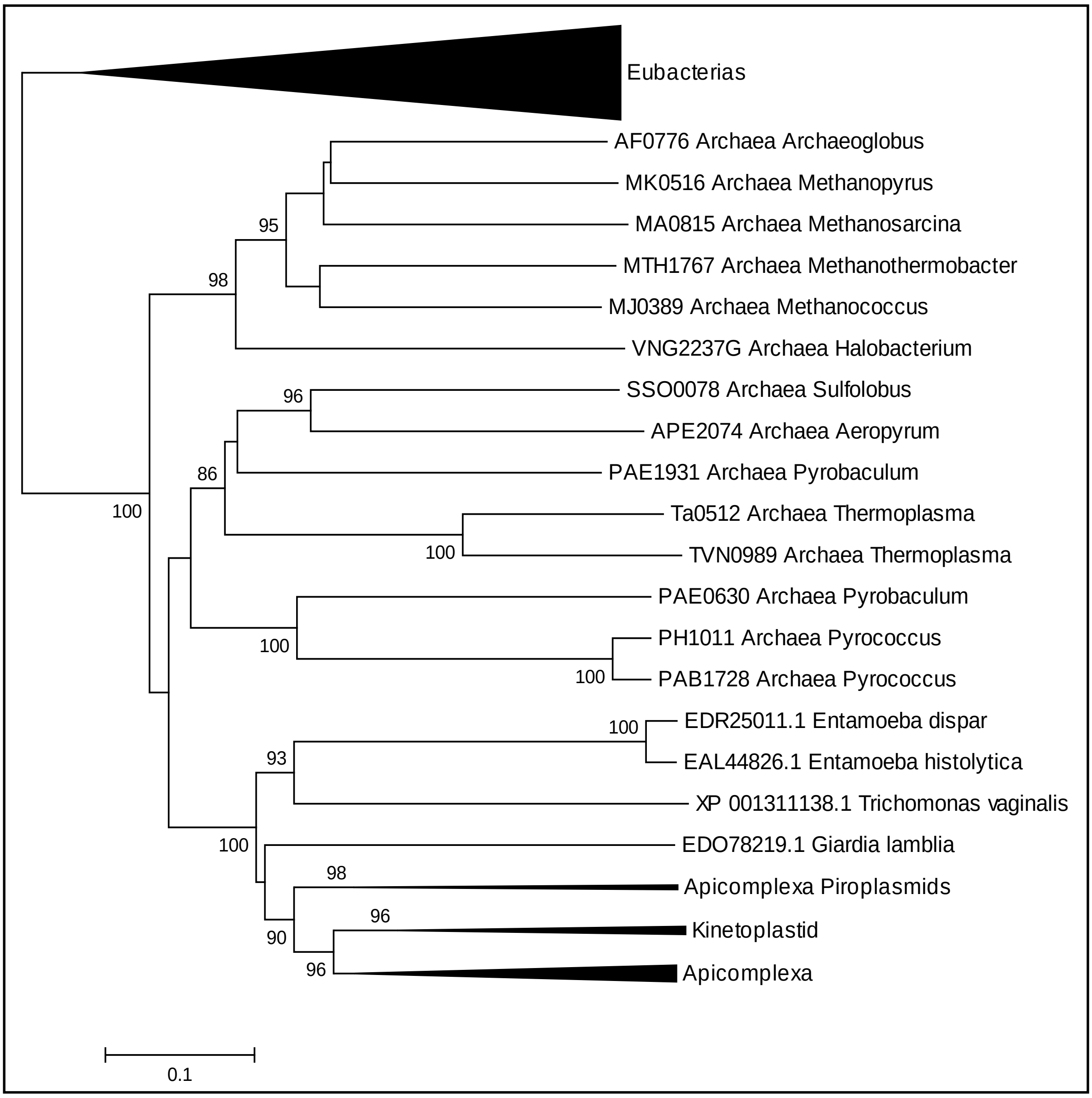

Eubacterias
AF0776 Archaea Archaeoglobus
MK0516 Archaea Methanopyrus
MA0815 Archaea Methanosarcina
MTH1767 Archaea Methanothermobacter
MJ0389 Archaea Methanococcus
VNG2237G Archaea Halobacterium
SSO0078 Archaea Sulfolobus
APE2074 Archaea Aeropyrum
PAE1931 Archaea Pyrobaculum
Ta0512 Archaea Thermoplasma
TVN0989 Archaea Thermoplasma
PAE0630 Archaea Pyrobaculum
PH1011 Archaea Pyrococcus
PAB1728 Archaea Pyrococcus
EDR25011.1 Entamoeba dispar
EAL44826.1 Entamoeba histolytica
XP 001311138.1 Trichomonas vaginalis
EDO78219.1 Giardia lamblia
Apicomplexa Piroplasmids
Kinetoplastid
Apicomplexa
95
98
96
86
100
100
100
100
100
93
100
98
96
90
96
100
0.1

Figure 6: Phylogenetic tree using Tyrosyl-tRNA synthetase (COG0162), constructed with MEGA, using Neighbor-Joining reconstruction, with p-distance, pairwise deletion and 1000 bootstrap replicates. The 22 protozoa species are condensed and represented by branches: Apicomplexa, Kinetoplastida and Apicomplexa Piroplasmid. COG Bacteria species used in this tree, are condensed and represented by branch: Eubacterias.

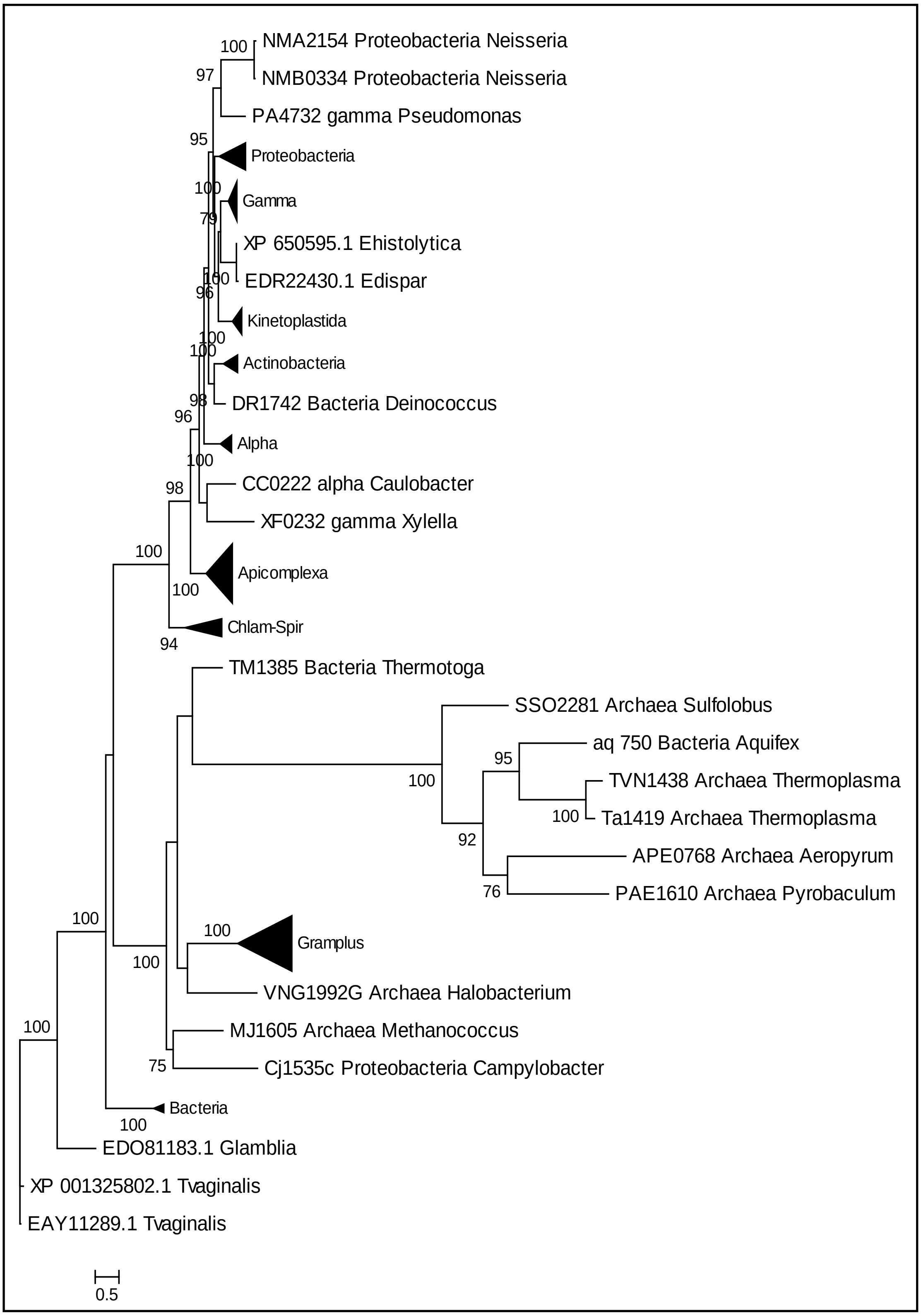

100
NMA2154 Proteobacteria Neisseria
97
NMB0334 Proteobacteria Neisseria
PA4732 gamma Pseudomonas
95
Proteobacteria
100
Gamma
79
XP 650595.1 Ehistolytica
100
EDR22430.1 Edispar
96
Kinetoplastida
100
100
Actinobacteria
98
DR1742 Bacteria Deinococcus
96
Alpha
100
98
CC0222 alpha Caulobacter
XF0232 gamma Xylella
100
Apicomplexa
100
Chlam-Spir
94
TM1385 Bacteria Thermotoga
SSO2281 Archaea Sulfolobus
aq 750 Bacteria Aquifex
95
100
TVN1438 Archaea Thermoplasma
100
Ta1419 Archaea Thermoplasma
92
APE0768 Archaea Aeropyrum
76
PAE1610 Archaea Pyrobaculum
100
100
Gramplus
100
VNG1992G Archaea Halobacterium
100
MJ1605 Archaea Methanococcus
75
Cj1535c Proteobacteria Campylobacter
Bacteria
100
EDO81183.1 Glamblia
XP 001325802.1 Tvaginalis
EAY11289.1 Tvaginalis
0.5

Figure 7: Phylogenetic tree using glucose-6-phosphate isomerase (COG0166), made with MEGA, using Neighbor-Joining reconstruction, with p-distance, pairwise deletion and 1000 bootstrap replicates. The 22 protozoa species are condensed and represented by branch: Kinetoplastida, Apicomplexa, Glamblia, Tvaginalis, Ehisto and Edispar. COG Bacteria and Archaea species used in this tree, has the suffix Archaea, Bacteria, Proteobacteria, gamma, alpha, Chlam-Spir, and some species are condensed and represented by branch:Gramplus, which has a big arrow.

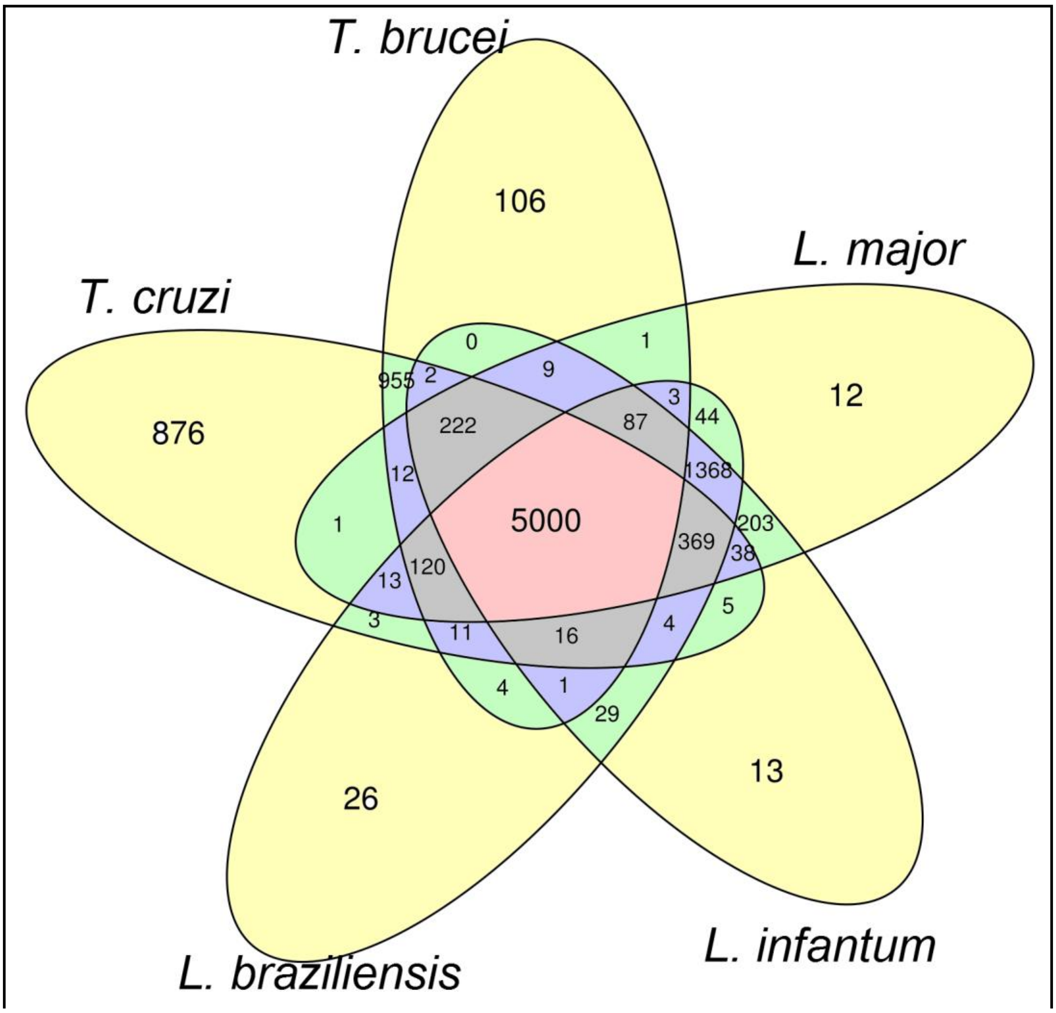


Figure 8: Venn diagram constructed for five species of protozoan Kinetoplastida. Figure shows the distribution of orthologous groups of *Trypanosoma brucei*, *Trypanosoma cruzi*, *Leishmania braziliensis*, *Leishmania infantum* and *Leishmania major* inferred by OrthoMCL. The yellow areas (106, 12, 13, 26 and 876) show proteins only from one species, green (0, 1, 44, 203, 5, 29, 4, 3, 1, 955) shared by groups two species, the blue area (9,3,1368,38,4,1,11,13,12 and 2) by three species and gray (87,369,16,120 and 222) and Red (5000) shared by four and five species, respectively

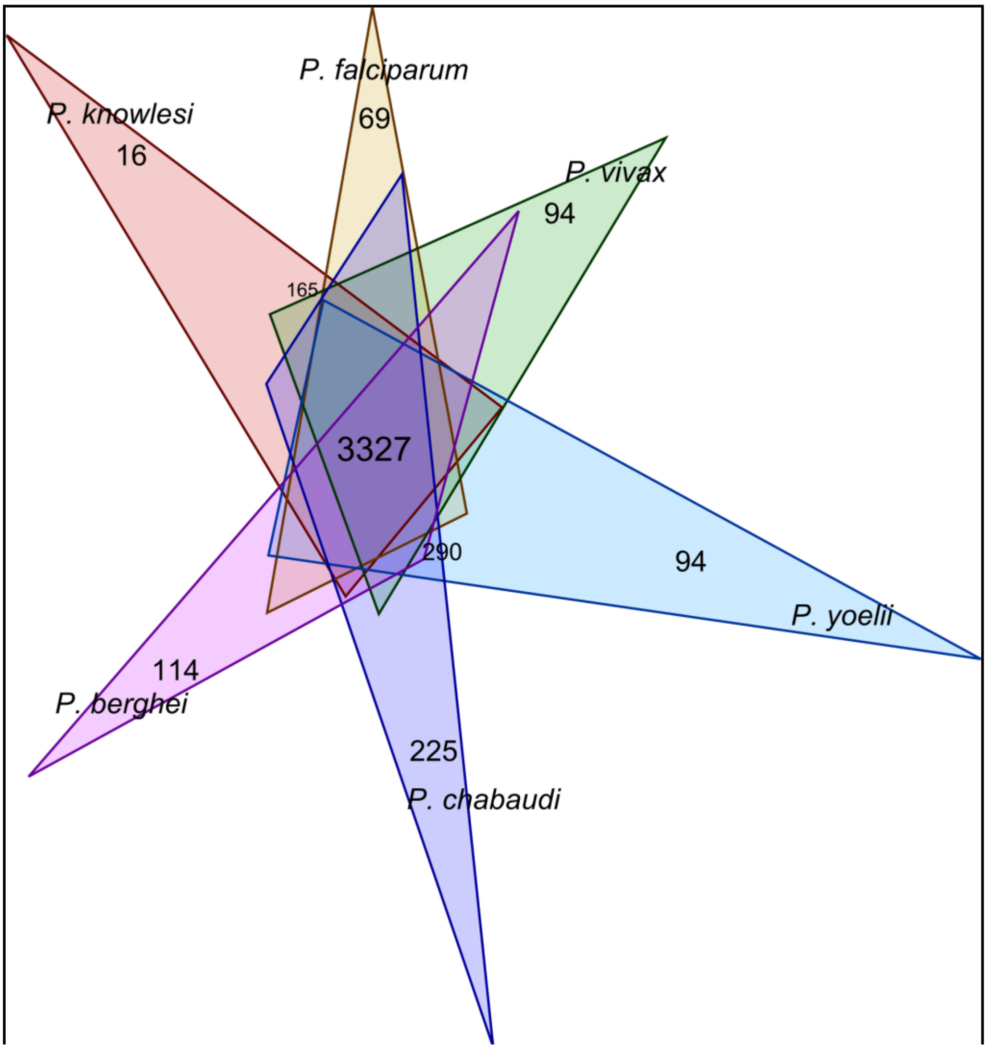


Figure 9: Venn diagram constructed for six species of Protozoan Apicomplexa, which shows the distribution of orthologous groups of *P. berguei*, *P.chaubadi*, *P. yoelli*, *P. falciparum*, *P. knowlesi* and *P. vivax*. We found 165 orthologous groups shared between *P. falciparum*, *P. knowlesi* and *P. vivax*. Furthermore, we found 290 orthologous groups shared between *P. berguei*, *P.chaubadi*, *P. yoelli.* Finally, 3327 orthologous groups, at figure center, were found shared by all six analyzed species.

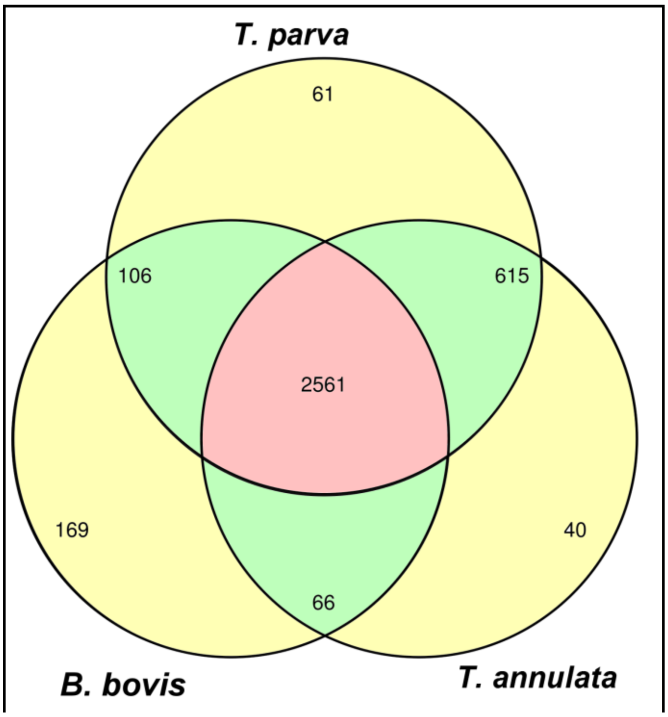


Figure 10: Venn diagram built for three species of protozoa piroplasmids which demonstrates the distribution of orthologous groups of *Babesia bovis*, *Theileria parva* and *Theileria annulata* inferred by OrthoMCL. The yellow areas indicate the proteins of only one species (61.40 and 169) in the green shared groups for two species (615, 66 and 106) and the red area (2561) for three species.

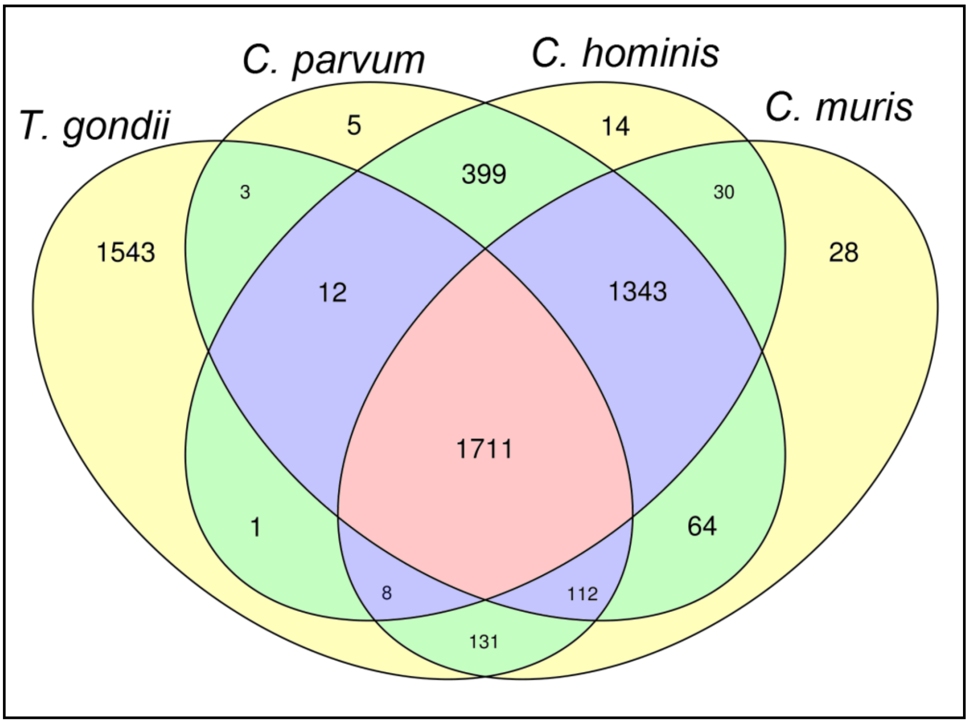


Figure 11: Venn diagram constructed for four species of protozoa coccidia, which demonstrates the distribution of orthologous groups of *Cryptosporidium parvum*, *Cryptosporidium hominis*, *Cryptosporidium muris* and *Toxoplasma gondii* inferred by OrthoMCL. The yellow areas (1543, 5,14 and 28) show a protein only from one species, green (30, 64, 131.1 and 3) groups shared by two species, the blue area (399, 1343, and 112.8 12) by three species and red (1711) shared by four species.

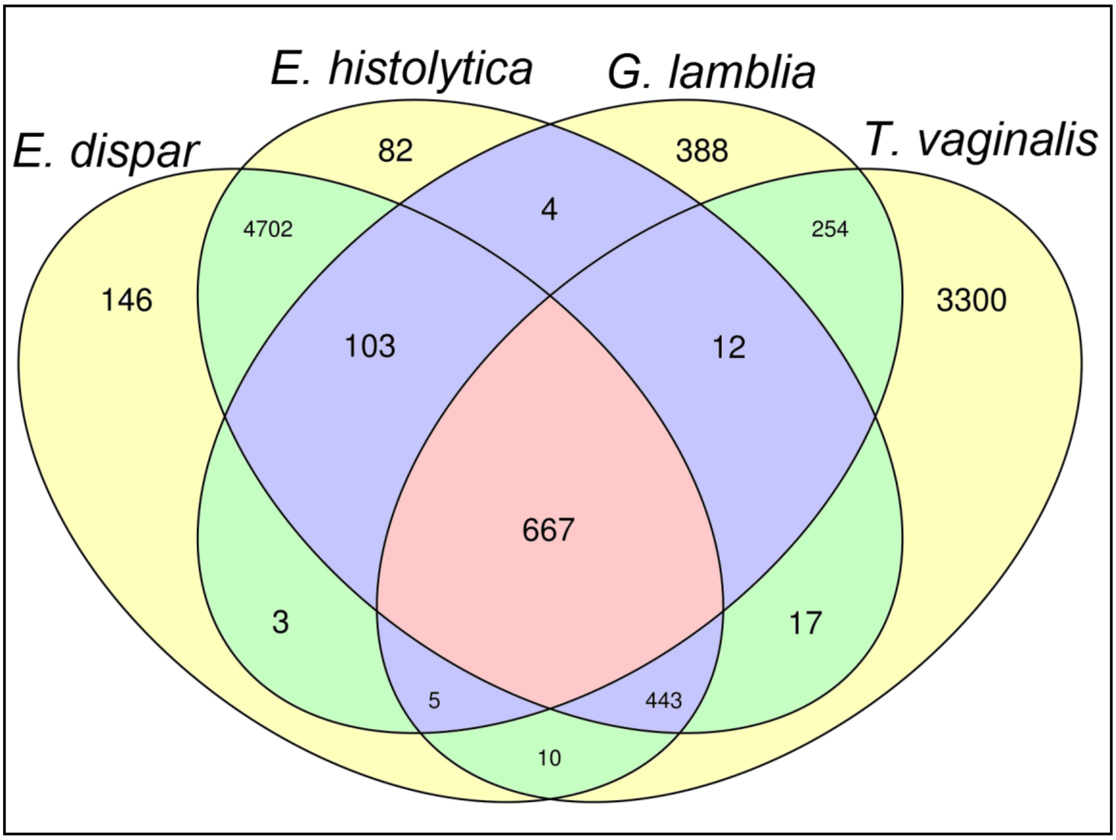


Figure 12: Venn diagram built to four species of protozoa, which shows the distribution of orthologous groups of *Entamoeba dispar*, *Entamoeba histolytica*, *Giardia lamblia* and *Trichomonas vaginalis* inferred by OrthoMCL. The areas in yellow (146, 82, 388 and 3300) indicate protein only from one species, green (4702, 254, 17, 10 and 3) groups shared by two species, the blue area (103, 4, 12, 443 and 5) of three species and red (667) shared by groups four species.

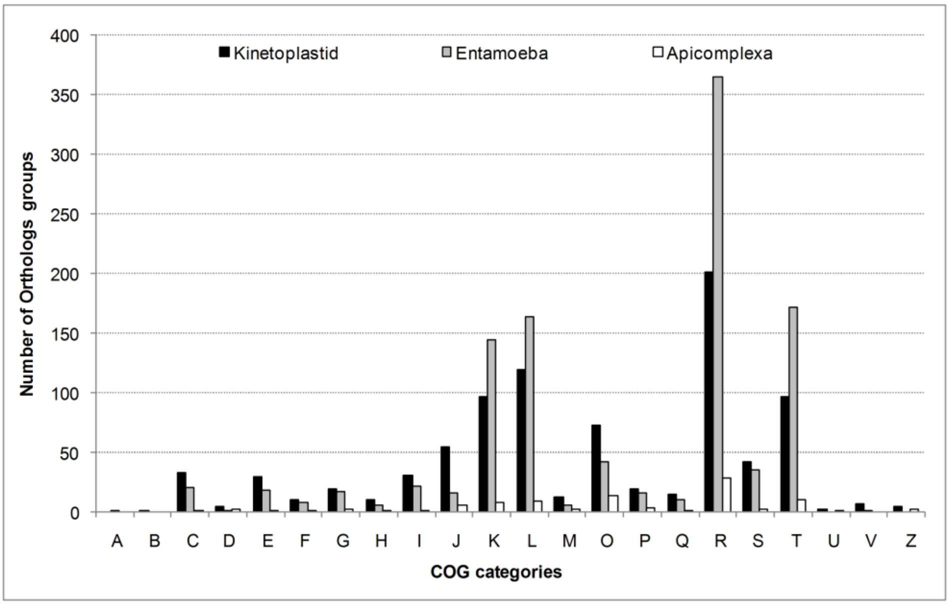


Figure 13. Functional characterization of the specifics proteins to Kinetoplastida, Apicomplexa and *Entamoeba*, using Blast search against COG/NCBI (Kinetoplastida N = 3396 groups; Apicomplexa N = 224 groups and *Entamoeba* N = 4441 groups). Functional characterization of Kinetoplastida Specific Proteins (*Trypanosoma cruzi, Trypanosoma brucei, Leishmania braziliensis, Leishmania infantum* e *Leishmania major*), Apicomplexa (ASP) (*Babesia bovis, Cryptosporidium parvum, Cryptosporidium hominis, Cryptosporidium muris, Plasmodium berghei, Plasmodium chabaudi, Plasmodium falciparum, Plasmodium vivax, Plasmodium yoelii, Plasmodium knowlesi, Theileria annulata,*

Theileria parva e *Toxoplasma gondii*) and *Entamoeba* (ESP) (*Entamoeba dispar*e *Entamoeba histolytica*) using COG/NCBI.

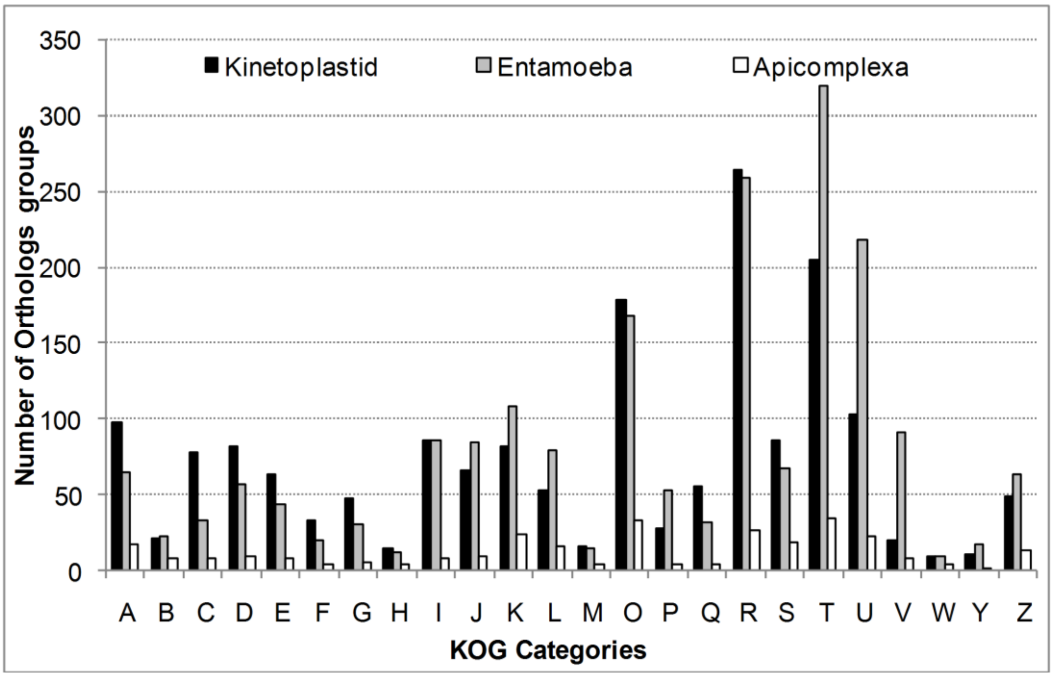


Figure 14. Functional characterization of the specific proteins of to Kinetoplastida, Apicomplexa and *Entamoeba*, using Blast search against KOG/NCBI. (Kinetoplastida N = 3396 groups; Apicomplexa N = 224 groups and *Entamoeba* N = 4441 groups). Functional characterization of Kinetoplastida Specific Proteins (*Trypanosoma cruzi, Trypanosoma brucei, Leishmania braziliensis, Leishmania infantum* e *Leishmania major*), Apicomplexa [ASP] (*Babesia bovis, Cryptosporidium parvum, Cryptosporidium hominis, Cryptosporidium muris, Plasmodium berghei, Plasmodium chabaudi, Plasmodium falciparum, Plasmodium vivax, Plasmodium yoelii, Plasmodium knowlesi, Theileria annulata, Theileria parva* e *Toxoplasma gondii*) and *Entamoeba* [ESP] (*Entamoeba dispar* and *Entamoeba histolytica*) using KOG/NCBI

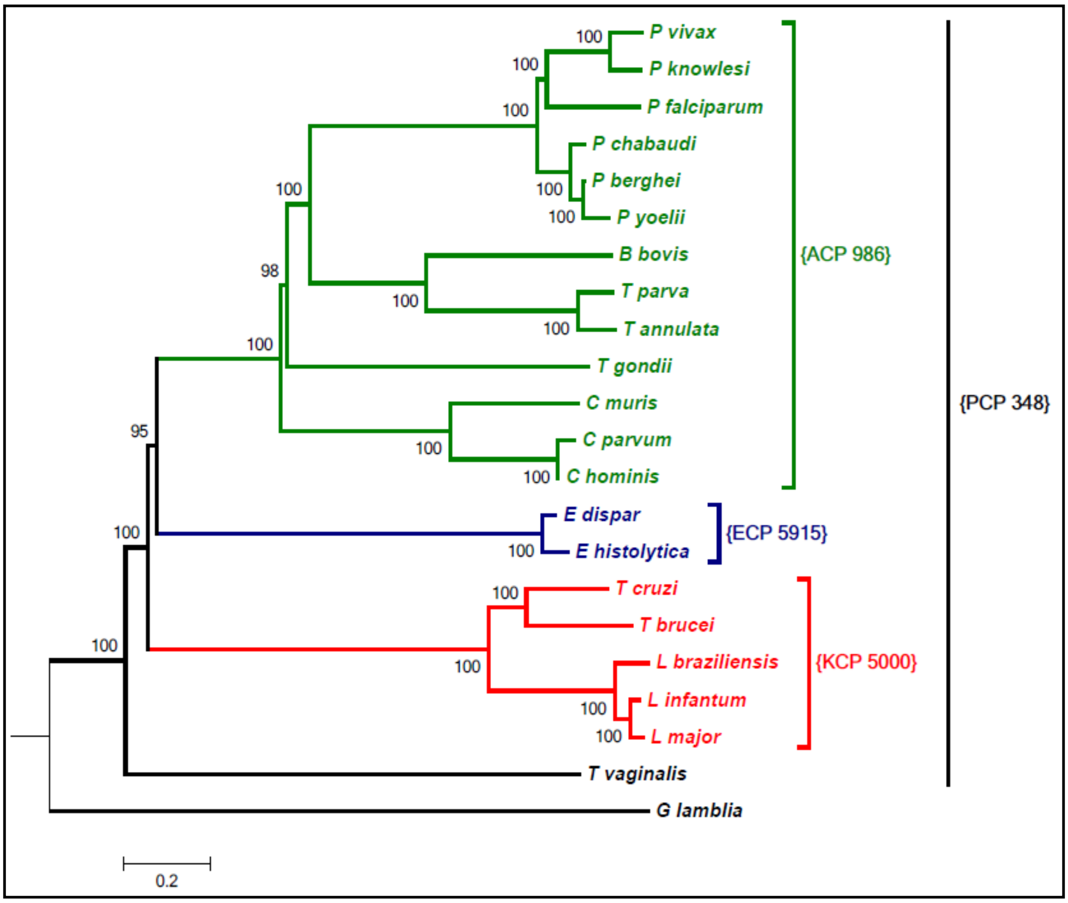


Fig 15. Phylogenomic analysis using the 348 PCP. Phylogenomic tree using MEGA bootstrap of 100 replicates and JTT model. PCP references the Protozoa Core Proteome (348 orthologs), ACP: Apicomplexa Core Proteome (986 orthologs) branch colored in green; ECP: *Entamoeba* Core Proteome (5915 orthologs), branch colored in blue; and KCP: Kinetoplastida Core Proteome represented by 5000 orthologs, and branch colored in red.

Tables

Table 1: Summary of dataset used in this study. Ordered by alphabetical name

Organism / Strain	**Non-redundant proteins**	**NCBI proteins**	**Literature proteins**	**%proteome used**	**% PCP X proteome size**	**Number proteins reference**
Babesia bovis T2Bo	3691	3713	3671	100.54	9.43	[8]
Crypstosporidium parvum Iowa II	3829	7667	3807	100.58	9.09	[5]
Cryptosporidium hominis TU502	3885	3885	3994	97.27	8.96	[11]
Cryptosporidium muris RN66	3930	7868	3934	99.9	8.85	site= Search=6061
Entamoeba dispar SAW760	8606	17622	8748	98.38	4.04	[22]
Entamoeba histolytica HM1:IMSS	7973	16372	8201	97.22	4.36	[20]
Giardia lamblia ATCC 50803	7163	13939	6470	110.71	4.86	[23]
Leishmania braziliensis MHOM BR 75 M2904	7825	7897	8153	95.98	4.45	[33]
*Leishmania infantum*JPCM5	7872	7992	8154	96.54	4.42	[33]
Leishmania major Friedlin	8003	16365	8298	96.44	4.35	[33]
Plasmodium berghei ANKA	9730	9840	5864	165.93	3.58	[16]
Plasmodium chabaudi chabaudi AS	14639	27045	5698	256.91	2.38	[49]
Plasmodium falciparum 3D7	5876	11669	5268	111.54	5.92	[15]
Plasmodium knowlesi strain H	5102	10213	5188	98.34	6.82	[19]
Plasmodium vivax SaI 1	5397	5413	5433	99.34	6.45	[17]
Plasmodium yoelii yoelii 17XNL	7303	7357	5878	124.24	4.77	[14, 109]
Theileria annulata Ankara	3790	3795	3792	99.95	9.18	[9]
Theileria parva Muguga	4050	4061	4035	100.37	8.59	[6]
Toxoplasma gondii ME49	7984	15983	8032	99.4	4.36	[110]
Trichomonas vaginalis G3	50189	119360	59681	84.1	0.69	[26]
Trypanosoma brucei treu927	8540	8768	9068	94.18	4.07	[36]
Trypanosoma cruzi CL Brener	19247	19644	12000	160.39	1.81	[35]

Name of the organism, and the strain used in this study. Total number of proteins utilized for each organism after remove the reduncancy with CD-HIT program. Total number of proteins obtained from NCBI (GenBank and RefSeq Dataset). Total number of proteins for each organism, described in literature. Proportion of the Protozoa Core Protein versus each proteome size. Reference of the source where the number of proteins was obtained in literature.

Table 2: Number of proteins used and number of homologs groups formed by OrthoMCL.

Organism	Proteins used by ORTHOMCL	Number of OrthoMCL Homologs Groups	%proteome used by OrthoMCL
Trichomonas vaginalis	44241	4708	88.15
Trypanosoma cruzi	18049	7647	93.78
Leishmania major	7937	7502	99.18
Leishmania infantum	7746	7366	98.4
Leishmania braziliensis	7602	7098	97.15
Entamoeba dispar	7498	6079	87.13
Trypanosoma brucei	7493	6549	87.74
Entamoeba histolytica	7152	6030	89.7
Plasmodium chabaudi	6683	5018	45.65
Plasmodium yoelii yoelii	6141	4679	84.09
Plasmodium berghei	6116	5002	62.86
Plasmodium falciparum	5604	4607	95.37
Plasmodium vivax	5214	4694	96.61
Plasmodium knowlesi	5062	4683	99.22
Toxoplasma gondii	4217	3521	52.82
Cryptosporidium parvum	3734	3649	97.52
Theileria parva	3683	3343	90.94
Cryptosporidium hominis	3623	3518	93.26
Theileria annulata	3606	3282	95.15
Cryptosporidium muris	3507	3427	89.24
Babesia bovis	3331	2902	90.25
Giardia lamblia	2857	1436	39.89

Organism name. Number of proteins used (which can be grouped) by OrthoMCL. Number of OrthoMCL homologs groups which each organism had.

Table 3: Number of homologs groups, orthologs and paralogs groups, for each species. Indecrescent order of the number of homologs groups

Species	Homologs groups	Total orthologs (%)	Paralogs (%)
Trypanosoma cruzi	7647	6833 (89.3)	814 (10.7)
Leishmania major	7502	7493 (99.9)	9 (0.01)
Leishmania infantum	7366	7358 (99.9)	8 (0.01)
Leishmania braziliensis	7098	7085 (99.8)	13 (0.02)
Trypanosoma brucei	6549	6454 (98.5)	95 (1.5)
Entamoeba dispar	6079	5943 (97.8)	136 (2.2)
Entamoeba histolytica	6030	5972 (99.0)	58 (1.0)
Plasmodium chabaudi	5018	4838 (96.4)	180 (3.6)
Plasmodium berghei	5002	4906 (98.1)	96 (1.9)
Trichomonas vaginalis	4708	1775 (37.7)	2933 (62.3)
Plasmodium vivax	4694	4654 (99.1)	40 (0.9)
Plasmodium knowlesi	4683	4667 (99.7)	16 (0.3)
Plasmodium yoeli	4679	4599 (98.3)	80 (1.7)
Plasmodium falciparum	4607	4553 (98.8)	54 (1.2)
Cryptosporidium parvum	3649	3646 (99.9)	3 (0.01)
Toxoplasma gondii	3521	3404 (96.7)	117 (3.3)
Cryptosporidium hominis	3518	3516 (99.9%)	2 (0.01)
Cryptosporidium muris	3427	3425 (99.9%)	2 (0.01)
Theileria parva	3343	3318 (99.3%)	25 (0.7)
Theileria annulata	3282	3266 (99.5%)	16 (0.5)
Babesia bovis	2902	2872 (99.0%)	30 (1.1)
Giardia lamblia	1436	1181 (82.2%)	255 (17.8)

Name of species. Number of homologs (orthologs and paralogs) groups assigned by OrthoMCL for each species. Number of orthologs groups assigned by OrthoMCL. Number of paralogs groups assigned by OrthoMCL divided by species.

Table 4: Function of some Protozoa Core Proteome orthologous proteins N = 348 orthologous groups

40S ribosomal protein S2/S3a/S6/S7/S8/S12/S16
60S ribosomal protein L3/L7/L12/L13a/L13/L17/L18/L32
Actin
Alpha tubulin
Asparaginyl-tRNA synthetase
Beta tubulin
Glucose-6-phosphate isomerase
Histone deacetylase
Histone H2A
Histone H3
Protein disulfide isomerase
Valyl tRNA synthetase
ATP binding protein
Proteasome subunit alpha 1
Eukaryotic translation initiation factor 2 gamma
NA polymerase B subunit RPB8
Isoleucyl-tRNA synthetase
Gamma tubulin
ATP-dependent DEAD/H RNA helicase
leucyl-tRNA synthetase
cysteine protease
DNA polymerase delta, catalytic subunit
zinc finger protein
DNA-directed RNA polymerase
5'-3' exonuclease

Function of some PCP proteins found in 22 protozoa studied

Table 5: Function and COG category of some PCP orthologous closer to Archaea or Bacteria

Archaea		Bacteria	
Protein function	**Category**	**Protein function**	**Category**
DNA-directed RNA polymerase	K	phospholipid-translocating P-type ATPase	P
RuvB DNA helicase	K	s-adenosylmethionine synthetase	H
transcription factor	K	molybdopterin synthase sulfurylase	H
60S rp L3/L13/L32	J	protein kinase	T
40S rp S12/S15	J	coatomer alpha subunit	T
Tyrosyl-tRNA synthetase	J	coatomer beta subunit	T
exodeoxyribonuclease	L	pyruvate kinase	G
DNA mismatch repair	L	glucose-6-phosphate isomerase	G
DNA primase large subunit,	L	cytidine and deoxycytidylate deaminase family	F
cell division cycle ATPase	O	guanylate kinase	F
proteasome subunit alpha	O	phospholipid-transporting ATPase	P

The function of these PCP orthologous were obtained trough protein annotation, and category trough a Blast analysis against COG database

Table 6: Number specific proteins for each taxa, proteomic core of these groups and the percentage of the core that is specific to each taxonomic group. Ordered from highest to lowest by percentage of specific groups.

Taxonomic group	Specifics	Proteomic Core	% Specific
Entamoeba	4441	5915	75.08
Kinetoplastida	3396	5000	67.92
Cryptosporidium	1194	3054	39.1
Plasmodium	965	3327	29.01
Apicomplexa	224	986	22.72
Piroplasmids	494	2561	19.29
Leishmania	1309	6824	19.18
Trypanosoma	900	6338	14.2
Coccids	174	1711	10.17
Rodent's *Plasmodium*	278	3763	7.39
Other	33	667	4.95
Human's *Plasmodium*	130	4334	3.0

Taxonomic group: represents a set of species. Specifics: Number of Orthologs groups created for each specific taxonomic group. Proteomic core: indicates the sum of orthologous groups shared by each taxonomic group. Specific percentage: indicates the fraction of proteins that are specific, that is shared by only the group

Table 7: Size of taxonomic groups, specific core percentage, hypothetical and R category groups number

Proteins	PCP (%)	ACP (%)	KCP (%)	ECP (%)
PCP size	348	986	5000	5915
Specific groups size:	--	224 (27.82)	3396 (67.92)	4441 (75,08)
Hypothetic groups size	--	92 (40.63)	1592 (46.29)	2905 (65,41)
R category size (KOG)	45 (12.93)	26 (9.15)	265 (15.17)	259 (13.28)

PCP represent Protozoa Core Proteome, ASP: Apicomplexa Specific Proteins, KSP: Kinetoplastida Specfic Protein and ESP: *Entamoeba* Specific Protein

Table 8: Function of some specific groups proteomic of Kinetoplastids (KSP), Apicomplexa (ASP) and *Entamoeba* (ESP). (KSP N = 3396; ASP N = 224; ESP N = 4441).

KSP functions	ASP functions	ESP functions
30S ribosomal protein S17	AAA family ATPase	60S ribosomal protein L22
40S ribosomal protein L14	Aspartyl proteinase	60S ribosomal protein L34
50S ribosomal protein L13	RhoGAP protein	60S ribosomal protein L7
60S ribosomal protein L28	S1/P1nuclease	ADP-ribosylation factor
ABC transporter	Actin	AIG1 family protein
Actin-like protein	apicomplexa specific secreted protein Pf (23508265), signal peptide	beta-amylas
ADP-ribosylation factor	aspartyl protease	cyst wall-specific glyco protein Jacob
Aminopeptidase	cGMP-dependent protein kinase	heat shock protein STI1
Calpain-like cysteine peptidase	deoxyuridine 5'-triphosphate nucleotidohydrolas	Hypothetical proteins
Chaperone protein DNAJ	DNAJ protein	lipase
Cytochrome p450 reductase	eukaryotic translation initiation factor 3 subunit 10	lysozyme
DNA replication licensing factor	Heat shock protein 70 (hsp70)	mannose-1-phosphate guanylyltransferase
DNA topoisomerase type IB small subunit	Heat shock protein 90	NADPH nitroreductase
Dynein arm light chain	Helicase	profilin
Fatty acid desnaturase	Histone binding protein	Rab GTPase activating protein
Glutathione peroxidase	Hypothetical proteins	rab10 family GTPase
glycosylphosphatidylinositol (GPI) anchor	K+ channel tetramerisation domain	rab-19
Hexose transporter	P-type ATPase	rab6 family GTPase
Hypothetical proteins	regulator of chromosome condensation protein	Ras guanine nucleotide exchange factor
Kinesin	RNA binding protein	Rho guanine nucleotide exchange factor
Lathosterol oxidase	RNA helicase	RNA 3'-terminal phosphate cyclase
Lipase	RNA methyltransferase	RNA-binding protein
p21 antigen protein	splicing factor	small GTPase Rab7
RNA binding protein-like protein	transporter protein CG10/chloroquine resistance transporter	sulfate adenylyltransferase
Zn-finger protein	Zinc finger protein	UDP-N-acetylglucosamine transporter

Description of the function of some proteins found in specific groups proteomic Kinetoplastida (KSP Kinetoplastida Specific Proteins), Apicomplexa (ASP: Apicomplexa Specific Proteins) and *Entamoeba* (ESP *Entamoeba* Specific Proteins). The function was obtained from the description given together with proteins found in each specific group.

Table 9: *Entamoeba* Specific Proteins without similarity against Refseq Database (N = 287 proteins).

AMP dependent ligase/synthetase
axoneme-associated protein mst101
caldesmon
centromeric protein E
cingulin
coiled-coil domain-containing protein
cylicin-2
cytoskeletal protein Sojo
DNA repair protein Rad-50
E3 ubiquitin protein ligase BRE1
early endsome antigen 1
epsin-2
ERC protein
F-box domain containing membrane protein
Filamin-A-interacting protein
formin 2,3 and collagen domain-containing protein
galactose-inhibitable lectin
glutamic acid-rich protein precursor
HMG box protein
leukocyte-endothelial cell adhesion molecule 3
major antigen
meiosis-specific nuclear structural protein
micronuclear linker histone polyprotein
M protein, serotype 5 precursor
myosin heavy chain
neurofilament medium polypeptide
nuclease sbcCD subunit C
Rho-associated protein kinase
Ring-infected erythrocyte surface antigen
serine-threonine-isoleucine rich protein
spindle assembly checkpoint component MAD1
structural maintenance of chromosomes protein
synaptonemal complex protein
Thioredoxin domain-containing protein 2
transcription initiation factor TFIID subunit
Troponin T, cardiac muscle isoforms
zinc finger protein Ran-binding domain-containing protein

4 – Artigo 2: Sobre as particularidades dos Protozoários: inferindo as expansões de famílias e proteínas órfãs

Tschoeke DA, Jardim R, Kotowoski Filho NP, Dávila AMR. **On the particularities of Protozoa: inferring family expansions and orphans.** Este manuscrito ainda encontra-se em preparação, contudo será submetido para revista *BMC Evolutionary Biology* assim que finalizado.

Este manuscrito corresponde aos objetivos específicos 2.2.5, 2.2.6 e 2.2.7: Identificar os parálagos para cada uma das espécies estudadas; Categorizar funcionalmente "*in silico*" os parálogos específicos e; Identificar as proteínas órfãs dos organismos estudados

O presente artigo faz um análise dos genes parálogos, utilizando os proteomas dos protozoários armazenados no ProtozoaDB 2.0. Observamos que as espécies que mais possuem duplicações são *T. vaginalis* e *T.* cruzi. Além disto, as maiores expansões são aquelas relacionadas, de maneira geral, a proteínas de superfície. Já as espécies que mais apresentaram proteínas órfãs foram *P. chaubadi* e *T. vaginalis*

On the particularities of Protozoa: inferring orphans and family expansions

Diogo A. Tschoeke[1,2], Rodrigo Jardim[1,2], Nelson Peixoto Kotowski Filho[1,2], Alberto M. R. Dávila *[1,2]

1 Computational and Systems Biology Pole, FIOCRUZ, Rio de Janeiro, Brazil **2** Computational and Systems Biology Lab, Oswaldo Cruz Institute, FIOCRUZ, Rio de Janeiro, Brazil *****Corresponding: davila@fiocruz.br

Abstract

Background: Protozoa is the common name given to unicellular Eukaryotes; it comprises about 200.000 species, extremely diverse and varied. Most of those species are free-living and nearly 10.000 are parasitic. The pathogenic species cause diseases such as: amoebiasis, Chagas disease, giardiasis, malaria, leishmaniasis and sleeping sickness. Paralogs identification plays an important role in genome characterization, as they usually undergo functional divergence by duplication, either via neofunctionalization and/or subfunctionalization.

Results: 204,624 non-redundant proteins from *Plasmodium*, *Entamoeba*, *Trypanosoma*, *Leishmania*, *Giardia*, *Theileria*, *Toxoplasma*, *Trichomonas* and *Cryptosporidium*, totalizing 22 species, were submitted to OrthoMCL, thus resulting in 26,101 homolog groups. Among them, 21,119 groups are orthologs, including 7,679 co-orthologs (groups that contain recent paralogs) and 4,982 are inparalogs. The Protozoa species which presented the highest number of inparalog groups was *Trichomonas vaginalis*, with 2,933 groups. Functional categorization has shown that the most abundant KOG category was the "T" category. *T. vaginalis* has also shown 948 co-orthologs, adding up to 3881 paralogs. However, *Trypanosoma cruzi* has shown the highest duplication number, totalizing 5777 paralogs, 4963 of them as co-orthologs and 814 as inparalogs groups, with “T” being the most abundant KOG category in inparalogs. When we look only for the larger expansions, that is, those families with at least 80 members, we found the following: *B. bovis* has a variant erythrocyte surface antigen-1; *G. lamblia* shows NEK and Variant-specific surface proteins; *P. chabaudi*: Pc-fam-2 protein; *P. falciparum*: rifin and var; *P. knowlesi*: SICA-like antigen; *P. vivax*: Vir24; *P. yoelii*: hypothetical protein; *T. brucei*: ESAG 4 protein and VSG; *T. cruzi*: trans-sialidase, mucin TcMUCII and retrotransposon hot spot protein; and *T. vaginalis* : ankyrin repeat protein, hypothetical protein TVAG_289600, TVAG_580790 and TVAG_235700. We were also able to detect orphan proteins (which showed no similarity within the cut-off values clustered by OrthoMCL), with the species that showed the highest number of orphans proteins being: *Plasmodium chabaudi*: 6961; and *Trichomonas vaginalis*, that, despite having almost 60,000 proteins, showed 4309 orphan proteins. It is important to notice that approximately 99% of orphan proteins in 22 species have been either described or annotated as hypothetical proteins or by having an unknown function.

Conclusions: *T. vaginalis* and *T. cruzi* showed higher expansions, considering the number of families and its sizes. Signal transduction mechanism was the most abundant category in inparalogs and proteins related to membrane usually have the largest expansions. Almost 50% of *Plasmodium chabaudi* proteins were classified as orphans.

Background

The Protozoa

Although a quite simplistic and enigmatic definition, Protozoa are defined as single-celled eukaryotic organisms, mainly due to its small-sized cells, the relative lack of uniform morphological features and the absence of its evolutionary history. Moreover, the act of classifying protozoa gets complex, as they stand between prokaryotic and higher eukaryotic organisms and also given the fact that they share characteristics from each of those [1].

In this sense, a wide range of definitions are available; Cavalier-Smith, in 1993 [2] stated: "Unicellular phagotrophic eukaryotes with mitochondria"; this would be a very simple definition, which would include the vast majority of Protozoa and exclude only a few; at the same time, it would include some Chromista. Such a diagnosis would not be accurate enough to define the kingdom's exact limits. The most simple, yet accurate phylogenetic definition of the Protozoa kingdom is as follows: eukaryotes, other than those that primitively lack mitochondria and peroxisomes (Archezoa), but which lack the shared characteristics that define the four higher derived kingdoms (Animalia, Fungi, Plantae, and Chromista) [2].

Electron microscopic discoveries eventually led to Bacteria being separated as a distinct kingdom and a five-kingdom system for eukaryotes: basal Protozoa and four derived kingdoms: the ancestrally heterotrophic Animalia and Fungi; and ancestrally phototrophic Plantae and Chromista [3, 4]. At last, the most conservative approach is to classify Protozoa as a subkingdom, without specifying whether it belongs to Animalia or Protista.

Paralogs and Co-Orthologs

Paralog genes are those generated via duplication event, in others words, are those duplicates genes. However, this definition does not include, in an explicit way, paralogs that reside in the same genome, because outparalogs are duplicated genes found in two different genomes. In other words, the duplication event occurs before a speciation event. Therefore, Co-Orthologs are defined as two or more genes in one lineage that are, collectively, orthologous to one or more genes in another lineage, due to a lineage-specific duplication [5–10].

For example, eukaryotic tubulins are co-orthologous to the single prokaryotic FtsZ proteins, just as actins are orthologous to the prokaryotic MreB. Within eukaryotes, however, several ortholog sets can be readily identified for each of these proteins [10]. Therefore, when studying paralogy, both functional diversity and specialization are general and persistent topics [7, 10, 11].

Gene copies creation may occur through unequal crossing-over, reverse transcription, or even by whole genome duplication. The duplication of a large region of a chromosome, resulting in either repeated sequences (tandem) or retrotranspons [9, 12] are also feasible events.

As a result of these processes, we may consider duplications as an important source for tracking evolutionary origin; it might ease speciation events, providing genetic material in order to speed up evolutionary rates, increasing enzymatic diversification, the number of cytoskeletal elements, gene expression patterns and more complex regulation, also including new gene functions, among other possibilities [8, 9].

These type of changes in the organisms genomes allow for: (i) the occurrence of events known as adaptive species radiation; (ii) the increase in the biological and genetic organism complexity and; (iii) a possible increase in genome size [6, 13, 14].

About the fate of these duplicated genes, the literature addresses three possibilities: (i) inactivation, by accumulating deleterious mutations, which would eventually turn it into a pseudogene, which occurs in 99% of these genes (non-functionalization); (ii) division of "ancestor" gene functions for the new formed genes, a process known as sub-functionalization; or (iii) the development of a new function (neo-functionalization). If this duplication is selectively advantageous, it will persist in the genome, although this fact is considered rare [5, 10, 11, 15–17].

Ohno (1970) also states that the gene duplication (paralog genes establishment) is primarily responsible for the emergence of functional novelties during evolution, most likely due to one of the newly duplicated genes escaping from purifying selection, therefor becoming free to accumulate mutations and develop a new function, while the other maintains the original gene function [15, 16, 18–20].

However, it seems that purifying selection, that is, the process of eliminating genes which carry mutations, disadvantageous not acts immediately after duplication, thus resulting in evolutional rate speedup for both paralog genes [10, 19]. These duplicate genes go through a stage where selection becomes softer and there may occur a faster nucleotide substitution rate [8].

According to Durand and Hoberman [21], this selection force leverages the remainder of this duplication; however, Kondrashov and Koonin [22] suggest that the initial gene fixation is due to the increased gene dosage. Also Gu [23], when comparing genomes, observed that some duplicate genes may easily modify the expression pattern, rather than single copy genes.

These observed relations between gene duplication and the expression diversity increase are considered to be important for two reasons: (i) the divergence in expression between duplicate genes can lead to functional specialization, therefore maintaining both gene copies in the genome; (ii) pairs of duplicated genes contribute to the expression diversity among strains. Studies also demonstrate that this functional redundancy exists between distant paralog genes.

Furthermore, during speciation and adaptation events, one may argue that: (i) functional redundant copies can offer more chances of adapting to new environmental conditions and physiology than single copy genes; (ii) changes in gene expression can lead to major phenotypic changes during evolution [23]. Thus, gene duplications carry the potential to generate new molecular substrates given rise to evolutionary novelties. Lynch and Conery [8] propose an average rate of gene duplication in the range of 0.01 genes per million years; it is then expected that 50% of all genes in a genome, about the size of *Caenorhabditis elegans*, will be duplicated, increasing their frequencies at least once in the next 35 to 350 million years, excluding the possibility of a whole genome duplication (polyploidy) [8].

Methods

Dataset

Protein sequences from 22 Protozoa species (*Babesia bovis, Crypstosporidium parvum, C. hominis, C. muris, Entamoeba dispar, E. histolytica, Giardia lamblia, Leishmania braziliensis, L. infantum, L. major, Plasmodium berghei, P. chabaudi, P. falciparum, P. knowlesi, P. vivax, P. yoelii, Theileria annulata, T. parva, Toxoplasma gondii, Trichomonas vaginalis, Trypanosoma brucei, T. cruzi*) were downloaded from GenBank (release 181.0) (http://www.ncbi.nlm.nih.gov/genbank/) and RefSeq (release 24.0) (http://www.ncbi.nlm.nih.gov/RefSeq/), in fasta format.

Redundancy removal

Identical protein sequences were removed using the software CD-HIT [24] with the following parameters: sequence identity threshold (1.00) and length difference cutoff (1.00).

Homology identification

The protein sequences from the 22 Protozoa species were used for homologs identification by OrthoMCL, with 1e-5 P-value cutoff [25]. Considering the huge amount of data and the substantial computational time to infer the relationship between proteins, our analysis was performed in a cluster at Coppe/UFRJ, with 640 dedicated cores, according to the protocol described by Coutinho [26].

The OrthoMCL output was used in order to infer ortholog and paralog protein groups among the 22 Protozoa species, as well as group-specific proteins and orphan proteins for each species. All obtained results were later loaded into ProtozoaDB [27], using in-house Perl/Ruby scripts, specifically developed for this purpose.

Homolog statistics

Homolog statistics data were obtained from OrthoMCL results, such as: (i) number of groups created, (ii) number of ortholog and paralog groups, (iii) and number of proteins used for each species. All statistics were performed using in-house scripts written in Perl, Ruby and R, as well as UNIX commands.

Paralogs identification and functional characterization

Paralog proteins analysis was performed using the OrthoMCL output file, then, we were able to identify paralogs among the 22 studied Protozoa. Paralogs protein function was assigned according to the given sequence annotation.

Functional categorization was performed through similarity analysis using

Blast and RpsBlast [28] against both prokaryote (COG) and eukaryote (KOG) ortholog genes database from NCBI. E-value was set up as 1e-5 in both applications in order to find out which functional category each one of ortholog group belongs to.

We followed the categorization proposed by Tatusov in 1997 [29] and 2003 [30]; however, sequences from unicellular eukaryotic species that are present in COG database were removed through an in-house script. At last, functional categories graphics were created with R software [31].

Co-ortholog proteins identification

Co-ortholog proteins identification was performed using an in-house Perl script, which uses the output file generated by the OrthoMCL software as input. This script identifies and classifies the number of proteins for each organism into co-orthologs groups, and then creates paralog groups for the species that contain duplicate entries of a given co-ortholog group. R software was used in order to generate family gene expansions graphics, for each organism and family.

Putative orphan proteins identification

Putative orphan proteins identification was performed using an in-house Ruby script, which uses the output file generated by the OrthoMCL software as input. It compares the list of identifiers submitted to OrthoMCL against the list of clustered identifiers. The identifiers submitted but not clustered were considered as putative orphans proteins.

In order to avoid any misclassification, we performed a Blast analysis with these putative orphan proteins against COG/KOG databases.

Finally, according to Domazet-Loso and Tautz [32], the term "orphan" originally had a double meaning: (i) coding regions without known function; and (ii) coding regions without matches to other genes in the Refseq database (release 56; 18,132,578 sequences); in our work we use the second one.

Results

Dataset and redundancy cleanup

A total of 346,468 proteins were obtained from NCBI (Genbank and RefSeq) databases for the 22 Protozoa species. Cd-hit was used in order to perform a redundancy cleanup, which generated the nr-Protozoa dataset, with a total of 204,624 proteins. Table 1 shows a summary on this process, pointing the amount of proteins obtained from NCBI, the effect of the redundancy cleanup with Cd-hit, the number of predicted proteins reported in the literature, the percentage of the proteome used and the adequate literature reference, separated by species.

Homologs identification

According to our calculations and based on data from literature, after redundancy cleanup, the maximum percentage of the proteome size used for homology analysis corresponds to *P. chabaudi* (259.9% of the predicted proteome size reported in the literature) and minimum to *T. vaginalis*, *with* 84.1% of its proteome (Table 1). 83.61% (171,096/204,624) of the nr-Protozoa grouped in some of those 26,101 homolog groups obtained by OrthoMCL. Of these groups, 4,982 (19.09%) are paralogs and 21,119 (80.91%) orthologs. Among the ortholog groups, 7,679 (36.36%) are co-orthologs (comprising recent paralogs) and 13,440 (63,64%) groups are ortholog-only proteins.

These results can be viewed at ProtozoaDB (Dávila et al., 2008) http://protozoadb.biowebdb.org website.

As shown in Table 2, *T. vaginalis* has the greatest amount of paralog groups 58.89% (2933/4982) of the total paralog numbers, followed by *T. cruzi* with 16.84% (814/4982). On the other hand, *C. hominis* and *C. muris* are the species with the least paralog groups, two (2) for each, respectively. Nevertheless, *T. vaginalis*, *T. cruzi* and *G. lamblia* show most duplicated groups (Figure 1) with 82%, 76% and 28% proportion, respectively.

When we observe the proportion of ortholog groups versus the number of co-ortholog groups with their recent paralogs, more than 70% of *T. cruzi* groups are duplicates (Table 3, 4 and Figure 2), while 70% of *T. vaginalis* groups are paralogs. Moreover, the species that had most duplication groups was *T. cruzi* with 5777 groups, of this 85.91% (4963/5777) are co-orthologs and 14.09% (814/5777) are paralogs (Table 4). *T. vaginalis* was the second species with most duplicated groups, 3881, while 75.57% (2933/3881) are paralogs and 24.43% (948/3881) are co-orthologs. The lowest duplicated proteome is of the *C. muris* species with two paralogs and 69 co-orthologs.

Finally, the number of paralog groups within the Plasmodium genus varies; as an example, paralog groups varied from 0.32% (16/4982) 16 (*P. knowlesi*) up to 3.61% (180/4982) (*P. chaubadi*).

Paralogous groups functional categorization

We performed the functional categorization of paralogs groups found in these species, in order to ascertain which category these groups belong to. In order to do so, we used both the prokaryotic (COG) and eukaryotic (KOG) orthologs from NCBI. To better visualize these results, we divided the organisms into groups: (i) Apicomplexa organisms; (ii) Kinetoplastida, and finally (iii) *Entamoeba* spp., *G. lamblia* and *T. vaginalis*.

The functional categorization of paralogs in Apicomplexa using COG (Figure 3) show that the most abundant category in *T. gondii* was “R”, with 33.33% (9/30) of the characterized paralogs. *P. berghei* shows one paralog in category “H” and one at “I”. *P. vivax* has one paralog in “E”, “N” and “R” category, respectively. *P. yoelii* has one paralog into “L” category. *T. annulata* shows one ortholog, classified as “U”. *C. parvum* has two paralogs in the following categories: “K”, “L”, “R” and “T”. Finally, *T. parva* has one paralog classified as “J” category.

Paralogs functional characterization using KOG in Apicomplexa species (Figure 4) shows that categories “R” and “T”, with 21.57% (11/51) each, were the two most abundant into *T. gondii*. *B. bovis* has one paralog into category “W”; *C. muris* one into “J”; *C. parvum* has two paralogs as “T” category; *P. vivax* one paralog into “C”, “F” and “M” category each; *P. yoelli* one into “A”, “D” and “Z”; *T. annulata* has one paralog belonging to “W” and another to “Z”; and *T. parva* one paralog into “J” category.

When Kinetoplastida species were functionally categorized through COG, *T. cruzi* (Figure 5) has shown “R” category with 24.72% (22/89) of the characterized paralogs, as its most abundant category. *L. braziliensis* has one paralog classified into “O”, while *L. infantum* has one into “P” and other in “S”1; *L. major* showed two paralogs, classified as “R”. Finally, *T. brucei* shows one paralog classified as “R” and other as “S”, respectively.

Using KOG, Kinetoplastida functional categorization shown that (Figure 6) “R” category, with 14.1% (22/156) of the characterized groups, was the most abundant one, in *T. cruzi*. For *Leishmania* species, *L. braziliensis* shows one paralog in categories: “M”, “O”, “P”, “R” and “V”; 1. *L. infantum* has one paralog classified into “R”, while *L. major* has four of them in the same category. Finally, *T. brucei* has one paralog into categories “A”, “E”, “I” and “O” respectively; two paralogs into “J” and “Q” and six classified into “R” category.

Functional categorization using COG in *E. dispar*, *E. histolytica*, *T. vaginalis* and *G. lamblia* paralogs all pointed to category “R”, with 54.9% (28/51), 57.9% (11/19), 40.5% (318/785) and 13.6% (20/147) of the characterized groups, respectively (Figure 7).

Functional categorization using KOG, for the species above, showed that “O” category was most representative into *E. dispar* and *E. histolytica* 30% (6/20) and 44.4% (4/9) respectively. On the other hand, R category was most abundant into *G. lamblia*, showing 13% (20/147) of the characterized groups and also into *T. vaginalis*, with 15% (243/1621) of the characterized groups in such category.

Family expansions (em desenvolvimento)

Putative Orphans proteins identification

We consider orphan proteins those which showed no similarity within the cutoff values to be clustered by OrthoMCL; in other words, they showed no similarity to other proteins of the protozoan subject to the program. The species that showed the highest number of orphan proteins was *P. chabaudi* with 6961 proteins, followed by *T. vaginalis,* that, despite having almost 60,000 proteins, pointed only 8.59% (4309/50189) orphans (Table 5). Nevertheless, *G. lamblia* showed the highest orphans proteins percentage with 50.05% (3585/7163), while *L. major* has the lowest orphan proteins percentage, 0.54% (43/8003).

Discussion (em desenvolvimento)

Conclusions (em desenvolvimento)

List of abbreviations

Competing interests

Authors' contributions

DAT: analyzed data and write the manuscript. RJ, NPKF: helps with analysis and revised the manuscript. AMRD: revised the manuscript and helps with scientific questions

Acknowledgements

To CAPES, FAPERJ and Fiocruz for the financial support.

References

1. Imam T: **The complexities in the classification of protozoa: a challenge to parasitologists**. *Bayero J Pure Appl Sci* 2009, **2**:159–164.

2. Cavalier-Smith T: *Kingdom Protozoa and Its 18 Phyla. Volume 57*; 1993:953–94.

3. Cavalier-Smith T: **Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree.** *Biol Lett* 2010, **6**:342–5.

4. Cavalier-Smith T: **Deep phylogeny, ancestral groups and the four ages of life.** *Philos Trans R Soc Lond B Biol Sci* 2010, **365**:111–32.

5. Veitia RA: **Paralogs in polyploids: one for all and all for one?** *Plant Cell* 2005, **17**:4–11.

6. Van de Peer Y: **Evolutionary genetics: when duplicated genes don’t stick to the rules.** *Heredity (Edinb)* 2006, **96**:204–5.

7. Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG, Koonin E V: **Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell.** *Nucleic Acids Res* 2005, **33**:4626–38.

8. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290**:1151–5.

9. Gogarten JP, Olendzenski L: **Orthologs, paralogs and genome comparisons.** *Curr Opin Genet Dev* 1999, **9**:630–6.

10. Koonin E V: **Orthologs, paralogs, and evolutionary genomics.** *Annu Rev Genet* 2005, **39**:309–38.

11. Wolfe KH, Li W-H: **Molecular evolution meets the genomics revolution.** *Nat Genet* 2003, **33 Suppl**:255–65.

12. Holland PW: **Gene duplication: past, present and future.** *Semin Cell Dev Biol* 1999, **10**:541–7.

13. Leveugle M, Prat K, Perrier N, Birnbaum D, Coulier F: **ParaDB: a tool for paralogy mapping in vertebrate genomes.** *Nucleic Acids Res* 2003, **31**:63–7.

14. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y: **Modeling gene and genome duplications in eukaryotes.** *Proc Natl Acad Sci U S A* 2005, **102**:5454–9.

15. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations.** *Genetics* 1999, **151**:1531–45.

16. Kondrashov FA, Rogozin IB, Wolf YI, Koonin E V: **Selection in the evolution of gene duplications.** *Genome Biol* 2002, **3**:RESEARCH0008.

17. Makarova KS, Koonin E V: **Comparative genomics of Archaea: how much have we learned in six years, and what's next?** *Genome Biol* 2003, **4**:115.

18. Escrivá García H, Laudet V, Robinson-Rechavi M: **Nuclear receptors are markers of animal genome evolution.** *J Struct Funct Genomics* 2003, **3**:177–84.

19. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH: **Why highly expressed proteins evolve slowly.** *Proc Natl Acad Sci U S A* 2005, **102**:14338–43.

20. Koonin E V, Galperin MY: *Sequence - Evolution - Function : Computational Approaches in Comparative Genomics*. Boston, Dordrecht, London: Kluwer academic publishers; 2002.

21. Durand D, Hoberman R: **Diagnosing duplications--can it be done?** *Trends Genet* 2006, **22**:156–64.

22. Kondrashov FA, Koonin E V: **A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications.** *Trends Genet* 2004, **20**:287–90.

23. Gu Z, Rifkin SA, White KP, Li W-H: **Duplicate genes increase gene expression diversity within and between species.** *Nat Genet* 2004, **36**:577–9.

24. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22**:1658–9.

25. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**:2178–89.

26. Coutinho F, Ogasawara E, Oliveira D, Braganholo V, Lima AAB, Dávila AMR, Mattoso M: **Many task computing for orthologous genes identification in protozoan genomes using Hydra**. *Concurr Comput Pract Exp* 2011:n/a–n/a.

27. Dávila AMR, Mendes PN, Wagner G, Tschoeke D a, Cuadrat RRC, Liberman F, Matos L, Satake T, Ocaña K a CS, Triana O, Cruz SMS, Jucá HCL, Cury JC, Silva FN, Geronimo G a, Ruiz M, Ruback E, Silva FP, Probst CM, Grisard EC, Krieger M a, Goldenberg S, Cavalcanti MCR, Moraes MO, Campos MLM, Mattoso M: **ProtozoaDB: dynamic visualization and exploration of protozoan genomes.** *Nucleic Acids Res* 2008, **36**(Database issue):D547–52.

28. Altschul SF, Madden TL, Schäffer a a, 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, **25**:3389–402.

29. Tatusov RL: **A Genomic Perspective on Protein Families**. *Science (80-)* 1997, **278**:631–637.

30. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, others: **The COG database: an updated version includes eukaryotes**. *BMC Bioinformatics* 2003, **4**:41.

31. R Core Team: **R: A Language and Environment for Statistical Computing**. 2013.

32. Domazet-Loso T, Tautz D: **An evolutionary analysis of orphan genes in Drosophila.** *Genome Res* 2003, **13**:2213–9.

33. Brayton K a, Lau AOT, Herndon DR, Hannick L, Kappmeyer LS, Berens SJ, Bidwell SL, Brown WC, Crabtree J, Fadrosh D, Feldblum T, Forberger H a, Haas BJ, Howell JM, Khouri H, Koo H, Mann DJ, Norimine J, Paulsen IT, Radune D, Ren Q, Smith RK, Suarez CE, White O, Wortman JR, Knowles DP, McElwain TF, Nene VM: **Genome sequence of Babesia bovis and comparative analysis of apicomplexan hemoprotozoa.** *PLoS Pathog* 2007, **3**:1401–13.

34. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto C a, Deng M, Liu C, Widmer G, Tzipori S, Buck G a, Xu P, Bankier AT, Dear PH, Konfortov B a, Spriggs HF, Iyer L, Anantharaman V, Aravind L, Kapur V: **Complete genome sequence of the apicomplexan, Cryptosporidium parvum.** *Science* 2004, **304**:441–5.

35. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, Pearson WR, Dear PH, Bankier AT, Peterson DL, Abrahamsen MS, Kapur V, Tzipori S, Buck GA: **The genome of Cryptosporidium hominis**. *Science (80-)* 2004, **431**(October):1107–1112.

36. Weedall GD, Hall N: **Evolutionary genomics of Entamoeba.** *Res Microbiol* 2011:1–9.

37. Lorenzi H a, Puiu D, Miller JR, Brinkac LM, Amedeo P, Hall N, Caler E V: **New assembly, reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features and protein content information.** *PLoS Negl Trop Dis* 2010, **4**:e716.

38. Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, Best A a, Cande WZ, Chen F, Cipriano MJ, Davids BJ, Dawson SC, Elmendorf HG, Hehl AB, Holder ME, Huse SM, Kim UU, Lasek-Nesselquist E, Manning G, Nigam A, Nixon JEJ, Palm D, Passamaneck NE, Prabhu A, Reich CI, Reiner DS, Samuelson J, Svard SG, Sogin ML: **Genomic minimalism in the early diverging intestinal parasite Giardia lamblia.** *Science* 2007, **317**:1921–6.

39. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, Kerhornou A, Ivens A, Fraser A, Rajandream M-A, Carver T, Norbertczak H, Chillingworth T, Hance Z, Jagels K, Moule S, Ormond D, Rutter S, Squares R, Whitehead S, Rabbinowitsch E, Arrowsmith C, White B, Thurston S, Bringaud F, Baldauf SL, et al.: **Comparative genomic analysis of three Leishmania species that cause diverse human disease.** *Nat Genet* 2007, **39**:839–47.

40. Hall N, Carlton J: **Comparative genomics of malaria parasites.** *Curr Opin Genet Dev* 2005, **15**:609–13.

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

42. 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 J a, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan M-S, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM a, et al.: **Genome sequence of the human malaria parasite Plasmodium falciparum.** *Nature* 2002, **419**:498–511.

43. Pain a, Böhme U, Berry a E, Mungall K, Finn RD, Jackson a P, Mourier T, Mistry J, Pasini EM, Aslett M a, Balasubrammaniam S, Borgwardt K, Brooks K, Carret C, Carver TJ, Cherevach I, Chillingworth T, Clark TG, Galinski MR, Hall N, Harper D, Harris D, Hauser H, Ivens a, Janssen CS, Keane T, Larke N, Lapp S, Marti M, Moule S, et al.: **The genome of the simian and human malaria parasite Plasmodium knowlesi.** *Nature* 2008, **455**:799–803.

44. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, Crabtree J, Angiuoli S V, Merino EF, Amedeo P, Cheng Q, Coulson RMR, Crabb BS, Del Portillo HA, Essien K, Feldblyum T V, Fernandez-Becerra C, Gilson PR, Gueye AH, Guo X, Kang'a S, Kooij TWA, Korsinczky M, Meyer EV-S, Nene V, Paulsen I, White O, Ralph SA, Ren Q, Sargeant TJ, et al.: **Comparative genomics of the neglected human malaria parasite Plasmodium vivax.** *Nature* 2008, **455**:757–63.

45. Carlton J, Silva J, Hall N: **The genome of model malaria parasites, and comparative genomics.** *Curr Issues Mol Biol* 2005, **7**:23–37.

46. Carlton JM, Angiuoli S V, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, Peterson JD, Pop M, Kosack DS, Shumway MF, Bidwell SL, Shallom SJ, van Aken SE, Riedmuller SB, Feldblyum T V, Cho JK, Quackenbush J, Sedegah M, Shoaibi A, Cummings LM, Florens L, Yates JR, Raine JD, Sinden RE, Harris M a, Cunningham D a, et al.: **Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii.** *Nature* 2002, **419**:512–9.

47. Pain A, Renauld H, Berriman M, Murphy L, Yeats C a, Weir W, Kerhornou A, Aslett M, Bishop R, Bouchier C, Cochet M, Coulson RMR, Cronin A, de Villiers EP, Fraser A, Fosker N, Gardner M, Goble A, Griffiths-Jones S, Harris DE, Katzer F, Larke N, Lord A, Maser P, McKellar S, Mooney P, Morton F, Nene V, O'Neil S, Price C, et al.: **Genome of the host-cell transforming parasite Theileria annulata compared with T. parva.** *Science* 2005, **309**:131–3.

48. Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM, Hall N, Ren Q, Paulsen IT, Pain A, Berriman M, Wilson RJM, Sato S, Ralph S a, Mann DJ, Xiong Z, Shallom SJ, Weidman J, Jiang L, Lynn J, Weaver B, Shoaibi A, Domingo AR, Wasawo D, Crabtree J, Wortman JR, Haas B, Angiuoli S V, Creasy TH, Lu C, Suh B, et al.: **Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes.** *Science* 2005, **309**:134–7.

49. Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X, Heiges M, Iodice J, Kissinger JC, Mackey AJ, Pinney DF, Roos DS, Stoeckert CJ, Wang H, Brunk BP: **ToxoDB: an integrated Toxoplasma gondii database resource.** *Nucleic Acids Res* 2008, **36**(Database issue):D553–6.

50. Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao Q, Wortman JR, Bidwell SL, Alsmark UCM, Besteiro S, Sicheritz-Ponten T, Noel CJ, Dacks JB, Foster PG, Simillion C, Van de Peer Y, Miranda-Saavedra D, Barton GJ, Westrop GD, Müller S, Dessi D, Fiori PL, Ren Q, Paulsen I, Zhang H, Bastida-Corcuera FD, Simoes-Barbosa A, Brown MT, Hayes RD, Mukherjee M, et al.: **Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis.** *Science* 2007, **315**:207–12.

51. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Böhme U, Hannick L, Aslett M a, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UCM, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth T-J, Churcher C, Clark LN, Corton CH, Cronin A, et al.: **The genome of the African trypanosome Trypanosoma brucei.** *Science* 2005, **309**:416–22.

52. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran A-N, Ghedin E, Worthey E a, Delcher AL, Blandin G, Westenberger SJ, Caler E, Cerqueira GC, Branche C, Haas B, Anupama A, Arner E, Aslund L, Attipoe P, Bontempi E, Bringaud F, Burton P, Cadag E, Campbell D a, Carrington M, Crabtree J, Darban H, da Silveira JF, de Jong P, Edwards K, et al.: **The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease.** *Science* 2005, **309**:409–15.

Figures legends

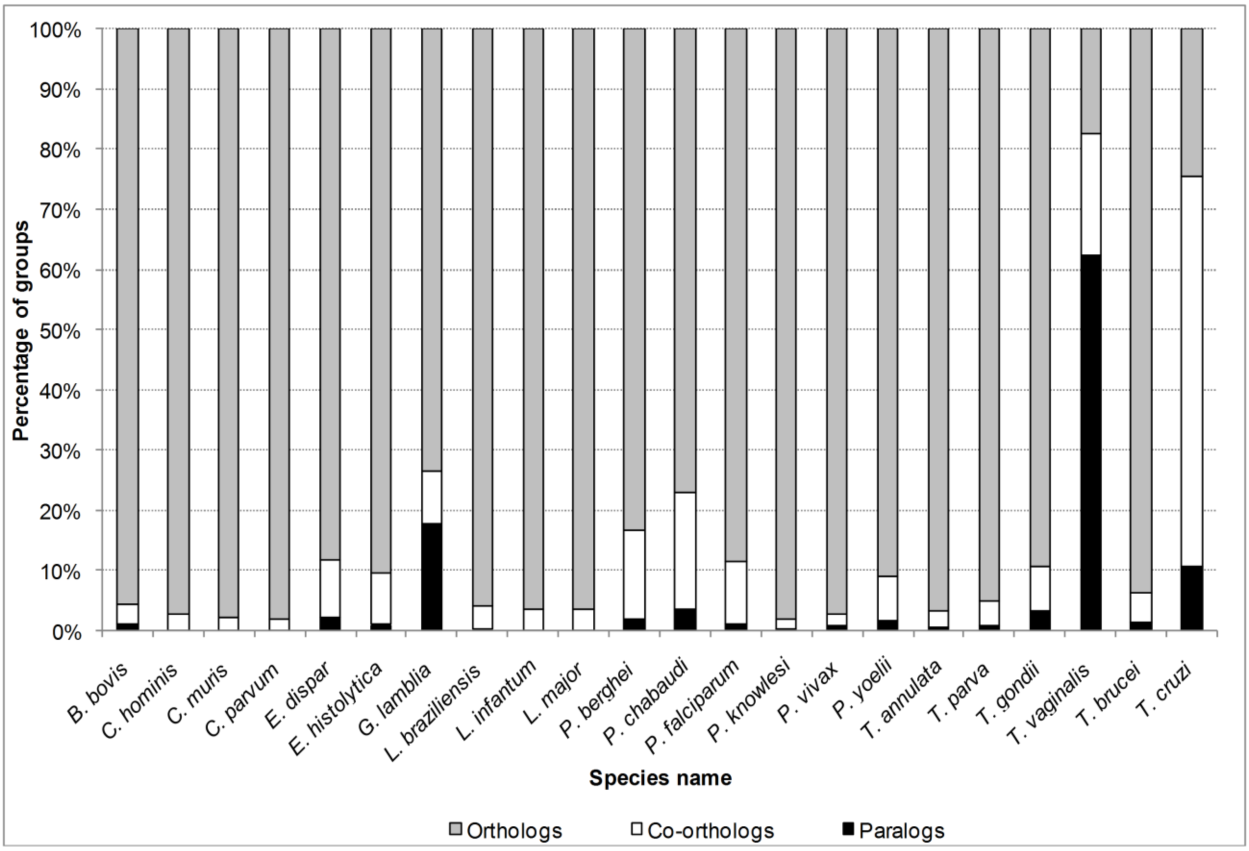


Figure 1: Homolog groups distribution (N = 26101) for each analyzed proteome species. Ortholog, Co-ortholog and paralog groups percentage found in each analyzed species, generated by OrthoMCL analysis.

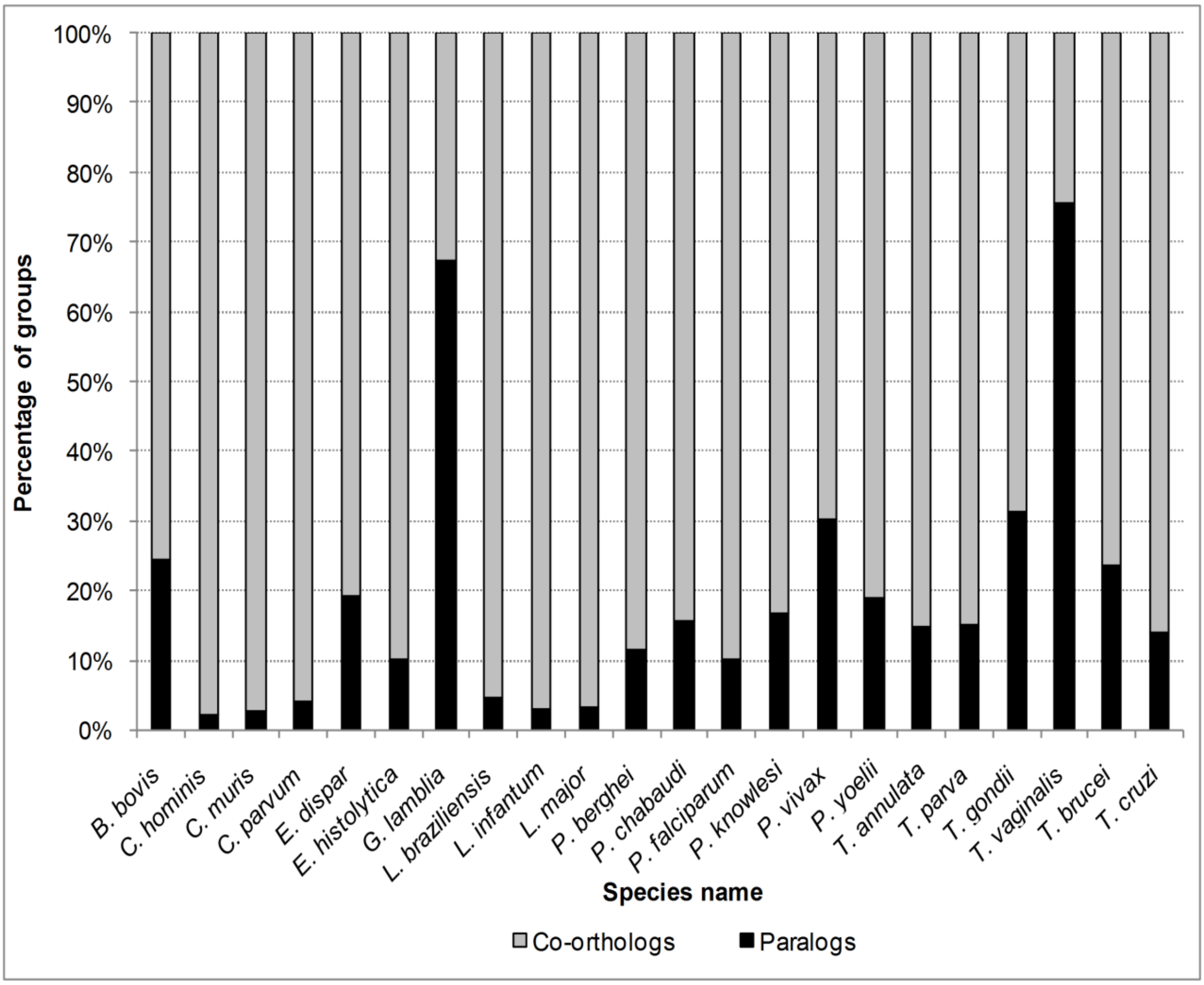


Figure 2: Duplicated groups distribution for the analyzed species (N = 12661). Co-ortholog and Paralog groups percentage found in each analyzed species.

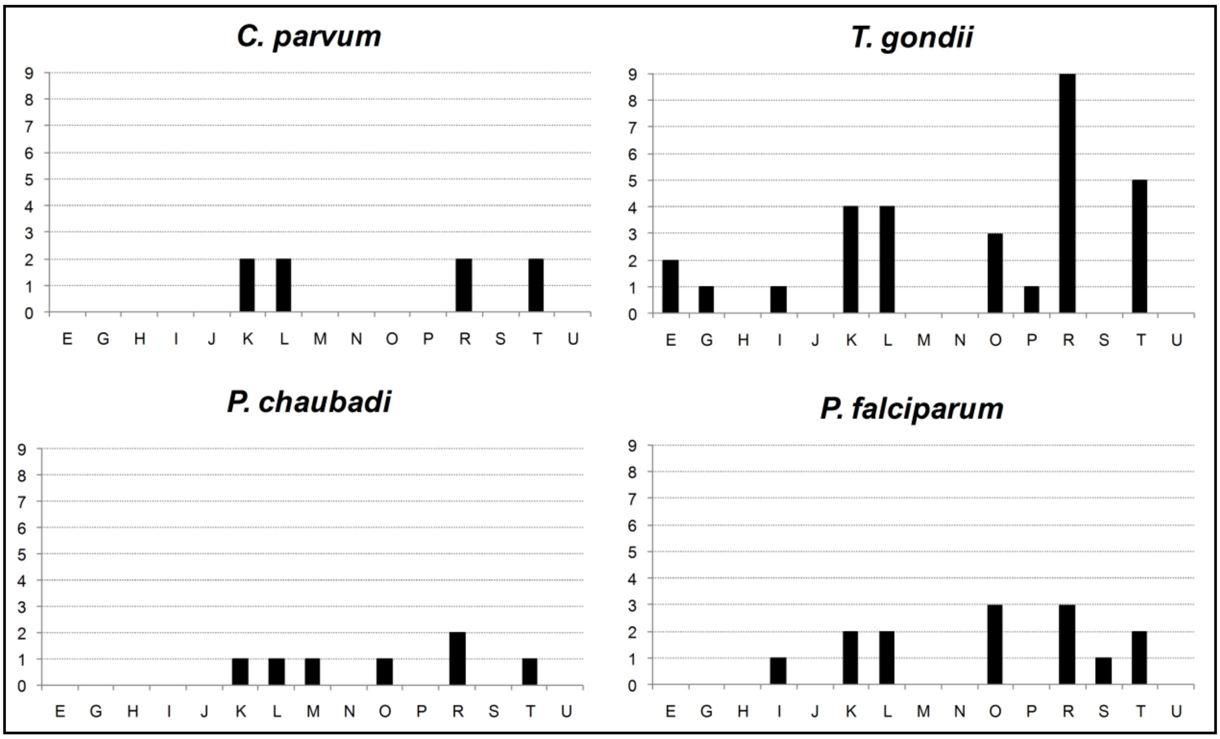


Figure 3. Functional characterization of the paralog proteins groups in Apicomplexa using COG/NCBI (*C. parvum* N = 3 groups; *T. gondii* N =117 groups; *P. chaubadi* N =180 groups; *P.*

falciparum N = 54 groups). Legend: Information Storage and Processing: [J] Translation, ribosomal structure and biogenesis; [K] Transcription; [L] Replication, recombination and repair; [T] Signal transduction mechanisms; [M] Cell wall/membrane/envelope biogenesis; [N] Cell motility; [U] Intracellular trafficking, secretion, and vesicular transport; [O] Posttranslational modification, protein turnover, chaperones. Metabolism: [G] Carbohydrate transport and metabolism; [E] Amino acid transport and metabolism; [I] Lipid transport and metabolism; [P] Inorganic ion transport and metabolism; Poorly characterized: [R] General function prediction only; [S] Function unknown.

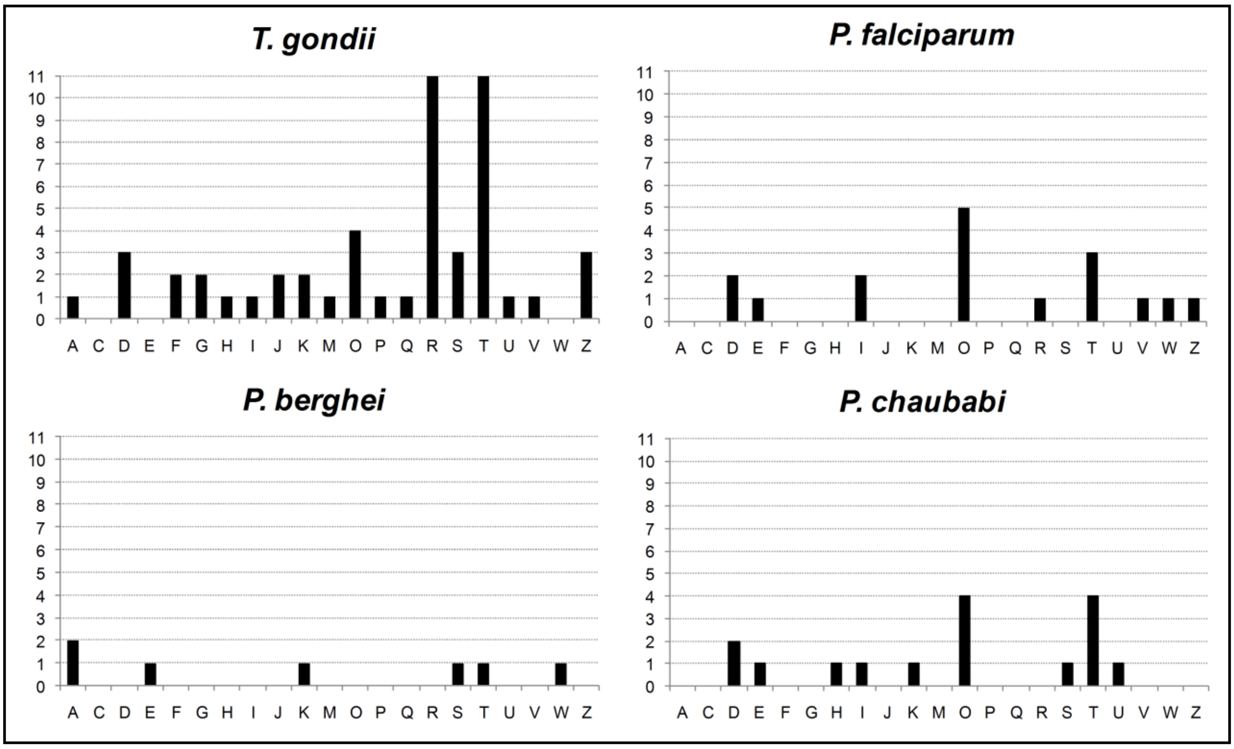


Figure 4. Functional characterization of the paralog proteins groups in Apicomplexa using KOG/NCBI. (*T. gondii* N = 117 groups; *P. berghei* N = 96 groups*; P. chaubadi* N = 180 groups; *P. falciparum* N = 54 groups) Legend: Information Storage and Processing: [J] Translation, ribosomal structure and biogenesis; [A] RNA processing and modification; [K] Transcription; [L] Replication, recombination and repair. Cellular Processes and Signaling: [D] Cell cycle control, cell division, chromosome partitioning; [V] Defense mechanisms; [T] Signal transduction mechanisms; [M] Cell wall/membrane/envelope biogenesis; [Z] Cytoskeleton; [W] Extracellular structures; [U] Intracellular trafficking, secretion, and vesicular transport; [O] Posttranslational modification, protein turnover, chaperones. Metabolism: [C] Energy production and conversion; [G] Carbohydrate transport and metabolism; [E] Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [H] Coenzyme transport and metabolism; [I] Lipid transport and metabolism; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport and catabolism. Poorly characterized: [R] General function prediction only; [S] Function unknown.

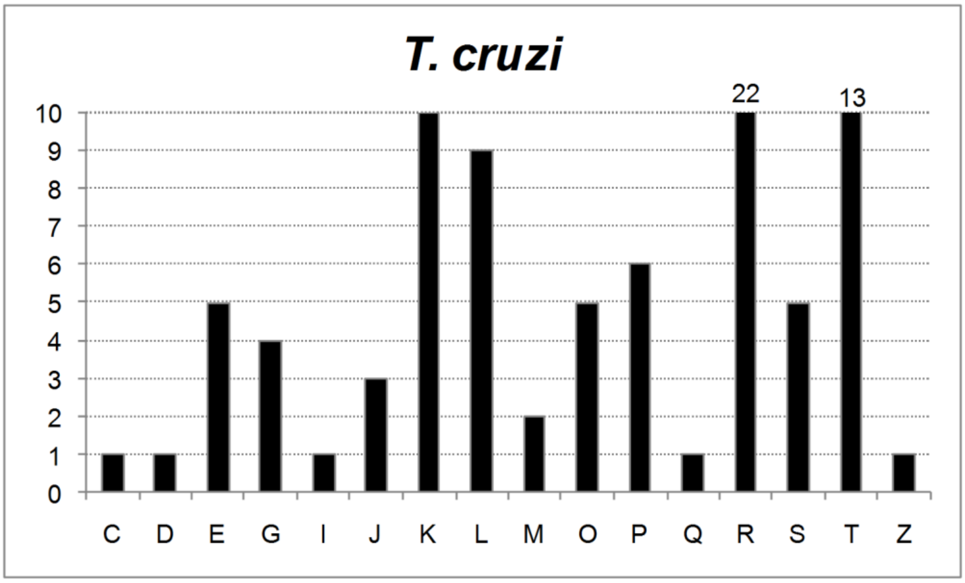


Figure 5. Functional characterization of the paralog proteins groups in *T. cruzi* using COG/NCBI. (N = 814 groups) Legend: Information Storage and Processing: [J] Translation, ribosomal structure and biogenesis; [K] Transcription; [L] Replication, recombination and repair. Cellular Processes and Signaling: [D] Cell cycle control, cell division, chromosome partitioning; [T] Signal transduction mechanisms; [M] Cell wall/membrane/envelope biogenesis; [Z] Cytoskeleton; [O] Posttranslational modification, protein turnover, chaperones. Metabolism: [C] Energy production and conversion; [G] Carbohydrate transport and metabolism; [E] Amino acid transport and metabolism; [I] Lipid transport and metabolism; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport and catabolism. Poorly characterized: [R] General function prediction only; [S] Function unknown.

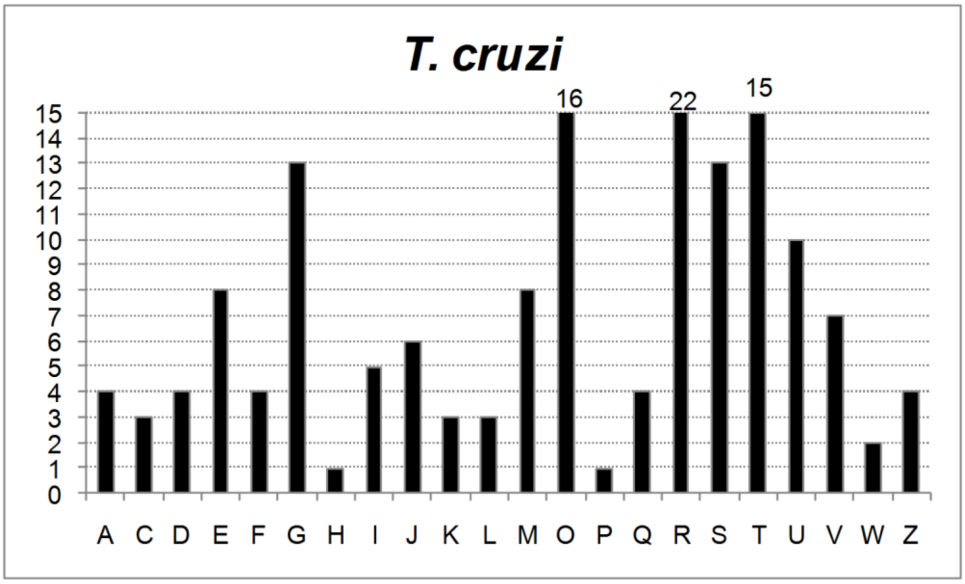


Figure 6. Functional characterization of the paralog proteins groups in *T. cruzi* using KOG/NCBI. (N = 814 groups). Legend Information Storage and Processing: [J] Translation, ribosomal structure and biogenesis; [A] RNA processing and modification; [K] Transcription; [L] Replication, recombination and repair. Cellular Processes and Signaling: [D] Cell cycle control, cell division, chromosome partitioning; [V] Defense mechanisms; [T] Signal transduction mechanisms; [M] Cell wall/membrane/envelope biogenesis; [Z] Cytoskeleton; [W] Extracellular structures; [U] Intracellular trafficking, secretion, and vesicular transport; [O] Posttranslational modification, protein turnover, chaperones. Metabolism: [C] Energy production and conversion; [G] Carbohydrate transport and metabolism; [E] Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [H] Coenzyme transport and metabolism; [I] Lipid transport and metabolism; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport and catabolism. Poorly characterized: [R] General function prediction only; [S] Function unknown.

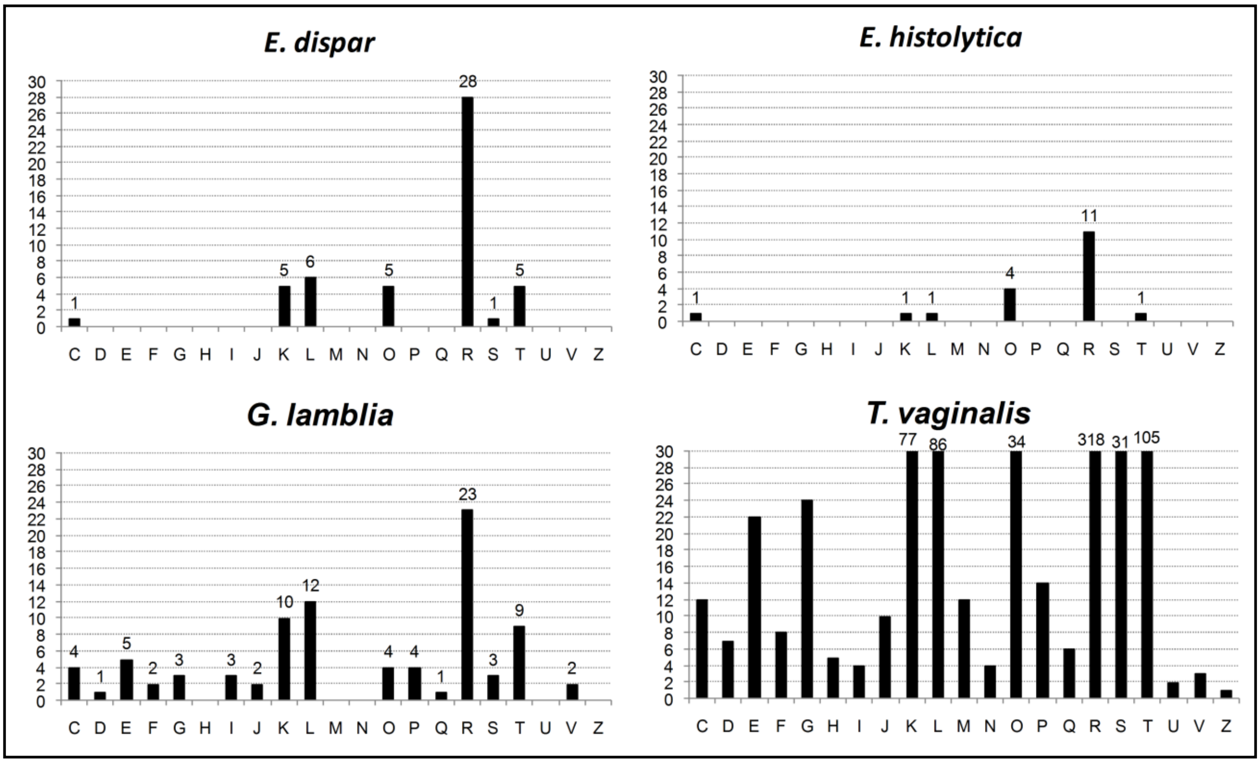


Figure 7. Functional characterization of the paralog proteins groups in *E. histolytica, E. dispar, G. lamblia* and *T. vaginalis* using COG/NCBI (*E. dispar* N = 136; *E. histolytica* N = 58; *G. lamblia* N = 255; *T. vaginalis* N = 2933). Legend: Information Storage and Processing: [J] Translation, ribosomal structure and biogenesis; [K] Transcription; [L] Replication, recombination and repair. Cellular Processes and Signaling: [D] Cell cycle control, cell division, chromosome partitioning; [V] Defense mechanisms; [T] Signal transduction mechanisms; [M] Cell wall/membrane/envelope biogenesis; [N] Cell motility; [Z] Cytoskeleton; [U] Intracellular trafficking, secretion, and vesicular transport; [O] Posttranslational modification, protein turnover, chaperones. Metabolism: [C] Energy production and conversion; [G] Carbohydrate transport and metabolism; [E] Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [H] Coenzyme transport and metabolism; [I] Lipid transport and metabolism; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport and catabolism. Poorly characterized: [R] General function prediction only; [S] Function unknown.

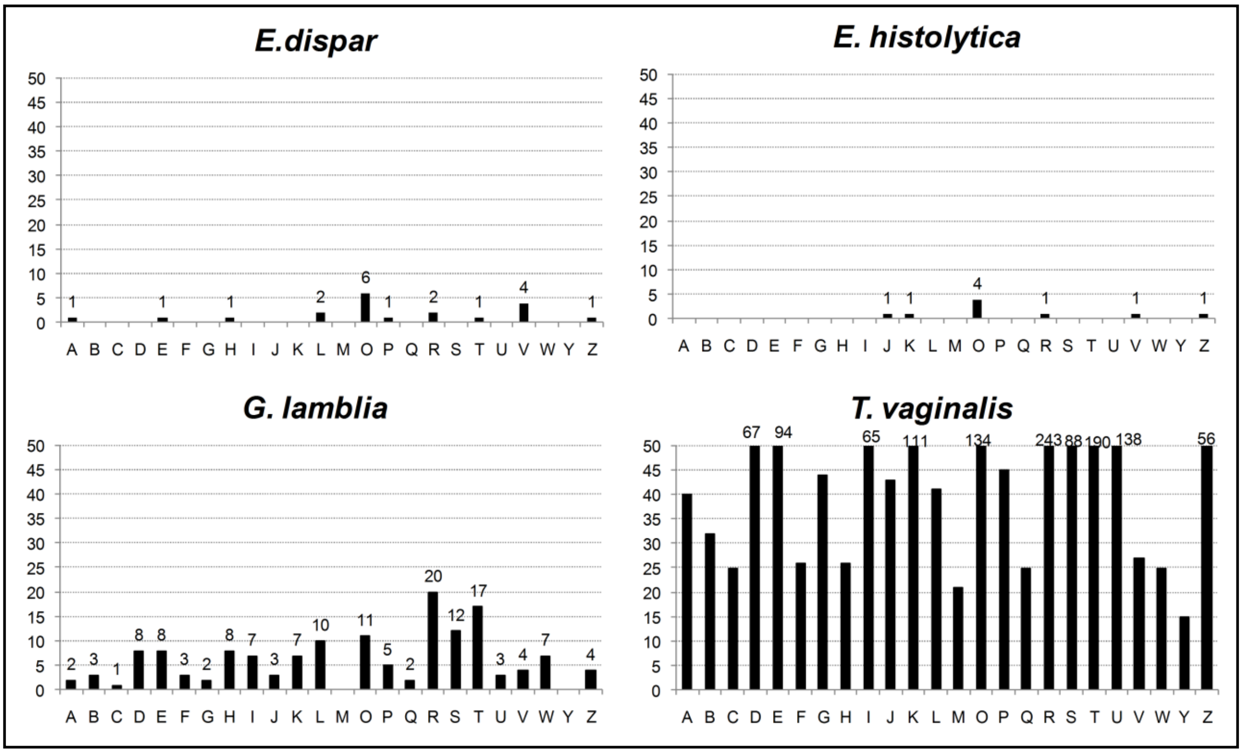


Figure 8. Functional characterization of the paralog groups in *E. histolytica, E. dispar, G. lamblia* and *T. vaginalis* using KOG/NCBI. (*E. dispar* N = 136; *E. histolytica* N = 58; *G. lamblia* N = 255; *T. vaginalis* N = 2933). Legend: Information Storage and Processing: [J] Translation, ribosomal structure and biogenesis; [A] RNA processing and modification; [K] Transcription; [L] Replication, recombination and repair; [B] Chromatin structure and dynamics. Cellular Processes and Signaling: [D] Cell cycle control, cell division, chromosome partitioning; [Y] Nuclear structure; [V] Defense mechanisms; [T] Signal transduction mechanisms; [M] Cell wall/membrane/envelope biogenesis; [Z] Cytoskeleton; [W] Extracellular structures; [U] Intracellular trafficking, secretion, and vesicular transport; [O] Posttranslational modification, protein turnover, chaperones. Metabolism: [C] Energy production and conversion; [G] Carbohydrate transport and metabolism; [E] Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [H] Coenzyme transport and metabolism; [I] Lipid transport and metabolism; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport and catabolism. Poorly characterized: [R] General function prediction only; [S] Function unknown.

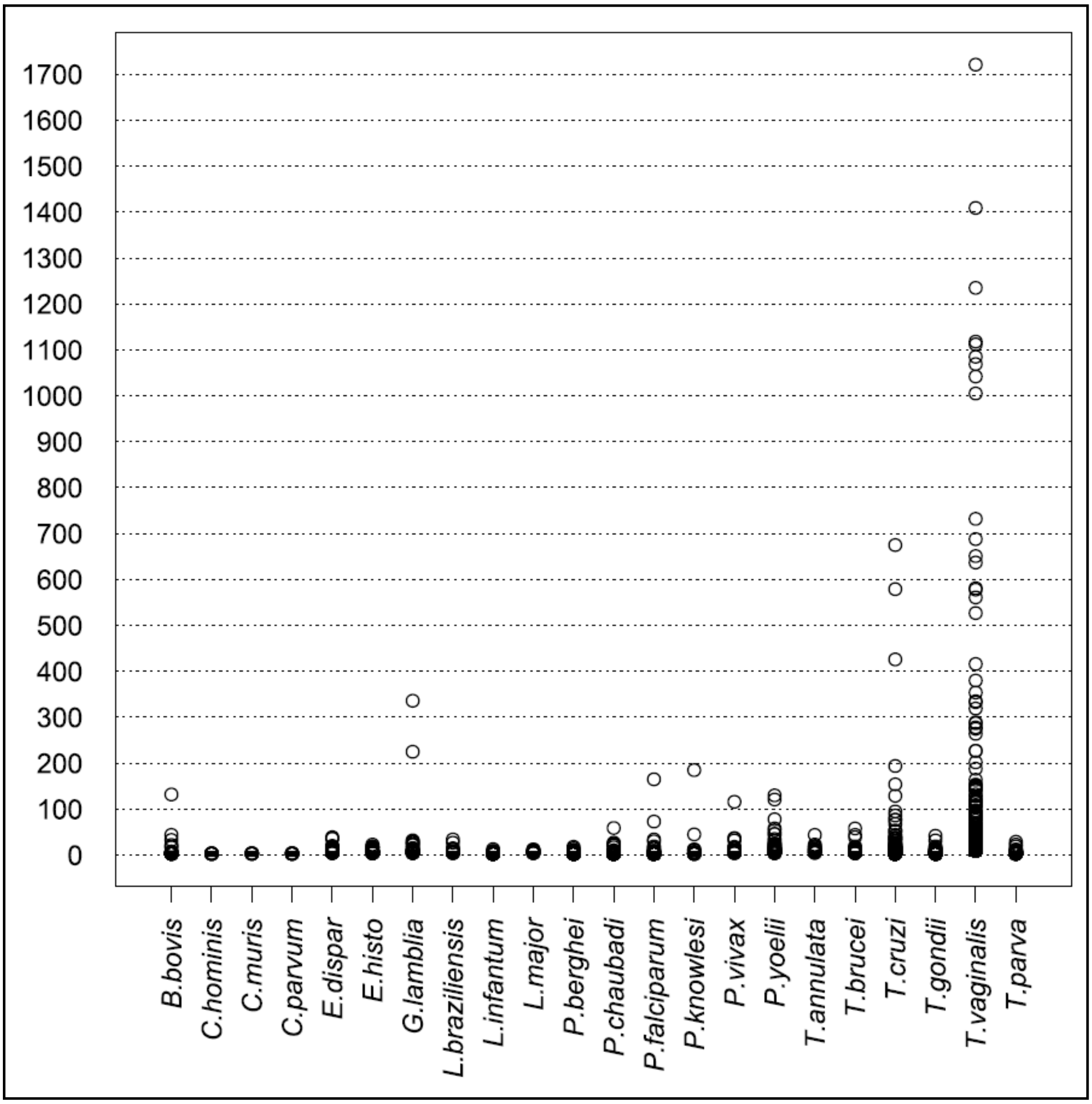


Figure 9: Number of paralog groups (co-orthologs and paralogs) identified for each species (N = 12661 families). Only the larger expansions, that is, those with at least 80 members have a function listed. *B. bovis*: variant erythrocyte surface antigen-1. *G. lamblia*: Kinase, NEK; Variant-specific surface protein. *P. chabaudi*: Pc-fam-2 protein, *P. falciparum*: rifin, var. *P. knowlesi*: SICA-like antigen. *P. vivax*: variable surface protein Vir24. *P. yoelii*: hypothetical protein. *T. brucei*: expression site associated gene (ESAG) 4 protein; variant surface glycoprotein (VSG). *T. cruzi*: trans-sialidase, mucin TcMUCII, dispersed gene family protein 1 (DGF-1) retrotransposon hot spot (RHS) protein. *T. vaginalis*: ankyrin repeat protein, hypothetical protein: TVAG_289600, TVAG_580790 e TVAG_235700.

Tables:

Table 1: Alphabetically ordered experiment dataset summary.

Species name / Strain	**Non-redundant proteins**	**NCBI proteins**	**Literature proteins**	**% proteome used**	**Number proteins reference**
Babesia bovis T2Bo	3691	3713	3671	100.54	[33]
Crypstosporidium parvum Iowa II	3829	7667	3807	100.58	[34]
Cryptosporidium hominis TU502	3885	3885	3994	97.27	[35]
Cryptosporidium muris RN66	3930	7868	3934	99.9	site= Search=6061
Entamoeba dispar SAW760	8606	17622	8748	98.38	[36]
Entamoeba histolytica HM1:IMSS	7973	16372	8201	97.22	[37]
Giardia lamblia ATCC 5080*3*	7163	13939	6470	110.71	[38]
Leishmania braziliensis MHOM BR 75 M2904	7825	7897	8153	95.98	[39]
*Leishmania infantum*JPCM5	7872	7992	8154	96.54	[39]
Leishmania major Friedlin	8003	16365	8298	96.44	[39]
Plasmodium berghei ANKA	9730	9840	5864	165.93	[40]
Plasmodium chabaudi chabaudi AS	14639	27045	5698	256.91	[41]
Plasmodium falciparum 3D7	5876	11669	5268	111.54	[42]
Plasmodium knowlesi strain H	5102	10213	5188	98.34	[43]
Plasmodium vivax SaI 1	5397	5413	5433	99.34	[44]
Plasmodium yoelii yoelii 17XNL	7303	7357	5878	124.24	[45, 46]
Theileria annulata Ankara	3790	3795	3792	99.95	[47]
Theileria parva Muguga	4050	4061	4035	100.37	[48]
Toxoplasma gondii ME49	7984	15983	8032	99.4	[49]
Trichomonas vaginalis G3	50189	119360	59681	84.1	[50]
Trypanosoma brucei treu927	8540	8768	9068	94.18	[51]
Trypanosoma cruzi CL Brene*r*	19247	19644	12000	160.39	[52]

Species name. Non-redundant protein number used for each species in this study. Protein number downloaded from NCBI. Proteome size according literature. Proteome size reference.

Table 2: Number of homolog, ortholog and paralog groups for each species, decreasing order, according to the number of homolog groups.

Species name	Homolog groups	Ortholog groups	Paralog groups
Trypanosoma cruzi	7647	6833 (89.3%)	814 (10.7%)
Leishmania major	7502	7493 (99.9%)	9 (0.01%)
Leishmania infantum	7366	7358 (99.9%)	8 (0.01%)
Leishmania braziliensis	7098	7085 (99.8%)	13 (0.02%)
Trypanosoma brucei	6549	6454 (98.5%)	95 (1.5%)
Entamoeba dispar	6079	5943 (97.8%)	136 (2.2%)
Entamoeba histolytica	6030	5972 (99.0%)	58 (1.0%)
Plasmodium chabaudi	5018	4838 (96.4%)	180 (3.6%)
Plasmodium berghei	5002	4906 (98.1%)	96 (1.9%)
Trichomonas vaginalis	4708	1775 (37.7%)	2933 (62.3%)
Plasmodium vivax	4694	4654 (99.1%)	40 (0.9%)
Plasmodium knowlesi	4683	4667 (99.7%)	16 (0.3%)
Plasmodium yoelii	4679	4599 (98.3%)	80 (1.7%)
Plasmodium falciparum	4607	4553 (98.8%)	54 (1.2%)
Cryptosporidium parvum	3649	3646 (99.9%)	3 (0.01%)
Toxoplasma gondii	3521	3404 (96.7%)	117 (3.3%)
Cryptosporidium hominis	3518	3516 (99.9%)	2 (0.01%)
Cryptosporidium muris	3427	3425 (99.9%)	2 (0.01%)
Theileria parva	3343	3318 (99.3%)	25 (0.7%)
Theileria annulata	3282	3266 (99.5%)	16 (0.5%)
Babesia bovis	2902	2872 (99.0%)	30 (1.1%)
Giardia lamblia	1436	1181 (82.2%)	255 (17.8%)

Species name. Number of homolog (orthologs and paralogs) groups assigned by OrthoMCL for each species. Number of ortholog groups assigned by OrthoMCL. Number of paralog groups assigned by OrthoMCL, by species.

Table 3: Number of total ortholog, orthologs groups and paralog groups for each species, decreasing order, according to the number of total ortholog groups.

Species name	Total ortholog groups	Ortholog groups (%)	Co-ortholog groups (%)
Leishmania major	7493	7237 (96.58%)	256 (3.42%)
Leishmania infantum	7358	7112 (96.66%)	246 (3.34%)
Leishmania braziliensis	7085	6816 (96.2%)	269 (3.8%)
Trypanosoma cruzi	6833	1870 (27.37%)	4963 (72.63%)
Trypanosoma brucei	6454	6147 (95.24%)	307 (4.76%)
Entamoeba histolytica	5972	5456 (91.36%)	516 (8.64%)
Entamoeba dispar	5943	5370 (90.36%)	573 (9.64%)
Plasmodium berghei	4906	4164 (84.88%)	742 (15.12%)
Plasmodium chabaudi	4838	3870 (79.99%)	968 (20.01%)
Plasmodium knowlesi	4667	4588 (98.31%)	79 (1.69%)
Plasmodium vivax	4654	4562 (98.02%)	92 (1.98%)
Plasmodium yoelii	4599	4258 (92.59%)	341 (7.41%)
Plasmodium falciparum	4553	4076 (89.52%)	477 (10.48%)
Cryptosporidium parvum	3646	3575 (98.05%)	71 (1.95%)
Cryptosporidium hominis	3516	3423 (97.35%)	93 (2.65%)
Cryptosporidium muris	3425	3356 (97.99%)	69 (2.01%)
Toxoplasma gondii	3404	3149 (92.51%)	255 (7.49%)
Theileria parva	3318	3179 (95.81%)	139 (4.19%)
Theileria annulata	3266	3175 (97.21%)	91 (2.79%)
Babesia bovis	2872	2779 (96.76%)	93 (3.24%)
Trichomonas vaginalis	1775	827 (46.59%)	948 (53.41%)
Giardia lamblia	1181	1057 (89.5%)	124 (10.5%)
Total	21119	13340	7679

Name of the species used in the study. Number of ortholog groups assigned by OrthoMCL, for each species. Total number and percentage of orthologs groups. Total number and percentage of orthologs and co-orthologs groups.

Table 4: Number of paralog groups assigned by OrthoMCL, decreasing order, according to the total number of paralog groups.

Species name	Total	Co-ortholog groups	Paralog groups
Trypanosoma cruzi	5777	4963 (85.91%)	814 (14.09%)
Trichomonas vaginalis	3881	948 (24.43%)	2933 (75.57%)
Plasmodium chabaudi	1148	968 (84.32%)	180 (15.68%)
Plasmodium berghei	838	742 (88.54%)	96 (11.46%)
Entamoeba dispar	709	573 (80.82%)	136 (19.18%)
Entamoeba histolytica	574	516 (89.90%)	58 (10.10%)
Plasmodium falciparum	531	477 (89.83%)	54 (10.17%)
Plasmodium yoelii	421	341 (81.00%)	80 (19.00%)
Trypanosoma brucei	402	307 (76.37%)	95 (23.63%)
Giardia lamblia	379	124 (32.72%)	255 (67.28%)
Toxoplasma gondii	372	255 (68.55%)	117 (31.45%)
Leishmania braziliensis	282	269 (95.39%)	13 (4.61%)
Leishmania major	265	256 (96.60%)	9 (3.40%)
Leishmania infantum	254	246 (96.85%)	8 (3.15%)
Theileria parva	164	139 (84.76%)	25 (15.24%)
Plasmodium vivax	132	92 (69.70%)	40 (30.30%)
Babesia bovis	123	93 (75.61%)	30 (24.39%)
Theileria annulata	107	91 (85.05%)	16 (14.95%)
Cryptosporidium hominis	95	93 (97.89%)	2 (2.11%)
Plasmodium knowlesi	95	79 (83.16%)	16 (16.84%)
Cryptosporidium parvum	74	71 (95.95%)	3 (4.05%)
Cryptosporidium muris	71	69 (97.18%)	2 (2.82%)
Total		4982	7679

Species name. Number of duplicated (co-orthologs and paralogs) groups assigned by OrthoMCL for each species. Number of co-ortholog groups assigned by OrthoMCL. Number of paralog groups assigned by OrthoMCL, by species

Table 5: Number of orphan proteins assigned by OrthoMCL, decreasing order, according to the number of orphan proteins.

Species name	Orphans	% Orphan / proteome
Plasmodium chabaudi	6961	47.55
Trichomonas vaginalis	4309	8.59
Giardia lamblia	3585	50.05
Plasmodium berghei	3134	32.21
Toxoplasma gondii	2997	37.54
Plasmodium yoelii	1045	14.31
Trypanosoma brucei	994	11.64
Trypanosoma cruzi	690	3.58
Entamoeba dispar	558	6.48
Entamoeba histolytica	442	5.54
Cryptosporidium muris	347	8.83
Theileria parva	318	7.85
Babesia bovis	306	8.29
Cryptosporidium hominis	221	5.69
Plasmodium falciparum	211	3.59
Plasmodium vivax	159	2.95
Theileria annulata	159	4.20
Leishmania braziliensis	142	1.81
Leishmania infantum	82	1.04
Crypstosporidium parvum	81	2.12
Leishmania major	43	0.54
Plasmodium knowlesi	38	0.74
Total	26822	

Species name. Total amount of orphan proteins for each species. Proportion between the total amount of orphan proteins versus species proteome size.

5 - Artigo 3: “Genômica comparativa do parasita *Leishmania amazonensis*”

Tschoeke DA, Nunes GL, Jardim R, Lima JA, Aline S. R. Dumaresq ASR,. Gomes RG, Pereira LM, Loureiro DR, Stoco PH, Guedes HLM, de Miranda AB, Ruiz J, Pitaluga AN, Floriano P. Silva Jr FP, Probst CM, Mottram J, Grisard EC, Dávila AMR. **The comparative genomics of *Leishmania amazonensis* parasite.**

O artigo foi submetido para a revista *Evolutionary Bioinformatics.*

Subject: Your Submission: Reply Required
Date: Sat, 30 Nov 2013 09:47:00 -0600 (CST)
From: Jan McIver<jan.mciver@libertasacademica.com>
Reply-To: Jan McIver <jan.mciver@libertasacademica.com>
To: davila@ioc.fiocruz.br

Dear Dr Davila
Thank you for your submission to Evolutionary Bioinformatics, titled The comparative genomics and phylogenomics of Leishmania amazonensis parasite .

Este artigo corresponde aos objetivos específicos 2.2.8, 2.2.9 e 2.2.10: Anotar o genoma de *Leishmania amazonensis;* Categorização funcional do genoma de *L. amazonensis* e; Analisar comparativamente o genoma de *Leishmania amazonensis* e de outras cinco espécies de *Leishmania* spp.

O presente artigo faz uma análise do genoma de *Leishmania amazoensis*, com o suporte do sistema de anotação STINGRAY, além de trazer uma análise comparativa do proteoma de *L. amazonensis* contra os proteomas de *L. braziliensis*, *L. donovani*, *L. infantum, L. major* e *L. mexicana*. Nesta análise observamos que o proteoma núcleo do gênero *Leishmania* é constituído por 7016 ortólogos e, além disso, apresenta poucas duplicações específicas e poucos genes espécie-específicos.

The comparative genomics and phylogenomics of *Leishmania amazonensis* parasite

14,408 words

Diogo A. Tschoeke[1,9§], Gisele L. Nunes[9§], Rodrigo Jardim[1,9], Joana Lima[9], Aline S. R. Dumaresq[9], Monete R. Gomes[9], Leandro de Mattos Pereira[9], Daniel R. Loureiro[1], Patricia H. Stoco[8], Herbert Leonel de Matos Guedes[2,7], Antonio Basilio de Miranda[1,9], Jeronimo Ruiz[3,1], André Pitaluga[4], Floriano P. Silva Jr[1,5], Christian M. Probst[6,1], Nicholas J. Dickens[7], Jeremy C. Mottram[7], Edmundo C. Grisard[8], Alberto M. R. Dávila[1,9*]

[1] Pólo de Biologia Computacional e Sistemas, Oswaldo Cruz Institute (FIOCRUZ/IOC), Rio de Janeiro, RJ, Brazil. E-mail: diogoat@ioc.fiocruz.br (D.A.T), gilopesnunes@gmail.com (G.L.N), davila@fiocruz.br (A.M.R.D)

[2] Laboratório de Inflamação Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro

[3] Instituto René Rachou, IRR, FIOCRUZ-MG

[4] Laboratório de Biologia Molecular de Parasitas e Vetores, Instituto Oswaldo Cruz, FIOCRUZ, Rio de Janeiro, RJ, Brazil

[5] Laboratório de Bioquímica de Proteínas e Peptídeos, Instituo Oswaldo Cruz, FIOCRUZ, Rio de Janeiro, RJ, Brazil

[6] Instituto Carlos Chagas, ICC - FIOCRUZ-PR

[7] Wellcome Trust Centre for Molecular Parasitology, Institute of Immunity, Infection and Inflammation, College of MVLS, University of Glasgow, 120 University Place, Glasgow, G12 8TA, UK

[8] Laboratório de Protozoologia, Universidade Federal de Santa Catarina, Florianópolis, SC, Brazil

[9] Laboratório de Biologia Computacional e Sistemas, Instituto Oswaldo Cruz, FIOCRUZ, Rio de Janeiro, RJ, Brazil

*Corresponding author. Mailing address: Laboratório de Biologia Computacional e Sistemas, FIOCRUZ/IOC, Av. Brasil, 4365, Manguinhos, CEP 21040-360, Rio de Janeiro, RJ, Brazil. Phone (+55 21) 2562-1025. E-mail: davila@fiocruz.br

[§] Authors contributed equally

We think that this manuscript is appropriate for this journal, because our *L. amazonensis* analysis was centered on comparative and phylogenomics analysis, which is one of the *Evolutionary Bioinformatics* focus. And open access *EB* gives us more visibility, since not all research have access to pay-per-read scientific magazines.

Abstract:

Leishmaniasis is an infectious disease caused by *Leishmania* spp. *Leishmania amazonensis* is considered a new world *Leishmania* species, belonging to the Mexicana complex, and can cause all types of leishmaniasis infections. The *L. amazonensis* reference strain MHOM/BR/1973/M2269 was sequenced in Solexa sequencer (Illumina) resulting in 29,670,588 bases assembled. Furthermore, 8802 CDS were identified and most of them have hypothetical function. Comparative analysis into *Leishmania* genus showed that these 6 species share 7016 orthologs, whilst *L. amazonensis* and *L. mexicana* has the biggest number of specific orthologs and *L. braziliensis* biggest number of inparalogs. Phylogenomic analysis showed the expected *L. amazonensis* taxonomic position together with *L. mexicana* forming the “Mexicana complex” and the New and Old *Leishmania* (L). spp. separation. Potential analogous proteins were found between *L. amazonensis* and *H. sapiens*, and also into the *L. amazonensis* proteome. Finally, RNAi analysis showed that *L. amazonensis*, probably, does not have a functional RNAi pathway.

Introduction

Leishmaniasis is an infectious disease caused by *Leishmania spp.* parasites and has worldwide impact with considerable morbidity and mortality rates, mainly in developing countries. The lack of vaccine and effective treatments is a major problem, since most of the drugs available, e.g. pentavalent antimony, are toxic or may cause side effects [1]. This disease is distributed in 88 countries and it is estimated that more than 12 million people are currently infected with *Leishmania sp.* Around 350 million people living in these areas of risk (poor rural and suburban zones) are vulnerable to infection [2,3]. Around 90% of global cases occur in the following countries: Afghanistan, Algeria, Brazil, India, Iran, Nepal, Peru, Saudi Arabia, Sudan and Syria [2].

At least 20 *Leishmania* species are described and can cause a variety and complex group of human diseases, characterized by symptoms that range from cutaneous lesions to fatal visceral leishmaniasis (VL) depending on the species and the host immune response [4–6]. The most severe form is VL, caused by *L. donovani* and *L. infantum*, where parasites affects mainly liver and spleen, resulting in host: (i) immunesuppression; (ii) progressive fever; (iii) weight loss and; (iv) anemia, that can be fatal in absence of an efficient treatment[2,7]. In cutaneous leishmaniasis (CL), the parasites (*L, major*, *L. tropica* and *L. aethiopica*) cause localized long-term ulceration, inducting a chronicity, latency and tendency to metastasize in the human host [8]. The mucocutaneous leishmaniasis (MCL), caused mainly by *L. braziliensis*, induces the destruction of nasopharyngeal tissue with hideous disfiguring lesions. The last type of leishmaniasis, diffuse cutaneous leishmaniasis (DCL), caused by parasites *L. amazonensis*, *L. guyanensis* and *L. aethiopica*, is a long-lasting disease due to a deficient cellular-mediated immune response presenting a progressive primary lesion and multiple metastatic lesions[9–11].

L. amazonensis is related to CL and DCL (rare manifestation), however, it was isolated from patients with MCL, VL and post kalaazar dermal leishmaniasis. All types of leishmaniasis can be caused by *L. amazonensis* infections [4]. This diversity of disease tropism emphasizes the importance of the availability of =its genome sequence. In Brazil, one of the most studied *L. amazonensis* strains is PH8 (IFLA/BR/67/PH8) because it is the component of Leishvaccine[12]. PH8 was isolated from sand fly then it was not chosen for this study because an isolate from human disease was needed. *Leishmania* (L.) *amazonensis* (MHOM/BR/71973/M2269) is a strain extracted from a single cutaneous lesion, the most common for a *L. amazonensis* infection then was chosen for this study.

The parasite's life cycle consists of two forms: promastigotes (found in the invertebrate host) and amastigotes (found in the vertebrate host). Sandflies become infected by ingesting cells with amastigotes forms during blood meals. In sandflies, amastigotes transform into promastigotes, which live within the midgut or hindgut of the sandfly vector that is differentiated into metacyclic forms in the invertebrate and transferred to vertebrate host by bites of female phlebotomine during blood meals. Promastigotes that reach the puncture wound are phagocytized by macrophages, and then differentiate into amastigotes inside the phagolysosome of host macrophages (vertebrate host) proliferating and infecting others macrophages cells[13,14].

In Brazil, CL is an endemic disease caused by at least six *Leishmania* species from the subgenus Viannia and *Leishmania*. The main agents of CL in the south of the Amazon basin are *L.* (Viannia) *braziliensis* and *L.* (Leishmania) *amazonensis*, showing no differences in clinical manifestations, however, the effects of the diseases are different [15,16]. *L. amazonensis* is considered a new world *Leishmania* species belonging to the *L. mexicana* complex that contains only 34 chromosomes (Chr) due to the fusion of chromosome 8 with 29 and 20 with 36 [17]. On the other hand, *L. major* (strain Friedlin) and *L. infantum*, considered old world leishmanias, have 36. Other new world *Leishmania* species such as *L. braziliensis*, which belongs to Viannia subgenus, are composed by 35 chromosomes (Chr 20+36 fused) [18].

Many advances have occurred during the last decade in the genomic area, mainly after the development of high-throughput sequencing. In the case of parasitic protozoans, several trypanosomatids genomes have been sequenced, among them: *Trypanosoma cruzi, T. brucei, L. major, L. infantum, L. braziliensis* and *L. mexicana* [19–25]. *L. major* was the first *Leishmania* genome sequenced showing 32.8 Mbp size and 8311 predicted protein-coding genes (CDS) [19]. A comparative analysis using three *Leishmania* species (*L. major, L infantum* and *L. braziliensis*) was done in 2007, revealing that the genomic organization among the studied species is highly conserved, showing a genome with an average of 8300 genes and highly syntenic for more than 99% of the genes. However, about 200 genes showed differential distribution between three evaluated genomes, with 47, 27 and 5 exclusive genes (or species-specific or unique) for *L. braziliensis, L. infantum* and *L. major*, respectively[22]. The most recently sequenced *Leishmania* genome was *L. mexicana* in 2007 and published in 2011[25] and in the middle of our study, *L. amazonensis* genome was published by Real and colleagues[26] in 2013.

The small number of species-specific differences detected between *Leishmania* genomes is generally related to differentially distributed genes or conserved genes differentially regulated [27]. Researchers have investigated those differences to reach a better understanding of virulence and pathogenicity of a particular species and observed that the differential expression of the genes mainly in amastigotes may be important for parasite survival and maintenance inside host[19,27,28].

Considering the Tri-Tryp genomes (*T. cruzi*, *T. brucei* and *L. major*), approximately 6,200 genes are conserved among the 3 species and show synteny in 94% of those genes [29]. Most of the species-specific genes appear on non-syntenic regions/chromosomes and consist in members of large surface antigen families [30]. This indicates that differences detected among these parasites and pathogenesis are associated with few species-specific genes, which are mostly described as uncharacterized [22,29]. The genomes of the *Leishmania* species contain a conserved number of genes, estimated at 8,200 [31]. While comparative genomic analysis revealed highly conserved gene synteny, an estimated divergence of 46 million years has been reported [22,32]. *L. braziliensis* shows a notable difference because it contains a putative RNA interference pathway and two types of transposons (TATES) and retroposons (SLACS) absent in *L. major* and *L. tarentolae* [31].

Here we analyzed the genome of the *L. amazonensis* in comparison with the *L. mexicana*, *L. major, L. infantum and L. braziliensis* genomes, all available in GenBank. The recently published *L. amazonensis* genome [26] was not analyzed because the data was not available at the time we finalized our analyses. Our study intends to provide further information on the *L. amazonensis* genome using comparative genomics and phylogenomics approaches, which are less explored in the recently published *L. amazonensis* study.

Material and Methods

DNA preparation and sequencing

Leishmania amazonensis reference strain MHOM/BR/1973/M2269, kindly provided by Dr Paul Bates, was selected for this study. DNA was extracted using a Quiagen QIAamp DNA Kit, according to the manufacturer's instructions. The extracted DNA was sequenced in Solexa sequencer (Illumina) using paired-end reads of 50 + 50 bases.

Assembly, sequence analysis and annotation

Genomic sequences were trimmed for platform dependent systematic errors then quality was evaluated using Phred (cutoff Q=26) [33,34]. High quality reads were assembled using Velvet version 0.7.55 software [35]. *Ab initio* and reference genome assembly strategies were applied using the publicly available *L. mexicana* genome (GenBank Assembly ID: GCA_000234665.4 and RefSeq Assembly ID: GCF_000234665.1) and *L. major* genome (GenBank Assembly ID: GCA_000002725.2 and RefSeq Assembly ID: GCF_000002725.2), then assemblies were merged using in house developed Perl scripts. Contigs were created by comparing assembled scaffolds and contigs with current available *Leishmania* spp genomes cited above.

The multi-fasta files from *L. amazonensis* genome assembled by us were submitted to STINGRAY pipeline (http://stingray.biowebdb.org) which is an improved version of the original GARSA [36] system. The semi-automatic annotation was performed using STINGRAY and a TblastX [37] approach by transferring *L. mexicana* [25] annotation to *L. amazonensis*, then this annotation was improved by the identification of conserved domains.

Protein families and domain identification

Pfam-A (v. 26.0) [38,39] and Hmmer 3.0 [40] were used against the 8802 predicted proteins inferred in *L. amazonensis*, by using hmmsearch program with e-value 1e-5 and other default parameters.

Gene Ontology inference

The 8802 *L. amazonensis* proteins were also analyzed using Gene Ontology (GO) [41]. Briefly, similarity analysis were performed using Blastp (v. 2.2.23)[37] against GO database (go_20130223-seqdb.fasta), and then proteins were classified within one of three GO categories, as follow: (i) Biological process refers to a biological objective to which the gene or gene product contributes, (ii) Molecular function is defined as the biochemical activity, including specific binding to ligands or structures, of a gene product, (iii) Cellular component refers to the place in the cell where a gene product is active.

Conserved domain identification

Conserved domains were identified using RpsBlast (v. 2.2.23) [37] analysis on the 8802 proteins inferred in *L. amazonensis* against seven databases simultaneously (CDD.v3.10 – Conserved Domain Database; COG.v1.0 - Cluster of Orthologous Groups; KOG.v1.0 - Cluster of Eukariotics Orthologous Groups; Pfam.v26.0 - Protein Family; PRK.v6.0; SMART.v6.0; TIGR.v13.0) with e-value 1e-05.

Identification of Orthologous and Paralogous groups

The identification of orthologous proteins was performed using results generated by the OrthoMCL v.1.4 software [42]. Orthologous proteins shared by all six *Leishmania* species (*L. major, L. infantum, L. donovani, L. braziliensis, L. mexicana* and *L. amazonensis*) were inferred. The protein function of those inferred orthologs was semi-automatically transferred from previously annotated *Leishmania* genomes. Inparalogs and recent paralogs proteins in *L. amazonensis* were also identified by extracting specific parts of the output file generated by the OrthoMCL software.

The orthologous proteins shared between the different *Leishmania* species as well as the inparalogous proteins from *L. amazonensis* and from others species, were used to generate a Venn diagram using R software [43].

Putative orphans proteins identification

To find putative orphan proteins, i.e., not homologous to any protein in this study, we generated a first list of proteins identifiers that were used as input to OrthoMCL and constructed a second list with proteins identifiers clustered by OrthoMCL (the OrthoMCL output). Then, these two lists (Submitted versus Clusterized) were compared using a script written in Ruby language to obtain the identifiers of the putative orphan proteins. Since those potential orphans are based on a universe of only 6 *Leishmania* genomes, and to minimize a possible misclassification, we performed a BlastP search (v. 2.2.28+)[37] with those putative orphans proteins against RefSeq database (r.56 18,132,578 sequences). This steps helped us to identify proteins that were classified as an putative orphans but had similarity to prokaryotic or other eukaryotic (non-*Leishmania*) protein. Finally, proteins without matches to Refseq database were considered orphan proteins in this study.

Phylogenomics to infer species tree

The phylogenomic tree was inferred based on the study by Ocaña & Dávila (2011)[44] and Ciccarelli and colleagues (2006) [45] for the selection and construction of the species tree. Thirty-one universal orthologous (UO) genes showing 1:1 orthologous relationships were used. Those UO were originally identified by Ciccarelli (2006)[45] showing the following characteristics: i) present in all complete genomes available at Genbank until 2006, ii) not involved in horizontal transfer. Since these 31 UO have direct correspondence in the protozoan genome data available at RefSeq and ProtozoaDB [46], they were mapped to the *Leishmania major* proteins using (a) the best blast hits (e-value, 1e-50), and (b) manual verification of the annotation (the RefSeq annotation of the best hits needed to match the UO annotation). Once mapped, the *L. major* protein sequences corresponding to those 31 UO were searched in the orthologous groups identified in the 6 *Leishmania* species by OrthoMCL, then each of those 31 orthologs mapped were exported as multifasta files and aligned using Mafft v5.861 [47] with default parameters.

A supermatrix tree was obtained using concatenated multiple alignments from entire protein sequences. The individual alignments were concatenated using an in-house perl script, resulting in a global supermatrix of 9,450 positions for the six species. The resulting supermatrix was used to generate the phylogenomic tree with MEGA 5 [48], inferred by Maximum Likelihood using 1000 bootstrap replicates. However, because it is not simple to use multiple models in a single (concatenated) alignment, we decided to adopt JTT that was also the model adopted in the phylogenomics studies of Ciccarelli (2006)[45] and Ocanã & Dávila (2011)[44]. JTT model assumes that there are two classes of sites, one class being invariable and the other class being free to change [49].

Intragenomic and intergenomic non-homologous isofunctional enzymes (NISE) identification

In order to identify possible cases of intragenomic NISE in the genome of *L. amazonensis* and intergenomic NISE between this genome and the genome of *Homo sapiens*, we applied methodologies described previously [50–52]. Briefly, protein sequences of enzymes with the same functional activity were downloaded and grouped in accordance to its functional activity as determined by the classification from the International Union of Biochemistry and Molecular Biology, the Enzyme Commission number (EC number) [53]. Protein sequences and functional classification were obtained from KEGG (Kyoto Encyclopedia of Genes and Genomes - version 58.1) [54]. After grouping, we performed a step to confirm the functional activity assigned by KEGG. First, from the 8802 predicted proteins for *L. amazonensis* we removed sequences with less than 60 amino acids, since they may represent protein fragments, forming a dataset with 8575 predicted proteins. Then, inside each group of functionally related proteins, their primary structures were compared in a pairwise, all-against-all manner, using Blastp. Functional activities were then confirmed via the AnEnPi's module [50], which classifies the enzymes in accordance to the EC number. This classification is obtained after parsing the results of Blastp, using the dataset of predicted proteins from *L. amazonensis* as query and the groups previously obtained as subject. A restrictive e-value of 10^{-20} was used as a threshold [51,52,55] to include a given primary structure in a certain group or cluster. Proteins were considered to be NISE if, inside each group of functionally related enzymes, they were allocated in different clusters after parsing the results from Blastp. Possible analogy cases were verified by the examination of their folding categories as classified by the SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/)[56] and SUPERFAMILY databases (supfam.cs.bris.ac.uk) [57]. Further refinement of confirmed NISEs was achieved by 3D structure prediction of *L. amazonensis* proteins by homology modeling and structural comparison with their human analogous counterparts (see below).

Homology modeling of L. amazonensis proteins and comparative structural analysis with human proteins

Modeling of the three-dimensional (3D) structure of the selected NISEs of *L.*

amazonensis was performed by the satisfaction of spatial restraints method implemented in the program Modeller 9v10 [58]. Global pair-wise sequence alignment between target *L. amazonensis* sequences and the respective templates was performed with the program needle (Needleman-Wunsch) program within the EMBOSS v.6.3.1 package [59]. The models were constructed using as templates the atomic coordinates of PDB ids listed in Table 5 for each of the selected analogy cases. Ten models were generated for each protein target sequence and the model with most favorable DOPE-score and the lowest Modeller objective function value was subjected to external assessment of the stereochemical and overall structural quality within the Structural Analysis and Verification Server (SAVES v.4) at the WWW address http://services.mbi.ucla.edu/SAVES/. All models selected for further analysis had at least 95% of residues in the most favorable and additionally allowed regions of Ramachandran plots along with other reasonable stereochemical quality parameters. Inspection of molecular structures and other structural analysis were performed in Sybyl X-1.3 software (Tripos L.P., St. Louis, MO).

L. amazonensis genome functional categorization

In order to briefly know the content of *L. amazonensis* we performed a functional categorization through similarity analysis using Blast and RpsBlast programs against the database of orthologous genes in Prokaryotes (COG/NCBI) and Eukaryotic orthologous genes (KOG/NCBI)[60,61], which are classified in functional categories (ftp://ftp.ncbi.nih.gov/pub/COG/COG/fun.txt). We use a cut-off e-value of 1e-5 in both programs and databases, in order to infer which functional category each protein belongs to. Plots of the functional categories were created with R software.

L. amazonensis proteome was also characterized by PFAM (v. 26.0) [39] and by Conserved Domain Database (v 3.10) through RpsBlast. A further analysis was performed using in-house perl scripts to identify (i) which genes were identified only by PFAM with Hmmer 3, (ii) which ones were identified only by “Conserved Domains” (CDD) and (iii) which ones were characterized by both of them (PFAM and CDD) .

The *Leishmania* core proteome identification

The *Leishmania* Core Proteins (LCP) were identified and analyzed among the orthologous groups and defined as orthologous proteins shared by the six species *Leishmania* studied. To find the LCP, the OrthoMCL results were analyzed, then only orthologs shared by the "*6 taxa*" were choosen. LCP functions were accessed through annotation provided with the sequences.

Leishmania amazonensis database

The contigs generated from the assembly of sequencing reads, the genes and proteins found from these contigs are all available for public consulting in the STINGRAY pipeline (http://stingray.biowebdb.org). Furthermore *L. amazonensis* contigs were submitted to GenBank under BioProject ID PRJNA221875, locus_tag prefix Q771.

The *Leishmania amazonensis* RNAi identification

L. amazonensis RNAi genes identification was performed thought Blast analysis against *L. amazonensis* genome, using as template/query RNAi genes from *Leishmania* species and *T. brucei* genes annotated as participants of the RNAi pathway in GeneDB database [62] (www.genedb.org). Genes related do RNAi machinery in *L. amazonensis* found using this strategy were then submitted to a phylogenetic analysis using MEGA 5 [48] and a tree was inferred by Neighbor-Joining using 1000 bootstrap replicates.

Results

Sequencing, assembly and genome characteristics

In the present study, the genome from a reference strain (MHOM/BR/1973/M2269) of *L. amazonensis,* isolated from patients with simple lesions, was sequenced by whole-genome shotgun sequencing using Solexa Genome Analyzer with a ten-fold coverage and pair-end reads (50+50 bases). Final sequences were produced by aligning contigs against the *L. mexicana* genome as reference to generate an assembled high-quality genome. The assembly resulted in 10305 contigs, with approximately 59% GC content. The smallest and largest detected contigs had 96 bases and 141211 bases, respectively, with a mean of 2879 bp and median of 853 bp (Table 1). We found 8802 protein-coding genes in the *L. amazonensis* genomes using TblastX against *L. mexicana* and Refseq database. The biggest coding regions had 19872 bp and the smallest only 66 bp with median and mean of 1637 bp and 1209 bp, respectively (Figure 1). The GC content for coding regions was 61.1%. Of these 8802 putative proteins, 887 did not was clusterized by OrthoMCL. However, Blast analysis was performed using the putative orphan proteins against Refseq database with e-value 1e-5 and resulted in 14 proteins classified as orphans in this study (Table 3). Furthermore we found some genes occurring in single copy, such as: Ribosomal Protein S2 (rpS2) and Ribosomal Protein L7. Moreover some genes exhibit multiple copies, such as: ABC transporter (50 copies) and Calpains (44 copies). Nonetheless, 63% of the CDS were annotated as proteins with hypothetical function (Table 2).

Functional analysis of *L. amazonensis* proteins

Among protein functions characterized with Gene Ontology, the most abundant molecular functions found were: protein binding (9%), nucleotide binding (8%), metal ion binding (5%), receptor activity (4%), DNA binding (4%), signal transducers activity (4%) and binding (4%) (Figure 4A); the most abundant functions related to Biological Process were: signal transduction (3%), transmembrane transport (3%), regulation of transcription, DNA-dependent (2%) and transport (2%) (Figure 4B); the last GO category, Cellular Component, had most abundant components related to: cytoplasm (12%), membrane (10%), nucleus (7%), intracellular (7%) and plasma membrane (6%) (Figure 4C). The most abundant protein coding genes detected in the *L. amazonensis* genome were ABC transporter, Kinesin, ATP-dependent RNA helicase, HSPs, Protein kinase, dynein heavy chain, Calpains, Amastin surface glycoprotein (Table 2). PFAM and Conserved Domains (CDD) analyses were performed in order to identify the families/domains in 8802 putative proteins. Of these, 3075 proteins were classified to the family level using PFAM, representing a total of 1004 different families in the *L. amazonensis* genome. The most abundant family assigned by PFAM was kinase that contains 69 entries for Pkinase_Tyr and 64 for Pkinase, approximately 2% from total families detected (Figure 2 and Figure 8). The families TPR, zf-C3Hc4_2, DnaJ, RRM_1, AAA22, Helicase C, URR1, URR6 and AAA25 range in size from 58 to 36 proteins and 617 families were represented by a single protein (Figure 2). Analysis with RpsBlast, used to find CDDs, characterized 6144 domains (Figure 3), approximately 1800 of which were found in single copy. Most frequent domains found in *L. amazonensis* proteins were: SMC_prok_B (chromosome segregation protein SMC) with 131 hits, PHA03247 (large tegument protein UL36) with 126 copies and PRK07003 (DNA polymerase III subunits gamma and tau) with 113 copies. These results confirm the great and unknown diversity of *Leishmania* sp functionality. When PFAM and CDD results (Figure 8) were combined, 2483 proteins were simultaneously assigned to some PFAM family and CDD, with the most frequent family found which has some CDD associated was Pkinase_tyr. Nevertheless, approximately 5500 *L. amazonensis* proteins do not have any functional annotation or were not assigned to any family. The functional analysis of *L. amazonensis* according to KOG and COG categories confirmed the specificity of its proteins, since R category (general function prediction only) was the most abundant category found (Figure 5).

Comparative *Leishmania* analysis

A comparative analysis to identify orthologous proteins between the six different *Leishmania* genomes (*L. mexicana*, *L. infantum*, *L. donovani*, *L. braziliensis* and *L. major*) was performed using the genome-scale algorithm, OrthoMCL. Most *L. amazonensis* proteins have orthologous with the five genomes evaluated. We found 7,016 (79.7%) orthologous groups among *L. amazonensis*, *L. donovani*, *L. mexicana*, *L. infantum*, *L. braziliensis* and *L. major* (Figure 7) (Suplementar Material, table S1). Within LCP, approximately 4,800 (68.4%) orthologs were annotated as hypothetical proteins, however among those who have a defined function, we found proteins like: Amastin, calpain-like cysteine peptidase, 40S ribosomal protein S16, RNA helicase, protein kinase, dynein heavy chain, activated protein kinase c receptor (LACK), ABC transporter, Tuzin and DNA primase large subunit. Considering genes shared between two *Leishmania* species, we found 18 orthologous protein groups between *L. amazonensis* and *L. mexicana*, which are closely related and belong to the *L. mexicana* complex (Table 4). Within those 18 orthologous groups, six proteins had an identified function (kinetoplast-associated protein, 3-hydroxyisobutyryl-coenzyme a hydrolase-like protein, viscerotropic *Leishmania*sis antigen, ribosomal protein L1a, amastin, viscerotropic *Leishmania*sis antigen and flagellar calcium-binding protein) and 12 were classified with hypothetical function. When the two most distant species inside *Leishmania* genus were compared, i.e., *L.*(L.) *amazonesis* and *L.* (V.) *braziliensis*, we found 9 proteins exclusive and shared by both of them, among which 4 had known function: heat shock 70-related protein 1, beta tubulin, tyrosine/dopa decarboxylase and oxidoreductase (Table 4). When inparalougous proteins were evaluated in *L. amazonensis*, one paralog was found: triacylglycerol lipase-like protein (Table 4).

To confirm the close relationships between *Leishmania* species, mainly involving *Leishmania* and Vianna subgenus, a phylogenomic analysis was performed based in 31 UO genes. Figure 6 shows the relationship between the species from *L. mexicana* (*L. amazonensis* and *L. mexicana*) and *L. donovani* (*L. donovani* and *L. infantum*) complexes. Even though these orthologous genes are very close the differences between groups could be observed in the generated dendogram (Figure 6). This result is supported by an alignment of one of the UOs (DNA-directed RNA polymerase; Figure 9), where *L. amazonensis* and *L. mexicana* have very similar sequences, whilst *L. braziliensis* has the most divergent sequence, indeed presenting a gap in the multiple alignment. Furthermore, *L. braziliensis* which belongs to Vianna subgenus is positioned in another clade highlighting the difference between them and reflecting the divergence observed in alignment (Figure 9).

Non-homologous isofunctional enzymes

After the initial clustering of 4215 ECs available in KEGG with AnEnPi, we detected 412 ECs with more than one cluster. This last group of 412 ECs was parsed for *L. amazonensis* sequences allocated in different clusters with the same enzymatic activity. With this procedure we could identify 25 potential cases of NISEs when we compared *L. amazonensis* with *Homo sapiens* (termed “intergenomic NISE”). Also, we identified 31 potential cases of NISEs when we searched on *L. amazonensis* protein sequences (termed “intragenomic NISE”). The presence of NISEs was detected in five of the six main EC classes, such as: Oxidoreductases (EC 1), Transferases (EC 2), Hydrolases (EC 3), Lyases (EC 4) and Isomerases (EC 5), but no cases of functional analogy on Ligases (EC 6) were found (Supplementary material - table S2 and S4).

Complementary analyses based on the SUPERFAMILY database for such potential NISEs excluded a few cases, where the status of “Predicted NISE” was given to cases with no significant hits on the SUPERFAMILY database, since although the sequences were allocated in different clusters, we could not confirm the structural differences (Supplementary material - table S2 and S3). Among 25 potential intergenomic NISEs, we could confirm 14 cases and 1 case was considered predicted NISE. Among 31 potential intragenomic NISEs we were able to confirm 15 cases and 1 was considered predicted NISE (table 5 and table 6). It is important to emphasize that the approach considered here is very restrictive since we took into account only NISEs that had totally different folds, not sharing any type of fold, under the same EC.

We selected for further structural characterization three confirmed intergenomic NISE cases where there was clear homology (above 30% sequence identity) to a protein with know 3D structure deposited in PDB and that at the same time there was a solved structure for its human analogous counterpart: LAJMNGS050H11.b.7960 (EC1.1.1.2 - "putative NADP-dependent alcohol dehydrogenase"), LAJMNGS010C07.b.1806 (EC 1.3.1.34 - " putative 2,4-dienoyl-coa reductase FADH1") and LAJMNGS034G09.b.5743 (EC 5.3.3.2 - "putative isopentenyl-diphosphate delta-isomerase"). Comparison of the 3D models for *L. amazonensis* proteins with the experimental structures of the respective human isofunctional enzymes confirmed the distinct folds adopted by the proteins and allowed for detailed characterization of the differences in catalytic sites employed by each analogue (Figure 11).

Finally, we performed a search in drug targets databases like TDR targets, TTD and

DrugBank, for the intergenomic NISE detected, verifying that some of the NISE identified in this work are already under study as potential drug targets against other pathogens. The complete list containing such targets and the pathogens is on Supplementary material (table S5).

RNA interference pathway in *L. amazonensis*

We found some RNAi pathway related genes present in L. amazonensis (Table 7). Dicer seem to be missing from trypanosomatids that lack a functional RNAi pathway like and we were unable to detect dicer in L. amazonensis genome, nor any sequence bearing the characteristic Rnc (dsRNA-specific ribonuclease) domain. Nine DEAD/H box RNA helicase and two Ribonuclease III were identified in L. amazonensis with putative relationship to RNAi pathway (Table 7). Although Dicer was not identified, some Dicer related genes were characterized. We were able to identify four ERI sequences in L. amazonensis genome data set (LAJMNGS009D01.b.1653, LAJMNGS023D01.b.3956, LAJMNGS034E11.b.5717, LAJMNGS035F02.b.5853) (Table 7). Two genes of the RISC (RNA-induced silencing complex, a major effector complex of the RNAi pathway) were also identified: tudor and piwi (argonaute family) (Table 7). The *L. amazonensis* argonaute-like gene identified (LlPWI1) is phylogenetically related to TbPWI1, which is not involved in RNAi. The full sequence of the LlPWI1 gene in *L. amazonensis* and its orthologs were subject to phylogenetic analysis (Figure 10). The neighbour-joining tree clearly distinguishes the two functionally different forms of argonaute family proteins based on *T. brucei* TbAGO1 and TbPWI1. Only *Leishmania* from subgenus Viannia (*L. braziliensis* and *L. guyanensis*) are related with TbAGO1, while the *Leishmania*s from subgenus *Leishmania* (*L. mexicana*, *L. major*, *L. donovani*, *L. infantum* and *L. amazonensis*) fall into TbPWI1 group. Besides *T. brucei*, only *L. braziliensis* possesses the two forms of argonaute family genes: ACI22628, related to TbAGO1 and XP_001564757, which is related to TbPWI1.

Discussion

Our results are similar to the ones obtained by Real and colleagues (2013)[26]. While our assembly resulted in 29,670,588 bases for the genome of *L. amazonensis*, 8802 putative CDS, with GC content of 59% for contigs and 61,12% for the CDS, Real and colleagues (2013)[26] found a genome size of 29.6 Mb with GC content of 58.5% for genome and 61% for CDS, and 8168 putatives genes.

Leishmania amazonensis contains multiple copies of different genes that encode proteins like ABC transporter and calpain like-cysteine (Table 2). Fifty copies of ABC transporter were found annotated in *L. amazonensis,* this large number of copies is expected because the superfamily of ATP-binding cassette (ABC) transporters is one of the largest families of proteins found in eukaryotes [63,64]. In *L. major* and *L. infantum* 42 ABC transporter genes were described, *T. cruzi* and *T. brucei* has 28 and 22 copies, respectively [64,65]. Further than the 50 ABC copies annotated by us in *L. amazonensis*, we observed 33 copies in *L. mexicana*, possibly some ABC transporter genes in *L. amazonensis* can be incomplete, overestimating this number, because instead having a complete copy of the gene, it was fragmented into two “genes” (contigs). Nevertheless these genes are very important because they are involved in drug resistance, infectivity and are related to treatment failure [64–66]. Among Calpains (another important gene related to cytoskeleton) 44 copies were found in *L. amazonensis*. Mottram and colleagues (2004)[67] found in *L. major* 27 calpain and Ersfeld and colleagues (2005)[68] found 24 copies in *T. cruzi* and 18 copies in *T. brucei.* Calpains are involved in the remodelling of cytoskeletal or membrane attachments and have mainly been found in invertebrates and lower eukaryotes. Therefore considering the importance of cytoskeleton remodelling during *Leishmania* differentiation can explain the many Calpain genes-copies in these parasites [67,68]. Calpain is essential for the parasite and it is a potential for drug target. It was demonstrated that MDL 28170, a calpain inhibitor, it shown a high antileishmanial activity against *L. amazonensis* [69]. The knowledge of these Calpain sequences can help for future studies on drug desing.

Other interesting genes found were Tuzins and Amastins. Eight Tuzins copies were found in the *L. amazonensis* genome with a moderate diversity. For comparison *L. mexicana* and *L. tarentolae* present four copies [31], *L. infantum 6* and *L. major* has the highest diversity with 28 copies [22]. Among the Tuzin copies *in L. amazonensis*, one form a ortholog group only with *L. mexicana* (Table 4), remembering that *L. amazonensis* and *L. mexicana* belongs to the same taxonomic complex [17,70], although its function is unknown, however we know that Tuzins are associated with Amastins. Although, the Amastin function is not known too, however could be an abundant surface antigen with stage-specific expression, a crucial transporter to life inside the vertebrate cell or a signal transducer allowing the parasite realize or manipulate the host environment [71,72]. Furthermore Amastins belong to a large family of surface proteins, unique to Kinetoplastids, which are expressed specifically in the amastigote stage of the parasite [71]. Among the 14 Amastins copies found in *L. amazonensis*, 10 of these copies are found in 8 orthologous genes shared with the other five analyzed species (*L. braziliensis*, *L. infantum*, *L. major*, *L. mexicana* and *L. donovanni*), one copy is shared with only between *Leishmania* subgenus. In addition *L. mexicana* shares exclusively with *L. amazonensis* two Amastins genes, remembering that *L. amazonensis* has a total of 28 genes of this family. These results are somewhat expected because Amastin family has 4 sub-families, among which we found some more conserved and more divergent, explaining the fact that we found some copies shared between LCP, or only within the Mexicana complex, that there must be one more specific sub-family [71,72]. Another interesting point is that there are studies which have demonstrated that amastin satisfies some antigenic criteria and are thus used for epitopic analysis [73], and may also be used as relevant biomarker for the VL serodiagnosis [74]. However despite belonging to the same complex [17] *L. amazonensis* and *L. mexicana* show differentiated epidemiologies, and it is interesting to note the presence of a amastin which can be used as a marker for the VL shared only by these two species, but it is known that on some occasions *L. amazonensis* can cause VL [75](Aleixo et al, 2006), as *L. mexicana* that can visceralize [76]. Rogers and collaborators (2011)[25] found a unique gene in *L. mexicana* which encode a predicted protein of unknown function that contains a predicted kelch actin binding domain (PFAM: PF01344). In our work we found a hypothetical protein shared only by *L. amazonensis* and *L. mexicana*, which also contains a predicted kelch actin binding domain, with the same PFAM mapped: PF01344, reinforcing the proximity between these two species, because these proteins were not found in the remaining four *Leishmania* species used in this study.

L. braziliensis presented the highest number of paralogous genes (15), and these results are similar to those obtained by Peacock and colleagues (2007)[22] and Rogers colleagues (2011)[25] when specific paralogous genes in *L. braziliensis* were analyzed. Genes related to telomerase activity and transposon were found such as: TATE DNA Transposon, SLACS like gene retrotransposon element, which as far as we know are unique, even when compared with the other five species of *Leishmania* examined, including the recently sequenced *L . amazonensis* presented in our study. Another notable difference is that *L. braziliensis* contains a functional putative RNA interference pathway, absent in the *L. major*, *L. tarentolae* [31] and *L. amazonensis* (figure 10). We also found some highly divergent copies of surface protein in *L. braziliensis*, not shared with others analyzed *Leishmania*, such as GP63, Amastin and surface antigen-like protein, corroborating previous studies [22,25]. It is known that GP63 protein is involved in *Leishmania* virulence, whose function is cell binding in host conferring parasite protection from complement-mediated lysis [22], for that reason it is considered to be a major virulence factor [77]. Interestingly there are studies showing that GP63 is under positive selection [78,79], and this incentive for changes may contribute to the functional variations of GP63 protease. It has been also described that GP63 is encoded by repeated gene cluster which seems to be enlarged fourfold in *L. braziliensis* compared with the Old World Leishmania [22,78]. *L. braziliensis* has 39 genes encoding GP63 while in *L. amazonensis* and *L. mexicana* we found only seven genes. It is interesting to note that even adding the previously published proteome of *L. donovani* [80] and the newly generated *L. amazonensis*, unique genes in *L. braziliensis* remained, although *L. amazonensis* and *L. braziliensis* have a geographical distribution similar. Interestingly, only the distribution is similar in these two species, because they have different vectors, different clinical manifestations, and they belong to different sub-genues [81]. This corroborates the similarity results between studies, besides being the most divergent species in these study [80,82]. Another important feature found in *L. braziliensis* is the presence of the largest number of paralogs, that is unique on the basis of sequence similarity, the New World *L. braziliensis* is clearly an outlier, consistent with its subgenus classification [22].

We found one highly divergent inparalog gene in *Leishmania infantum*, a Amastin. Rogers and colleagues[25] found 19 highly divergent inparalogs. Our study corroborated the presence of a Amastin, which has a unique highly divergent subfamily of the genus *Leishmania*, besides some sites were found positive selective under in Amastin [72], which may explain the presence of this single paralog in *L. infantum*. Comparatively we found a smaller number of *L. infantum* paralogs than in previous studies [22,72] because this two studies compared *L. infantum* with *L. major*, however the average amino acid identity within *L. infantum*/*L.major* is 92% [22], and Downing et al [80] show that *L. infantum* and *L. donovani* species are much closer than *L. infantum*/*L.major*, belonging to the same complex (*L. infantum*/*L. donovani* or Donovani complex) [83]. Another study shows the same picture demonstrating that these two species (*L. infantum*/*L. donovani*) are phylogenetically very close, with *L. braziliensis* being more divergent inside the genus, although the analysis has been carried out with only one gene, HSP70 [84]. The genomes of those two species has not been analyzed comparatively aiming to identify genus or species-specific (unique) genes, that is probably the reason previous studies found about 26 unique genes in *L. infantum* [22]. However in our study by analyzing these two closely related species together, 25 orthologs shared by them (*L. infantum*/*L. donovani*) were found, and only one highly divergent paralog, (Amastin-like), was found in *L. infantum*. The fact that *L. infantum* belongs to the same *L. donovani* complex, and together share 7619 orthologous groups (93.5% of the *L. infantum* proteome), while with *L. braziliensis* it shares 7401 (90.8% of its proteome) orthologs, could explain this scenario. The same thing happens for *L. amazonensis*, only one inparalog found is probably, because it belongs to the same *L. mexicana* complex and is very close to this species. For example, *L. amazonensis* and *L. mexicana* share several orthologous, 7380 in total (85.8%), whereas *L. amazonensis* and *L. braziliensis,* shares 7162 groups (83.3% of *L. amazonensis* proteome). As expected, we noted that the closer the species, the greater the number of genes shared by them. This is especially true when comparing species within the same complex that show the higher number of shared genes. However when the species among the Viannia subgenus are compared, the number of shared genes is smaller, which may explain the higher number of unique paralogs found in *L. braziliensis* and the small number found in other species. Some large gene families present in *L. amazonensis* may have in common only one conserved domain, then the remaining of their sequences are so divergent that sub-families or classes are identified. As an example we can mention Amastin, which is found in all the six *Leishmania* species analyzed, that present the signatures C-[IVLYF]-[TS]-[LF]-[WF]-G-X-[KRQ]-X-[DENT]-C, however some Amastin are so divergent that can be classified in four sub-families or classes: α, β, γ and δ [72].

Phylogenomics using universal genes

Mauricio et al (1999)[83] using the gene mspC3 as a marker, in species of *Leishmania* subgenus recapitulated a phylogeny very close to the found in this study, keeping *L. infantum* and *L. donovani* in the same branch, with *L. major* and *L. mexicana* more distant, which makes sense because, according to him the species (*L. infantum* and *L. donovani)* belong to the same complex. Although in this study of Mauricio and colleagues (1999)[83] was not possible to observe the separation of subgenus *Leishmania* in New and Old world species. On the other hand, Simpson and colleagues (2004)[84], using HSP70 genes, achieved a reasonable and consistently separation of the subgenus Viannia and *Leishmania*, furthermore inside *Leishmania* *s*ub-genus it was useful to separate the Old and New world *Leishmania* species, but sometimes the support branch value was low. Mauricio and colleagues (2007)[78] using the GP63 gene, which is under positive selection [78,79], achieved a good separation between the subgenus, as well as be able to identify those species originating from the New or Old world, although depending on which copy is used the classification of the genus can get confusing. *L. donovani* complex formation (*L. donovani* and *L. infantum*) was constant in the majority of studies, as well as the formation of *L. mexicana* complex, besides the correct separation of subgenus, and within the *Leishmania* subgenus the separation between Old and New World species [80,83–87]. Fraga and colleagues (2013)[87], using the HSP20 gene separated the Old /New world *Leishmania* subgenus with a bootstrap of 89. However, when they used the HSP20 and HSP70 genes concatenated the bootstrap support value of this separation improved up to 99, and 100 to support the separation of the sub-genus. This example demonstrates the advantage of concatenating genes to infer phylogenomic-based species trees. Our approach of species tree by using 31 UOGs and phylogenomic-based approach was robust showing a minimum bootstrap support of 98. This analysis showed the expected and known separation of this genus for all six species analyzed. *L.* (V.) *braziliensis* as outgroup, *L.* (L.) *infantum* and *L.* (L.) *donovani* very close, reflecting the complex formed by them. As expected, this complex is closer to *L.* (L.) *major* recapitulating phylogeny of the Old Word species inside *Leishmania* subgenus. Nevertheless, *L. mexicana* and *L. amazonensis* are placed together on the same clade, reflecting *L. mexicana* complex, corroborating classical phylogeny showing that those two species represents New World *Leishmania* subgenus species. It should be noted that the bootstrap values were higher than that observed in other works (bootstrap value 100), although the taxonomic position of these species remained mostly the same [32,78,80,83,84,87]. Although the *L. amazonensis* taxonomic position is already known, the phylogenomic species tree obtained using 31 UOG, proved to be a good approach for robust species tree inference using multiple genes, and also a good option to avoid the bias of extrapolating single-gene phylogenies.

Intergenomic and intragenomic NISE as possible drug targets

In this work we have identified a set of Non-homologous Isofunctional Enzymes (NISE – also known as analogous enzymes) between *Leishmania amazonensis* and *Homo sapiens* (table 5), and inside the proteome of *L. amazonensis* (table 6). Such enzymes display the same functional activity, but are unrelated from an evolutionary point of view, since no significant similarity is found either between their primary or between their tertiary structures, which indicates their different ancestries. These structural differences found between NISE could be exploited for the design of drugs that would be active against the parasite´s enzyme, but with no effect to the host´s enzyme. NISE may therefore represent a rather unexploited gene reservoir for the identification of potential drug targets. In fact, some drug targets in study, such as trypanothione-disulfide reductase, present analogy between the enzyme of *Leishmania donovani* and the enzyme of *Homo sapiens* [88,89].

Among the list of intergenomic NISE identified in this work (table 5 and S2), a few interesting cases appeared after an initial literature survey, such as phosphomevalonate kinase (EC: 2.7.4.2), exodeoxyribonuclease III (EC: 3.1.11.2) and isopentenyl-diphosphate Delta-isomerase (EC: 5.3.3.2). Exodeoxyribonuclease III participates in DNA repair [90], a very important activity for the survival of the organism. In fact, Exodeoxyribonuclease III has already been proposed as a drug target candidate against Tri-tryps [52] and cancer [91]. The two other enzymes are involved in the isoprenoid biosynthetic pathway, a chemically diverse pathway responsible for the production of a very large number of natural metabolites such as sterols, carotenoids, dolichols, ubiquinones and some important classes of prenylated proteins: phosphomevalonate kinase, is involved in the biosynthesis of isopentenyl diphosphate (IPP), the building block of all isoprenoids, while IPP isomerase is a key enzyme which catalyzes an essential activation step in isoprenoid biosynthesis by isomerization of the carbon-carbon double bond of IPP to create its electrophilic allylic isomer dimethylallyl diphosphate (DMAPP). Inhibition of this pathway offers potential for the development of antibiotics against bacteria [92] and *P. falciparum* [93]. Relevant information about other enzymatic activities, particularly when considering trypanosomatids, is scarce. An example is 2-alkenal reductase (EC: 1.3.1.74), a defensive role has been shown for this enzyme in some plants, apparently by protecting them from oxidative stress by catalyzing the reduction of reactive carbonyls [94,95], but no information about its biological role has been found for trypanosomatids.

On the other hand, the identification of NISEs inside *L. amazonensis'* proteome (intragenomic NISE) could provide new insights about alternative biochemical pathways and the meaning of functional redundancy inside a genome. Among the NISEs inside *L. amazonensis'* proteome (Table 6 and S2), Carbonate dehydratase (EC 4.2.1.1), DNA-(apurinic or apyrimidinic site) lyase (EC 4.2.99.18) and phosphoglycerate mutase (EC 5.4.2.1) could proposed as potential drug targets. Carbonate dehydratase catalyzes the interconversion of CO2 and HCO3-, this enzymatic function is present in animals, plants, yeast, archaea, bacteria and parasites [96]. Studies have proposed this enzyme as a candidate drug target in *P. falciparum;* inhibition of this enzyme affects the pathway of pyrimidine biosynthesis [96,97]. DNA-(apurinic or apyrimidinic site) lyase is involved in the repair of abasic sites caused by oxidative stress, external agents (chemical or physical), with spontaneous hydrolysis resulting in purine or pyrimidine loss [98]. Phosphoglycerate mutase catalyzes the interconversion of 2-phosphoglycerate (2PG) and 3-phosphoglycerate (3PG) in the glycolytic and gluconeogenic pathways. Phosphoglycerate mutase (PGAM) was structurally characterized in *L. mexicana,* and been proposed as a possible drug target, since the enzymatic form in the parasite is structurally different of the host and has different properties [99], an earlier example of analogy found by an experimental approach.

An integrative approach will be employed in the future to obtain a more complete understanding of the biological role of the intergenomic and intragenomic NISE detected in this work.

RNAi machinery

One of the first organisms where functional RNAi pathway was described was the Trypanosomatidae *T. brucei* [100], since then several trypanosomatids were subject to

RNAi characterization through direct analysis or genome sequencing [22,101,102]. RNA silencing pathways play critical roles in gene regulation, virus infection, and transposon control. RNA interference (RNAi) is mediated by small interfering RNAs (siRNAs), which are liberated from double-stranded (ds)RNA precursors by Dicer and guide the RNA-induced silencing complex (RISC) to degenerates sequence-specific mRNA targets. Phylogenetic analysis suggests the presence of the RNAi pathway in the last common ancestor of eukaryotes with putative important role in defense responses against genomic parasites such as transposable elements and viruses [103].

The RNAi pathway related genes present in different trypanosomatid protozoans [102] were used to identify orthologous genes in *L. amazonensis* genome (Table 7). A key step of RNAi pathway is Dicer activity, which converts double-stranded RNA (dsRNA) into small interfering RNA (siRNA). Dicer has been identified in *T. brucei* (Tb927.8.2370) and a protein with a similar architecture domain, bearing the two RNAse III–like domains, was characterized in *L. braziliensis* (LbrM23_V2.0390). Such proteins seem to be missing from trypanosomatids that lack a functional RNAi pathway like *T. cruzi* and *L. major* [104]. Genomic analysis of *L. infantum*, *L. braziliensis* and *L. major* have demonstrated the presence of dicer only in *L. braziliensis* and otherwise shows synteny for the others *Leishmania*s [22]. We were unable to detect dicer in *L. amazonensis* genome, nor any sequence bearing the characteristic Rnc (dsRNA-specific ribonuclease) domain of *L. braziliensis* putative dicer gene [22]. Since Dicer activity might be performed by a combination of different proteins bearing typical RNAi domains like DEAD-box RNA helicase and Ribonuclease III [22], such domains were subject of analysis in *L. amazonensis* genome data set. Nine DEAD/H box RNA helicase and two Ribonuclease III were identified in *L. amazonensis* with putative relationship to RNAi pathway (Table 7). Although Dicer was not identified, some Dicer related genes were characterized. ERI proteins are another important components of RNAi pathway involved in the formation of the ERI/DICER complex [105]. We were able to identify four ERI sequences in *L. amazonensis* genome data set (LAJMNGS009D01.b.1653, LAJMNGS023D01.b.3956, LAJMNGS034E11.b.5717, LAJMNGS035F02.b.5853) (Table 7). Two genes of the RISC (RNA-induced silencing complex, a major effector complex of the RNAi pathway) were also identified: tudor and piwi (argonaute family) (Table 7). Several argonaute family genes have been described in trypanosomatids.

In *T. brucei*, two argonaute-like genes were identified (TbAGO1 and TbPWI1). Both forms are expressed in the procyclic culture stages but only TbAGO1 is involved in RNAi [106]. Previous data have demonstrated the presence of RNAi key genes argonaute and/or dicer in *Leishmania* subgenus Viannia (*L. braziliensis*, *L. guyanensis* and *L. panamensis*) but not in the subgenus *Leishmania* (*L. mexicana*, *L. major* and *L. donovani*) [102]

Here we describe the first evidence through genomic analysis of RNAi pathway absence of a new-world cutaneous *Leishmania* subgenus *Leishmania* (*Leishmania*) *amazonensis*. So far, experimental evidences pointed out the absence of a functional RNAi pathway in whole subgenus *Leishmania* [22,102], corroborated by the analysis of *L. amazonensis* genome data set. Several arguments have been elegantly raised by Lye et al., (2010)[102] in attempt to understand this phenomena, they describes the viral infections, genome plasticity and phenotype selection as the major players of RNAi lost event. The identified sequences related to RNAi pathway in *L. amazonensis*, might reflect the remains of an erstwhile ancient functional RNAi pathway. The remaining functional genes might me present today because an association with different pathways required for parasite survival. It might be the case of ERI sequences where its dual role in rRNA processing and RNAi [107] might have prevented its loss.

Comparative genome analysis shows that most likely, the last common eukaryote, possess two copies of argonaute related genes suggesting the presence of two distinct silencing machinery. The argonaute-like proteins might had been involved in transcriptional regulation by targeting RNAm in cytoplasm, while piwi-like proteins would act in nucleus targeting transposons [108]. Contrary to most of eukaryotes, in which the argonaute duplication followed by function diversification is common [108], trypanosomatids have, so far, no more than one copy of each argonaute like genes (ago and piwi) per species. Indeed, trypanosomatids with functional RNAi (*T. brucei*, *L. braziliensis*, *L. guyanensis*) has both genes however species with non functional RNAi pathway (*T. cruzi*, *L. amazonensis*, *L. major*, *L. mexicana*) possess only the piwi version of the argonaute family [22,102]. The main difference in the protein domain architecture between the two argonaute family proteins is the absent of a PAZ domain in piwi-like proteins [108]. PAZ domain consist in two sub-domains, with a oligonucleotide/oligosaccharide binding region which is responsible for 3’ ends ssRNA recognition typically found in 3’ overhangs of the small interfering RNAs (siRNAs) [108]. In early work on RNAi characterization in trypanosomatids, two argonaute like genes were identified in *T. brucei* termed TbAGO1 and TbPWI1 [106]. After functional analysis the authors shows that TbAGO1, but not TbPWI1, is involved in RNAi. *L. amazonensis* do not have the ago-like gene, and the piwi-gene (LaPWI1) is homologs to TbPWI1 with orthologs group in subgenus *Leishmania*. Recently, Padmanabhan et al. (2012)[109] identified putative functions for PIWI-like in *L. infantum* and *L. major*. Like *T. brucei*, *Leishmania* piwi-like protein is not related to RNAi pathway neither to siRNA biogenesis. Piwi-like gene are expressed in both parasite forms, but piwi mutation affect the amastigote infection delaying the pathology and increasing apoptosis susceptibility. The authors raised the hypothesis about piwi-like protein role: located in the parasite single mitochondrion, it might act as an apoptotic sensor [109].

The absence of post-transcriptional control of the RNAi might help to explain also the differences observed among the *Leishmania* and Viannia subgenus related with pathogenicity in mammalian host, insect vector relationship and distinct surface glycocalyx structure [110,111].

Conclusions

The *L. amazonensis* genome assembly resulted in approximately 29 million base pairs. The smallest contig had 96 bases and largest 141,211 bases. The annotation resulted in 8802 codifying sequences (CDS), where the biggest coding regions had 19872 bases and the smallest only 66 bases with median and mean of 1637 bp and 1209 bp, respectively. Of these *L. amazonensis* CDS, 63.1% (5554/8802) were annotated as “Hypothetic protein” and 79.71% (7016/8802) were grouped into *Leishmania* Core Proteome. Our work is the first to propose a *Leishmania* Core Proteome using the six sequenced *Leishmania*, while previous studies performed similar comparative analysis using up to five *Leishmania* species. Within *Leishmania* Core Proteome, in general, we found housekeeping proteins as: 40S ribosomal protein S16, RNA helicase, protein kinase, dynein heavy chain, activated protein kinase c receptor (LACK), ABC transporter, calpain-like cysteine peptidase and DNA primase. These LCP genes, could be potentially explored as molecular markers either for diagnosis or genotyping *Leishmania* populations, since all *Leishmania* species studied here have these genes and show minimum differences. Furthermore, some genes related to membrane surface were found: GP63, Amastin and Tuzin. *L. amazonensis* and *L. mexicana* showed the biggest number of specific shared orthologs, 18, most of them without a defined function. However, divergent amastin-like protein and viscerotropic leishmaniasis antigen were found as a ortholog only between these two species and it may be possible to use these are a complex marker. These specific *L. amazonensis/L. mexicana* orthologs are potential specific “Mexicana complex” markers, since are unique to them. The orphans genes found could be explored as markers for species-specific diagnosis since they are present uniquely in this one species. Our original phylogenomic tree using 31 UOGs, confirmed the position of *L. amazonensis* closer to *L. mexicana* and belonging to the New World *Leishmania* subgenus. Probably RNAi pathway in *L. amazonensis* is not functional since key genes are missing in its genome. Finally, we present new information, not described in previous studies, about the NISE search in *L. amazonesis* genome. The NISE search resulted in 25 potential analogous between *L. amazonensis* and *Homo sapiens*. Also, 31 potential analogous were found into *L. amazonensis* protein sequences. Five of the six main EC classes showed potential NISEs: Oxidoreductases (EC 1), Transferases (EC 2), Hydrolases (EC 3), Lyases (EC 4) and Isomerases (EC 5). These NISE findings are new and represents potential drug targets because analogous proteins perform the same function using different proteins and 3D structures. In other words, an analogous protein in *L. amazonensis* can be silenced without affecting the host.

Acknowledgments

We would like to thank Prof. Jeffrey Shaw, Dr. Aldina Barral, Dr. Elisa Cupolillo Prof. Elvira Saraiva, Prof. Bartira Rossi Bergmann, Prof. Salvatore Giovanni de Simone for participation in the choice of the *Leishmania amazonensis* strain for this study.

Author Contributions

Assembled the genome: JR Analysed the data: DAT, GLN, RJ, JL, ASRD, MRG, LMP, DRL, PHS, ANP, FPS. Wrote the first draft of the manuscript: DAT, GLN. Contributed to the writing of the manuscript: MRG, LMP, ANP, HLMG, CMP, ECG, FPS. Jointly developed the structure and arguments for the paper: DAT, HLMG, ABM, FPS, AMRD. Made critical revisions and approved final version: DAT, GLN, RJ, JL, ASRD, MRG, LMP, DRL, PHS, HLMG, ABM, JR, ANP, FPS, CMP, ECG, AMRD. All authors reviewed and approved of the final manuscript

Competing Interests

Authors disclose no potential conflicts of interest

Disclosures and Ethics

As a requirement of publication authors have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Choi J, El-Sayed NM. Functional genomics of trypanosomatids. *Parasite Immunol*. 2012;34(2-3):72–9.

2. Desjeux P. Leishmaniasis. Public health aspects and control. *Clin Dermatol*. 1996;14(5):417–23.

3. WHO. World Health Statistics.http://www.who.int/whosis/whostat/2010/en/. Published 2010. Accessed April 8, 2013. Available at: http://www.who.int/whosis/whostat/2010/en/. Accessed April 8, 2013.

4. Barral A, Pedral-Sampaio D, Grimaldi Júnior G, et al. Leishmaniasis in Bahia, Brazil: evidence that Leishmania amazonensis produces a wide spectrum of clinical disease. *Am J Trop Med Hyg*. 1991;44(5):536–46.

5. Grimaldi G, McMahan-Pratt D. Leishmaniasis and its etiologic agents in the New World: an overview. *Prog Clin Parasitol*. 1991;2:73–118.

6. Weigle K, Saravia NG. Natural history, clinical evolution, and the host-parasite interaction in New World cutaneous Leishmaniasis. *Clin Dermatol*. 1996;14(5):433–50.

7. Chappuis F, Sundar S, Hailu A, et al. Visceral leishmaniasis: what are the needs for diagnosis, treatment and control? *Nat Rev Microbiol*. 2007;5(11):873–82.

8. Akilov OE, Khachemoune A, Hasan T. Clinical manifestations and classification of Old World cutaneous leishmaniasis. *Int J Dermatol*. 2007;46(2):132–42.

9. Desjeux P. Leishmaniasis: current situation and new perspectives. *Comp Immunol Microbiol Infect Dis*. 2004;27(5):305–18.

10. Marsden PD. Mucosal leishmaniasis (“espundia” Escomel, 1911). *Trans R Soc Trop Med Hyg*. 1986;80(6):859–76.

11. Cupolilo SMN, Souza CSF, Abreu-Silva AL, Calabrese KS, Goncalves da Costa SC. Biological behavior of Leishmania (L.) amazonensis isolated from a human diffuse cutaneous leishmaniasis in inbred strains of mice. *Histol Histopathol*. 2003;18(4):1059–65.

12. Mayrink W, Mendonça-Mendes A, de Paula JC, et al. Cluster randomised trial to evaluate the effectiveness of a vaccine against cutaneous leishmaniasis in the Caratinga microregion, south-east Brazil. *Trans R Soc Trop Med Hyg*. 2013;107(4):212–9.

13. Alexander J, Russell DG. The interaction of Leishmania species with macrophages. *Adv Parasitol*. 1992;31:175–254.

14. CDC - Center for Disease Control. Parasites – Leishmaniasis.http://www.cdc.gov/parasites/leishmaniasis/biology.html. Published 2013. Accessed September 24, 2013. Available at: http://www.cdc.gov/parasites/leishmaniasis/biology.html. Accessed September 24, 2013.

15. Passos VM, Fernandes O, Lacerda PA, et al. Leishmania (Viannia) braziliensis is the predominant species infecting patients with American cutaneous leishmaniasis in the State of Minas Gerais, Southeast Brazil. *Acta Trop*. 1999;72(3):251–8.

16. TDR. Leishmaniasis.http://www.who.int/tdr/diseases-topics/Leishmaniasis/en/index.html. Published 2013. Accessed April 11, 2013. Available at: http://www.who.int/tdr/diseases-topics/Leishmaniasis/en/index.html. Accessed April 11, 2013.

17. Stiles JK, Hicock PI, Shah PH, Meade JC. Genomic organization, transcription, splicing and gene regulation in Leishmania. *Ann Trop Med Parasitol*. 1999;93(8):781–807.

18. Wincker P, Ravel C, Blaineau C, et al. The Leishmania genome comprises 36 chromosomes conserved across widely divergent human pathogenic species. *Nucleic Acids Res*. 1996;24(9):1688–94.

19. Ivens AC, Peacock CS, Worthey EA, et al. The genome of the kinetoplastid parasite, Leishmania major. *Science*. 2005;309(5733):436–42.

20. Berriman M, Ghedin E, Hertz-Fowler C, et al. The genome of the African trypanosome Trypanosoma brucei. *Science*. 2005;309(5733):416–22.

21. El-Sayed NM, Myler PJ, Bartholomeu DC, et al. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science*. 2005;309(5733):409–15.

22. Peacock CS, Seeger K, Harris D, et al. Comparative genomic analysis of three Leishmania species that cause diverse human disease. *Nat Genet*. 2007;39(7):839–47.

23. Laurentino EC, Ruiz JC, Fazelinia G, et al. A survey of Leishmania braziliensis genome by shotgun sequencing. *Mol Biochem Parasitol*. 2004;137(1):81–6.

24. Denise H, Poot J, Jiménez M, et al. Studies on the CPA cysteine peptidase in the Leishmania infantum genome strain JPCM5. *BMC Mol Biol*. 2006;7:42.

25. Rogers MB, Hilley JD, Dickens NJ, et al. Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. *Genome Res*. 2011;21(12):2129–42.

26. Real F, Vidal RO, Carazzolle MF, et al. The Genome Sequence of Leishmania (Leishmania) amazonensis: Functional Annotation and Extended Analysis of Gene Models. *DNA Res*. 2013.

27. Depledge DP, Evans KJ, Ivens AC, et al. Comparative expression profiling of Leishmania: modulation in gene expression between species and in different host genetic backgrounds. *PLoS Negl Trop Dis*. 2009;3(7):e476.

28. Adaui V, Castillo D, Zimic M, et al. Comparative gene expression analysis throughout the life cycle of Leishmania braziliensis: diversity of expression profiles among clinical isolates. *PLoS Negl Trop Dis*. 2011;5(5):e1021.

29. El-Sayed NM, Myler PJ, Blandin G, et al. Comparative genomics of trypanosomatid parasitic protozoa. *Science*. 2005;309(5733):404–9.

30. Teixeira SM, de Paiva RMC, Kangussu-Marcolino MM, Darocha WD. Trypanosomatid comparative genomics: Contributions to the study of parasite biology and different parasitic diseases. *Genet Mol Biol*. 2012;35(1):1–17.

31. Raymond F, Boisvert S, Roy G, et al. Genome sequencing of the lizard parasite Leishmania tarentolae reveals loss of genes associated to the intracellular stage of human pathogenic species. *Nucleic Acids Res*. 2012;40(3):1131–47.

32. Lukes J, Mauricio IL, Schönian G, et al. Evolutionary and geographical history of the Leishmania donovani complex with a revision of current taxonomy. *Proc Natl Acad Sci U S A*. 2007;104(22):9375–80.

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

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

35. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 2008;18(5):821–9.

36. Dávila AMR, Lorenzini DM, Mendes PN, et al. GARSA: genomic analysis resources for sequence annotation. *Bioinformatics*. 2005;21(23):4302–3.

37. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25(17):3389–402.

38. Bateman A, Birney E, Cerruti L, et al. The Pfam protein families database. *Nucleic Acids Res*. 2002;30(1):276–80.

39. Punta M, Coggill PC, Eberhardt RY, et al. The Pfam protein families database. *Nucleic Acids Res*. 2012;40(Database issue):D290–301.

40. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res*. 2013;41(12):e121.

41. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25–9.

42. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13(9):2178–89.

43. R Core Team. R: A Language and Environment for Statistical Computing. 2013.

44. Ocaña KACS, Dávila AMR. Phylogenomics-based reconstruction of protozoan species tree. *Evol Bioinform Online*. 2011;7:107–21.

45. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. *Science*. 2006;311(5765):1283–7.

46. Dávila AMR, Mendes PN, Wagner G, et al. ProtozoaDB: dynamic visualization and exploration of protozoan genomes. *Nucleic Acids Res*. 2008;36(Database issue):D547–52.

47. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*. 2005;33(2):511–8.

48. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 2011;28(10):2731–9.

49. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*. 1992;8(3):275–82.

50. Otto TD, Guimarães ACR, Degrave WM, de Miranda AB. AnEnPi: identification and annotation of analogous enzymes. *BMC Bioinformatics*. 2008;9:544.

51. Alves-Ferreira M, Guimarães ACR, Capriles PV da SZ, Dardenne LE, Degrave WM. A new approach for potential drug target discovery through in silico metabolic pathway analysis using Trypanosoma cruzi genome information. *Mem Inst Oswaldo Cruz*. 2009;104(8):1100–10.

52. Gomes MR, Guimarães ACR, de Miranda AB. Specific and nonhomologous isofunctional enzymes of the genetic information processing pathways as potential therapeutical targets for tritryps. *Enzyme Res*. 2011;2011:543912.

53. Barrett AJ. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). *Eur J Biochem*. 1997;250(1):1–6.

54. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*. 2012;40(Database issue):D109–14.

55. Capriles PVSZ, Guimarães ACR, Otto TD, Miranda AB, Dardenne LE, Degrave WM. Structural modelling and comparative analysis of homologous, analogous and specific proteins from Trypanosoma cruzi versus Homo sapiens: putative drug targets for chagas' disease treatment. *BMC Genomics*. 2010;11:610.

56. Andreeva A, Howorth D, Chandonia J-M, et al. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res*. 2008;36(Database issue):D419–25.

57. Wilson D, Pethica R, Zhou Y, et al. SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res*. 2009;37(Database issue):D380–6.

58. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 1993;234(3):779–815.

59. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. 2000;16(6):276–7.

60. Tatusov RL, Koonin E V, Lipman DJ. A genomic perspective on protein families. *Science*. 1997;278(5338):631–7.

61. Tatusov RL, Fedorova ND, Jackson JD, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. 2003;4:41.

62. Logan-Klumpler FJ, De Silva N, Boehme U, et al. GeneDB--an annotation database for pathogens. *Nucleic Acids Res*. 2012;40(Database issue):D98–108.

63. Saurin W, Hofnung M, Dassa E. Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. *J Mol Evol*. 1999;48(1):22–41.

64. Sauvage V, Aubert D, Escotte-Binet S, Villena I. The role of ATP-binding cassette (ABC) proteins in protozoan parasites. *Mol Biochem Parasitol*. 2009;167(2):81–94.

65. Leprohon P, Légaré D, Girard I, Papadopoulou B, Ouellette M. Modulation of Leishmania ABC protein gene expression through life stages and among drug-resistant parasites. *Eukaryot Cell*. 2006;5(10):1713–25.

66. Torres DC, Adaui V, Ribeiro-Alves M, et al. Targeted gene expression profiling in Leishmania braziliensis and Leishmania guyanensis parasites isolated from Brazilian patients with different antimonial treatment outcomes. *Infect Genet Evol*. 2010;10(6):727–33.

67. Mottram JC, Coombs GH, Alexander J. Cysteine peptidases as virulence factors of Leishmania. *Curr Opin Microbiol*. 2004;7(4):375–81.

68. Ersfeld K, Barraclough H, Gull K. Evolutionary relationships and protein domain architecture in an expanded calpain superfamily in kinetoplastid parasites. *J Mol Evol*. 2005;61(6):742–57.

69. d'Avila-Levy CM, Marinho FA, Santos LO, Martins JL, Santos ALS, Branquinha MH. Antileishmanial activity of MDL 28170, a potent calpain inhibitor. *Int J Antimicrob Agents*. 2006;28(2):138–42.

70. Eresh S, McCallum SM, Barker DC. Identification and diagnosis of Leishmania mexicana complex isolates by polymerase chain reaction. *Parasitology*. 1994;109 (Pt 4:423–33.

71. Rochette A, McNicoll F, Girard J, et al. Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in Leishmania spp. *Mol Biochem Parasitol*. 2005;140(2):205–20.

72. Jackson AP. The evolution of amastin surface glycoproteins in trypanosomatid parasites. *Mol Biol Evol*. 2010;27(1):33–45.

73. Grover A, Kumar Shakyawar S, Saudagar P, Kumar Dubey V, Sundar D. Epitopic analysis of Potential Vaccine Candidate in Leishmania infantum for Development of Human Vaccine. *Lett Drug Des Discov*. 2012;9(7):698–705.

74. Rafati S, Hassani N, Taslimi Y, Movassagh H, Rochette A, Papadopoulou B. Amastin peptide-binding antibodies as biomarkers of active human visceral leishmaniasis. *Clin Vaccine Immunol*. 2006;13(10):1104–10.

75. Aleixo JA, Nascimento ET, Monteiro GR, et al. Atypical American visceral leishmaniasis caused by disseminated Leishmania amazonensis infection presenting with hepatitis and adenopathy. *Trans R Soc Trop Med Hyg*. 2006;100(1):79–82.

76. Mestra L, Lopez L, Robledo SM, Muskus CE, Nicholls RS, Vélez ID. Transfusion-transmitted visceral leishmaniasis caused by Leishmania (Leishmania) mexicana in an immunocompromised patient: a case report. *Transfusion*. 2011;51(9):1919–23.

77. Yao C, Donelson JE, Wilson ME. The major surface protease (MSP or GP63) of Leishmania sp. Biosynthesis, regulation of expression, and function. *Mol Biochem Parasitol*. 2003;132(1):1–16.

78. Mauricio IL, Gaunt MW, Stothard JR, Miles MA. Glycoprotein 63 (gp63) genes show gene conversion and reveal the evolution of Old World Leishmania. *Int J Parasitol*. 2007;37(5):565–76.

79. Ma L, Chen K, Meng Q, et al. An evolutionary analysis of trypanosomatid GP63 proteases. *Parasitol Res*. 2011;109(4):1075–84.

80. Downing T, Imamura H, Decuypere S, et al. Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res*. 2011;21(12):2143–56.

81. Sharma U, Singh S. Insect vectors of Leishmania: distribution, physiology and their control. *J Vector Borne Dis*. 2008;45(4):255–72.

82. Croan DG, Morrison D a, Ellis JT. Evolution of the genus Leishmania revealed by comparison of DNA and RNA polymerase gene sequences. *Mol Biochem Parasitol*. 1997;89(2):149–59.

83. Mauricio IL, Howard MK, Stothard JR, Miles MA. Genomic diversity in the Leishmania donovani complex. *Parasitology*. 1999;119 (Pt 3:237–46.

84. Simpson AGB, Gill EE, Callahan HA, Litaker RW, Roger AJ. Early evolution within kinetoplastids (euglenozoa), and the late emergence of trypanosomatids. *Protist*. 2004;155(4):407–22.

85. Widmer G, Comeau AM, Furlong DB, Wirth DF, Patterson JL. Characterization of a RNA virus from the parasite Leishmania. *Proc Natl Acad Sci U S A*. 1989;86(15):5979–82.

86. Lukes J, Mauricio IL, Schönian G, et al. Evolutionary and geographical history of the Leishmania donovani complex with a revision of current taxonomy. *Proc Natl Acad Sci U S A*. 2007;104(22):9375–80.

87. Fraga J, Montalvo AM, Van der Auwera G, Maes I, Dujardin J-C, Requena JM. Evolution and species discrimination according to the Leishmania heat-shock protein 20 gene. *Infect Genet Evol*. 2013;18:229–37.

88. Cunningham ML, Fairlamb AH. Trypanothione reductase from Leishmania donovani. Purification, characterisation and inhibition by trivalent antimonials. *Eur J Biochem*. 1995;230(2):460–8.

89. Ilari A, Baiocco P, Messori L, et al. A gold-containing drug against parasitic polyamine metabolism: the X-ray structure of trypanothione reductase from Leishmania infantum in complex with auranofin reveals a dual mechanism of enzyme inhibition. *Amino Acids*. 2012;42(2-3):803–11.

90. Lindahl T, Demple B, Robins P. Suicide inactivation of the E. coli O6-methylguanine-DNA methyltransferase. *EMBO J*. 1982;1(11):1359–63.

91. Sultana R, McNeill DR, Abbotts R, et al. Synthetic lethal targeting of DNA double-strand break repair deficient cells by human apurinic/apyrimidinic endonuclease inhibitors. *Int J Cancer*. 2012;131(10):2433–44.

92. Doun SS, Burgner JW, Briggs SD, Rodwell VW. Enterococcus faecalis phosphomevalonate kinase. *Protein Sci*. 2005;14(5):1134–9.

93. Wiesner J, Jomaa H. Isoprenoid biosynthesis of the apicoplast as drug target. *Curr Drug Targets*. 2007;8(1):3–13.

94. Yin L, Mano J, Wang S, Tsuji W, Tanaka K. The involvement of lipid peroxide-derived aldehydes in aluminum toxicity of tobacco roots. *Plant Physiol*. 2010;152(3):1406–17.

95. Mano J, Belles-Boix E, Babiychuk E, et al. Protection against photooxidative injury of tobacco leaves by 2-alkenal reductase. Detoxication of lipid peroxide-derived reactive carbonyls. *Plant Physiol*. 2005;139(4):1773–83.

96. Krungkrai SR, Krungkrai J. Malaria parasite carbonic anhydrase: inhibition of aromatic/heterocyclic sulfonamides and its therapeutic potential. *Asian Pac J Trop Biomed*. 2011;1(3):233–42.

97. Reungprapavut S, Krungkrai SR, Krungkrai J. Plasmodium falciparum carbonic anhydrase is a possible target for malaria chemotherapy. *J Enzyme Inhib Med Chem*. 2004;19(3):249–56.

98. Vidal AE, Harkiolaki M, Gallego C, et al. Crystal structure and DNA repair activities of the AP endonuclease from Leishmania major. *J Mol Biol*. 2007;373(4):827–38.

99. Nowicki MW, Kuaprasert B, McNae IW, et al. Crystal structures of Leishmania mexicana phosphoglycerate mutase suggest a one-metal mechanism and a new enzyme subclass. *J Mol Biol*. 2009;394(3):535–43.

100. Ngô H, Tschudi C, Gull K, Ullu E. Double-stranded RNA induces mRNA degradation in Trypanosoma brucei. *Proc Natl Acad Sci U S A*. 1998;95(25):14687–92.

101. Robinson KA, Beverley SM. Improvements in transfection efficiency and tests of RNA interference (RNAi) approaches in the protozoan parasite Leishmania. *Mol Biochem Parasitol*. 2003;128(2):217–28.

102. Lye L-F, Owens K, Shi H, et al. Retention and loss of RNA interference pathways in trypanosomatid protozoans. *PLoS Pathog*. 2010;6(10):e1001161.

103. Cerutti H, Casas-Mollano JA. On the origin and functions of RNA-mediated silencing: from protists to man. *Curr Genet*. 2006;50(2):81–99.

104. Shi H, Tschudi C, Ullu E. An unusual Dicer-like1 protein fuels the RNA interference pathway in Trypanosoma brucei. *RNA*. 2006;12(12):2063–72.

105. Pavelec DM, Lachowiec J, Duchaine TF, Smith HE, Kennedy S. Requirement for the ERI/DICER complex in endogenous RNA interference and sperm development in Caenorhabditis elegans. *Genetics*. 2009;183(4):1283–95.

106. Durand-Dubief M, Bastin P. TbAGO1, an argonaute protein required for RNA interference, is involved in mitosis and chromosome segregation in Trypanosoma brucei. *BMC Biol*. 2003;1:2.

107. Gabel HW, Ruvkun G. The exonuclease ERI-1 has a conserved dual role in 5.8S rRNA processing and RNAi. *Nat Struct Mol Biol*. 2008;15(5):531–3.

108. Hutvagner G, Simard MJ. Argonaute proteins: key players in RNA silencing. *Nat Rev Mol Cell Biol*. 2008;9(1):22–32.

109. Padmanabhan PK, Dumas C, Samant M, Rochette A, Simard MJ, Papadopoulou B. Novel features of a PIWI-like protein homolog in the parasitic protozoan Leishmania. *PLoS One*. 2012;7(12):e52612.

110. Bates PA. Transmission of Leishmania metacyclic promastigotes by phlebotomine sand flies. *Int J Parasitol*. 2007;37(10):1097–106.

111. Bañuls A-L, Hide M, Prugnolle F. Leishmania and the leishmaniases: a parasite genetic update and advances in taxonomy, epidemiology and pathogenicity in humans. *Adv Parasitol*. 2007;64:1–109.

Tables

Table 1: Summary of the *Leishmania amazonensis* assembly and genome

Contigs	10305
Sum of consensus sequences length	29,670,588 bases
Number of scaffolds > 1K nt	4827 (46.8%)
Number of scaffolds > 10K nt	732 (7.1%)
Number of scaffolds > 100K nt	2 (0.02%)
Coding genes: CDS	8802
Chromosome	34
%GC content: Contigs/CDS	59% / (61.125%)
Size: Contigs/CDS	
Max (bases)	141211 / (19872)
Min (bases)	96 / (66)
Mean (bases)	2879 / (1637)
Median (bases)	853 / (1209)
N50 scaffold length	8346
CDS Ontology	
Molecular Function	4065
Biological Process	4007
Cellular Component	4054
Protein families (PFAM)	3075
Conserved Domains (CDD)	6144
Annotated as “Hypothetic protein”	5554
Putative Orphans (orthoMCL)	887

Table 2: Resume table of most abundant and single copy genes/domains found in *Leishmania amazonensis* genome analysis.

Most abundant genes/domains	Single copy genes/domains
ABC transporter	rpS2
Amastin surface glycoprotein	rpS5
ATP-dependent RNA helicase	rpS8
Calpains	rpS10
dynein heavy chain	rpS12
Heat Shock Proteins (HSPs)	rpL7
Kinesin	rpL12
Protein kinase	rpL13
WD40	rpL19
chaperone DNAJ	rpL23

Most abundant genes/domains found in the initial *Leishmania amazonensis* genome analysis. Genes/domains found in single copy during the analysis. 40S ribomosomal proteins (rpS) and 60S ribosomal proteins (rpL)

Table 3: List of orphans proteins found in *Leishmania amazonensis* with their respective identification, description and length (aa).

Identification	**Description**	**Length**
LAJMNGS001H06.b.195	unspecified product	98
LAJMNGS002H09.b.421	unspecified product	150
LAJMNGS005H02.b.1027	hypothetical protein, conserved	79
LAJMNGS006F03.b.1178	hypothetical protein, unknown function	771
LAJMNGS018E09.b.3196	carboxypeptidase, putative	325
LAJMNGS018H07.b.3264	hypothetical protein	951
LAJMNGS027A04.b.4532	hypothetical protein	167
LAJMNGS030G04.b.5103	unspecified product	48
LAJMNGS031F02.b.5255	unspecified product	212
LAJMNGS038C10.b.6191	unspecified product	37
LAJMNGS038E01.b.6205	unspecified product	94
LAJMNGS051A11.b.7995	hypothetical protein	139
NODE_5216_1	hypothetical protein, unknown function	68
NODE_20256_1	unspecified product	81

Table 4. Identification of orthologous groups between *L. amazonensis* and *Leishmania* species and inparalogous from *L. amazonensis*

ORTHOMCL	***L. amazonensis*** accession	**Pfam annotation**	**Cdd annotation**	**Protein Description**	**L. am**	**L. me**	**L. do**	**L. br**	**L. ma**	**L in**
ORTHOMCL7819	LAJMNGS015A07.b.2588 LAJMNGS029B11.b.4872			triacylglycerol lipase	X					
ORTHOMCL7720	NODE_11447_1 gi|401430294 gi|401430407		PTZ00494, tuzin	unspecified product	X	X				
ORTHOMCL7785	NODE_9861_1 gi|401424225			kinetoplast-associated protein	X	X				
ORTHOMCL7787	NODE_2310_4 gi|401430466			unspecified product	X	X				
ORTHOMCL7789	NODE_20602_1 gi|401430272		pfam07344, Amastin	unspecified product	X	X				
ORTHOMCL7794	NODE_11369_4 gi|401427459		pfam13766, ECH_C, 2-enoyl-CoA Hydratase	3-hydroxyisobutyryl-coenzyme a hydrolase	X	X				
ORTHOMCL7799	LAJMNGS049C04.b.7651 gi|401414155			unspecified product	X	X				
ORTHOMCL7802	LAJMNGS046H11.b.7373 gi|401414833			viscerotropic *Leishmania*sis antigen,	X	X				
ORTHOMCL7803	LAJMNGS046G07.b.7351 gi|401418572		cd00051, EFh, EF-hand, calcium binding motif	flagellar calcium-binding protein, putative	X	X				
ORTHOMCL7804	LAJMNGS045F11.b.7202 gi|401417934			hypothetical protein, conserved	X	X				
ORTHOMCL7805	LAJMNGS043C01.b.6864 gi|401414282			unspecified product	X	X				
ORTHOMCL7807	LAJMNGS035F02.b.5854 gi|401416024			hypothetical protein	X	X				
ORTHOMCL7808	LAJMNGS033H06.b.5610 gi|401428307		PTZ00201, amastin surface glycoprotein	amastin-like protein	X	X				
ORTHOMCL7810	LAJMNGS031F10.b.5268 gi|401415800			unspecified product	X	X				
ORTHOMCL7813	LAJMNGS029D07.b.4914 gi|401427209	PF13415.1Kelch_3	pfam01344, Kelch_1	hypothetical protein	X	X				
ORTHOMCL7814	LAJMNGS029D03.b.4903 gi|401430342	PF00806.14PUF	cd07920, Pumilio	unspecified product	X	X				
ORTHOMCL7822	LAJMNGS010A05.b.1767 gi|401415906		PTZ00428, 60S ribosomal protein L4	ribosomal protein L1a, putative	X	X				
ORTHOMCL7826	LAJMNGS006D01.b.1132			unspecified product	X	X				

	gi|401414193									
ORTHOMCL7827	LAJMNGS003F09.b.554 gi|401417308			hypothetical protein, unknown function	X	X				
ORTHOMCL7717	NODE_33600_1 gi|398019921 gi|398019923			amino acid permease	X		X			
ORTHOMCL7786	NODE_27330_4 gi|398020509			hypothetical protein, conserved	X		X			
ORTHOMCL7788	NODE_21871_1 gi|398010889		PTZ00201, amastin	amastin-like protein	X		X			
ORTHOMCL7790	NODE_20189_1 gi|398010239		cd03213, ABCG_EPDR	ATP-binding cassette protein subfamily G, member 1	X		X			
ORTHOMCL7791	NODE_16342_1 gi|398013201			hypothetical protein, conserved	X		X			
ORTHOMCL7792	NODE_12712_1 gi|398023645		PTZ00263, protein kinase A	protein kinase A catalytic subunit isoform 2	X		X			
ORTHOMCL7795	NODE_10493_4 gi|398023914			phosphoglycan beta 1,3 galactosyltransferase 4	X		X			
ORTHOMCL7796	NODE_10388_4 gi|398020023			hypothetical protein, conserved	X		X			
ORTHOMCL7800	LAJMNGS047C07.b.7405 gi|398019480		COG1788, Acyl CoA:acetate/3-ketoacid	succinyl-coa:3-ketoacid-coenzyme a transferase-like protein	X		X			
ORTHOMCL7801	LAJMNGS047B12.b.7397 gi|398015472		PTZ00243, ABC transporter	multidrug resistance protein, putative,p-glycoprotein, putative,ABC transporter	X		X			
ORTHOMCL7806	LAJMNGS038F02.b.6219 gi|398014585			unspecified product	X		X			
ORTHOMCL7809	LAJMNGS033B09.b.5509 gi|398014545			calpain-like cysteine peptidase	X		X			
ORTHOMCL7811	LAJMNGS030F01.b.5087 gi|398010628			vacuolar-type Ca2 -ATPase, putative	X		X			
ORTHOMCL7823	LAJMNGS008G11.b.1562 gi|398015632		COG1621, SacC, Beta-fructosidases	beta-fructosidase, invertase,sucrose hydrolase	X		X			
ORTHOMCL7824	LAJMNGS008E09.b.1520 gi|398013197			hypothetical protein, conserved	X		X			
ORTHOMCL7714	NODE_8040_1 gi|389603588 gi|389603590			hypothetical protein, unknown function	X			X		

ORTHOMCL7719	NODE_11708_1 gi|154341831 gi|154341835		PTZ00186, heat shock 70 kDa	heat shock 70-related protein 1, mitochondrial precursor, putative	X			X		
ORTHOMCL7793	NODE_1256_4 gi|154343852		PLN00220, tubulin beta chain	beta tubulin	X			X		
ORTHOMCL7798	LAJMNGS050C12.b.7841 gi|389602223		PLN02880, tyrosine decarboxylase	tyrosine/dopa decarboxylase	X			X		
ORTHOMCL7812	LAJMNGS029E12.b.4945 gi|154346778			hypothetical protein	X			X		
ORTHOMCL7815	LAJMNGS027C01.b.4566 gi|154334177		pfam00201 , UDP-glucoronosyl and glucosyl transferase	hypothetical protein, conserved	X			X		
ORTHOMCL7818	LAJMNGS015D10.b.2693 gi|154331940		PTZ00261, acyltransferase	unspecified product	X			X		
ORTHOMCL7821	LAJMNGS010C09.b.1810 gi|154339794			hypothetical protein	X			X		
ORTHOMCL7825	LAJMNGS008E08.b.1511 gi|154344054	PF00107.21ADH_zinc_N	cd08250, Mgc45594_like	oxidoreductase-like protein	X			X		
ORTHOMCL7733	LAJMNGS011E12.b.2013			hypothetical protein, conserved	X				X	
	LAJMNGS013H02.b.2370 gi|157867225									
ORTHOMCL7816	LAJMNGS026B05.b.4415 gi|157865501			hypothetical protein, conserved	X				X	
ORTHOMCL7817	LAJMNGS018C09.b.3168 gi|157865347			unspecified product	X				X	
ORTHOMCL7820	LAJMNGS013E02.b.2302 gi|157865279			hypothetical protein	X				X	
ORTHOMCL7797	LAJMNGS050E01.b.7873 gi|339899089			hypothetical protein, conserved	X					X

Table 5: Intergenomic NISEs, their official enzyme names, sequencies Ids, Uniport Ids for human sequencies, PDB structures and the identity for each sequence.

EC	Enzyme Name (Oficial)	Organism	Sequencies_IDs: L. amazonensis	Uniprot Access	PDB [Best hit] (*)	Identity (PDB)
1.1.1.1	Alcohol dehydrogenase	*L. amazonensis*	LAJMNGS020G03.b.3554	N/A	3OWO	133/371 (36%)
		H. sapiens	hsa:127	P08319	3COS	Structure solved
			hsa:128	P11766	1M6H (+)	Structure solved
				Q6IRT1		
			hsa:130	P28332	1EE2	224/366 (61%)
				Q8IUN7		
			hsa:131	P40394	1AGN	Structure solved
1.1.1.2	Alcohol dehydrogenase (NADP(+))	*L. amazonensis*	LAJMNGS050H11.b.7960	N/A	1UUF	160/332 (48%)
		H. sapiens	hsa:10327	P14550	2ALR	Structure solved
1.3.1.34	2,4-dienoyl-CoA reductase (NADPH)	*L. amazonensis*	LAJMNGS010C07.b.1806	N/A	1PS9	294/730 (40%)
			LAJMNGS024B09.b.4107	N/A		198/658 (30%)
		H. sapiens	hsa:1666	Q16698	1W6U	Structure solved
			hsa:26063	Q9NUI1	4FC6	Structure solved
1.3.1.74	2-alkenal reductase	*L. amazonensis*	LAJMNGS036G08.b.6014	N/A	4GBY	139/482 (29%)
		H. sapiens	hsa:22949	Q14914	1ZSV (+)	Structure solved
1.3.3.4	Protoporphyrinogen oxidase	*L. amazonensis*	LAJMNGS030H07.b.5136	N/A	N/A	N/A
		H. sapiens	hsa:5498	P50336	3NKS (+)	Structure solved
2.1.1.63	Methylated-DNA--[protein]-cysteine S-methyltransferase	*L. amazonensis*	NODE_15705_1	N/A	3ZEY	164/268 (61%)
		H. sapiens	hsa:4255	B4DEE8	1EH6	Structure solved
2.7.1.31	Glycerate 3-kinase	*L. amazonensis*	LAJMNGS044H03.b.7072	N/A	N/A	N/A
		H. sapiens	hsa:132158	Q8IVS8	N/A	N/A
2.7.4.2	Phosphomevalonate kinase	*L. amazonensis*	LAJMNGS005E09.b.95	N/A	N/A	N/A
		H. sapiens	hsa:10654	Q15126	3CH4	Structure solved
				Q6FGV9		

3.1.3.1	Alkaline phosphatase	*L. amazonensis*	LAJMNGS050E08.b.7900	N/A	N/A	N/A
		H. sapiens	hsa:248	P09923	1EW2	430/494(87%)
			hsa:249	P05186	1EW2	273/475 (57%)
			hsa:250	P05187	1EW2 (+)	Structure solved
			hsa:251	P10696	1EW2	468/493 (95%)
3.1.11.2 (Predicted NISE)	Exodeoxyribonuclease III	*L. amazonensis*	LAJMNGS001G08.b.166	N/A	N/A	N/A
		H. sapiens	hsa:5810	O60671	3G65 (+)	Structure solved
			hsa:5883	Q99638	3GGR (+)	Structure solved
			hsa:11219	Q9BQ50	1Y97	Structure solved
			hsa:11277	Q9NSU2	3U6F	178/304 (59%)
				Q5TZT0		
3.4.11.5	Prolyl aminopeptidase	*L. amazonensis*	LAJMNGS001H01.b.176	N/A	1Q0R	77/312 (25%)
			LAJMNGS033A05.b.5486	N/A	1Q0R	76/296 (26%)
			LAJMNGS039D03.b.6327	N/A	1Q0R	90/313 (29%)
			LAJMNGS044G01.b.7054	N/A	1Q0R	92/343 (27%)
			LAJMNGS044G01.b.7055	N/A	1Q0R	97/372 (26%)
		H. sapiens	hsa:51056	P28838	1BLL	450/487 (92%)
3.6.1.17	Bis(5'-nucleosyl)-tetraphosphatase (asymmetrical)	*L. amazonensis*	LAJMNGS022B08.b.3759	N/A	3KSV	128/138 (93%)
		H. sapiens	hsa:318	P50583	3U53 (+)	Structure solved
4.2.1.2	Fumarate hydratase	*L. amazonensis*	LAJMNGS009H05.b.1751	N/A	2ISB	69/196 (35%)
		H. sapiens	hsa:2271	P07954	3E04	457/467 (98%)
				B1ANK7		
5.3.1.6	Ribose-5-phosphate isomerase	*L. amazonensis*	LAJMNGS039D01.b.6325	N/A	3K7O	70/156 (45%)
		H. sapiens	hsa:22934	P49247	1XTZ	110/240 (46%)
5.3.3.2	Isopentenyl-diphosphate Delta-isomerase	*L. amazonensis*	LAJMNGS034G09.b.5743	N/A	2ZRU	118/352 (34%)
		H. sapiens	hsa:91734	Q9BXS1	2PNY	Structure solved
			hsa:3422	Q13907	2ICJ	Structure solved

(*) The (+) signal on "PDB [Best hit]" column represent that there are more structures solved for this sequence.

Table 6: Intragenomic NISEs, their official enzyme names, sequencies Ids, PDB structures identified and the identity for each sequence.

EC	Enzyme Name (Oficial)	Original Annotation	Sequencies_IDs: *L. amazonensis* (*)	PDB/Best hit	Identity
1.1.1.95	Phosphoglycerate dehydrogenase	GTP-binding protein-like protein	LAJMNGS002H05.b.408	4DCS	132/467 (28%)
	Phosphoglycerate dehydrogenase	D-3-phosphoglycerate dehydrogenase-like protein	LAJMNGS029A10.b.4851	1UAY	206/398 (52%)
1.3.1.74	2-alkenal reductase	sugar transporter, putative	LAJMNGS036G08.b.6014	4GBY	139/482 (29%)
	2-alkenal reductase	heat shock protein 70-related protein	LAJMNGS013H10.b.2384 (6)	1YUW	201/328 (61%)
1.6.5.5	NADPH:quinone reductase	hypothetical protein, conserved	LAJMNGS006D12.b.1149 (11)	4DUP	126/335 (38%)
	NADPH:quinone reductase	unspecified product	LAJMNGS005A12.b.838 (2)	3OOX	90/309 (29%)
1.6.99.3	NADH dehydrogenase	NADH dehydrogenase, putative	LAJMNGS030C06.b.5029	4G6G	139/468 (30%)
	NADH dehydrogenase	nitroreductase-like protein	LAJMNGS035C11.b.5805	3BEM	57/203 (28%)
	NADH dehydrogenase	acyl carrier protein, putative	LAJMNGS023D05.b.3963	2EHS	34/70 (49%)
2.3.1.48	Histone acetyltransferase	acetyltransferase, putative	LAJMNGS001H02.b.186 (3)	2OU2	79/213 (37%)

	Histone acetyltransferase	histone acethyltransferase-like protein	LAJMNGS013C07.b.2270 (2)	N/A	N/A
2.3.1.51_1	1-acylglycerol-3-phosphate O-acyltransferase	1-acyl-sn-glycerol-3-phosphateacyltransferase- like protein, putative	LAJMNGS011E07.b.2000	N/A	N/A
	1-acylglycerol-3-phosphate O-acyltransferase	hypothetical protein, conserved	LAJMNGS014A12.b.2407	N/A	N/A
	1-acylglycerol-3-phosphate O-acyltransferase	hypothetical protein, conserved	LAJMNGS048A06.b.7502	N/A	N/A
2.1.1.17 (Predicted NISE)	Phosphatidylethanolamine N-methyltransferase	phosphatidylethanolaminen-methyltransferase-like protein	LAJMNGS021H05.b.3716	N/A	N/A
	Phosphatidylethanolamine N-methyltransferase	phosphatidylethanolaminen-methyltransferase-lik e protein	LAJMNGS004D11.b.699	N/A	N/A
3.1.1.3	Triacylglycerol lipase	hypothetical protein, conserved	LAJMNGS005H07.b.1038 (2)	3PIK	27/103(26%)
	Triacylglycerol lipase	lipase, putative	LAJMNGS004H08.b.800	4TGL	79/217 (36%)
3.1.1.47	1-alkyl-2-acetylglycerophosphocholine esterase.	hypothetical protein, conserved	LAJMNGS004E06.b.714	2YMU	97/286 (34%)
	1-alkyl-2-acetylglycerophosphocholine esterase.	phospholipase A2-like protein, putative	LAJMNGS003B12.b.479	3D59	76/270 (28%)
3.6.1.43	Dolichyldiphosphatase	PAP2 family protein, putative	LAJMNGS004C09.b.679	N/A	N/A
	Dolichyldiphosphatase	hypothetical protein, conserved	LAJMNGS014E10.b.2497	1XTP	225/250

					(90%)
3.1.3.16	Phosphoprotein phosphatase	hypothetical protein, conserved	LAJMNGS008F12.b.1543 (9)	2ISN	41/135 (30%)
	Phosphoprotein phosphatase	dual-specificity protein phosphatase, putative	LAJMNGS025D10.b.4296 (3)	2G6Z	53/148 (36%)
	Phosphoprotein phosphatase	hypothetical protein, conserved	LAJMNGS002E02.b.313 (21)	1MF8	76/373 (20%)
3.6.1.23	dUTP diphosphatase	coronin, putative	LAJMNGS027B10.b.4560	2AQ5	142/397 (36%)
	dUTP diphosphatase	deoxyuridine triphosphatase, putative,dUTP diphosphatase	LAJMNGS025E10.b.4319	2CJE	257/268 (96%)
4.2.1.1	Carbonate dehydratase	carbonic anhydrase-like protein	LAJMNGS019E05.b.3366	4G7A	53/164 (32%)
	Carbonate dehydratase	carbonic anhydrase family protein, putative	LAJMNGS035D05.b.5816	1I6O	97/229 (42%)
4.2.99.18	DNA-(apurinic or apyrimidinic site) lyase	endonuclease III, putative	LAJMNGS002A05.b.218 (2)	1P59	66/194 (34%)
	DNA-(apurinic or apyrimidinic site) lyase	endonuclease/exonuclease protein-like protein	LAJMNGS041H02.b.6678 (2)	2ISI	37/106 (35%)
5.2.1.8	Peptidylprolyl isomerase	cyclophilin, putative	LAJMNGS001G02.b.135 (18)	2HQJ	119/180 (66%)
	Peptidylprolyl isomerase	peptidyl-prolyl cis-trans isomerase, putative	LAJMNGS006G08.b.1207 (5)	3H9R	58/110 (53%)

5.4.2.1	Phosphoglycerate mutase	phosphoglycerate mutase protein, putative	LAJMNGS013E01.b.2299	4IJ5	45/152 (30%)
	Phosphoglycerate mutase	2,3-bisphosphoglycerate-independent phosphoglycerate mutase,2,3-bisphosphoglycerate-independentphosphoglyce ra temutase	LAJMNGS025H05.b.4375 (2)	3IGY	497/552 (90%)
(*) The numbers between parenthesis on "Sequencies_IDs: *L. amazonensis*" column, represent the number of copies of this enzyme.					

Table 7 RNAi pathway related sequences in *L. amazonensis*
Piwi-AGO
LAJMNGS037G03.b.6124
Tudor
LAJMNGS051A10.b.7989
ERI-1
LAJMNGS009D01.b.1653
LAJMNGS023D01.b.3956
LAJMNGS034E11.b.5717
LAJMNGS035F02.b.5853
DEAD-Box RNA helicase
LAJMNGS002E10.b.336
LAJMNGS005H09.b.1043
LAJMNGS016A07.b.2785
LAJMNGS018H11.b.3270
LAJMNGS021E12.b.3675
LAJMNGS024C05.b.4124
LAJMNGS042E06.b.6755
LAJMNGS045F10.b.7200
LAJMNGS046A05.b.7244
Ribonuclease III/Dicer
LAJMNGS020H09.b.3587
LAJMNGS021A10.b.3613

Figure legends

Insert figure legends here

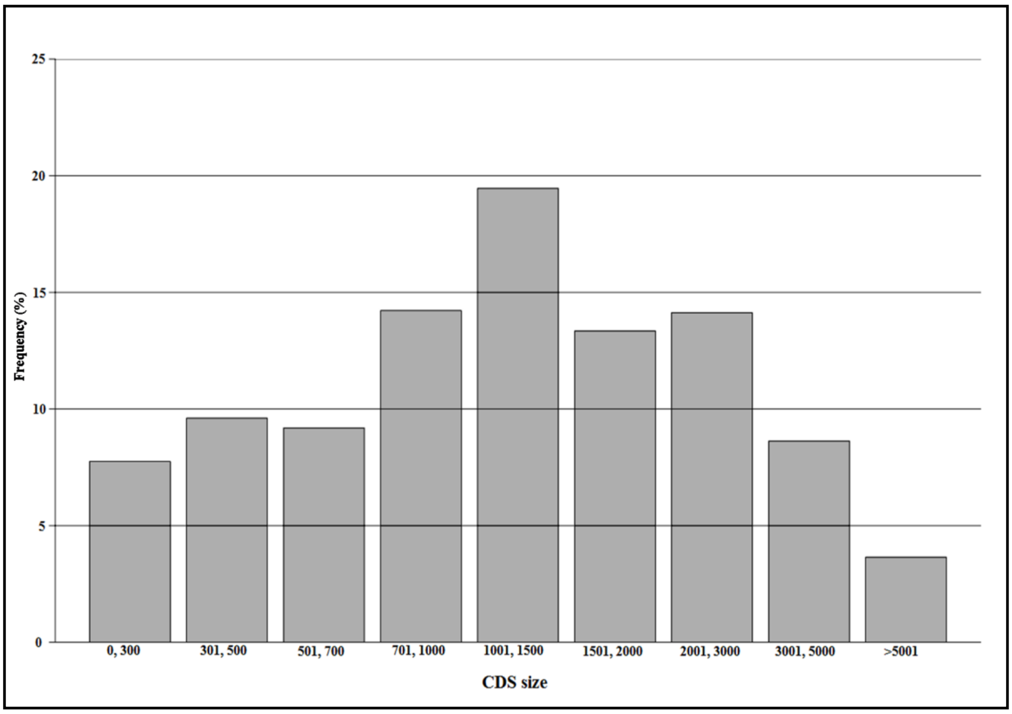


Figure 1: Average size (bases) from putative CDS identified in *L. amazonensis* genome.

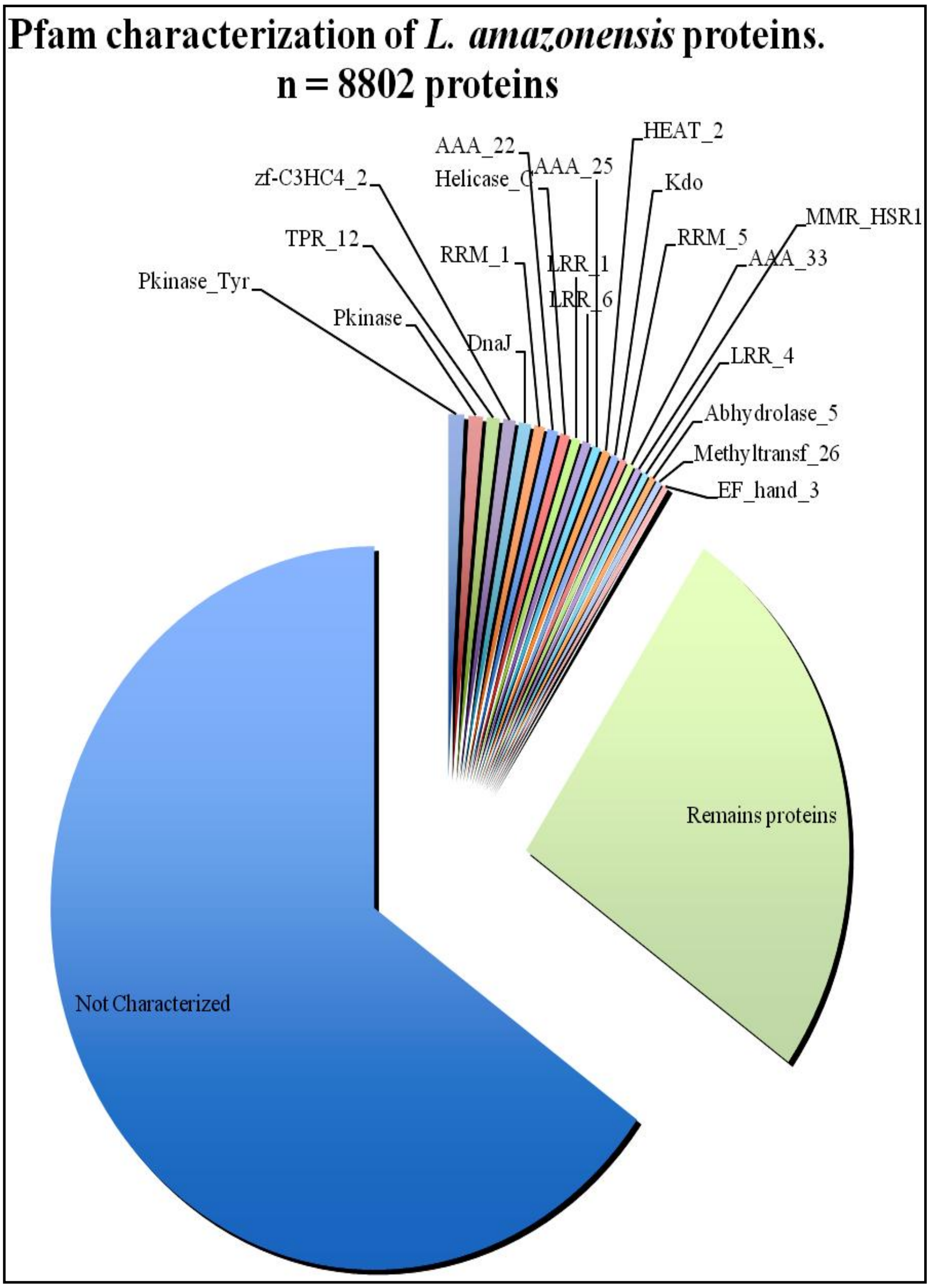


Figure 2: Proteins families identification generated by PFAM database. Only 20 most abundant families were represented in the figure. Remaining families are grouped into green square and not characterized proteins are in blue.

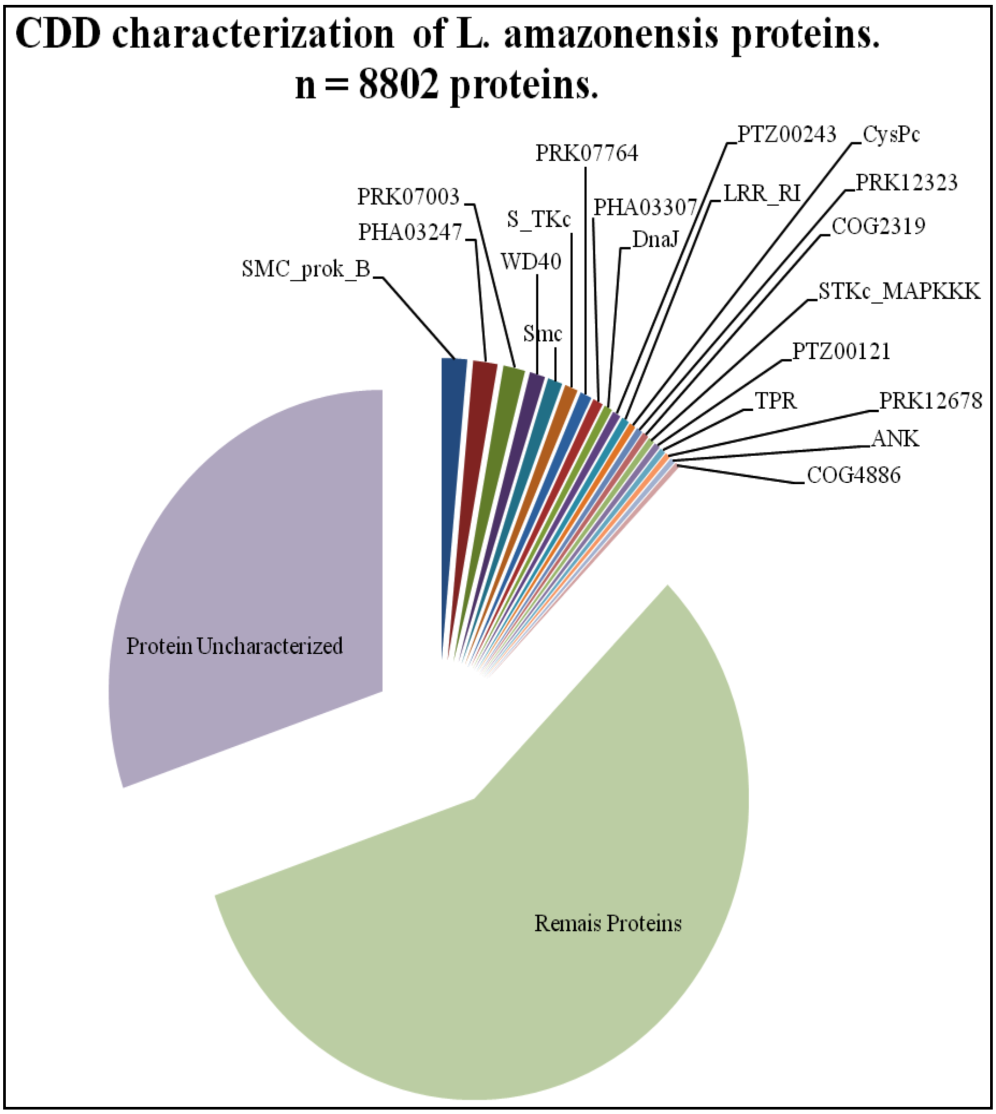
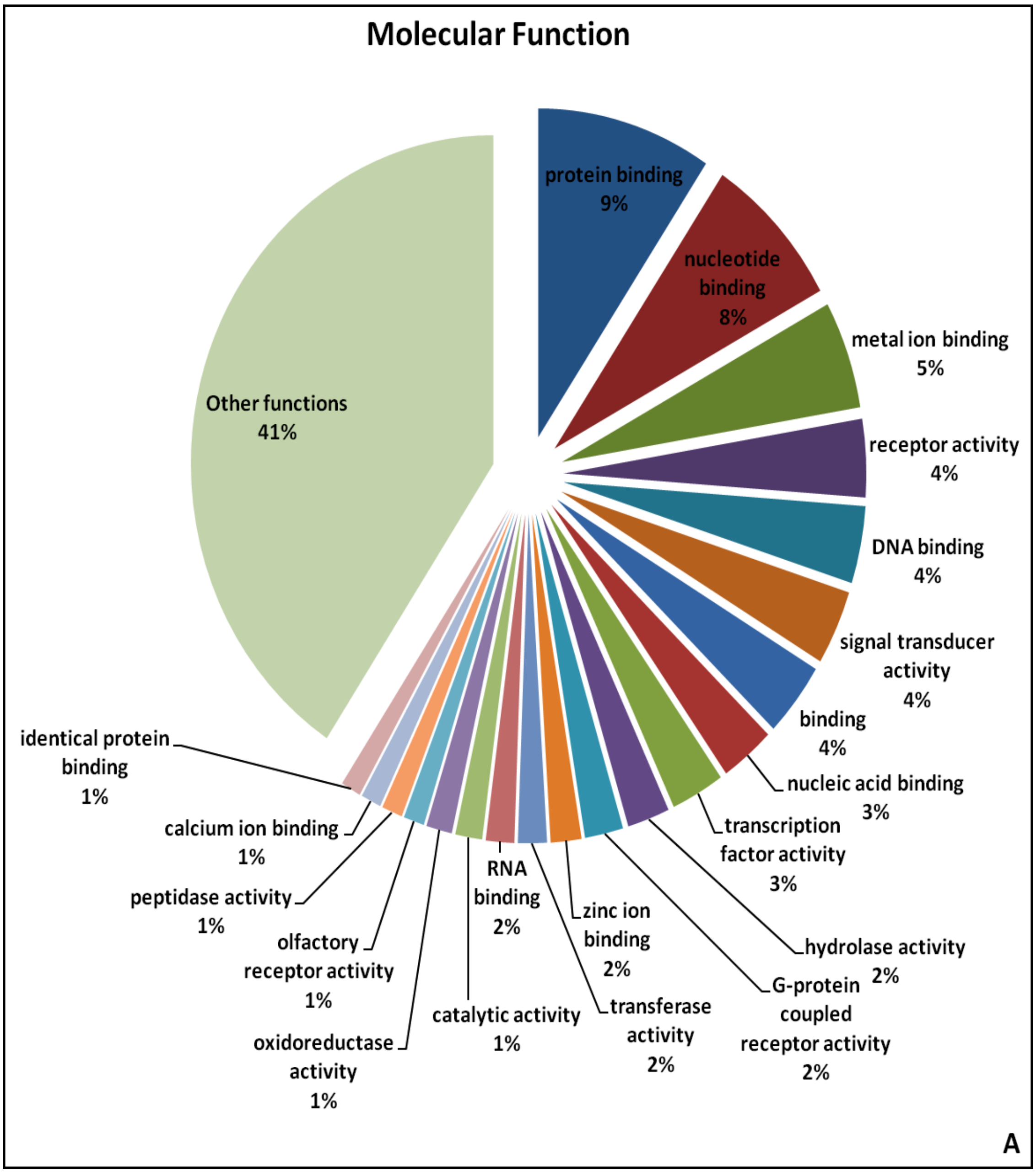
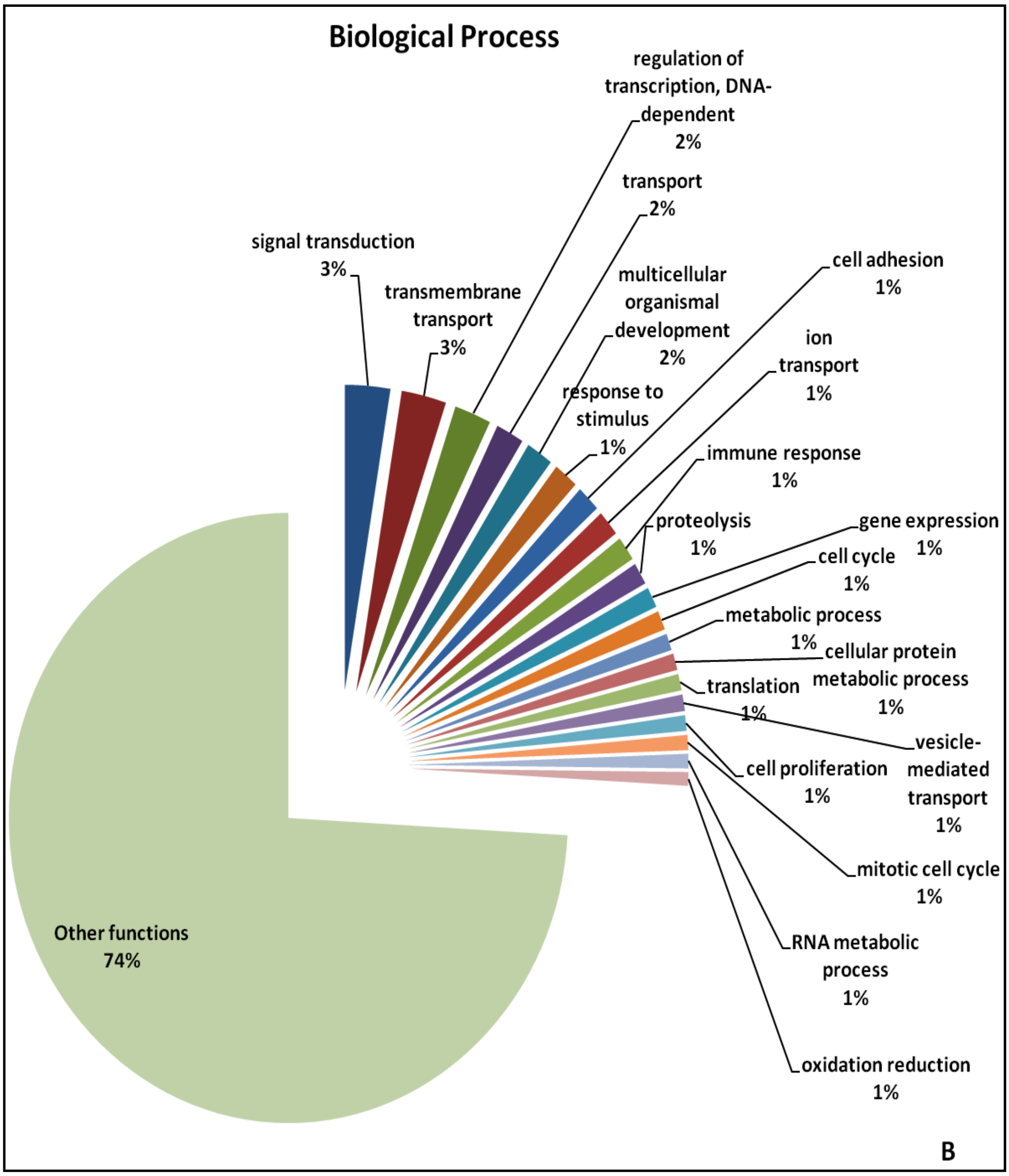


Figure 3: Conserved domains identification generated by RPSBlast with CDD database. Only 20 most domains were represented in the legend. Remaining families are grouped into green square and uncharacterized proteins are in purple.


Molecular Function
protein binding 9%
nucleotide binding 8%
metal ion binding 5%
receptor activity 4%
DNA binding 4%
signal transducer activity 4%
binding 4%
nucleic acid binding 3%
transcription factor activity 3%
hydrolase activity 2%
G-protein coupled receptor activity 2%
transferase activity 2%
zinc ion binding 2%
RNA binding 2%
catalytic activity 1%
oxidoreductase activity 1%
olfactory receptor activity 1%
peptidase activity 1%
calcium ion binding 1%
identical protein binding 1%
Other functions 41%
A


Biological Process
regulation of transcription, DNA-dependent 2%
transport 2%
signal transduction 3%
transmembrane transport 3%
multicellular organismal development 2%
cell adhesion 1%
ion transport 1%
response to stimulus 1%
immune response 1%
proteolysis 1%
gene expression 1%
cell cycle 1%
metabolic process 1%
cellular protein metabolic process 1%
translation 1%
vesicle-mediated transport 1%
cell proliferation 1%
mitotic cell cycle 1%
RNA metabolic process 1%
oxidation reduction 1%
Other functions 74%
B

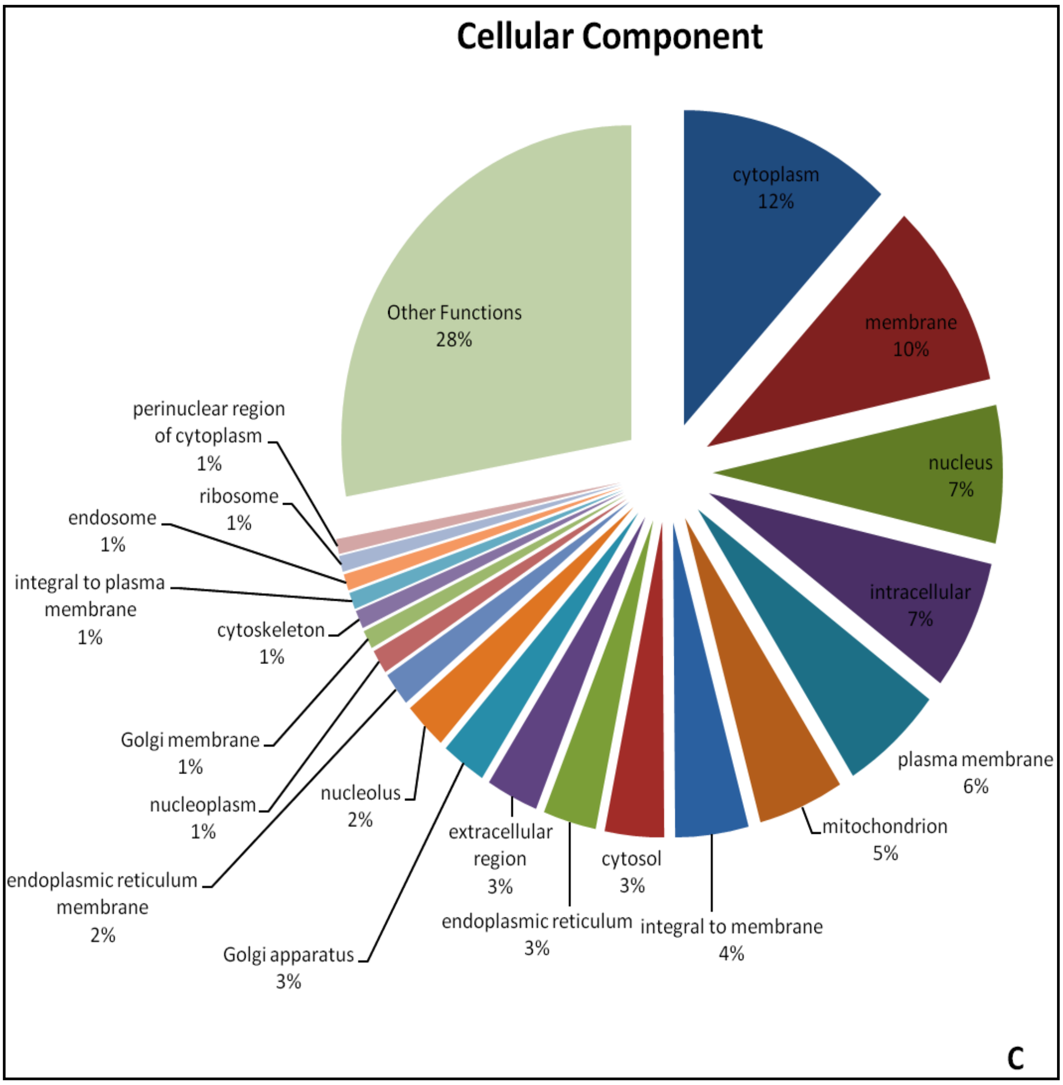


Figure 4 – Gene ontology results to protein characterization in level of Molecular Function (A), Biological Process (B) and Cellular Component (C). Only the 20 most abundant characteristics were listed here.

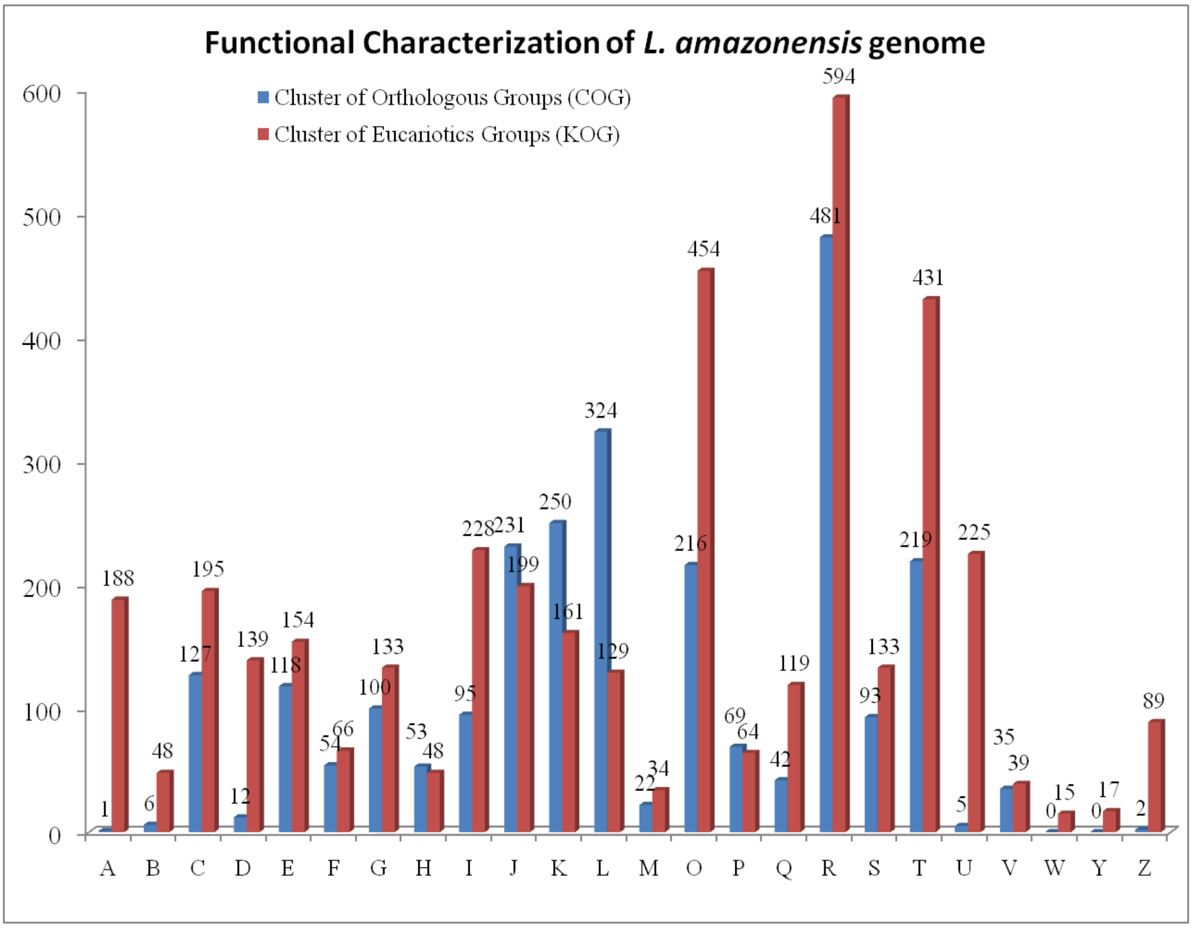


Figure 5: - Functional category by KOG and COG for *Leishmania amazonensis* proteins INFORMATION STORAGE AND PROCESSING: [J] Translation, ribosomal structure and biogenesis, [A] RNA processing and modification, [K] Transcription, [L] Replication, recombination and repair, [B] Chromatin structure and dynamics

CELLULAR PROCESSES AND SIGNALING: [D] Cell cycle control, cell division, chromosome partitioning, [Y] Nuclear structure, [V] Defense mechanisms, [T] Signal transduction mechanisms,

[M] Cell wall/membrane/envelope biogenesis, [N] Cell motility, [Z] Cytoskeleton, [W] Extracellular structures, [U] Intracellular trafficking, secretion, and vesicular transport, [O] Posttranslational modification, protein turnover, chaperones

METABOLISM: [C] Energy production and conversion,[G] Carbohydrate transport and metabolism,

[E] Amino acid transport and metabolism, [F] Nucleotide transport and metabolism, [H] Coenzyme transport and metabolism, [I] Lipid transport and metabolism, [P] Inorganic ion transport and metabolism, [Q] Secondary metabolites biosynthesis, transport and catabolism

POORLY CHARACTERIZED: [R] General function prediction only, [S] Function unknown

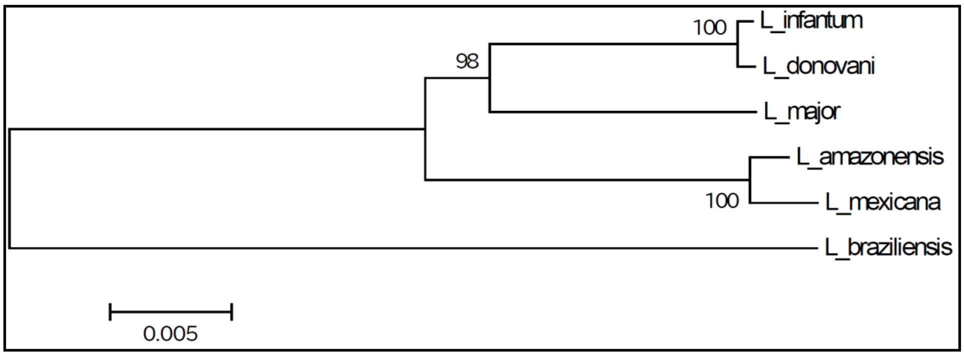


Figure 6 Phylogenomics analysis tree for all six *Leishmania* species, inferred by Maximum Likehood with 1000 boostrap replicates, based on thirty-one universal orthologous (UO) genes.

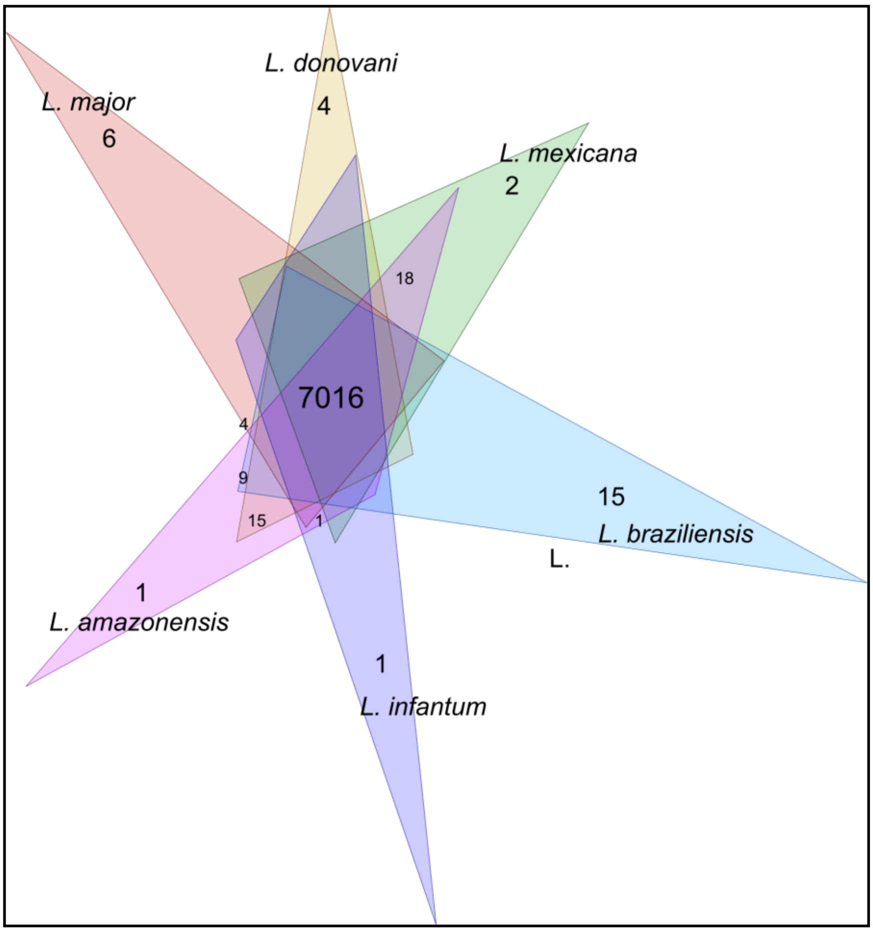
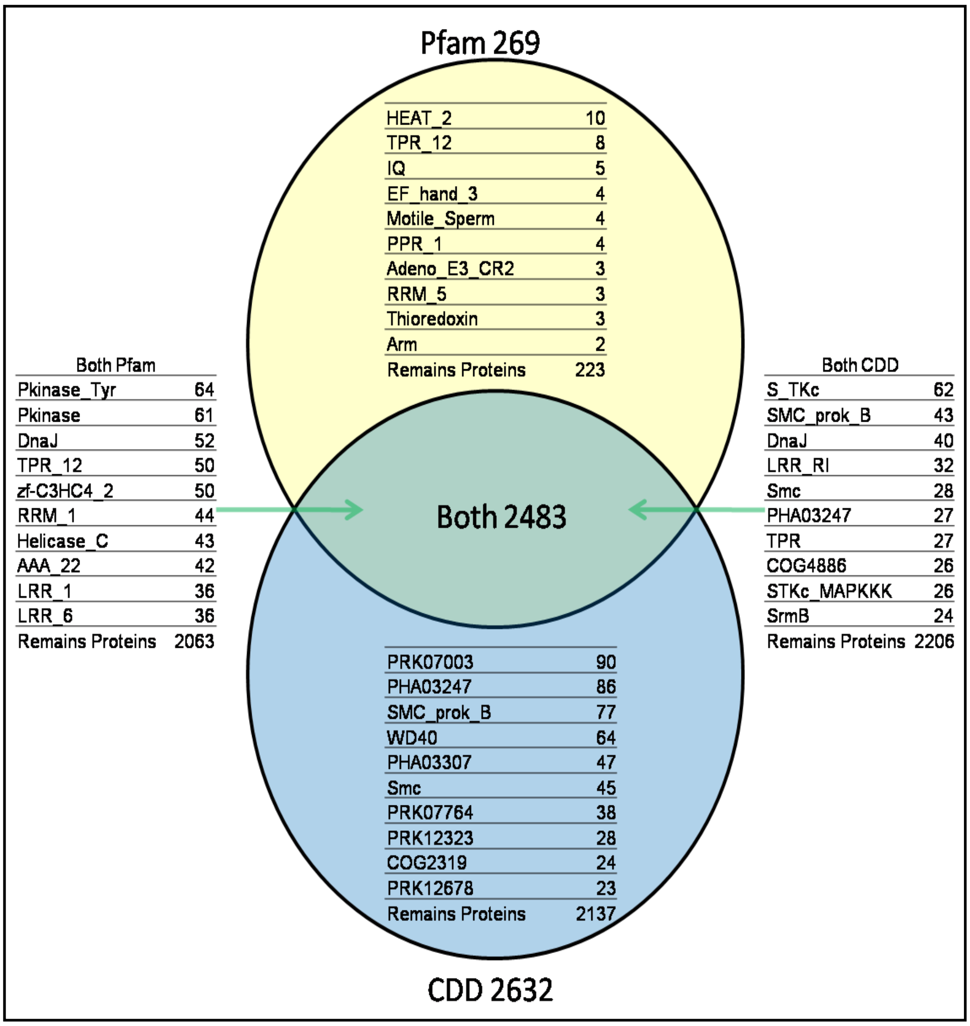


Figure 7. Comparative analysis of species *Leishmania* using orthologous and paralogous protein groups generated by OrthoMCL. The colors represent the number of protein shared between the species. blue (intern paralogous into specie); green orthologous groups between 2 species (*L. amazonensis* and L.mexicana: 18; *L. amazonensis* and L. donovani: 15; L. amazonensis and *L. braziliensis*: 9; *L. amazonensis* and *L. major:* 4; *L. amazonensis* and L.infantum: 1); and red: 7026 orthologous groups shared between all six *Leishmania* species. Orthologous groups shared between 3, 4 and 5 species are yellow.



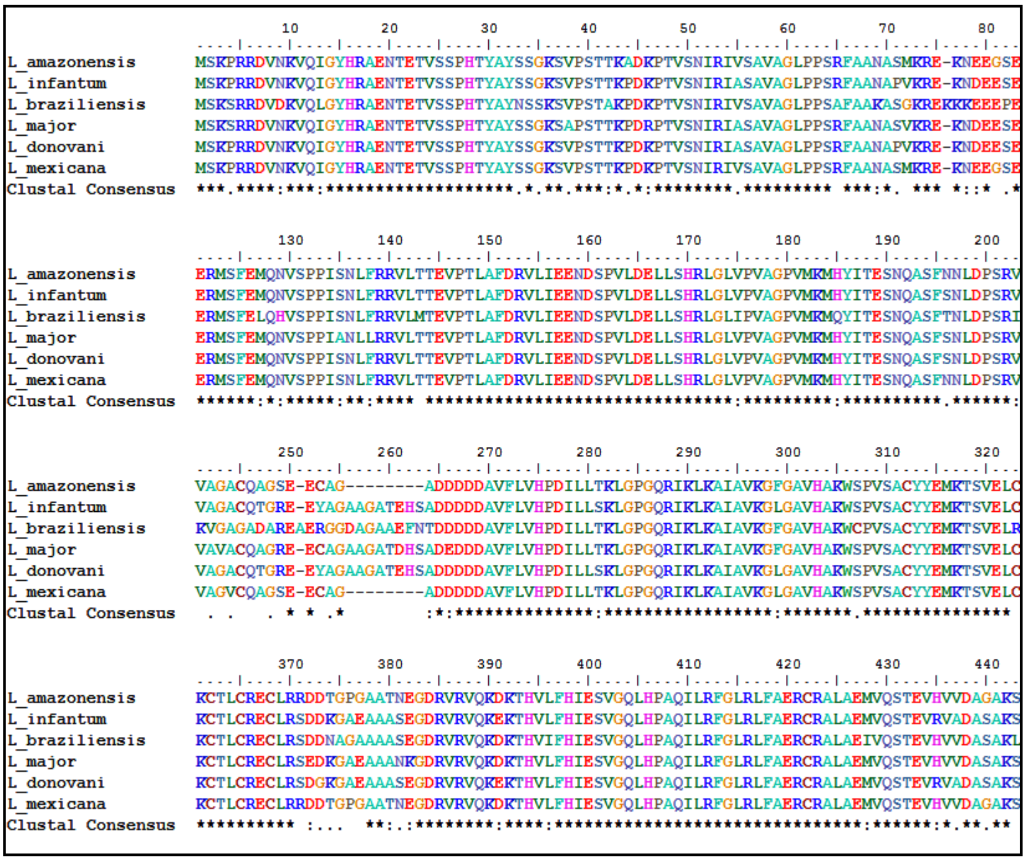
HEAT_2	10
TPR_12	8
IQ	5
EF_hand_3	4
Motile_Sperm	4
PPR_1	4
Adeno_E3_CR2	3
RRM_5	3
Thioredoxin	3
Arm	2
Remains Proteins	223

Both Pfam	
Pkinase_Tyr	64
Pkinase	61
DnaJ	52
TPR_12	50
zf-C3HC4_2	50
RRM_1	44
Helicase_C	43
AAA_22	42
LRR_1	36
LRR_6	36
Remains Proteins	2063

Both CDD	
S_TKc	62
SMC_prok_B	43
DnaJ	40
LRR_RI	32
Smc	28
PHA03247	27
TPR	27
COG4886	26
STKc_MAPKKK	26
SrmB	24
Remains Proteins	2206

PRK07003	90
PHA03247	86
SMC_prok_B	77
WD40	64
PHA03307	47
Smc	45
PRK07764	38
PRK12323	28
COG2319	24
PRK12678	23
Remains Proteins	2137

Figure 8: In green area, a total of 2483 *L. amazonensis* proteins identified by both Conserved Domains Database (RPSBlast-CDD) and Protein Families Database (HMMER-PFam). In lateral tables we visualize most frequents Families (Pfam) and Domains (CDD). 269 proteins was identified only by Protein Families Database, and inside yellow area we show 10 most frequent families found by Pfam. A total of 2634 proteins was identified only by Conserved Domains Database (CDD), and in blue area the 10 most common domains assigned by CDD in *L. amazonensis*.



```
                           10        20        30        40        50        60        70        80
                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|...
L_amazonensis      MSKPRRDVNKVQIGYHRAENTETVSSPHTYAYSSGKSVPSTTKADKPTVSNIRIVSAVAGLPPSRFAANASMKRE-KNEEGSE
L_infantum         MSKPRRDVNKVQIGYHRAENTETVSSPHTYAYSSGKSVPSTTKPDKPTVSNIRIASAVAGLPPSRFAANAPVKRE-KNDEESE
L_braziliensis     MSKSRRDVDKVQLGYHRAENTETVSSPHTYAYNSSKSVPSTAKPDKPTVSNIRIVSAVAGLPPSAFAAKASGKREKKKEEEPE
L_major            MSKSRRDVNKVQIGYHRAENTETVSSPHTYAYSSGKSAPSTTKPDRPTVSNIRIASAVAGLPPSRFAANASVKRE-KNDEESE
L_donovani         MSKPRRDVNKVQIGYHRAENTETVSSPHTYAYSSGKSVPSTTKPDKPTVSNIRIASAVAGLPPSRFAANAPVKRE-KNDEESE
L_mexicana         MSKPRRDVNKVQIGYHRAENTETVSSPHTYAYSSGKSVPSTTKPDKPTVSNIRIVSAVAGLPPSRFAANASMKRE-KNEEGSE
Clustal Consensus  ***.****:***:*******************.*.**.***:*.*:********.********* ***:*. *** *::* .*

                          130       140       150       160       170       180       190       200
                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|...
L_amazonensis      ERMSFEMQNVSPPISNLFRRVLTTEVPTLAFDRVLIEENDSPVLDELLSHRLGLVPVAGPVMKMHYITESNQASFNNLDPSRV
L_infantum         ERMSFEMQNVSPPISNLFRRVLTTEVPTLAFDRVLIEENDSPVLDELLSHRLGLVPVAGPVMKMHYITESNQASFSNLDPSRV
L_braziliensis     ERMSFELQHVSPPISNLFRRVLMTEVPTLAFDRVLIEENDSPVLDELLSHRLGLIPVAGPVMKMQYITESNQASFTNLDPSRI
L_major            ERMSFEMQNVSPPIANLLRRVLTTEVPTLAFDRVLIEENDSPVLDELLSHRLGLVPVAGPVMKMHYITESNQASFSNLDPSRV
L_donovani         ERMSFEMQNVSPPISNLFRRVLTTEVPTLAFDRVLIEENDSPVLDELLSHRLGLVPVAGPVMKMHYITESNQASFSNLDPSRV
L_mexicana         ERMSFEMQNVSPPISNLFRRVLTTEVPTLAFDRVLIEENDSPVLDELLSHRLGLVPVAGPVMKMHYITESNQASFNNLDPSRV
Clustal Consensus  ******:*:*****:**:**** *******************************:*********:**********.******:

                          250       260       270       280       290       300       310       320
                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|...
L_amazonensis      VAGACQAGSE-ECAG--------ADDDDDAVFLVHPDILLTKLGPGQRIKLKAIAVKGFGAVHAKWSPVSACYYEMKTSVELC
L_infantum         VAGACQTGRE-EYAGAAGATEHSADDDDDAVFLVHPDILLSKLGPGQRIKLKAIAVKGLGAVHAKWSPVSACYYEMKTSVELC
L_braziliensis     KVGAGADAREAERGGDAGAAEFNTDDDDDAVFLVHPDILLTKLGPGQRIKLKAIAVKGFGAVHAKWCPVSACYYEMKTSVELR
L_major            VAVACQAGRE-ECAGAAGATDHSADEDDDAVFLVHPDILLTKLGPGQRIKLKAIAVKGFGAVHAKWSPVSACYYEMKTSVELC
L_donovani         VAGACQTGRE-EYAGAAGATEHSADDDDDAVFLVHPDILLSKLGPGQRIKLKAIAVKGLGAVHAKWSPVSACYYEMKTSVELC
L_mexicana         VAGVCQAGSE-ECAG--------ADDDDDAVFLVHPDILLTKLGPGQRIKLKAIAVKGLGAVHAKWSPVSACYYEMKTSVELC
Clustal Consensus   . .   . * * .*        :*:**************:*****************:*******.***************

                          370       380       390       400       410       420       430       440
                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|...
L_amazonensis      KCTLCRECLRRDDTGPGAATNEGDRVRVQKDKTHVLFHIESVGQLHPAQILRFGLRLFAERCRALAEMVQSTEVHVVDAGAKS
L_infantum         KCTLCRECLRSDDKGAEAAASEGDRVRVQKEKTHVLFHIESVGQLHPAQILRFGLRLFAERCRALAEMVQSTEVRVADASAKS
L_braziliensis     KCTLCRECLRSDDNAGAAAASEGDRVRVQKDKTHVIFHIESVGQLHPAQILRFGLRLFAERCRALAEIVQSTEVHVVDASAKL
L_major            KCTLCRECLRSEDKGAEAAANKGDRVRVQKDKTHVLFHIESVGQLHPAQILRFGLRLFAERCRALAEMVQSTEVHVVDASAKS
L_donovani         KCTLCRECLRSDGKGAEAAASEGDRVRVQKEKTHVLFHIESVGQLHPAQILRFGLRLFAERCRALAEMVQSTEVRVADASAKS
L_mexicana         KCTLCRECLRRDDTGPGAATNEGDRVRVQKDKTHVLFHIESVGQLHPAQILRFGLRLFAERCRALAEMVQSTEVHVVDAGAKS
Clustal Consensus  ********** :...  **:.:********:****:*******************************:******:*.**.**
```

Fig. 9 Alignment of DNA-directed RNA polymerase, alpha subunit, sequences between the 6 *Leishmania* species. We color-coded the sites with identical residues with the same color, and used asterisks to indicate the conserved residues in all species.

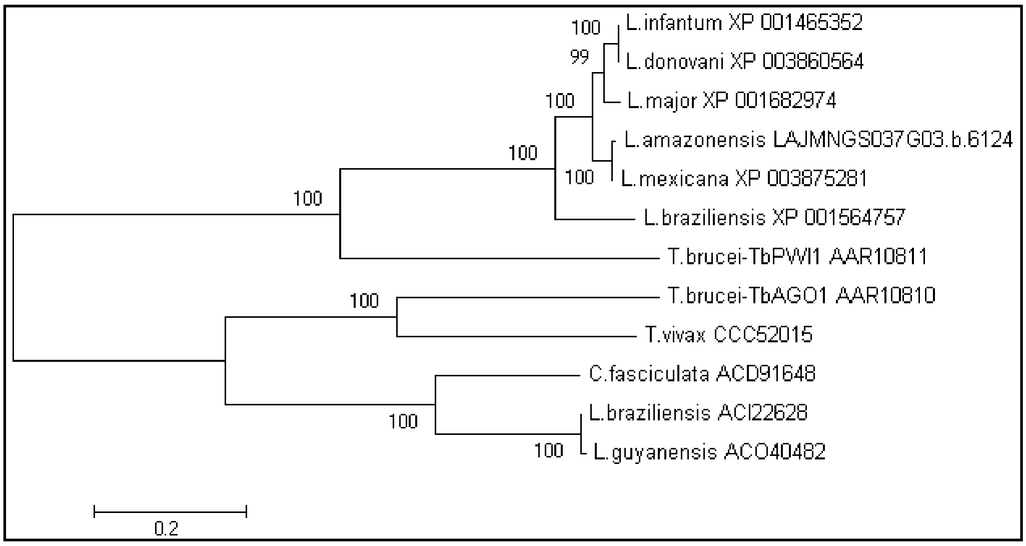


Figure 10. Phylogenetic relationship among argonaute-like genes in Trypanosomatids, constructed by Neighbor-Joinning with 1000 bootstrap replicates.

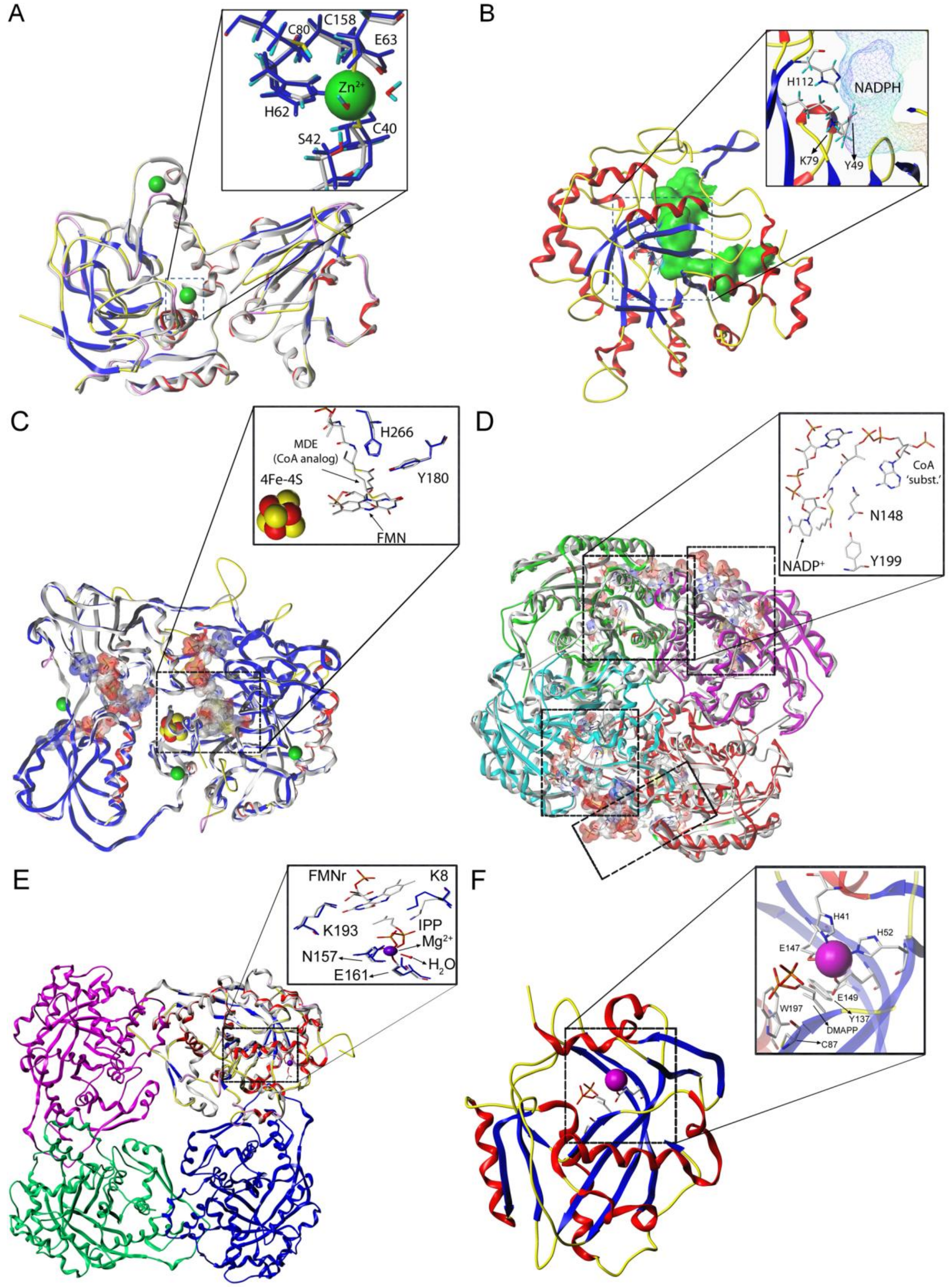


Figure 11. Structural comparison of selected intergenomic NISE cases between *L. amazonensis* and Human. Top panel (EC 1.1.1.2): A LAJMNGS050H11.b.7960 "putative NADP-dependent alcohol dehydrogenase" from *L. amazonensis* (A) and human Aldehyde reductase (PDB 2ALR) (B). Middle panel (EC 1.3.1.34): LAJMNGS010C07.b.1806 "putative 2,4-dienoyl-coa reductase FADH1" from *L.*

amazonensis (C) and human mitochondrial 2,4-dienoyl-CoA reductase (PDB 1W6U) (D). Bottom panel (EC 5.3.3.2): LAJMNGS034G09.b.5743 "putative isopentenyl-diphosphate delta-isomerase" from L. amazonensis (E) and human Isopentenyl-diphosphate Delta-isomerase (PDB 2ICK) (F). Models for all proteins are presented as ribbons. Parasite proteins are colored by secondary structure and presented superposed on their templates (gray ribbons) used in homology modeling. Human analogs are colored by secondary structure, except for 1W6U, which is colored by chain and presented superposed on the peroxisomal isoform (PDB) shown as gray ribbons. The insets show details of the proposed catalytic residues and co-factors for each analogous enzyme. Residues colored blue belong to the parasite enzymes while residues from human analogs are color-coded by atom type.]

Supplementary material

Table S1:OrthoMCL identifier for each LCP ortholog, identifier for protein into LCP Ortholog, protein function and species name for each sequence.

Table S2: Intergenomic NISEs, original annotation and oficial enzime names, fold and superfamily classification based on SUPERFAMILY database and the function of each enzyme.

Table S3: Intragenomic NISEs, original annotation and oficial enzime names, folds and superfamilies classification based on SUPERFAMILY database and the function of each enzyme.

Table S4: Total of enzimatic activities with NISE cases detected by our methodology.

Table S5: Intergenomic NISE as potential drug target searched among three drug target databases (TDR target, TTD, DrugBank).

.

6- Discussão

Genômica comparativa dos 22 protozoários

Neste ponto, discutiremos os resultados destas análises de genômica comparativa obtidos com o programa OrthoMCL, que correspondem aos 22 protozoários estudados nos capítulos quatro e cinco desta tese. Os dados dos genomas e proteomas foram armazenados no ProtozoaDB 2.0, bem como os resultados da análise de homologia. Através da análise de homologia, observamos que a espécie com o maior número de proteínas agrupadas pelo OrthoMCL foi *T. vaginalis* com 44241, dado que é a espécie com o maior proteoma no estudo, com cerca de 60 mil proteínas de acordo com Carlton e colaboradores (Carlton et al., 2007). Entretanto, ela apresenta poucos grupos ortólogos identificados neste estudo, 1775, que somam 8591 proteínas ortólogas. Este baixo valor de grupos ortólogos e o elevado valor de grupos parálogos, 2933, possivelmente devem-se ao fato desta ser a única espécie pertencente ao táxon *Parabasalia*; ou devido ao massivo processo de expansão que este organismo sofre, como observado por Carlton e colaboradores (Carlton et al., 2007). A segunda espécie que exibiu mais proteínas agrupadas foi *T. cruzi*, com 18049 proteínas (5088 parálogas e 12961 ortólogas) em 6833 grupos ortólogos e 814 parálogos. Contudo, dos 6833 ortólogos, 72.63% (4963/6833) destes grupos são co-ortólogos. Este grande número de proteínas co-ortólogas pode ser explicado pelo fato de que a cepa sequenciada, CL-Brenner, é híbrida e possui, geralmente, os dois alelos de cada gene. Além disso, esta espécie possui grandes expansões de proteínas de superfície e membrana (El-Sayed et al., 2005a), corroborando nossos resultados, uma vez que encontramos muitos grupos co-ortológos e parálogos, principalmente relacionados às proteínas de superfície. Quanto ao número elevado de grupos ortólogos encontrados em *T. cruzi* quando comparado a outras espécies, deve ser levado em consideração que existem mais quatro espécies de Kinetoplastida neste estudo que compartilham muitos genes entre si (El-Sayed et al., 2005b; Peacock et al., 2007) e possuem características muito específicas deste grupo (Simpson et al., 2006; Deschamps et al., 2011). Quando analisamos o número de grupos ortólogos formados, observamos que eles permanecem constantes dentro dos gêneros, como por exemplo, para o gênero *Leishmania*. Este fato é esperado, uma vez que estas espécies são intimamente relacionadas (Grimaldi & Tesh, 1993; Peacock et al., 2007). Até mesmo o número de parálogos internos (espécie-específicos) das espécies do gênero *Leishmania* é pequeno, onde *L. major* apresenta nove parálogos, *L. infantum* oito e *L. braziliensis*

13 parálogos. Outro fator interessante é que devido à proximidade entre estas espécies, o número de proteínas agrupadas pelo OrthoMCL ficou muito próximo, com poucas proteínas classificadas como órfãos. Entretanto, *L. braziliensis* possui um número ligeiramente inferior de grupos ortólogos, em comparação com as outras duas espécies do subgênero *Leishmania,* uma vez que pertence ao subgênero *Viannia* (Grimaldi & Tesh, 1993; Peacock et al., 2007).

Analisando o núcleo proteômico dos 22 protozoários estudados, encontramos 348 grupos ortólogos compartilhados por estas espécies. Observamos que as proteínas deste núcleo proteômico são geralmente relacionadas à manutenção celular e processamento da informação, uma vez que foram encontradas várias proteínas ribossomais, histonas, proteínas do citoesqueleto e tRNA sintetase. Além disso, de acordo com a classificação funcional do KOG, a maioria dos ortólogos é distribuída nas categorias funcionais "J", "A" e "O", confirmando a idéia de que esses ortólogos são mais relacionados à manutenção e processamento, subdividindo-se em duas seções: "Armazenamento e Processamento de Informação" e "Processo e sinalização Celular", conforme observado por Ciccarelli e colaboradores (Ciccarelli et al., 2006) e Liu e colaboradores (Liu et al., 2012). Prosseguindo as análises no núcleo proteômico dos protozoários, conseguimos categorizar funcionalmente 275 ortólogos deste núcleo utilizando os genes ortólogos de procariotos (COG). Sendo que dois grupos destes 275 ortólogos são mais similares aos Procariotos do que aos Eucariotos (methionil-tRNA sintetease e histona acetil-transferase), sugerindo que estas proteínas sejam comuns tanto a procariotos quanto a eucariotos; ou que alguns destes genes foram transferidos. Cavalier-Smith (Cavalier-Smith, 2010) argumenta que a LGT (transferência lateral de genes do inglês *Lateral Gene Transfer*) é importante evolutivamente entre bactérias e protozoários, corroborando resultados de trabalhos anteriores que descrevem que os genomas de: (i) *Cryptospodirium hominis* (Xu et al., 2004), (ii) das espécies do gênero *Leishmania* (Ivens et al., 2005; Peacock et al., 2007), (iii)*T. brucei* (Berriman et al., 2005) e (iv) *E. histolytica* (Loftus et al., 2005) possuem genes de origem bacteriana, contribuindo para algumas das diferenças metabólicas encontradas nestes protozoários. Estes estudos relatam que é possível a existência de transferência lateral entre procariotos e eucariotos, sendo que estes eventos podem ser responsáveis pelo mosaico de genes encontrado nos protozoários. Um exemplo de gene transferido que possivelmente foi adquirido de bactérias, o piruvato ferredoxina oxidoredutase está relacionado à fermentação em *G. lamblia, Entamoeba* spp. e *T. vaginalis* (Loftus et al., 2005; Carlton et al., 2007; Morrison et al., 2007). Em nosso estudo, encontramos este gene compartilhado por estas quatro espécies e também por *C. muris* e *C. parvum*.

O número de ortólogos aumenta consideravelmente quando analisamos organismos mais próximos. Por exemplo, ao analisarmos os organismos *T. cruzi*, *T. brucei* e *L. major*, que são próximos, encontramos 5739 grupos ortólogos compartilhados entre eles, enquanto El-Sayed e colaboradores (El-Sayed et al., 2005b), utilizando uma metodologia diferente, encontraram 6158. Quando analisamos os cinco proteomas dos Kinetoplastida (*T. brucei*, *T. cruzi*, *L. major*, *L. infantum* e *L. braziliensis*), observamos que compartilham cinco mil grupos ortólogos, contrastando com o grupo artificial formado pelos quatro protozoários *E. histolytica*, *E. dispar*, *G. lamblia* e *T. vaginalis* que compartilham somente 667 grupos ortólogos em comum. Entretanto, ao analisarmos as proteínas compartilhadas pelo grupo monofilético *E. histolytica* e *E. dispar* (Lorenzi et al., 2010), observamos que eles compartilham 5915 grupos ortólogos. Em outro exemplo de agrupamento por similaridade, Brayton e colaboradores (Brayton et al., 2007) compararam os proteomas de *B. bovis*, *T. parva* e *P. falciparum* encontrando 1945 ortólogos comuns as três espécies e 2650 ortólogos compartilhados entre *B. bovis* e *T. parva*. Em nossas análises, observamos que o número de proteínas compartilhadas entre estas duas espécies é de 2667 grupos e de 1947 para as três espécies. Esta diferença de valores encontrada entre os ortólogos compartilhados pelas três espécies e por *B. bovis* e *T. parva*, pode ser explicada pelo fato de que embora estes três organismos pertençam ao filo Apicomplexa, eles são um pouco distantes evolutivamente. *B. bovis* e *T. parva* são classificados como piroplasmídeos (Sato, 2011), enquanto *P. falciparum* é classificado como *Haemosporida.* Todavia, quando comparamos organismos que pertencem ao mesmo gênero, dentro do filo Apicomplexa, como as seis espécies do gênero *Plasmodium* causadores da malária (três parasitas de humanos: *P. falciparum*, *P. knowlesi* e *P. vivax*; e três parasitas de roedores: *P. berghei*, *P. chabaudi* e *P. yoelii*), geramos 3327 agrupamentos de proteínas ortólogas compartilhadas, valor próximo aos 3336 grupos encontrados por Carlton e colaboradores (Carlton et al., 2008). Analisando a quantidade de proteínas compartilhadas entre organismos que pertencem ao mesmo gênero, encontramos um número maior do que quando comparamos organismos menos relacionados, como foi o caso da comparação entre os piroplasmídeos (*B. bovis* e *T. parva*) e Haemosporida (*P. falciparum*) e dos quatro protozoários não relacionados (*E.*

histolytica, *E. dispar*, *G. lamblia* e *T. vaginalis)*. Portanto, nossos resultados corroboram a hipótese de que organismos filogeneticamente próximos possuem mais genes ortólogos compartilhados entre eles. Estes resultados são esperados, uma vez que organismos filogeneticamente relacionados possuem um maior número de genes conservados e compartilhados. Porém, quando a distância aumenta o número de genes compartilhados diminui, conforme podemos observar neste estudo.

Análisando os grupos de ortólogos específicos, observamos que entre as proteínas Kinetoplastida específicas apresentam funções tais como: (i) transportador ABC, relacionado à resistência às drogas (Torres et al., 2010); (ii) metabolismo de carboidratos e; (iii) transportador de hexose. Sendo que a presença de genes específicos do metabolismo de carboidratos corrobora a idéia que os Cinetoplastídeos apresentam uma maneira diferenciada de metabolizá-los (Simpson et al., 2006; Cáceres et al., 2007). Entre os grupos específicos de Apicomplexa, observa-se que alguns são específicos inclusive na anotação, como por exemplo, uma proteína de membrana apicomplexa-específica. Além disto, observamos que eles apresentam uma forma particular de processamento DNA/RNA, uma vez que existem proteínas relacionadas a esta função, tais como: RNA helicase, RNA metiltransferase e proteínas de ligação de histona, corroborando com Xu e colaboradores (Xu et al., 2004) que encontraram várias proteínas relacionadas à manipulação de nucleotídeos. Dentre as proteínas específicas de *Entamoeba*, podemos observar que ela possui como proteínas específicas: AIG1 com função desconhecida, embora apresente varias cópias em seu genoma; e os genes da família Rab GTPases, específicos de *Entamoeba* neste estudo. Apesar de *T. vaginalis* também apresentar Rab GTPases, estas não possuem similaridade suficiente para serem consideradas ortólogas, contrariando Weedall e Hall (Weedall & Hall, 2011). Por fim, a maior parte dos grupos ortólogos específicos de Kinetoplastida (46,29%; 1592/3396), Apicomplexa (40,63%; 92/224) e *Entamoeba* (65,41%; 2905/4441) têm função hipotética, o que é esperado, pois devido à natureza destes ortólogos muitas vezes não é possível a transferência de anotação por similaridade, dado que não existem genes similares, com funções descritas, em outros organismos previamente estudados.

Estudando os grupos parálogos inferidos pelo programa OrthoMCL, encontramos 4982 grupos de parálogos internos para os 22 protozoários. *T. vaginalis* foi a espécies que apresentou mais parálogos internos, 2933, sendo que o maior parálogo apresentou 1721 cópias. Enquanto que *E. histolytica* apresentou 58 parálogos internos e *L. major* 9 parálogos internos, porém suas expansões são pequenas. A maior expansão em *E. histolytica* é uma “*mucin-like*” com 15 cópias, já em *L. major* as maiores expansões são Tuzina e proteína de antígeno de superfície, com 5 cópias cada. *T. cruzi* possui 5088 cópias distribuídas em 814 parálogos internos e as maiores expansões neste organismo encontradas em um único gene são a Transialidase e mucina TcMUCII com 675 e 579 cópias, respectivamente. Este grande número de duplicações que observamos em *T. vaginalis* e *T. cruzi* é esperado porque estes organismos estudados possuem altas taxas de duplicação gênica (Li et al., 2003; El-Sayed et al., 2005a; Carlton et al., 2007). O gênero *Plasmodium* contém expansões de famílias gênicas (conhecidas como famílias PIR) que codificam proteínas envolvidas na variação antigênica e evasão à resposta do sistema imunedo hospedeiro. Encontramos em trabalhos anteriores que as famílias gênicas *var*, *rifin*, *stevor*, em *P. falciparum* apresentam 60, 140 e 25 cópias, respectivamente. (Carlton et al., 2005; Hall & Carlton, 2005). Assim como nestes trabalhos, também encontramos várias cópias destes parálogos em *P. falciparum* (73 cópias de *var*, 165 de*rifin* e34 de *stevor*). Além disso, em *P. vivax,* identificamos 349 cópias de *vir*, distribuídas em 20 genes, sendo que a maior expansão do gene *vir* possui 116 cópias. Carlton e colaboradores (Carlton et al., 2008) encontraram 346 cópias de *vir*, porém neste trabalho não é mencionado em quantos genes estas cópias estão distribuídas. Observando as expansões de *kir*, em *P. knowlesi,* outra proteína relacionada à variação antigênica, encontramos 61 cópias de *kir* e a maior expansão apresentou 45 cópias. Comparando nossos resultados com o de Pain e colaboradores (Pain et al., 2008), obtivemos um valor um pouco abaixo, uma vez que foram encontradas 68 cópias de *kir* por Pain e colaboradores. Contudo, a maior expansão neste organismo foi relacionada ao gene *SICA* que apresentou 185 cópias em um parálogo, um número superior as 107 encontradas por Pain e colaboradores (Pain et al., 2008). Finalmente, em *P. chabaudi* encontramos 27 cópias de CIR em um único gene e 34 cópias de BIR em *P. berghei,* distribuídas em 13 genes. Estes valores são menores do que os encontrados anteriormente (Carlton et al., 2008), uma vez que foi relatada a presença de 245 cópias de BIR em *P. berghei* e 135 cópias de CIR em *P. chabaudi.* Entretanto, nossos resultados corroboram que destas famílias PIR a mais diversa e extensa é a *vir* encontrada em *P. vivax* (Carlton et al., 2008).

Os *Tritryps* utilizam diferentes estratégias de evasão do sistema imune. *L. major,* por exemplo, altera a função do macrófago que infecta; *T. cruzi* expressa uma complexa variedade de antígenos de superfície a partir das células infectadas e *T. brucei*, finalmente, permanece no meio extracelular contornando a resposta imune do hospedeiro pela troca periódica da maioria da sua proteína de superfície. As glicoproteínas variantes de superfície (VSG) são restritas a *T. brucei*, porém *T. cruzi* possui uma expansão dos domínios de transialidases e mucinas, enquanto *L. major* apresenta um grande arranjo de amastinas de superfície (El-Sayed et al., 2005b). Em nossos resultados, encontramos valores abaixo dos descritos na literatura para estes genes responsáveis pela fuga do sistema imune. Por exemplo, em *T. brucei* encontramos 166 cópias de VSG distribuídas em 16 parálogos, enquanto Berriman e colaboradores encontraram 806 cópias distribuídas em um grande arranjo telomérico (Berriman et al., 2005). Em *T. cruzi* encontramos 697 cópias de transialidase, 919 cópias de MASP, 564 de TcMUCII e 218 cópias do gene GP63, contudo de acordo com a literatura (El-Sayed et al., 2005a) 50% de seu genoma formado por sequências repetidas, e tal montagem apresentou 1430 cópias do gene da Transialidases, incluindo 693 pseudogenes, 1377 de MASP (com 433 pseudogenes), 863 de Mucinas (201 pseudogenes) e 425 cópias de gp63 (251 pseudogenes). *L. major* possui 57 cópias de amastina (Ivens et al., 2005), considerado um número baixo de genes duplicados quando comparado com o genoma de *T. cruzi* e *T. brucei*. Porém, em nosso estudo encontramos três cópias de amastina somente em *L. major*, oito cópias específicas de *L. infantum* e três em *L. braziliensis*. Além disso, é interessante citar que encontramos seis genes co-ortólogos de amastina compartilhado por estas três espécies do gênero *Leishmania*. Esta diferença entre o número de genes duplicados descritos na literatura e o número encontrado em nosso estudo pode ser devido ao fato de trabalharmos com as proteínas, enquanto estes estudos utilizaram os genes, permitindo a eles avaliar a presença de pseudogenes.

A duplicação gênica é importante para a geração de diversidade antigênica, particularmente em *T. brucei*, *T. cruzi* e *Plasmodium*. Portanto, com base nessa perspectiva, é possível argumentar que a quantidade reduzida de cópias de proteínas de superfície encontradas em *Leishmania* e *Entamoeba*, em nosso estudo, pode refletir as suas diferentes estratégias de evasão do sistema imune, uma vez que ambos parasitas passam pouco tempo no tecido sanguíneo, ficando menos expostos ao sistema imune. Esta pode ser uma das razões para a pouca quantidade de proteínas/genes parálogos internos encontrados. De fato, a maioria dos genes duplicados está relacionada com a resposta contra o sistema imune, como observado nos trabalhos de El-Sayed e colaboradores (El-Sayed et al., 2005b), Berriman e colaboradores (Berriman et al., 2005) e Hall e Carlton (Hall & Carlton, 2005), embora também tenhamos observado parálogos relacionados a mecanismos de transdução de sinal.

Em relação à categorização funcional observada para os genes parálogos em *T. vaginalis,* após a categoria "R" (somente função descrita), a segunda mais abundante foi "T" (Mecanismo de Transdução de Sinal) e encontramos também proteínas pertencentes à categoria "M" (Biogênese de membrana). Como descrito na literatura, muitas cópias auxiliam o organismo a produzir uma grande quantidade de proteínas (Gu et al., 2004; Nielsen, 2006). Portanto, uma maior diversidade no repertório de proteínas de superfície, com o objetivo de evadir do sistema imune do hospedeiro, pode auxiliar o parasita nesta tarefa. Além disso, o fato de encontrarmos uma grande quantidade de parálogos relacionados a mecanismos de transdução de sinal podem indicar um complexo sistema de sinalização e percepção do ambiente. Em *T. cruzi* é interessante notar que as categorias "M" e "T" estão presentes nesta categorização, uma vez que é conhecido que esta espécie possui grandes famílias de proteínas de superfície, como Mucinas entre outras. Deste modo, tais duplicações são esperadas porque aumentam o repertório de defesa possibilitando a troca destas proteínas, corroborando com os resultados encontrados no presente estudo. É interessante ressaltar que foram encontradas várias proteínas espécie-específicas do tipo cisteína em *E. dispar* e *E. histolytica*. Porém, encontramos em *E. histolytica* uma cisteína peptidase duplicada como parálogo interno não encontrada em *E. dispar*. Sabe-se que este gene é um fator chave da virulência em *Entamoeba* (Weedall & Hall, 2011), o que pode explicar o fato de não termos encontrado cisteínas peptidases duplicadas no protozoário não patogênico *E. dispar.*

Com isto, finalizamos a discussão da genômica comparativa dos 22 protozoários e passaremos a discutir, a partir deste ponto, as funcionalidades do sistema de anotação STINGRAY e seu uso na anotação do genoma de *L. amazonensis*, bem como sua análise comparativa com as outras espécies do gênero *Leishmania.*

Analisando e comparando o genoma de *L. amazonensis*

Depois de anotarmos o genoma de *L. amazonensis* utilizando o STINGRAY, procedemos com as análises deste genoma e com as análises comparativas utilizando o programa OrthoMCL e os proteomas de *L. braziliensis*, *L. donovani*, *L. major*, *L. mexicana*, *L. infantum*, juntamente com o proteoma predito de *L. amazonensis* gerado no transcorrer desta tese.

A montagem do genoma de *L. amazonensis* resultou em 29.670.588 bases, 8802 possíveis regiões codificantes de sequências (CDS) com um conteúdo de GC de 59% para os contigs e 61,12% para a CDS. Real e colaboradores (Real et al., 2013) verificaram que o tamanho do genoma de *L. amazonensis* é de 29,6 Mb com conteúdo GC de 58,5% para o genoma e 61% para as CDS, e 8.168 putativos genes. Em nossa análise, encontramos alguns genes duplicados como, por exemplo, 44 cópias de calpaina, uma cisteíno peptidase dependente de cálcio, envolvida na remodelação do citoesqueleto importante durante a diferenciação de *Leishmania* ou adesão na membrana (Mottram et al., 2004; Ersfeld et al., 2005). Além disto, este gene é um potencial alvo para drogas, demonstrado por d'Avila-Levy e colaboradores (d'Avila-Levy et al., 2006). Encontramos também outros genes duplicados como, por exemplo, tuzinas e amastinas. No genoma de *L. amazonensis,* encontramos sete cópias de tuzinas, sendo uma delas ortóloga somente à *L. mexicana*. O fato de estas duas espécies pertencerem ao mesmo complexo taxonômico (Eresh et al., 1994; Stiles et al., 1999) pode justificar tal resultado. A diversidade de tuzinas em *L. amazonensis* é moderada quando comparada a *L. major* que possui a maior diversidade de tuzinas com 28 cópias (Peacock et al., 2007). Sua função é desconhecida, porém sabe-se que estão associados a amastinas, também de função desconhecida, mas com expressão abundante na forma amastigota (Rochette et al., 2005; Jackson, 2010). As amastinas pertencem a uma grande família de proteínas de superfície única nos Kinetoplastídeos (Rochette et al., 2005), corroborando nosso estudo de genômica comparativa dos 22 protozoários, e são utilizadas como marcadores para a leishmaniose visceral (Rafati et al., 2006). Embora *L. amazonensis* e *L. mexicana* pertençam ao mesmo complexo (Stiles et al., 1999), causando doenças diferenciadas, já foi verificado que ambas podem causar a leishmaniose visceral (Aleixo, 2006; Mestra et al., 2011); e interessantemente, encontramos uma amastinacompartilhada apenas por estas duas espécies. Este e outros resultados, como um gene hipotético compartilhado somente entre estas duas espécies, que para Rogers (Rogers et al., 2011) é *L. mexicana*-específico, reforçam a idéia de que estas duas espécies são muito próximas.

Analisando comparativamente, *L. braziliensis* apresenta o maior número de genes únicos, 15; menor do que encontrado por Peacock, 47 (Peacock et al., 2007) que comparou *L. braziliensis*, *L. infantum* e *L. major.* Mesmo com a adição dos proteomas de *L. donovani* (Downing et al., 2011) e de *L. amazonensis*, *L. braziliensis* continuou apresentando a maior quantidade de genes únicos. Porém, o número de genes únicos em *L. infantum* diminuiu, pois encontramos quatro genes únicos enquanto Rogers e Peacock encontraram 19 e 27, respectivamente (Peacock et al., 2007; Rogers et al., 2011). Finalmente, ao analisarmos seis espécies de *Leishmania* encontramos um gene único em *L. amazonensis,* enquanto Real (Real et al., 2013) analisando cinco espécies encontrou 23*.* Portanto, quanto mais espécies de *Leishmania* são analisadas paralelamente, menor o número de genes únicos.

Quando analisamos duas espécies de *Leishmania* do Novo Mundo, *L. braziliensis* e *L. amazonensis*, apenas a distribuição geográfica entre elas é semelhante, pois ambas apresentam vetores diferentes (*L. amazonensis*: *Lutzomyia flaviscutellata* e *Lu. longipalpis*; *L. braziliensis*: *Lu. wellcomei*), manifestações clínicas diferenciadas e pertencem a sub-genêneros diferentes (Sharma & Singh, 2008), o que justifica a semelhança entre nossos resultados com estudos anteriores, onde *L. braziliensis* é sempre a espécie mais divergente (Croan et al., 1997; Downing et al., 2011). Como esperado, observou-se que quanto mais próximas as espécies maior o número de genes compartilhados entre elas, principalmente quando se comparam espécies do mesmo complexo. Contudo, quando comparamos espécies dos subgêneros *Viannia* e *Leishmania,* o número de genes comuns é menor, o que pode explicar o maior número de genes únicos encontrados em *L. braziliensis* e o pequeno número encontrado nas outras espécies.

Nossa árvore filogênomica mostrou a separação esperada e conhecida deste gênero para todas as seis espécies analisadas. A árvore apontou que *L. braziliensis* é a espécie mais distante dentre as analisadas e ainda conseguiu recapitular a posição taxonômica conhecida de *L. amazonensis* e das outras cinco espécies de *Leishmania*. A utilização de 31 genes ortólogos universais mostrou-se uma abordagem boa e robusta para inferência de árvore de espécie usando vários genes, evitando-se assim a extrapolação da filogenia de espécie através da utilização de um único gene. Nossa análise foi capaz de agrupar *L. infantum* e *L. donovani* no mesmo ramo assim como no estudo de Mauricio e colaboradores (Mauricio et al., 1999). Simpson e colaboradores (Simpson et al., 2004) conseguiram uma separação razoável do subgênero *Leishmania* e *Viannia*, igualmente a nossa análise. A recapitulação do complexo *L. donovani* (*L donovani* e *L. infantum*) foi constante na maioria dos estudos, bem como a formação de complexo *L. mexicana*; além da separação correta dos subgêneros e da separação das espécies do Velho e do Novo Mundo no subgênero *Leishmania* (Widmer et al., 1989; Mauricio et al., 1999; Simpson et al., 2004; Lukes et al., 2007; Downing et al., 2011; Fraga et al., 2013). Entretanto, estes estudos utilizam no máximo dois genes agrupados e nós ao utilizarmos 31 genes ortológos universais obtivemos um valor mínimo de *bootstrap* de 98, superior aos valores observados em outros trabalhos (Mauricio et al., 1999, 2007; Simpson et al., 2004; Lukes et al., 2007; Downing et al., 2011; Fraga et al., 2013). Finalizando, *L. (V.) braziliensis*foi a espécie mais externa, enquanto que *L. (L.) infantum* e *L. (L.) donovani* ficaram muito próximas, mesmo clado, formando o complexo *L. donovani*. Este complexo, por sua vez, ficou mais próximo à *L. (L.) major*, outra espécie do Velho Mundo que também pertence ao subgênero *Leishmania*. Enquanto que as outras duas espécies do subgênero *Leishmania, L. (L.) mexicana* e *L. (L.) amazonensis,* aparecem formando o complexo *L. mexicana*, são um pouco mais distantes das outras três espécies do subgênero *Leishmania,* uma vez que representam espécies do Novo Mundo.

7 - Conclusões

- Encontramos dentre as proteínas dos 22 protozoários 21119 grupos ortólogos, dos quais 348 são compartilhadas por todas as espécies e formam o núcleo proteômico destes organismos. Sendo que este é um dos primeiros trabalhos a propor um conjunto de genes compartilhados por todos os protozoários e estes genes nos ajudam a entender o que estes organismos apresentam em comum

- A categorização funcional do proteoma núcleo dos 22 protozoários mostrou que a maior parte deste ortólogos está relacionada à manutenção da célula

- *Entamoeba* apresenta o maior número de ortólogos grupo-específicos (4441), seguido pelo Kinetoplastídeos com 3396 ortólogos grupo-específicos. As espécies de Apicomplexa apresentam baixo número de ortólogos grupo-específicos (224) quando comparado a estes outros dois grupos. De certa forma, este resultado é esperado, uma vez que estes organismos apresentam características específicas e não compartilhadas por outros organismos e estes ortólogos podem ser potenciais genes alvos para marcação ou desenho de fármacos, pois caracterizam unicamente um grupo

- *A* categorização funcional dos ortólogos grupo-específicos mostrou que em *Entamoeba*, Kinetoplastida e Apicomplexa aproximadamente 50% destes ortólogos possuem “somente função predita”. Ou seja, a categoria “R” do KOG apresenta funções grupo-específicas que outros organismos não apresentam, podendo ser potencial fonte para marcadores genéticos ou alvos-específicos para drogas

- As espécies que apresentaram o maior número de parálogos foram *T. cruzi* e *T. vaginalis*, respectivamente. Contudo, em número absoluto de cópias *T. vaginalis* foi a espécie que mais apresentou duplicações e estes resultados corroboram trabalhos anteriores que mostraram que estes dois organismos passam por grandes processos de duplicações gênicas, possivelmente para uma melhor adaptação aos seus nichos

- A categorização funcional destes parálogos demonstrou que a categoria funcional “T” (mecanismos de transdução de sinal) do KOG é a mais abundante em *T. cruzi* e *T. vaginalis,* refletindo a importância para estes organismos da percepção e propagação de sinais, principalmente externos, e consequentemente aumentando a sobrevivência dentro de seus hospedeiros evitando assim ambientes desfavoráveis

- A espécie que apresenta o maior número de proteínas órfãs é P. *chabaudi*, seguida de *T. vaginalis.* Não esperávamos este grande número de proteínas órfãs em *P. chabaudi*, uma vez que, teoricamente, a maior parte de seus genes é compartilhada com o gênero *Plasmodium*. Porém, o preditor de genes utilizado em *P. chaubadi*, de acordo com a literatura, fragmentou demasiadamente suas proteínas, podendo explicar este alto valor encontrado. Já o grande número de órfãos em *T. vaginalis* é esperado, pois ela é o único representante de sua ordem, além de apresentar o nicho mais específico de todos

- A anotação do genoma de *L. amazoensis* demonstrou que esta espécie apresenta aproximadamente 8500 genes, sendo que boa parte deles possui função hipotética, o que é esperado, pois levando em consideração que os outros genomas previamente anotados do mesmo gênero também apresentam quase metade de seus genes com função hipotética. Além disso, estes genes representam potencias fontes de novidades evolutivas e necessitam de mais estudos

- A categorização funcional do genoma de *L. amazonesis* demonstrou, como esperado, que a categoria “R” do KOG é a mais abundante, uma vez que a biologia deste organismo apresenta uma grande especificidade de funções, não sendo possível descobrir sua categoria através de inferência por ortologia. A categorização com o PFAM demonstrou que a família mais abundante é a das Tirosina-Kinases, conforme esperado, pois o processo de sinalização celular depende desta família e a percepção do ambiente e sua sinalização são cruciais para a sobrevivência deste parasita

- A análise comparativa do gênero *Leishmania* demonstrou que estas espécies compartilham 7016 ortólogos comuns a todas as seis espécies. O maior número de ortólogos genes compartilhados por duas espécies foi entre *L. amazonensis* e *L. mexicana* (18 ortólogos), conforme esperado, pois estas duas espécies são muito próximas e pertencem ao mesmo complexo*.* Enquanto que o maior número de genes espécie-específicos foi encontrado em *L. braziliensis*, como esperado, uma vez que é a espécie mais distante desta análise e pertence a outro subgênero

8 - Referências

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto C a, et al. Complete genome sequence of the apicomplexan, Cryptosporidium parvum. Science . 2004 Apr 16;304(5669):441–5.

Adam RD. The Giardia lamblia genome. Int J Parasitol . 2000 Apr 10;30(4):475–84.

Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, Barta JR, et al. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol . 2005;52(5):399–451.

Akilov OE, Khachemoune A, Hasan T. Clinical manifestations and classification of Old World cutaneous leishmaniasis. Int J Dermatol . 2007 Feb;46(2):132–42.

Aleixo J. Atypical American visceral leishmaniasis caused by disseminated Leishmania amazonensis infection presenting with hepatitis and adenopathy. Trans … . 2006;

Arango E, Carmona-Fonseca J, Blair S. [In vitro susceptibility of Colombian Plasmodium falciparum isolates to different antimalarial drugs]. Biomedica . 2008 Jun;28(2):213–23.

Arevalo J, Ramirez L, Adaui V, Zimic M, Tulliano G, Miranda-Verástegui C, et al. Influence of Leishmania (Viannia) species on the response to antimonial treatment in patients with American tegumentary leishmaniasis. J Infect Dis . 2007 Jun 15;195(12):1846–51.

Atwood JA, Weatherly DB, Minning TA, Bundy B, Cavola C, Opperdoes FR, et al. The Trypanosoma cruzi proteome. Science . 2005 Jul 15;309(5733):473–6.

Barral A, Pedral-Sampaio D, Grimaldi Júnior G, Momen H, McMahon-Pratt D, Ribeiro de Jesus A, et al. Leishmaniasis in Bahia, Brazil: evidence that Leishmania amazonensis produces a wide spectrum of clinical disease. Am J Trop Med Hyg . 1991 May;44(5):536–46.

Barrett MP, Burchmore RJS, Stich A, Lazzari JO, Frasch AC, Cazzulo JJ, et al. The trypanosomiases. Lancet . 2003 Nov;362(9394):1469–80.

Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome Trypanosoma brucei. Science . 2005 Jul 15;309(5733):416–22.

Brayton K a, Lau AOT, Herndon DR, Hannick L, Kappmeyer LS, Berens SJ, et al. Genome sequence of Babesia bovis and comparative analysis of apicomplexan hemoprotozoa. PLoS Pathog . 2007 Oct 19;3(10):1401–13.

Brown D, Sjölander K. Functional classification using phylogenomic inference. PLoS Comput Biol . 2006 Jun 30;2(6):e77.

Burri C, Keiser J. Pharmacokinetic investigations in patients from northern Angola refractory to melarsoprol treatment. Trop Med Int Health . 2001 May;6(5):412–20.

Cáceres AJ, Quiñones W, Gualdrón M, Cordeiro A, Avilán L, Michels P a M, et al. Molecular and biochemical characterization of novel glucokinases from Trypanosoma cruzi and Leishmania spp. Mol Biochem Parasitol . 2007 Dec;156(2):235–45.

Carlton J. The Plasmodium vivax genome sequencing project. Trends Parasitol . 2003 May;19(5):227–31.

Carlton J, Silva J, Hall N. The genome of model malaria parasites, and comparative genomics. Curr Issues Mol Biol . 2005 Jan;7(1):23–37.

Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, et al. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature . Nature Publishing Group; 2008 Oct 9;455(7214):757–63.

Carlton JM, Angiuoli S V, Suh BB, Kooij TW, Pertea M, Silva JC, et al. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. Nature . 2002 Oct 3;419(6906):512–9.

Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao Q, et al. Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. Science . 2007 Jan 12;315(5809):207–12.

Cavalier-Smith T. Kingdom protozoa and its 18 phyla. . Microbiol. Rev. 1993. p. 953–94.

Cavalier-Smith T. Predation and eukaryote cell origins: a coevolutionary perspective. Int J Biochem Cell Biol . 2009 Feb;41(2):307–22.

Cavalier-Smith T. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. Biol Lett . 2010 Jun;6(3):342–5.

Chappuis F, Sundar S, Hailu A, Ghalib H, Rijal S, Peeling RW, et al. Visceral leishmaniasis: what are the needs for diagnosis, treatment and control? Nat Rev Microbiol . Nature Publishing Group; 2007;5(11):873–82.

Choi J, El-Sayed NM. Functional genomics of trypanosomatids. Parasite Immunol . 2012;34(2-3):72–9.

Ciccarelli FDFD, Doerks T, von Mering C, Creevey CJCJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. Science . AAAS; 2006 Mar 3;311(5765):1283–7.

Conte MG, Gaillard S, Droc G, Perin C. Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants. BMC Genomics . 2008 Jan;9:183.

Croan DG, Morrison D a, Ellis JT. Evolution of the genus Leishmania revealed by comparison of DNA and RNA polymerase gene sequences. Mol Biochem Parasitol . 1997 Nov;89(2):149–59.

Cupolilo SMN, Souza CSF, Abreu-Silva AL, Calabrese KS, Goncalves da Costa SC. Biological behavior of Leishmania (L.) amazonensis isolated from a human diffuse cutaneous leishmaniasis in inbred strains of mice. Histol Histopathol . 2003 Oct;18(4):1059–65.

d'Avila-Levy CM, Marinho FA, Santos LO, Martins JL, Santos ALS, Branquinha MH. Antileishmanial activity of MDL 28170, a potent calpain inhibitor. Int J Antimicrob Agents . 2006 Aug;28(2):138–42.

Daubin V, Gouy M, Perrière G. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res . 2002 Jul;12(7):1080–90.

Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet . 2005 May;6(5):361–75.

Deschamps P, Lara E, Marande W, López-García P, Ekelund F, Moreira D. Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids. Mol Biol Evol . 2011 Jan;28(1):53–8.

Desjeux P. Leishmaniasis. Public health aspects and control. Clin Dermatol . 1996;14(5):417–23.

Desjeux P. Leishmaniasis: current situation and new perspectives. Comp Immunol Microbiol Infect Dis . 2004 Sep;27(5):305–18.

Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in Drosophila. Genome Res . 2003 Oct;13(10):2213–9.

Doolittle WF. Phylogenetic classification and the universal tree. Science . 1999 Jun 25;284(5423):2124–9.

Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton JA, et al. Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. Genome Res . 2011 Dec;21(12):2143–56.

Dutilh BE, Huynen MA, Bruno WJ, Snel B. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J Mol Evol . 2004 May;58(5):527–39.

Eisen JA. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res . 1998 Mar;8(3):163–7.

Eisen JA, Fraser CM. Phylogenomics: intersection of evolution and genomics. Science . 2003 Jun 13;300(5626):1706–7.

Eisen JA, Wu M. Phylogenetic analysis and gene functional predictions: phylogenomics in action. Theor Popul Biol . 2002 Jun;61(4):481–7.

Elmore S a, Jones JL, Conrad P a, Patton S, Lindsay DS, Dubey JP. Toxoplasma gondii: epidemiology, feline clinical aspects, and prevention. Trends Parasitol . Elsevier Ltd; 2010 Apr;26(4):190–6.

El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran A-N, et al. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. Science . 2005 a Jul 15;309(5733):409–15.

El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, et al. Comparative genomics of trypanosomatid parasitic protozoa. Science . 2005 b Jul 15;309(5733):404–9.

Eresh S, McCallum SM, Barker DC. Identification and diagnosis of Leishmania mexicana complex isolates by polymerase chain reaction. Parasitology . 1994 Nov;109 (Pt 4:423–33.

Ersfeld K, Barraclough H, Gull K. Evolutionary Relationships and Protein Domain Architecture in an Expanded Calpain Superfamily in Kinetoplastid Parasites. J Mol Evol. 2005;742–57.

Fraga J, Montalvo AM, Van der Auwera G, Maes I, Dujardin J-C, Requena JM. Evolution and species discrimination according to the Leishmania heat-shock protein 20 gene. Infect Genet Evol . 2013 Aug;18:229–37.

Gardner MJ, Bishop R, Shah T, de Villiers EP, Carlton JM, Hall N, et al. Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes. Science . 2005 Jul 1;309(5731):134–7.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature . 2002 Oct;419(6906):498–511.

Graham WV, Tcheng DK, Shirk AL, Attene-Ramos MS, Welge ME, Gaskins HR. Phylomat: an automated protein motif analysis tool for phylogenomics. J Proteome Res . 2004;3(6):1289–91.

Grimaldi G, McMahan-Pratt D. Leishmaniasis and its etiologic agents in the New World: an overview. Prog Clin Parasitol . 1991 Jan;2:73–118.

Grimaldi G, Tesh RB. Leishmaniases of the New World: current concepts and implications for future research. Clin Microbiol Rev . 1993 Jul;6(3):230–50.

Gu Z, Rifkin SA, White KP, Li W-H. Duplicate genes increase gene expression diversity within and between species. Nat Genet . 2004 Jun;36(6):577–9.

Hall N, Carlton J. Comparative genomics of malaria parasites. Curr Opin Genet Dev . 2005 Dec;15(6):609–13.

Hall N, Karras M, Raine JD, Carlton JM, Kooij TW a, Berriman M, et al. A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. Science . 2005 Jan 7;307(5706):82–6.

Hardison RC. Comparative genomics. PLoS Biol . 2003 Nov;1(2):E58.

Imam T. The complexities in the classification of protozoa: a challenge to parasitologists. Bayero J Pure Appl Sci . 2009 Feb 23;2(2):159–64.

Ivens AC, Peacock CS, Worthey E a, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, Leishmania major. Science . 2005 Jul 15;309(5733):436–42.

Jackson AP. The evolution of amastin surface glycoproteins in trypanosomatid parasites. Mol Biol Evol . 2010 Jan;27(1):33–45.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. Phylogenomics: the beginning of incongruence? Trends Genet . 2006 Apr;22(4):225–31.

Koonin E V. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet . 2005 Jan;39:309–38.

Koonin E V, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol . 2004;5(2):R7.

Koonin E V, Galperin MY. Sequence - Evolution - Function : computational approaches in comparative genomics . Boston, Dordrecht, London: Kluwer academic publishers; 2002.

Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res . 2003 Sep;13(9):2178–89.

Liu W, Fang L, Li M, Li S, Guo S, Luo R, et al. Comparative Genomics of Mycoplasma: Analysis of Conserved Essential Genes and Diversity of the Pan-Genome. Fairhead C, editor. PLoS One . 2012 Apr 20;7(4):e35698.

Loftus B, Anderson I, Davies R, Alsmark UCM, Samuelson J, Amedeo P, et al. The genome of the protist parasite Entamoeba histolytica. Nature . 2005 Feb 24;433(7028):865–8.

Lorenzi H a, Puiu D, Miller JR, Brinkac LM, Amedeo P, Hall N, et al. New assembly, reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features and protein content information. PLoS Negl Trop Dis . 2010 Jan;4(6):e716.

Lukes J, Mauricio IL, Schönian G, Dujardin J, Soteriadou K, Dedet J, et al. Evolutionary and geographical history of the Leishmania donovani complex with a revision of current taxonomy. Proc Natl Acad Sci U S A . 2007;104(22):9375–80.

Makarova KS, Sorokin A V, Novichkov PS, Wolf YI, Koonin E V. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. Biol Direct . 2007 Jan;2:33.

Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG, Koonin E V. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. Nucleic Acids Res . 2005;33(14):4626–38.

Marsden PD. Mucosal leishmaniasis (“espundia” Escomel, 1911). Trans R Soc Trop Med Hyg . 1986 Jan;80(6):859–76.

Mauricio IL, Gaunt MW, Stothard JR, Miles MA. Glycoprotein 63 (gp63) genes show gene conversion and reveal the evolution of Old World Leishmania. Int J Parasitol . 2007 Apr;37(5):565–76.

Mauricio IL, Howard MK, Stothard JR, Miles MA. Genomic diversity in the Leishmania donovani complex. Parasitology . 1999 Sep;119 (Pt 3:237–46.

Miles MA, Llewellyn MS, Lewis MD, Yeo M, Baleela R, Fitzpatrick S, et al. The molecular epidemiology and phylogeography of Trypanosoma cruzi and parallel research on Leishmania: looking back and to the future. Parasitology . 2009;136(12):1509–28.

Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, et al. Genomic minimalism in the early diverging intestinal parasite Giardia lamblia. Science . 2007 Sep 28;317(5846):1921–6.

Mottram JC, Coombs GH, Alexander J. Cysteine peptidases as virulence factors of Leishmania. Curr Opin Microbiol . 2004 Aug;7(4):375–81.

Murray HW, Berman JD, Davies CR, Saravia NG. Advances in leishmaniasis. Lancet . 2005;366(9496):1561–77.

Nielsen R. Comparative genomics: difference of expression. Nature . 2006 Mar 9;440(7081):161.

Pain a, Böhme U, Berry a E, Mungall K, Finn RD, Jackson a P, et al. The genome of the simian and human malaria parasite Plasmodium knowlesi. Nature . 2008 Oct 9;455(7214):799–803.

Pain A, Renauld H, Berriman M, Murphy L, Yeats C a, Weir W, et al. Genome of the host-cell transforming parasite Theileria annulata compared with T. parva. Science . 2005 Jul 1;309(5731):131–3.

Passos VM, Fernandes O, Lacerda PA, Volpini AC, Gontijo CM, Degrave W, et al. Leishmania (Viannia) braziliensis is the predominant species infecting patients with American cutaneous leishmaniasis in the State of Minas Gerais, Southeast Brazil. Acta Trop . 1999 Apr 30;72(3):251–8.

Pays E, Vanhamme L, Pérez-Morga D. Antigenic variation in Trypanosoma brucei: facts, challenges and mysteries. Curr Opin Microbiol . 2004 Aug;7(4):369–74.

Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, et al. Comparative genomic analysis of three Leishmania species that cause diverse human disease. Nat Genet . 2007 Jul;39(7):839–47.

Rafati S, Hassani N, Taslimi Y, Movassagh H, Rochette A, Papadopoulou B. Amastin peptide-binding antibodies as biomarkers of active human visceral leishmaniasis. Clin Vaccine Immunol . 2006 Oct;13(10):1104–10.

Real F, Vidal RO, Carazzolle MF, Mondego JMC, Costa GGL, Herai RH, et al. The Genome Sequence of Leishmania (Leishmania) amazonensis: Functional Annotation and Extended Analysis of Gene Models. DNA Res . 2013 Jul 15;

Rochette A, McNicoll F, Girard J, Breton M, Leblanc E, Bergeron MG, et al. Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in Leishmania spp. Mol Biochem Parasitol . 2005 Apr;140(2):205–20.

Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, Depledge DP, et al. Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. Genome Res . 2011 Dec;21(12):2129–42.

Dos Santos F, Lyra M, Alves L, da Silva K, Rolim L, Gomes T, et al. Pesquisa , desenvolvimento e inovação para o controle das doenças negligenciadas. Rev Ciências Farm Básica e Apl. 2012;33(1):37–47.

Sato S. The apicomplexan plastid and its evolution. Cell Mol Life Sci . 2011 Apr;68(8):1285–96.

Sharma U, Singh S. Insect vectors of Leishmania : distribution , physiology and their control. 2008;(December):255–72.

Simpson AGB, Gill EE, Callahan HA, Litaker RW, Roger AJ. Early Evolution within Kinetoplastids (Euglenozoa), and the Late Emergence of Trypanosomatids. Small. 2004;155(December):407–22.

Simpson AGB, Stevens JR, Lukes J. The evolution and diversity of kinetoplastid flagellates. Trends Parasitol . 2006 Apr;22(4):168–74.

Sobel JD, Nagappan V, Nyirjesy P. Metronidazole-resistant vaginal trichomoniasis--an emerging problem. N Engl J Med . 1999 Jul 22;341(4):292–3.

Sonnhammer ELL, Koonin E V. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet . 2002 Dec;18(12):619–20.

Stiles JK, Hicock PI, Shah PH, Meade JC. Genomic organization, transcription, splicing and gene regulation in Leishmania. Ann Trop Med Parasitol . 1999 Dec;93(8):781–807.

Subileau M, Barnabé C, Douzery EJP, Diosque P, Tibayrenc M. Trypanosoma cruzi: new insights on ecophylogeny and hybridization by multigene sequencing of three nuclear and one maxicircle genes. Exp Parasitol . 2009;122(4):328–37.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin E V, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics . 2003 Sep 11;4:41.

Tatusov RL, Galperin MY, Natale D a, Koonin E V. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res . 2000 Jan;28(1):33–6.

Tatusov RL, Koonin E V, Lipman DJ. A Genomic Perspective on Protein Families. Science (80-) . 1997 Oct 24;278(5338):631–7.

Tatusov RL, Natale DA, Garkavtsev I V, Tatusova TA, Shankavaram UT, Rao BS, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res . 2001;29(1):22–8.

TDR. Leishmaniasis . 2013 [cited 2013 Apr 11]. Available from: http://www.who.int/tdr/diseases-topics/Leishmaniasis/en/index.html

Telford MJ. Phylogenomics. Curr Biol . 2007 Nov 20;17(22):R945–6.

Thompson RCA. The Impact of Giardia on Science and Society. In: Ortega-Pierres, Guadalupe; Cacciò, Simone M; Fayer, Ronald; Mank, Theo G; Smith, Huw V; Thompson R, editor. GIARDIA CRYPTOSPORIDIUM From Mol to Dis. CAB International, 2009; 2009. p. 1–11.

Torres DC, Adaui V, Ribeiro-Alves M, Romero GAS, Arévalo J, Cupolillo E, et al. Targeted gene expression profiling in Leishmania braziliensis and Leishmania guyanensis parasites isolated from Brazilian patients with different antimonial treatment outcomes. Infect Genet Evol . 2010 Aug;10(6):727–33.

Vallejo GA, Guhl F, Carranza JC, Lozano LE, Sánchez JL, Jaramillo JC, et al. kDNA markers define two major Trypanosoma rangeli lineages in Latin-America. Acta Trop . 2002 Jan;81(1):77–82.

Weedall GD, Hall N. Evolutionary genomics of Entamoeba. Res Microbiol . Elsevier Masson SAS; 2011 Feb;1–9.

Weigle K, Saravia NG. Natural history, clinical evolution, and the host-parasite interaction in New World cutaneous Leishmaniasis. Clin Dermatol . 1996;14(5):433–50.

Weiss LM, Dubey JP. Toxoplasmosis: A history of clinical observations. Int J Parasitol . 2009 Jul 1;39(8):895–901.

WHO. World Health Statistics . 2010 [cited 2013 Apr 8]. Available from: http://www.who.int/whosis/whostat/2010/en/

Widmer G, Comeau AM, Furlong DB, Wirth DF, Patterson JL. Characterization of a RNA virus from the parasite Leishmania. Proc Natl Acad Sci U S A . 1989 Aug;86(15):5979–82.

Widmer, Giovanni; London, Eric; Zhang, Linghui; Ge GT, Saul; Carlton, Jane; da Silva J. Preliminary Analysis of the Cryptosporidium muris Genome. In: Ortega-Pierres, Guadalupe; Cacciò, Simone M; Fayer, Ronald; Mank, Theo G; Smith, Huw V; Thompson RCA, editor. GIARDIA CRYPTOSPORIDIUM From Mol to Dis. CAB International; 2009. p. 320–7.

Wincker P, Ravel C, Blaineau C, Pages M, Jauffret Y, Dedet JP, et al. The Leishmania genome comprises 36 chromosomes conserved across widely divergent human pathogenic species. Nucleic Acids Res . 1996 May 1;24(9):1688–94.

Winters DJ. Expanding global research and development for neglected diseases. Bull World Health Organ . 2006 May;84(5):414–6.

Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. Genome Biol . 2008 Jan;9(10):R151.

Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, et al. The genome of Cryptosporidium hominis. Science (80-). 2004;431(October):1107–12.

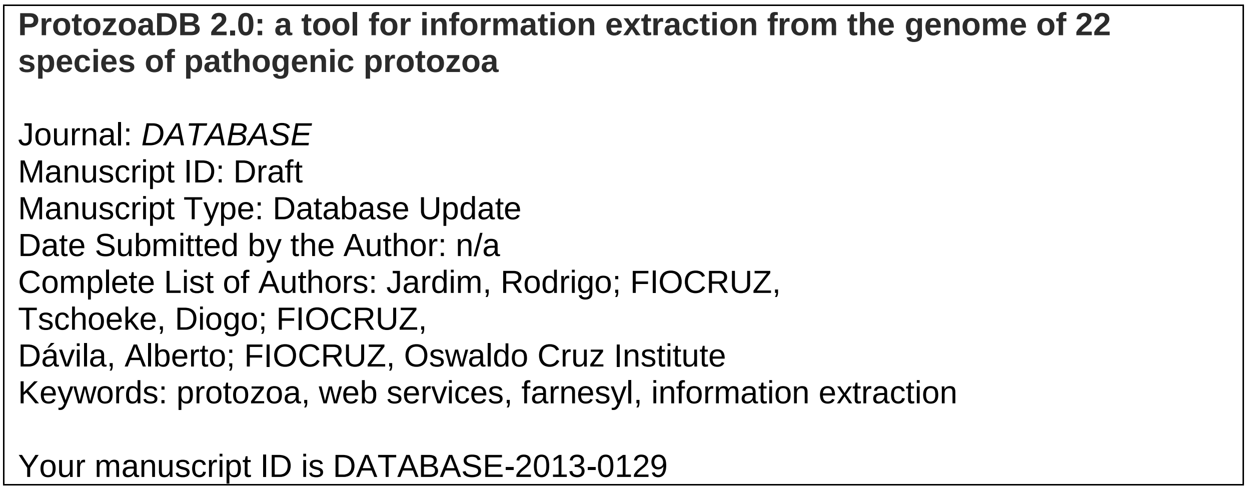
Yamey G, Torreele E. The world’s most neglected diseases. BMJ . 2002 Jul 27;325(7357):176–7.

9: Anexos

9.1- Artigo 4: ProtozoaDB 2.0: uma ferramenta para a extração de informações a partir do genoma de 22 espécies de protozoários patogênicos.

Jardim R, **Tschoeke DA**, Davila AMR. **ProtozoaDB 2.0: a tool for information extraction from the genome of 22 species of pathogenic protozoa.**

Este artigo foi submetido para a revista *DATABASE Oxford*.



ProtozoaDB 2.0: a tool for information extraction from the genome of 22 species of pathogenic protozoa

Journal: *DATABASE*
Manuscript ID: Draft
Manuscript Type: Database Update
Date Submitted by the Author: n/a
Complete List of Authors: Jardim, Rodrigo; FIOCRUZ,
Tschoeke, Diogo; FIOCRUZ,
Dávila, Alberto; FIOCRUZ, Oswaldo Cruz Institute
Keywords: protozoa, web services, farnesyl, information extraction

Your manuscript ID is DATABASE-2013-0129

No presente artigo apresentamos o banco de dados ProtozoaDB 2.0, responsável pelo armazenamento das sequências utilizadas nas análises de homologia dos 22 protozoários (cap. 3 e 4) e guarda dos resultados gerados por estas análises, permitindo a visualização dos mesmo. Finalmente, o ProtozoaDB 2.0 promove uma integração dos dados obtidos nestas análises de homologia com outras fontes de dados como o PDB, KEGG e Superfamily.

ProtozoaDB 2.0: a tool for information extraction from the genome of 22 species of pathogenic protozoa

Rodrigo Jardim1, Diogo A Tschoeke1 and Alberto MR Davila1,2

1 Computacional and Systems Biology Laboratory, Oswaldo Cruz Institute, Fiocruz., Av. Brasil, Rio de Janeiro, Brazil

Full list of author information is available at the end of the article
*Correspondence: davila@fiocruz.br

Content

Background

New species of Protozoa were recently sequenced and deposited in GenBank [1]. While the availability of primary genome sequence is a good starting point for the community to contribute with further analysis, including identification and functional annotation of coding sequences as well as comparative genomics analysis, in order to infer new informations on the biology of these organisms. Analyzing large amount of data generated by genomics experiments, especially using Next Generation Sequencing (NGS) is not a trivial task, mainly considering that researchers were used to deal or focus their studies in a few genes or gene families or even small genomes. The ongoing improvement of NGS technologies makes the sequencing of more and more eukaryotic genomes a reality, then yielding to new paradigms, either for the development and improvement of semi-automatic analysis/annotation systems for this huge ammount of data or even for an object-view concept where raw reads are the main, fixed object, and assemblies with their annotations take a role of dinamically changing and modificable views of the object [2].

Similar to the life cycle of software, the processes involved in sequencing and preparation of genomic information could be represented in the same way (Figure 1). The first step is data acquisition that can be performed by several steps such as: (i) download from public databases, (ii) sequencing across multiple platforms, like Sanger and/or NGS (Illumina, 454, Ion Torrent and/or Pacific BioScience). The second step, called preprocessing, formats and stores genomic data for subsequent use. The third step refers to the use of a number of computational tools to transform raw data on information and knowledge. The fourth and last step is the distribution and availability of all this information to the community for further analysis and inferences.

Therefore, in order to facilitate the information extraction, we developed the ProtozoaDB [3] database system, that in its first version included five genomes of Protozoa (*Entamoeba histolytica*, *Leishmania major*, *Plasmodium falciparum*, *Trypanosoma cruzi* and *T. brucei*) and a set of tools for searching and analyzing data, including phylogeny inference. In the present study we present a new version of ProtozoaDB, called from now ProtozoaDB 2.0 (http://protozoadb.biowebdb.org), that as above described would fit into the third and final steps of the bioinformatics cycle: transforming row data into information then distribution and availability. The development of new generation databases as ProtozoaDB is being encouraging by the community, especially in the context of the BioCreative initiative (http://www.biocreative.org/) and reviewed by [4].

The system has been fully remodelled to allow for new tools and a more expanded view of data, using advanced computational techniques and providing a wider range of information for users. Now with the genomes of 22 pathogenic Protozoa, this new version includes a number of analysis such as: (i) similarities against other bases (*Homo sapiens*, model organisms, Conserved Domains Database - CDD and Protein Data Bank - PDB); (ii) visualization of metabolic pathways of Kyoto Encyclopedia of Genes and Genomes - KEGG [5]; (iii) protein structures by PDB [6]; (iv) homology studies, using results from OrthoMCL [7], KEGG Orthology (KO) and OrthoSearch [8]; (v) search for related publications at PubMed; (vi) Superfamily classification [9]; and (vii) phenotype inferences based on comparisons with model organisms, particularly with Saccharomyces cerevisae.

The source code of ProtozoaDB 2.0 was completely rewritten in another programming language and with more elaborate techniques. It uses now a framework for developing Web applications known as Rails (http://rubyonrails.org/). It was developed in layers allowing a separation of the business object code of the pages displayed to users, making maintenance easier and consequently access to its pages lighter and faster. Furthermore, there is a specific layer to treat with data to be fetched from other sources. The Ruby language, suitable for the use of the Rails, was adopted for this version together with BioRuby library [10], enabling the developing of pages with less code and better reuse of functions. ProtozoaDB 2.0 was also implemented using concepts of Object Orientation and Design Patterns. This made the application lighter, safer and simpler to maintain. The use of the JQuery library made possible the web pages to work with Asynchronous Javascript and XML (AJAX), creating a more friendly user interface. Now it is possible to view on one page all the information provided by ProtozoaDB 2.0.

The new system uses the concept of Web Services to access all internal and external databases. Thus, the application focuses only on usability and user-friendly information. All bases are queried simultaneously allowing a response time considered to be satisfactory for the application. The ProtozoaDB 2.0 allows queries by several methods, including: Genbank Identifier (GI), Accession Number, Description, Blast, Motif and Phenotype.

With the information all in one place, it is possible to understand the biology and biological systems of the organisms studied, in particular protozoa. Added to this, the ProtozoaDB 2.0 has information inferred from phenotypes. According to the literature orthology helps to transfer phenotype information based on genotypes. According to [21] it is also possible to transfer functional information based on similar phenotypes and a specialized database called PhenomicDB was developed using this concept [22].

Implementation

Web Services

ProtozoaDB 2.0 supports RESTful (REpresentational State Transfer) [11] Web Services to make data access easier. Available in http://services.biowebdb.org/howtouse, these services were written in Ruby language using Ruby on Rails (RoR), allowing accessing the information about proteome of Protozoa available in ProtozoaDB 2.0 web application.

New source code

The source code was rewritten in Ruby using Ruby on Rails, which allowed the de velopment of three layers: view, with web pages based on Asynchronous JavaScript and XML (AJAX); model with search algorithms in remote web services, and the controller, which is an interface between the view and model.

Data acquisition

Primary dataset of genome and proteome of 22 Protozoa species (Table 1), including families Babesiidae, Cryptosporidiidae, Entamoebidae, Hexamitidae, Plasmodium, Theileriidae, Trichomonadidae and Trypanosomatidae were downloaded from Genbank [1] in Genbank Flat File (GBFF) format.

Preprocessing

Primary data were stored locally in a server with a Database Manage Systems (DBMS) PostgreSQL version 8.4 (http://www.postgres.org), through the Genomics Unified Schema (GUS) version 3.5 [12] framework.

Transformation

Some analyses were performed to incorporate new informations to the primary dataset. Inferences about homologies were performed on Protozoa data and the results locally stored, namely: (i) orthology inference using OrthoMCL [7] and (ii) OrthoSearch [8] which uses Hidden Markov Model (HMM) through HMMER (version 3.0) and best reciprocal hits; and (iii) BLAST-based similarity results.

Phenotypes were retrieved from the Saccharomyces Database [13] and stored locally. Mappings between (i) KEGG Orthology (KO), (ii) Saccharomyces cerevisae proteins and (iii) Protozoa proteins, were performed and stored locally, aiming to infer phenotypes in Protozoa.

Analysis

Web services technologies are used to access several databases worldwide to complement existing information (Table 2), among them:

1. Similarity against the human proteome: the system performs a query, through web service, the human proteome is locally stored in the database, updated every six months, and returns the top ten hits, independent of the score or e-value.
2. Search for conserved domains: by running RPSBlast against Conserved Domain Database (CDD) [14], the system returns the top ten results independent of the score or e-value.
3. Superfamily classification: the Superfamily database [9, 15] has structural, functional and evolution of proteins from different genomes, including Protozoa. The new system performs a query through web service retrieving graphical information on the classification of superfamily.

4. Similar protein structure: the system retrieves information from 2D and 3D similarity by performing BLAST [16] and FASTA [17] directly in Protein Data Bank (PDB) [18, 19].

5. Metabolic pathways: the system performs a query to KEGG Pathway database [5] to retrieve the metabolic pathways where a given protein participates, showing the maps and their interactions with other proteins participanting of that pathway.

6. Publications: finally, the system performs two queries in Pubmed (http://www.ncbi.nlm.nih.gov/pubmed) to retrieve the original publication of protein and other publications having relevance with organism and product.

Results

Protozoa genomic data

The ProtozoaDB 2.0 provides descriptive, quantitative, qualitative and comparative information on genomes and proteomes of 22 protozoa (Table 1), thus allowing a more detailed analysis of each organism in addition to the relationships between them. The new version contains 193,559 genes, 218,100 proteins, 26,101 homologous groups of 22 pathogens (21,119 orthologous groups and 4,982 paralogous groups) obtained by OrthoMCL analysis (Figure 2) and 195 phenotypes infered by across information on Saccharomyces Database.

Proteome

Information about the proteins of 22 different Protozoa are complemented by results obtained by real-time queries, performed in several remote databases through the use of Web Services. Two similarity analyses are performed using PDB [6]: (i) BLAST [16] and FASTA [17]. The FASTA results also allows a visual comparison of the 3D structures of similar proteins, while the BLAST results were improved by us allowing the users to select any or all hits retrieved and export them to fasta format (Figure 3). (ii) Furthermore, conserved domains analysis using CDD [20] and similarity against the human proteome are also allowed, this information is displayed showing the top 10 results (Figure 4).

Homology

Preliminary analysis of homology between the 22 Protozoa were performed and results are available for queries. The orthologous groups were inferred using the methodology implemented in OrthoMCL (Figure 5) and OrthoSearch, using either Blast-based and Hmmer-based algorithm, respectively.

Metabolic Pathways

The system perform a webservices-based query to retrieve metabolic maps available on KEGG, showing the involvement of given protein in that pathway (Figure 6).

Phenotypes

ProtozoaDB 2.0 allows webservices-based queries through the phenotypes mapped from Saccharomyces Database [13], retrieving proteins from 22 Protozoa that could potentially provide such features. This information was made possible by mapping the proteome of the 22 species with information from the KEGG orthologous groups (Kegg Orthology - KO) as part of the “transformation” step described in the “construction and content” (Figure 7).

How to search

The new system retrieves the information through various search engines. Based on the previous version, the system will search for the description of the protein or part of the description, Accession number, Genbank Identifier (GI) and organism name. In addition to these mechanisms, th new version also allows quer by phenotype and by **Blast**.

How to search using web service

Addition we available also a set of web services functions to retrieve information available in ProtozoaDB 2.0 web application. In http://services.biowebdb.org/howtouse are the information about how to use available services with source code examples. Functions to search some protozoa protein by Accession Number, Genbank Identifier, description (annotation), organism, phenotype and blast, well as details of located protein like orthologous groups, similarities results, KEGG pathways and phenotypes are available from queries with web services.

Information extraction case study - Find potentials drug targets

To demonstrate the usefulness of ProtozoaDB 2.0 for information extraction, a study using phenotypes in Kinetoplastid species was conducted. Through the search field system the option Phenotypes was choosen and the keyword ”inviable” used with Kinetoplastid database. This phenotype may indicate, depending on the experiment, a situation of impossibility of survival of the organism [13]. The system returns a list containing a wide range of proteins that have this phenotype based on orthology with Saccharomyces cerivisae. From the obtained list, the first hit meeting the following requirements was choosen: (i) low similarity against human proteome; (ii) high similarity against bacterial species; and (iii) a pathway available in KEGG. The choosen hit was XP_844041.1 corresponding to the farnesyltransferase (PFT) alpha subunit preotein, from Trypanosoma brucei, showing a high similarity against bacterial prenyltransferase group (Figure 8).

Comparisson with other information extraction tool

We performed a comparative analysis with EuPathDB (Aurrecoechea et al., 2010) to evaluate the similarities and differences between these two information extraction tools (Table 3). We observed that both tools offers similar functionalities, such as: similarity search against Protozoa species, homology search using OrthoMCL, and search by phenotype and product name. EuPathDB allow users to perform search by SNP characteristics, genomic position, cellular location and transcript expression, not offered by ProtozoaDB 2.0. On the other hand, ProtozoaDB 2.0 allows users to perform searches in PDB Data Bank, retrieving PDB fasta sequences and 3D PDB structures, search articles related to sequences in the PubMed, homology search using OrthoMCL and OrthoSearch methodologies; and pathways from KEGG. ProtozoaDB 2.0 also provides SuperFamily-based analisys and similarities searches using BLAST against CDD database, PDB database, model organisms and Homo sapiens proteome.

Discussion

In the previous version of ProtozoaDB only five pathogenic protozoa were available together with some analyses tools. ProtozoaDB 2.0 has additional 17 Protozoa species, comprising 22 genomes and proteomes. New analysis tools were added, such as: homology analysis among the 22 organisms, using two different approaches; and, phenotypes inference through orthology with a model organism. Furthemore, to allow a more comprehensive information about these organisms, several real time queries are allowed, retrieving information about the proteome of organisms.

The use of webservices implemented allows for a flexible and unique system capable of: (i) integrating a range of related information, (ii) allowing direct access to information in their original sources, and (iii) avoiding to locally store data from third parties and its periodic update. These advantages allow the system to be always updated, since most of the information are queried directly in sources databases and in real time.

Using a AJAX-based framework enables the ProtozoaDB 2.0 performs all queries through web services and simultaneously taking the response time queries quite suitable for online.

The new search engines, particularly through BLAST, allows researchers to query the base of ProtozoaDB 2.0 directly by protein or gene of interest, viewing several aggregate information.

Thus, it was possible to found a potencial drug target just browsing through the system and using the information provided.

Case study

Farnesyltransferase is one enzyme of the prenyltransferase group, which attach a 15-carbon isoprenoid farnesyl group to proteins with CAAX motif: a four-amino acid sequence at the carboxyl terminus of a protein [23]. Farnesylation is a type of prenylation, a post-translational modification of proteins which binds a isoprenyl group (15-carbon isoprenoid) to a cysteine residue. In other words, protein farnesylation involves protein farnesyltransferase (PFT)-that catalyze a attachment of the farnesyl group from farnesyl pyrophosphate (FPP) to cysteine SH of the C-terminal sequence motif CAAX, where C is cysteine usually but not always an aliphatic residue. The terminal amino acid is determinant of farnesylation because FTase is preferentially active on protein substrates with CAAX [18]. This is an important process to mediate protein-protein interactions and protein-membrane interactions [23, 24]. Prenylation (farnesylation) and subsequent modifications are essential for correct membrane targeting and cellular functioning of a number of proteins in eukaryotic cells such as Ras superfamily GTPases [25]. The enzyme farnesyltransferase is heterodimeric and has two subunits: alpha (α) and beta (β) subunit. The α subunit is a double layer paired alpha helices pilled up in parallel, which enfold partly the beta subunit like a mantle. The helices of the β subunit form a barrel and the active site is created by the center of the β subunit, flanked by part of the α subunit. Junction of the α - and β -subunits, bind the CaaX protein, and a hydrophobic cleft within the α and β barrel structure of the beta-subunit, which bind the farnesyl diphosphate substrate [26]. In other words, PFT drives a zinc cation on its β subunit at the edge of the active site, and has a hydrophobic binding pocket for farnesyl diphosphate, the lipid donor molecule. Besides, the PFT substrates have a cysteine as their fourth-to-last residue [25]. The condensation of two molecules of farnesyl diphosphate produce squalene which are involved in sterol biosynthesis [27]. This PFT and farnesylation are a potentially good drug target for trypanosomatids [24] because inhibitors have potent activity against cultured forms and are less toxic to mammalian cells than parasite cells and in T. brucei substrates specificities and inhibitor selectivity are distinct from mammalian.

Comparison between ProtozoaDB 2.0 and EupathDB

Both tools for information extraction evaluated have several features that allow a wider analysis on the organism studied. The EuPathDB allows a more comprehensive view of the characteristics of the protein investigated, whereas the ProtozoaDB 2.0 focuses its analysis to infer and/or confirm the functional annotation of a given protein, based on its primary annotation deposited in Genbank. Furthermore, ProtozoaDB 2.0 also allows a view of the biological role played by the protein in biological systems, including related literature. Through ProtozoaDB 2.0 it is possible to re-annotate some of the proteins identified as "hypothetical" through similarity-based programs as well as SuperFamily-based classification. Finally, using the tools provided by ProtozoaDB 2.0 it is also possible to infer potential drug targets, as described in the Case Study.

Conclusions

The new version of ProtozoaDB 2.0 allows a more detailed analysis of the object of study than the previous version, as well as expanding the number of genomes and proteomes available to the scientific community. In our case study we found a group of protein prenyltransferase just browsing through the results provided by webservcies-based tools developed for the new system version. This protein is already described in the literature [24] as a good drug target for trypanosomatids for the following reasons: (i) its inhibitors have potent activity against cultured forms of these parasites and these inhibitors are more toxic against parasite cells than mammalian cells; (ii) for *T. brucei* PFT (TbPFT), the substrate specificities and inhibitor selectivity are distinct from mammalian PFT; and (iii) effort in pharmaceutical industry to develop small molecule inhibitors of mammalian PFTs for anti-cancer purposes creates an abundance of compounds that can be screened for selective activity against parasites.

We were able identify this potential drug target using only a "In Silico"-based strategy and the information available in public databases integrated into ProtozoaDB 2.0.

Competing interests

The authors declare that they have no competing interests.

Author’s contributions

All authors contributed equally to this work.

Acknowledgements

CNPq, CAPES, Fiocruz for financial supports.

References

1. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W.: GenBank. Nucleic acids research 41(Database issue), 36{42 (2013)
2. Munoz, J.F., Gallo, J.E., Misas, E., McEwen, J.G., Clay, O.K.: The eukaryotic genome, its reads, and the unfinished assembly. fFEBSg Letters 587(14), 2090{2093 (2013)
3. Daavila, A.M.R., Mendes, P.N., Wagner, G., Tschoeke, D.A., Cuadrat, R.R.C., Liberman, F., Matos, L., Satake, T., Oca~na, K.A.C.S., Triana, O., Cruz, S.M.S., Juc_a, H.C.L., Cury, J.C., Silva, F.N., Geronimo, G.A., Ruiz, M., Ruback, E., Silva, F.P., Probst, C.M., Grisard, E.C., Krieger, M.A., Goldenberg, S., Cavalcanti, M.C.R., Moraes, M.O., Campos, M.L.M., Mattoso, M.: ProtozoaDB: dynamic visualization and exploration of protozoan genomes. Nucleic acids research 36(Database issue), 547{52 (2008)
4. Krallinger, M., Valencia, A., Hirschman, L.: Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol. 9(2) (2008)
5. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M.: KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research 40(Database issue), 109{14 (2012)
6. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Research 28(1), 235{242 (2000)
7. Li, L., Stoeckert, C.J., Roos, D.S., Jr, C.J.S.: OrthoMCL : Identi_cation of Ortholog Groups for Eukaryotic Genomes OrthoMCL : Identi_cation of Ortholog Groups for Eukaryotic Genomes. Genome Research, 2178{2189 (2003)
8. da Cruz, S.M.S., Batista, V., Silva, E., Tosta, F., Vilela, C., Cuadrat, R., Tschoeke, D., D_avila, A.M.R., Campos, M.L.M., Mattoso, M.: Detecting distant homologies on protozoans metabolic pathways using scienti_c workows. International journal of data mining and bioinformatics 4(3), 256{80 (2010)
9. Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., Gough, J.: SUPERFAMILY{sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic acids research 37(Database issue), 380{6 (2009)

10. Goto, N., Prins, P., Nakao, M., Bonnal, R., Aerts, J., Katayama, T.: BioRuby: bioinformatics software for the Ruby programming language. Bioinformatics (Oxford, England) 26(20), 2617{9 (2010)
11. Fielding, R.T.: Architectural styles and the design of network-based software architectures. PhD thesis (2000)
12. Davidson, S.B., Crabtree, J., Brunk, B., Schug, J., Tannen, V., Overton, C., Stoeckert, C.: K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources (2000)
13. Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Karra, K., Krieger, C.J., Miyasato, S.R., Nash, R.S., Park, J., Skrzypek, M.S., Simison, M., Weng, S., Wong, E.D.: Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic acids research 40(Database issue), 700{5 (2012)
14. Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M.K., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., Lu, F., Lu, S., Marchler, G.H., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D., Bryant, S.H.: Cdd: conserved domains and protein three-dimensional structure. Nucleic Acids Research 41(D1), 348{352 (2013)
15. de Lima Morais, D.A., Fang, H., Rackham, O.J.L., Wilson, D., Pethica, R., Chothia, C., Gough, J.:
SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic acids research 39(Database issue), 427{34 (2011)
16. Altschul, S.F., Madden, T.L., Sch□a_er, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25(17), 3389{402 (1997)
17. Pearson, W.R.: Rapid and sensitive sequence comparison with fFASTPg and fFASTAg. Methods in Enzymology, vol. 183, pp. 63{98. Academic Press (1990).
18. Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D., Zardecki, C.: The Protein Data Bank. Acta crystallographica. Section D, Biological crystallography 58(Pt 6 No 1), 899{907 (2002)
19. Berman, H.M., Kleywegt, G.J., Nakamura, H., Markley, J.L.: The Protein Data Bank at 40: reecting on the past to prepare for the future. Structure (London, England : 1993) 20(3), 391{6 (2012)
20. Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Lu, F., Marchler, G.H., Mullokandov, M., Omelchenko, M.V., Robertson, C.L., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D., Zhang, N., Zheng, C., Bryant, S.H.: Cdd: a conserved domain database for the functional annotation of proteins. Nucleic Acids Research 39(suppl 1), 225{229 (2011)
21. Groth, P., Weiss, B., Pohlenz, H.-D., Leser, U.: Mining phenotypes for gene function prediction. BMC Bioinformatics 136(9) (2008)
22. Groth, P., Pavlova, N., Kalev, I., Tonov, S., Georgiev, G., Pohlenz, H.-D., Weiss, B.: Phenomicdb: a new cross-species genotype/phenotype resource. Nucleic Acids Research 35(suppl 1), 696{699 (2007)
23. Ayong, L., DaSilva, T., Mauser, J., Allen, C.M., Chakrabarti, D.: Evidence for prenylation-dependent targeting of a Ykt6 SNARE in Plasmodium falciparum. Molecular and biochemical parasitology 175(2), 162{8 (2011)

24. Buckner, F.S., Eastman, R.T., Nepomuceno-Silva, J.L., Speelmon, E.C., Myler, P.J., Van Voorhis, W.C., Yokoyama, K.: Cloning, heterologous expression, and substrate speci_cities of protein farnesyltransferases from Trypanosoma cruzi and Leishmania major. Molecular and biochemical parasitology 122(2), 181{8 (2002)

25. Brunner, T.B., Hahn, S.M., Gupta, A.K., Muschel, R.J., McKenna, W.G., Bernhard, E.J.: Farnesyltransferase inhibitors: an overview of the results of preclinical and clinical investigations. Cancer research 63(18), 5656{68 (2003)

26. Trueblood, C.E., Boyartchuk, V.L., Rine, J.: Substrate speci_city determinants in the farnesyltransferase beta-subunit. Proceedings of the National Academy of Sciences of the United States of America 94(20), 10774{9 (1997)

27. P_erez-Moreno, G., Sealey-Cardona, M., Rodrigues-Poveda, C., Gelb, M.H., Ruiz-P_erez, L.M., Castillo-Acosta, V., Urbina, J.A., Gonz_alez-Pacanowska, D.: Endogenous sterol biosynthesis is important for mitochondrial function and cell morphology in procyclic forms of Trypanosoma brucei. International journal for parasitology 42(11), 975{89 (2012)

Figures

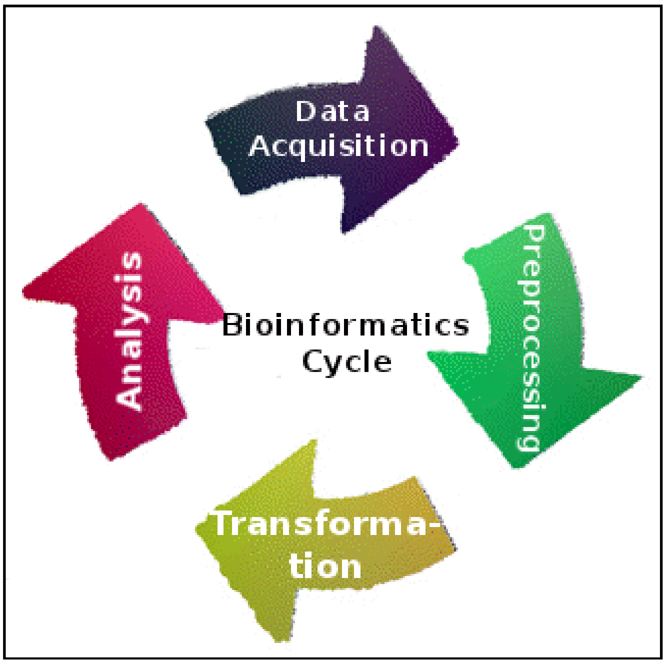
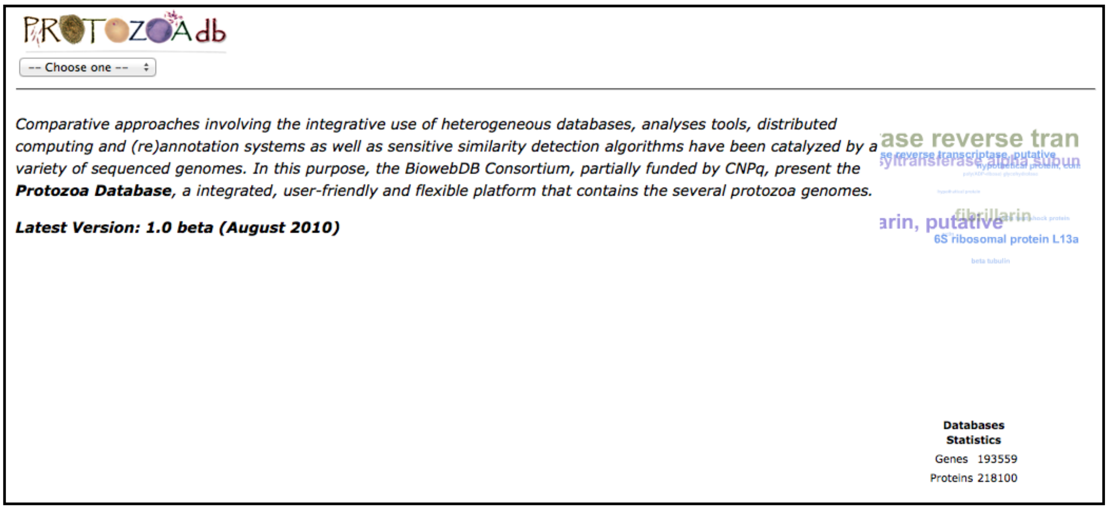


Figure 1: Bioinformatics data lifecycle. Example for data lifecycle in bioinformatics. The lifecycle begins with data acquisition, through the pre-processing data, data transformation and, finally the analysis information generated by all process.


PROTOZOAdb
-- Choose one --
ase reverse tran
se reverse transcriptase, putative
syltransferase alpha subun
hypothetical protein, con
arin, putative
fibrillarin
6S ribosomal protein L13a
beta tubulin

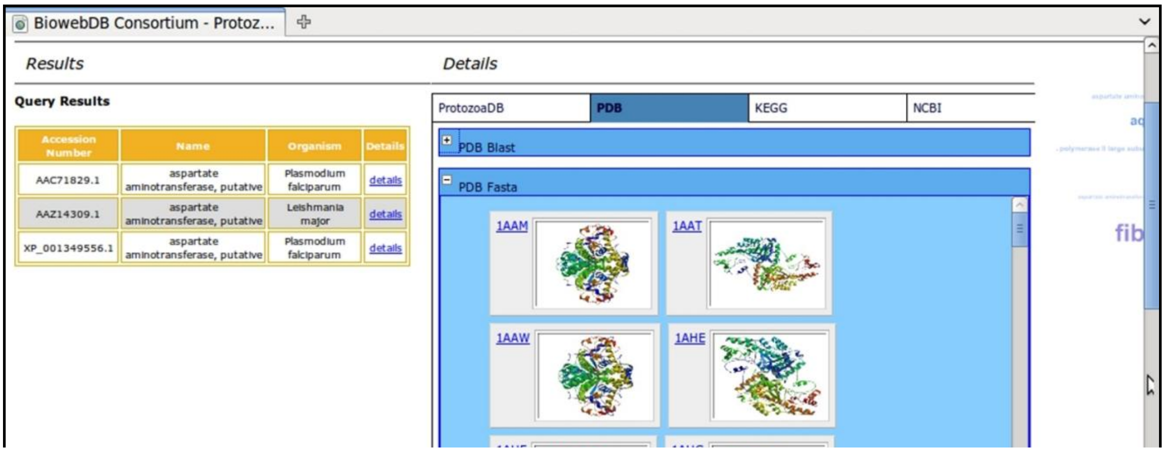
Comparative approaches involving the integrative use of heterogeneous databases, analyses tools, distributed computing and (re)annotation systems as well as sensitive similarity detection algorithms have been catalyzed by a variety of sequenced genomes. In this purpose, the BiowebDB Consortium, partially funded by CNPq, present the ***Protozoa Database****, a integrated, user-friendly and flexible platform that contains the several protozoa genomes.*

Latest Version: 1.0 beta (August 2010)

Databases Statistics

Genes	193559
Proteins	218100

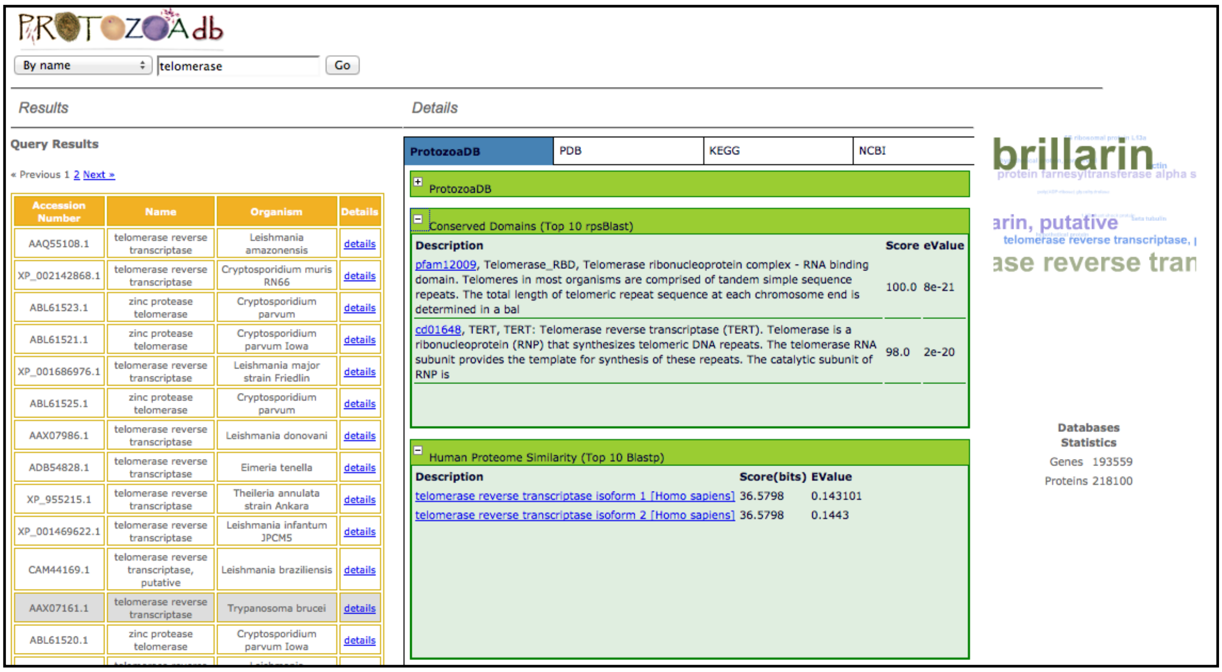
Figure 2: Main page. First page of ProtozoaDB 2.0 web page with database statistics information, field for search and cloud tags.



Query Results

Accession Number	Name	Organism	Details
AAC71829.1	aspartate aminotransferase, putative	Plasmodium falciparum	details
AAZ14309.1	aspartate aminotransferase, putative	Leishmania major	details
XP_001349556.1	aspartate aminotransferase, putative	Plasmodium falciparum	details

Figure 3: PDB information. PDB information shown similarities 2D (blast) and 3D (fasta) against Protein Data Bank (PDB). The figure show similarities 3D to aspartate aminotransferase of *L. major. Clicking on each figure the system show the complete information in original web site.*



Query Results

« Previous 1 2 Next »

Accession Number	Name	Organism	Details
AAQ55108.1	telomerase reverse transcriptase	Leishmania amazonensis	details
XP_002142868.1	telomerase reverse transcriptase	Cryptosporidium muris RN66	details
ABL61523.1	zinc protease telomerase	Cryptosporidium parvum	details
ABL61521.1	zinc protease telomerase	Cryptosporidium parvum Iowa	details
XP_001686976.1	telomerase reverse transcriptase	Leishmania major strain Friedlin	details
ABL61525.1	zinc protease telomerase	Cryptosporidium parvum	details
AAX07986.1	telomerase reverse transcriptase	Leishmania donovani	details
ADB54828.1	telomerase reverse transcriptase	Eimeria tenella	details
XP_955215.1	telomerase reverse transcriptase	Theileria annulata strain Ankara	details
XP_001469622.1	telomerase reverse transcriptase	Leishmania infantum JPCM5	details
CAM44169.1	telomerase reverse transcriptase, putative	Leishmania braziliensis	details
AAX07161.1	telomerase reverse transcriptase	Trypanosoma brucei	details
ABL61520.1	zinc protease telomerase	Cryptosporidium parvum Iowa	details

Conserved Domains (Top 10 rpsBlast)

Description	Score	eValue
pfam12009, Telomerase_RBD, Telomerase ribonucleoprotein complex - RNA binding domain. Telomeres in most organisms are comprised of tandem simple sequence repeats. The total length of telomeric repeat sequence at each chromosome end is determined in a bal	100.0	8e-21
cd01648, TERT, TERT: Telomerase reverse transcriptase (TERT). Telomerase is a ribonucleoprotein (RNP) that synthesizes telomeric DNA repeats. The telomerase RNA subunit provides the template for synthesis of these repeats. The catalytic subunit of RNP is	98.0	2e-20

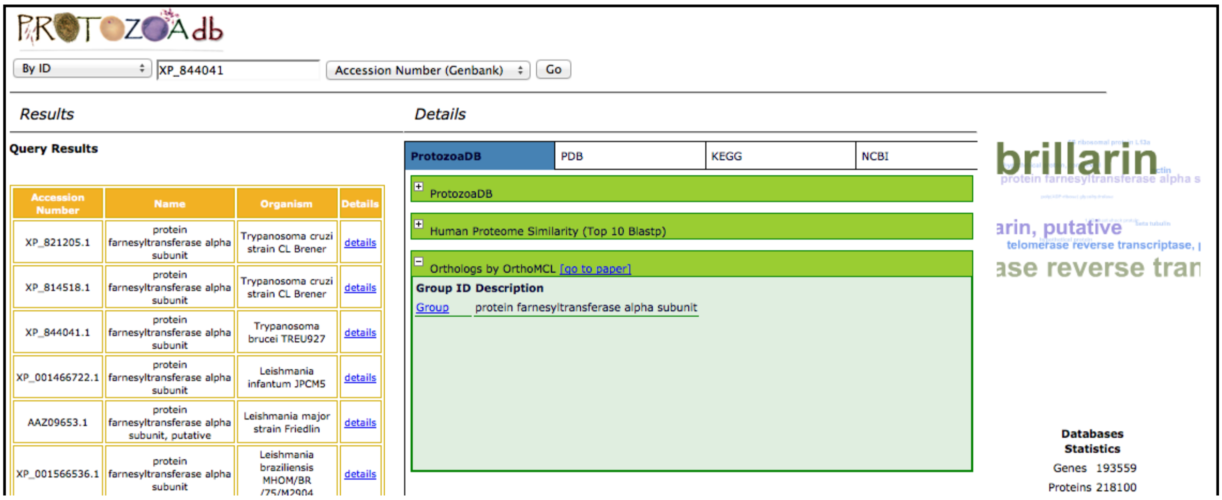
Human Proteome Similarity (Top 10 Blastp)

Description	Score(bits)	EValue
telomerase reverse transcriptase isoform 1 [Homo sapiens]	36.5798	0.143101
telomerase reverse transcriptase isoform 2 [Homo sapiens]	36.5798	0.1443

Databases Statistics

Genes	193559
Proteins	218100

Figure 4: Similarities information. Results for similarities searches against *Homo sapiens* proteome and Conserved Domains Database. Only top ten results are showed. Clicking on links in blue the system open a new window in original web site of information



By ID
Accession Number (Genbank)
Go
Results
Details
ProtozoaDB
PDB
KEGG
NCBI
brillarin
protein farnesyltransferase alpha s
arin, putative
telomerase reverse transcriptase,
ase reverse tran

XP_844041

Query Results

Accession Number	Name	Organism	Details
XP_821205.1	protein farnesyltransferase alpha subunit	Trypanosoma cruzi strain CL Brener	details
XP_814518.1	protein farnesyltransferase alpha subunit	Trypanosoma cruzi strain CL Brener	details
XP_844041.1	protein farnesyltransferase alpha subunit	Trypanosoma brucei TREU927	details
XP_001466722.1	protein farnesyltransferase alpha subunit	Leishmania infantum JPCM5	details
AAZ09653.1	protein farnesyltransferase alpha subunit, putative	Leishmania major strain Friedlin	details
XP_001566536.1	protein farnesyltransferase alpha subunit	Leishmania braziliensis MHOM/BR /75/M2904	details

⊞ ProtozoaDB

⊞ Human Proteome Similarity (Top 10 Blastp)

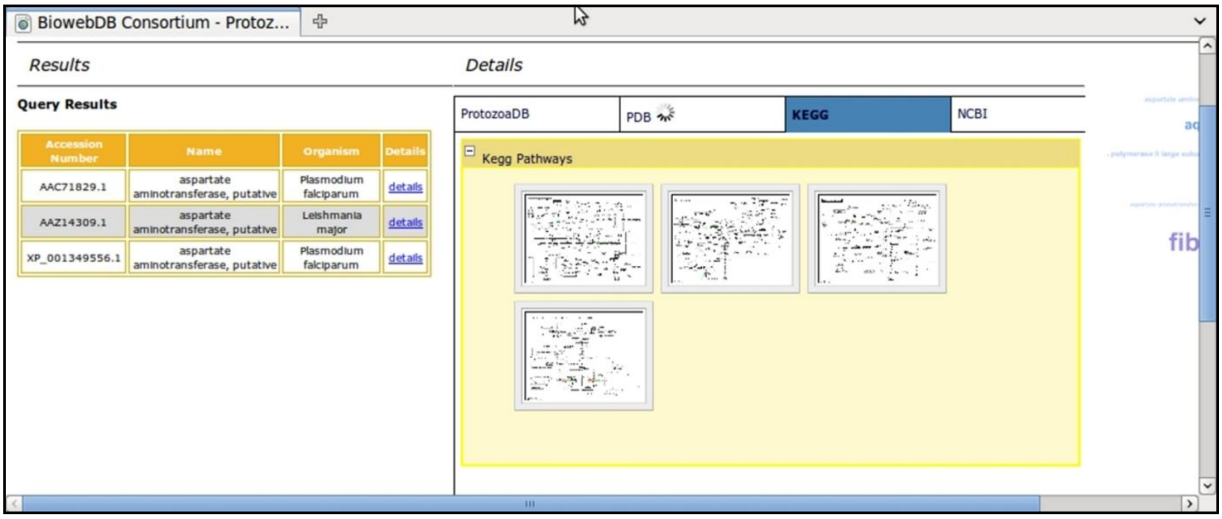
⊟ Orthologs by OrthoMCL [go to paper]

Group ID	Description
Group	protein farnesyltransferase alpha subunit

Databases Statistics

Genes	193559
Proteins	218100

Figure 5: Orthologous information. Group of orthologous using OrthoMCL methodology. Clicking at word group and all proteins of group are shown in left panel (Query Results).

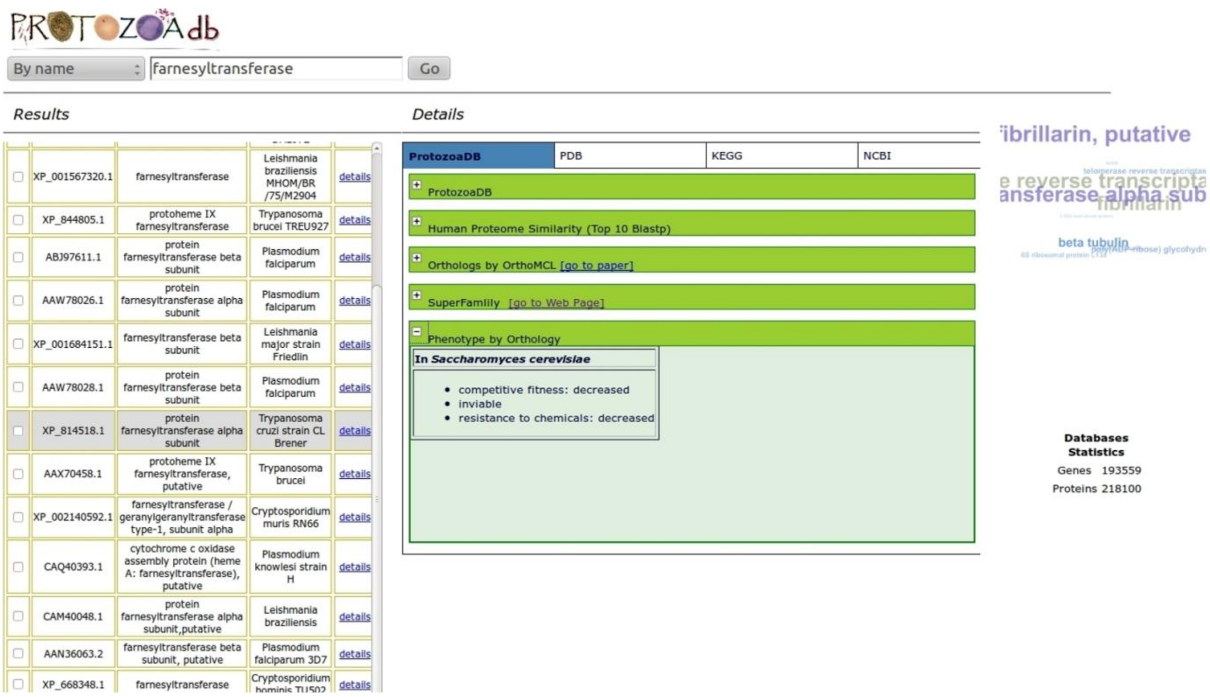


Query Results

Accession Number	Name	Organism	Details
AAC71829.1	aspartate aminotransferase, putative	Plasmodium falciparum	details
AAZ14309.1	aspartate aminotransferase, putative	Leishmania major	details
XP_001349556.1	aspartate aminotransferase, putative	Plasmodium falciparum	details

⊟ Kegg Pathways

Figure 6: Pathways information. Metabolic pathways from KEGG. The figure show all metabolics pathways that include aspartate aminotrasnferase. Clicking in map the system open a new window in original web site of information (KEGG).



By name: farnesyltransferase

Results

□	XP_001567320.1	farnesyltransferase	Leishmania braziliensis MHOM/BR /75/M2904	details
□	XP_844805.1	protoheme IX farnesyltransferase	Trypanosoma brucei TREU927	details
□	ABJ97611.1	protein farnesyltransferase beta subunit	Plasmodium falciparum	details
□	AAW78026.1	protein farnesyltransferase alpha subunit	Plasmodium falciparum	details
□	XP_001684151.1	farnesyltransferase beta subunit	Leishmania major strain Friedlin	details
□	AAW78028.1	protein farnesyltransferase beta subunit	Plasmodium falciparum	details
□	XP_814518.1	protein farnesyltransferase alpha subunit	Trypanosoma cruzi strain CL Brener	details
□	AAX70458.1	protoheme IX farnesyltransferase, putative	Trypanosoma brucei	details
□	XP_002140592.1	farnesyltransferase / geranylgeranyltransferase type-1, subunit alpha	Cryptosporidium muris RN66	details
□	CAQ40393.1	cytochrome c oxidase assembly protein (heme A: farnesyltransferase), putative	Plasmodium knowlesi strain H	details
□	CAM40048.1	protein farnesyltransferase alpha subunit,putative	Leishmania braziliensis	details
□	AAN36063.2	farnesyltransferase beta subunit, putative	Plasmodium falciparum 3D7	details
□	XP_668348.1	farnesyltransferase	Cryptosporidium hominis TU502	details

Details

ProtozoaDB | PDB | KEGG | NCBI

+ ProtozoaDB

+ Human Proteome Similarity (Top 10 Blastp)

+ Orthologs by OrthoMCL [go to paper]

+ SuperFamily [go to Web Page]

− Phenotype by Orthology

In *Saccharomyces cerevisiae*

- competitive fitness: decreased
- inviable
- resistance to chemicals: decreased

Databases Statistics

Genes 193559

Proteins 218100

Figure 7: Phenotypes information. The figure show the phenotypes found by orthology with *Saccharomyces cerevisiae to protein farnesyltransferase alpha subunit.*

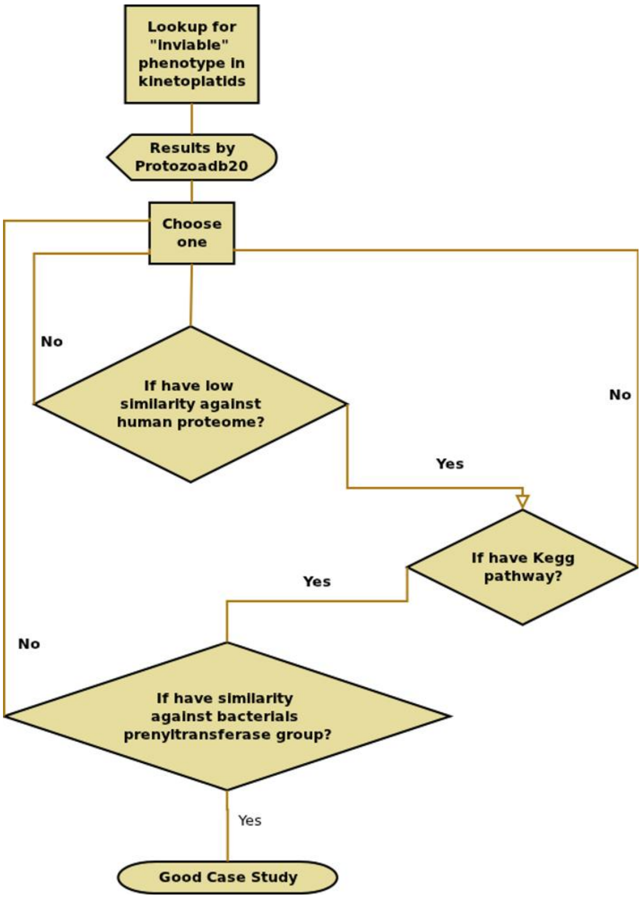

Lookup for "inviable" phenotype in kinetoplatids
Results by Protozoadb20
Choose one
No
If have low similarity against human proteome?
No
Yes
If have Kegg pathway?
Yes
No
If have similarity against bacterials prenyltransferase group?
Yes
Good Case Study

Figure 8: Case study. Flow chart showing the options and choices for the identification of a good case study.

Tables

Table 1: Species List of organisms loaded in ProtozoaDB 2.0.

Organism / Strain
Babesia bovis T2Bo
Crypstosporidium parvum Iowa II
Cryptosporidium hominis TU502
Cryptosporidium muris RN66
Entamoeba dispar SAW760
Entamoeba histolytica HM1:IMSS
Giardia lamblia ATCC 50803
Leishmania braziliensis MHOM BR 75 M2904
Leishmania infantum JPCM5
Leishmania major Friedlin
Plasmodium berghei ANKA
Plasmodium chabaudi chabaudi AS
Plasmodium falciparum 3D7
Plasmodium knowlesi strain H
Plasmodium vivax Sal 1
Plasmodium yoelii yoelii 17XNL
Theileria annulata Ankara
Theileria parva Muguga
Toxoplasma gondii ME49
Trichomonas vaginalis G3
Trypanosoma brucei treu927
Trypanosoma cruzi CL Brener

Table 2: Web services accessed

Information	URL
ProtozoaDB	http://services.biowebdb.org/howtouse
PDB	http://www.rcsb.org/pdb/software/rest.do
Kegg	http://www.kegg.jp/kegg/docs/keggapi.html
PubMed(NCBI)	http://www.ncbi.nlm.nih.gov/books/NBK55693/
Superfamily	http://supfam.cs.bris.ac.uk/SUPERFAMILY/web_services.html

Table 3: Comparing funcionalities between ProtozoaDB 2.0 and EuPathDB

Functionalities	**ProtozoaDB 2.0**	**EuPathDB**
Blast similarities against *Homo sapiens*	X	
Blast similarities against model organisms	X	
Blast similarities against protozoa species	X	X
Blast similarities against CDD	X	
Blast similarities against PDB	X	
Similarities against Intepro Domains		X
KEGG metabolic pathways	X	
Protein structures by PDB	X	
Homology study: OrthoMCL	X	X
Homology study: KEGG orthologous	X	
Homology study: OrthoSearch	X	
Publications at PubMed	X	
Superfamily classification	X	
Phenotype Search	X	X
SNP Characteristics Search		X
Genomic Position Search		X
Cellular Location Search		X
Transcript Expression Search		X

Additional Files

Additional file 1 — Source code of Protozoadb 2.0

Complete source code of ProtozoaDB 2.0 including library developed by our group writing in Ruby and RoR (Ruby in Rails). All files are in text format, except images that are in JPG format.

9.2- Artigo 5: “STINGRAY: Sistema Integrado para Analises de Recursos Genômicos”

Wagner G, Jardim M, **Tschoeke DA**, Loureiro DR, Ocaña KACS, Ribeiro ACB, Emmel VE, Probst CM, Pitaluga AN, Grisard EC, Cavalcanti MC, Campos MLM, Mattoso M, Dávila AMR. STINGRAY: System for Integrated Genomic Resources and Analysis.

Este artigo foi submetido para a revista *BMC Research Notes.*

Article title: STINGRAY: System for Integrated Genomic Resources and Analysis
MS ID: 1134080514110841
Authors: Glauber Wagner, Rodrigo Jardim, **Diogo A Tschoeke**, Daniel R Loureiro, Kary A C S Ocana, Antonio C B Ribeiro, Vanessa E Emmel, Christian M Probst, Andre N Pitaluga, Edmundo C Grisard, Maria C Cavalcanti, Maria L M Campos, Marta Mattoso and Alberto M R Davila
Journal: BMC Research Notes

O presente artigo trata sobre as funcionalidades e facilidades do sistema de anotação STINGRAY, que foi utilizado nas análises dos dados, sequências e anotação do genoma de *Leishmania amazonensis*.

STINGRAY: System for Integrated Genomic Resources and Analysis

Glauber Wagner [1,2,3], Rodrigo Jardim [1,4], Diogo A Tschoeke [1,4], Daniel R Loureiro [1,4], Kary ACS Ocaña [1,4], Antonio CB Ribeiro [1,4,5] Vanessa E Emmel [1,6], Christian M Probst [4,7,] André N Pitaluga [5], Edmundo C Grisard [2], Maria C Cavalcanti [8], Maria LM Campos [9], Marta Mattoso [10], Alberto MR Dávila [1,4 §].

[1] Laboratório de Biologia Computacional e Sistemas, Instituto Oswaldo Cruz (IOC), Fundação Oswaldo Cruz (FIOCRUZ). Avenida Brasil 4365, 21040-900, Rio de Janeiro, Rio de Janeiro, Brazil.

[2] Laboratório de Protozoologia, Departamento de Microbiologia, Imunologia e Parasitologia (MIP), Centro de Ciências Biológicas (CCB), Universidade Federal de Santa Catarina (UFSC). Campus Universitário, Setor F, Bloco A, Trindade 88040-970, Caixa Postal 476, Florianópolis, Santa Catarina, Brazil.

[3] Laboratório de Doenças Infecciosas e Parasitárias (LDIP), Área de Ciências Biológicas e da Saúde (ACBS), Universidade do Oeste de Santa Catarina (Unoesc). Avenida Getúlio Vargas 2125, Flor da Serra 89600-000, Joaçaba, Santa Catarina, Brazil.

[4] Pólo de Biologia Computacional e Sistemas, Instituto Oswaldo Cruz (IOC), Fundação Oswaldo Cruz (FIOCRUZ).

[5] Laboratório de Biologia Molecular de Parasitas e Vetores, Instituto Oswaldo Cruz (IOC), Fundação Oswaldo Cruz (FIOCRUZ).

[6] Laboratório de Genética Molecular de Microrganismos, Instituto Oswaldo Cruz (IOC), Fundação Oswaldo Cruz (FIOCRUZ).

[7] Laboratório de Bioinformática, Instituto Carlos Chagas (ICC), Fundação Oswaldo Cruz (FIOCRUZ). Avenida Algacyr Munhoz Mader, 3775, Cidade Industrial 81350-010, Curitiba, Paraná, Brazil.

[8] Instituto Militar de Engenharia (IME), Seção de Engenharia de Computação (SE-8). Praça General Tibúrcio 80, Praia Vermelha, Urca 22290-270, Rio de Janeiro, Rio de Janeiro, Brazil.

[9] Universidade Federal do Rio de Janeiro, Instituto de Matemática, Departamento de Ciência da Computação. Bloco C, CCMN, Sala E-2206, Ilha do Fundão 21945-970, Rio de Janeiro, Rio de Janeiro, Brazil.

[10] Universidade Federal do Rio de Janeiro, Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia (COPPE). P.O. Box 68511, Ilha do Fundão 21941-972, Rio de Janeiro, Rio de Janeiro, Brazil.

[§] Corresponding Author

Email addresses:

GW: glauber.wagner@unoesc.edu.br
RJ: rodrigo_jardim@fiocruz.br
DAT: diogoat@ioc.fiocruz.br
DRL: drsloureiro@ioc.fiocruz.br
KACSO: kary@cos.ufrj.br
ACBR: acbellor@ioc.fiocruz.br
VEE: vemmel@inca.gov.br
CMP: cprobst@fiocruz.br
ANP: pitaluga@ioc.fiocruz.br
ECG: edmundo.grisard@ufsc.br
MCC: yoko@ime.eb.br
MLMC: mluiza@ufrj.br
MM: marta@cos.ufrj.br
AMRD: davila@fiocruz.br

Abstract

Background: The STINGRAY system has been conceived to ease the tasks of integrating, analyzing, annotating and presenting genomic and expression data from Sanger and Next Generation Sequencing (NGS) platforms.

Results: STINGRAY includes: (a) a complete and integrated workflow (more than 20 bioinformatics tools) ranging from functional annotation to phylogeny; (b) a MySQL database schema, suitable for data integration and user access control; and (c) a user-friendly graphical web-based interface that makes the system intuitive, facilitating the tasks of data analysis and annotation.

Conclusion: STINGRAY showed to be an easy to use and complete system for analyzing sequencing data. While both Sanger and NGS platforms are supported, the system could be faster using Sanger data, since the large NGS datasets could potentially slow down the MySQL database usage. STINGRAY is available at http://stingray.biowebdb.org.

Keywords: Genome; Annotation; Workflow; Next Generation Sequencing; Sanger; Data integration.

Background

With the expansion of genomic, transcriptomic and proteomic data, the availability for both intra and inter-specific analyses of nucleotide and protein sequences has raised new levels of difficulty for scientists to understand, integrate and compare this ever increasing information. An important and long lasting problem is how to process and deal with large complex sequence files with distinct formats and using different tools that do not easily exchange data with each other. Thus, researchers must deal with dozens of sequence formats and a variety software packages to analyze nucleotide or protein sequences.

In order to simplify such tasks, researchers have been using alternative strategies such as the development of custom *ad-hoc* scripts, sometimes even ignoring pre-existing generic modules (e.g. Bioperl, Biopython, Bioruby, Biojava). It has been widely used and has proved its efficacy for simple environments, however *ad-hoc* scripting often results in redundant work and code, difficulties to adapt, which reduces efficiency and is a more error-prone development. Furthermore, the intermediate files generated throughout the process are usually not properly stored and organized, generating a large number of files and versions that can potentially lead to errors in data processing, analyses and/or inferences.

Alongside, the use of database management systems have facilitated several tasks by enforcing integrity constraints, supporting transaction management, concurrent access control, structuring and integrating data into a single schema, and providing structured query languages (SQL), among others.

Another common problem faced by many researchers is the difficulty to handle the installation and working with Unix/Linux-based software, as well as the integration of them. Therefore, the development of user-friendly applications is becoming more common providing a uniform user interface to integrate all these programs with their inputs/outputs in scientific workflows making the annotation and functional analysis process painless to users.

There are several sequence and expression analysis workflows described, such as the EST (Expressed Sequence Tags) pipeline system [1], SABIA [2], GARSA [3], GATO [4], JUICE [5], and others for Next Generation Sequencing (NGS) data analysis, such as NGSPE [6], WEP [7] and DDBJ Pipeline [8]. However, none of these systems were designed to deal at once with EST or GSS (Genome Survey Sequences) data or from different sequencing platform as NGS or Sanger technologies in the same system. Furthermore, those available systems usually don’t include protein, phylogenetic and ontology-based analyze such as STINGRAY does. Available workflows usually require some adaptation to optimize performance for each user. For this reason we have designed a flexible workflow in which researchers can use or combine its different sequenced data (subsets of functionalities), according to their needs.

Implementation

STINGRAY purpose, development and management

Considering the previously mentioned challenges, plus the increase of available sequences and multi-team-based projects involving laboratories that are usually geographically dispersed, STINGRAY was conceived as an environment aiming to facilitate the storage, analysis, integration and presentation of genomic and gene expression information. This system integrates several bioinformatics tools and sequence databases, using a flexible and user-friendly approach.

STINGRAY workflow (Figure 1) was built upon the previous and smaller scale GARSA workflow and sustained significantly improvement as: (a) a larger number of bioinformatics programs; (b) automatic functional prediction and annotation; (c) improvement of phylogenetic analysis; (d) larger and more flexible workflow; (e) the use of a more comprehensive database schema; (f) connection with remote servers for intensive computing; (g) NGS datasets analysis; and (h) a user-friendly configuration interface, resulting in a new and comprehensive system.

The underlying STINGRAY platform includes Perl, Bioperl, CGI, Apache, MySQL, and several Linux-based bioinformatics packages (Table 1). In its current version, the system is able to handle EST, ORESTES and GSS Sanger, as well as NGS (454, SOLiD and Illumina) data, accepting as inputs: (a) Sanger-based chromatograms; (b) NGS-based 454's flowgrams, Illumina's FASTQ and SOLiD's color space; (c) nucleotide or protein FASTA sequences from GenBank [9] (Additional data file 1); (d) nucleotide or protein FASTA sequences stored locally; or (e) a combination of all of these inputs. Also, STINGRAY is able to analyze protein sequences, accepting both locally or downloaded sequences from GenBank, and to perform comprehensive sequence and genome analysis, distant homology detection and phylogenetic analysis.

The STINGRAY system is being offered as a web server (i.e. CGI-based), so that common users do not need to deal with a large number of dependencies. A web-based setup page is available to configure dependency paths and other features (Additional data file 1), thus eliminating the need for interacting directly with the Linux/Unix server. All programs (Table 1) can be configured to run locally (e.g. in the same server where STINGRAY is installed) or remotely, in a different server, like the structure available at FIOCRUZ (Additional data file 2).

Nowadays many researchers collaborate in the annotation process in different locations leading to control the different access grants for each user in order to avoid data loss, simultaneous modification, conflicts and security issues. STINGRAY system has data access control for six different user profiles: (a) system administrator; (b) project administrator; (c) "write" users (which can run programs and annotate sequences); (d) "read” users (which are not allowed to run programs or to do annotation); (e) "guest" users (which can only view non-sensitive data and low level of annotation details, and do not have permission to download/upload sequences); and (f) "statistics" users (which can only access statistical data about a project, e.g., total number of sequences analyzed).

Results and discussions

STINGRAY Workflow

To provide an integrated view and execution of required tools, the current STINGRAY workflow has two major sections, (I) nucleotide and (II) protein. Both of them share the same initial configuration section (Figure 1). For nucleotide section, STINGRAY workflow uses the Phred [10] package to process chromatograms from Sanger technology, evaluating the traces quality and removing any occasional vector contamination. Following, Repeat Masker [11] is used to find and mask repeated sequences, and CAP3 [12] for clustering the sequences (reads) into a consensus sequence (clusters).

To deal with NGS datasets, MIRA [13] package is used to process 454 flowgrams [14] and Illumina reads, while ABI SOLiD™ System *de novo* Accessory Tools 2.0 package and VELVET [15] deal with the color-space dataset. These packages enable STINGRAY to perform the *de novo* assembly routines, provided by the short-read assemblers generating contigs data set and the output data can be loaded in STINGRAY databases that will consider the each sequence as cluster.

Gene predictions for prokaryote and eukaryote genomes are performed using Glimmer [16] and GlimmerHMM [17], respectively. Furthermore, users can continue with subsequently analysis using (a) all clustered sequences, (b) the Open Reading Frames (ORFs) sequences predicted by gene finders or (c) both.

To estimate G+C content and codon usage, STINGRAY uses the EMBOSS Geecee and Cusp packages [18], respectively. Clusters or ORFs are then submitted to standalone BLAST [25] for similarity searches against user-defined datasets, downloaded and updated by the server administrator, using an intuitive interface.

To use protein section of STINGRAY workflow, users need to upload the amino acid sequences in FASTA format and the system will consider each entry as a unique sequence. It is important to mention that STINGRAY automatically recognizes the project type (nucleotide or protein).

STINGRAY offers to user a phylogenetic module that all three steps typically necessary for molecular phylogenetic analysis, (1) retrieval/inference of homologous sequences, (2) creation of multiple sequence alignments and, (3) phylogenetic tree construction can be performed in STINGRAY. Besides, STINGRAY allowed user to infer phylogenetic trees using either full cluster, ORF or high-scoring segment pairs (HSP) obtained automatically by BLAST, then multiple sequence alignments generated using ClustalW [19], MAFFT [20] or ProbCons [21] packages and alignments can be presented in ClustalW, PHYLIP, and WebLogo formats. Phylogenetic trees are built using SeqBoot, Dnadist, Protdist, Neighbor and Consense software from PHYLIP package [22], as well as Weighbor [23] or Ancescon [24] algorithms. Phylogenetic trees outputs are presented in PHYLIP, NEXUS and NEWICK formats, which are available for users visualization and download (Figure 2).

STINGRAY Schema

STINGRAY uses the MySQL Database Management System (DBMS) to store all data in order to improve the performance, data security and management. A relational schema was specially designed to register and straightforward future reference of all data produced by the workflow execution. The schema is able to register data from projects and their users and also permits to maintain data about project-specific configuration and access restrictions. In addition, the STINGRAY data schema also provides some data provenance. For instance, the Cluster table, in the core of the STINGRAY data schema (Additional Data File 3), records the cluster sequence under investigation and these are connected to the reads used to assemble it by the provenance registered in the Clusters_Reads table. The clustering/assembly software execution (e.g. CAP3, MIRA or VELVET) is also registered in the Clustering table.

Furthermore, each BLAST analysis is registered in the “Blast_Search” table, which stores information about the parameters values used for that BLAST analysis, such as the sequence database (e.g. NCBI nr, Swissprot), BLAST algorithm (e.g. blastx, blastn) and all BLAST similarity feature, as hit sequences, accessions, score, e-value, consensus sequences, identity and positive values among others are sorted in the Blast_Hit table. So it is possible to link the results with a specific BLAST analysis. The Additional Data File 4 shows the complete relational schema.

Functional annotation

Once sequence data is uploaded or assembled, clusters are generated and used for subsequent analyses. All sequences and their analyses results can be assessed (depending on the user access level) and viewed by a straightforward Web interface (Additional Data File 5 and 6). One example is the Cluster View page where all data analyses for a chosen cluster are summarized (Additional Data File 7). The user can also compare sequences from a specific library, compare common sequences among all libraries or even obtain library-specific sequences. BLAST [25], InterProScan [26], PSI-BLAST [25], HMMER [26], tRNA-Scan [27], WolfPsort [28], SignalP [29], Gene Prediction and Automatic Function Prediction (AFP) (by using Glimmer [16] or GlimmerHMM [17]) results are presented in specific tables in a unique interface that holds all necessary information for user analysis (Additional Data File 8). Besides that, in this interface, the user select sequences to perform phylogenetic analysis as described before (Figure 2). Another important feature provided by STINGRAY is to allow ontology-based functional annotation using Gene Ontology (GO) terms (http://www.geneontology.org/) as part of AFP. This feature was implemented based on similarity results with databases like Seqdblite (GO), UniProtKB (Swiss-Prot and TrEMBL) [30], and InterPro [31] results as formerly proposed. Briefly, the GO descriptions, associated to sequences from each quoted databases, are used for semi-automatic annotation of clusters, ORFs or proteins. The methodology that scores the terms through accordance and distance methods was incorporated into STINGRAY [34]. The system suggests which terms are more "related" to the protein sequence being analyzed, allowing the user to define the best functional annotation for a sequence during the manual annotation process.

For manual annotation, the user must inform the region of the cluster that corresponds to a coding sequence (CDS) or select one ORF in the list of the cluster. STINGRAY then estimates the G+C content, predicts the amino acid sequence for CDS, sub-cellular location (using Wolf-Psort) and peptide cleavage signals (using SignalP). When available, information about (i) Enzyme Commission Code (Union of Biochemistry and Molecular Biology), (ii) Monica Riley classification numbers, (iii) similar organism, (iv) BLAST similarity, (v) domain, and (vi) notes about sequences, are automatically included as part of the annotation process (Additional Data File 8). After one sequence has been annotated by AFP, the user can manually verify the results through an interface to confirm or update the automatically annotation.

Since STINGRAY is a multi-user system, the confidentially and maintenance integrity of the data are important. To achieve this level of security only "project administrator" and "write" users can modify and run the programs needed to perform the annotation of the sequences. Nevertheless, only project administrators have permission to remove any data.

Once a project in STINGRAY is finished/published, the project administrator allows, upon user request, the data and analyses to become public, and then the scientific community (or “read” users) can view some project details (Additional data file 9), statistics (Additional data file 10) and graphics.

A common concern in sequencing projects, especially EST and GSS projects, is the GenBank submission of the annotated sequences. STINGRAY provides an intuitive interface where all sequences are formatted and a file generated according to GenBank recommendations in order to facilitate and enable the submission; however, the annotator or project administrator *per se* must perform the submission process externally or manually.

Pre-assembly and automatic functional prediction test

In order to test the STINGRAY workflow, genomes from bacterial *Escherichia coli* K12, *Neisseria meningitides, Streptococcus pneumoniae* GA17457, and eukaryote *Phlebotomus papatasi* (NCBI Sequence Read Archive number SRX000353, ERX005963, SRX028097, and SRX027131, respectively) were processed, pre-assembled then AFP was performed. The time of pre-assembly, GO-based annotations, and numbers of contigs obtained by pre-assembly for each of the four data sets are listed in Table 2. Assembly quality is strongly dependent on quantity (coverage) and quality of data/reads as well as “fine tuning” of the many parameters available in the genome assembler software, then the pre-assembly performed was done with the only purpose to illustrate the STINGRAY functionalities.

Conclusions

STINGRAY is currently under use by more than 20 different projects, among them: *T. vivax* (GSS and EST) [35], *Bothrops jararaca* (EST) [36], *Lutzomyia longipalpis* (EST) [37, 38], *Taenia solium* (EST) [39] and *Trypanosoma rangeli* (GSS, EST and ORESTES) [40]. The main advantage of STINGRAY over GARSA and related systems is its larger and flexible workflow on which advanced users or annotators are able to fine-tune the parameters of some programs to extract the maximum of valuable information and knowledge from their sequences. The Stingray ipeline is able to manipulate Sanger and NGS sequence data (table 3), whereas other (recently developed) pipelines do not longer center on Sanger technology. Since this data format is still used in several institutions, a system being able to deal with these two technologies should be seen as an advantage. Furthermore, Stingray is a complete annotation pipeline, allowing the user to perform automatic, semi-automatic and manual annotation, while others pipelines like RATT [41], WEP [6] and NGSPE [7] perform only automatic annotation. Stingray is also capable to edit annotations and as far as we know Artemis [42, 43] is the only other system with this feature. Finally, Stingray is the only pipeline which allows the use of Intepro Search, Phylogeny analysis, Codon Usage Analysis and tRNA sequence prediction integrated into a unique system, then being a web-based platform with friendly interface is a plus. Future developments and improvements include the use of "cloud-based" applications as part of its workflow, either using private clouds or even commercial ones as Amazon's (http://aws.amazon.com/ec2/). In the current context of high-throughput sequence generation and many more genomic and metagenomic projects being planned, the use of cloud computing is the way forward. Larger and improved database schemas as GUS (Genomics Unified Schema) (http://www.gusdb.org/) could be potentially used to content different datasets and sequence features. Data integration using LOD (Linked Open Data) technology is also planned for the next version, as it is now clear that connecting local data with many other sources (curated, non-curated or even complementary) in the LOD cloud (http://richard.cyganiak.de/2007/10/lod/) might help to accelerate knowledge extraction. Online documentation for installation using STINGRAY and technical information are available at http://stingray.biowebdb.org.

Availability and requirements

Project name: Stingray

Project home page: http://stingray.biowebdb.org

Operating system(s): Unix

Programming language: Perl

Other requirements: Perl, Apache, MySQL

License: GNU GPLv2.

Any restrictions to use by non-academics: license

List of abbreviations

AFP	…	Automatic Function Prediction
BLAST	…	Basic Local Alignment Search Tool
CDS	…	Coding Sequences
CGI	…	Common Gateway Interface
DBMS	…	Database Management System
DDBJ	…	DNA Data Bank of Japan
EMBOSS	…	The European Molecular Biology Open Software Suite
EST	…	Expressed Sequence Tags
GARSA	…	Genomic Analysis Resources for Sequence Annotation
GATO	…	Gene Annotation Tool
GSS	…	Genomic Sequence Survey
GO	…	Gene Ontology
GUS	…	Genomics Unified Schema
LOD	…	Linked Open Data
MPI	…	Message Passing Interface
NCBI	…	National Center for Biotechnology Information
NFS	…	Network File System
NGS	…	Next Generation Sequencing
ORESTES	…	Open Reading frame ESTs
ORF	…	Open Reading Frames
PERL	…	Practical Extraction and Report Language

PSI-BLAST	…	Reversed Position Specific BLAST
RPS-BLAST	…	Position specific iterative BLAST
SABIA	…	System for Annotation Bacterial (genome) Integrated Annotation
SQL	…	Structured Query Language
STINGRAY	…	System for Integrated Genomic Resources and Analysis
tRNA	…	Ribonucleic Acid Transporter

Competing interests

All other authors declare that they have no competing interests.

Authors' contributions

GW, DAT and DRL were responsible for programming, development and tests with Sanger data. KACSO participated in the programming and development of some bioinformatics tasks and drafted the manuscript. RJ, ACBR and VEE participated in the assembly development and tested the system. CMP, ANP, ECG, MCC, MLMC, MM, AMRD conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

To Brazilian sponsors CNPq, FAPERJ and CAPES for financial support. To Pablo Mendes, Sérgio Serra, Milene Guimarães, Raphael Tavares, Guilherme Geronimo, Michel Abanto and Ana C. P. Vicente for useful discussions and help either during the early stage of development or the final testing of the system.

References

1. Xu H, He L, Zhu Y, Huang W, Fang L, Tao L, Zhu Y, Cai L, Xu H, Zhang L, Xu H, Zhou Y: **EST pipeline system: detailed and automated EST data processing and mining**. *Genomics Proteomics Bioinformatics* 2003, **1**:236–242.

2. Almeida LGP, Paixão R, Souza RC, Costa GC da, Barrientos FJA, Santos MT dos, Almeida DF de, Vasconcelos ATR: **A System for Automated Bacterial (genome) Integrated Annotation--SABIA**. *Bioinformatics* 2004, **20**:2832–2833.

3. Dávila AMR, Lorenzini DM, Mendes PN, Satake TS, Sousa GR, Campos LM, Mazzoni CJ, Wagner G, Pires PF, Grisard EC, Cavalcanti MCR, Campos MLM: **GARSA: genomic analysis resources for sequence annotation**. *Bioinformatics* 2005, **21**:4302–4303.

4. Fujita A, Massirer KB, Durham AM, Ferreira CE, Sogayar MC: **The GATO gene annotation tool for research laboratories**. *Braz. J. Med. Biol. Res.* 2005, **38**:1571–1574.

5. Latorre M, Silva H, Saba J, Guziolowski C, Vizoso P, Martinez V, Maldonado J, Morales A, Caroca R, Cambiazo V, Campos-Vargas R, Gonzalez M, Orellana A, Retamales J, Meisel LA: **JUICE: a data management system that facilitates the analysis of large volumes of information in an EST project workflow**. *BMC Bioinformatics* 2006, 7:513.

6. Huang K, Yellapantula V, Baier L, Dinu V: **NGSPE: A pipeline for end-to-end analysis of DNA sequencing data and comparison between different platforms**. *Comput Biol Med.* 2013, **43**:1171-6.

7. D'Antonio M, D'Onorio De Meo P, Paoletti D, Elmi B, Pallocca M, Sanna N, Picardi E, Pesole G, Castrignanò T: **WEP: a high-performance analysis pipeline for whole-exome data.** *BMC Bioinformatics* 2013; **14**:S11

8. Nagasaki H, Mochizuki T, Kodama Y, Saruhashi S, Morizaki S, Sugawara H, Ohyanagi H, Kurata N, Okubo K, Takagi T, Kaminuma E, Nakamura Y: **DDBJ Read Annotation Pipeline: A Cloud Computing-Based Pipeline for High-Throughput Analysis of Next-Generation Sequencing Data**. *DNA Res.* 2013, **20**:383-90.

9. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res.* 2012, **40**:D48–53.

10. Machado M, Magalhães WC, Sene A, Araújo B, Faria-Campos AC, Chanock SJ, Scott L, Oliveira G, Tarazona-Santos E, Rodrigues MR: **Phred-Phrap package to analyses tools: a pipeline to facilitate population genetics re-sequencing studies**. *Investig Genet* 2011, **2**:3.

11. Smit, AFA, Hubley, R, Green, P: *RepeatMasker*. 2010.

12. Huang X, Madan A: **CAP3: A DNA sequence assembly program**. *Genome Res.* 1999, **9**:868–877.

13. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs**. *Genome Res.* 2004, **14**:1147–1159.

14. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**:376–380.

15. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Res.* 2008, **18**:821–829.

16. Delcher AL, Bratke KA, Powers EC, Salzberg SL: **Identifying bacterial genes and endosymbiont DNA with Glimmer**. *Bioinformatics* 2007, **23**:673–679.

17. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders**. *Bioinformatics* 2004, **20**:2878–2879.

18. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet.* 2000, **16**:276–277.

19. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: **Clustal W and**

Clustal X version 2.0. *Bioinformatics* 2007, **23**:2947–2948.

20. Katoh K, Toh H: **Parallelization of the MAFFT multiple sequence alignment program**. *Bioinformatics* 2010, **26**:1899–1900.

21. Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S: **ProbCons: Probabilistic consistency-based multiple sequence alignment**. *Genome Res.* 2005, **15**:330–340.

22. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2)**. *Cladistics* 1989, **5**:164–166.

23. Bruno WJ, Socci ND, Halpern AL: **Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction**. *Mol. Biol. Evol.* 2000, **17**:189–197.

24. Cai W, Pei J, Grishin NV: **Reconstruction of ancestral protein sequences and its applications**. *BMC Evol. Biol.* 2004, **4**:33.

25. Altschul SF, Madden TL, Schäffer 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, **25**:3389–3402.

26. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro**. *Bioinformatics* 2001, **17**:847–848.

27. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence**. *Nucleic Acids Res.* 1997, **25**:955–964.

28. Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: **WoLF PSORT: protein localization predictor**. *Nucleic Acids Res.* 2007, **35**:W585–587.

29. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nat. Methods* 2011, **8**:785–786.

30. Apweiler R: **UniProt: the Universal Protein knowledgebase**. *Nucleic Acids Research* 2004, **32**:115D–119.

31. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, *et al.*: **InterPro in 2011: new developments in the family and domain prediction database**. *Nucleic Acids Research* 2011, **40**:D306–D312.

32. Crooks GE, Hon G, Chandonia J-M, Brenner SE: **WebLogo: a sequence logo generator**. *Genome Res.* 2004, **14**:1188–1190.

33. Keane TM, Creevey CJ, Pentony MM, Naughton TJ, Mclnerney JO: **Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified**. *BMC Evol. Biol.* 2006, **6**:29.

34. Jones CE, Baumann U, Brown AL. **Automated methods of predicting the function of biological sequences using GO and BLAST**. *BMC Bioinformatics* 2005, **6**:27

35. Guerreiro LT, Souza SS, Wagner G, De Souza EA, Mendes PN, Campos LM, Barros L, Pires PF, Campos ML, Grisard EC, Dávila AM: **Exploring the genome of Trypanosoma vivax through GSS and in silico comparative analysis**. *OMICS* 2005, **9**:116-28.

36. Cidade DA, Simão TA, Dávila AM, Wagner G, Junqueira-de-Azevedo IL, Ho PL, Bon C, Zingali RB, Albano RM: **Bothrops jararaca venom gland transcriptome: analysis of the gene expression pattern**. *Toxicon* 2006, **48**:437-61.

37. Pitaluga AN, Beteille V, Lobo AR, Ortigão-Farias JR, Dávila AM, Souza AA, Ramalho-Ortigão JM, Traub-Cseko YM: **EST sequencing of blood-fed and Leishmania-infected midgut of Lutzomyia longipalpis, the principal visceral leishmaniasis vector in the Americas**. *Mol Genet Genomics* 2009, **282**:307-17.

38. Azevedo RV, Dias DB, Bretãs JA, Mazzoni CJ, Souza NA, Albano RM, Wagner G, Davila AM, Peixoto AA: **The transcriptome of Lutzomyia longipalpis (Diptera: Psychodidae) male reproductive organs**. *PLoS One,* 2012, **7**:e34495.

39. Almeida CR, Stoco PH, Wagner G, Sincero TC, Rotava G, Bayer-Santos E, Rodrigues JB, Sperandio MM, Maia AA, Ojopi EP, Zaha A, Ferreira HB, Tyler KM, Dávila AM, Grisard EC, Dias-Neto E: **Transcriptome analysis of Taenia solium cysticerci using Open Reading Frame ESTs (ORESTES).** *Parasit Vectors* 2009, **2**:35.

40. Grisard EC, Stoco PH, Wagner G, Sincero TC, Rotava G, Rodrigues JB, Snoeijer CQ, Koerich LB, Sperandio MM, Bayer-Santos E, Fragoso SP, Goldenberg S, Triana O, Vallejo GA, Tyler KM, Dávila AM, Steindel M: **Transcriptomic analyses of the avirulent protozoan parasite Trypanosoma rangeli.** *Mol Biochem Parasitol.* 2010, **174**:18-25.

41. Otto TD, Dillon GP, Degrave WS, Berriman M: **RATT: Rapid Annotation Transfer Tool**. *Nucleic Acids Res*. 2011, **39**(9): e57.

42. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics.* 2000,**16**;10;944-5.

43. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA: **Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data.** *Bioinformatics.* 2012, **28**;4;464-9.

Figure legends

Figure 1

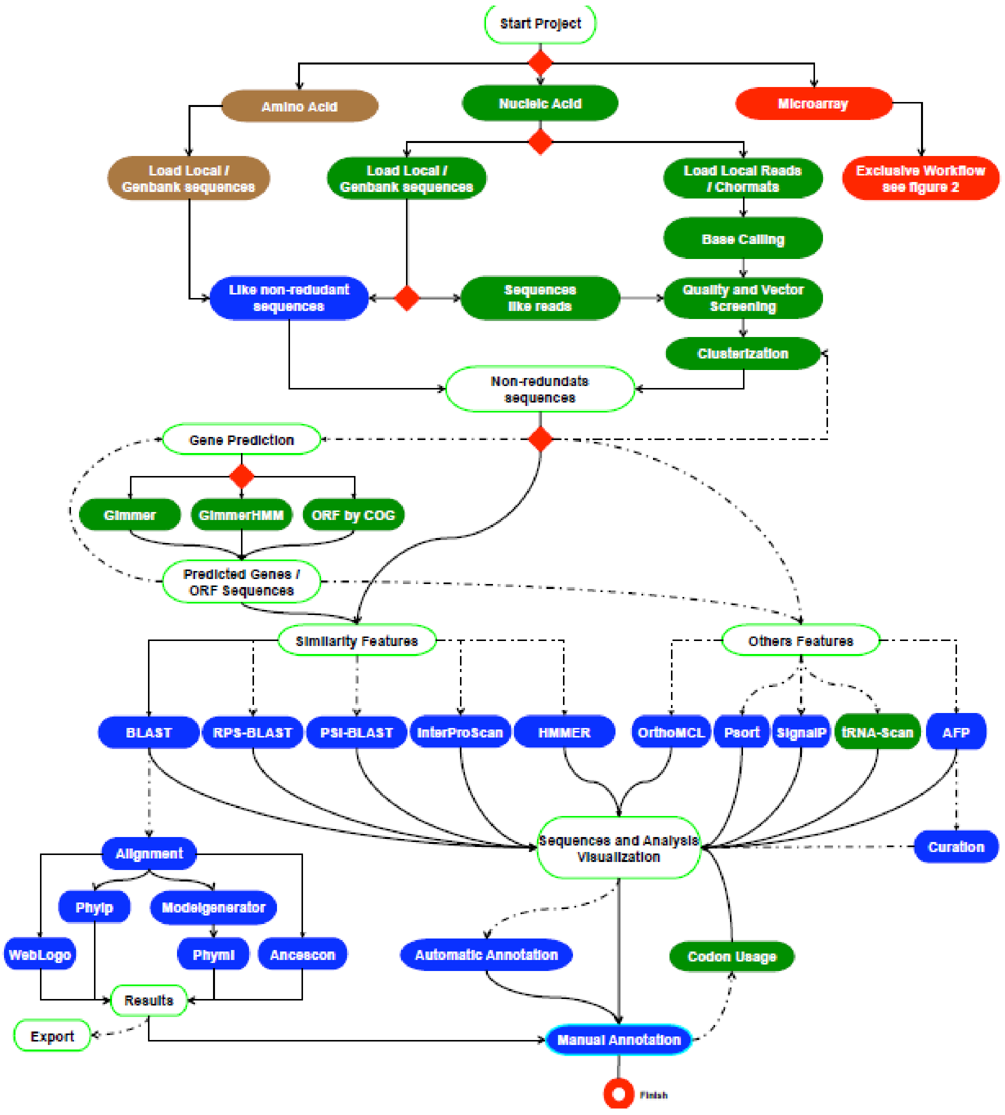
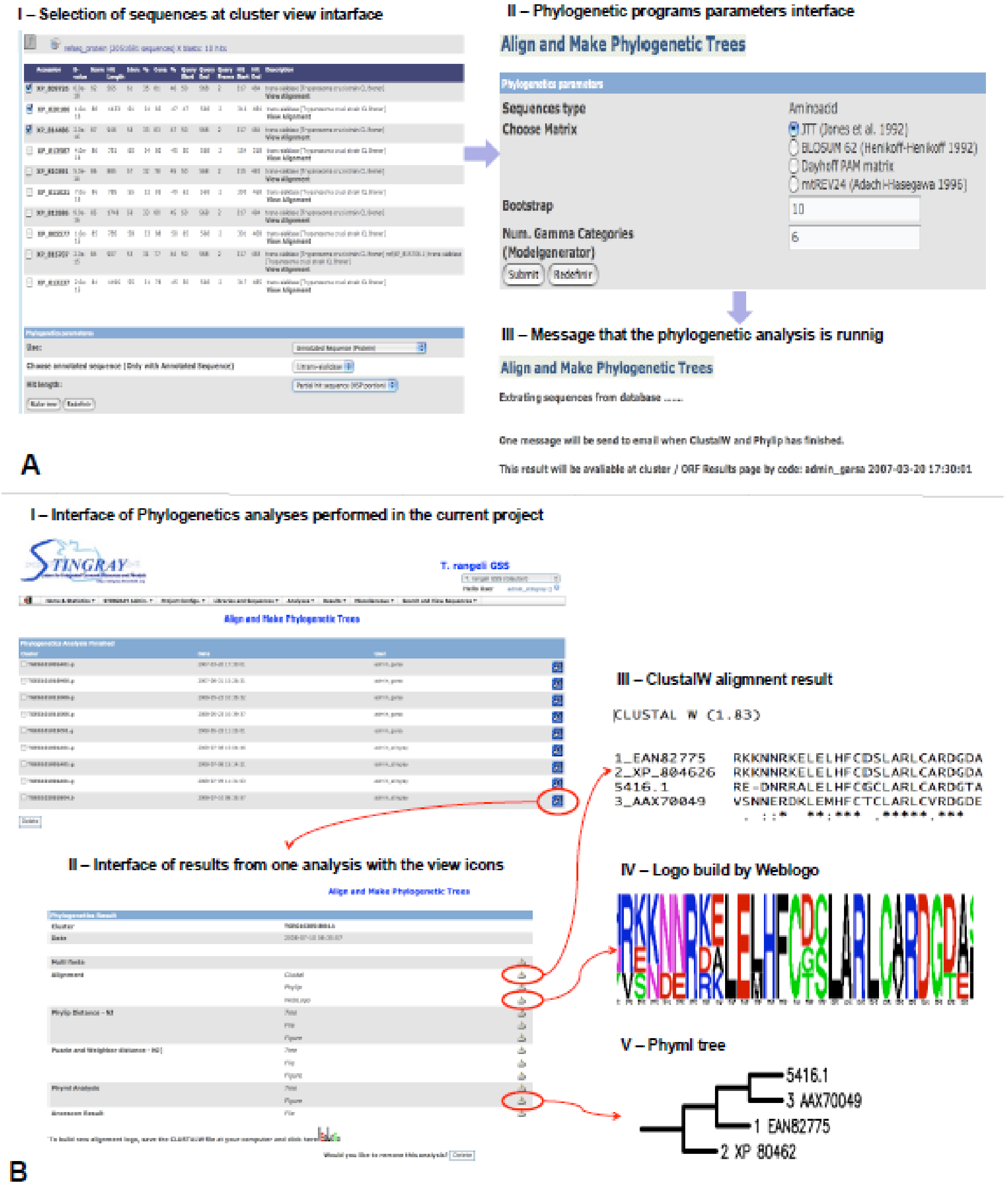


Figure 1 - Schematic representation of the STINGRAY workflow

The white boxes show the analyses categories and respective visualization interface. Green boxes indicate software or features for nucleic acid analyses, brown boxes for protein analyses and blue boxes are for both. Red diamonds indicate the user decision or inputs points. Full lines represent minimal analyses and discontinuous lines are the alternative or non-obligatory analyses. The AFP box is the automatic functional prediction based on Gene Ontology.

Figure 2



I – Selection of sequences at cluster view intarface

II – Phylogenetic programs parameters interface

Align and Make Phylogenetic Trees

Phylogenetics parameters

Sequences type: Aminoacid

Choose Matrix:
● JTT (Jones et al. 1992)
○ BLOSUM 62 (Henikoff-Henikoff 1992)
○ Dayhoff PAM matrix
○ mtREV24 (Adachi-Hasegawa 1996)

Bootstrap: 10

Num. Gamma Categories (Modelgenerator): 6

III – Message that the phylogenetic analysis is runnig

Align and Make Phylogenetic Trees

Extrating sequences from database ......

One message will be send to email when ClustalW and Phylip has finished.

This result will be avaliable at cluster / ORF Results page by code: admin_garsa 2007-03-20 17:30:01

I – Interface of Phylogenetics analyses performed in the current project

II – Interface of results from one analysis with the view icons

III – ClustalW alignment result

```
CLUSTAL W (1.83)

1_EAN82775     RKKNNRKELELHFCDSLARLCARDGDA
2_XP_804626    RKKNNRKELELHFCDSLARLCARDGDA
5416.1         RE-DNRRALELHFCGCLARLCARDGTA
3_AAX70049     VSNNERDKLEMHFCTCLARLCVRDGDE
                .  ::*  **:***  .*****.***
```

IV – Logo build by Weblogo

V – Phyml tree

Figure 2 - Phylogenetic workflow of STINGRAY

A.I) Blast hits and project sequence selection; A.II) interface to define the evolutionary and phylogenetic parameters used by Phylip; A.III) phylogenetic progress analyses information. When the phylogenetic execution is concluded, the previous code is presented in cluster/ORF view page (B.I); list of all phylogenetic results (B.II); ClustalW alignment (B.III); alignment logo was performed by WebLogo program [32] (B.IV) and phylogenetic tree was performed by Phylip program (B.V).

Tables

Table1-Bioinformatics software or packages incorporated on STINGRAY workflow.

Software/Packages	Workflow Function	**Ref.**
MIRA	Assembly	[13]
Velvet	Assembly	[15]
Phred	Reads quality match	[10]
Crossmatch	Vectormask	[10]
RepeatMasker	Repeatsequencemask	[11]
CAP3	Sequencesclusterization	[12]
Glimmer3	Prokaryoticgeneprediction	[16]
GlimmerHMM	Eukaryoticgeneprediction	[17]
Geecee	G+Ccontent calculation	[18]
Cusp	Codonusagecalculation	[18]
tRNA-Scan	tRNAsearch	[27]
BLAST	Similaritysearch	[25]
Rps-BLAST	Conserveddomainsearch	[25]
Psi-BLAST	Similaritysearch	[25]
Signalp	Signalpeptidecleavagesites prediction	[29]
Wolf-Psort	Proteinlocalization	[28]
MAFFT	Multiple sequence alignmentconstruction	[20]
ProbCons	Multiple sequence alignmentconstruction	[21]
WebLogo	Alignmentlogos generation	[32]
Ancescon	Ancestorsequenceprediction	[24]
Phylip	Phylogenetic tree construction	[22]
Weighbor	Phylogenetic tree construction	[23]
ModelGenerator	Evolutionarymodel search	[33]

Table2- Pre-assembly and automatic functional prediction test

Organism	Genome size	SRA[1]	Sequencing technology	Totalreads	Nºofcontigs	Timeof pre-assembly	Number of sequences with one GO-basedannotations at least			
							MF[2]	BP[3]	CC[4]	WA[5]
*E.coli*K12	4,7Mb	SRR001354	SOLiD	25162805	11011	30min	1593	1577	1562	9328
N.meningitidis	2Mb	ERR015596	Illumina	5418859	4815	3h4min	1055	1055	995	2812
*S.pneumoniae*GA17457	2Mb	SRR068304	454GSFLX	252646	11317	2h52min	5899	5942	5393	5340
P.papatasi	~170Mb	SRR066482	454Titanium	498629	9836	5h18min	554	553	458	9993

1 - SRA: Sequence Read Archive

2 - MF: Molecular Function

3 - BF: Biological Process

4 - CC: Cellular component

5 - WA: Without annotation

Table 3: Features comparison between Stingray and others annotation pipelines

Features	Stingray	Ratt	Artemis	WEP	NGSPE
Preproccesing sequencing output data files	X			X	X
Proccesing Sanger output files	X				
Proccesing NGS output files	X			X	X
Gene Prediction	X	X	X		
Similarity Search Blast	X		X		
Similarity Search Hmmer	X		X		
Similarity Search RPSBlast	X		X		
Similarity Search Interpro	X				
Similarity Search PSIBlast	X				
Homologs identification	X	X			
Phylogeny analysis	X				
Codon Usage analysis	X				
tRNA prediction	X				
Manual annotation	X		X		
Semi-automatic annotation	X				
Automatic annotation	X	X		X	X
Friendly Interface	X		X		
Web platform	X			X	
Use of SGBD	X			X	
Applet			X		
Browse Genome Visualization			X		
Jalview Visualization			X		
Generate GBFF file	X		X		
Generate SeqIn file	X		X		

Additional files

Additional Data File 1 – Screenshot from configuration interface

With this intuitive interface the system manager can configure all programs paths and options of software include on STINGRAY workflow, as well as some project parameters.

Additional Data File 2 – Schema of FIOCRUZ servers where the STINGRAY is installed

To improve STINGRAY performance the system platform (i.e. CGI/Perl scripts) was installed on the web-server and software as Phred, CAP3, BLAST, InterProScan, among others were installed on "process server" and MySQL were located on database server. The requested program stared by user STINGRAY on web server is forward to the process server using in-house scripts and after the program has finished the output file located on Network File System (NFS) partition is parsed and the results are stored at MySQL database.

Additional Data File 3 – The core of relational STINGRAY database schema.

This figure shows the resume of relational STINGRAY database schema. The boxes represent the SQL tables and the lanes the relation between the tables.

Additional Data File 4 – The complete STINGRAY database schema.

This figure shows the complete relational STINGRAY database schema. The boxes represent the SQL tables and the lanes the relation between the tables.

Additional Data File 5 – Screenshot of search sequence interface

In this interface the users can search sequences by the identification of the reads, clusters or ORF or even by BLAST/InterPro/HMMER, annotations or Gene Ontology descriptions.

Additional Data File 6 – Screenshot of BLAST results search interface

Using this interface the user can view the all similarity BLAST results. Notice the other results interfaces are available at the upper menu.

Additional Data File 7 – Screenshot of cluster view interface

This intuitive interface shows all cluster features, like length, reads and similarity results obtain by BLAST, InterProScan and HMMER results. The ORF view interface is similar.

Additional Data File 8 – Screenshot of annotation (CDS) interface

This interface allowed user to annotate the sequence and insert other important information.

Additional Data File 9 – Screenshot of a current available project

This is the specific project page, with the information about the project and number of sequences.

Additional Data File 10 – Statistic reports interface screenshot

In this interface the user can view the summary of the current project data.

Additional Data File 11 – STINGRAY source code

Al the scripts and web pages needed to setup STINGRAY are available in this compressed file