André Guilherme da Costa Martins

Genes codificadores de proteínas implicadas na relação de espécies do gênero *Trypanosoma* com seus hospedeiros: diversidade, transferência horizontal e relações filogenéticas.

Tese apresentada ao Programa de Pós-Graduação em Biologia da Relação Patógeno Hospedeiro do Instituto de Ciências Biomédicas da Universidade de São Paulo, para obtenção do Título de Doutor em Ciências.

São Paulo
2016

André Guilherme da Costa Martins

Genes codificadores de proteínas implicadas na relação de espécies do gênero *Trypanosoma* com seus hospedeiros: diversidade, transferência horizontal e relações filogenéticas.

Tese apresentada ao Programa de Pós-Graduação em Biologia da Relação Patógeno Hospedeiro do Instituto de Ciências Biomédicas da Universidade de São Paulo, para obtenção do Título de Doutor em Ciências.

Área de concentração: Biologia da Relação Patógeno Hospedeiro

Orientadora: Profa. Dra. Marta Maria Geraldes Teixeira

Versão original

São Paulo
2016

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

Martins, André Guilherme da Costa.
Genes codificadores de proteínas implicadas na relação de espécies do gênero Trypanosoma com seus hospedeiros: diversidade, transferência horizontal e relações filogenéticas / André Guilherme da Costa Martins. -- São Paulo, 2016.

Orientador: Marta Maria Geraldes Teixeira.

Tese (Doutorado) – Universidade de São Paulo. Instituto de Ciências Biomédicas. Departamento de Parasitologia. Área de concentração: Biologia da Relação Patógeno-Hospedeiro. Linha de pesquisa: Filogênia e taxônomia de tripanossomatídeos.

Versão do título para o inglês: Genes encoding proteins implicated in the relationship of Trypanosoma species with their hosts: diversity, horizontal transfer and phylogenetic relationships.

1. Trypanosoma 2. Proteínas do choque térmico 3. HSP70 4. Alpha-Cristalino 5. Filogênia 6. I. Teixeira, Marta Maria Geraldes II. Universidade de São Paulo. Instituto de Ciências Biomédicas. Programa de Pós-Graduação em Biologia da Relação Patógeno Hospedeiro III. Título.

ICB/SBIB070/2016

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Candidato(a): André Guilherme da Costa Martins.

Título da Tese: Genes codificadores de proteínas implicadas na relação de espécies do gênero Trypanosoma com seus hospedeiros: diversidade, transferência horizontal e relações filogenéticas.

Orientador(a): Marta Maria Geraldes Teixeira.

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

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

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

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

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

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

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

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Cidade Universitária "Armando de Salles Oliveira"
Av. Prof. Lineu Prestes, 2415 – CEP. 05508-000 São Paulo, SP – Brasil
Telefone :(55) (11) 3091-7733 - telefax : (55) (11) 3091-8405
e-mail: cep@icb.usp.br

Comissão de Ética em Pesquisa

CERTIFICADO DE ISENÇÃO

Certificamos que o Protocolo CEP-ICB Nº **590/13** referente ao projeto intitulado: "*Genes codificadores de proteínas de choque térmico (HSP) de espécies do genero Trypanosoma: repertório, polimorfismo, organização genômica e relações filogenéticas*" sob a responsabilidade de **André Guilherme da Costa Martins,** foi analisado na presente data pela CEUA - COMISSÃO DE ÉTICA NO USO DE ANIMAIS e pela CEPSH- COMISSÃO DE ÉTICA EM PESQUISA COM SERES HUMANOS, tendo sido deliberado que o referido projeto não utilizará animais que estejam sob a égide da lei 11.794 de 8 de outubro de 2008, nem envolverá procedimentos regulados pela Resolução CONEP nº196 de 1996.

São Paulo, 26 de abril de 2013.

PROF. DR. WOTHAN TAVARES DE LIMA
Coordenador da CEUA - ICB/USP

PROF. DR. PAOLO M.A ZANOTTO
Coordenador da CEPsh - ICB/USP

AGRADECIMENTOS

A Profa. Dra. Marta Maria Geraldes Teixeira agradeço por confiar a mim essa pesquisa, pela orientação, aprendizado, incentivo e paciência durante todos esses anos de trabalho e também pelo exemplo de pesquisador e professor que trarei como norte em toda minha carreira profissional.

Ao Prof. Dr. João Marcelo Pereira Alves, pela co-orientação, pelos ensinamentos, pela paciência, pelo exemplo profissional e pela grande dedicação a este trabalho.

Ao Prof. Gregory Buck e Dra. Myrna Serrano pelo sequenciamento e montagem de grande parte dos genomas utilizados nesse trabalho.

À Profa. Dra. Silvia Alfieri, pela colaboração nos trabalhos, pela grande ajuda, amizade e ensinamentos ao longo de todos estes anos.

Ao Prof. Dr. Erney Camargo pelo incentivo e por me ensinar a buscar constantemente conhecimento e aprendizado.

À Carmen Takata, à Marta Campaner e à Tania Matsumoto pela paciência, colaboração em todos os trabalhos e pela amizade ao longo de todos estes anos.

Ao Robson Ferreira e ao Herakles Garcia pela amizade, paciência e aprendizado e pelo exemplo de profissionalismo ao longo desses anos.

Aos amigos do laboratório: Adriana Fuzato, Andernice, Anderson, Arlei, Bruno, Carla “Caca”, Charlotte, Flávia, Juliana, Laerte, Luciana, Lyslaine, Omar, Oneida, Paola, Priscilla, Tarcilla e Zuleima pelo companheirismo e amizade durante todos esses anos.

Aos amigos e funcionários deste departamento que de forma direta ou indireta foram muito importantes para realização deste trabalho, com muito especial carinho à Marinete, Manuel, Valnice, Emília, José Mário e Márcio.

À minha mãe Tereza e aos meus irmãos, Andrea e Marcelo, pelo amor, incentivo e apoio agora e ao logo de toda minha vida.

Ao meu pai Antônio e minha irmã Deborah que partiram, mas que sempre estarão em meus pensamentos.

À Universidade de São Paulo e CNPq e FAPESP pelo apoio as pesquisas.

Muito obrigado,

"Ter coragem diante de qualquer coisa na vida, essa é a base de tudo."

Teresa de Ávila

Este trabalho contou com apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnólogico (CNPq) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

RESUMO

Costa-Martins AG. Genes codificadores de proteínas implicadas na relação de espécies do gênero *Trypanosoma* com seus hospedeiros: diversidade, transferência horizontal e relações filogenéticas. [tese (Doutorado em Parasitologia)]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2016.

O gênero Trypanosoma representa um grupo diversificado de endoparasitas de invertebrados e vertebrados que exibem ciclos de vida complexos, com uma grande variabilidade de hospedeiros, formas infectantes e multiplicativas, e estratégias de transmissão e de evasão das defesas dos hospedeiros. Nos hospedeiros vertebrados, os tripanossomas em geral se proliferam na corrente sanguínea, mas alguns são capazes de invadir tecidos e espaços extra-vasculares, e algumas espécies se desenvolvem intracelularmente. Essa diversidade apresentada pelos tripanossomas depende de um amplo arsenal de proteínas que regulam uma variedade de processos complexos envolvidos nas relações entre hospedeiros e parasita, incluindo enzimas, chaperonas, proteínas de superfície celular e receptores. Diversas proteínas com uma variedade de funções têm sido extensivamente estudadas nas espécies patogênicas para o homem, T. cruzi e T. brucei. No entanto, essas espécies representam apenas uma pequena fração de diversidade do gênero Trypanosoma. Visando contribuir para o conhecimento sobre as proteínas envolvidas em estratégias adaptativas dos tripanossomas, espécies patogênicas foram comparadas com espécies não-patogênicas representativas de linhagens filogenéticas distantes, que exibem diversas diferenças em suas relações com os vetores e hospedeiros vertebrados. Com este intuito, neste estudo foram examinados os genomas de vários tripanossomas a fim de avaliar o repertório genético, a ocorrência de transferência horizontal de genes, a organização genômica e as relações filogenéticas dos genes que codificam as seguintes proteínas: chaperonas das famílias de proteínas de choque térmico (Hsp70 e proteínas p23/alfa-cristalinas) e enzimas codificadoras das enzimas Prolina racemase (PRAC) e catepsina-L like (CATL-like).

Os membros da superfamília Hsp70 são chaperonas que desempenham uma série de funções fisiológicas e em situações de estresse, como mudanças de temperatura e pH, ambientes extra- e intracelulares e estresse oxidativo contribuindo, assim, para a adaptação, diferenciação e sobrevivência dos tripanossomatídeos tanto em invertebrados quanto em vertebrados de todas as classes. O repertório de Hsp70 presente em tripanossomas foi definindo nesse estudo, com base em análises de polimorfismo, arquitetura de domínios e relações filogenéticas, com as seguinte famílias: Famílias de Hsp70 conservadas presentes no citoplasma (CAN Hsp70, Hsp70.4 and Hsp70.c and Hsp110), retículo endoplasmático (Grp78, Grp170) e mitocôndria (mtHsp70/Lc2.2); uma família mitocondrial extra (mtHsp70/Lc2.1) cuja presença em todos os gêneros de tripanossomatídeos exceto Trypanosoma foi comprovada neste estudo; as famílias recém-descritas, Hsp70.a e Hsp70.b, sem ortólogos definidos em outros organismos, aqui demostrada como ubíquas nos cinetoplastídeos. A análise filogenética sugeriu que a controversa família Hsp70.b foi adquirida com um único e antigo evento de transferência horizontal (HGT) de uma DnaK-like de um doador bacteriano. Todas as famílias de TryHsp70, exceto Grp78, foram localizadas em regiões altamente sintênicas. Finalmente, dois parálogos bastante divergentes, DnaK e DnaK-like, foram identificados exclusivamente nos genomas dos endossimbiontes abrigados pelos tripanossomatídeos de insetos do gênero Angomonas e Strigomonas. Análises filogenéticas com os ortólogos de cada família corroboram, de um modo geral, os principais clados suportados pela filogenia tradicional de Trypanosomatidae, com relações filogenéticas no gênero Trypanosoma melhor apoiadas nas inferências obtidas com genes das famílias Hsp70.a, CAN Hsp70, Hsp70.b e Hsp110 indicando, assim, que genes Hsp70 são excelentes marcadores filogenéticos independentemente de sua origem evolutiva.

Os domínios p23/α-cristalino (ACD) representam uma estrutura antiga e conservada, presente em muitas famílias de chaperonas eucarióticas, inclusive proteínas de choque térmico de baixo peso molecular (sHSP) presentes eucariotos e procariotos, que atuam como chaperonas em diversas vias celulares tais como proliferação, diferenciação, sinalização, modulação imunológica e resistência à patógenos. Entretanto, pouco se sabe sobre o repertório e as funções destas proteínas em protozoários parasitas. Nossa pesquisa em vários genomas de Trypanosoma spp. corroborou uma única sHSPs em todos os tripanossomatídeos investigados e, além disso, revelou um repertório compartilhado em todas as espécies de Trypanosoma, formado por oito proteínas ACD. Ortólogos de todas essas proteínas foram também encontrados em Leishmania spp., vários gêneros de tripanossomatídeos de insetos e Bodo spp. Sete proteínas ACD (TryDYX1C1, TrySGT1, Tryp23A, Tryp23B, TryNudC1, TryNudC2 e HSP20) são conservadas em eucariotos em geral, enquanto duas (TryACDP e TryACD-TPR) não apresentaram ortólogos nos bancos de dados disponíveis para outros organismos. Para

discutir possíveis funções das proteínas ACD em tripanossomas, a arquitetura dos domínios e motivos funcionais foi comparada com os respectivos arquétipos com funções conhecidas em outros organismos.

Este estudo inclui também análises de genes codificadores das enzimas prolina racemase (PRAC) e Catepsina L-like (CATL) de Trypanosoma spp. As enzimas CATL desempenham um papel importante nos processos de infecção, diferenciação celular, patogenicidade, virulência e evasão das defesas do hospedeiro em tripanossomas. Neste estudo, pela primeira vez, a principal CATL de T. congolense - congopaina, foi caracterizada em isolados dos três subgrupos genéticos (Savannah, Forest e Kilifi), que variam na virulência, patogenicidade e distribuição geográfica. O polimorfismo, organização estrutural e genealogia revelaram que o repertório de genes de congopaina divergiu de forma espécies-específica para cada subgrupo de T. congolense, com Savannah exibindo sequências altamente polimórficas, inter- e intra-isolados, enquanto os grupos Forest e Kilifi apresentaram diversidade moderada e limitada. Além disso, um PCR baseado nos genes de congopaina foi desenvolvido para um diagnóstico sensível e específico de todos os subgrupos de T. congolense. As descobertas deste estudo demonstram que os genes de congopaina são úteis para o diagnóstico, genotipagem e inferências filogenéticas e taxonômicas de T. congolense.

A enzima PRAC é responsável pela interconversão entre L- e D-prolina livre, presente em um grupo restrito de bactérias agindo como um fator de virulência. O primeiro gene eucariótico de PRAC foi descrito em T. cruzi e, em seguida, em T. vivax, e estudos funcionais e imunológicos provaram que essa proteína promove ativação policlonal de células B em animais infectado, o que leva a um atraso na resposta imunológica específica do hospedeiro e garante a sobrevivência do parasita. Neste estudo, as análises genômicas de homólogos da PRAC revelaram uma única cópia em 12 das 15 espécies de tripanossomas investigadas: nos parasitas de mamíferos T. cruzi, T. cruzi marinkellei, T. dionisii, T. erneyi, T. rangeli, T. conorhini e T. lewisi; e nos tripanossomas de anuros, serpentes, crocodilos, lagartos e pássaros. T. rangeli apresentou apenas pseudogenes; T. brucei, T. congolense e espécies relacionadas, exceto T. vivax que é filogeneticamente mais distante, perderam completamente o gene PRAC. A genealogia dos homólogos PRAC suportada uma história evolutiva congruente com a filogenia de Trypanosoma. Este dado, juntamente com dados de sintenia, das relações filogenéticas com PRAC procariotas, e a ausência de genes PRAC em tripanossomatídeos dos outros gêneros, bodonídeos e euglenídeos, sugerem que um ancestral comum de Trypanosoma adquiriu o gene PRAC, em um evento único de HGT, de uma bactéria Firmicutes ancestral, mais relacionada ao gênero Gemella e outros bacilos do que a Clostridium como previamente sugerido.

Palavras chaves: *Trypanosoma*. Proteínas do Choque térmico. Catepsina L. Genômica comparativa. Transferência horizontal de genes. Filogenia. Evolução.

ABSTRACT

Costa-Martins AG. Genes encoding proteins implicated in the relationship of *Trypanosoma* species with their hosts: diversity, horizontal transfer and phylogenetic relationships. [Ph.D. thesis (Parasitology)]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2016.

The genus Trypanosoma represents a diverse group of obligate endoparasites of invertebrates and vertebrates, displaying a complex life cycle in which host range, life cycles, infective and multiplicative forms, transmission, and infection and evasion of host defense strategies varies broadly according to species. Inside its vertebrate hosts, trypanosomes proliferate in the bloodstream, but some species invade extra-vascular tissues, and a few species develop in intracellular compartments. The Trypanosoma diversity relies on a vast arsenal of proteins that regulate an assortment of complex processes involved in host-parasite relationships, including enzymes, chaperones, cell surface proteins and receptors, which have been extensively studied in the human trypanosomes T. cruzi and T. brucei. However, these species represent a small fraction of the trypanosome diversity. To contribute to the knowledge about some proteins involved in trypanosome adaptive strategies, pathogenic species were compared with a wide set of non-pathogenic species varying in many traits on the vector and vertebrate-host relationships, and representing distant phylogenetic lineages. With this purpose, in this study we surveyed several trypanosome genomes to assess genetic repertoires, gene horizontal transfers, genome organization and phylogenetic relationships based on genes encoding the following proteins: chaperones of Heat Shock Proteins (Hsp70 and p3/alpha-crystalline proteins) and enzymes of Proline racemase (PRAC) and Cathepsin-L like (CATL-like).

The members of Hsp70 superfamily are chaperones that play a range of physiological roles, and contribute to the trypanosomatid adaptation, differentiation, and surviving in invertebrates and all classes of vertebrates under stressful situations such as shifts of temperature and pH, extra- and intracellular environments, and oxidative stress. Trypanosome Hsp70 repertoire defined in this study through polymorphism, architecture domain, phylogenetic and evolutionary analyses included: Cytosolic (CAN Hsp70, Hsp70.4 and Hsp70.c and Hsp110), ER (Grp78, Grp170) and mitochondrial (mtHsp70/Lc2.2) evolutionarily conserved Hsp70 families; an extra mitochondrial member (mtHsp70/Lc2.1) herein identified in trypanosomatids of several genera except Trypanosoma; and the last discovered TryHsp, Hsp70.a and Hsp70.b, ubiquitous in kinetoplastids with no clear orthologs in other eukaryotes. The phylogenetic analyses suggested that the controversial Hsp70.b family was acquired through a single and ancient HGT from a bacterial DnaK-like donor. All TryHsp70 families, except Grp78, were located in highly syntenic genome regions. Finally, divergent DnaK or DnaK-like paralogs were identified exclusively in the genomes of the endosymbionts harbored by the insect trypanosomatids of the genera Angomonas and Strigomonas. Phylogenetic analyses of orthologs from each family corroborated the main clades supported by the traditional phylogeny of Trypanosomatidae. The best resolved phylogenies of the genus Trypanosoma were obtained with Hsp70.a, CAN Hsp70, Hsp70.b, and Hsp110, indicating that Hsp70 genes are valuable phylogenetic markers regardless of their evolutionary origin.

The p23/α-crystallin domain (ACD) represent an ancient and conserved structure present in many eukaryotic chaperone families, including the small heat shock proteins (sHSP) of eukaryotes and prokaryotes, acting as molecular chaperone in diverse cellular pathways such as cell proliferation, differentiation, signalization, immunological modulation and pathogen resistance. However, little is known about the repertoires and functions of these proteins in parasitic protozoans. In the preset study, we also searched for ACD proteins including the sHSP in several genomes and found a single sHSP in the genomes of all trypanosomatids investigated. In addition, we uncovered a shared repertoire formed by eight additional ACD proteins in all Trypanosoma spp. regardless of the remarkable differences in hosts, life cycles, cell compartments and mechanisms employed to survive in stressful situations. Orthologues of all ACD proteins were also identified in Leishmania, several genera

of insect trypanosomatids, and Bodo spp. Seven identified ACD proteins (TryDYX1C1, TrySGT1, Tryp23A, Tryp23B, TryNudC1, TryNudC2, HSP20) are found in eukaryotes in general, whereas two (TryACDP, TryACD-TPR) were not previously identified in other organisms. The domain architecture and functional motifs of ACD proteins in trypanosomes were compared with those of ACD archetypes.

Additionally, this study includes the analyses of proline racemase (PRAC) and Cathepsin L-like (CATL) enzymes of Trypanosoma spp. The CATL enzymes are encoded by multiple gene families, and play important roles in infection processes, cell differentiation, pathogenicity, virulence, and host defense of Trypanosoma spp. Here, for the first time, congopain, the major CATL of the livestock pathogenic T. congolense, was characterized in isolates of the three genetically groups (Savannah, Forest, and Kilifi), which vary in virulence, pathogenicity and geographical distribution. The polymorphism, structural organization, and genealogy revealed that the repertoire of congopain genes diverged to become specific for each group of T. congolense, with Savannah exhibiting highly polymorphic sequences inter- and intra-isolates, whereas Forest and Kilifi groups showed moderate and limited diversity respectively. In addition, a PCR targeting congopain was developed for specific and sensitive diagnosis of all T. congolense groups. Findings from this study demonstrated that congopain genes are valuable targets for the diagnosis, genotyping, phylogenetic and taxonomic inferences of T. congolense.

The PRAC enzyme is responsible for the interconversion of free L- and D-proline in a restricted group of bacteria, playing a relevant role as a virulence factor. The first eukaryotic PRAC was described in T. cruzi and then in T. vivax, and proved to promote polyclonal activation of B cell in infected animals, then delaying host specific immunological response against the parasites, and ensuring parasite survival in vertebrate hosts. Herein, the genomic search of PRAC homologs revealed single copy PRAC genes in 12 of 15 Typanosoma spp. investigated: T. cruzi and T. cruzi marinkellei, T. dionisii, T. erneyi, T. rangeli, T. conorhini and T. lewisi, all parasites of mammals; trypanosomes from anurans, snakes, crocodiles, lizards and birds. T. rangeli possesses only PRAC pseudogenes, maybe in the process of being lost. Our findings corroborated that T. brucei, T. congolense and their allied species, except the more distantly related T. vivax, have completely lost PRAC genes. The genealogy of PRAC homologs supported an evolutionary history congruent with the Trypanosoma phylogeny. This finding, together with the synteny of PRAC loci, the relationships with prokaryotic PRAC inferred by taxon-rich phylogenetic analysis, and the absence in trypanosomatids of any other genera, bodonids and euglenids suggested that a common ancestor of Trypanosoma gained PRAC gene by a single and ancient HGT from a Firmicutes bacterium more closely related to Gemella and other Bacilli spp. than to Clostridium as previously suggested.

Keywords: *Trypanosoma*. Heat Shock Proteins. Cathepsin L. Comparative genomics. Horizontal Gene Transference. Phylogeny. Evolution.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

µg	Micrograma
µL	Microlitro
AToL	*Assembling the tree of life AToL*
BSA	Albumina bovina sérica
CATL	Catepsina L
CDD	Conseved Domain Database
dATP	Desoxiadenosina-trifosfato
dCTP	Desoxicitosina-trifosfato
dGTP	Desoxiguanosina-trifosfato
DMSO	Dimetil-sulfóxido
DNA	Ácido desoxiribonucleico
dTTP	Desoxitimidina-trifosfato
EDTA	Ácido etileno diamino tetracético
ES	*Expression Sites*
ESAG	*Expression Site Associated Genes*
gGAPDH	gliceraldeído-3-fosfato desidrogenase glicossomal
HGT	Horizontal gene Transfer
HSP	Proteínas de Choque Térmico
kDNA	DNA do cinetoplasto
LIT	Meio de cultura Infuso de fígado
M	Molar
Mb	Mega base
mg	Miligrama
min	Minutos
mL	Mililitro
mM	Milimolar
mRNA	Ácido ribonucléico mensageiro
N2	Nitrogênio líquido
PBS	Salina em tampão fosfato
PCR	Reação em cadeia da polimerase
PRAC	Prolina Racemase
PSSM	*Position-specific scoring matrix*
RNA	Ácido ribonucléico

LISTA DE ABREVIATURAS

µg	Microgranma
µL	Microlitro
AToL	*Assembling the tree of life AToL*
BSA	Albumina bovina sérica
CATL	Catepsina L
CDD	Conseved Domain Database
dATP	Desoxiadenosina-trifosfato
dCTP	Desoxicitosina-trifosfato
dGTP	Desoxiguanosina-trifosfato
DMSO	Dimetil-sulfóxido
DNA	Ácido desoxiribonucleico
dTTP	Desoxitimidina-trifosfato
EDTA	Ácido etileno diamino tetracético
ES	*Expression Sites*
ESAG	*Expression Site Associated Genes*
gGAPDH	gliceraldeído-3-fosfato desidrogenase glicossomal
HGT	Horizontal gene Transfer
HSP	Proteínas de Choque Térmico
kDNA	DNA do cinetoplasto
LIT	Meio de cultura Infuso de fígado
M	Molar
Mb	Mega base
mg	Miligrama
min	Minutos
mL	Mililitro
mM	Milimolar
mRNA	Ácido ribonucléico mensageiro
N2	Nitrogênio líquido
PBS	Salina em tampão fosfato
PCR	Reação em cadeia da polimerase
PRAC	Prolina Racemase
PSSM	*Position-specific scoring matrix*
RNA	Ácido ribonucléico

rRNA	Ácido ribonucléico ribossômico
Sarkosil	Lauril sarcosinato de sódio
SDS	Dodecil sulfato de sódio
SE	Solução salina Tris-EDTA
SFB	Soro fetal bovino
SRA	*Serum Resistance-Associated gene*
SSU	*Small Subunit ribossomal*
TAE	Tampão Tris acetato-EDTA
TCC	Coleção de culturas de tripanossomatídeos da Universidade de São Paulo
TE	Tampão tris-EDTA
UV	Luz ultravioleta
VSG	*Variant Surface Glycoprotein*

SUMÁRIO

1 INTRODUÇÃO

1.1 O filo Euglenozoa

O filo Euglenozoa (Eukaryota; Excavata) compreende eucariotos unicelulares flagelados que apresentam ampla diversidade morfológica e genética. O filo inclui tanto espécies de vida livre (fagotróficas, osmotróficas e fotossintéticas) quanto espécies comensais e parasitas obrigatórios e facultativos (Moreira et al., 2004, 2007; Simpson, Roger, 2004). Características morfológicas e inferências filogenéticas permitiram a classificação dos organismos pertencentes ao filo Euglenozoa em três classes: Kinetoplastea, Diplonemea e Euglenoidea (Busse, Preisfeld, 2002; Moreira et al., 2001, 2004). A recente inclusão de amostras obtidas de ambientes aquáticos com baixas concentrações de oxigênio revelou a existência de mais um grupo, nomeado como Symbiontida. Este é composto por flagelados de vida livre que possuem uma estreita relação com epibiontes bacterianos aderidos a matriz extracelular (Yubuki et al., 2009). Além da diversidade de espécies, o filo Euglenozoa destaca-se por singularidades que incluem desde a organização celular até a genômica (Simpson, Roger, 2004). Em geral, o filo Euglenozoa é marcado pela presença de maquinarias celulares e genômicas ditas "barrocas", que incluem: transcrição simultânea de dezenas de genes em tripanossomatídeos (Myler et al., 1999); edição pós-transcricional massiva do RNA mitocondrial em cinetoplastídeos (Estévez, Simpson, 1999); íntrons dentro de íntrons em euglenídeos (Thompson et al., 1997); presença de uma base hipermodificada (base J) no genoma nuclear dos tripanossomas (Borst, Sabatini, 2008); e trans-splicing entre o pré-mRNA e o gene *Spliced Leader*, que é encontrado ao longo de todo o filo Euglenozoa (Sturm et al., 2001).

Os euglenídeos (classe Euglenoidea) compreendem protozoários fototróficos, fagotróficos e osmotróficos (Breglia et al., 2007; Preisfeld et al., 2001). As espécies fototróficas, assim como algas verdes, apresentam cloroplastos com clorofila A e B nas membranas tilacóides (Nisbet, 1984). Apesar da semelhança fisiológica e estrutural entre euglenídeos e algas verdes restringir-se ao cloroplasto, a similaridade entre plastídios levou, no passado, à classificação artificial das espécies fotossintéticas como "algas verdes". A grande diferença morfológica entre euglenídeos e algas e a presença de uma membrana tripla envolvendo o cloroplasto de euglenídeos sugeriu sua origem a partir de um evento de endossimbiose secundária entre euglenídeos e algas verdes (Gibbs, 1981). A hipótese de endossimbiose secundária foi mais tarde apoiada por inferências filogenéticas obtidas a partir da subunidade menor do gene ribossômico (SSU rRNA) e genes codificantes de proteínas (Karnkowska et al., 2015; Linton et al., 1999; Sogin et al., 1986) que posicionaram os euglenídeos fotossintéticos próximos aos demais euglenídeos, diplonemídeos e cinetoplastídeos. O sequenciamento do genoma

plastidial de euglenídeos permitiu confirmar a hipótese da endossimbiose secundária (Hrdá et al., 2012; Turmel et al., 2009).

O grupo formado pelos diplonemídeos (classe Diplonemea) reune organismos biflagelados fagotróficos de vida livre, encontrados no sedimento e em superfícies submersas marinhas, e por parasitas facultativos de crustáceos (Kent et al., 1987; Lara et al., 2009). Os diplonemídeos são divididos em apenas dois gêneros, *Diplonema* e *Rhynchopus*, que se destacam dentro de Euglenozoa por não possuírem plastídios e cinetoplasto, este último encontrado nos cinetoplastídeos (Simpson, Roger, 2004). Estudos moleculares baseados na SSU rRNA sugeriram inicialmente que euglenídeos e diplonemídeos pudessem ser grupos irmãos (Preisfeld et al., 2001). No entanto, inferências baseadas em genes codificantes de proteínas de choque térmico (HSP70 e HSP90), e que incluem um número maior de *taxa*, contradizem o relacionamento inicial e indicam uma maior proximidade dos diplonemídeos com cinetoplastídeos (Simpson, Roger, 2004).

A classe Kinetoplastea, ordem Kinetoplastida, compreende as famílias Trypanosomatidae e Bodonidae. A principal característica dos cinetoplastídeos é a presença de uma única mitocôndria modificada que contém uma região rica em moléculas de DNA (kDNA) denominada cinetoplasto (Simpson et al., 2006; Stevens, 2008). O kDNA é constituído por um grande número moléculas de DNA dupla-fita e circulares, conhecidas como maxicírculos e minicírculos. Nas espécies da ordem Trypanosomatidae, essas moléculas estão concatenadas de maneira a formar uma rede única e densa, o que não é observado em Bodonidae (Simpson et al., 2002). Análises filogenéticas baseadas em 18S rRNA dividiram os cinetoplastídeos em duas subclasses: Prokinetoplastina, que compreende uma única ordem (Prokinetoplastida) e Metakinetoplastina, que é dividida em quatro ordens: Eubodonida, Parabodonida e Neobodonida, que correspondem aos bodonídeos; e Trypanosomatida, que compreende os tripanossomatídeos (Moreira et al., 2004). Os bodonídeos são em geral organismos biflagelados e de vida livre, exceto os dos gêneros *Cryptobia* e *Trypanoplasma,* que são parasitas de peixes. As espécies de vida livre são encontradas tanto no solo quanto em ambientes aquáticos, de água doce e salgada, e atuam como micropredadores que auxiliam no equilíbrio microbiológico destes ambientes (D'Avila-Levy et al., 2009). Os bodonídeos são divididos em vários gêneros, entre eles: *Ichthyobodo*, *Perkinsiella*, *Bodo*, *Neobodo*, *Parabodo*, *Trypanoplasma, Dimastigella*, *Rhynchobodo*, *Rhynchomonas* e *Cryptobia* (López-García et al., 2003; Moreira et al., 2004; von der Heyden, Cavalier-Smith, 2005).

1.2 A família Trypanosomatidae

A família Trypanosomatidae (ordem Kinetoplastidada) é constituída por parasitas obrigatórios de insetos, insetos fitófagos e plantas, e artrópodes hematófagos ou sanguessugas e vertebrados (Simpson et al., 2006; Stevens, 2008; Teixeira, 2010). O sucesso evolutivo dos tripanossomatídeos como parasitas é inquestionável, pois, depois dos nematóides, representam o grupo de eucariotos com a maior variedade de hospedeiros e distribuição geográfica (Simpson et al., 2006; Stevens et al., 2001). As espécies de tripanossomatídeos apresentam profundas diferenças biológicas, estruturais e fisiológicas entre si. O ciclo de vida desses organismos é complexo, com a presença de diversas formas que variam de acordo com a espécie e com a fase do ciclo de vida. A classificação das diferentes formas é feita em função da posição do cinetoplasto em relação ao núcleo e da presença ou não de flagelo livre e membrana ondulante. As principais formas encontradas nos tripanossomatídeos são: tripomastigota, amastigota, epimastigota, opistomastigota e promastigota. Com base na morfologia e em marcadores moleculares, a família Trypanosomatidae está dividida em 18 gêneros, sendo 14 compostos de parasitas monoxênicos de insetos (*Sergeia*, *Crithidia*, *Blastocrithidia*, *Wallacemonas*, *Leptomonas*, *Herpetomonas, Lafontella, Angomonas, Kentomonas, Strigomonas, Blechomonas, Jaenimonas, Lotmaria* e *Paratrypanosoma*), três heteroxênicos (*Endotrypanum*, *Leishmania, Trypanosoma*), que circulam entre artrópodes hematófagos e vertebrados, e um gênero heteroxênico (*Phytomonas*) encontrado em insetos fitófagos e plantas (Borghesan et al., 2013; Flegontov et al., 2013; Kostygov et al., 2014; Lukeš et al., 2014; Maslov et al., 2013; Merzlyak et al., 2001; Svobodova et al., 2007; Teixeira et al., 2011; Votýpka et al., 2013, 2014;). Alguns gêneros de tripanossomatídeos possuem uma estreita relação com endossimbiontes bacterianos, como é o caso de *Angomonas, Strigomonas* e *Kentomonas* (Teixeira et al., 2011; Votýpka et al., 2014). Partículas virais também foram descritas em associação com *Phytomonas, Angomonas* e *Leishmania* (Marche et al., 1993; Molyneux, 1974; Motta et al., 2003). Desse modo, incluindo o genoma nuclear e do cinetoplasto, podem estar presentes em alguns casos até quatro genomas em uma mesma célula.

Filogenias moleculares baseadas tanto em marcadores ribossomais quanto em genes codificantes suportam a estreita relação entre tripanossomatídeos e bodonídeos, não estando, porém, ainda bem definido o posicionamento filogenético dos dois grupos dentro dos cinetoplastídeos (Moreira et al., 2004; Simpson et al., 2006; Stevens, 2008). Um estudo recente com filogenias baseadas em 64 genes codificantes corroborou a posição dos tripanossomatídeos dentro da subclasse Metakinetoplastina, como grupo irmão da ordem Eubodonida, o que torna os bodonídeos per se um grupo parafilético (Moreira et al., 2004). A posição filogenética dos tripanossomatídeos entre os

bodonídeos parasitas e de vida livre levantou muitas questões sobre o aparecimento do parasitismo e da evolução dos grupos monoxênicos e heteroxênicos (Figura 1).

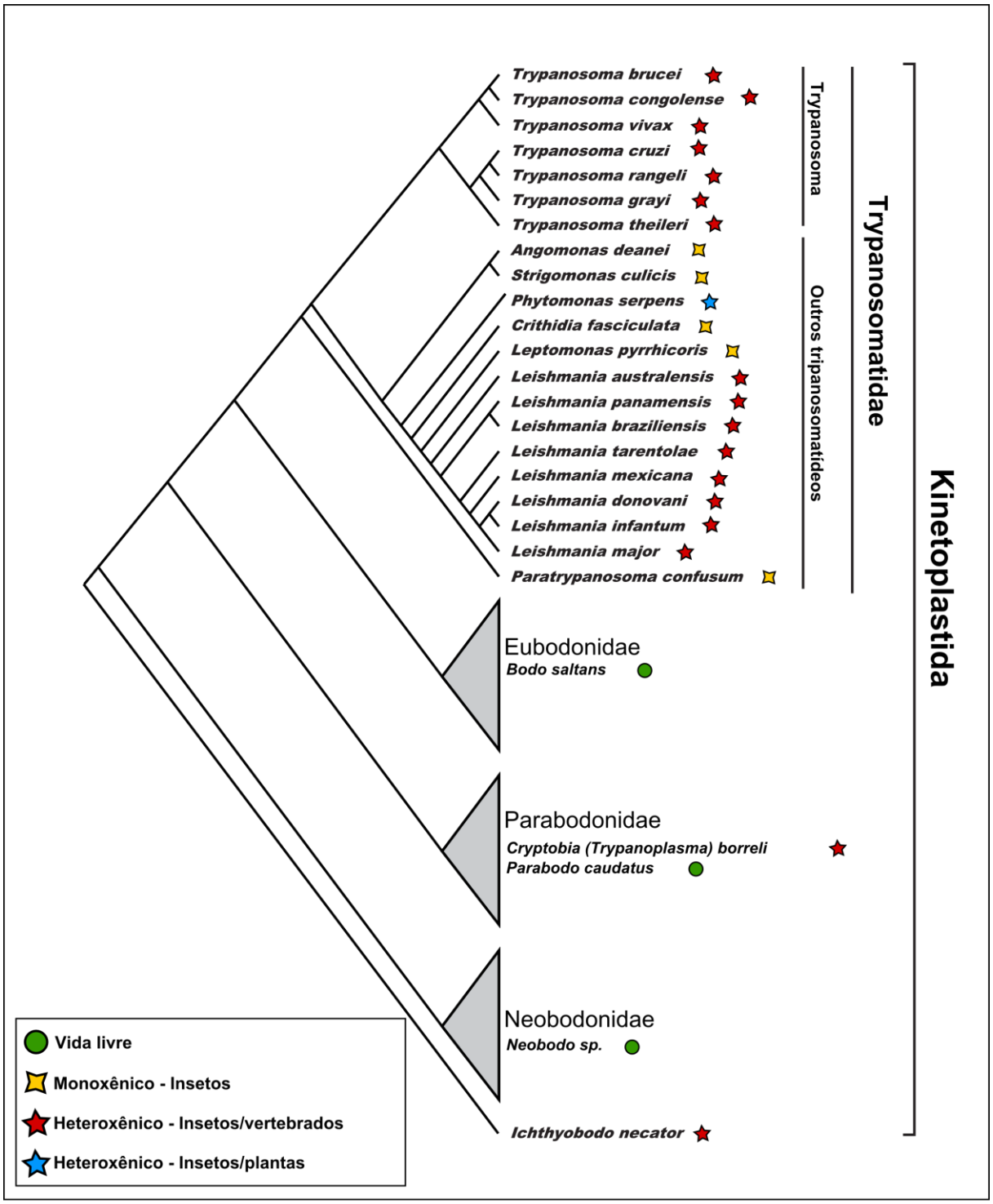

Figura 1 - Relações filogenéticas e variação do modo de vida entre os principais clados de cinetoplastídeos: Os triângulos representam um grande número de espécies condensadas em uma única ramificação, sendo algumas espécies do grupo nomeadas. Figura adaptada de Jackson, 2015.

A hipótese mais aceita defende o surgimento dos tripanossomatídeos a partir de um ancestral de vida livre e aquático, relacionado ao grupo atual dos eubodonídeos, sendo *Bodo saltans* a

espécie mais próxima conhecida (Deschamps et al., 2011). O ancestral de vida livre, ao ser ingerido por insetos, teria se adaptado ao ambiente intestinal, dando origem a um organismo monoxênico de insetos. Este, por sua vez, teria divergido dando origem aos diversos grupos de tripanossomatídeos monoxênicos de insetos. O padrão de ramificação filogenético dos gêneros que compõem a família Trypanosomatidae sugere que o salto do parasitismo monoxênico para heteroxênico ocorreu múltiplas vezes e em diferentes momentos na evolução da família (Figura1) (Stevens 2008; Stevens et al., 2001; Vickerman, 1994). Nesses saltos, ancestrais monoxênicos, ao colonizarem o trato digestório de insetos hematófagos e fitófagos, podem ter sido inoculados nos tecidos de vertebrados e plantas, adaptando-se e dando origem aos gêneros heteroxênicos (Teixeira, 2010). A recente descrição de um novo gênero aparentemente monoxênico, *Paratrypanosoma*, que se posiciona na base dos demais gêneros em inferências baseadas nos marcadores gGAPDH e SSU rRNA, reforça a hipótese de múltiplos saltos para o modo de vida heteroxênico (Flegontov et al., 2013).

1.3 O gênero *Trypanosoma*

O gênero *Trypanosoma* é composto por espécies que alternam o ciclo de vida entre hospedeiros invertebrados e vertebrados. A diversidade dos tripanossomas está refletida em ambos hospedeiros. Conforme o tripanossoma, os vetores podem ser dípteros, hemípteros, pulgas, carrapatos e sanguessugas, e os hospedeiros vertebrados podendo compreender aves, répteis, anfíbios, peixes e mamíferos (Hamilton et al., 2007; Hoare, 1972; Simpson et al., 2006; Stevens, 2008). A maior parte das espécies se desenvolve somente ciclicamente no vetor. A exceção ocorre em *Trypanosoma evansi, Trypanosoma equiperdum e Trypanosoma vivax.* As duas primeiras são somente transmitidas mecanicamente por insetos hematófagos e pelo coito respectivamente, enquanto *T. vivax* é transmitido tanto pela via mecânica por moscas hematófagas quanto ciclicamente pela mosca tsé-tsé. A adaptação à transmissão mecânica é responsável pela disseminação destas espécies para fora do cinturão africano de ocorrência da mosca tsé-tsé, considerada vetor natural desses tripanossomas. A transmissão exclusivamente mecânica observada em *T. evansi* e em *T. equiperdum* é possivelmente responsável pela perda parcial ou total do cinetoplasto, essencial para o desenvolvimento cíclico no inseto vetor (Lun et al., 2010). A maior parte dos estudos sobre os tripanossomas está restrita a espécies patogênicas que possuem interesse médico humano e veterinário. A ampla distribuição geográfica, variedade de nichos ecológicos, ecótopos e mecanismos de transmissão das espécies estudadas sugerem que a diversidade ainda é subestimada.

Até o momento, são conhecidos três tripanossomas naturalmente infectantes para o homem: *Trypanosoma cruzi, Trypanosoma brucei,* que inclui duas subespécies – *T. b. gambiense, T. b. rhodesiense*, e *Trypanosoma rangeli*. As duas primeiras espécies são respectivamente os agentes

etiológicos da doença de Chagas e do sono, enquanto *T. rangeli*, embora capaz de infectar humanos, até agora não é considerado patogênico. Por sua vez, *T. vivax*, *T. evansi*, *T. equiperdum*, *Trypanosoma congolense*, *T. brucei brucei* e *Trypanosoma simiae* são agentes de tripanossomíases animais. Essas espécies são todas de origem africana e patogênicas para bovinos, bufalinos, ovinos, caprinos, equinos e suínos, sendo responsáveis por perdas econômicas em países da África, Ásia e Américas Central e do Sul (Morrison, MacLeod, 2011). A maior parte dos tripanossomas circula no ambiente silvestre como enzootias, transmitidos por vetores associados com seus hospedeiros e respectivos ecótopos. Recentemente, junto ao surgimento de técnicas de diagnóstico molecular tem sido reportado um aumento no número infecções humanas atípicas por tripanossomas, como *Trypanosoma lewisi, T. vivax, T. evansi, T. congolense e T. brucei brucei* (Truc et al., 2013, 2014). Apesar dos casos serem até agora restritos a indivíduos imunossuprimidos ou desnutridos, a possibilidade do aparecimento de doenças emergentes tem redirecionado o foco dos novos estudos para espécies antes negligenciadas (Truc et al., 2013).

Ao longo do ciclo de vida, os tripanossomas podem apresentar-se sob as formas epimastigota, tripomastigota (sanguíneo ou metacíclico) e, em alguns casos, amastigota. As formas infectantes presentes nos vertebrados são denominadas tripomastigotas sanguíneos. Estas, quando ingeridas pelos vetores durante o repasto sanguíneo, sofrem alterações morfológicas, bioquímicas e fisiológicas, diferenciando-se em epimastigotas, que são as formas multiplicativas adaptadas ao novo ambiente. Uma vez cessada a fase multiplicativa, as formas epimastigotas diferenciam-se em tripomastigotas metacíclicos, que são as formas infectantes para o hospedeiro vertebrado. O local onde se dá a metaciclogênese no vetor determina a via de transmissão, se contaminativa (caso das espécies de *Trypanosoma* cujos tripomastigotas metacíclicos desenvolvem-se nas porções posteriores do tubo digestório e são eliminadas juntamente com as fezes do vetor) ou inoculativa (metaciclogênese nas porções anteriores e glândulas salivares do vetor e consequente inoculação com a saliva), durante o repasto sanguíneo (Hoare, 1972). Com exceção de *T. rangeli,* transmitido por ambas as vias, as demais espécies utilizam apenas uma das estratégias de transmissão (Gilles, 1999). Com base no desenvolvimento no vetor e, consequentemente, na via de eliminação das formas infectantes, os subgêneros que infectam mamíferos foram separados em Secções Stercoraria e Salivaria (Hoare, 1964). A taxonomia das espécies do gênero *Trypanosoma* era tradicionalmente baseada em caracteres morfológicos existentes, nas formas encontradas no sangue de vertebrados e também no inseto vetor. Além da morfologia, o comportamento biológico, hospedeiro e localização geográfica também eram considerados na classificação. Centenas de espécies foram descritas em mamíferos e classificadas em oito subgêneros com base nesses critérios (Hoare, 1972).

A Secção Salivaria compreende apenas espécies de origem africana, a maioria transmitida ciclicamente por moscas tsé-tsé (gênero *Glossina*). O local de desenvolvimento dentro da mosca varia conforme a espécie, ocorrendo nas porções anteriores do tubo digestivo e glândulas salivares (*T. brucei* spp.), apenas na porção anterior do tubo digestivo (*T. congolense)*, ou restrito à probóscide no caso de *T. vivax* (Hamilton et al., 2007; Hoare, 1972; Stevens, 2008; Stevens et al., 2001). A Secção Stercoraria compreende três subgêneros: *Schizotrypanum* cuja espécie-tipo é *T. cruzi; Herpetosoma (*espécie-tipo *T. lewisi*) e *Megatrypanum* (espécie-tipo *Trypanosoma theileri*) (Hoare, 1964, 1972). As espécies pertencentes a esta Secção se desenvolvem no tubo digestivo do inseto vetor e são transmitidas pela contaminação com as fezes dos vetores deixadas durante o repasto sanguíneo. *T. lewisi* e *T. theileri* apresentam restrição quanto aos hospedeiros vertebrados (roedores e ungulados respectivamente) e possuem ampla distribuição geográfica. Por outro lado, *T. cruzi* e *T. rangeli* infectam uma grande diversidade de mamíferos e são restritos às Américas Central e do Sul (Guhl, Vallejo, 2003; Hoare, 1972).

Embora a monofilia do gênero *Trypanosoma* tenha sido contestada em estudos iniciais baseados em SSU rRNA (Maslov et al., 1996), trabalhos que incluem um número maior de *taxa* e adicionam outros marcadores confirmam o gênero como monofilético (Hamilton et al., 2004; Leonard et al., 2011; Simpson et al., 2006; Stevens et al., 2001). A visão atual das relações filogenéticas entre as espécies de tripanossomas é tradicionalmente baseada nos genes de SSU rRNA e gGAPDH (Hamilton et al., 2004, 2007; Hannaert et al., 1998; Stevens, 2008). O gênero *Trypanosoma* possui uma divisão em duas linhagens filogenéticas conhecidas como clados aquático e terrestre (Figura 2) (Hamilton et al., 2004, 2007; Stevens 2008; Stevens et al., 2001). O clado aquático compreende tripanossomas isolados de sanguessugas e vertebrados de hábitos aquáticos ou semiaquáticos, principalmente peixes de água doce e salgada e anfíbios. Também foram encontrados isolados de tartaruga, camaleão e ornitorrinco que pertencem a esse grupo (Dvořáková et al., 2015; Hamilton et al., 2007; Paparini et al., 2014). As espécies do clado aquático são transmitidas principalmente por sanguessugas aquáticas; contudo, flebotomíneos foram relatados como vetores de tripanossomas de anuros (Ferreira et al., 2008). Por sua vez, os tripanossomas de vertebrados terrestres são isolados de mamíferos, aves, crocodilianos, cobras e lagartos e são transmitidos por uma ampla diversidade de vetores que incluem carrapatos, moscas, mosquitos, hemípteros e pulgas (Hamilton et al., 2007). Inferências filogenéticas baseadas nos marcadores tradicionais e incluindo diferentes conjuntos de *taxa* revelam a existência vários clados entre as espécies da linhagem terrestre (Figura 2): os clados *T. brucei; T. cruzi; T. lewisi; T. theileri; T. rangeli-T. conorhini; T. avium; T. corvi* e clados de tripanossomas de crocodilianos e de

lagartos e serpentes (Cavazzana et al., 2010; Hamilton et al., 2004, 2005, 2007, 2008, 2009; Lima et al., 2012b; Maia da Silva et al., 2010; Rodrigues et al., 2006; Stevens, 2008; Stevens et al., 2001).

O clado *T. brucei* compreende uma grande diversidade de tripanossomas de origem africana, pertencentes à Secção Salivaria e naturalmente transmitidos pela mosca tsé-tsé (Adams et al., 2010; Hamilton et al., 2007; Hoare, 1972). As espécies desse clado são classificadas em quatro subgêneros: *Trypanozoon* (*T. brucei* sspp*. T. evansi* e *T. equiperdum), Duttonella* (*T. vivax); Pycnomonas* (*T. suis)* e *Nannomonas* (*T. congolense e T. simiae*) (Hoare, 1972). O clado *T. theileri* corresponde ao subgênero *Megatrypanum*, sendo formado por tripanossomas restritos a ungulados (ordem Artiodactyla). *T. theileri* é cosmopolita, não patogênico ao hospedeiro vertebrado e transmitido por moscas hematófagas da família Tabanidae (Garcia et al., 2011; Rodrigues et al., 2006, 2010). O clado *T. lewisi* corresponde ao subgênero *Herpetosoma*, que compreende espécies cosmopolitas encontradas principalmente em roedores silvestres e domésticos (Hoare, 1972; Maia da Silva et al., 2010). Inicialmente, com base na morfologia das formas sanguíneas, *Herpetosoma* incluía também *T. rangeli*. No entanto, inferências filogenéticas baseadas em diversos marcadores moleculares permitiram a revisão desse subgênero que agora compreende apenas as espécies do clado *T. lewisi* (Hamilton et al., 2007; Maia Da Silva et al., 2007; Stevens, Gibson, 1999). O clado *T. cruzi* compreende o subgênero *Schizotrypanum, o clado T. rangeli-T. conohrini, Trypanosoma livingstonei* (Lima et al., 2013)*, Trypanosoma wauwau* (Lima et al., 2015), isolados de morcegos (*Trypanosoma* sp. Bat) e carnívoros (T*rypanosoma* sp. NanDoum1) da África (Cavazzana et al., 2010; Hamilton et al., 2009; Stevens et al., 2001), e de marsupiais australianos incluindo um isolado de canguru (*Trypanosoma* sp. H25) (Stevens et al., 1999) (Figura 2).

Todas as espécies do clado *T. cruzi* desenvolvem-se nas porções posteriores do trato digestório do inseto vetor e pertencem à Secção Stercoraria. O subgênero *Schizotrypanum* inclui *T. cruzi*, *T. cruzi marinkellei, Trypanosoma dionisii, Trypanosoma erneyi, Trypanosoma vespertilionis, Trypanosoma hastatus, Trypanosoma hedricki, Trypanosoma myoti* e *Trypanosoma phyllostomus*. Embora várias espécies tenham sido descritas nesse clado, somente *T. c. marinkellei T. dionisii* e *T. erneyi* foram confirmadas por marcadores moleculares (Cavazzana et al., 2010). As espécies de *Schizotrypanum* apresentam forte restrição quanto ao hospedeiro vertebrado, sendo encontradas somente em morcegos, exceto *T. cruzi* que naturalmente infecta mamíferos de diversas ordens (Cavazzana et al., 2010; Hoare 1972; Marcili et al., 2009; Molyneux, 1991).

O clado *T. rangeli-T. conorhini* posiciona-se como grupo irmão do subgênero *Schizotrypanum* e compreende *T. rangeli, T. conorhini*, *Trypanosoma* sp. Bat e *Trypanosoma* sp. NanDoum1 (Hamilton et al., 2009). Basal à ramificação formada por *T. rangeli, T. conorhini* e *Schizotrypanum* posiciona-se o

grupo formado por isolados de marsupiais australianos e *T. wauwau*, uma nova espécie encontrada em morcegos da América do Sul (Lima et al., 2015). *T. livingstonei,* parasita de morcegos africanos, representa a espécie mais basal pertencente ao clado *T. cruzi* (Lima et al., 2013).

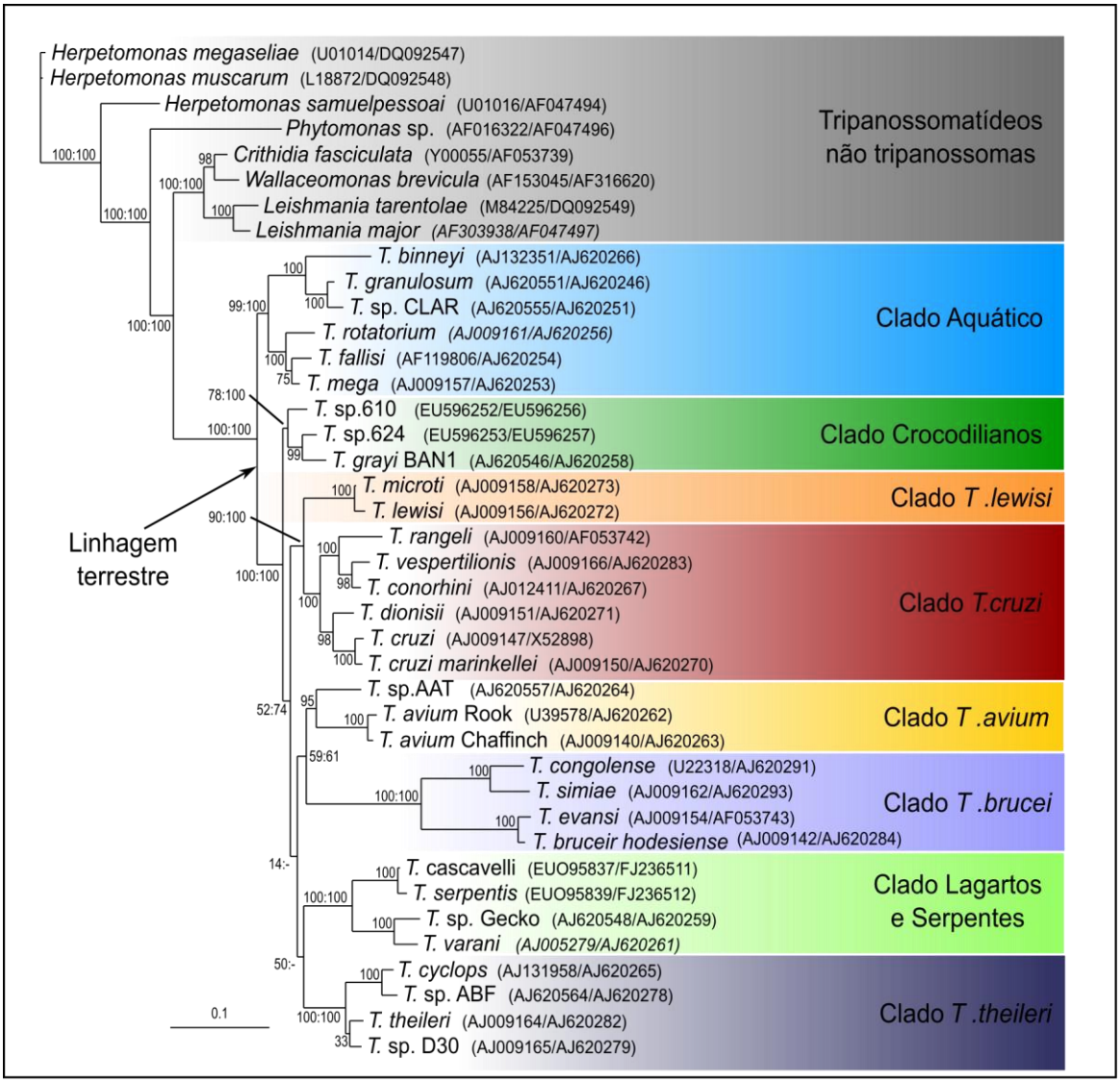

Figura 2 - Inferência filogenética do gênero Trypanosoma. Análise concatenada dos genes SSU rRNA e gGAPDH incluindo 32 espécies de tripanosomas que representam os principais clados. Figura adaptada de Viola et al., 2009.

A presença de tripanossomas restritos a morcegos da Europa, África, América em vários pontos da ramificação do clado *T. cruzi* sugere que as espécies desse grupo evoluíram de um tripanossoma ancestral exclusivo de morcegos (Figura 3). Ao longo da evolução e expansão dos morcegos, esse ancestral divergiu, originando as espécies atuais, que foram dispersas junto dos morcegos através dos continentes. A dispersão dessas espécies restritas a morcegos foi acompanhada de saltos desses tripanossomas para outros mamíferos. Essa hipótese para o surgimento do clado *T. cruzi* ficou conhecida como *bat seeding* (Hamilton et al., 2012b). O padrão de ramificação filogenético das espécies encontradas em mamíferos não morcegos sugere que o salto entre hospedeiros ocorreu

de maneira independente e inúmeras vezes ao longo desse processo (Figura 3) (Hamilton et al., 2012a, 2012b; Lima et al., 2012b, 2013, 2015).

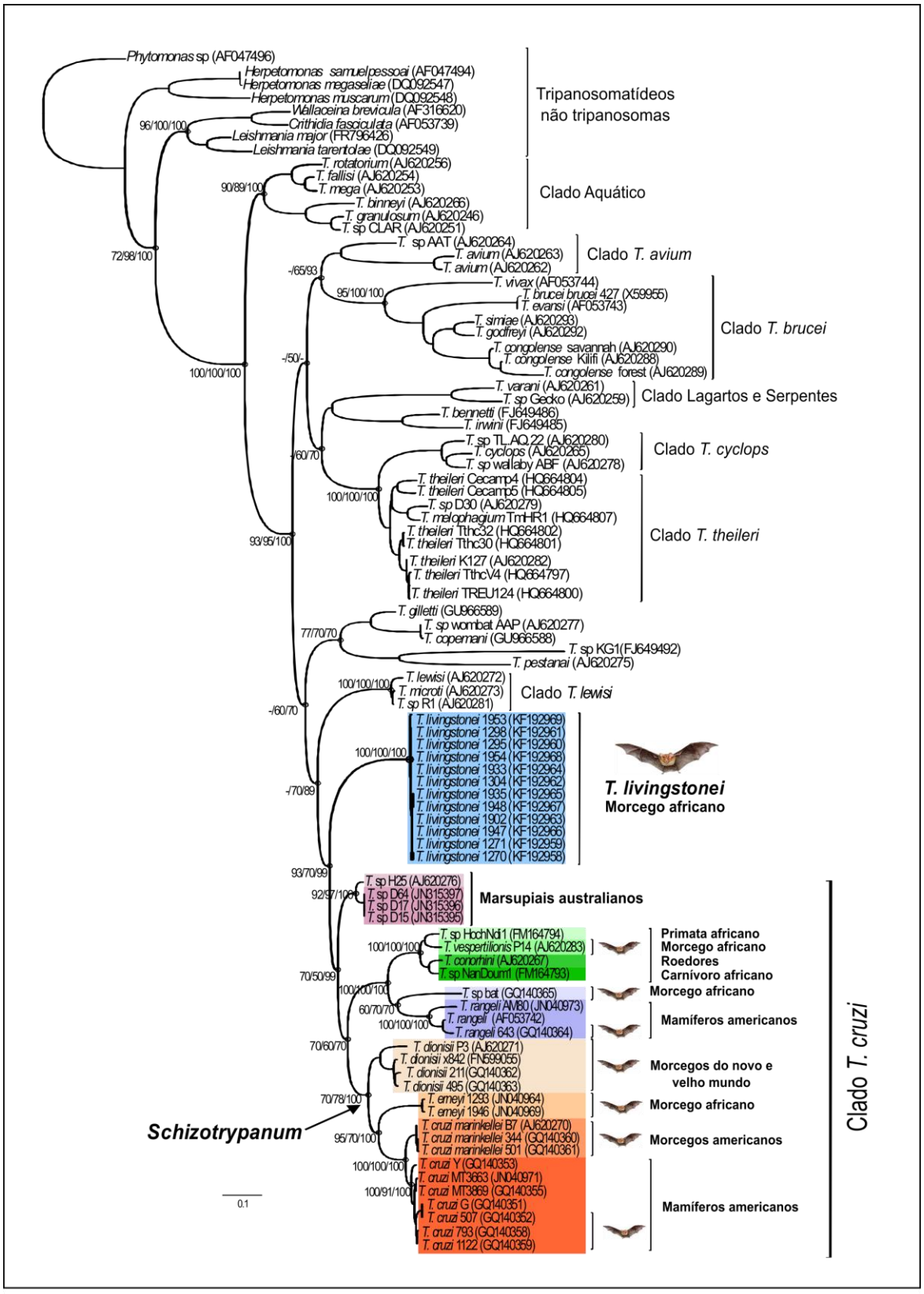

Figura 3: Inferência filogenética do gênero Trypanosoma com ênfase no clado T. cruzi. Análise baseada no gene gGAPDH. Figura adaptada de Lima et al., 2013.

Dentro da linhagem terrestre, os tripanossomas de vertebrados não mamíferos compreendem três clados: tripanossomas de crocodilianos, de aves e de lagartos e serpentes (Hamilton et al., 2007). Embora muitas espécies de tripanossomas tenham sido descritas infectando esses hospedeiros, poucas foram confirmadas por técnicas moleculares e posicionadas em filogenias, o que faz deste grupo de tripanossomas o menos estudado. Além disso, os critérios taxonômicos tradicionais, como morfologia das formas sanguíneas e em cultura, espécie do hospedeiro de origem e localização geográfica se mostraram pouco confiáveis para o gênero *Trypanosoma.* Isto ocorre pela quantidade de infecções mistas e pela variabilidade de restrição ao hospedeiro e distribuição geográfica entre as espécies de tripanossomas.

O clado dos tripanossomas de crocodilianos é formado por três espécies: *Trypanosoma grayi,* de crocodilos africanos e transmitido pela mosca tsé-tsé, *Trypanosoma ralphi* e *Trypanosoma terena,* ambos isolados de jacarés da América do Sul e cujo vetor ainda é desconhecido (Fermino et al., 2013). Inferências atuais baseadas nos genes de gGAPDH e SSU rRNA e que incluem as três espécies posicionam esse clado na base da linhagem terrestre (Fermino et al., 2013, 2015). Recentemente, outro estudo mostrou que *Trypanosoma clandestinus,* encontrado em jacarés e possivelmente transmitido por sanguessugas, está mais relacionado com as espécies do clado aquático do que com as do clado de tripanossomas de crocodilianos (Fermino et al., 2015). O clado dos tripanossomas de aves apresenta ao menos três subclados: *T. corvi*, *T. avium* e *T. benetti* (Hamilton et al., 2007; Votýpka et al., 2002). Embora as relações filogenéticas dentro de cada subclado pareçam consistentes usando os marcadores tradicionais, as relações entre os subclados ainda não estão claras, sendo possível que não formem um clado monofilético (Hamilton et al., 2007). Por último, o clado dos tripanossomas de lagartos e serpentes agrupa espécies obtidas de vertebrados da ordem Squamata (lagartos e serpentes) (Viola et al., 2009). Até o momento, as posições filogenéticas dos subclados de tripanossomas de aves e do clado de tripanossomas de lagartos e serpentes não estão bem estabelecidas dentro do gênero *Trypanosoma*.

1.4 Evolução genômica em tripanossomas e outros tripanossomatídeos

Ao longo da última década, o sequenciamento de nova geração, seguido pela diminuição gradual dos custos, revolucionou o estudo da parasitologia. O surgimento dessa nova tecnologia ampliou o conhecimento sobre as interações entre parasita e hospedeiro e, consequentemente, mudou a visão sobre os mecanismos de infecção, invasão, escape imunológico, evolução, resistência a drogas e desenvolvimento de novos alvos terapêuticos (Lv et al., 2015). Filogenias, antes limitadas pela escassez de caracteres, agora contam com um vasto repertório de marcadores moleculares, essenciais para estudos micro e macroevolutivos (Jackson 2015b). Mesmo quando o objetivo não é

diretamente filogenético, como o estudo funcional de genes, a possibilidade de comparar genes e vias metabólicas entre espécies relacionadas torna o pensamento evolutivo necessário.

Parasitas tendem a compartilhar alguns fenômenos na evolução de seus genomas, que conforme a espécie, possuem um papel maior ou menor na sobrevivência dos organismos (Jackson, 2015b). Entre esses, pode se citar a redução do conteúdo genômico, aparecimento de genes espécie específicos, recrutamento de genes para novas funções e transferências horizontal de genes (Capewell et al., 2015; Jackson, 2015b; Lv et al., 2015; Wijayawardena et al., 2013). Onze anos após a publicação dos genomas de *T. cruzi* (El-Sayed et al., 2005a), *T. brucei* (Berriman et al., 2005) *e Leishmania major* (Ivens et al., 2005), os genomas nucleares de diversas outras espécies e subespécies foram sequenciados, estando disponíveis em bancos públicos (Tabela 1). Embora a maioria dos genomas seja de espécies com importância médica humana e veterinária, um esforço recente tem sido feito para aumentar a representatividade e preencher as lacunas evolutivas entre as espécies patogênicas. Entre os genomas recém-sequenciados de tripanossomatídeos considerados não patogênicos para o homem e animais domésticos estão os de *Crithidia fasciculata, Endotrypanhum monterogeii, T. grayi, T. cruzi marinkellei* e *T. rangeli* (Kelly et al., 2014; Schwarz et al., 2015; Stoco et al., 2014). A comparação entre os genomas dos tripanossomatídeos revelou uma série de singularidades, tais como a conservação da sintenia de grande parte do genoma, mesmo entre espécies com grande tempo de divergência, um arsenal complexo de proteínas de superfície e uma grande quantidade de genes que vieram de transferência horizontal (Ghedin et al., 2004; Jackson, 2015a; Opperdoes, Michels, 2007).

Tabela 1 - Informações gerais sobre o genoma de organismos parasitas. Adaptado de Lv et al., 2015.

Taxon	Tamanho (Mb)	Cromossomos	Conteúdo GC%	Genes
Trematoídes				
Clonorchis sinensis	516	56	44.8	13634
Opisthorchis viverrini	634.5	12	43.7	16379
Schistosoma haematobium	385	16	34.3	13073
Schistosoma japonicum	398	16	34.1	13469
Schistosoma mansoni	363	16	35.3	11809
Nematoídes				
Brugia malayi	93.7	10	30.5	11515
Loa loa	91.4	12	31	14907
Necator americanus	244	12	40.2	19151
Trichinella spiralis	64	12	33.9	15808
Trichuris suis	80	–	43.9	14781
Cestoídes				
Echinococcus granulosus	151.6	18	42.1	11325
Echinococcus multilocularis	115	18	42.2	10345

Taenia solium	122.3	–	42.9	12490
Protozoários				
Babesia bovis	8.2	6	41.8	2228
Babesia microti	6.5	3	36	3500
Cryptosporidium hominis	9.2	8	31.7	3994
Eimeria falciformis	44	14	52.9	5879
Entamoeba histolytica	23.75	–	22.4	9938
Giardia intestinalis	11.7	10	49	4889
Leishmania braziliensis	32	35	57.76	8314
Leishmania donovani	32.4	36	59.06	8195
Leishmania major	32.8	36	62.5	8272
Plasmodium falciparum	23.3	14	19.4	5403
Plasmodium vivax	26.8	14	42.28	5433
Trichomonas vaginalis	176.44	6	32.7	59681
Trypanosoma brucei	26.08	22	50.9	9068
Trypanosoma cruzi	55	28	50.9	12000
Microsporídeos				
Encephalitozoon intestinalis	2.3	–	–	1833
Enterocytozoon bieneusi	6	–	–	3804
Encephalitozoon cuniculi	2.9	–	–	1999

1.5 Estrutura e organização do genoma em tripanossomatídeos

Em geral, ao se adaptar a sobreviver do espólio de outro organismo, os parasitas sofrem um processo de simplificação fisiológica e fenotípica, decorrente da redução nas pressões seletivas que atuam sobre as características, antes necessárias ao modo de vida livre. Esse processo leva a perda massiva de genes em vias metabólicas e regulatórias, o que cria uma dependência fisiológica do parasita ao seu hospedeiro. Por exemplo, *Cryptosporidium hominis* precisa obter a maior parte dos aminoácidos e nucleotídeos da célula hospedeira, pois não é mais capaz de sintetizá-los (Xu et al., 2004). Desse modo, espera-se que a redução no número de genes seja refletida no tamanho do genoma. No caso extremo de *Encephalitozoon intestinalis,* o genoma foi reduzido a apenas 2,3 Mpb, que contém o mínimo necessário para manter a homeostase celular (Corradi, Slamovits, 2011). O número crescente de parasitas com o genoma sequenciado tem revelado que, embora exista a tendência à redução do genoma, esta característica não é ubíqua entre parasitas (Tabela 2).

Antes da disponibilidade dos genomas, era conhecido que os tripanossomatídeos apresentavam algumas deficiências metabólicas. A publicação dos primeiros genomas confirmou que

as limitações são decorrentes da perda de genes, tais como os envolvidos na biossíntese do heme, de purinas e de tetrahidrobiopterina (Korený et al., 2010; Marr et al., 1978; Ouellette et al., 2002;).

Tabela 2 - Genomas de tripanossomatídeos disponíveis em bancos públicos.

Taxon	Cepa	Número de acesso
Herpetomonas muscarum	TCC 001E	GenBank AUXJ01000000
Phytomonas serpens	P9	GenBank AIHY00000000.1
Phytomonas sp.	HART1	GenBank CAVR000000000.2
Phytomonas sp.	EM1	GenBank CAVQ000000000.1
Angomonas deanei	TCC 036E	GenBank AUXM01000000
Angomonas desouzai	TCC 079E	GenBank AUXL01000000
Strigomonas culicis	TCC 012E	GenBank AUXH01000000
Strigomonas galati	TCC 219	GenBank GCA_000482125
Strigomonas oncopelti	TCC 290E	GenBank AUXK01000000
Leptomonas pyrrhocoris	H10	TriTrypDB
Leptomonas seymouri	ATCC 30220	TriTrypDB
Lotmaria passim	SF	GeneBank AHIJ00000000
Crithidia fasciculata	Cf-CI	TriTrypDB
Endotrypanum monterogeii	LV88	TriTrypDB
Leishmania enrietti	LEM3045	TriTrypDB
Leishmania martiniquensis	MARLEM2494	TriTrypDB
Leishmania (Viannia) braziliensis	MHOMBR75M2904	TriTrypDB
Leishmania (Viannia) panamensis	MHOM/COL/81/L13	TriTrypDB
Leishmania (Leishmania) aethiopica	L147	TriTrypDB
Leishmania (Leishmania) amazonensis	MHOM/BR/71973/M2269	TriTrypDB
Leishmania (Leishmania) arabica	LEM1108	TriTrypDB
Leishmania (Leishmania) donovani	BPK282A1	TriTrypDB
Leishmania (Leishmania) gerbilli	LEM452	TriTrypDB
Leishmania (Leishmania) infantum	JPCM5	TriTrypDB
Leishmania (Leishmania) major	Friedlin	TriTrypDB
Leishmania (Leishmania) mexicana	MHOM/GT/2001/U1103	TriTrypDB
Leishmania (Leishmania) tropica	L590	TriTrypDB
Leishmania (Leishmania) turanica	LEM423	TriTrypDB
Leishmania (Sauroleishmania) tarentolae	ParrotTarII	TriTrypDB
Trypanosoma evansi	STIB 805	TriTrypDB
Trypanosoma brucei brucei	Lister 427	TriTrypDB
Trypanosoma brucei brucei	TREU927	TriTrypDB
Trypanosoma brucei gambiense	DAL972	TriTrypDB
Trypanosoma congolense	IL30000	TriTrypDB
Trypanosoma grayi	ANR4	TriTrypDB
Trypanosoma vivax	Y486	TriTrypDB
Trypanosoma rangeli	SC58	TriTrypDB
Trypanosoma cruzi cruzi	Dm28c	TriTrypDB
Trypanosoma cruzi cruzi	CL Brener	TriTrypDB

Trypanosoma cruzi cruzi	Sylvio X10.6	TriTrypDB
Trypanosoma cruzi marinkellei	B7	TriTrypDB

Assim como em outros parasitas, essas deficiências são supridas pelos hospedeiros e, em alguns casos, também por endossimbiontes bacterianos (Jackson, 2015a; Klein et al., 2013; Korený et al., 2010). Embora os tripanossomatídeos apresentem perdas de vários genes, o genoma haploide possui entre 25 a 35 Mb e densidade de genes variando entre 2,8 a 4,6 Kb/gene, o que é semelhante ao encontrado em organismos de vida livre, como *Saccharomyces cerevisiae* (12,5 Mb e 2,9 Kb/gene) e *Bodo saltans* (39,9 Mb e 2,1 Kb/gene) (El-Sayed et al., 2005b; Jackson, 2015a; Jackson et al., 2016). Em contraponto à perda de genes, os tripanossomatídeos desenvolveram algumas novidades metabólicas, como novos mecanismos de controle da homeostase redox, um sistema não canônico de transporte mitocondrial, ausência de proteínas ligantes nos extremos teloméricos, expansões de algumas famílias multigênicas por duplicação e aquisição de novos genes a partir de transferências horizontais (Krauth-Siegel, Comini, 2008; Jackson 2015a; Lira et al., 2007; Pusnik et al., 2009).

1.6 Organização sintênica em tripanossomatídeos

A comparação entre os genomas de diversas espécies de tripanossomatídeos revelou uma organização sintênica de genes e sob uma mesma orientação. Essa característica está presente nas regiões centrais de praticamente todos os cromossomos. A sintenia dos genes é observada em todas as espécies estudadas até o momento e inclui também *Bodo saltans,* o representante de vida livre filogeneticamente mais relacionado à família Trypanosomatidae (Figura 4) (Jackson et al., 2016).

A conservação da sintenia, mesmo após um longo tempo de divergência, está provavelmente relacionada a altas pressões seletivas (Ghedin et al., 2004). Estudos abrangentes de genômica comparativa mostram uma forte tendência à proximidade e conservação da ordem entre genes relacionados funcional ou fisiologicamente (Dávila López et al., 2010) No entanto, não há evidências suficientes de que esse fenômeno seja a única causa da sintenia extrema em tripanossomatídeos. A orientação única dentro dos blocos sintênicos é provavelmente fruto da pressão seletiva imposta pela transcrição policistrônica que ocorre nesses organismos; no entanto, a aquisição dessa característica dentro dos cinetoplastídeos permanece inexplicada (Jackson, 2015a). Enquanto a sintenia apresenta-se mais conservada no interior dos cromossomos, as regiões subteloméricas são pouco conservadas e concentram as famílias multigênicas especializadas que exibem altas taxas evolutivas e envolvidas no escape imunológico (El-Sayed et al., 2005b). Essas regiões ficaram conhecidas como zonas de contingência.

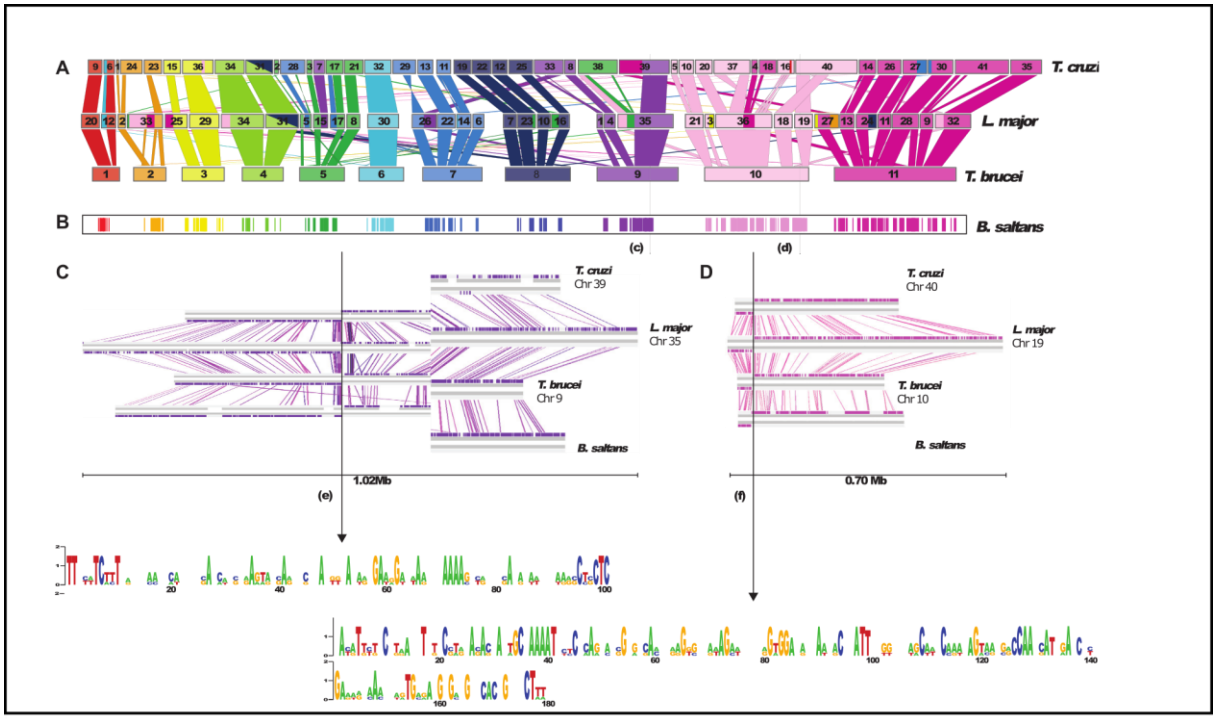

Figura 4 - Representação esquemática da organização genômica sintênica conservada ao longo de toda a família Trypanosomatidae e em Bodo saltans. A conservação da origem da replicação do bloco policistrônico é mostrada pelas setas. Adaptado de Jackson et. al., 2016.

1.7 Repertório de genes especializados

A existência de genes espécie-específicos, geralmente implicados na interação com o hospedeiro, é uma característica comum a vários grupos de parasitas, como tripanossomatídeos, oomicetos, cestoides e plasmódios (Adhikari et al., 2013; Carlton et al., 2008; Jackson et al., 2012, 2013; Tsai et al., 2013; Tyler et al., 2006). A duplicação gênica é um dos principais mecanismos que permitem o surgimento de inovações (Zhang, 2003). Segundo a teoria de nascimento e morte de genes, uma vez duplicadas, as cópias parálogas podem adquirir novas funções (neofuncionalização) ou funções semelhantes sob condições diferentes (subfuncionalização) ou degenerar, tornando-se pseudogenes (Lynch, Conery, 2000). Em tripanossomatídeos, a presença de múltiplas cópias dispostas em tandem é comum em várias famílias multigênicas, como proteínas de choque térmico e cisteína proteases (Folgueira, Requena, 2007; Sajid, McKerrow, 2002). Esse fenômeno é possivelmente decorrente da expressão policistrônica, na qual o número de cópias pode servir como controle da quantidade de mRNA transcrito (Jackson, 2015a).

Em geral, as cópias em tandem de um mesmo gene tendem a evoluir em concerto, de modo a exibirem pouco polimorfismo entre elas. Isso ocorre devido à recombinação não homóloga entre as cópias, processo conhecido como conversão gênica (Jackson, 2007). Embora prevalente no genoma dos tripanossomatídeos, a evolução em concerto não impede o surgimento de parálogos

especializados. Isso ocorre devido à quebra na estrutura de repetição em tandem por rearranjos cromossômicos e também pelo fato de que o processo de recombinação, além de permitir a evolução em concerto, também é capaz de promover diversificação entre cópias, que podem ser selecionadas de modo a apresentar novas funções (Jackson, 2007). A prolina racemase de *T. cruzi* é um exemplo de especialização funcional entre cópias, uma vez que TcPRACA e TcPRACB são expressas em etapas diferentes do ciclo do parasita e exercem funções diferentes (Chamond et al., 2003). Outro exemplo ocorre no clado *T. brucei* cujo receptor de transferrina, responsável pelo sequestro do ferro circulante do hospedeiro, evoluiu de um tipo de glicoproteína de superfície, responsável por promover a variação antigênica e escape do sistema imune (Jackson et al., 2012, 2013).

Outro mecanismo que promove a evolução genômica é a transferência horizontal de genes (HGT – *horizontal gene transfer*). A transferência horizontal (ou lateral) de genes se caracteriza pela transmissão de informação genética não herdada por descendência reprodutiva entre o receptor e o doador (Cooper, 2014; Keeling, Palmer 2008). Esse processo foi originalmente reconhecido na década de 60 em cepas de *Shigella,* durante um surto de disenteria no Japão (Akiba et al., 1960). Com o crescente aumento no número de genomas disponíveis, ficou evidente que, embora mais frequente em microrganismos, este processo é ubíquo e representa uma importante fonte de adaptação. São inúmeros os exemplos do aparecimento de adaptações decorrentes de transferência horizontal. Em mamíferos, a proteína conhecida como sincitina, que coordena a formação do sinciciotrofoblasto, responsável pela troca de nutrientes entre o feto e o útero, é de origem retroviral e foi adquirida em múltiplos eventos independentes ao longo da evolução dos mamíferos (Cornelis et al., 2015). A ocorrência de transferência horizontal é mais comum entre organismos como parasitas e endossimbiontes, que possuem estreita relação biológica com seus hospedeiros e organismos fagotróficos.

Em tripanossomatídeos, sabe-se que vários genes foram transferidos de várias fontes para esses parasitas, incluindo: enzimas envolvidas no metabolismo de carboidratos em *Leishmania* e *Trypanosoma* e genes relacionados a mobilização de cálcio e invasão celular em *T. cruzi* (origem bacteriana em ambos os casos); um sistema com seis polimerases de DNA responsáveis pela replicação do kDNA (origem viral); e enzimas e chaperonas similares às encontradas em plantas e cianobactérias (Folgueira, Requena, 2007; Hirt et al., 2015; Keeling, Palmer, 2008; Oliveira et al., 2014; Opperdoes, Michels, 2007; Silva et al., 2013). A presença de genes candidatos a terem sido transferidos de plantas ou cianobactérias para tripanossomatídeos levantou a hipótese de que a endossimbiose secundária, que gerou o cloroplasto presente nas espécies fotossintéticas de euglenídeos, possa ter ocorrido antes da divergência entre euglenídeos e cinetoplastídeos (Hannaert et

al., 2003). Entretanto, o padrão de distribuição desses genes entre os grupos que compõem o filo Euglenozoa indica que a aquisição do cloroplasto em euglenídeos é recente e que os genes encontrados em tripanossomatídeos foram adquiridos posteriormente.

A maior parte dos genes espécie-específicos está relacionada com escape imunológico, virulência e invasão de tripanossomas. A evolução desses genes apresenta características próprias, pois estão submetidos a pressões positivas de seleção e, desse modo, apresentam altas taxas de evolução, o que possibilita o rápido aparecimento de novas funções. *T. b. rhodesiense* é um exemplo, pois a resistência ao soro humano ocorre graças à presença da proteína conhecida como SRA (*Serum Resistance-Associated gene*), uma variante truncada de VSG (De Greef, Hamers, 1994; Jackson et al., 2012). Esses genes tendem a localizar-se em regiões próximas aos telômeros, que se tornaram regiões especializadas do controle da expressão e diversificação desses genes em tripanossomas. Cada clado apresenta seu próprio arsenal gênico, que é expresso em estágios específicos do ciclo de vida, garantindo a sobrevivência dos organismos tanto no hospedeiro vertebrado quanto no vetor. Na forma tripomastigota sanguínea, os genes das famílias das Mucinas e Trans-sialidases são destaque no clado *T. cruzi* (Cerqueira et al., 2008), enquanto que no clado *T. brucei* o foco está nos genes de VSGs (*Variant Surface Glycoprotein*) e ESAGs (*Expression Site Associated Genes*) (Jackson et al., 2012).

Os tripanossomatídeos apresentam regiões especializadas no genoma que dão suporte a esse amplo arsenal de genes espécie-específicos. Essas regiões encontram-se segregadas do genoma principal, situadas adjacentes aos telômeros. Elas oferecem um ambiente flexível com mecanismos epigenéticos distintos do resto do genoma e estrutura que permite altas taxas de recombinação (Figueiredo et al., Jackson, 2015ª; 2009; Rudenko 2010). Devido à alta taxa de diversificação dos genes presentes nesses *loci*, essas regiões ficaram conhecidas como zonas de contingência (Deitsch et al., 1997; Jackson, 2015a). Embora o surgimento de zonas de contingência seja fenômeno observável em vários grupos de parasitas, esta característica está acentuada em tripanossomatídeos, principalmente em tripanossomas. A variação antigênica é uma estratégia fundamental para a sobrevivência do parasita dentro do hospedeiro e as zonas de contingência fornecem as características essenciais para incubação, regulação e diversificação dos genes envolvidos no fenômeno (Barry et al., 2003). Neste contexto, as VSGs presentes nos tripanossomas do clado *T. brucei* são um bom exemplo da importância evolutiva do surgimento das zonas de contingência (Jackson et al., 2012, 2013; Jackson, 2015a).

As VSGs são glicoproteínas expressas na superfície dos tripanossomas do clado *T. brucei* de maneira a formar uma monocamada que atua como uma barreira para os anticorpos produzidos

pelo hospedeiro. A expressão desses genes é coordenada por um sítio de expressão ES (*expression sites*) de modo que cada célula expressa somente uma variante de VSG por vez, que é substituída por outra quando o tripanossoma volta a expressar proteínas de superfície (Figura 5) (Jackson et al., 2012, 2013; Jackson, 2015a). Isso faz com que a monocamada protetora seja trocada periodicamente e exiba, a cada troca, uma combinação antigênica diferente. Isso possibilita que o tripanossoma se torne temporariamente invisível para a resposta imune humoral hospedeiro (Jackson, 2015a). A função do ES consiste em assegurar a expressão de uma única cópia de VSG e o silenciamento das demais. A troca entre as diferentes VSGs ocorre por recombinação não homóloga ou silenciamento do ES ativo seguido da ativação de outro (Horn, McCulloch, 2010; Rudenko, 2010). A comparação do ES entre as espécies do clado *T. brucei* revelou uma sequência conservada para todo o clado e a existência de sequências repetitivas promotoras da recombinação e genes que codificam proteínas conhecidas como ESAGs (Figura 5) (Berriman et al., 2002; Graham et al., 1999; Hertz-Fowler et al., 2008).

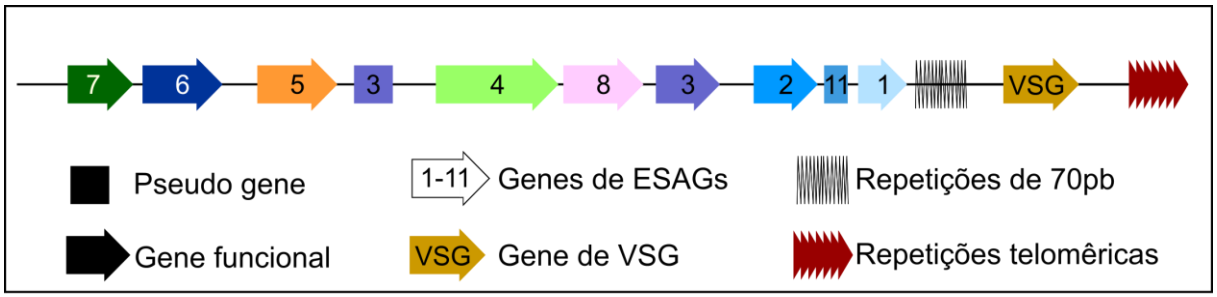

Figura 5 - Representação esquemática do sítio de expressão de T. b. brucei isolado Lister 427. Adaptado de Hertz-Fowler et al., 2008

Os genes que codificam ESAG6 e ESAG7 estão envolvidos na captação do ferro, essencial para a manutenção de diversas funções fisiológicas em tripanossomas. A captação do ferro ocorre via captação da transferrina circulante do hospedeiro por um receptor formado por hetorodímeros das ESAG6 e ESAG7 (Schell et al., 1991; Steverding, 2000, 2003). Os genes das ESAG6s e ESAG7s são homólogos entre si e ambos evoluíram a partir de uma variante de VSG presente no tripanossoma ancestral de *T. brucei* e *T. congolense* (Jackson et al., 2012, 2013). Apesar da similaridade entre as sequências desses genes, não são observadas recombinações entre os genes do receptor de transferrina e as VSGs (Jackson, 2015a).

As várias cópias dessas ESAGs mostraram-se muito polimórficas, com pelo menos 20 variantes descritas em *T. brucei* spp., e pouco polimórficas em *T. evansi* (Isobe et al., 2003; Witola et al., 2005). Estudos sugerem uma relação entre variabilidade de ESAG6 e ESAG7 e a manutenção da captação de ferro, mesmo na presença de anticorpos antiESAG6-7 que competem com a transferrina pela ligação ao receptor (Steverding, 2003). O polimorfismo das ESAG6 e ESAG7 também pode desempenhar um papel na capacidade de infectar hospedeiros diferentes, pois permite captação

contínua de ferro mesmo entre transferrinas de várias espécies de vertebrados (Young et al., 2008). Além da importância fisiológica para o parasita, os diferentes níveis de polimorfismo exibidos pelos tripanossomas africanos podem ser usados em estudos populacionais, identificação de genótipos e coevolução entre parasita e hospedeiro (Witola et al., 2005)

Além dos genes implicados no escape imunológico, os tripanossomas contam com um vasto arsenal de chaperonas e co-chaperonas, moléculas cruciais tanto para as transformações celulares sofridas ao longo do ciclo, quanto para assegurar a sobrevivência durante a troca do hospedeiro. O estresse celular não somente faz parte do ciclo de vida desses organismos como também desempenha um papel de gatilho que desencadeia profundas mudanças no padrão de expressão proteica e promove a diferenciação celular e progressão do ciclo dos parasitas (Folgueira, Requena, 2007; Pérez-Morales, Espinoza, 2015). Entre as diversas famílias de chaperonas presentes nos tripanossomas, as proteínas de choque térmico (HSP) destacam-se pela diversidade do repertório. Algumas famílias de HSP em tripanossomas apresentam genes especializados não encontrados em outros organismos e cuja função ainda é desconhecida, como a família HSP70, enquanto que, comparativamente a outros organismos, a família HSP40 possui ao menos duas vezes a quantidade de genes (Folgueira, Requena, 2007).

1.7.1 Proteínas de choque térmico

As proteínas de choque térmico (*Heat Shock Proteins* – HSPs) atuam como chaperonas moleculares que protegem proteínas recém-sintetizadas, desnaturadas ou danificadas, impedindo a formação irreversível de aglomerados proteicos (Burdon, 1986; Goloubinoff, De Los Rios, 2007; Lindquist, Craig 1988). Embora inicialmente descritas como específicas de condições de choque térmico, diversas condições de estresse celular (variação de pH, ação de toxinas, hipóxia, infecções, radiação e radicais livres) são capazes de desencadear a expressão de HSPs (Burdon, 1986; Lindquist, Craig, 1988). Isto ocorre devido ao aparecimento de proteínas na forma não nativa em condições de estresse celular (Feder, Hofmann, 1999). Embora vitais para assegurar a sobrevivência da célula em condições de estresse, elas também são constitutivamente expressas. As HSPs prestam assistência na síntese, ativação e transporte proteico entre compartimentos celulares (Folgueira, Requena, 2007; Genevaux et al., 2007; Goloubinoff, De Los Rios, 2007). As HSPs são ubíquas e fazem parte das proteínas com estrutura mais conservada ao longo da evolução. São encontradas sob a forma de um ou mais parálogos em todos os compartimentos da célula (Tutar, Tutar, 2010). A classificação das HSPs é feita conforme o peso molecular, sendo divididas atualmente em seis famílias e superfamílias: HSP90, HSP104/CLpB, HSP60/HSP10, HSP40, sHSP (Small heat shock proteins) e

HSP70, que incluiu a família HSP110 por homologia da sequência (Burdon, 1986; Easton et al., 2000; Lindquist, Craig, 1988; Tutar, Tutar, 2010).

Os parasitas correspondem a um caso peculiar, pois o estresse celular é parte crucial do ciclo de vida desses organismos. Neste caso, as HSPs, além de manter a homeostase, são usadas como estratégia de adaptação durante a troca entre hospedeiros. Para a grande maioria dos parasitas, a troca entre microambientes é acompanhada de variações de temperatura, concentração de micronutrientes e pH (Figura 6), alterações acompanhadas por uma forte indução de HSPs (Maresca, Carratu, 1992). Em tripanossomatídeos, análises proteômicas detalhadas durante a diferenciação de *Leishmania donovani* e *T. cruzi* mostraram aumento na síntese de HSP60, HSP70 (citosólica e mitocondrial) e HSP90 (Bente et al., 2003; Paba et al., 2004). Porém, ainda não é claro se as HSPs estão apenas relacionadas com o estresse celular momentâneo ou se atuam diretamente no processo de diferenciação celular (Folgueira, Requena, 2007; Maresca, Carratu, 1992).

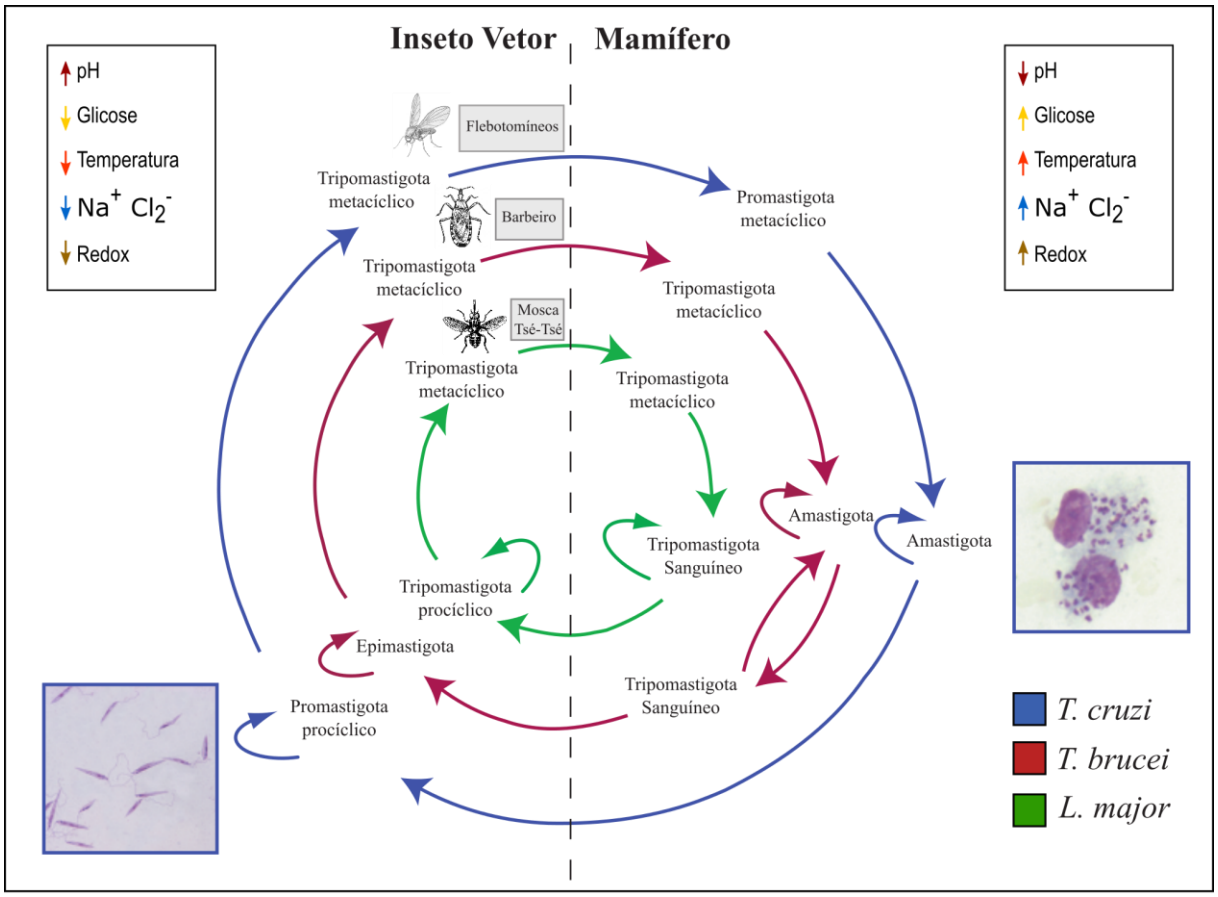

Figura 6 - Diferenças entre o microambiente encontrado no hospedeiro vertebrado e no vetor Adaptado de Folgueira e Requena, 2007 e Pérez-Morales e Espinoza 2015.

A família HSP104 é responsável por inibir a formação de agregados proteicos e também participa da reestruturação da estrutura terciária de peptídeos. Esta última característica faz com que essa família seja importante no controle da formação de *príons* (Shorter; Lindquist, 2008). Os

representantes dessa família são conhecidos como ClpB (bactérias), HSP104 (leveduras) e HSP101 (plantas). Estudos recentes mostraram a presença de um gene de HSP100 (membro das HSP104) em *Leishmania donovani*; indicando uma provável função no desenvolvimento normal do parasita, mutantes para os dois alelos apresentam amastigotas com morfologia aberrante (Krobitsch, Clos, 1999).

A família HSP60/HSP10, também conhecida como Chaperoninas do tipo I, está presente em bactérias e em organelas eucarióticas de origem procariótica, como mitocôndria e cloroplasto (Hirtreiter et al., 2009). A HSP60 necessita formar um complexo que inclui várias subunidades da HSP10, de modo a atuar como uma jaula, conhecida como gaiola de Anfinsen (Hartl, Hayer-Hartl, 2002), que retém proteínas recém-sintetizadas. Em tripanossomas, a concentração da HSP60 nas formas encontradas no inseto é de três a quatro vezes maior quando comparada à presente nas formas sanguíneas. Isso provavelmente decorre da adaptação do metabolismo mitocondrial durante a passagem entre os hospedeiros (Bringaud et al., 1995).

As HSP90s estão presentes em todos os compartimentos da célula e, embora tenham sido descritas como proteínas de estresse térmico, em condições normais representam de 1 a 2% de todas as proteínas da célula (Montalvo-Alvarez et al., 2008). Essa família, junto com a HSP70, é responsável pela reciclagem e pelo controle do tráfego de proteínas entre os compartimentos celulares. A maior parte das proteínas clientes da HSP90 está envolvida no controle do ciclo celular e transdução de sinal (Richter, Buchner, 2001; Pratt, Toft, 2003; Whitesell, Lindquist, 2005). Em tripanossomatídeos, vários genes que codificam HSP90 foram descritos, sobretudo para *T. b. brucei*, *T. cruzi* e *Leishmania* spp.: HSP83 no citoplasma (Dragon et al., 1987; Mottram et al., 1989; Shapira, Pedraza, 1990), Grp94 (Larreta et al., 2000; Folgueira, Requena, 2007) no retículo e mtHSP70 na mitocôndria (Folgueira, Requena, 2007). Na maior parte dos casos, foram encontradas várias cópias de cada gene dispostas em tandem (Folgueira, Requena, 2007).

Assim como as HSP90, os membros da família HSP70 são ubíquos e estão envolvidos no transporte através de membrana entre organelas, degradação de peptídeos instáveis, excreção, transdução de sinal, regulação do ciclo celular, diferenciação e apoptose (Sharma, Masison, 2009; Zylicz, Wawrzynow, 2001). A família HSP110, embora classificada inicialmente como uma família a parte devido ao peso molecular, foi reclassificada com base na homologia da sequência como pertencente à família HSP70 (Easton et al., 2000). As HSP110s exibem baixa afinidade por proteínas clientes, carregando-as momentaneamente, função conhecida como *holdase*. No entanto, a principal função dessa família é atuar como co-chaperonas, estimulando a troca de ADP por ATP (*nucleotide exchange factor* – Nef) durante o mecanismo de ação, dependente de ATP, das demais HSP70s

(Dragovic et al., 2006). A arquitetura da família HSP70 é conservada, sendo composta por um sítio de ligação ao ATP, situado na porção amino-terminal, e outro de ligação às proteínas clientes, na porção carboxi-terminal, ambos ligados por uma região conservada denominada *Linker* (Figura 7) (Goloubinoff, De Los Rios, 2007).

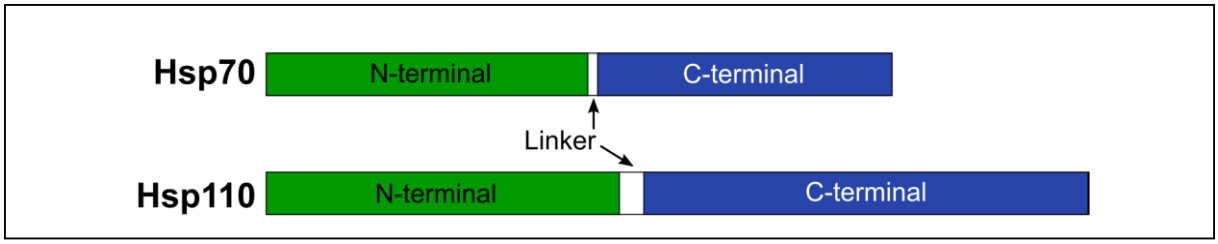

Figura 7 - Arquitetura da família Hsp70. A arquitetura e sequência são conservadas entre as Hsp110 e Hsp70.

Para desempenhar essa gama de funções, é necessária a associação com outras HSPs, como a HSP90 e HSP40, que parecem ter evoluído de maneira conjunta aos genes da superfamília HSP70 (Kabani, 2009). Proteínas da família HSP70 estão presentes em todos os compartimentos, podendo apresentar múltiplos membros em um mesmo compartimento (Kabani, Martineau, 2008). Em eucariotos, análises filogenéticas têm mostrado a existência de quatro grupos monofiléticos de HSP70, que correspondem ao citoplasma, retículo endoplasmático, mitocôndria e cloroplasto (Gupta, Golding, 1993; Gupta et al., 1994). A hipótese mais aceita para o surgimento desses grupos defende que as HSP70s citoplasmáticas e do retículo surgiram a partir de uma duplicação de uma HSP70 primitiva, presente no ancestral comum a todos os eucariotos, e se adaptaram aos compartimentos celulares em formação (Germot, Philippe, 1999). Os grupos residentes do cloroplasto e mitocôndria são de origem procariótica e foram transferidos para o núcleo pelo ancestral endossimbionte durante o processo de formação dessas organelas (Boorstein et al., 1994; Germot, Philippe, 1999; Gupta, Golding, 1993). Os genes de HSP70 têm sido usados em estudos filogenéticos de vários grupos de eucariotos, inclusive parasitas de importância médica, como *Cryptosporidium* spp. e *Babesia* spp., e, mais recentemente, para *Leishmania* spp. (Fraga et al., 2010; Sulaiman et al., 2000; Yamasaki et al., 2002).

Diversos genes de HSP70 foram descritos em tripanossomas e leishmanias no citoplasma (HSP70, HSP70.4, HSP70.c) (Folgueira, Requena, 2007; Glass et al., 1986; Lee et al., 1988, 1990; Requena et al., 1988); retículo (Grp78) (Bangs et al., 1993; Jensen et al., 2001; Tibbetts et al., 1994); mitocôndria (mtp70 – Lc2.1 e Lc2.2) (Campos et al., 2008; Engman et al., 1989); e HSP110s (Grp170 e HSP110) (Folgueira, Requena, 2007). Além desses genes, a comparação entre os genomas de *T. cruzi, T. b. brucei* e *L. major* revelou a existência de dois parálogos divergentes, HSP70.a e HSP70.b, sendo o primeiro sem ortólogos conhecidos fora de tripanossomatídeos e o segundo, semelhante ao encontrado em algumas cianobactérias, uma possível aquisição por transferência horizontal (Folgueira, Requena, 2007).

Os integrantes da família da HSP40 possuem sequência diversificada, embora todas compartilhem um domínio conhecido como Domínio J, com cerca de 70 resíduos, geralmente localizado na porção amino-terminal (Walsh et al., 2004). Esse domínio é fundamental para a parceria funcional entre as HSP40s e as HSP70s. Foram descritas inicialmente quatro HSP40 em *T. cruzi,* chamadas de TcJ1-4 (Tibbetts et al., 1998), mas a busca por genes contendo o domínio J em tripanossomatídeos mostrou a existência de pelo menos 67 genes presentes em *L. major*, *T. b. brucei* e *T. cruzi* (Folgueira, Requena, 2007). Embora exista uma expansão dessa família em tripanossomatídeos em relação a outros parasitas, a família HSP40 permanece como a menos estudada das HSPs nesses organismos.

As proteínas de choque térmico de baixo peso molecular (*Small Heat Shock Proteins* – sHSP) representam a família menos conservada, porém, amplamente difundida entre todos os reinos (de Miguel et al., 2009; Haslbeck et al., 2005;). Estruturalmente, todas as sHSPs possuem somente um domínio característico, considerado assinatura dessa família e relativamente conservado em tripanossomatídeos (Folgueira, Requena, 2007). Esse domínio recebe o nome de α-cristalino, pois as sHSPs são uma das classes de proteínas mais abundantes no cristalino de vertebrados. Estudos filogenéticos sugerem que as sHSPs divergiram bem cedo na história evolutiva, estando presente no último ancestral comum entre procariotos e eucariotos (Fu et al., 2006). O número de parálogos parece aumentar conforme o grau de complexidade do organismo, o que sugere a existência de funções especializadas relacionadas a tecidos (Haslbeck et al., 2005). Apesar da função importante em organismos parasitas (Pérez-Morales, Espinoza, 2015), a análise genômica *in silico* revelou um único gene, HSP20, em *T. brucei*, *L. major* e *T. cruzi* (Folgueira, Requena, 2007). Esse gene foi mais tarde caracterizado bioquimicamente em *T. cruzi* (Perez-Morales et al., 2009). Recentemente, o gene da HSP20 tem sido usado combinado ao da HSP70 para identificação de espécies de *Leishmania* (Fraga et al., 2013; Montalvo et al., 2014).

1.7.2 Catepsinas L-like de tripanossomas

A Catepsina L (CATL) é uma cisteína protease envolvida nos processos de invasão, multiplicação e diferenciação em tripanossoma, estando também envolvida na modulação da resposta imune do hospedeiro (Garcia et al. 2011a, 2011b; Jefferson et al. 2016). A CATL é sintetizada na forma de pré-pro-catepsina no retículo endoplasmático, em seguida direcionada para os lisossomos, quando o pré-domínio (peptídeo sinal) é removido, gerando a pro-catepsina. Ao chegar aos lisossomos sofre a conversão pelo pH ácido para a forma madura com atividade de endopeptidase. A CATL-like de tripanossomas possui quatro domínios: pré-domínio, que é o peptídeo sinal, pró-domínio, domínio catalítico e extensão carboxi-terminal. O domínio catalítico é responsável pela atividade da enzima,

sendo conservado em todas as espécies de tripanossoma. A comparação entre os genomas disponíveis mostrou uma grande quantidade de cópias dispostas em tandem (Campetella et al., 1992; Eakin et al., 1992). A enzima arquétipo, descrita em *T. cruzi* (TcrCATL), é codificada por pelo menos 130 genes que apresentam níveis variados de polimorfismo, contrastando com *T. brucei* e *T. rangeli,* que possuem cerca de 20 e 75 cópias, respectivamente (Campetella et al., 1992; Eakin et al., 1992). Desse modo, em tripanossomas a CATL é uma família multigênica cujos membros desempenham papéis diferentes conforme a espécie e o estágio do ciclo de vida e dessa forma são fundamentais para a sobrevivência do parasita. Estudos sugerem que todas as cisteína proteases similares com Catepsina L são homologas e divergiram, originando diversas variantes ao longo do processo de especiação (Sajid, McKerrow, 2002). Os genes de Catepsina L de tripanossomas têm sido usados com sucesso para diagnóstico, genotipagem, alvos terapêuticos e estudos evolutivos em *T. theileri*, *T. rangeli*, *T. cruzi, T. brucei* e *T. vivax* (Cortez et al. 2009; Ettari et al. 2016; Garcia et al. 2011a, 2011b; Jefferson et al. 2016; Lima et al. 2012a; Nakayima et al. 2013; Rodrigues et al. 2010, 2015; Ortiz et al. 2009). A grande diversidade de funções exibidas por essas proteases pode ter contribuído para complexidade do ciclo de vida e o grande número de hospedeiros e vetores observados em *Trypanosoma* (Alvarez et al., 2012; Atkinson et al., 2009; Lima et al., 2012a). Evidencias apontam que os blocos que compreendem as cópias de CATL são sintênicos e podem estar evoluindo em concerto; no caso dos tripanossomatídeos essa característica é compartilhada entre genes que se comportam como bons marcadores filogenéticos (Jackson, 2007). Nesse contexto, estudos apontam que os genes de CATL possuem potencial como marcadores para diagnóstico, para a determinação de genótipos e, também, estudos filogenéticos (Ruszczyk et al., 2008; Sakanari et al., 1997; Ortiz et al., 2009).

1.7.2 Prolina racemase

A prolina racemase (PRAC) pertence ao grupo das enzimas que promovem a interconversão entre os aminoácidos estereoisômeros levógiros (L) e dextrógiros (D), produzindo uma mistura racêmica. As enzimas pertencentes à classe das aminoácido-racemases são associadas a funções metabólicas, nutricionais e imunológicas em patógenos (Yoshimura, Esak, 2003). Essas enzimas estão presentes em uma grande variedade de procariotos e em alguns grupos de eucariotos cujo gene foi adquiro por transferência horizontal (Fitzpatrick et al., 2008; Visser et al., 2012; Yoshimura, Esak 2003). Em bactérias patogênicas, as aminoácido-racemases são responsáveis pela presença de D-aminoácidos na parede celular, o que confere proteção contra a ação enzimática e os componentes proteolíticos do sistema imune dos hospedeiros (Coatnoan et al., 2009; Thompson et al., 1998).

Originalmente descrita em *Clostridium sticklandii* (Stadtman, Elliott, 1957), *T. cruzi* foi o primeiro eucarioto descrito com um gene funcional de prolina racemase (Reina-San-Martín et al.,

2000). Mais tarde, foi comprovada a existência do gene da PRAC em *T. vivax* (Chamond et al., 2009). Em *T. cruzi,* os estudos sobre a PRAC mostraram a existência de duas cópias, TcPRACA e TcPRACB, que são expressas diferencialmente e desempenham funções essenciais para a sobrevivência do parasita (Chamond et al., 2003, 2005). TcPRACA possui cerca de 45KDa e é secretada tanto por tripomastigotas metacíclicos quanto sanguíneos, graças à presença de um peptídio sinal, ausente na TcPRACB. A TcPRACB é menor devido à ausência de peptídeo sinal (cerca de 39kDa). Desse modo, a TcPRACB não é secretada, sendo expressa principalmente na forma epimastigota. Embora as duas difiram pela presença do peptídeo sinal e função, ambas compartilham cerca de 96% de identidade na sequência de aminoácidos (Chamond et al., 2003).

Estudos com inibidores da TcPRAC evidenciam que a inativação da enzima é responsável pela diminuição da proliferação e viabilidade do parasita, o que evidencia o potencial dessa enzima como alvo quimioterapêutico no tratamento da Doença de Chagas (Conti et al., 2011; Coutinho et al., 2009). Estudos imunológicos indicam que a TcPRACA secretada pelas formas sanguíneas é capaz de induzir a ativação policlonal dos linfócitos B e, consequentemente, um atraso na resposta imune humoral (Buschiazzo et al., 2006; Chamond et al., 2005; Reina-San-Martín et al., 2000). Também foi comprovada que TcPRAC é capaz de induzir a liberação de IL-10, interleucina responsável pela supressão de algumas vias imunológicas, deixando o hospedeiro susceptível a infecção por *T. cruzi* (Bryan, Norris, 2010; Chamond et al., 2003, 2005; Reina-San-Martín et al., 2000). Do mesmo modo que em bactérias, também foi sugerido que TcPRAC pudesse participar da incorporação de D-aminoácidos em peptídeos de superfície, o que tornaria *T. cruzi* resistente a ação proteolítica e enzimática do hospedeiro (Coatnoan et al., 2009; Reina-San-Martín et al., 2000). A procura por genes de PRAC em *T. vivax* revelou um único gene, TvPRAC, semelhante à TcPRACB por não possuir o peptídeo sinal e exibir parâmetros cinéticos e imunológicos semelhantes aos encontrados em *T. cruzi* (Chamond et al., 2009). Embora a TvPRAC não apresente sinalização para ser secretada, experimentos *in vivo* mostraram um aumento no número de linfócitos B, como também altos títulos de anticorpos no soro de camundongos (Chamond et al., 2009). Desse modo, foi levantada a possibilidade da TvPRAC ser secretada via bolso flagelar ou liberada na lise do parasita junto com o conteúdo intracelular (Chamond et al., 2009).

A comparação genômica entre as espécies de tripanossoma revelou que o *locus* do gene da PRAC é sintênico entre *T. vivax* e *T. cruzi.* Contudo, nenhum ortólogo da PRAC foi encontrado em *T. brucei brucei,* enquanto que em *T. congolense* foram encontrados apenas vestígios do que teria sido um pseudogene no loci homologo ao da TcPRAC e TvPRAC (Chamond et al., 2009). Nos genomas de *L. major*, *Leishmania braziliensis* e *Leishmania infantum,* o locus da PRAC parece ter sido ocupado por

um gene que codifica uma proteína da família Tub, ausente no loci correspondente em *Trypanosoma* (Chamond et al., 2003, 2009).

2 JUSTIFICATIVA E OBJETIVOS

A complexidade dos ciclos de vida das espécies do gênero *Trypanosoma* é surpreendente. Durante o desenvolvimento em seus respectivos hospedeiros vertebrados e vetores, os tripanossomas apresentam diversos estágios em diferentes combinações, variáveis de acordo com a espécie, sendo as formas tripomastigotas as únicas encontradas em todas as espécies. Embora muitas espécies tenham sido descritas, se conhece muito pouco sobre o desenvolvimento e mecanismos de infecção dos tripanossomas em hospedeiros não mamíferos. A maior parte dos estudos concentra-se apenas nas espécies que parasitam o homem e animais de importância econômica. Estes estudos descrevem uma grande variedade de famílias multigênicas que são, direta ou indiretamente, responsáveis pela complexidade observada entre os tripanossomas. Além da complexidade do ciclo de vida, o surgimento de repertórios de genes especializados moldou de maneira independente a evolução dos diversos clados em *Trypanosoma.* Diferentes estratégias de infecção e escape imunológico surgiram nesses parasitas, o que, sem dúvida, é refletido na enorme quantidade de hospedeiros vertebrados (animais aquáticos, terrestres e voadores de todas as classes) e de vetores (sanguessugas, moscas, mosquitos, hemípteros, pulgas e carrapatos). Desse modo, para entender tanto a evolução do repertório gênico quanto dos tripanossomas é essencial que novos estudos incluam uma maior representatividade, incluindo espécies não patogênicas e que infectam animais que não possuem importância econômica (Jackson, 2015a).

Essas famílias multigênicas possuem diferentes papéis na sobrevivência desses parasitas e, desse modo, são submetidas a diferentes pressões seletivas e, consequentemente, apresentam histórias evolutivas diferentes. Alguns genes foram adquiridos por transferência horizontal e rapidamente evoluiram de forma distinta em diferentes espécies como a PRAC em *T. vivax* e *T. cruzi* (Chamond et al., 2003, 2009). Outras famílias sofreram expansão e rápida diversificação dos parálogos, como as cisteína peptidases observadas em diferente número entre os diversos clados (Lima et al., 2012a; Ortiz et al., 2009). As proteínas de choque térmico (HSP) desempenham papéis muito importantes na capacidade de invadir, sobreviver, multiplicar e diferenciar de espécies de *Trypanosoma* e *Leishmania* (Giambiagi-deMarval et al., 1996; Hubel et al., 1997; Krobitsch, Clos, 1999; Naderer, McConville, 2011; Wiesgigl, Clos, 2001; Raina, Kaur, 2011). A comparação entre o repertório de HSPs encontrado nos genomas de *T. brucei* e *T. cruzi e L. major* indica que cada família evoluiu de maneira diferente nesses organismos, apresentando expansão, redução e possíveis genes de transferência horizontal (Folgueira, Requena, 2007). Nesse trabalho, é nosso objetivo explorar o repertório e a evolução de famílias gênicas presentes nos genomas de diversas espécies e que estão envolvidas na sobrevivência do parasita. O principal enfoque é o estudo de espécies de tripanossomas,

com outros tripanossomatídeos utilizados para comparação e filogenia. Pretendemos investigar essas famílias gênicas nos genomas de diversas espécies de tripanossomas, a fim de se entender a evolução desses genes e explorar o potencial destes como marcadores filogenéticos, de genotipagem e diagnóstico. Os objetivos específicos são:

a) Investigar o repertório e a evolução de famílias multigênicas selecionadas

b) Avaliar o potencial desses genes como marcadores filogenéticos para o gênero *Trypanosoma*

c) Comparar as filogenias obtidas com as baseadas nos marcadores tradicionais (genes SSU rRNA e gGAPDH) e inferir filogenias com sequências combinadas

d) Investigar a organização genômica e a sintenia dos genes em diferentes espécies de tripanossomas

f) Explorar a utilidade dos genes como marcadores da variabilidade genética dos tripanossomas para genotipagem e diagnóstico.

3 MATERIAIS E MÉTODOS

3.1 Organismos

Os isolados utilizados neste estudo, assim como seus respectivos hospedeiros de origem e áreas geográficas de isolamento estão detalhados nos trabalhos anexados e que compõem o corpo da tese. Todas as culturas de tripanossomatídeos são mantidas criopreservadas por N_2 líquido em meio LIT (Camargo, 1964) com 10% de SFB acrescido de 20% de DMSO e fazem parte da coleção de culturas de tripanossomatídeos (TCC: "*Trypanosomatidae Culture Collection*") do Departamento de Parasitologia, ICB II – USP.

3.2 Obtenção de DNA genômico e sequenciamento dos genomas

O DNA genômico total foi extraído pelo método de fenol clorofórmio. Foram aplicados os métodos de depleção do kDNA para minimizar a presença deste tipo de molécula, como previamente descrito (Alves et al., 2011). Após a retirada do kDNA, cerca de 5 µg de DNA foram submetidos à técnica de sequenciamento por *shotgun* (Roche 454) de acordo com os protocolos do fabricante. Os genomas sequenciados possuem coberturas estimadas entre 15 a 23 vezes (considerando um genoma de aproximadamente 30 Mpb). As sequências foram montadas usando o montador Newbler versão 2.3, fornecido pela Roche. Os genomas resultantes pertencem ao projeto ATOL – *Assembling the Tree of Life* (NSF-USA) estão disponíveis na tabela 3.

Tabela 3 - Genomas utilizados neste trabalho

Organismos	Isolados	Hospedeiros	Projeto e número de acesso
Euglenida			
Discoplastiss patirhyncha	SAG 1224.42	vida livre	ATOL
Euglena gracilis	-	vida livre	ATOL
Eutreptia viridis	SAG 1226-1c	vida livre	ATOL
Ploeotia costata	ATCC 1264	vida livre	ATOL
Diplonemida			
Diplonema ambulator	ATCC 50223	vida livre	ATOL
Rhynchopus sp.	-	vida livre	ATOL
Bodonida			
Bodo sp.	ATCC 50149	vida livre	ATOL
Parabodo caudatus	ATCC 30905	vida livre	ATOL
Dimastigella trypaniformis	-	vida livre	ATOL
Trypanosomatida			
Herpetomonas wanderley	TCC 1982	*Cochliomy macellaria*	ATOL
Herpetomonas muscarum	TCC 001E	*musca dosmestica*	ATOL AUXJ01000000

Phytomonas sp.	TCC 418	*Pachycoris torridus*	ATOL
Phytomonas sp.	TCC 066E	*Jatropha macrantha*	ATOL
Angomonas deanei	TCC 036E	*Zelus leucogrammus*	ATOL AUXM01000000
Angomonas desouzai	TCC 079E	*Ornidia obesa*	ATOL AUXL01000000
Angomonas ambiguus	TCC 2435	*Chrysomya albiceps*	ATOL
Strigomonas culicis	TCC 012E	*Aedes vexans*	ATOL AUXH01000000
Strigomonas galati	TCC 219	*Lutzomya almerioi*	ATOL
Strigomonas oncopelti	TCC 290E	*Oncopeltus fasciatus*	ATOL AUXK01000000
Leptomonas costaricensis	TCC 169E	*Zelinae* sp.	ATOL
Crithidia acantocephali	TCC 037E	*Acanthocephala femorata*	ATOL
Crithidia luciliae thermophila	TCC 050E	*Zelus leucogrammus*	ATOL
Endotrypanum schaudinni	TCC 224	*Choloepus hoffmani*	ATOL
Trypanosoma cyclops	TCC 052	*Macaca ira*	ATOL
Trypanosoma livingstonei	TCC 1270	*Rhinolophus landeri*	ATOL
Trypanosoma wauwau	TCC 1873	-	ATOL
Trypanosoma ralphi	TCC 1838	*Melanosuchus niger*	ATOL
Trypanosoma serpentis	TCC 1052	*Pseudoboa nigra*	ATOL
Trypanosoma theileri	TCC 165	*Bubalus bubalis*	ATOL
Trypanosoma lewisi	TCC 34	*Rattus rattus*	ATOL
Trypanosoma rangeli	AM80	*Homo sapiens*	ATOL
Trypanosoma conorhini	TCC 025E	*Rattus rattus*	ATOL
Trypanosoma dionisii	TCC 211	*Eptesicus brasiliensis*	ATOL
Trypanosoma erneyi	TCC 1946	*Mops condylurus*	ATOL
Trypanosoma cruzi cruzi	G	*Didelphis marsupialis*	ATOL
Trypanosoma cruzi marinkellei	TCC 344	*Carollia perspicillata*	ATOL
Trypanosoma sp.	TCC 2045	-	ATOL
Trypanosoma sp.	TCC 339	*Rhinella marina*	ATOL
Trypanosoma sp.	TCC 878	*Mabuya frenata*	ATOL
Trypanosoma sp.	TCC 1825	*Ramphocelus nigrogularis*	ATOL
Trypanosoma sp.	TCC 16	*Macropus giganteus*	ATOL
Trypanosoma sp.	TCC 17	*Vombatus ursinus*	ATOL
Trypanosoma sp.	TCC 1307	*Ptychadena mossambica*	ATOL
Trypanosoma sp.	TCC 2186	*Leptodactylus latraus*	ATOL

3.3 Busca genômica das sequências, alinhamento e análises filogenéticas

A busca dos genes foi realizada por similaridade através do programa BLAST (BLASTn, BLASTp, tBLASTn e BLASTx) e rpsBLAST, usando como isca sequências previamente descritas na literatura ou matrizes de pontuação por posição específica (PSSM – P*osition-specific scoring matrix*) obtidas do banco de domínios conservados CDD-NCBI, para o BLAST e rpsBLAST respectivamente.

Os genes pertencentes a cada família foram classificados como parálogos ou ortólogos com base na relação filogenética, similaridade de sequência e comparação de motivos. Para verificar a existência de genes frutos de transferência horizontal, uma análise abrangente foi realizada incluindo sequência no NR (ftp://ftp.ncbi.nih.gov/blast/db/). Para reduzir a demanda computacional, as sequências obtidas do NR foram agrupadas usando o programa USEARCH (Edgar 2010) usando 95%

de similaridade e 90% de cobertura como parâmetros. Sequências parciais com menos de 70% do tamanho médio do gene ou que apresentam domínios proteicos não relacionados foram excluídas das análises.

Os alinhamentos múltiplos de sequências foram feitos utilizando o programa MUSCLE v3.8 (Edgar 2004) e editados usando o Gblocks v0.91b (Castresana 2000), a fim de eliminar posições com alinhamento espúrio. Os modelos de substituição usados nas análises de aminoácido foram escolhidos pelo critério de informação de Akaike (AIC) usando ProtTest3 (Abascal et al., 2005), enquanto que o modelo GTR com taxa de distribuição gama foi utilizado para análises de nucleotídeos. As análises de máxima verossimilhança foram feitas usando o RAxML v8.2.0 (Stamatakis 2014). O suporte de ramos foi estimado com 500 replicatas pelo método de *bootstrap,* implementado pelo programa. Devido ao alto tempo computacional as análises que incluem sequências do NR foram realizadas pelo método de máxima verossimilhança aproximada implementado pelo programa FastTree v2.1 (Price et al., 2010) com o modelo de WAG e taxa distribuição gama. As árvores filogenéticas obtidas foram visualizadas e editadas com o programa Figtree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/), Treegraph2 (Stöver e Müller 2010) e dendroscope v3.2.4 (Huson e Scornavacca 2012).

3.4 Amplificação de DNA por PCR – *Polymerase Chain Reaction*

As reações de PCR foram realizadas a partir da mistura de reação: 100 ng de DNA genômico; 100 ng de cada iniciador; 200 mM de cada dNTP (dATP, dCTP, dGTP, dTTP); 5 µL de tampão (200 mM Tris-HCl, pH 8,4, 500 mM KCl e 1,5 - 3 mM $MgCl_2$); 2,5 U de Taq DNA polimerase e água bidestilada deionizada e autoclavada (qsp 50 µl). Os ciclos de amplificação e as temperaturas foram definidos de acordo com os iniciadores empregados (Tabela 4).

3.5 Eletroforese e purificação de fragmentos de DNA.

Alíquotas das reações de PCR foram misturadas com solução de tampão de amostra (50% de glicerol; 0,4% de azul de bromofenol; 0,4% de xilenocianol) e submetidas à eletroforese em géis de agarose (1,5 a 2,5%) em tampão TAE (40 mM de Tris-acetato; 2 mM de EDTA, pH 8,0.) a 70 V/100 mA. Como marcador de peso molecular foi utilizado DNA Leader Mix (MBI Fermentas). Após a eletroforese, os géis foram corados com GelRed® *Nucleic acid Gel Stain* e fotografados em um ImageQuant 350 (GE Healthcare®) sob luz UV. Os fragmentos obtidos foram cortados do gel com uma lâmina estéril e purificados em coluna Spin X (Costar®). Os produtos purificados foram clonados ou submetidos diretamente às reações de sequenciamento.

Tabela 4 - Ciclos de amplificação e temperaturas utilizadas nas diferentes reações de PCR

Gene e oligonucletídeos ("primers") empregados	Condições de amplificação
CATL (Catepsina L) domínio catalítico	
DTO154 (5' ACA GAA TTC CAG GGC CAA TGC GGC TCG TGC TGG 3')	1 ciclo: 3 min 95 °C, 39 ciclos: 1 min 94 °C; 1 min 55 °C; 1 min 72 °C, 1 ciclo: 10 min 72 °C
DTO155 (5' TTA AAG CTT CCA CGA GTT CTT GAT GAT CCA GTA 3')	
ESAGs (Gene associado a sítios de expressão de VSG)	
ESAG7-455s (5´ CAT TCC AGC AGG AGT TGG AGG 3´)	1 ciclo: 3 min 94 °C, 30 ciclos: 1 min 94 °C; 30 seg 58 °C; 1 min 72 °C, 1 ciclo: 10 min 72 °C
ESAG6-1045as (5´ TTG TTC ACT CAC TCT CTT TGA CAG 3´)	
HSP20 – Gene completo	
FsH-AFR (5' ATGT GGG ACC CTT TTC GTG ATA 3')	
FsH-RAN (5' ATG TGG GAC CCG TTT CGT GAC 3')	
FsH-SCH (5' ATG TGG GAC CCG TTT CGT GAC A 3')	1 ciclo: 3 min 94 °C, 30 ciclos: 1 min 94 °C; 30 seg 54 °C; 1 min 72 °C, 1 ciclo: 10 min 72 °C
RsH-SCH (5' ATT GAT CTT CAC GGA GAT CCC C 3')	
RsH-RAN (5' ATC TTC ACA GAG GTT CCT GAG C 3')	
RsH-VIX (5' CTG AAT ATT CAC TGA CGA ACC TG 3')	
RsH-TZO (5' CTG AAT CGT TAC GGA TGT TGC AG 3')	
HSP70 – Porção carboxiterminal	
HSP70F (5' ATG TGG GAC CCG TTT CGT GAC 3')	1 ciclo: 3 min 94 °C, 30 ciclos: 1 min 94 °C; 30 seg 56 °C; 1 min 72 °C, 1 ciclo: 10 min 72 °C
HSP70R (5' CTG GTA CAT CTT CGT CAT GAT G 3')	
Reações de sequenciamento	
M13F (5' GTA AAA CGA CGG CCA G 3´)	1 ciclo: 1 min 96 °C, 30 ciclos: 15 seg 96 °C; 15 seg 50 °C; 4 min 60 °C, 1 ciclo: 5 min 72 °C
M13R (5' CAG GAA ACA GCT ATG AC 3´)	

3.6 Sequenciamento de fragmentos amplificados por PCR

Os fragmentos obtidos foram submetidos a reações de sequenciamento utilizando o kit Big Dye Terminator (Perkin Elmer®), de acordo com especificações do fabricante, em sequenciador automático ABI PRISM 3100 Genetic Analyzer (Perkin Elmer®). As reações foram efetuadas empregando os mesmos iniciadores utilizados nas reações de PCRs ou os iniciadores universais M13F/M13R para sequenciamento de fragmentos de DNA clonados em vetor pCR 2.1.

4 RESULTADOS E DISCUSSÃO

Os resultados obtidos neste trabalho serão apresentados e discutidos na forma de manuscritos prontos para submissão em periódicos internacionais. Os resultados que já se encontram publicados são apresentados na forma de resumos nesta sessão e como artigos originais na sessão – **ANEXOS.**

4.1 Genome-wide phylogenetic analyses of genetic repertoires and evolution of the Hsp70 superfamily across Trypanosomatidae (Kinetoplastea: Euglenozoa) parasites of invertebrates and vertebrates

4.2 The expanding superfamily of protein sharing α-crystallin domain (ACD), including the Small Heat Shock Protein (sHSP) family, revealed by surveys in the genomes of a large diversity of Trypanosoma species

ANEXO A - Congopain genes diverged to become specific to Savannah, Forest and Kilifi subgroups of *Trypanosoma congolense*, and are valuable for diagnosis, genotyping and phylogenetic inferences.

ANEXO B - Phylogenetic and syntenic data support a single horizontal transference to a Trypanosoma ancestor of a prokaryotic proline racemase implicated in parasite evasion from host defences.

ANEXO C - Endosymbiosis in trypanosomatids: the genomic cooperation between bacterium and host in the synthesis of essential amino acids is heavily influenced by multiple horizontal gene transfers.

Genome-wide phylogenetic analyses of genetic repertoires and evolution of the Hsp70 superfamily across Trypanosomatidae (Kinetoplastea: Euglenozoa) parasites of invertebrates and vertebrates

Costa-Martins A. G.[a]; Alves, J. M. P[a]; Ortiz, P. A.[a]; Lima, L.[a]; Fermino, B.[a]; Serrano, M. G.[b]; Camargo, E. P.[a]; Buck G. A.[b]; Marta M. G. Teixeira[a]*

[a] Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, SP, Brazil

[b] Department of Microbiology and Immunology and the Center for the Study of Biological Complexity, Virginia Commonwealth University

Abstract

Background: The Trypanosomatidae are one of the most diverse and widespread families of obligate endoparasites of invertebrates and vertebrates. Relevant part of the success of these early-branching protists relies on their large repertoires of heat shock proteins (Hsp). The members of Hsp70 superfamily are chaperones that play a range of physiological roles, and contribute to the trypanosomatid adaptation, differentiation, and surviving in invertebrates and all classes of vertebrates under stressful situations such as shifts of temperature and pH, extra- and intracellular environments, and oxidative stress.

Methods: We compared the whole Hsp70 superfamily genetic repertoires of *Trypanosoma brucei, Trypanosoma cruzi*, and *Leishmania* spp. with Hsp70 paralogs from 19 draft genomes of 19 *Trypanosoma* spp. from cold and warm-blooded hosts, plus genomes from *Leishmania* spp., *Endotrypanum*, and 21 species of nine genera of insect trypanosomatids.

Results: Along with the Hsp70 (canonical), the repertoire of trypanosomatid Hsp70 (TryHsp70) superfamily comprises one mitochondrial (Lc2.2) and one reticular (Grp78) family, all homologs to Hsp70 of eukaryotes in general. In addition, the cytosolic families HSP70.4 andHSP70.c, which have no clear relatives in humans, are also ubiquitous in the kinetoplastids. The analyses of 618 TryHsp70 paralogs revealed that the newly discovered Hsp70.a and Hsp70.b families are highly divergent in the trypanosomatids and ubiquitous in the kinetoplastids, and corroborated that Hsp70.a originated by gene duplication in the lineage of the Grp170 (Hsp70ER) and Hsp110 families. In contrast to the ubiquitous mtHsp70/Lc2.2 family originated from the Alphaproteobacteria that gave rise to the mitochondria, neither trypanosomes nor bodonids share the mtHsp70/Lc2.1 paralogs identified in species of *Leishmania*, *Endotrypanum, Leptomonas*, *Crithidia*, *Angomonas*, *Strigomonas, Phytomonas, Sergeia* and *Herpetomonas.* Our analysis supported that the controversial Hsp70.b family was acquired through a single and ancient HGT from a DnaK-like donor. The TryHsp70 genes are located in highly syntenic regions, the only exception being the Grp78 genes, which are located in regions that are partially syntenic among species of distantly related trypanosomes. Finally, divergent DnaK or DnaK-like paralogs were identified exclusively in the genomes of the endosymbionts harbored by the insect trypanosomatids of the genera *Angomonas* and *Strigomonas*.

Conclusions: Our taxon-rich and wide genomic phylogenetic analyses supported a complex of ten families in the TryHsp70 superfamily. Evolutionarily conserved families of eukaryotic Hsps are shared by all trypanosomatids (CAN Hsp70, mtHsp70/Lc2.2, Grp78, Grp170 and Hsp110); Hsp70.4 and Hsp70.c group with CY Hsp70 paralogs from other eukaryotes, the mtHsp/Lc2.1 family is present in all genera of trypanosomatids investigated except *Trypanosoma*, and the recently discovered Hsp70.a and Hsp70.b families are ubiquitous in trypanosomatids, identified in other kinetoplastids, but absent in other eukaryotes. Phylogenetic analyses of orthologs from each family corroborated, in general, the main clades supported by traditional phylogeny of Trypanosomatidae, with the best resolved phylogenies of *Trypanosoma* obtained for Hsp70.a, CAN Hsp70, Hsp70.b, and Hsp110, thus indicating that Hsp70 genes are valuable phylogenetic markers regardless of their evolutionary origin.

Keywords: Heat Shock Proteins, Trypanosomatidae, chaperones, phylogeny, evolution, comparative genomics.

INTRODUCTION

The Trypanosomatidae (Euglenozoa: Kinetoplastidea) are a successful and widespread family of early-branching unicellular eukaryotes (protists), all obligate endoparasites of invertebrates and vertebrates, which evolved from free-living bodonids (Simpson et al., 2004; Jackson et al. 2008, 2016).*Trypanosoma* comprises species that develop extracellularly and/or intracellularly with complex and largely different life cycles in vertebrates of all classes, and are transmitted by a range of vectors such as leeches, ticks, fleas, hemipterans, and dipterans (Stevens et al., 2001; Hamilton et al., 2007). In contrast, *Leishmania* spp. are intracellular parasites exclusively of mammals and lizards, transmitted by sand flies. The Trypanosomatidae family also harbors several genera of monoxenous species of insects and species of the genus *Phytomonas*, which comprises plant parasites transmitted by phytophagous hemipterans (Camargo, 1999; Teixeira et al., 2011; Borghesan et al., 2013; Maslov et al., 2013; Lukes et al., 2014).

The great majority of trypanosomatids are not harmful to their usual hosts in natural transmission cycles. A few species of trypanosomatids are pathogenic to their vertebrate hosts; for example, *T. brucei* and *T. cruzi,* respectively the etiologic agents of Chagas disease, and African sleeping sickness in humans and Nagana in livestock. *Leishmania* spp. cause human diseases through the world and some *Phytomonas* spp. are plant pathogens. Host range studies suggested relevant restriction of most species of trypanosomes to classes, orders, genera or even species of vertebrate hosts, even though an increasing number of species have been reported in phylogenetically distant hosts (Hamilton et al., 2007, 2012; Garcia et al., 2011; Truc et al., 2013; Fermino et al., 2015; Lima et al., 2015).

The switching from aquatic to parasitic life and the adaptation to invertebrates and cold and warm-blooded vertebrates rely on the capacity of the kinetoplastids to respond to drastic environmental changes. The success of the heteroxenous trypanosomatids depends on the overcoming of an assortment of stressful conditions during their passage from invertebrate to vertebrate hosts. Heat shock proteins play important roles in their adaptation to extreme environmental changes in their life, such as shifts in temperature and pH, nutritional stress, hypoxia, oxidative stress, and process of host defenses, which are all potential triggers of cellular stress (Folgueira and Requena, 2007; Louw et al., 2010; Requena et al., 2015). To deal with stressful challenges, trypanosomatids respond using a plethora of mechanisms, including those involving the evolutionarily conserved Hsps (Folgueira and Requena, 2007; Louw et al., 2010; Requena et al., 2015; Tyc et al., 2015).

The eukaryotes possess multiple Hsps that cooperate in a variety of cellular processes and are located in the cytosol (CY), endoplasmic reticulum (ER), mitochondria (MT), and chloroplast (CP). The Hsps are classified, according to amino acid sequence homology and molecular weight, as Hsp100, Hsp90, Hsp70, Hsp60, Hsp40, and small Hsps (Young et al., 2004; Kriehuler et al., 2010). The members of the Hsp70 superfamily are abundant and important molecular chaperones that require the coordinated action of the Hsp40 and Hsp90 chaperones. Studies on the Hsp70s functions have emphasized their essential participation in protein homeostasis at both physiological and stressful conditions, promoting proper folding and assembly of

newly synthesized proteins, refolding and repairing of proteins, and acting in the proteolytic degradation and membrane translocation of organellar and secretory proteins (Young et al., 2004; Mayer and Bukau 2005; Kabani and Martineau, 2008; Kriehuler et al., 2010).

The number of Hsp70s varies widely in the eukaryotes, from 17 in humans to six in *Plasmodium falciparum* (Daugaard et al. 2007; Przyborski et al. 2015). Similar to other eukaryotes, the Hsp70 of the trypanosomatids is a superfamily comprising constitutive and inducible members localized in the cytoplasm, mitochondria, and endoplasmic reticulum. Hsp70s are known to be expressed in response to marked changes during the life cycles of *T. brucei*, *T. cruzi*, and *L. major* (collectively known as the Tritryps) triggered by the drastic environmental differences between insect vector and human host. Despite the small number of trypanosomatids examined, an expanded repertoire of Hsp70 families has been unveiled by genomic analyses on the Tritryps. Several families are homologues of major eukaryotic Hsp70s: the cytoplasmatic CAN Hsp70 (the prototypical/canonical Hsp70 referred in this study as CAN) and Hsp110 families, the mtHsp70/Lc2.2 family resident in the MT, and those resident in the ER assigned to the Grp78 and Grp170 families. The families Hsp70.4, Hsp70.c, which clustered with prototypical cytoplasmic CAN HSP70 and Grp78 and have no clear orthologous in humans, are also ubiquitous in the kinetoplastid. In addition, recent studies have revealed a number of paralogs, new structural features, and novel families in the Tritryps, such as Hsp70.a, Hsp70.b and the mtHsp70/Lc2.1 (Folgueira and Requena 2007; Campos et al., 2008; Louw et al., 2010; Shonhai et al., 2011; Tyc et al., 2015).

Proteomics analyses have identified Hsps as one of the major protein groups in the kinetoplastids and highlighted the complexity of their expression patterns and putative functions. TryHsp70s play a range of roles in innate and adaptive immunity of trypanosomatids infective to mammals (Urmenyi et al., 2014). Hsp70 of the fish parasite *T. carasii* up regulates the expression of nitric oxide synthase and induced a strong nitric oxide response of fish macrophages (Oladiran and Belosevic 2009). TryHsp70s are highly immunogenic, and have been investigated as targets for vaccines and serological diagnosis (Krautz et al., 1998; Levy et al., 1992; Flechas et al. 2009). In addition, uniqueness in the structure of the TryHsp70s has provided opportunities for the discovery of new drug targets (Requena et al. 2012; Urmenyi et al., 2012; 2014). In addition, inter- and intra-specific polymorphisms of CAN Hsp70s genes in *Trypanosoma* and *Leishmania* have been explored for molecular diagnosis. Recently, CAN Hsp70 family have proposed the use of Hsp70 for phylogenetic inferences of trypanosomes and leishmanias (Fraga et al. 2010, 2012, 2014, 2016).

Systematic analyses of the genetic repertoire of the TryHsp superfamily are so far restricted to the Tritryps. Although separated by a large phylogenetic distance, *T. brucei* and *T. cruzi* do not represent trypanosomatid diversity. The majority of trypanosome species differ from *T. brucei* and *T. cruzi* in many traits of their life cycle, vertebrate hosts, vectors, cell compartments, metabolism, and host-parasite-vector interactions. Data about Hsp genes and functions in the non-pathogenic species infecting mammals are very limited; and almost nothing is known regarding Hsps of the most basal trypanosomes of the Aquatic clade of *Trypanosoma* (Stevens et al., 2001; Hamilton et al., 2007; Ferreira

et al., 2008; Fermino et al., 2015). In addition, Hsp families in the genomes of trypanosomatids restricted to insects were not systematically investigated or analyzed in a phylogenetic framework.

Extensive genomic surveys of the Hsp70 superfamily in species representing the evolutionary diversity of the Trypanosomatidae, detailed analysis of the structural features, and classification of all identified Hsp70 paralogs by phylogenetic inferences are required to unveil the whole repertoire and properly hypothesize on the evolutionary processes that shaped the unique TryHsp70 repertoire. Our main goal in this study was to expand the knowledge on the genetic repertoire of the trypanosomatids representing several genera, with emphasis in the genus *Trypanosoma.* With this purpose, we characterized a large number of Hsp70 paralogs retrieved from the most comprehensive set of trypanosomatid genomes compared to date. The data set was employed to evaluate the composition of each Hsp70 family through phylogenetic inferences, which together with synteny analyses provides new and relevant insights to the complex evolutionary history of the TryHsp70 superfamily.

MATERIALS AND METHODS

Organisms and genomes

The draft genomes of the following organisms were used in this study (TCC- Trypanosomatid Culture Collection codes) (Table 1): *T. serpentis* (TCC1052), *T. theileri* (TCC165), *T. rangeli* (TCCAM80), *T. lewisi* (TCC34), *T. conorhini* (TCC025E), *T.dionisii* (TCC211), *T. erneyi* (TCC1946), *T. wauwau* (TCC1873), *T. livingstonei* (TCC1270), *T. ralphi* (TCC1838), *T. cruzi marinkellei* (B7), *T. cruzi marinkellei* (TCC344), *T. cyclops* (TCC52) *Trypanosoma* sp. (TCC1825), *Trypanosoma* sp. (TCC878), *Trypanosoma* sp. (TCC339), *Trypanosoma* sp. (TCC16 – H25), *Trypanosoma* sp. (TCC1307), *Trypanosoma* sp. (TCC2186), *T. cruzi* (G), *T. brucei* (TREU 927), *T. brucei* (Lister 427), *T. brucei gambiense* (DAL972), *T. vivax* (Y486) and *T. congolense* (IL3000), *T. grayi* (ANR4), and *T. evansi* (STIB805).

As out-groups of *Trypanosoma*, the phylogenetic analyses of TryHsp70 repertoire included: *Leishmania major (*MHOM/IL/80/Friedlin*), Leishmania braziliensis* (MHOM/BR/75/M2904*), Leishmania enriettii* (LEM3045), *L. martiniquensis* (MARLEM2494), *L. tarentolae* (ParrotTarII), *Angomonas deanei* (TCC036E), *Angomonas desouzai* (TCC079E), *Strigomonas culicis* (TCC012E), *Strigomonas galati* (TCC219), *Strigomonas oncopelti* (TCC290E), *Herpetomonas muscarum* (TCC001E), *Crithidia acanthocephali* (TCC037E), *Crithidia luciliae thermophila* (TCC050E), *Crithidia fasciculata* (Cf-Cl), *Lotmaria passim* (SF), *Endotrypanum schaudinni* (TCC224), *Endotrypanum monterogeii* (LV88), *Leptomonas costaricensis* (TCC169E), *Phytomonas* sp. (TCC066E), and *Phytomonas* sp. (HART1).

The genomes of *Angomonas deanei*, *Angomonas desouzai*, *Strigomonas culicis*, *Strigomonas galati*, *Strigomonas oncopelti*, *Herpetomonas muscarum*, and *Crithidia acanthocephali* were sequenced as previously described (Alves et al., 2013). *Endotrypanum schaudinni*, *Leptomonas costaricensis, T. cruzi*, *T. rangeli*, *T. lewisi*, *T. conorhini*, *T.dionisii*, *T. erneyi*, *T. cruzi marinkellei, Trypanosoma* sp. (TCC339) were sequenced using standard pyrosequencing shotgun methodology according to Roche 454 protocols. *T. serpentis* (TCC1052), and *T. theileri* (TCC165), *T. wauwau* (TCC1873), *T. livingstonei* (TCC1270), *T. ralphi*

(TCC1838), *T. cyclops* (TCC52) *Trypanosoma* sp. (TCC1825), *Trypanosoma* sp. (TCC878), *Trypanosoma* sp. (TCC16 – H25), *Trypanosoma* sp. (TCC1307), *Trypanosoma* sp. (TCC2186) were sequenced using the MiSeq Illumina plataform (paired-end reads of 2 x 300 bp), and assembled using Newbler software (version 2.9).

As outgroup of the Trypanosomatidae, the analyses included *Bodo* sp. (ATCC50149), *Parabodo caudatus* (ATCC30905), and *Bodo saltans* (Welcome Trust Sanger Institute).

All trypanosomatids are cryopreserved at the Trypanosomatid Culture Collection of the University of São Paulo (TCC-USP) and were grown in LIT or TC100 media supplemented with 2-5% FBS (Fetal Bovine Serum). Information regarding genomes used in this study was summarized in Table 1.

Table 1: Trypanosomes, other trypanosomatids and free living kinetoplastids and euglenids genomes used in this study.

Organism	Sample code	Host	Data origin
Trypanosomatida			
Trypanosoma cyclops	TCC 052	*Macaca ira*	ATOL - USP
Trypanosoma livingstonei	TCC 1270	*Rhinolophus landeri*	ATOL - USP
Trypanosoma wauwau	TCC 1873	*Pteronotus personatus*	ATOL - USP
Trypanosoma ralphi	TCC 1838	*Melanosuchus niger*	ATOL - USP
Trypanosoma serpentis	TCC 1052	*Pseudoboa nigra*	ATOL - USP
Trypanosoma evansi	STIB 805	*Bubalus bubalis*	TriTrypDB
Trypanosoma brucei brucei	Lister 427	*Ovis aries*	TriTrypDB
Trypanosoma brucei brucei	TREU927	*Glossina pallidipes*	TriTrypDB
Trypanosoma brucei gambiense	DAL972	*Homo sapiens*	TriTrypDB
Trypanosoma congolense	IL30000	*Bos* sp.	TriTrypDB
Trypanosoma grayi	ANR4	*Glossina palpalis*	TriTrypDB
Trypanosoma vivax	Y486	*Bos taurus*	TriTrypDB
Trypanosoma theileri	TCC 165	*Bubalus bubalis*	ATOL
Trypanosoma lewisi	TCC 34	*Rattus rattus*	ATOL
Trypanosoma rangeli	AM80	*Homo sapiens*	ATOL
Trypanosoma conorhini	TCC 025E	*Rattus rattus*	ATOL
Trypanosoma dionisii	TCC 211	*Eptesicus brasiliensis*	ATOL
Trypanosoma erneyi	TCC 1946	*Mops condylurus*	ATOL
Trypanosoma cruzi cruzi	G	*Didelphis marsupialis*	ATOL
Trypanosoma cruzi marinkellei	TCC 344	*Carollia perspicillata*	ATOL
Trypanosoma cruzi marinkellei	B7	*Phyllostomus discolor*	TriTrypDB
Trypanosoma sp.	TCC 339	*Rhinella marina*	ATOL
Trypanosoma sp.	TCC 878	*Mabuya frenata*	ATOL - USP
Trypanosoma sp.	TCC 1825	*Ramphocelus nigrogularis*	ATOL - USP
Trypanosoma sp.	TCC 16	*Macropus giganteus*	ATOL - USP
Trypanosoma sp.	TCC 1307	*Ptychadena mossambica*	ATOL - USP
Trypanosoma sp.	TCC 2186	*Leptodactylus latrans*	ATOL - USP
Herpetomonas muscarum	TCC 001E	*Musca domestica*	ATOL (AUXJ01000000)
Phytomonas sp.	HART1	*Cocos nucifera*	GenBank CAVR000000000.2
Phytomonas sp.	TCC 066E	*Jatropha macrantha*	ATOL
Angomonas deanei	TCC 036E	*Zelus leucogrammus*	ATOL (AUXM01000000)
Angomonas desouzai	TCC 079E	*Ornidia obesa*	ATOL (AUXL01000000)
Angomonas ambiguus	TCC 2435	*Chrysomya albiceps*	ATOL
Strigomonas culicis	TCC 012E	*Aedes vexans*	ATOL (AUXH01000000)
Strigomonas galati	TCC 219	*Lutzomya almerioi*	ATOL
Strigomonas oncopelti	TCC 290E	*Oncopeltus fasciatus*	ATOL (AUXK01000000)
Leptomonas costaricensis	TCC 169E	*Zelinae* sp.	ATOL
Lotmaria passim	SF	*Apis mellifera*	AHIJ00000000
Crithidia acanthocephali	TCC 037E	*Acanthocephala femorata*	ATOL (AUXI00000000)
Crithidia luciliae thermophila	TCC 050E	*Zelus leucogrammus*	ATOL - USP
Crithidia fasciculata	Cf-CI		TriTrypDB

Endotrypanum monterogeii	LV88	*Choloepus hoffmani*	TriTrypDB
Endotrypanum schaudinni	TCC 224	*Choloepus hoffmani*	ATOL
Leishmania enriettii	LEM3045	*Cavia porcellus*	TriTrypDB
Leishmania martiniquensis	MARLEM2494	*Homo sapiens*	TriTrypDB
Leishmania braziliensis	MHOMBR75M2904	*Homo sapiens*	TriTrypDB
Leishmania donovani	BPK282A1	*Homo sapiens*	TriTrypDB
Leishmania gerbilli	LEM452	*Homo sapiens*	TriTrypDB
Leishmania major	Friedlin	*Homo sapiens*	TriTrypDB
Leishmania tarentolae	ParrotTarII	*Tarentola mauritanica*	TriTrypDB
Bodonida			
Parabodo caudatus	ATCC 30905	free living	ATOL
Bodo saltans	Lake Konstanz	free living	Wellcome Trust Sanger
Bodo sp.	ATCC 50149	free living	ATOL

Genomic search, sequence alignment, and phylogenetic analyses.

Families belonging to TryHsp70 repertoires were identified in the genomes present in TriTrypDB databank, and in draft genomes by similarity searches using BLAST algorithms. Previously reported sequences from Hsp70 families of *T. brucei* were used as queries (Folgueira and Requena, 2007). The genes were classified as paralogous or orthologous based on inferred phylogenetic relationships, sequence similarity, and comparison of functional domains and sequence motifs. The paralogs were designated according to the classification previously defined by Folgueira and Requena (2007).

To check for the existence of horizontal gene transfer, a comprehensive analysis was performed including Hsp70 sequences present in NR (Non-redundant protein sequences) (ftp://ftp.ncbi.nih.gov/blast /db/ as available in February 2015). A blastp search was performed with a maximum expect value threshold of 1e-20, using the sequences found in trypanosomatids as queries. The retrieved sequences were checked for Hsp70 domains using RPSBLAST against the Conserved Domain Database (http://www.ncbi.nlm.nih.gov/ Structure/bwrpsb/bwrpsb.cgi).

To reduce computational demand in comprehensive analysis, highly similar sequences representing closely related strains or species were removed using the USERACH (Edgar, 2010) program with 95% similarity and 90% length coverage thresholds. Sequences with less than 500 amino acids (70% of the average Hsp70 size) and larger than 1,300 amino acids, or possessing conserved domains not related to Hsp70 super-family, were also removed. Access numbers of sequences included in the comprehensive analysis are included in Table S2. In all TryHsp70 analyses, partial sequences with less than 500 amino acids were not included in the presented inferences. Sequences presenting frame shifts or multiple stop codons were also discarded. Multiple sequence alignment was performed using MUSCLE v3.8 (Edgar, 2004) and edited using Gblocks v0.91b (Castresana, 2000) to eliminate poorly aligned positions. The substitution model was chosen by the Akaike Information Criterion (AIC) using ProtTest3 (Abascal et al., 2005). Maximum likelihood analyses of the trypanosomatid-only dataset were performed using the LG substitution model, gamma-distributed heterogeneity rate categories, and estimated empirical residue frequencies (PROTGAMMALGF), as implemented in RAxML v8.2.0 (Stamatakis, 2014). Due to computational time constraints, the comprehensive Hsp70 superfamily phylogeny including NR sequences was performed with the approximately-maximum-likelihood method of FastTree v2.1 (Price

et al., 2010), using the WAG model and gamma-distributed rates. Phylogenetic trees were visualized using Figtree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/) and Dendroscope v3.2.4 (Huson and Scornavacca, 2012).

The recombination analysis done for mtHsp70 (Lc2.2 and Lc2.1) paralogs was carried out using RDP4 v4.24 software (Martin et al. 2010). Only events detected by at least three methods and with a p-value lower than 0.05 were considered as valid.

RESULTS AND DISCUSSION

1. Phylogenetic support for TryHsp70 families and new insights into evolutionary histories

Previous studies focused on the whole TryHsp70 superfamily repertoires were restricted to the TriTryp genomes and did not include phylogenetic analysis of most families (Folqueira and Requena, 2007, Louw et al., 2010). Here, we compared the whole repertoires of the Hsp70 superfamily reported for the TriTryp with those retrieved from the currently available genomes of 23 additional species of *Trypanosoma,* including 15 complete and annotated genomes from public data banks, and 28 draft genomes that have been sequenced in our laboratories (Table 1). Extensive surveys were also performed on the genomes of *Leishmania* spp, monoxenous trypanosomatids of eight genera, and *Phytomonas*. The phylogenetic analyses included all the 618 TryHsp70 paralogs retrieved from the genomes of trypanosomatids, in addition to paralogs from 1.108 species of other Although some TryHsp70 families present multiple identical copies organized in head-to-tail tandem organization (Folgueira and Requena, 2007), our searches revealed one sequence for each family in each genome. The existence of more than one identical copy of some families cannot be excluded due to the technical limitations (collapsing) in the assembling of tandem repetitive conserved genes using NGS methods. In addition, sequencing of several of the genomes used here is still in progress, thus better coverages and new assemblies will need to be checked regarding other Hsp70 paralogs undetected so far. eukaryotes and 4.303 from prokaryotes from NR database (Fig. 1).

The phylogenetic analyses of all TryHsp70 paralogs corroborated the remarkable split between the eukaryotic Hsp70s and the bacterial counterpart, DnaK.

Although some eukaryotic Hsp70 sequences clustered together with DnaK gene sequences, no prokaryotic sequence nested within the eukaryotic clades. A few virus sequences were found within the eukaryotic Hsp70 repertoire, and they appear to have been transferred many times to different lineages (Fig. 1).

Our phylogenetic analysis is consistent with the evolutionary history proposed for the eukaryotic Hsp families (Young et al., 2004; Mayer and Bukau, 2005; Kabani and Martineau, 2008; Kriehuler et al., 2010). The systematic and wide-ranging searches of TryHsp70 paralogs performed in the present study supported that most TryHsp70 families evolved from the eukaryotic ancestor by gene duplication, including the typical mtHsp70 (Folgueira and Requena, 2007) (Fig. 1).

Although located in the mitochondrial matrix, members of the conserved mtHsp70 family are encoded by nuclear genes and phylogenetically more closely related to DnaK than to any eukaryotic Hsp70 sequence (Germot and Philippe 1999; Folgueira and Requena, 2007). The accepted evolutionary hypothesis is that this family originated by HGT from the Alphaproteobacteria ancestor that gave rise to the modern mitochondria of extant eukaryotes (Gupta

and Golding, 1993; Gupta et al, 1994; Boorstein et al., 1994; Germot and Philippe, 1999). Our findings corroborate this evolutionary hypothesis, since all paralogs of trypanosomatid MT Hsp70 sequences (named mtHsp70 for trypanosomes and Lc.2.2 for leishmanias) clustered into a single monophyletic group adjacent to DnaK sequences from Alphaproteobacteria (Fig. 1). In contrast, paralogs of mtHsp70 Lc2.1 reported in *L. chagasi/L. infantum* (Campos et al., 2008) were not identified in any

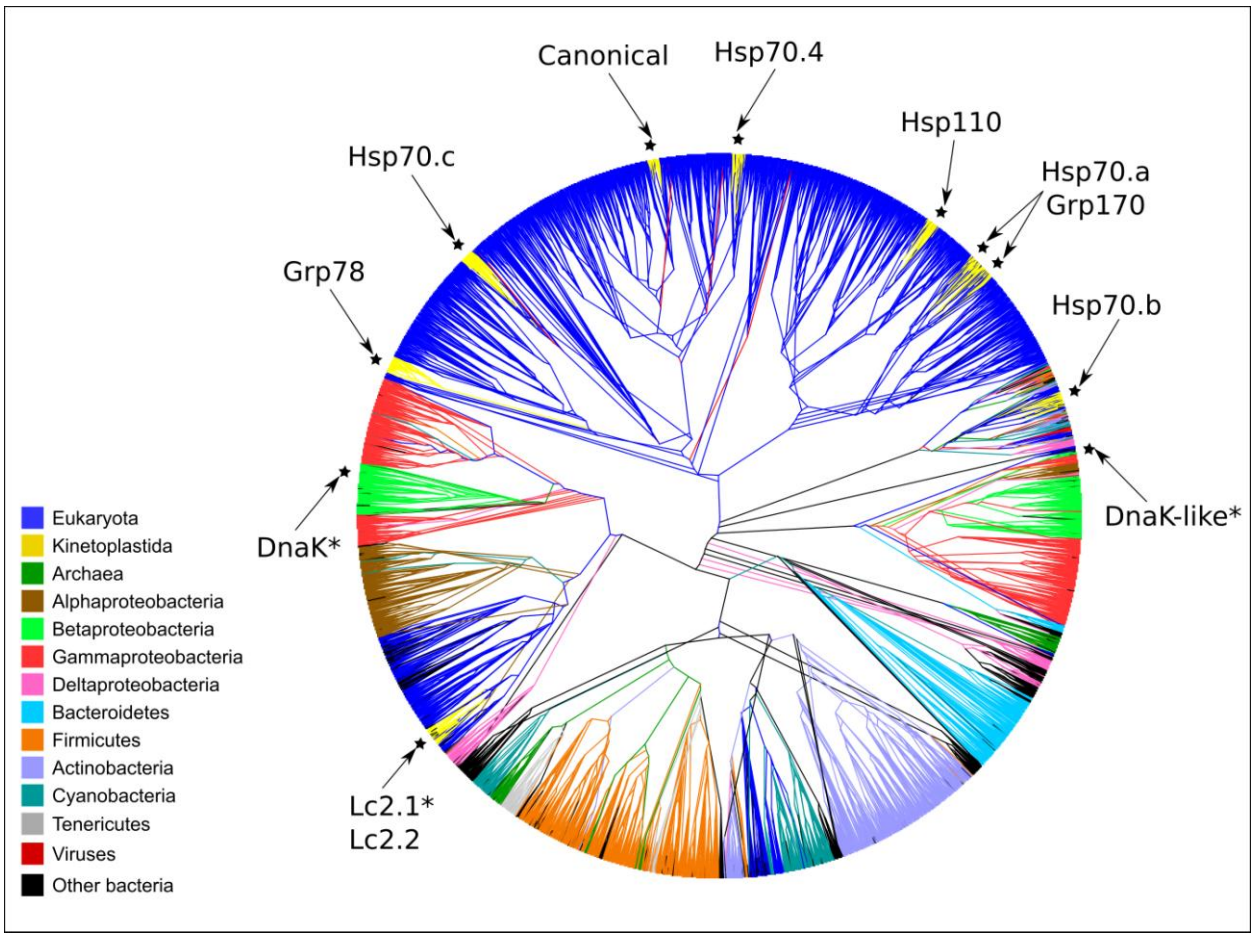

Figure 1: Comprehensive phylogenetic analysis including TryHsp70 families – Maximum Likelihood analysis using FastTree software (WAG + G substitution model) includes 9.733 Hsp70/DnaK sequences from 1.108 species of other eukaryotes and 4.303 from prokaryotes from NR database. Branches are painted according taxonomic groups. The * symbol in indicates Hsp70 families found in other trypanosomatids, but not found in Trypanosoma. The access number of TryHsp70 and NR Hsp70 sequences employed in the phylogenetic analysis are listed in Table S1 and S2 respectively.

trypanosome species, but seem to be present in all non-trypanosome trypanosomatids investigated. Therefore, Lc2.1 is a new TryHsp70 family that, according to our phylogenetic analysis, likely originated by gene duplication from Lc2.2 (Fig. 2) (see Section 6).

The recently identified Hsp70.a and Hsp70.b (Folgueira and Requena, 2007; Louw et al., 2010) were herein seen as ubiquitous in the Trypanosomatidae, and also present in other kinetoplastids. Our analysis supported that the Hsp70.a family derived from duplications in the Hsp lineage that gave origin to the Hsp110 and Grp170 families. Taking in account its absence in other eukaryotic lineages, Hsp70.a was most likely the last family to diverge in the NEFs (Nucleotide exchange factors – Hsp110/Grp170) lineage of the kinetoplastids. It had been previously suggested that

the Hsp70.b is more similar to cyanobacterial DnaK than to the mit Hsp70s, then suggesting the existence of chloroplast in an ancestor of the trypanosomatids (Martin et al., 2003; Hannaert et al., 2003; Folgueira and Requena, 2007; Requena et al., 2015). This hypothesis was questioned by Louw et al. (2010), suggesting that Hsp70.b could be related to DnaK

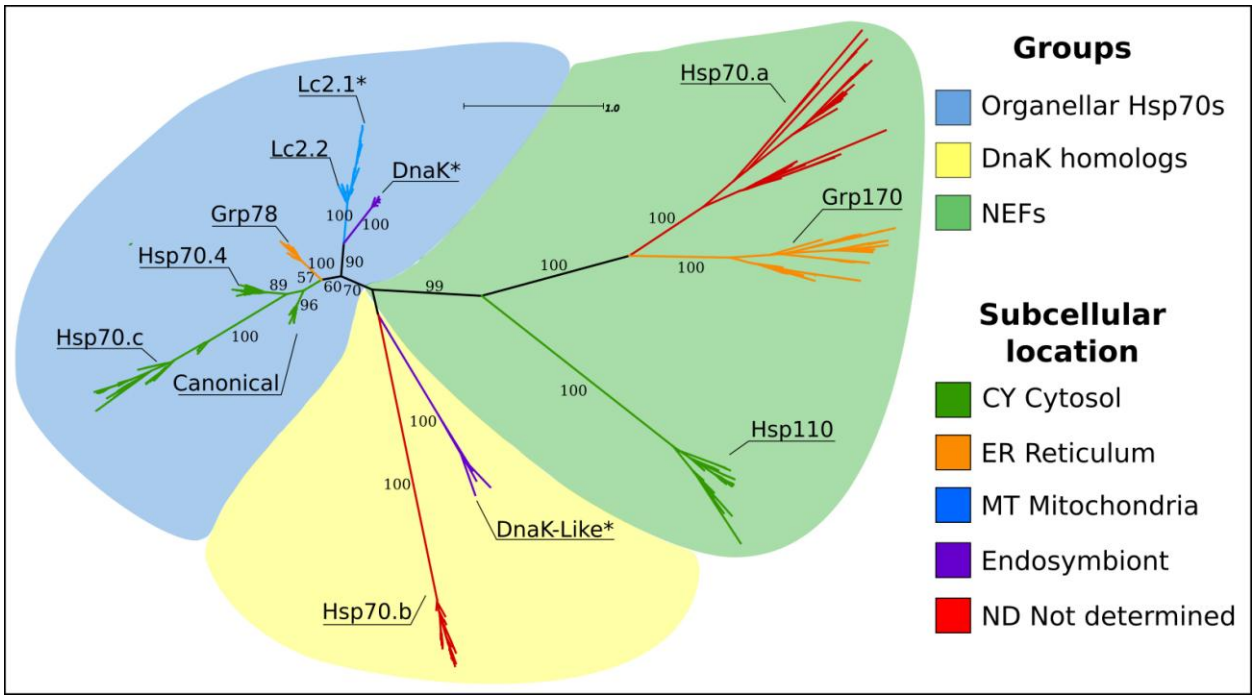

Figure 2: Phylogenetic relationships of the TryHsp70 families – Maximum Likelihood radial tree resulting from an analysis using RaxML version 8.1.15. The best substitution model was chosen by the Akaike Information Criterion (AIC) using the PROTGAMMAAUTO setting implemented in the program. Support was calculated using 300 bootstrap replicates. Groups of TryHsp70 families are highlighted according to function and branches according cellular localization

Our analysis of Hsp70 paralogs from eukaryotes and prokaryotes demonstrated for the first time that Hsp70.b paralogs are more closely related to DnaK-like genes than to the typical DnaK, and separated from any other TryHsp70 family. The TryHsp70.b branched with the DnaK-like from a wide range of bacterial groups, virus, flatworms, and single-celled eukaryotes. The clustering of Hsp70.b with DnaK from *Giardia* and *Entamoeba*, considered to be among some of the earliest-diverging lineages of extant protists, Fungi (*Microsporidium* and Ascomycetes) and green algae (Fig. 1, S1), suggested that Hsp70.b likely resulted from a very ancient HGT from a prokaryotic gene into a primitive eukaryote. The existence of Hsp70 genes more similar to bacterial DnaK than to eukaryotic genes was previously reported in *Giardia, Entamoeba* and microsporidians, and these odd DnaK homologs were supposed to be vestiges of the lost mitochondria (Aggarwal et al., 1990; Bakatselou et al., 2000; Morrison et al., 2001; Arisue et al., 2002). However, in our phylogenetic analysis, TryHsp70.b and these previously reported eukaryotic DnaK homologs were placed far from Alphaproteobacteria. Indeed, the phylogenetic position of these sequences together with Hsp70.b indicates a possible new bacterial DnaK group shared by single celled eukaryotes. However, the great diversity of bacteria distributed across the largely unresolved branch did not allow estimating neither number of HGT events nor donors.

Here, phylogenetic analyses including all TryHsp70 genes strongly supported that TryHsp70

sequences clustered in one of the three main Hsp70 monophyletic groups of eukaryotes in general: organellar, DnaK, and NEF (Young et al., 2004; Mayer and Bukau 2005; Kabani and Martineau, 2008; Kriehuler et al., 2010) (Figs 1, 2).

The TryHsp70 repertoire includes cytosolic and closely related families of the organellar group (CAN Hsp70, Hsp70.4 and Hsp70.c), separated from the also cytosolic Hsp110 of the NEF group (Fig. 2). Similarly, the Hsp70 families coding for ER residents (Grp78 and Grp170) clustered in different groups, organellar and NEF, respectively. Our analysis provides strong support to the hypothesis that Hsp70.a (unknown location) may have originated by gene duplication in the lineage that also originated Hsp110 (CY) and Grp170 (ER). The scenario observed in the NEF group is similar to that found in Hsp70s of the organellar group, where related paralogs were also segregated in different (CY and ER) cellular compartments (Fig. 2).

In conclusion, ten TryHsp70 families were identified in the trypanosomatids, of which seven have conserved homologs in eukaryotes in general, the exceptions beingHsp70.a, Hsp70.b and Lc2.1; nine of the families are shared by all trypanosomatids (the exception being Lc2.1); and one family (Lc2.1) is present in a range of non-trypanosome trypanosomatids, but is lacking from all trypanosomes (Fig. 2). The extensive and broad analyses of TryHsp70 paralogs strongly suggests that the two recently discovered and highly divergent Hsp70.a and Hsp70.b families are ubiquitous in the trypanosomatids. Preliminary data from our laboratories have suggested that Hsp70.a and Hsp70.bhomologues are present in all kinetoplastids. Further systematic studies are required to investigate the Hsp70 repertoires in free-living kinetoplastids, euglenids, and diplonemids, which are the nearest relatives of the kinetoplastids (Simpson et al., 2004; Jackson et al. 2008, 2016).

Finally, our searches also disclosed two DnaK paralogs exclusively in the genomes of the bacterial endosymbiont harbored by trypanosomatids of the genera *Angomonas* and *Strigomonas* (Fig. 1) (see section 9).

3. Sequence structures and motifs typical of each trypanosomatid Hsp70 family

The members of the Hsp70 superfamily typically contain three conserved domains: At the N-terminal lies the 45-kDa ATP binding/ATPase or nucleotide-binding domain (NBD), followed by the peptide-binding domain or substrate-binding domain (SBD) at the C-terminal, containing the conserved EEVD motif necessary for association with co-chaperones (Mayer and Bukau, 2005; Kominek et al., 2013). Here, multiple alignments of TryHsp70 gene sequences reveals 22 conserved blocks shared by all paralogs. Twelve blocks are located in the NBD, nine in the SBD, and between the binding domains is the Linker. Conservation decreases along the molecule, being higher in the N-terminus. The Linker and C-terminal motifs enabled an easy visual differentiation among the paralogs of the different Hsp70 families (Fig. 3).

The CY Hsp70 with the C-terminal EEVD motif shared by stress inducible Hsp70s from eukaryotes in general is considered the prototypical Hsp70 (CAN Hsp70), and represents the most studied family of TryHsp70 (Freeman et al., 1995; Louw et al., 2010). As expected, the trypanosomatids share highly conserved CAN Hsp70 characterized by the EEVD motif at the end of the C-terminal, and show a variable number of GGMP repeats at the C-terminal. CAN Hsp70 and Hsp70.4 show an identical LLLLD

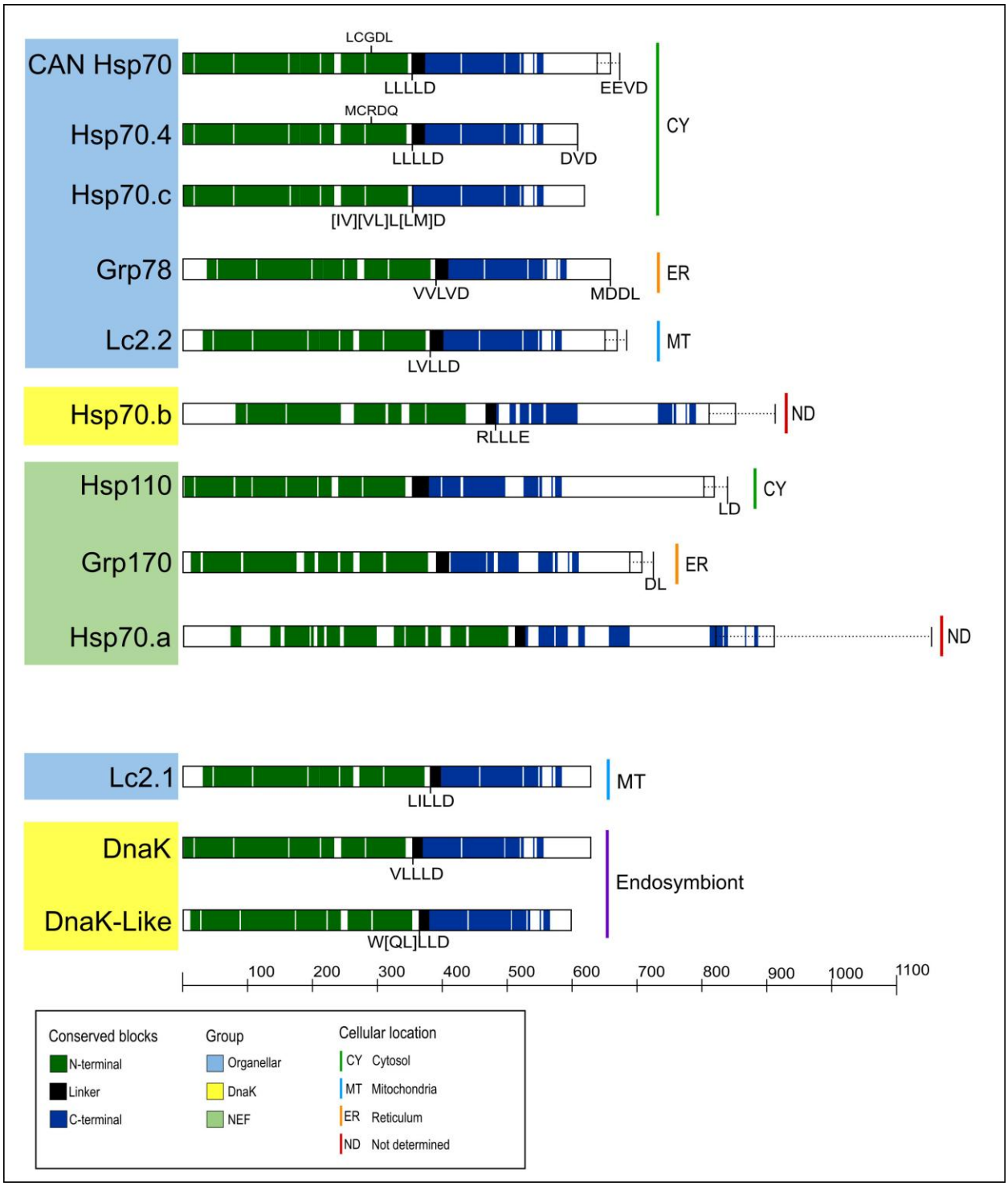

Figure 3: Architecture of TryHsp70s families – Schematic representation of conserved blocks, motifs and domains of TryHsp70 family found in *Trypanosoma* and related proteins found in other trypanosomatids The NBD and SBD conserved blocks are shown in green and blue respectively, the Linker regions are shown in black. Insertions in relation of typical organellar Hsp70s are painted according conservation degree. Sequence motifs are also indicated using amino acids single letter code. The * symbol after names indicates Hsp70 families exclusive from non trypanosome trypanosomatids. Boxes indicate the corresponding group which each family was classified in phylogenetic analysis Figure 2.

Linker motif in the C-terminus while neighboring motifs, LCGDL or MCRDQ, distinguish between the sequences of the two families. Instead of EEVD, Hsp70.4 presents a conserved DVD motif at the end of the C-terminus, and lacks GGMP or similar repetitions (Fig. 3). Differently from other CY

paralogs, Hsp70.c sequences display polymorphisms in substrate binding domains, lack the conserved C-terminus motif, and exhibit a polymorphic [VI][VL]L[ML]D Linker. The ER resident Grp78 paralogs show a conserved VVLVD Linker motif, which is different from those of all other *Try*Hsp70 families, and exhibit in the end of the C-terminus the ER retention motif MDDL (Fig. 3), which is responsible for the retention of the Grp78 proteins at the ER lumen (Bangs et al., 1996).

Regarding the MT members of the TryHsp70 families, Lc.2.2 and Lc2.1, a short Linker allowed to separate both Lc.2.1 and Lc2.2 from all other *Try*Hsp70s. In addition, there are clear differences between Lc2.2 and Lc2.1: Lc2.2 has a conserved LVLLD motif in the Linker region, whereas Lc2.1 presents an LILLD motif, and Lc2.2, but not Lc2.1, has a variable number of SN and Q repeats at the C-terminal end (Fig. 3). Therefore, the results from the analysis of a large number of paralogs from several trypanosomatid genera agreed that the main difference between Lc2.2 and Lc2.1 relied on the longer C-terminal extension of Lc2.2 homologs, as reported before for *Leishmania* (Campos et al., 2008).

As reported for other eukaryotes (Easton et al., 2000), the structural organizations of the Hsp70 families are highly different among the members of the group NEF. The Hsp110 displays a conserved LD motif at the end of the C-terminus and a large and polymorphic Linker region. Like the CY Hsp110, the RE resident Grp170 displays a large and poorly conserved Linker region and a DL motif at the end of the C-terminus, which fits the final DL motif present in the also RE resident Grp78 nested into the organellar group (Fig. 2). Like other NEFs, Hsp70.a possesses large insertions in the substrate and nucleotide bind domains (Fig. S2). Although the functions of HspP110 and Grp170 remained elusive, it is believed that they do not act as protein folding chaperones, acting as a holdases for misfolded proteins instead, preventing their aggregation until the native fold state is recovered by typical Hsp70s (Easton et al, 2000; Bracher and Verghese, 2015).

The analysis of Hsp70.b orthologs, which formed the DnaK-like group within the TryHsp70 superfamily, revealed a conserved RLLLE Linker motif, similar to those found in members of the organellar group: CAN Hsp70, Grp78 and Lc2.2. Indeed, Hsp70.b Linker sequences exhibits a leucine rich repetition and a negatively charged amino acid (usually aspartic or glutamic acids) in the last position. On the other hand, Hsp70.b exhibited large insertions typical of the NEF group in the NBD and SBD (Fig. 3).

4. Phylogenetic analyses of the cytosolic CAN Hsp70, Hsp70.4 and Hsp70.c of trypanosomatids

Homologues of the three conserved CY Hsp70s families, CAN Hsp70, Hsp70.4 and Hsp70.c, previously identified in the TriTryp genomes, were identified in all trypanosomatid genomes examined herein. Due to high evolutionary conservation, genes encoding CAN Hsp70 genes can be easily isolated, and were one of the first Hsp70 genes identified in the trypanosomatids (Glass et al., 1986; Engman et al., 1989, 1992; Requena et al., 1992). In *T. cruzi*, CY Hsp70s are expressed in all parasite stages, and migrate to the nucleus upon heat shock and when cells reach the stationary phase (Requena et al., 1992; Martin et al., 1993; Folgueira et al., 2005; Urményi et al., 2012, 2014; Nazer et al., 2012). *T. brucei* and *T. cruzi* were described as having multiple copies of CAN Hsp70 in head-to-tail tandem arrangements, in addition to a few isolated genes (Requena et al., 1988; Folgueira et al., 2005;

Folgueira and Requena, 2007). In our searches, only one or two complete copies showing identical CAN Hsp70 sequences were recovered from the genomes, certainly because of the collapsing of identical tandem repeats.

The Hsp70.4 was first identified as a single copy gene in *L. major* (Searle et al., 1989), and then reported in the genomes of *T. brucei* (single copy) and *T. cruzi* (2-3 copies) (Folgueira and Requena, 2007). The cytosolic HSP70.4 is highly enriched in amastigotes but undetectable in trypomastigotes of *T. cruzi* (Atwood et al., 2005). Homologues of Hsp70.c were identified in *T. brucei* (single copy) (Lee et al., 1990), and the expression levels were increased in the parasite by heat stress (Burger et al., 2014). Conserved homologs of Hsp70.c were identified by genome searches in the TriTryps (Folgueira and Requena, 2007).

A single copy of each Hsp70.4 and Hsp70.c genes was identified in our analyses. Genes encoding CAN Hsp70, Hsp70.4, and Hsp70.c clustered into a monophyletic group. In agreement with the branching patterns of the genealogies, CAN Hsp70 sequences were closer to those of Hsp70.4 family (~ 67% amino acid sequence similarity) than to those of Hsp70.c paralogs (~46%). Also supported by the phylogenetic analysis, Hsp70.c is the most divergent of the CY Hsp70 families that nested into the organellar group (Fig. 2), diverging ~ 44% in their amino acid sequences from both CAN Hsp70 and Hsp70.4 gene sequences. Within each family, the similarities among the orthologs ranged from 95% for CAN Hsp70, ~85% for Hsp70.4, and ~72% for Hsp70.c.

CAN Hsp70 sequences from the same organism share nearly identical sequences that cluster tightly together forming species-specific clades, thus indicating concerted evolution in this family. The genealogies of CAN Hsp70, Hsp70.4 and Hsp70.c (Fig. 4A, B, C) supported the two major clades within the *Trypanosoma*, Aquatic and Terrestrial, as well as the relationships among the clades being in general congruent with those supported for the Trypanosomatidae using traditional phylogenetic markers (Hamilton et al., 2007, 2012; Garcia et al., 2011; Maslov et al., 2013; Fermino et al., 2013, 2015; Lukes et al., 2014; Lima et al., 2015). The analysis based on conserved CAN Hsp70 genes was highly concordant, corroborating the positioning of all major clades. However, better resolved intra-clade relationships and highly supported topologies were inferred using the more divergent Hsp70.4 and Hsp70.c (Fig. 4) (see section 10). In addition, CAN Hsp70, Hsp70.4 and Hsp70.c genealogies were largely consistent with the main clades formed by the monoxenous trypanosomatids of insects and *Leishmania/Endotrypanum*, confirming CAN Hsp70 (Fig. 4A) as suitable to infer relationships within *Leishmania* (Fraga et al., 2010, 2012).

5. Phylogenetic analyses of the trypanosomatid Grp78 family members

The ER resident members of the HSP70 family (ER Hsp70) in eukaryotes are often referred to as Grp78 or BiP (Easton et al., 2000). The first Grp78 homolog in kinetoplastids was identified in *T. brucei* (Bangs et al., 1993) and then in *T. cruzi* and not induced by heat shock (Tibbetts et al., 1994). The role played by Grp78 is not clear. Recently, Grp78 was shown binding and assisting in the folding and delivery of the lysosomal cysteine protease cruzipain to calreticulin (Labriola et al. 2011). In our searches, only a single Grp78 sequence was recovered from most trypanosome genomes. The exceptions were

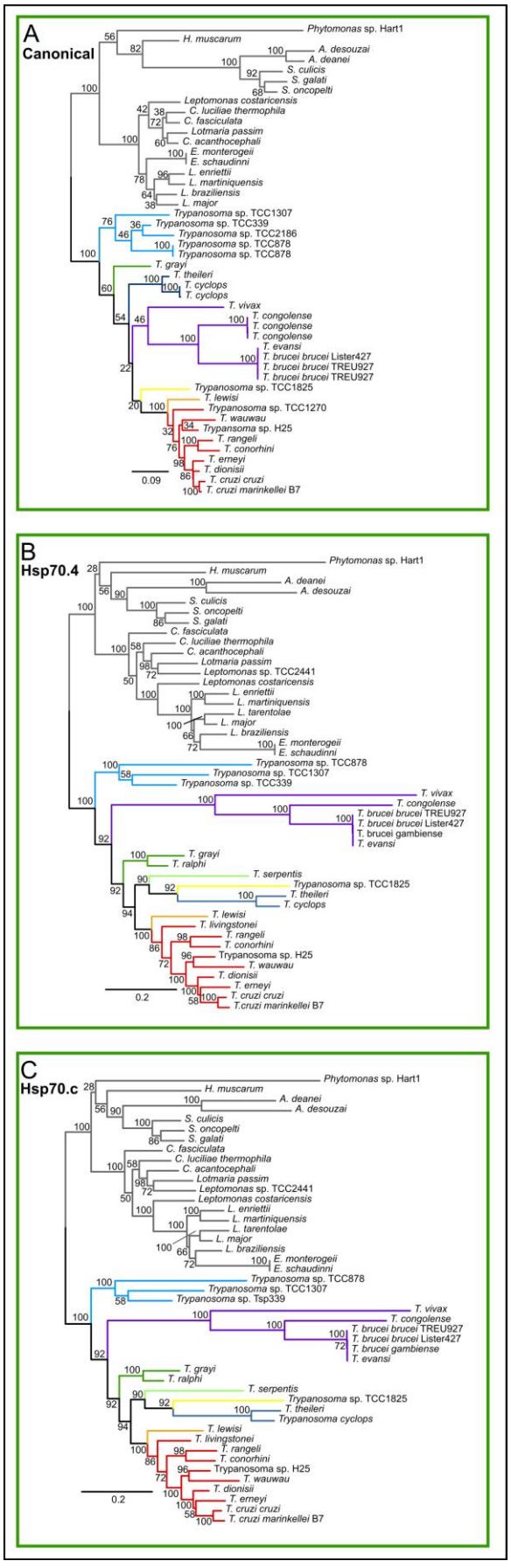

Figure 4: ML CY Hsp70 gene trees rooted using non-trypanosome trypanosomatids. Values at nodes are ML bootstrap values using 500 replicates. Branch colors represent the *Trypanosoma* clade, in which each OTU is classified according classical molecular markers gGAPDH and SSU rRNA. Letters A, B and C represent CAN Hsp70, Hsp70.4 and Hsp70.c respectively.

T. b. brucei, *T. b. gambiense*, *T. congolense*, and *T. conorhini,*all showing at least two nearly identical copies, located in the same scaffold, but not in tandem. Genome searches revealed two copies in *T. brucei* and a single copy in *T. cruzi* (Folgueira and Requena, 2007).

Amino acid sequences of Grp78 orthologs were highly homogeneous (~93% similarity) and closest (~63% similarity) to CAN Hsp70, in agreement with their positioning in the phylogenetic analysis (Fig. 3). All Grp78 paralogs obtained in our surveys were used for phylogenetic analysis of trypanosomatids. The topology of the inferred genealogy showed several incongruences compared to traditional phylogeny (Fig. 5) (see section 10). However, Grp78 genealogy strongly supported the recently inferred phylogeny of the *Leishmaniinae* subfamily (Fig. 5)

6. Phylogenetic analyses of trypanosomatid mtHsp70: Lc2.2 and Lc2.1 families

In trypanosomatids, orthologs of the MT Hsp70 were initially described in the mitochondria of *T. cruzi* in association with kinetoplast DNA (Engman et al., 1989), and afterward in *T. brucei* and *Leishmania* major (Klein et al., 1995). It is well know that mtHsp70 takes part in essential processes such as protein import into the mitochondrion across the organellar double membrane. A recent study demonstrated that the mtHsp70/mtHsp40 machinery of *T. brucei* plays a crucial role in DNA replication and maintenance in kinetoplast, a function likely retained from their bacterial ancestors (Engman et al., 1989; Tyc et al., 2015). Genome analysis of the TriTryps revealed

three identical mtHsp70 genes in *T. brucei,* whereas only one complete copy was identified in the *T. cruzi* genome, and five genes were identified in the genome of *L. major* (Folgueira and Requena, 2007). Here, tandem repeats could not be found in genome assemblies of *T. cruzi* (G) as showed before for other strains (Engman et al., 1989)

In our surveys, highly conserved orthologs of mtHsp70 (Lc.2.2) were identified across the Trypanosomatidae family sharing ~95% amino acid sequence similarity. All genes from the same genome cluster tightly together forming a monophyletic assemblage as shown for *T. brucei*, *T. congolense*, *T. cruzi* and *Leishmania* spp. (Fig. 6A), thus indicating that mtHsp70 has been homogenized by concerted evolution. As expected, the question on the number of copies remains unanswered for most genomes.

Two homologs of the mtHsp70 from *L. chagasi* and *L. infantum,* both distributed throughout the mitochondrial matrix (Campos et al., 2008), were designated as Lc2.2 (homologous to the prototypical mtHsp70) and Lc2.1 (a new Hsp70 family). Trypanosomatids possess multiple copies of the mtHsp70/Lc2.2, which are expressed in *Leishmania* spp. promastigotes and *T. cruzi* epimastigotes. In contrast, a single copy of the Lc2.1 paralog was identified in *L. chagasi* and *L. infantum*, presumably expressed in amastigotes, thus suggesting novel mtHsp70 functions evolving within the genus *Leishmania* (Campos et al., 2008).

Our findings reveal that, in addition to *T. brucei* and *T. cruzi* (Campos et al., 2008), the genomes of the 23 species of trypanosomes examined herein lack homologs of Lc2.1. In contrast, Lc2.1 was identified in the genomes of all species of *Leishmania*, *Endotrypanum*, *Leptomonas*, *Crithidia*, *Angomonas*, *Strigomonas*, *Phytomonas, Sergeia*, and *Herpetomonas.* In the inferred genealogy, mtHsp70 Lc2.2 and Lc2.1 paralogs were separated in two well-supported clades, justifying their classification in two different families (Figs. 2, 6). Sequences of Lc2.2 and Lc2.1 share ~79% amino acid similarity. The absence of Lc2.1 or Lc2.1-like paralogs in the genus *Trypanosoma,* and also in *Bodo saltans*, *Bodo* sp., and *Parabodo caudatus,* suggests that Lc2.1 may have emerged in one ancestor of the lineage comprising the non-trypanosome trypanosomatids, favoring duplication in an ancestor of this lineage rather than a gene loss in the genus *Trypanosoma* and other kinetoplastids.

In accord with data showing that the Lc2.2 family presents homogeneous copies in the trypanosomatid genomes, the genealogy based on Lc2.2 orthologs shows relevant support values for most clades evidenced in the traditional species-tree of *Trypanosoma* (see section 10). In the analysis of combined Lc2.2 and Lc2.1 sequences, Lc2.1 emerged as a branch of the clade formed by *Leishmania* Lc2.2 sequences (data not shown). However, corroborating the origin of the Lc2.1 family in a common ancestor of the non-trypanosome trypanosomatid lineage, the genealogy of Lc2.1 homologs from species of *Leishmania* and several other genera was largely congruent with those inferred using Lc2.2. (Fig 6B).

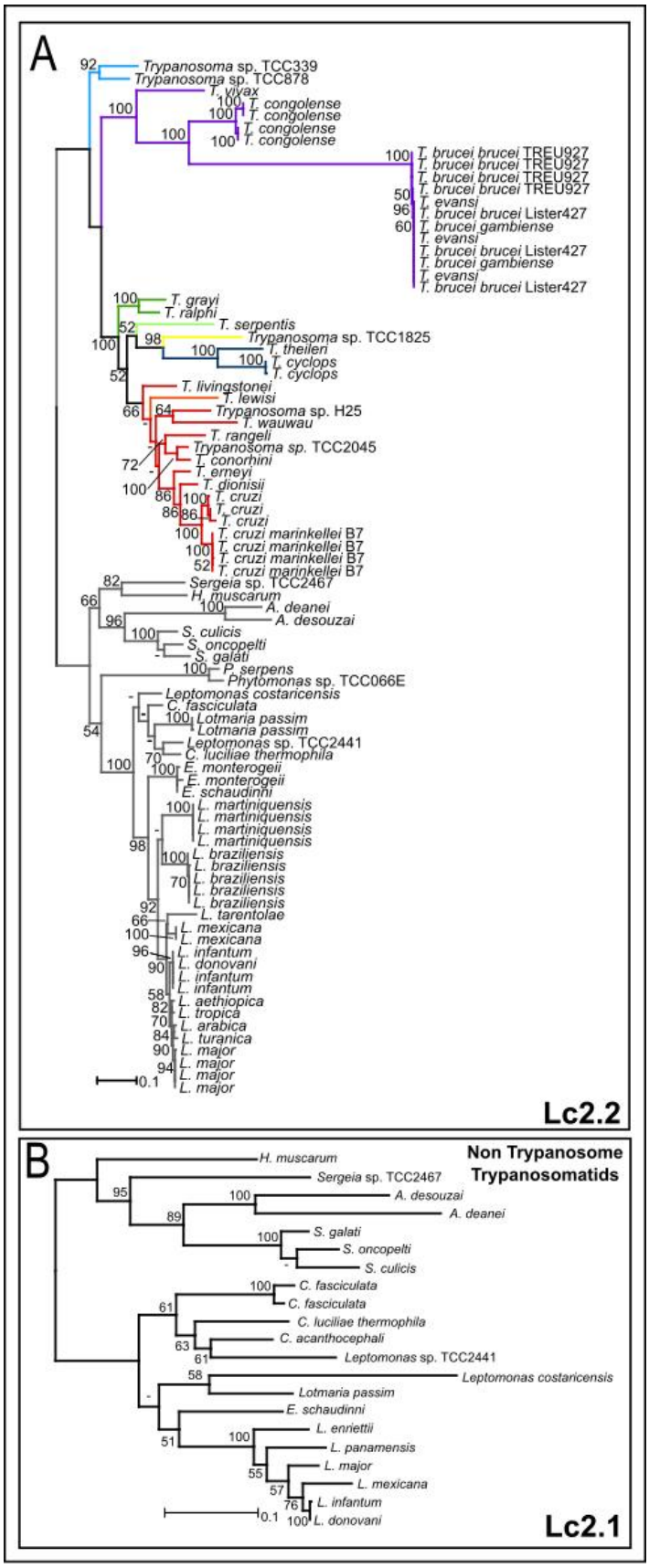

Figure 5. ML gene tree of MT Hsp70s, rooted using non-trypanosome trypanosomatids. Values at nodes are ML bootstrap values using 500 replicates. Branch colors represent the Trypanosoma clade, in which each OTU is classified according classical molecular markers gGAPDH and SSU rRNA. Letters A and B represent mtHsp70 families Lc2.2 and Lc2.1 respectively.

7. Phylogenetic analyses of the trypanosomatid large Hsps: the Hsp70.a, Hsp110, and Grp170 families

The large Hsp70s, i.e., Hsp110 and glucose-regulated protein 170 (Grp170) families, contain specialized proteins that are largely distantly related and structurally different from the other Hsp70s. These proteins have been described as nucleotide exchange factors (NEFs) of typical Hsp70s. Although closely related, these families differ in their location; while Hsp110 occurs in the nucleus and cytosol, the Grp170 is characterized by ER signals that confine the proteins into that cellular compartment. Hsp110s are known to inhibit protein aggregation by serving as substrate holdases (Easton et al., 2000; Andreasson et al., 2008; Inoue and Tsai, 2016).

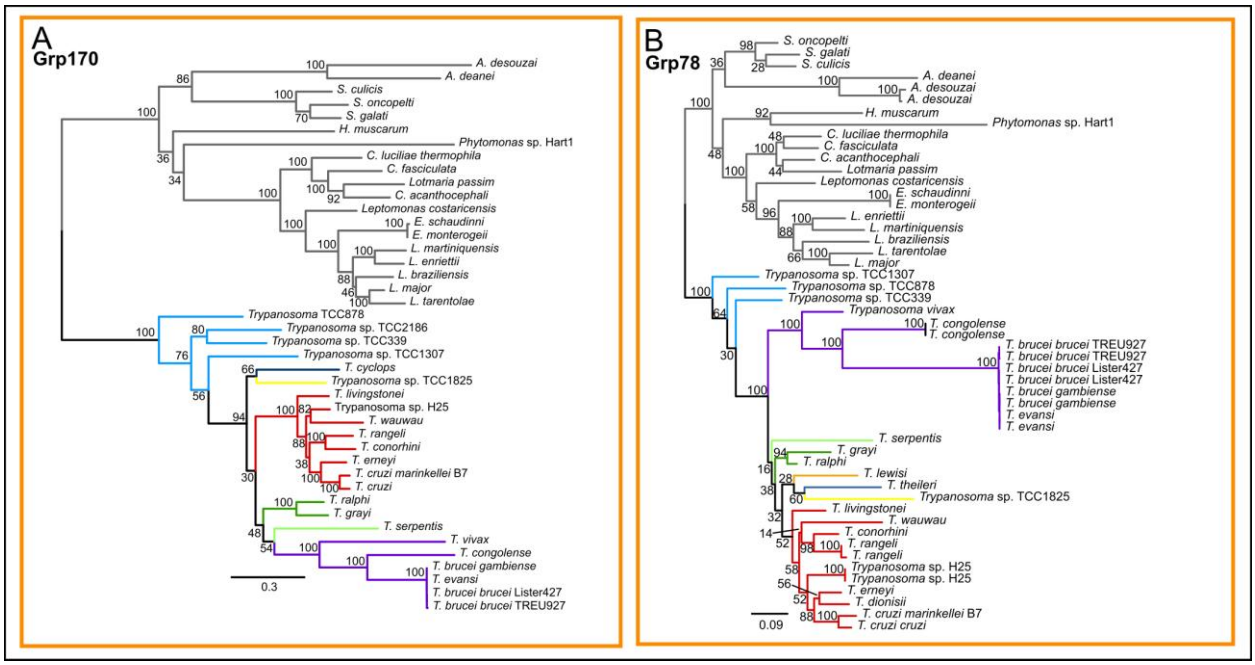

Figure 6. ML gene tree of ER Hsp70s, rooted using non-trypanosome trypanosomatids. Values at nodes are ML bootstrap values using 500 replicates. Branch colors represent the Trypanosoma clade, in which each OTU is classified according classical molecular markers gGAPDH and SSU rRNA. Letters A and B represent Grp170 and Grp78 families respectively.

The large sizes of these Hsp70s are due to the loop structure, which results in their remarkable ability to bind to protein and non-protein ligands, such as microbial molecules, favoring amplification of innate and adaptive immune responses (Zuo et al., 2016). Despite multifaceted functions under physiological or pathological conditions, the roles played by these large Hsp70s in trypanosomatids are virtually unknown.

Before the present study, the knowledge on Hsp70 families of the NEF group in trypanosomatids included only structural features of sequences from the TriTryp: Grp170 (ER), Hsp110 (CY), and the recently discovered Hsp70.a paralogs (Folgueira and Requena, 2007; Louw et al., 2010). In contrast to the ubiquitous Grp170 and Hsp110 homologs in eukaryotes in general, our taxon-rich and comprehensive analysis demonstrated that Hsp70.a is ubiquitous in trypanosomatids, and also present in other kinetoplastids. There is no information about the Hsp70.a location, but its positioning closest to Grp170 in the NEF group may suggest an ER location. However, Hsp70.a lacks any ER retention motif, suggesting an alternative cellular location. *In silico* prediction, and demonstration that the packaging of proteins into exosomes of *Leishmania,* known to

prime host cells for *Leishmania* invasion, depends on Hsp100, suggests that Hsp70.a may play a role in secretory pathways in trypanosomatids (Silverman et al., 2010).

The phylogenetic relationships inferred using all NEF Hsp families suggest that Hsp70.a may have originated by gene duplication followed by large divergence, resulting in a new Hsp70 family (Fig. 3). Hsp70.a represents the less conserved TryHsp70 family, with the orthologs sharing only ~54% amino acid similarity due to the presence of several highly polymorphic blocks of amino acids (Fig. 3). The similarity between the Hsp70.a and other NEFs range from ~29 to ~22%, compared to Grp170 (~68% amino acid similarity among the orthologs) and Hsp110 (~76% amino acid similarity). Therefore, the trypanosomatid Hsp110 and Grp170, as observed in other eukaryotes, are less conserved than the other organellar Hsp70 families.

Notably well-resolved topologies were obtained using both Hsp110 and Hsp70.a for inferences of the phylogenetic relationships among the trypanosomatids (Fig. 7), strongly supporting all main clades of trypanosomes and their relationships, as supported by traditional phylogenetic markers (see section 10). In addition, the relationships among the non-trypanosome trypanosomatids were also well resolved using both Hsp110 and Hsp70.a orthologs. In contrast, the genealogy of the polymorphic ER resident members of the Grp170 family revealed several relevant incongruences in major clades of both *Trypanosoma* and non-trypanosome trypanosomatids (Fig. 7).

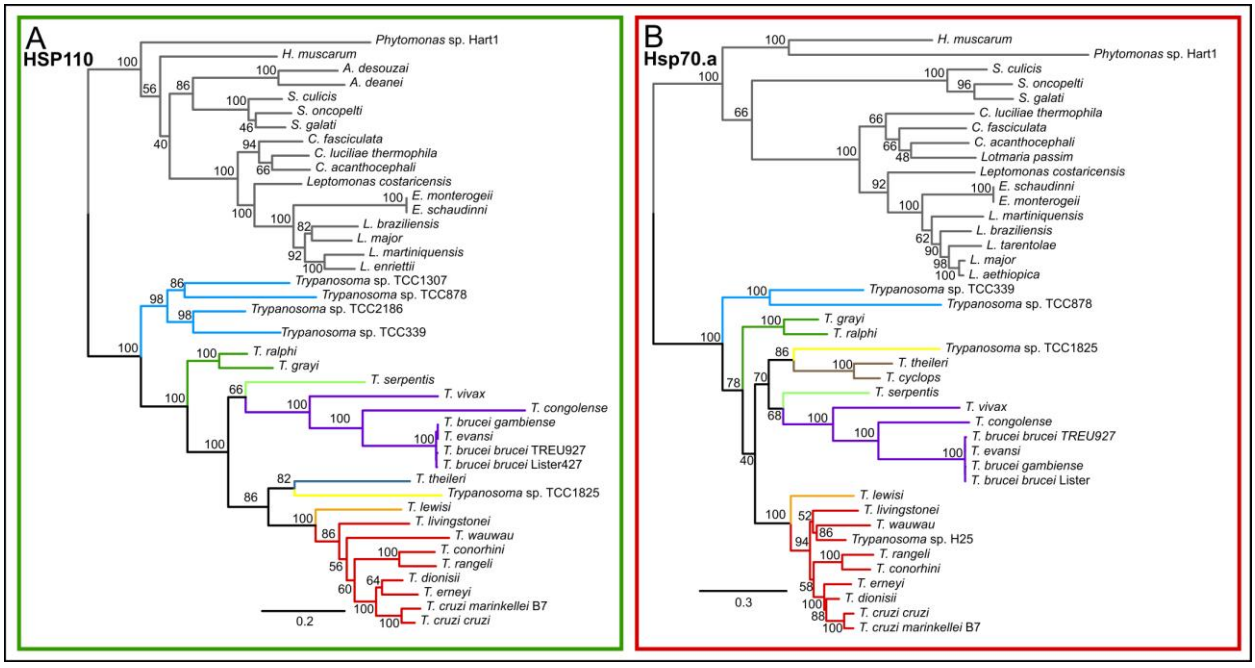

Figure 7. ML gene tree of Hsp110 and Hsp70.a, rooted using non-trypanosome trypanosomatids. Values at nodes are ML bootstrap values using 500 replicates. Branch colors represent the Trypanosoma clade, in which each OTU is classified according classical molecular markers gGAPDH and SSU rRNA. Letters A and B represent Hsp110 and Hsp70.a families respectively.

8. Phylogenetic analyses of the trypanosomatid Hsp70.b family

Our analysis demonstrated that Hsp70.b paralogs are present in all trypanosomatids, forming a monophyletic assemblage separated from both the typical Hsp70 and NEF groups (Fig. 2). The intermediary position of Hsp70.b between organellar Hsp70s and NEFs agrees with its structural features: Hsp70.b shows a conserved RLLLE Linker motif, similar to those found in CAN Hsp70, Grp78 and mtHsp70/Lc2.2.However, like the NEF members, Hsp70.b exhibited large insertions at the NBD and SBD (Fig. 3).

As discussed above (section 3), the trypanosomatid Hsp70.b family was positioned in our phylogenetic analysis in a branch formed by DnaK-like genes from a wide range of single-celled eukaryotes and bacterial groups. *In silico* prediction of subcellular localization carried out in this study suggests that Hsp70.b might be a mitochondrial protein.

The Hsp70.b orthologs share ~82% amino acid similarity, and were separated from other TryHsp70 families by large sequence divergences. Notably, the genealogy based on Hsp70.b orthologs (Fig. 8) strongly supported the traditional phylogeny of the *Trypanosoma*, with small intra-clade incongruences (see section 10). However, there are relevant inconsistencies in the phylogenetic relationships inferred for the non-trypanosome trypanosomatids (Fig. 8).

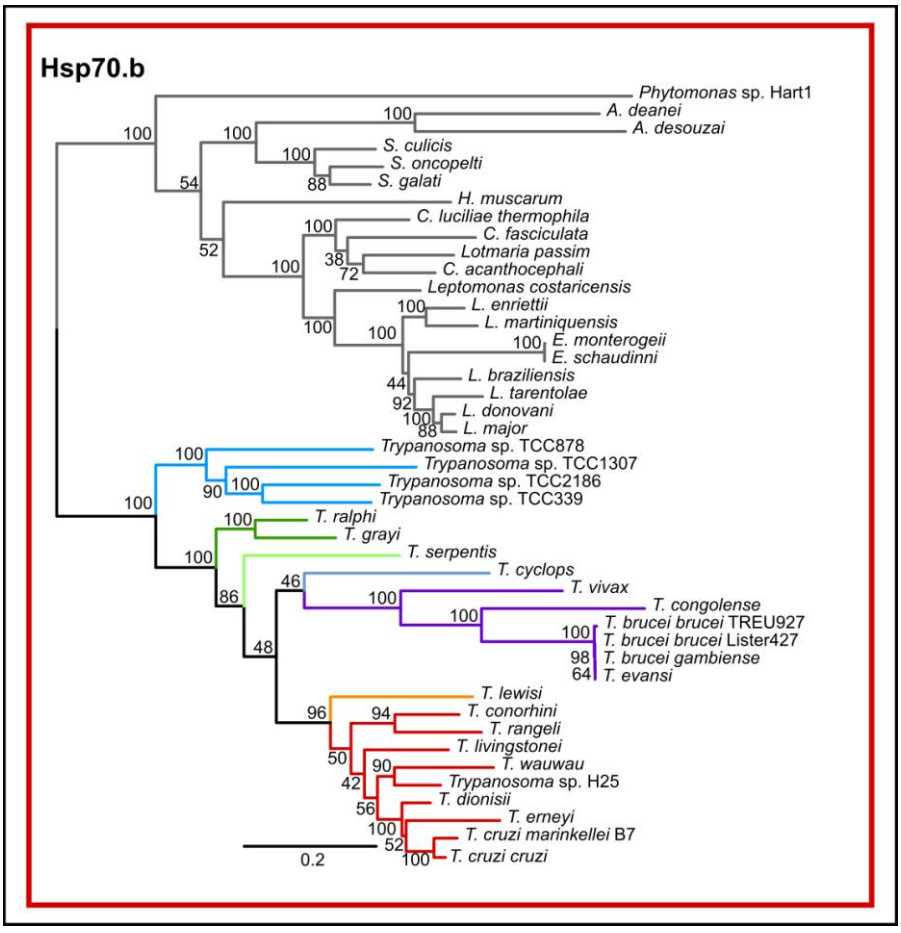

Figure 8. ML gene tree of Hsp70.b, rooted using non-trypanosome trypanosomatids. Values at nodes are ML bootstrap values using 500 replicates. Branch colors represent the *Trypanosoma* clade, in which each OTU is classified according classical molecular markers gGAPDH and SSU rRNA.

9. Phylogenetic analysis of DnaK paralogs from the endosymbionts of *Angomonas* and *Strigomonas*

In our analysis, sequences of two DnaK families (DnaK and DnaK-like) were identified in the genomes of the bacterial endosymbionts harbored by the insect trypanosomatids of the genera *Angomonas* and *Strigomonas* (Teixeira et al., 2011). The DnaK sequences recovered belong to the genomes of the Betaproteobacterial endosymbionts classified as *Candidatus* Kinetoplastibacterium spp. (Alves et al., 2013) and they clustered with Betaproteobacterial sequences in our wide phylogenetic analyze (Fig. 1), and with DnaK-paralogs of mtHsp70 (Lc2.2/Lc2.1) genes (Fig. 3) derived from Alphaproteobacteria forming a cluster of genes sharing typical motifs of the archetypical DnaK. The other DnaK paralog identified in the genomes of the endosymbionts was assigned to a DnaK-like group (Fig. 3), which clusters with the Hsp70.b family (in a branch without phylogenetic support) likely because, although distantly related, genes for all members of this branch lack motifs typical of the Dnak family (Fig. 2).

10. Syntenic organization of Hsp70 families in trypanosomatids

Synteny analyses were done for all TryHsp70 families in all genomes analyzed herein. Although some draft genomes are highly fragmented, hampering the analysis of long genomic segments, genomes representing a wide evolutionary range of the Trypanosomatidae family could be checked regarding the regions harboring the Hsp70 genes. Genomic organization of each Hsp70 family in the very well assembled genome of *T. brucei,* and in other genomes of either closely or distantly related trypanosome species, were selected to illustrated the results obtained (Fig. 9).

The genomic organization of CAN Hsp70, Hsp70.,4 and Hsp70.c is highly conserved in all genomes investigated, with total synteny of their flanking genes, which often can be Hsp70 and Hsp40 homologs (Fig. 9). The highly conserved mtHsp70/Lc2.2 genes are also totally syntenic across the Trypanosomatidae. In *T. brucei*, upstream from the Lc2.2 tandem block, there is another copy of Lc2.2 (Fig. 9). In the genome of *L. major,* the single copy Lc2.1 gene is flanked upstream by a tandem block of Lc2.2 copies and downstream by a putative Hsp40 gene. Taking into account the tandem organization of the Lc2.2 and Lc2.1 paralogs and data from our analysis suggesting recombination between Lc2.2 copies as well as between Lc2.2 and Lc2.1 (data not shown), recombinant events may be responsible for the origin and polymorphisms of the Lc2.1 family. The identification in *L. panamensis* of a new paralog more structurally related to Lc2.1, but clustering within the Lc2.2 clade, provides additional support for the idea of recombination producing new mtHsp70 sequences (data not shown).

Regarding the NEF group, all Hsp70 families exhibited highly syntenic organization in all trypanosomatid genomes. However, the regions flanking these genes did not exhibit other Hsp70 genes (Fig. 9).

In contrast to all other Hsp70 families, the genomic organization of Grp78 genes is not totally syntenic across the Trypanosomatidae. Although total synteny was conserved among trypanosome species from each phylogenetic clade, our analyses revealed only partial synteny among distantly related trypanosomatid species. The regions flanking Grp78 of *T. brucei* and *T. cruzi* were include in the Figure 9.

11. *Trypanosoma* phylogenetic and taxonomic inferences based on Hsp70 families

We compared the usefulness of each Hsp70 family to the inference of phylogenetic relationships among the species of the genus *Trypanosoma*. In general, all Hsp70 families corroborated the main clades supported by traditional phylogeny based on SSU rRNA and gGAPDH (Fig. S1). Nevertheless, not all Hsp70s produced well-resolved phylogenies for the whole Trypanosomatidae family, and support values were highly variable depending on the family. The best-resolved topologies were inferred using CAN Hsp70, Hsp70.b (DnaK-like), Hsp70.a and Hsp110 (Figs 4A, 7AB, 8), and were highly congruent with the traditional *Trypanosoma* phylogenetic tree. In contrast, mtHsp70/Lc2.2 and above all Grp78 and Grp170 analyses produced relevant incongruences and low support values (Figs 5A, 6AB).The genus *Trypanosoma* harbors two deep-split phylogenetic lineages called the Aquatic and Terrestrial clades. The Aquatic clade comprehends trypanosomes of aquatic and semi-aquatic vertebrates, such as fishes, turtles, anurans, platypus and crocodilians. In addition, this clade also harbors one trypanosome of a chameleon from Madagascar and another *Trypanosoma* sp. (TCC878) from *Mabuya* sp., a Neotropical lizard. Leeches in aquatic environments, and sandflies, frog-biting flies and culicids transmit the trypanosomes of the Aquatic clade.

Recent studies have unveiled the high genetic diversity of this clade, supporting fish and anuran trypanosomes separated in two main clades, and the close relationships between trypanosomes of caimans and fishes, evidencing the complex evolutionary history of the Aquatic clade (Hamilton et al., 2007; Ferreira et al., 2007, 2008; Fermino et al., 2013, 2015; Lemos et al., 2015; Attias et al., 2016). For the first time in this study, genomes of trypanosomes of the aquatic clade were sequenced from three trypanosomes of anurans (TCC339, TCC1307 and TCC 2186), which were previously assigned to different phylogenetic lineages (Ferreira et al., 2008; Attias et al., 2016), and the *Trypanosoma* from Mabuya (Table 1). In all TryHsp70 based phylogenies, the trypanosomes of the Aquatic clade were strongly supported as the most basal and monophyletic lineage, in contrast with the phylogenies based on Grp78, Grp170 and Lc2.2, which did not support the monophyly of this clade (Figs 5A, 7AB). The basal position of the fish parasite *T. carasii* was corroborated in a recent phylogenetic study restricted to CAN Hsp70, as well as in concatenated SSU rRNA, gGADPH and Hsp70 sequences (Fraga et al., 2016).

The Terrestrial clade comprehends trypanosomes from mammals, birds, snakes, lizards and crocodiles transmitted by fleas, ticks, dipterans and hemipterans. Trypanosomes from mammals clustered into the strongly supported clades *T. brucei, T. cruzi, T. lewisi* and *T. theileri*. The clade *T. cruzi* comprises *T. cruzi* and the closely related bat-restricted *T. cruzi marinkellei*, *T. erneyi* and *T. dionisii,* all forming the subgenus *Schizotrypanum* (Hamilton et al., 2012; Lima et al., 2012, 2013, 2015). All Hsp70 inferences strongly supported the subgenus *Schizotrypanum* as a monophyletic group, and placed *T. c. marinkellei* sister to *T. cruzi*. However, either *T. erneyi* or *T. dionisii,* which were separated from *T. cruzi* by comparable Hsp70 sequence divergence, could be the closest relatives of *T. cruzi* (Figs. 4, 5A, 7B, 8), while in some analyses *T. erneyi* appeared as sister species to *T. dionisii* (Figs. 6B, 7A).

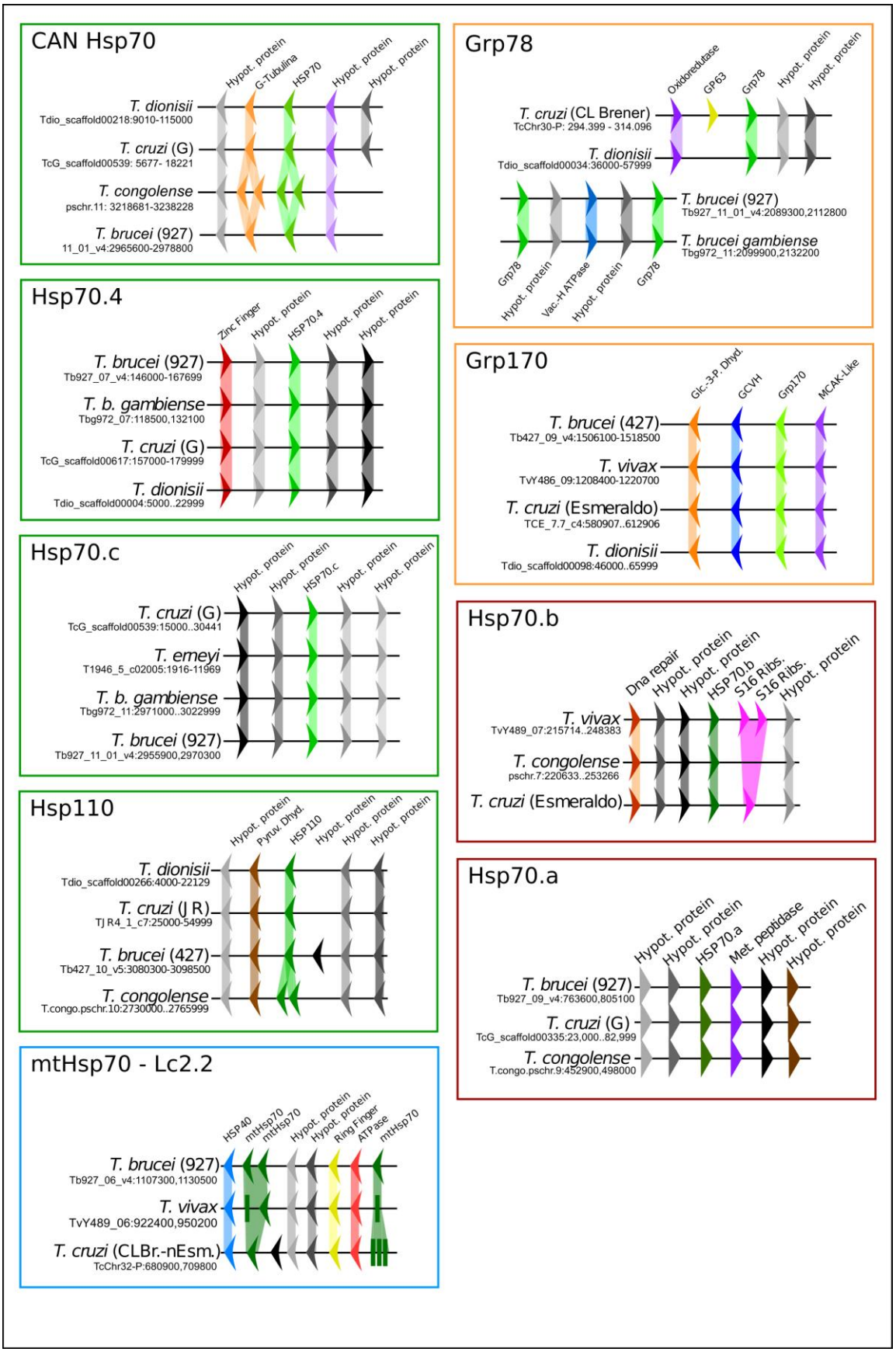

Figure 9. Synteny analysis showing the genomic organization of TryHsp70 orthologous genes arranged in the genomes trypanosome species. The TryHsp70 genes are located in highly syntenic regions, the only exception being the Grp78 genes, which are located in regions that are partially syntenic among species of distantly related trypanosomes. Box colors corresponding cellular localization indicating in figure 2 and 3.

Along with the species of *Schizotrypanum,* the clade *T. cruzi* also includes the subclade *T. rangeli-T. conorhini,* strongly supported as the sister group to the clade *Schizotrypanum* as inferred by traditional phylogeny. This relationship was also supported in phylogenies based on the genes coding for CatL enzymes, as well as SL rRNA and PRAC gene sequences (Caballero et al., 2014). *T. rangeli,* likewise *T. cruzi,* is a generalist parasites of mammals, whereas *T. conorhini* infects rats and monkeys, while the three these species are transmitted by triatomines. The clade *T. rangeli-T. conorhini* was recovered with high bootstrap values in all *Try*Hsp70 analyses. In addition, this clade was sister to *Schizotrypanum* using Hsp70.a, Grp170, Hsp110, CAN Hsp70 and mtHsp70/Lc.2.2 (Figs. 4A, 5A, 6A, 7), which is consistent with data from traditional phylogenies. However, the analyses of Hsp70.c, Hsp70.4, Grp78 and Hsp70.b positioned *T. wauwau* and its closely related *T. sp.* H25 (Lima et al., 2015) as sister group to the clade *Schizotrypanum* (Figs. 4BC, 6B, 8). In the traditional phylogeny, the clade *T. wauwau* is basal to the major assemblage formed by the clades *Schizotrypanum* and *T. rangeli-T. conorhini*, whereas the most basal species of the *T. cruzi* whole clade is *T. livingstonei* (Lima et al., 2013, 2015). All Hsp70 analyses support this positioning for *T. livingstonei*, except Grp78 and mtHsp70/Lc2.2 (Figs 5A, 6B). *T. lewisi* is the type species of the subgenus *Herpetosoma* (Maia da Silva et al., 2010), which is formed mainly by trypanosomes of rodents transmitted by fleas through the world. The placement of *T. lewisi* as the closest outgroup of the whole *T. cruzi* clade, as supported by traditional phylogeny, was corroborated in all phylogenies excepting those inferred from genes sequences of the families Grp78 and Lc2.2 (Figs. 5A, 6B).

The genomes of trypanosomes belonging to *T. brucei* clade included in our analysis are from *T. b. brucei*, *T. b. gambiense*, *T. evansi, T. congolense* and *T. vivax*. *T. brucei, T. congolense* and *T. vivax,* which are cyclically transmitted by tsetse flies in Africa. Other hematophagous flies in Africa and Central and South America can mechanically transmit *T. vivax*, whereas *T. evansi* is transmitted exclusively mechanically. The clade *T. brucei* (subgenus *Trypanozoon* - type species *T. brucei*) was recovered in all Hsp70 inferences exhibiting, as in traditional phylogenetic markers, the longest branches due to evolutionary rates higher than those of other trypanosomes. In agreement with several other markers (Hamilton et al., 2007; Cortez et al., 2008; Rodrigues et al. 2008), all inferences based on Hsp70 families strongly support *T. vivax* (type species of the subgenus *Duttonella*) as the most divergent and basal species of the *T. brucei* clade, positioned at the edge of the clade. Also in agreement with other analyses (Hamilton et al., 2007; Rodrigues et al., 2014), *T. congolense* (type species of the subgenus *Nannomonas*) is positioned sister to the clade *Trypanozoon.* The trypanosome species placed closest to the clade *T. brucei* largely diverged depending on the Hsp70 family and taxa included in the analyses (Figs 5-8).

T. theileri, the type-species of the subgenus *Megatrypanum,* which comprises exclusively trypanosomes of ruminants (Rodrigues et al., 2006; Garcia et al., 2011), when included in the analysis (some genes could be not identified in the genome draft of this species) always clustered with *T. cyclops* from Asian monkeys as supported by all previously inferred phylogenies. The positioning of the clade *T. theileri* was not resolved by any Hsp70 phylogenies.

Trypanosomes of non-mammalian hosts of the terrestrial clade are from birds, lizards, snakes and crocodilians. Recent studies have supported the Crocodilian clade, formed by trypanosomes from caimans and crocodiles, as the sister group to the major clade comprising all other terrestrial trypanosomes (Fermino et al., 2013, 2015). We examined draft genomes form two species of the clade Crocodilian, *T. grayi* from African crocodiles (Kelly et al., 2014), and *T. ralphi* from South American caimans (Fermino et al., 2013). All Hsp70 inferences recovered this clade, and its positioning as sister group to all other terrestrial clades was supported by CAN Hsp70, mtHsp70/Lc2.2, Hsp110, Hsp70.a and Hsp70.b analyses (Figs. 4A, 5A, 7, 8). The only genome of a bird trypanosome (TCC1825) included in this study belongs to the *T. corvi* clade (Hamilton et al., 2007; Votypka et al., 2004). Interestingly, in several analyses this trypanosome was sister to the clade *T. theileri/T. cyclops* (Fig. 4BC, 5A, 6, 7). The position of *T. serpentis* (Viola et al., 2009) varies broadly according to the Hsp70 family used for phylogenic inferences. However, in traditional phylogenies, trypanosomes from snakes cluster with trypanosomes from lizards forming the Lizard/Snake clade nesting with trypanosomes transmitted by sandflies (Viola et al., 2009). The resulting phylogenies based on Hsp70 inferences do not permit any conclusive positioning of *T. serpentis* and the bird trypanosome, most likely because only one representative of each, the Avian and Lizard/Snake clades, were included in the analyses.

Although this study includes the most taxon-rich sampling to date, the taxon coverage still represents only a fraction of the *Trypanosoma* diversity, and the lack or low representation of some parasite lineages hampers the achievement of well-resolved phylogenies for the genus *Trypanosoma* because the positioning of clades represented by a small number of species is not resolved.

CONCLUSIONS

The results from the present study yield a comprehensive inventory of the different Hsp70 families present in the genomes of the Trypanosomatidae. Wide genome surveys have unveiled an expanded repertoire of TryHsp70 families, beyond the conserved typical Hsp70. Before this study, systematic analyses of TryHsp70 superfamily were restricted to the Tritryps (*T. brucei*, *T. cruzi* and *L. major*). The large set of genomes from cold and warm-blooded trypanosomes and insect-restrict trypanosomatids examined in the present study revealed a great number of Hsp70 paralogs. These paralogs were phylogenetically characterized permitting to uncover the repertoires of Hsp70s from each trypanosomatid species, and the composition of each TryHsp70 family. Our analyses largely expanded the knowledge on the TryHsp70 families by comparing their degree of conservation on domain architecture, amino acid sequences and genome organization. In addition, highly comprehensive phylogenetic analyses provided new and relevant insights to better hypothesize on the evolutionary origin of each TryHsp70 families. Nine out 10 families nested into the TryHsp70 superfamily have homologs in all trypanosomatids: CAN Hsp70, Hsp70.4, Hsp70.c, mtHsp70/Lc2.2, Grp78, Grp170, Hsp110, Hsp70.a and Hsp70.b. The family Hsp70.b, herein supported as a Dnak-like family, is ubiquitous to trypanosomatids, present in other kinetoplastids, but is absent in eukaryotes in general. The family mtHsp/Lc2.1 is present in all genera of

trypanosomatids investigated except *Trypanosoma.* Orthologues of the TryHsp70 families largely varied on degrees of sequence divergence: CAN Hsp70 and mtHsp70/Lc.2.2 are the most conserved, and the less conserved are Hsp70.a and Grp170. Detailed knowledge on polymorphisms and structural organization of each Hsp70 families provides a framework for studies addressing the roles of each TryHsp70 family trypanosomatids. Altogether, these data are crucial to future studies addressing Hsp70 location, functions and co-partnerships. Finally, taxon-rich phylogenetic inferences based on each Hsp70 family corroborated the main clades supported by the traditional phylogeny of the Trypanosomatidae family regardless of their evolutionary origin. The best-resolved *Trypanosoma* phylogenies were inferred using Hsp70.a, CAN Hsp70, Hsp70.b and Hsp110, despite the variable positioning of some still poorly represented lineages. Improved taxa coverage including sequences from a large number of species, obtained by PCR-sequencing or from new genomes, is being currently investigated to evaluate the power of Hsp70 genes as phylogenetic markers for trypanosomes and other kinetoplastids.

REFERENCES

Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinforma Oxf Engl. 2005 May 1;21(9):2104–5.

Aggarwal A, de la Cruz VF, Nash TE. A heat shock protein gene in *Giardia lamblia* unrelated to HSP70. Nucleic Acids Res. 1990 Jun 11;18(11):3409.

Alves JMP, Serrano MG, Maia da Silva F, Voegtly LJ, Matveyev AV, Teixeira MMG, et al. Genome evolution and phylogenomic analysis of Candidatus Kinetoplastibacterium, the betaproteobacterial endosymbionts of *Strigomonas* and *Angomonas*. Genome Biol Evol. 2013;5(2):338–50.

Andréasson C, Fiaux J, Rampelt H, Druffel-Augustin S, Bukau B. Insights into the structural dynamics of the Hsp110-Hsp70 interaction reveal the mechanism for nucleotide exchange activity. Proc Natl Acad Sci U S A. 2008 Oct 28;105(43):16519–24.

Arisue N, Sánchez LB, Weiss LM, Müller M, Hashimoto T. Mitochondrial-type hsp70 genes of the amitochondriate protists, Giardia intestinalis, *Entamoeba histolytica* and two microsporidians. Parasitol Int. 2002 Mar;51(1):9–16.

Attias M, Sato LH, Ferreira RC, Takata CSA, Campaner M, Camargo EP, et al. Developmental and Ultrastructural Characterization and Phylogenetic Analysis of *Trypanosoma herthameyeri* n. sp of Brazilian Leptodactilydae Frogs. J Eukaryot Microbiol. 2016 Mar 2;

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.

Bakatselou C, Kidgell C, Graham Clark C. A mitochondrial-type hsp70 gene of *Entamoeba histolytica*. Mol Biochem Parasitol. 2000 Sep;110(1):177–82.

Bangs JD, Brouch EM, Ransom DM, Roggy JL. A soluble secretory reporter system in *Trypanosoma brucei.* Studies on endoplasmic reticulum targeting. J Biol Chem. 1996 Aug 2;271(31):18387–93.

Bangs JD, Uyetake L, Brickman MJ, Balber AE, Boothroyd JC. Molecular cloning and cellular localization of a BiP homologue in *Trypanosoma brucei.* Divergent ER retention signals in a lower eukaryote. J Cell Sci. 1993 Aug;105 (Pt 4):1101–13.

Boorstein WR, Ziegelhoffer T, Craig EA. Molecular evolution of the HSP70 multigene family. J Mol Evol. 1994 Jan;38(1):1–17.

Borghesan TC, Ferreira RC, Takata CSA, Campaner M, Borda CC, Paiva F, et al. Molecular phylogenetic redefinition of *Herpetomonas* (Kinetoplastea, Trypanosomatidae), a genus of insect parasites associated with flies. Protist. 2013 Jan;164(1):129–52.

Bracher A, Verghese J. The nucleotide exchange factors of Hsp70 molecular chaperones. Front Mol Biosci. 2015;2:10.

Burger A, Ludewig MH, Boshoff A. Investigating the Chaperone Properties of a Novel Heat Shock Protein, Hsp70.c, from *Trypanosoma brucei*. J Parasitol Res. 2014;2014:172582.

Caballero ZC, Costa-Martins AG, Ferreira RC, P Alves JM, Serrano MG, Camargo EP, et al. Phylogenetic and syntenic data support a single horizontal transference to a Trypanosoma ancestor of a prokaryotic proline racemase implicated in parasite evasion from host defences. Parasit Vectors. 2015;8:222.

Camargo EP. Phytomonas and other trypanosomatid parasites of plants and fruit. Adv Parasitol. 1999;42:29–112.

Campos RM, Nascimento M, Ferraz JC, Pereira MM, Rocha PO, Thompson GM, et al. Distinct mitochondrial HSP70 homologues conserved in various *Leishmania* species suggest novel biological functions. Mol Biochem Parasitol. 2008 Aug;160(2):157–62.

Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000 Apr;17(4):540–52.

Cortez AP, Ventura RM, Rodrigues AC, Batista JS, Paiva F, Añez N, et al. The taxonomic and phylogenetic relationships of *Trypanosoma vivax* from South America and Africa. Parasitology. 2006 Aug;133(Pt 2):159–69.

Daugaard M, Rohde M, Jäättelä M. The heat shock protein 70 family: Highly homologous proteins with overlapping and distinct functions. FEBS Lett. 2007 Jul 31;581(19):3702–10.

Easton DP, Kaneko Y, Subjeck JR. The hsp110 and Grp1 70 stress proteins: newly recognized relatives of the Hsp70s. Cell Stress Chaperones. 2000 Oct;5(4):276–90.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinforma Oxf Engl. 2010 Oct 1;26(19):2460–1.

Engman DM, Fehr SC, Donelson JE. Specific functional domains of mitochondrial hsp70s suggested by sequence comparison of the trypanosome and yeast proteins. Mol Biochem Parasitol. 1992 Mar;51(1):153–5.

Engman DM, Kirchhoff LV, Donelson JE. Molecular cloning of mtp70, a mitochondrial member of the hsp70 family. Mol Cell Biol. 1989a Nov;9(11):5163–8.

Engman DM, Sias SR, Gabe JD, Donelson JE, Dragon EA. Comparison of HSP70 genes from two strains of *Trypanosoma cruzi*. Mol Biochem Parasitol. 1989b Dec;37(2):285–7.

Fermino BR, Paiva F, Soares P, Tavares LER, Viola LB, Ferreira RC, et al. Field and experimental evidence of a new caiman trypanosome species closely phylogenetically related to fish trypanosomes and transmitted by leeches. Int J Parasitol Parasites Wildl. 2015 Dec;4(3):368–78.

Fermino BR, Viola LB, Paiva F, Garcia HA, de Paula CD, Botero-Arias R, et al. The phylogeography of trypanosomes from South American alligatorids and African crocodilids is consistent with the geological history of South American river basins and the transoceanic dispersal of *Crocodylus* at the Miocene. Parasit Vectors. 2013;6(1):313.

Ferreira RC, Campaner M, Viola LB, Takata CSA, Takeda GF, Teixeira MMG. Morphological and molecular diversity and phylogenetic relationships among anuran trypanosomes from the Amazonia, Atlantic Forest and Pantanal biomes in Brazil. Parasitology. 2007 Oct;134(Pt 11):1623–38.

Ferreira RC, De Souza AA, Freitas RA, Campaner M, Takata CS, Barrett TV, et al. A phylogenetic lineage of closely related trypanosomes (Trypanosomatidae, Kinetoplastida) of anurans and sand flies (Psychodidae, Diptera) sharing the same ecotopes in brazilian amazonia. J Eukaryot Microbiol. 2008 Sep;55(5):427–35.

Flechas ID, Cuellar A, Cucunubá ZM, Rosas F, Velasco V, Steindel M, et al. Characterising the KMP-11 and HSP-70 recombinant antigens' humoral immune response profile in chagasic patients. BMC Infect Dis. 2009;9:186.

Folgueira C, Quijada L, Soto M, Abanades DR, Alonso C, Requena JM. The translational efficiencies of the two Leishmania infantum HSP70 mRNAs, differing in their 3'-untranslated regions, are affected by shifts in the temperature of growth through different mechanisms. J Biol Chem. 2005 Oct 21;280(42):35172–83.

Folgueira C, Requena JM. A postgenomic view of the heat shock proteins in kinetoplastids. FEMS Microbiol Rev. 2007 Jul;31(4):359–77.

Fraga J, Fernández-Calienes A, Montalvo AM, Maes I, Deborggraeve S, Büscher P, et al. Phylogenetic analysis of the *Trypanosoma* genus based on the heat-shock protein 70 gene. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2016 May 12;

Fraga J, Fernandez-Calienes A, Montalvo AM, Maes I, Dujardin J-C, Van der Auwera G. Differentiation between *Trypanosoma cruzi* and *Trypanosoma rangeli* using heat-shock protein 70 polymorphisms. Trop Med Int Health TM IH. 2014 Feb;19(2):195–206.

Fraga J, Montalvo AM, De Doncker S, Dujardin JC, Van der Auwera G. Phylogeny of *Leishmania* species based on the heat-shock protein 70 gene. Infect Genet Evol. 2010 Mar;10(2):238–45.

Fraga J, Veland N, Montalvo AM, Praet N, Boggild AK, Valencia BM, et al. Accurate and rapid species typing from cutaneous and mucocutaneous leishmaniasis lesions of the New World. Diagn Microbiol Infect Dis. 2012 Oct;74(2):142–50.

Freeman BC, Myers MP, Schumacher R, Morimoto RI. Identification of a regulatory motif in Hsp70 that affects ATPase activity, substrate binding and interaction with HDJ-1. EMBO J. 1995 May 15;14(10):2281–92.

Garcia HA, Kamyingkird K, Rodrigues AC, Jittapalapong S, Teixeira MM, Desquesnes M. High genetic diversity in field isolates of *Trypanosoma theileri* assessed by analysis of cathepsin L-like sequences disclosed multiple and new genotypes infecting cattle in Thailand. Vet Parasitol. 2011 Aug 25;180(3–4):363–7.

Germot A, Philippe H. Critical analysis of eukaryotic phylogeny: a case study based on the HSP70 family. J Eukaryot Microbiol. 1999 Mar;46(2):116–24.

Glass DJ, Polvere RI, Van der Ploeg LH. Conserved sequences and transcription of the hsp70 gene family in *Trypanosoma brucei*. Mol Cell Biol. 1986 Dec;6(12):4657–66.

Gupta RS, Aitken K, Falah M, Singh B. Cloning of Giardia lamblia heat shock protein HSP70 homologs: implications regarding origin of eukaryotic cells and of endoplasmic reticulum. Proc Natl Acad Sci U S A. 1994 Apr 12;91(8):2895–9.

Gupta RS, Golding GB. Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria, and eukaryotes. J Mol Evol. 1993 Dec;37(6):573–82.

Hannaert V, Saavedra E, Duffieux F, Szikora J-P, Rigden DJ, Michels PAM, et al. Plant-like traits associated with metabolism of *Trypanosoma* parasites. Proc Natl Acad Sci U S A. 2003 Feb 4;100(3):1067–71.

Hamilton PB, Cruickshank C, Stevens JR, Teixeira MMG, Mathews F. Parasites reveal movement of bats between the New and Old Worlds. Mol Phylogenet Evol. 2012 May;63(2):521–6.

Hamilton PB, Gibson WC, Stevens JR. Patterns of co-evolution between trypanosomes and their hosts deduced from ribosomal RNA and protein-coding gene phylogenies. Mol Phylogenet Evol. 2007 Jul;44(1):15–25.

Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 2012 Dec 1;61(6):1061–7.

Inoue T, Tsai B. The Grp170 nucleotide exchange factor executes a key role during ERAD of cellular misfolded clients. Mol Biol Cell. 2016 Mar 30;

Jackson AP, Otto TD, Aslett M, Armstrong SD, Bringaud F, Schlacht A, et al. Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. Curr Biol CB. 2016 Jan 25;26(2):161–72.

Jackson AP, Quail MA, Berriman M. Insights into the genome sequence of a free-living Kinetoplastid: Bodo saltans (Kinetoplastida: Euglenozoa). BMC Genomics. 2008;9:594.

Kabani M, Martineau CN. Multiple hsp70 isoforms in the eukaryotic cytosol: mere redundancy or functional specificity? Curr Genomics. 2008;9(5):338–248.

Kominek J, Marszalek J, Neuvéglise C, Craig EA, Williams BL. The complex evolutionary dynamics of Hsp70s: a genomic and functional perspective. Genome Biol Evol. 2013;5(12):2460–77.

Kriehuber T, Rattei T, Weinmaier T, Bepperling A, Haslbeck M, Buchner J. Independent evolution of the core domain and its flanking sequences in small heat shock proteins. FASEB J Off Publ Fed Am Soc Exp Biol. 2010 Oct;24(10):3633–42.

Labriola CA, Giraldo AMV, Parodi AJ, Caramelo JJ. Functional cooperation between BiP and calreticulin in the folding maturation of a glycoprotein in *Trypanosoma cruzi.* Mol Biochem Parasitol. 2011 Feb;175(2):112–7.

Lee MG, Polvere RI, Van der Ploeg LH. Evidence for segmental gene conversion between a cognate hsp 70 gene and the temperature-sensitively transcribed hsp70 genes of *Trypanosoma brucei.* Mol Biochem Parasitol. 1990 Jun;41(2):213–20.

Lemos M, Fermino BR, Simas-Rodrigues C, Hoffmann L, Silva R, Camargo EP, et al. Phylogenetic and morphological characterization of trypanosomes from Brazilian armoured catfishes and leeches reveal high species diversity, mixed infections and a new fish trypanosome species. Parasit Vectors. 2015;8(1):573.

Lima L, Espinosa-Álvarez O, Pinto CM, Cavazzana M, Pavan AC, Carranza JC, et al. New insights into the evolution of the *Trypanosoma cruzi* clade provided by a new trypanosome species tightly linked to Neotropical Pteronotus bats and related to an Australian lineage of trypanosomes. Parasit Vectors. 2015;8:657.

Louw CA, Ludewig MH, Mayer J, Blatch GL. The Hsp70 chaperones of the Tritryps are characterized by unusual features and novel members. Parasitol Int. 2010 Dec;59(4):497–505.

Lukeš J, Skalický T, Týč J, Votýpka J, Yurchenko V. Evolution of parasitism in kinetoplastid flagellates. Mol Biochem Parasitol. 2014 Jul;195(2):115–22.

Maia da Silva F, Marcili A, Ortiz PA, Epiphanio S, Campaner M, Catao-Dias JL, et al. Phylogenetic, morphological and behavioural analyses support host switching of *Trypanosoma (Herpetosoma) lewisi* from domestic rats to primates. Infect Genet Evol. 2010 May;10(4):522–9.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. RDP3: a flexible and fast computer program for analyzing recombination. Bioinforma Oxf Engl. 2010 Oct 1;26(19):2462–3.

Martin F, Requena JM, Martin J, Alonso C, López MC. Cytoplasmic-nuclear translocation of the Hsp70 protein during environmental stress in *Trypanosoma cruzi.* Biochem Biophys Res Commun. 1993 Nov 15;196(3):1155–62.

Martin W, Borst P. Secondary loss of chloroplasts in trypanosomes. Proc Natl Acad Sci U S A. 2003 Feb 4;100(3):765–7.

Maslov DA, Votýpka J, Yurchenko V, Lukeš J. Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. Trends Parasitol. 2013 Jan;29(1):43–52.

Mayer MP, Bukau B. Hsp70 chaperones: cellular functions and molecular mechanism. Cell Mol Life Sci. 2005 Mar;62(6):670–84.

Morrison HG, Roger AJ, Nystul TG, Gillin FD, Sogin ML. Giardia lamblia expresses a proteobacterial-like DnaK homolog. Mol Biol Evol. 2001 Apr;18(4):530–41.

Názer E, Verdún RE, Sánchez DO. Severe heat shock induces nucleolar accumulation of mRNAs in *Trypanosoma cruzi.* PloS One. 2012;7(8):e43715.

Oladiran A, Belosevic M. Trypanosoma carassii hsp70 increases expression of inflammatory cytokines and chemokines in macrophages of the goldfish (*Carassius auratus* L.). Dev Comp Immunol. 2009 Oct;33(10):1128–36.

Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PloS One. 2010;5(3):e9490.

Przyborski JM, Diehl M, Blatch GL. Plasmodial HSP70s are functionally adapted to the malaria parasite life cycle. Front Mol Biosci. 2015;2:34.

Requena JM, Chicharro C, García L, Parrado R, Puerta CJ, Cañavate C. Sequence analysis of the 3'-untranslated region of HSP70 (type I) genes in the genus *Leishmania*: its usefulness as a molecular marker for species identification. Parasit Vectors. 2012;5:87.

Requena JM, Jimenez-Ruiz A, Soto M, Assiego R, Santarén JF, Lopez MC, et al. Regulation of hsp70 expression in *Trypanosoma cruzi* by temperature and growth phase. Mol Biochem Parasitol. 1992 Jul;53(1–2):201–11.

Requena JM, Lopez MC, Jimenez-Ruiz A, Morales G, Alonso C. Complete nucleotide sequence of the hsp70 gene of T. cruzi. Nucleic Acids Res. 1989 Jan 25;17(2):797.

Requena JM, Lopez MC, Jimenez-Ruiz A, de la Torre JC, Alonso C. A head-to-tail tandem organization of hsp70 genes in *Trypanosoma cruzi.* Nucleic Acids Res. 1988 Feb 25;16(4):1393–406.

Requena JM, Montalvo AM, Fraga J. Molecular Chaperones of Leishmania: Central Players in Many Stress-Related and -Unrelated Physiological Processes. BioMed Res Int. 2015;2015:301326.

Rodrigues AC, Neves L, Garcia HA, Viola LB, Marcili A, Da Silva FM, et al. Phylogenetic analysis of *Trypanosoma vivax* supports the separation of South American/West African from East African isolates and a new T. vivax-like genotype infecting a nyala antelope from Mozambique. Parasitology. 2008 Sep;135(11):1317–28.

Rodrigues AC, Ortiz PA, Costa-Martins AG, Neves L, Garcia HA, Alves JMP, et al. Congopain genes diverged to become specific to Savannah, Forest and Kilifi subgroups of *Trypanosoma congolense,* and are valuable for diagnosis, genotyping and phylogenetic inferences. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2014 Apr;23:20–31.

Searle S, Campos AJ, Coulson RM, Spithill TW, Smith DF. A family of heat shock protein 70-related genes are expressed in the promastigotes of *Leishmania major*. Nucleic Acids Res. 1989 Jul 11;17(13):5081–95.

Shonhai A, Maier AG, Przyborski JM, Blatch GL. Intracellular protozoan parasites of humans: the role of molecular chaperones in development and pathogenesis. Protein Pept Lett. 2011 Feb;18(2):143–57.

Silverman JM, Clos J, Horakova E, Wang AY, Wiesgigl M, Kelly I, et al. Leishmania exosomes modulate innate and adaptive immune responses through effects on monocytes and dendritic cells. J Immunol Baltim Md 1950. 2010 Nov 1;185(9):5011–22.

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

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinforma Oxf Engl. 2014 May 1;30(9):1312–3.

Stevens JR, Noyes HA, Schofield CJ, Gibson W. The molecular evolution of Trypanosomatidae. Adv Parasitol. 2001;48:1–56.

Teixeira MM, Borghesan TC, Ferreira RC, Santos MA, Takata CS, Campaner M, et al. Phylogenetic validation of the genera Angomonas and Strigomonas of trypanosomatids harboring bacterial endosymbionts with the description of new species of trypanosomatids and of proteobacterial symbionts. Protist. 2011 Jul;162(3):503–24.

Tibbetts RS, Kim IY, Olson CL, Barthel LM, Sullivan MA, Winquist AG, et al. Molecular cloning and characterization of the 78-kilodalton glucose-regulated protein of *Trypanosoma cruzi.* Infect Immun. 1994 Jun;62(6):2499–507.

Truc P, Büscher P, Cuny G, Gonzatti MI, Jannin J, Joshi P, et al. Atypical human infections by animal trypanosomes. PLoS Negl Trop Dis. 2013;7(9):e2256.

Truc P, Nzoumbou-Boko R, Desquesnes M, Semballa S, Vincendeau P. [Atypical human trypanosomoses]. Médecine Santé Trop. 2014 Sep;24(3):249–52.

Týč J, Klingbeil MM, Lukeš J. Mitochondrial heat shock protein machinery hsp70/hsp40 is indispensable for proper mitochondrial DNA maintenance and replication. mBio. 2015;6(1).

Urményi TP, Rodrigues DC, Silva R, Rondinelli E (2012) The stress response of *Trypanosoma cruzi*. In: Requena JM (ed) Stress response in microbiology. Caister Academic Press, Norwich, pp 345–373

Urményi TP, Silva R, Rondinelli E. The heat shock proteins of *Trypanosoma cruzi*. Subcell Biochem. 2014;74:119–35.

Votypka J., Lukes J., Obornık M. (2004). Phylogenetic relationship of *Trypanosoma corvi* with other avian trypanosomes. Acta Protozool 43, 225–231.

Young JC, Agashe VR, Siegers K, Hartl FU. Pathways of chaperone-mediated protein folding in the cytosol. Nat Rev Mol Cell Biol. 2004 Oct;5(10):781–91.

Zuo D, Subjeck J, Wang X-Y. Unfolding the Role of Large Heat Shock Proteins: New Insights and Therapeutic Implications. Front Immunol. 2016;7:75.

The expanding superfamily of protein sharing α-crystallin domain (ACD), including the Small Heat Shock Protein (sHSP) family, revealed by surveys in the genomes of a large diversity of *Trypanosoma* species

Costa-Martins, A.G.[a]; Alves, J. M. P[a], Serrano, M. G.[b], Camargo, E. P[a], Buck G. A.[b]; Marta M. G. Teixeira[a]*

[a]Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, SP, Brazil
[b]Department of Microbiology and Immunology and the Center for the Study of Biological Complexity, Virginia Commonwealth University

Abstract

Background: *Trypanosoma* are a widespread and successful group of protists, all obligate parasites of a range of invertebrates and vertebrates, showing highly variable life cycles coordinated by sophisticated regulatory machineries such as the complex cooperation of Heat Shock Proteins (HSP). Previous studies revealed a single sHSP sharing the conserved α-crystallin domain (ACD) in the human pathogens *Trypanosoma cruzi*, *Trypanosoma brucei*, and *Leishmania* spp. ACD proteins are involved in avoiding aggregation of unstable/misfolded proteins, coordination of cell division, and resistance to stressful conditions such as high temperature and oxidative stress. Here, we investigated the repertoire of ACD proteins in a large number of species representatives of the main clades of the genus *Trypanosoma*.

Methods: RPS-BLAST based searches were employed to assess the repertoire of proteins containing the ACD in the genomes of *T. cruzi, T. brucei*, and other 25 species of trypanosomes from mammals, lizard, snake, bird, frogs, and crocodiles. We further searched for homologues of the ACD proteins in genomes of other kinetoplastids. Taking into account that p23/α-crystallin proteins share a single conserved domain and very low levels of similarity, the ACD regions (~100 residues) were compared using structural data, and sequence divergences were compared in a 2D MDS plot.

Results: All sequences retrieved from the genomes were analyzed regarding similarity and domain architecture by comparison with orthologs from other organisms. The genomic survey for ACD proteins in *Trypanosoma* reveals nine proteins: *Try*DYX1C1, *Try*NudC1, *Try*NudC2, *Try*p23A, *Try*P23B, HSP20, *Try*SGT1, *Try*ACDP, and *Try*ACD-TPR. Besides the sHSP present in eukaryotes, archaea, and bacteria, six additional ACD proteins identified in the trypanosomatid genomes were assigned to previously known eukaryotic ACD protein families. In addition, results disclosed two new trypanosome ACD families with no clear orthologs identified in other organisms.

Conclusions: Our genome searches revealed a single sHSP in the genomes of all trypanosomatids investigated, and a conserved repertoire formed by eight additional ACD proteins in all species of *Trypanosoma* regardless of their remarkable differences in hosts, life cycles, cell compartments, and mechanisms employed to survive in stressful situations. Orthologs of all ACD proteins were also identified in *Leishmania,* insect-restricted trypanosomatids, and *Bodo spp.* Seven ACD were conserved in eukaryotes in general, whereas two were novel. To uncover some clues about putative functions of ACD proteins in trypanosomes, domain architecture and functional motifs were compared with those of ACD archetypes.

Keywords: Trypanosoma, protist, α-crystallin domain, ACD proteins, Small Heat Shock Protein.

INTRODUCTION

Trypanosoma (Euglenozoa, Kinetoplastida) are successful, widespread parasites found infecting vertebrates of all classes and being transmitted by hematophagous insects, ticks, and leeches (Stevens et al. 2001; Hamilton et al. 2007). Although some trypanosomes are pathogenic to their vertebrate host, such as *T. cruzi* and *T. brucei*, responsible respectively for human Chagas disease and African Sleep sickness, and *T. congolense*, *T. brucei, T. vivax*, and *T. evansi*, which are pathogenic to livestock, the majority of trypanosomes are harmless to their host (Stevens et al. 2001; Hamilton et al. 2004). Trypanosome species exhibit remarkably different behaviors while infecting both vertebrate hosts and vectors. Inside its vertebrate hosts, the majority of trypanosomes proliferate exclusively in the bloodstream, but a few can also invade extra-vascular tissues, such as the central nervous system (*T. brucei* and *T. vivax*) and skeletal and cardiac muscles (*T. cruzi*), and some species develop in intracellular compartments (Stevens et al. 2001; Hamilton et al. 2007). Trypanosomes also develop different strategies to proliferate and keep their vectors infected. While some species invade and multiply in insect salivary glands, being transmitted by inoculation during vector feeding, others develop in different parts of the vector gut, infecting a new vertebrate host by contamination with the feces, eliminated by the vectors after hematophagic feeding (Hoare 1972; Hamilton et al. 2007).

The success of the different species of trypanosomes in their complex and variable life cycles relies on their capacity of deal with stress conditions caused by environmental changes in their transition between invertebrate and vertebrate hosts. Indeed, environmental changes usually trigger parasite differentiation, promoting adaptation and progression in their life cycles. Accumulated evidence indicates that Hsp chaperone systems play important roles in parasite adaptation and proliferation. The number of sHSP paralogs varies broadly according to organism, with a few in single celled eukaryotes and up to dozens in animals and plants (Haslbeck et al. 2005)

The α-crystallin and p23 domains represent an ancient and conserved structure present in many eukaryotic protein families, and in small heat shock proteins ubiquitous in eukaryotes and prokaryotes (Garcia-Ranea et al. 2002). In its turn, p23 is an important HSP90 partner, participating in the folding of different regulatory proteins and other cellular processes (Felts and Toft 2003). Many eukaryotic protein families containing the p23 domain have been recently characterized in fungi, metazoans, and plants. These families include p23, SGT1, melusin, Rar1, CacyBP/SIP, NudC, DYX1C1, and B5+B5R NAD(P)H oxidoreductase B (Garcia-Ranea et al. 2002).

Functional studies have shown that ACD (p23/α-crystallin) proteins act as molecular chaperones in diverse cellular pathways, such as coordination of cell proliferation, differentiation, migration, signalization, formation of cytoskeleton, and also immunological modulation in mammals, and activation of pathogen resistance mechanisms in plants (Felts and Toft 2003; Spiechowicz and Filipek 2005; Mayor et al. 2007; Meldau et al. 2011; Bondino et al. 2012; LoTurco and Tarkar 2013; Fu et al. 2016).

Despite low similarity of amino acid sequence, p23 and α-crystallin domains comprise a conserved structure of seven β-strands forming a compact antiparallel β-sandwich fold. Moreover, the shared chaperone activities and proteins containing both

domains suggest a common origin. Since sHSPs are present in eukaryotes, bacteria, and archaea, it seems that the p23 domain has diverged later from an α-crystallin ancestor, at least after eukaryotes and prokaryotes separated, and rapidly diverged to cover new functions (Garcia-Ranea et al. 2002).

Although studies of α-crystallin and p23 domains in yeast and metazoans reveal that these proteins are involved in many complex and vital cellular functions, little is known about the repertoire and functions of these proteins in parasitic protozoans. Among the HSPs that constitute the large repertoire of chaperones/co-chaperones characterized in the trypanosomatids, a single sHSP (HSP20), and no other protein families containing the ACD (p23/α-crystallin), was identified in the genomes of *T. brucei* and *T. cruzi* (Folgueira and Requena, 2007; Ürményi et al., 2014). In *Leishmania*, HSP20 was also the sole sHSP reported to date (Folgueira et al., 2015). Functional studies disclosed p23 proteins in *Leishmania* spp., and searches using p23 co-chaperone sequences disclosed two additional ACD-containing proteins in *L. donovani* essential for the parasite viability and infectivity of host cells, and then investigated as potential immunoprophylactics (Perez-Morales et al. 2009; Batista et al. 2015; Hombach et al., 2015). HSP20 has been proven to be a useful molecular marker to distinguish between *Leishmania* species (Fraga et al. 2013; Montalvo et al. 2014).

Our main goal in the present study was to use *in silico* approaches to systematically investigate the genomic repertoire of p23/α-crystallin proteins in a highly comprehensive set of trypanosome species comprehending representatives of the main phylogenetic clades. Phylogenetically distant trypanosome species differ in many traits on host-parasite-vector interactions, varying in vertebrate hosts, vectors, life cycle, metabolism, cell compartments, and responses adopted to deal with stressful situations. In addition, we compared sequences from the archetypical and trypanosomes ACD proteins, aiming to assess differences in domain architectures and functional motifs.

MATERIALS AND METHODS

Organisms and genomic sequences

The draft genomes of the following organisms were used in this study: *T. serpentis* (TCC1052), *T. theileri* (TCC165), *T. rangeli* (TCCAM80), *T. lewisi* (TCC34), *T. conorhini* (TCC025E), *T. dionisii* (TCC211), *T. erneyi* (TCC1946), T. wauwau (TCC1873), *T. livingstonei* (TCC1270), *T. ralphi* (TCC1838), *T. cruzi marinkellei* (B7), *T. cyclops* (TCC52), *Trypanosoma* sp. (TCC1825), *Trypanosoma* sp. (TCC878), *Trypanosoma* sp. (TCC339), *Trypanosoma* sp. H25 (TCC16), *Trypanosoma* sp. (TCC1307), *Trypanosoma* sp. (TCC2186), *T. brucei* (TREU 927), *T. brucei* (Lister 427), *T. brucei gambiense* (DAL972), *T. vivax* (Y486), *T. congolense* (IL3000), *T. grayi* (ANR4), *T. evansi* (STIB805), *Bodo saltans* (Lake Konstanz), *Bodo* sp. (ATCC50149), *Herpetomonas muscarum* (TCC001E), and *Leishmania major* (Friedlin). Organisms with TCC culture collection identifiers are cryopreserved at the Trypanosomatid Culture Collection of the University of São Paulo (TCC-USP) and were grown in LIT media (Camargo, 1964) supplemented with 2% FBS (Fetal Bovine Serum).

Genomic search and sequence alignment.

All proteins analyzed here share a common signature represented by the ACD. Due to low levels of similarity in the whole sequence, searches were

conducted with RPS-BLAST using a maximum expect value (E-value) threshold of 1e-5 against genome open reading frames using getorf program of EMBOSS package v6.6.0.0 with standard parameters. As RPS-BLAST database, we have used the available position-specific score matrices (PSSMs) for ACD in the Conserved Domain Databases (CDD -NCBI http://www.ncbi.nlm.nih.gov/cdd). The list of ACD PSSMs, including description and accession numbers, is supplied in Supplementary file 2. A BLASTp search was then performed against NR with a maximum E-value threshold of 1e-20, using the sequences found in RPS-BLAST as queries. The retrieved sequences were annotated according to the presence of other conserved domains and similarity with sequences in NR. The PSI-BLAST searches were conducted using initial E-value thresholds of 1e-5 and 1e-3, with only subjects showing at least 50% query coverage included in the consecutive rounds. The searches were interrupted in the fourth round and any subjects found were analyzed.

Analyses of ACD signature and alignment of the repertoire found in trypanosomes were conducted using GLAM2 (Gapped Local Alignment of Motifs) (Frith et al. 2008) including in MEME software suit version 3.5.4 (Bailey et al. 2009). The tertiary protein structures of trypanosome ACD proteins were constructed using the SWISS-MODEL homology-modeling server (Biasini et al. 2014). Models were visualized and aligned using the SwissPdbViewer v4.1 program (Guex and Peitsch 1997). Average amino acid identity between trypanosome ACD proteins was calculated with MEGA software v6.0 using the P-distance method and pairwise deletion for missing data. To provide a visual representation of the ACD distance matrix, we used the multidimensional scaling (MDS) plot with two dimensions (2D). The MDS statistical analysis and graphing were performed using the Bios2mds package of the R statistical language environment for statistical computing (Pelé et al. 2012).

RESULTS AND DISCUSSION

1. ACD protein repertoire in *Trypanosoma*

Our genomic survey for genes encoding proteins belonging to the p23/α-crystallin superfamily reveals the existence of nine proteins. The α-crystallin domain was named after the α-crystallin chaperone, a member of the sHSP family and the major constituent of the vertebrate eye lens, essential for preventing aggregation of denatured proteins and for keeping lens transparency (Horwitz 1992; Haslbeck et al. 2005). All sequences retrieved were classified according to domain architecture and sequence similarity with previously known ACD proteins. Seven trypanosome ACD genes were assigned to five known ACD protein groups reported in other organisms: HSP20(1), p23(2), NudC(2), DYX1C1(1), and SGT1(1). These proteins were reported to be involved as molecular chaperones in diverse cellular and systemic pathways, which include proliferation, expression control, cell migration, cytoskeleton structure, and immunological responses (Table 1). In addition, two new proteins were found, named here as *Try*ACDP and *Try*ACD-TPR.

The nine proteins identified (HSP20, *Tryp*23A, *Try*P23B, *Try*DYX1C1, *Try*NudC1, *Try*NudC2, *Try*SGT1, *Try*ACDP and *Try*ACDTPR) are shared by all surveyed trypanosome species, regardless of host range, cell cycle or infective strategies (Figure 1). The comparison with other trypanosomatids and free-living bodonids reveals that all ACD proteins are most likely present across the

Table 1: General functions of ACD/p23 proteins

Gene	Bacteria	Trypanosomatids	Yeast	Plants	Vertebrates	References
HSP20	Chaperone and cochaperone activities Thermotolerance	Chaperone and cochaperone activities Thermotolerance Protection to acidic pH, ethanol and redox stresses	Chaperone and cochaperone activities Thermotolerance	Chaperone and cochaperone activities Thermotolerance Regulation of development	Keep eye lens transparency Chaperone and cochaperone activities Thermotolerance	(Bondino, Valle, e Ten Have 2012; X. Fu 2014; PérezMorales e Espinoza 2015)
p23	Absent	Protects parasites against HSP90 inhibitors Cochaperone activities	Chaperone and cochaperone activities Transcriptional activation	Regulation of root development	Chaperone and cochaperone activities Prostaglandin synthase activity Lowers the ATPase activity rate of HSP90 Steroid receptor chaperoning	(D'Alessandro et al. 2015; Felts e Toft 2003; Hombach et al. 2015)
SGT1	Absent	Unknown	Kinetochore assembly Chaperone and cochaperone activities	Pathogen defences via R proteins Regulation of development	Essential for inflammasome activity	(Bansal et al. 2009; Mayor et al. 2007; Meldau, Baldwin, e Wu 2011)
DYX1C 1	Absent	Unknown	Unknown	Unknown	Neuronal migration Cilia structure and function Estrogen receptor transport Human dyslexia pathogenesis Putative chaperone and cochaperone activities	(Massinen et al. 2009; Tammimies et al. 2013)
NudC family	Absent	Unknown	Dynein-dependent nuclear migration Chaperone and cochaperone activities	Heat shock protection and chaperone activity (BOBBER)	Neuronal migration Megakaryocytopoiesis and thrombopoiesis Ciliogenesis Monocyte to macrophage differentiation Dynein-dependent nuclear migration Chaperone and cochaperone activities	(Q. Fu et al. 2016)

Kinetoplastea, and probably also present at least in a common ancestor of parasitic and free-living kinetoplastids. Interestingly, the repertoire in *Bodo* comprehends at least three other genes encoding ACD proteins, so far seemingly absent in *Trypanosoma*.

The comparison of the ACD domains (~100 amino acid sequence) shows that they are highly polymorphic, showing identity values ranging from ~10 up to ~25% between protein sequences, e. g. NudC and p23 share ~16% amino acid identity. Nevertheless, the structural signature of ACD was recovered from all sequences (Figure 2A). The multiple sequence alignment of ACD obtained using MEME and confirmed by structure homology reveals a few conserved positions across the whole repertoire (Figure 2B). The MDS analysis of the ACD alignment reveals that each gene family forms a tight group of orthologous sequences, with the exception of the NudC family. *Try*NudC1 and *Try*NudC1 displayed an overlapped distribution, indicating that the ACD is not able to clearly discriminate between these two paralogs (Figure 3). However, the whole protein domain architecture is different for both (Figure 1). The relative position of the HSP20 family and its ubiquitous distribution in bacteria, archaea, and

eukaryotes favors the hypothesis that the small heat shock proteins are the ancient ACD chaperones and that the remaining ACD proteins diverged and diversified after in the eukaryotes. The specific features of each ACD protein are discussed in sections below.

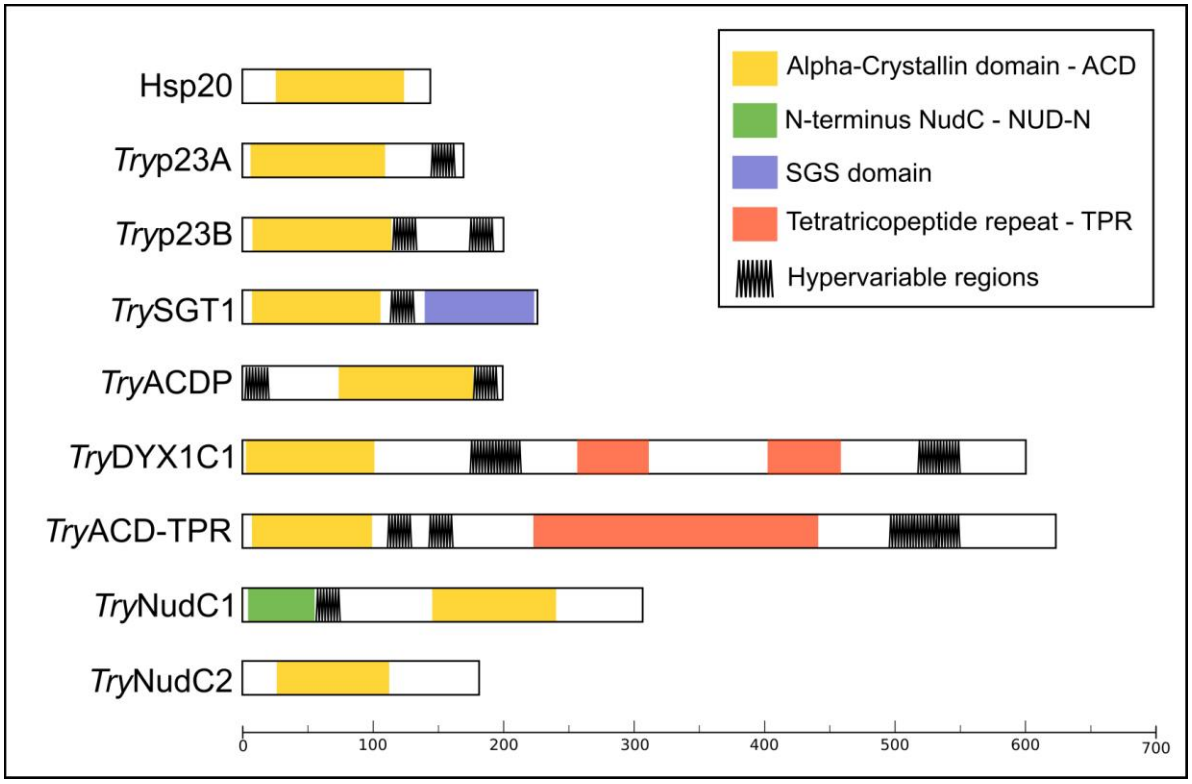

Figure 1: Domain architecture of trypanosome ACD proteins – Schematic representation showing protein length, conserved domains, and hypervariable regions found in trypanosome ACD proteins.

2. Small heat shock proteins

The sHSPs are one of the major groups of heat shock proteins being found throughout the Archaea, Bacteria, and Eukaryota superkingdoms. Phylogenetic analyses suggest that this family diverged very early in evolution and quickly originated many paralogs (Haslbeck et al. 2005). Although these proteins share a common signature represented by the α-crystallin domain, sequence similarity is the lowest compared to other families of HSPs, with an overall amino acid sequence identity of less than 50% (Pérez-Morales and Espinoza 2015). Despite their sequence and size diversities, sHSPs tend to form large oligomers and act as molecular chaperones suppressing protein aggregation and keeping protein homeostasis in stress and physiological conditions (Haslbeck et al. 2005; Pérez-Morales and Espinoza 2015).

Parasites often display complex life cycles, where each life stage faces a new environmental condition and a stress challenge. Indeed, sHSPs are overexpressed in the life cycle of many parasites, participating in the differentiation process and adaptation to new conditions (Maresca & Carratu 1992; Pérez-Morales and Espinoza 2015).

Our searches corroborate that HSP20 is the only sHSP in the trypanosomatids (Folgueira e Requena 2007; Perez-Morales et al. 2009; Requena et al., 2015). The comparison with the kinetoplastid *Bodo* suggested that the sHSP family remains unchanged at least since the free-living ancestor of

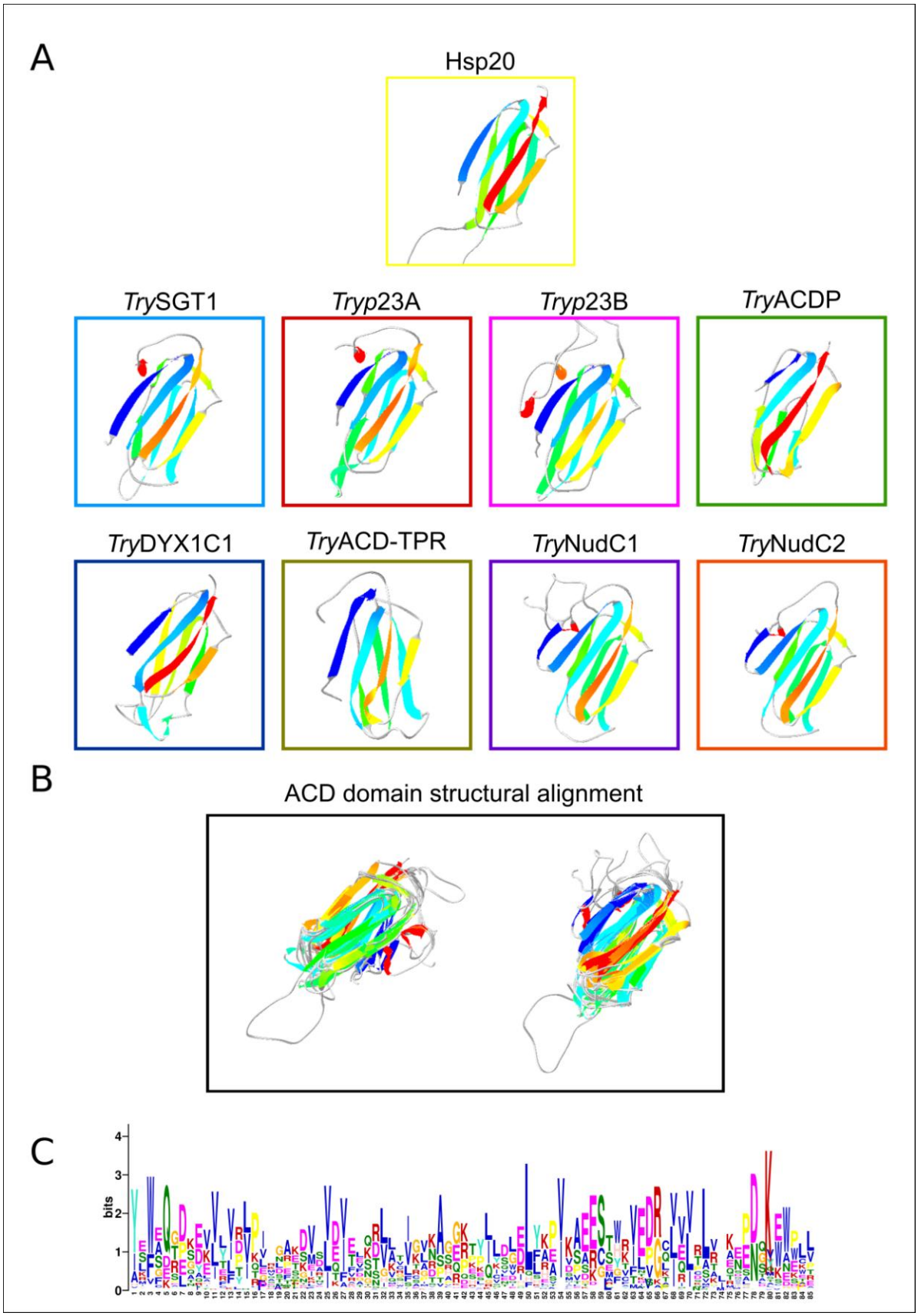

Figure 2: Structural conservation of the alpha-crystallin domain in trypanosome ACD proteins. A – Three-dimensional structure model of p23/alpha-crystallin domains from each trypanosome ACD protein. Models were obtained using the SWISS MODEL server (http://swissmodel.expasy.org/) and visualized with Swiss Pdb-Viewer. B – ACD domain structural alignment, using Swiss Pdb-Viewer, of all models. C – Amino acid logo representing the multiple sequence alignment of the p23/alpha-crystallin domain in trypanosome ACD proteins.

Table 2: Amino acid identity along the alpha-crystallin domain in trypanosome ACD proteins

	HSP20	*Tryp*23A	*Tryp*23B	*Try*SGT1	*Try*DYX1C1	*Try*ACDP	*Try*ACD-TPR	*Try*NudC1	*Try*NudC2
HSP20	-								
***Tryp*23A**	13.4	-							
***Tryp*23B**	16.2	21.2	-						
***Try*SGT1**	10.9	15.0	20.5	-					
***Try*DYX1C1**	16.2	16.1	20.6	19.2	-				
***Try*ACDP**	11.0	20.8	18.4	15.7	23.0	-			
***Try*ACD-TPR**	11.9	13.5	16.0	16.6	25.9	21.0	-		
***Try*NudC1**	15.0	13.2	20.3	16.8	22.9	19.4	24.4	-	
***Try*NudC2**	14.9	13.0	17.3	16.7	20.5	20.5	23.4	42.5	-

bodonids and trypanosomatids. The HSP20 gene encodes a ~141 amino acid protein with no hypervariable regions or domains other than the ACD (Figure 1). Although HSP20s are quite conserved, showing around 60% identity between the orthologs from distantly related species, they varied between 10.9 and 16.2% identity compared to other ACD families (Table 2, Figure 3).

3. p23 family

The p23 family is exclusive to eukaryotes and displays a highly conserved sequence from fungi to vertebrates (Garcia-Ranea et al. 2002). It was first discovered as a co-chaperone participating in the HSP90 complex with progesterone (Johnson et al. 1994). p23 is one of the smallest partners of the HSP90 machinery and possesses a simple structure formed by a single domain, the ACD, and by a flexible tail at the C-terminus (Felts & Toft 2003). Although it displays a simple architecture, this co-chaperone presents a complex spectrum of activities, being involved in several cellular regulatory pathways. The multifunctional role played by p23 is possible through its interaction with HSP90 and a diverse range of client proteins, including polymerases such as telomerases, steroid receptors, transcription factor Hsf1, nitric oxide synthase, and a variety of protein kinases (Garcia-Ranea et al. 2002; Felts and Toft 2003). Besides its participation in the HSP90 machinery, p23 also presents chaperoning activity by itself, suppressing the aggregation of denatured proteins *in vitro* (Freeman et al. 1996).

Two homologs of p23 were reported in the genome of pathogenic trypanosomatids *L. major*, *T. cruzi*, and *T. brucei* (Folgueira and Requena 2007), and then recently characterized in *L. braziliensis* and named as Lbp23A and Lbp23B (Batista et al. 2015). Our searches reveal two p23 homologs in all trypanosomes. These p23 genes were named *Tryp*23A and *Tryp*23B to correspond to *L. braziliensis* homologs Lbp23A and Lbp23B, respectively. *Tryp*23A encodes a ~180 amino acid protein, while *Tryp*23B shows ~200 residues. Both homologs present no conserved domain other than the ACD. While *Tryp*23A displays a hypervariable region, which is rich in GG[VLM], DD, and EE repetitions, *Tryp*23B shows two hypervariable regions with fewer repetitions of the same motifs (Figure 1). The ACDs in *Tryp*23A and *Tryp*23B share 21.2% amino acid identity, and diverged from 10.8 to 20.8% from other repertoire members (Table 2).

4. SGT1

The suppressor of G2 allele of SKP1 (SGT1) is highly conserved in eukaryotes, being found in plants,

animals, and yeast (Spiechowicz e Filipek 2005; Meldau et al. 2011). SGT1 was first identified in yeast (*Saccharomyces cerevisiae*) as a suppressor of SKP1 and essential for kinetochore assembly (Kitagawa et al. 1999). The SGT1 architecture includes three conserved domains: a tetratricopeptide repeat (TPR) in the N-terminus, followed by the ACD and by the SGS domain in the C-terminus. In addition, the ACD is separated from adjacent domains by two hypervariable regions (V1 and V2) (Azevedo et al. 2002). The ACD and SGS are responsible for the interaction with HSP90 and HSP70 respectively, while TPR and ACD are the binding sites for SKP1 (Kitagawa et al. 1999; Noël et al. 2007; Yan et al. 2012). Interactions with heat shock proteins were described in yeast, mammals, and plants, indicating that SGT1 plays roles in the chaperone assembly machinery of diverse protein complexes. In addition, the SGT1 expression pattern matches that found in the well known heat shock proteins, thus being up-regulated by heat shock and stress conditions (Zabka et al. 2008). The double knock-out for both copies of STG1 present in *Arabidopsis* is lethal while knockdown of tomato orthologs leads to different degrees of growth defects, indicating that it is important for plant development (Azevedo et al. 2002; Bhattarai et al. 2007). SGT1 and HSP90 together are also involved in regulation of plant defense by interacting with a large number of disease-resistance proteins (R proteins) (Meldau et al. 2011). R proteins are responsible for detecting pathogen-associated proteins and for triggering complex defense pathways, which include tissue reparation and cell death. Moreover, mammals possess intracellular immune proteins that are structurally related to R proteins, known as Nod-like receptors (NLRs), essential for inflammasome assembly (Mayor et al. 2007). Recent evidence suggests that the complex SGT1/HSP90 interacts with NLRs using the same mechanism observed in SGT1/HSP90 and R proteins in plant, and SGT1/HSP90 and SKP1 during the kinetochore assembly in yeast. This suggests that the role played by the SGT1/HSP90 in immunological response in mammals, pathogen defense in plants, and kinetochore assembly in yeast convergently evolved using a common mechanism (Mayor et al. 2007).

Our searches revealed a member of this family, the *Try*SGT1, in the trypanosomatids. In the majority of trypanosome genomes, only a single copy was recovered. Exceptions are *Trypanosoma* spp. TCC2045, from bat, and TCC878 from frog, which show multiple nearly identical copies. The *Try*SGT1 gene encodes a ~215 amino acid protein displaying 64.9% amino acid identity between the orthologs. *Try*SGT1 is shorter than the homologs found in yeast and metazoans, and lacks the TPR and the first hyperveriable region in the N-terminus (Figure 1). The SGT1 gene found in free-living bodonids encodes a large protein with 292 amino acid residues due to an increase in the N-terminus but, as seen in *Trypanosoma*, the TPR domain is absent. Although both ACD and TPR are important to SGT1 dimerization and SPK1 interaction, mutations in the TPR domain are reported to prevent the SGT1 and SPK1 binding *in vivo,* but not *in vitro* (Bansal et al. 2009). This suggests that *Try*SGT1 might present alternative function, partners or interactions. The ACD present in *Try*SGT1 is more similar to the ones in *Tryp*23B (20.5%) and *Try*DYXC1 (19.2%) than to those in other repertoire members (~15.3%) (Table 2 and Figure 3). The absence of TPR, essential to SKP1 binding *in vivo,* rises questions about SKP1 homologs in the trypanosomes. To investigate the

presence of this homolog, we have used a yeast sequence (accession P52286) as query for BLAST search. A single SKP1 gene was recovered from the searched genomes, which displays ~45% amino acid identity with the query sequence. The *T. brucei* candidate (Tb11.02.3990) was previously annotated as SKP1-like protein, and the RNAi knockdown inhibited bloodstream-form growth (Benz & Clayton 2007). Since TPR domains are absent from *Try*SGT1, the interaction with SPK1 in trypanosomes remains unclear.

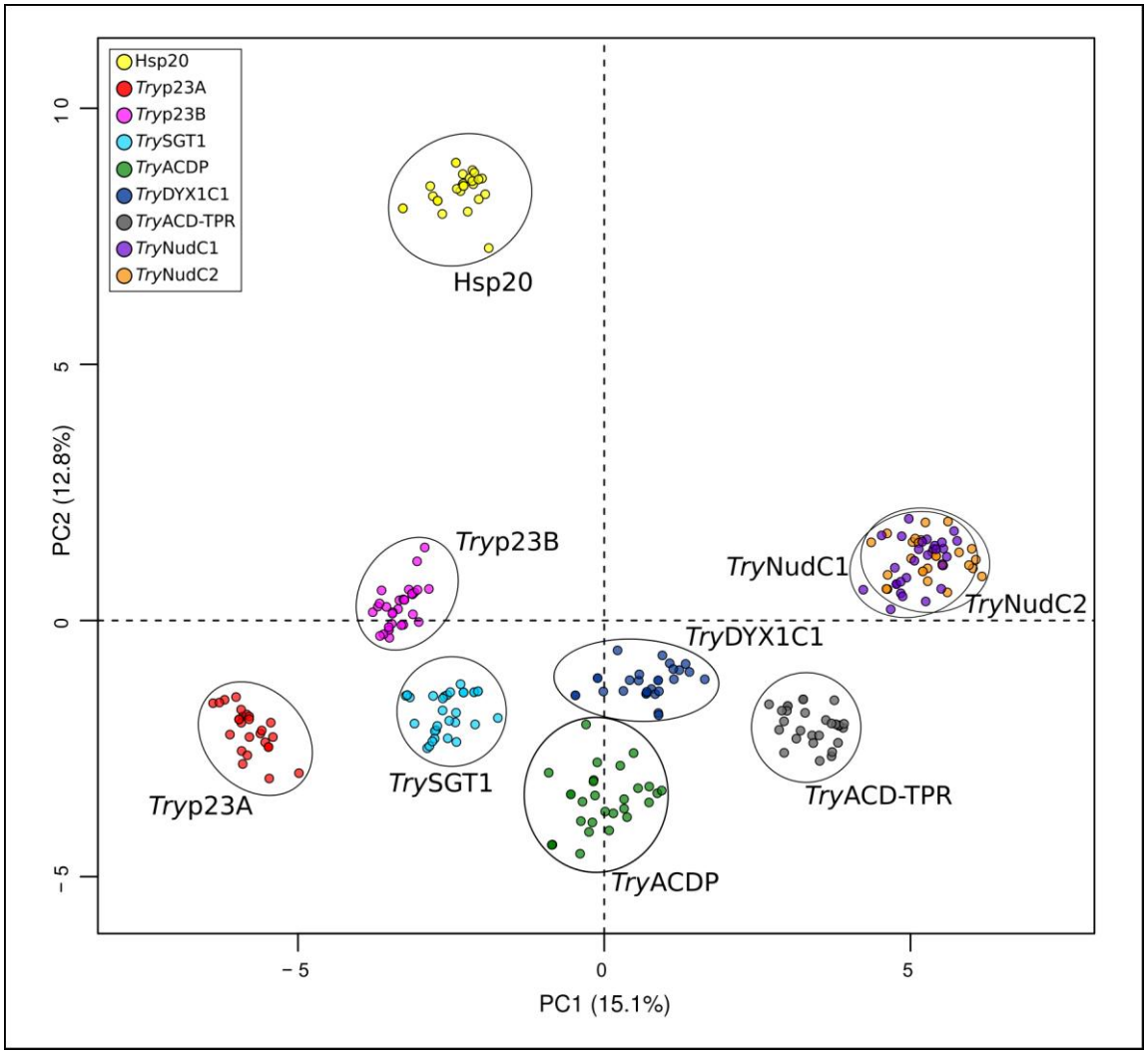

Figure 3: Multidimensional scaling (MDS) plot of two dimensions representing the visual amino acid distance between p23/alpha-crystallin domains of trypanosome ACD proteins.

5. NudC family

The Nuclear distribution C (NudC) proteins were first described as associated with dynein-dependent nuclear migration in the reproductive cycle of *Aspergillus nidulans* (Osmani et al. 1990). The depletion of NudC in this organism causes severe morphological and compositional changes in cell wall, leading to death (Chiu et al. 1997). The deacetylation of NudC regulates mitotic progression, and its failure is responsible for chromosome misalignment and missegregation (Chuang et al. 2013). In addition, either downregulation or overexpression of NudC is capable of inducing cytokinesis defects (Aumais et al. 2003; Zhou et al. 2003). Besides its multiple roles in cell cycle progression, accumulated evidence has shown that these proteins have acquired new

functions in metazoans, being implicated in neuronal migration, regulation of the inflammatory response, and thrombopoiesis (Fu et al. 2016). Proteomic and histological studies of human midbrain stem cells have shown that hNUDC is highly expressed during neuronal differentiation. Because of its interaction with LIS1, which is associated with Miller–Dieker disease, NudC genes may also play roles in neuronal differentiation (Morris et al. 1998).

The interaction NudC/LIS1 is also associated with regulation of inflammatory pathways via PAF (Riera et al. 2007). It is thought that acquisition of these new functions in animals is due to the presence of conserved motifs in the N-terminus of NudC, which is absent in microbial orthologs (Riera & Lazo 2009). As expected by the presence of the α-crystallin domain, accumulated evidence indicates that NudC possesses chaperone activity or acts independently as a holdase, stabilizing or enhancing folding of client proteins and acting as a co-chaperone of HSP90, modulating its ATPase activity (Faircloth et al. 2009; Zhu et al. 2010).

Highly conserved NudC homologs have been identified in several eukaryotes, and four paralogs were recognized in this family: NudC (Osmani et al. 1990; Matsumoto and Ledbetter 1999); NudC-like (NudCL) (Zhou et al. 2006); NudC-like 2 (NudCL2) (Yang et al. 2010); and the divergent NudC domain-containing protein 1 (NudCD1) (Yang et al. 2001). Recently, a new member belonging to this family, BOBBER1, was described in *Arabidopsis* sp., and shown to act as a non-canonical sHSP (Perez et al. 2009). Although family members display differences in sizes and sequences, all posses a common α-crystallin domain, similar to those found in p23 and sHSPs (Garcia-Ranea et al. 2002; Fu et al. 2016). NudC and NudCL, but not NudCL2, possess a conserved domain in the N-terminus called NUDC-N.

Our search in the trypanosome genomes reveals a single copy of each of two members belonging to NudC family, here referred to as *Try*NudC1 and *Try*NudC2. The comparison of the ACD from the NudC homologs shows that they are more similar to each other (42.5%) than to any other repertoire member (identity ranging from 13.0 to 24.4%), and present an overlapped distribution in the MDS plot (Figure 3). The comparison of complete trypanosome NudC protein sequences with other eukaryotic homologs reveals that *Try*NudC1 is more similar to human hNudC (45%) than to the archetype described in *Aspergillus* (32.5%). In addition, *Try*NudC1 also presents a conserved domain called NUD-N, present only in NudC, but not NudC-like paralogs. On the other hand, *Try*NudC2 proves to be more similar to the fungal archetype (28.9%) than to vertebrate paralogs: NudC (19.6%), NudCL (13.8%), and NudCL2 (19.6%). In contrast to *Try*NudC1 and as observed in the fungal archetype, the NUD-N domain is absent from *Try*NudC2 (Figure 1). Moreover, comparison between *Try*NudC1 and animal NudC proves that both display a similar size (304 and 330 amino acids, respectively) and a conserved N-terminus stretch in relation to hNudC, which is recognized as a modulator of macrophage inflammatory response. In contrast, in smaller fungal orthologs (198 residues) the N-terminal extension is absent. *Try*NudC1 also displays a hypervariable region between residues 66-88, absent in TryNudC2. The novel functions present in animal NudC may be due to the N-terminal extension proprieties or by the protein moonlighting phenomenon. Thus, the similarity of trypanosome and animal NudC sequences and the inhibitory effect of NudC on PAF

pathways, and consequent reduction of monocyte differentiation, rises some questions about the functions of this gene in trypanosomes.

6. DYXC1

The dyslexia susceptibility 1 candidate 1 (DYX1C1) is implicated in several cellular process, such as cytoskeletal structure formation, protein folding, hormonal pathways, cilia structure and function, and neural development (Massinen et al. 2009; Chandrasekar et al. 2013; Tammimies et al. 2013). Although DYX1C1 was first related to dyslexia, the mechanisms and SNP polymorphisms correlating gene and disease have not been well established (LoTurco & Tarkar 2013; Tammimies et al. 2013). A molecular network study indicates that DYX1C1 modulates expression of genes involved in cell migration and nervous system development, and suggests its possible involvement in the etiology of dyslexia (Tammimies et al. 2013). DYX1C1 displays the ACD domain in the N-terminus and a variable number of tetratricopeptide repeats (TPR) in the C-terminus. TPR domains are found in several eukaryotic and prokaryotic proteins and are usually present in those involved in chaperone machinery, cell-cycle, transcription, and protein transport complexes. DYX1C1 homologs have been found in several eukaryotes, ranging from single celled protists to vertebrates. In addition, only a single gene is usually found and phylogenetic analyses indicate that DYX1C1 is highly conserved and evolution of the gene is compatible with expected eukaryotic species evolution (Tammimies et al. 2013).

Our genomic survey recovered a single DYX1C1 homolog in trypanosomes, referred to here as *Try*DYX1C1. Its gene encodes a ~600 (579 to 630) amino acid protein, which is bigger than the human archetype (420 aa) and displaying around 25% amino acid identity to the full length human homolog sequence. Except for *T. lewisi*, *T. ralphi*, and *Trypanosoma* spp. (TCC2045 and TCC878), where two or more copies coding identical or nearly identical proteins were found, in most trypanosomes a single copy was recovered. The comparison of *Try*DYX1Cs reveals a highly conserved sequence, with 73.5% amino acid identity between trypanosome orthologs. The alignment also reveals two hypervariable regions between residues 170-218 and 533-550 in the alignment (Figure 1). Size variation among *Try*DYX1C1 is almost exclusively due to differences in the two hypervariable regions. The ACDs present in *Try*DYX1C1 are more similar to the *Try*ACDP and *Try*ACD-TPR orphans (23.0 and 25.9%, respectively) than to other ACD superfamily members found in *Trypanosoma* (Figure 3).

***7. Try*ACDP and *Try*ACD-TPR**

Our search uncovered two new ACD proteins in *Trypanosoma*, named according to their domain architecture as *Try*ACDP and *Try*ACD-TPR. *Try*ACDP orthologs were also recovered from non-trypanosome trypanosomatid and bodonid genomes. *Try*ACDP is on average 200 amino acids long and is the most polymorphic ACD protein in *Trypanosoma*, exhibiting ~56% amino acid identity among the orthologs in the genus. In contrast, *Try*ACD-TPR is more than three time as long (~620 residues) and displays ~59% identity between trypanosome species. Despite a similar domain architecture with p23 and HSP20 (Figure 1), the ACD sequences from *Try*ACDP are more similar to those of *Try*DYX1C1 (23.0%) than to the remaining repertoire (Table 2). On the other hand, *Try*ACD-TPR possesses a *Try*DYX1C1-like architecture presenting TPR motifs

flanked by hypervariable regions and also presents an ACD similar to *Try*DYX1C1 (25.9%)(Figure 1 and Table 2). However, the sequence similarities between *Try*DYX1C1 and *Try*ACD-TPR are restricted to the ACD and TPR domains.

To investigate the existence of proteins related to *Try*ACDP and *Try*ACD-TPR in other organisms, we conducted a PSI-BLAST search against NR using both proteins from *T. brucei* as queries. Using *Try*ACDP, we recovered uncharacterized proteins from pathogenic oomycetes *Saprolegnia*, *Phytophthora*, *Aphanomyces*, and *Plasmopara* (accession numbers XP_009521390.1, XP_002905176.1, XP_008896317.1, CEG43972.1, XP_008605824.1, and XP_009829214.1) after the fourth search round. These proteins are similar in architecture and size and but share only ~28% identity with *Try*ACDP and sequence similarity restricted to ACD segment. Thus no clear ortholog was recovered for *Try*ACDP in oomycetes. The *Try*ACD-TPR search resulted in several protein families containing the TPR domain, but no similarity outside this domain and with no ACD domain identified.

CONCLUSIONS

The trypanosome ACD repertoire presents seven proteins assigned to groups known in other organisms (*Try*DYX1C1, *Try*NudC1, *Try*NudC2, *Try*p23A, *Try*P23B, HSP20, *Try*SGT1) and two proteins shared only by kinetoplastids (*Try*ACDP and *Try*ACD-TPR). The comparison of the genomic repertoires of other trypanosomatids and the closely related free-living bodonids reveals that this repertoire was probably present in a common ancestor of the kinetoplastids. Furthermore, the comparison with the family archetypes in other organisms reveals that unexpected molecular mechanisms and functions may exist in *Try*SGT1 and *Try*NudC1. Further functional studies are needed to understand the roles played by these proteins in the life cycles displayed in the genus *Trypanosoma.*

REFERENCES

Aumais JP, Williams SN, Luo W, Nishino M, Caldwell KA, Caldwell GA, et al. Role for NudC, a dynein-associated nuclear movement protein, in mitosis and cytokinesis. J Cell Sci. 2003;116(Pt 10):1991–2003.

Azevedo C, Sadanandom A, Kitagawa K, Freialdenhoven A, Shirasu K, Schulze-Lefert P. The RAR1 interactor SGT1, an essential component of R gene-triggered disease resistance. Science. 2002;295(5562):2073–6.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37(Web Server issue):W202-208.

Bansal PK, Nourse A, Abdulle R, Kitagawa K. Sgt1 dimerization is required for yeast kinetochore assembly. J Biol Chem. 2009;284(6):3586–92.

Batista FAH, Almeida GS, Seraphim TV, Silva KP, Murta SMF, Barbosa LRS, et al. Identification of two p23 co-chaperone isoforms in *Leishmania braziliensis* exhibiting similar structures and Hsp90 interaction properties despite divergent stabilities. FEBS J. 2015;282(2):388–406.

Benz C, Clayton CE. The F-box protein CFB2 is required for cytokinesis of bloodstream-form *Trypanosoma brucei*. Mol Biochem Parasitol. 2007;156(2):217–24.

Bhattarai KK, Li Q, Liu Y, Dinesh-Kumar SP, Kaloshian I. The MI-1-mediated pest resistance requires Hsp90 and Sgt1. Plant Physiol. 2007;144(1):312–23.

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014;42(Web Server issue):W252-258.

Bondino HG, Valle EM, Ten Have A. Evolution and functional diversification of the small heat shock protein/α-crystallin family in higher plants. Planta. 2012;235(6):1299–313.

Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17(4):540–52.

Chandrasekar G, Vesterlund L, Hultenby K, Tapia-Páez I, Kere J. The zebrafish orthologue of the dyslexia candidate gene DYX1C1 is essential for cilia growth and function. PloS One. 2013;8(5):e63123.

Chiu YH, Xiang X, Dawe AL, Morris NR. Deletion of nudC, a nuclear migration gene of *Aspergillus nidulans*, causes morphological and cell wall abnormalities and is lethal. Mol Biol Cell. 1997;8(9):1735–49.

Chuang C, Pan J, Hawke DH, Lin S-H, Yu-Lee L. NudC deacetylation regulates mitotic progression. PloS One. 2013;8(9):e73841.

Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinforma Oxf Engl. 2011;27(8):1164–5.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

Faircloth LM, Churchill PF, Caldwell GA, Caldwell KA. The microtubule-associated protein, NUD-1, exhibits chaperone activity in vitro. Cell Stress Chaperones. 2009;14(1):95–103.

Felts SJ, Toft DO. p23, a simple protein with complex activities. Cell Stress Chaperones. 2003;8(2):108–13.

Folgueira C, Requena JM. A postgenomic view of the heat shock proteins in kinetoplastids. FEMS Microbiol Rev. 2007;31(4):359–77.

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 J Mol Epidemiol Evol Genet Infect Dis. 2013;18:229–37.

Freeman BC, Toft DO, Morimoto RI. Molecular chaperone machines: chaperone activities of the cyclophilin Cyp-40 and the steroid aporeceptor-associated protein p23. Science. 1996;274(5293):1718-20.

Frith MC, Saunders NFW, Kobe B, Bailey TL. Discovering sequence motifs with arbitrary insertions and deletions. PLoS Comput Biol. 2008;4(4):e1000071.

Fu Q, Wang W, Zhou T, Yang Y. Emerging roles of NudC family: from molecular regulation to clinical implications. Sci China Life Sci. 2016;

Garcia-Ranea JA, Mirey G, Camonis J, Valencia A. p23 and HSP20/alpha-crystallin proteins define a conserved sequence domain present in other eukaryotic protein families. FEBS Lett. 2002;529(2–3):162–7.

Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis. 1997;18(15):2714–23.

Hamilton PB, Gibson WC, Stevens JR. Patterns of co-evolution between trypanosomes and their hosts deduced from ribosomal RNA and protein-coding gene phylogenies. Mol Phylogenet Evol. julho de 2007;44(1):15–25.

Hamilton PB, Stevens JR, Gaunt MW, Gidley J, Gibson WC. Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA. Int J Parasitol. 2004;34(12):1393–404.

Haslbeck M, Franzmann T, Weinfurtner D, Buchner J. Some like it hot: the structure and function of small heat-shock proteins. Nat Struct Mol Biol. 2005;12(10):842–6.

Hoare CA. The trypanosomes of mammals: a zoological monograph. Oxford,: Blackwell Scientific Publications; 1972.

Horwitz J. Alpha-crystallin can function as a molecular chaperone. Proc Natl Acad Sci U S A. 1992;89(21):10449–53.

Johnson JL, Beito TG, Krco CJ, Toft DO. Characterization of a novel 23-kilodalton protein of unactive progesterone receptor complexes. Mol Cell Biol. 1994;14(3):1956–63.

Kitagawa K, Skowyra D, Elledge SJ, Harper JW, Hieter P. SGT1 encodes an essential component of the yeast kinetochore assembly pathway and a novel subunit of the SCF ubiquitin ligase complex. Mol Cell. 1999;4(1):21–33.

LoTurco JJ, Tarkar A. DYX1C1 placed in a molecular context. Biol Psychiatry. 2013;73(6):497–8.

Maresca B, Carratu L. The biology of the heat shock response in parasites. Parasitol Today. 1992;8(8):260–6.

Massinen S, Tammimies K, Tapia-Páez I, Matsson H, Hokkanen M-E, Söderberg O, et al. Functional interaction of DYX1C1 with estrogen receptors suggests involvement of hormonal pathways in dyslexia. Hum Mol Genet. 2009;18(15):2802–12.

Matsumoto N, Ledbetter DH. Molecular cloning and characterization of the human NUDC gene. Hum Genet. 1999;104(6):498–504.

Mayor A, Martinon F, De Smedt T, Pétrilli V, Tschopp J. A crucial function of SGT1 and HSP90 in inflammasome activity links mammalian and plant innate immune responses. Nat Immunol. 2007;8(5):497–503.

Meldau S, Baldwin IT, Wu J. For security and stability: SGT1 in plant defense and development. Plant Signal Behav. 2011;6(10):1479–82.

Montalvo AM, Fraga J, Rodríguez O, Blanco O, Llanos-Cuentas A, García AL, et al. [Detection of *Leishmania* spp. based on the gene encoding HSP20]. Rev Peru Med Exp Salud Pública. 2014;31(4):635–43.

Morris SM, Albrecht U, Reiner O, Eichele G, Yu-Lee LY. The lissencephaly gene product Lis1, a protein involved in neuronal migration, interacts with a nuclear movement protein, NudC. Curr Biol CB. 1998;8(10):603–6.

Noël LD, Cagna G, Stuttmann J, Wirthmüller L, Betsuyaku S, Witte C-P, et al. Interaction between SGT1 and cytosolic/nuclear HSC70 chaperones regulates Arabidopsis immune responses. Plant Cell. 2007;19(12):4061–76.

Osmani AH, Osmani SA, Morris NR. The molecular cloning and identification of a gene product specifically required for nuclear movement in *Aspergillus nidulans*. J Cell Biol. 1990;111(2):543–51.

Pelé J, Bécu J-M, Abdi H, Chabbert M. Bios2mds: an R package for comparing orthologous protein families by metric multidimensional scaling. BMC Bioinformatics. 2012;13:133.

Perez DE, Hoyer JS, Johnson AI, Moody ZR, Lopez J, Kaplinsky NJ. BOBBER1 is a noncanonical Arabidopsis small heat shock protein required for both development and thermotolerance. Plant Physiol. 2009;151(1):241–52.

Pérez-Morales D, Espinoza B. The role of small heat shock proteins in parasites. Cell Stress Chaperones. 2015;20(5):767–80.

Perez-Morales D, Ostoa-Saloma P, Espinoza B. Trypanosoma cruzi SHSP16: Characterization of an alpha-crystallin small heat shock protein. Exp Parasitol. 2009;123(2):182–9.

Requena JM, Montalvo AM, Fraga J. Molecular Chaperones of Leishmania: Central Players in Many Stress-Related and -Unrelated Physiological Processes. BioMed Res Int. 2015;2015:301326.

Riera J, Lazo PS. The mammalian NudC-like genes: a family with functions other than regulating nuclear distribution. Cell Mol Life Sci CMLS. 2009;66(14):2383–90.

Riera J, Rodríguez R, Carcedo MT, Campa VM, Ramos S, Lazo PS. Isolation and characterization of nudC from mouse macrophages, a gene implicated in the inflammatory response through the regulation of PAF-AH(I) activity. FEBS Lett. 2007;581(16):3057–62.

Spiechowicz M, Filipek A. The expression and function of Sgt1 protein in eukaryotic cells. Acta Neurobiol Exp (Warsz). 2005;65(2):161–5.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinforma Oxf Engl. 2014;30(9):1312–3.

Stevens JR. Kinetoplastid phylogenetics, with special reference to the evolution of parasitic trypanosomes. Parasite Paris Fr. 2008;15(3):226–32.

Stevens JR, Gibson WC. The evolution of pathogenic trypanosomes. Cad Saude Publica 1999;15(4):673–84.

Stevens JR, Noyes HA, Schofield CJ, Gibson W. The molecular evolution of Trypanosomatidae. Adv Parasitol. 2001;48:1–56.

Tammimies K, Vitezic M, Matsson H, Le Guyader S, Bürglin TR, Ohman T, et al. Molecular networks of DYX1C1 gene show connection to neuronal migration genes and cytoskeletal proteins. Biol Psychiatry. 2013;73(6):583–90.

Urményi TP, Silva R, Rondinelli E. The heat shock proteins of *Trypanosoma cruzi*. Subcell Biochem. 2014;74:119–35.

Yan J, Zhang Y, Ding Y. Binding mechanism between Hsp90 and Sgt1 explored by homology modeling and molecular dynamics simulations in rice. J Mol Model. 2012;18(10):4665–73.

Yang XF, Wu CJ, McLaughlin S, Chillemi A, Wang KS, Canning C, et al. CML66, a broadly immunogenic tumor antigen, elicits a humoral immune response associated with remission of chronic myelogenous leukemia. Proc Natl Acad Sci U S A. 2001;98(13):7492–7.

Yang Y, Yan X, Cai Y, Lu Y, Si J, Zhou T. NudC-like protein 2 regulates the LIS1/dynein pathway by stabilizing LIS1 with Hsp90. Proc Natl Acad Sci U S A. 2010;107(8):3499–504.

Zabka M, Leśniak W, Prus W, Kuźnicki J, Filipek A. Sgt1 has co-chaperone properties and is up-regulated by heat shock. Biochem Biophys Res Commun. 2008;370(1):179–83.

Zhou T, Aumais JP, Liu X, Yu-Lee L-Y, Erikson RL. A role for Plk1 phosphorylation of NudC in cytokinesis. Dev Cell. 2003;5(1):127–38.

Zhou T, Zimmerman W, Liu X, Erikson RL. A mammalian NudC-like protein essential for dynein stability and cell viability. Proc Natl Acad Sci U S A. 2006;103(24):9039–44.

Zhu X-J, Liu X, Jin Q, Cai Y, Yang Y, Zhou T. The L279P mutation of nuclear distribution gene C (NudC) influences its chaperone activity and lissencephaly protein 1 (LIS1) stability. J Biol Chem. 2010;285(39):29903–10.

4.3 O Repertorio de genes de congopaína divergiu para se tornar subgrupo específico e um valioso marcador para o diagnostico e genotipagem de isolados dos grupos Savannah, Forest e Kilifi de *Trypanosoma congolense*

Trypanosoma congolense é o agente causal mais importante da nagana, ou tripanossomíase animal, responsável por grandes perdas nos rebanhos de gado da África subsaariana. Esta espécie é formada por um complexo de três subgrupos (Savannah, Forest e Kilifi) que diferem na virulência, resistência às drogas, vetores, e distribuição geográfica. As Congopainas representam as cisteína proteases (Catepsinas L) de *T. congolense*. Essas peptidases têm sido extensivamente investigadas como factor de virulência e alvo para drogas e vacinas. Entretanto, o conhecimento sobre esta enzima está principalmente restringido ao isolado de referência IL3000, o qual pertence ao o subgrupo Savannah.

Neste trabalho, os genes de Congopainas presentes no genoma do isolado de referência de *T. Congolense* (IL3000) são comparados com sequências obtidas de isolados dos três grupos de *T. congolense*. Os resultados demonstraram que esses genes divergiram em três subclados, consistentes com os três subgrupos dentro *T. congolense*. Isolados de laboratório e de campo, pertencentes ao do grupo Savannah, exibem um repertório altamente polimórfico de Congopainas, tanto dentro do mesmo genoma quanto entre isolados. As sequências que sequências que compartilham a tríade catalítica típica do arquetípico estão posicionadas nos grupos SAV1-SAV3, enquanto as sequências que apresentaram substituições incomuns na tríade catalítica (variantes) estão atribuídas a SAV4 ou não puderam ser posicionadas em nenhum dos grupos. Por outro lado, os genes homólogos de Congopaina obtidos de isolados dos grupos Forest e Kilifi apresentaram, respectivamente, diversidade moderada e restrita. Na árvore filogenética com base nas sequências de Congopaina, o grupo Savannah se posiciona mais relacionado à Forest do que à Kilifi. Todos os grupos de *T. congolense* se agrupam em um único clado monofilético, que, em conjunto o grupo irmão formado por de *Trypanosoma godfreyi* e *Trypanosoma simiae* formam o subgênero *Nannomonas*.

Um PCR único com genes de Congopain foi desenvolvido para o diagnóstico de dos três grupos de *T. congolense.* Nossos resultados demonstraram que os genes Congopaina são alvos valiosos para o diagnóstico, genotipagem, e inferências filogenéticas e taxonômicas entre isolados de *T. congolense* e outros membros do subgênero *Nannomonas*.

4.4 Dados filogenéticos e genômicos suportam uma única transferência horizontal do gene de Prolina Racemase procariótica, implicadas na evasão das defesas do hospedeiro pelo parasita, para o ancestral dos tripanossomas.

A enzima Prolina racemase (PRAC) de *Trypanosoma cruzi* (TcPRAC) e de *Trypanosoma vivax* (TvPRAC) têm sido implicadas a imunossupressão do hospedeiro e a evasão das defesas do hospedeiro pelo parasita através da ativação policlonal das células B. A semelhança com PRAC procariótica e a ausência dessa enzima em *Trypanosoma brucei* e *Trypanosoma congolense* levanta muitas perguntas sobre a origem, a evolução e as funções da PRAC (TryPRAC) em tripanossomas.

Neste trabalho, homólogos *Try*PRAC foram identificados como genes de cópia única por genoma haplóide em 12 das 15 espécies de *Trypanosoma*, incluindo *T. cruzi* e *T. cruzi marinkellei*, *T. dionisii*, *T. erneyi*, *T. rangeli*, *T. conorhini* e *T. lewisi*, todos os parasitas de mamíferos. O polimorfismo encontrado na TcPRAC coincide com a diversidade de genótipos de *T. cruzi.* Desse modo, as DTUs TCI a TCIV e Tcbat apresentam uma única cópia do gene, enquanto as linhagens híbridas TCV e TcVI possuem duas cópias cada (TcPRACA e TcPRACB) dos parentais TCII e TCIII, respectivamente. Além de *T. cruzi,* ortólogos da PRAC foram identificados em tripanossomas de anuros, cobras, crocodilos, lagartos e pássaros. A maioria dos tripanossomas apresenta o gene da PRAC intacto, porém os isolados de *T. rangeli* possuem apenas um pseudogene, possivelmente em processo de degeneração. As espécies do clado de tripanossomas africanos pesquisadas, exceto *T. vivax,* perderam completamente genes PRAC.

A inferência filogenética da *Try*PRAC suporta uma história evolutiva congruente com a filogenia do gênero *Trypanosoma*. A concordância entre as filogenias, junto à sintênia dos loci de *Try*PRAC, à proximidade da *Try*PRAC com a PRAC procariótica inferida pela análise filogenética abrangente e à ausência de ortólogos da PRAC em tripanossomatídeos de quaisquer outros gêneros e em bodonídeos e euglenídeos sugerem que um ancestral comum de Trypanosoma ganhou gene PRAC através de um evento de transferência horizontal de genes único e antigo, cuja origem remete a uma bactéria firmicute mais relacionado com *Gemella* e outras espécies de bacilos do que a *Clostridium*, como sugerido anteriormente. Nosso estudo filogenético abrangente permitiu a investigar a evolução dos genes de *Try*PRAC em escalas de tempo longas e curtas. Os genes *Try*PRAC divergiram junto às espécies de tripanossoma de modo a se tornarem espécie e genótipos específicos como visto em *T. cruzi* e *T. rangeli* e consequentemente, apresentando inferências congruentes com as obtidas usando genes verticalmente herdados. O inventário de genes de *Try*PRAC descrito aqui é o primeiro passo para a compreensão dos papéis da enzima PRAC em tripanossomas diferentes do ciclo de vida, virulência e com diferentes estratégias de infecção e de evasão do sistema imune.

4.5 Endossimbiose em tripanossomatídeos: A Cooperação genômica entre o simbionte bacteriano e hospedeiro na síntese dos aminoácidos essenciais é em grande parte influenciada por múltiplas transferências horizontais de genes.

As espécies de tripanossomatídeos dos gêneros *Angomonas* e *Strigomonas* possuem vivem em uma associação mutualista com endossimbiontes betaproteobacterianos. Essa associação é caracterizada por uma ampla cooperação metabólica entre o endossimbionte obrigatório e o hospedeiro. Embora o papel metabólico desempenhado pelo simbionte tenha sido mostrado de forma indireta por estudos bioquímicos, nenhuma evidencia genômica tinha sido demonstrada. Os tripanossomatídeos que abrigam endossimbiontes (SHT - *Symbiont-harboring trypanosomatids*) requerem menos nutrientes do ambiente, sendo autotrófico para os aminoácidos essenciais, o que não é observado nas demais espécies de tripanossomatídeos sem endossimbiontes. Neste trabalho, as vias de síntese de aminoácidos presentes nos genomas do endossimbionte e de tripanossomatídeos com e sem endossimbionte foram comparadas a fim de se evidenciar a cooperação genômica entre simbionte e hospedeiro.

As análises revelaram que a maior parte dos genes que codificam as enzimas das vias de biossíntese de aminoácidos essenciais está contida no genoma do endossimbionte, enquanto que o genoma dos tripanossomatídeos hospedeiros contém um número menor de genes. Nossas análises indicaram que cerca da metade dos genes do hospedeiro é proveniente transferência horizontal de genes de diferentes grupos de bactérias, sendo apenas um possivelmente transferido do endossimbionte (ornitina ciclodeaminase, EC: 4.3.1.12). Os dados de requerimento nutricionais, enzimáticos e genômicos foram analisados em conjunto para construir uma visão integrada do metabolismo dos aminoácidos essenciais presente nas espécies de tripanossomatídeos que contém endossimbionte. Esta abordagem abrangente mostrou concordância perfeita entre todos esses dados e revelou que o simbionte contém os genes que codificam para as enzimas que completam rotas biossintéticas essenciais para a produção de aminoácidos do hospedeiro, o que explica a baixa exigência nutricional dos SHT em relação aos demais tripanossomatídeos. As análises filogenéticas mostram que a cooperação entre simbiontes e seus hospedeiros é complementada por muitos eventos de transferência horizontal de genes de diversas linhagens bacterianas para o genoma nuclear dos tripanossomatídeos no curso de sua evolução.

Neste trabalho, revelamos as bases genéticas e evolutivas da biossíntese de aminoácidos essenciais em vários tripanossomatídeos com e sem endossimbiontes, explicando e complementando décadas de resultados experimentais. Nossos resultados mostram a plasticidade notável na evolução das vias de biossíntese de aminoácidos essenciais nesses protozoários e demonstram a grande influência de eventos de transferência horizontal de genes, de bactérias para o genoma nuclear dos tripanossomatídeos.

REFÊRENCIAS[1]

Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinforma Oxf Engl. 1 de maio de 2005;21(9):2104–5.

Adams ER, Hamilton PB, Gibson WC. African trypanosomes: celebrating diversity. Trends Parasitol. julho de 2010;26(7):324–8.

Adhikari BN, Hamilton JP, Zerillo MM, Tisserat N, Lévesque CA, Buell CR. Comparative genomics reveals insight into virulence strategies of plant pathogenic oomycetes. PloS One. 2013;8(10):e75072.

Akiba T, Koyama K, Ishiki Y, Kimura S, Fukushima T. On the mechanism of the development of multiple-drug-resistant clones of Shigella. Jpn J Microbiol. abril de 1960;4:219–27.

Alvarez VE, Niemirowicz GT, Cazzulo JJ. The peptidases of Trypanosoma cruzi: digestive enzymes, virulence factors, and mediators of autophagy and programmed cell death. Biochim Biophys Acta. janeiro de 2012;1824(1):195–206.

Alves JMP, Voegtly L, Matveyev AV, Lara AM, da Silva FM, Serrano MG, et al., Identification and phylogenetic analysis of heme synthesis genes in trypanosomatids and their bacterial endosymbionts. PloS One. 2011;6(8):e23518.

Atkinson HJ, Babbitt PC, Sajid M. The global cysteine peptidase landscape in parasites. Trends Parasitol. dezembro de 2009;25(12):573–81.

Bangs JD, Uyetake L, Brickman MJ, Balber AE, Boothroyd JC. Molecular cloning and cellular localization of a BiP homologue in Trypanosoma brucei. Divergent ER retention signals in a lower eukaryote. J Cell Sci. agosto de 1993;105 (Pt 4):1101–13.

Barry JD, Ginger ML, Burton P, McCulloch R. Why are parasite contingency genes often associated with telomeres? Int J Parasitol. janeiro de 2003;33(1):29–45.

Bente M, Harder S, Wiesgigl M, Heukeshoven J, Gelhaus C, Krause E, et al., Developmentally induced changes of the proteome in the protozoan parasite Leishmania donovani. Proteomics. setembro de 2003;3(9):1811–29.

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

Berriman M, Hall N, Sheader K, Bringaud F, Tiwari B, Isobe T, et al., The architecture of variant surface glycoprotein gene expression sites in Trypanosoma brucei. Mol Biochem Parasitol. julho de 2002;122(2):131–40.

Boorstein WR, Ziegelhoffer T, Craig EA. Molecular evolution of the HSP70 multigene family. J Mol Evol. janeiro de 1994;38(1):1–17.

Borghesan TC, Ferreira RC, Takata CSA, Campaner M, Borda CC, Paiva F, et al., Molecular phylogenetic redefinition of Herpetomonas (Kinetoplastea, Trypanosomatidae), a genus of insect parasites associated with flies. Protist. janeiro de 2013;164(1):129–52.

Borst P, Sabatini R. Base J: discovery, biosynthesis, and possible functions. Annu Rev Microbiol. 2008;62:235–51.

Breglia SA, Slamovits CH, Leander BS. Phylogeny of phagotrophic euglenids (Euglenozoa) as inferred from hsp90 gene sequences. J Eukaryot Microbiol. janeiro de 2007;54(1):86–92.

[1] De acordo com:
International Committee of Medical Journal Editors.
Uniform requirements for manuscripts submitted to
Biomedical Journal: sample references.
Available from: http://www.icmje.org
[updated 2011 Jul 15].

Bringaud F, Peyruchaud S, Baltz D, Giroud C, Simpson L, Baltz T. Molecular characterization of the mitochondrial heat shock protein 60 gene from Trypanosoma brucei. Mol Biochem Parasitol. outubro de 1995;74(1):119–23.

Bryan MA, Norris KA. Genetic immunization converts the trypanosoma cruzi B-Cell mitogen proline racemase to an effective immunogen. Infect Immun. fevereiro de 2010;78(2):810–22.

Burdon RH. Heat shock and the heat shock proteins. Biochem J. 1 de dezembro de 1986;240(2):313–24.

Buschiazzo A, Goytia M, Schaeffer F, Degrave W, Shepard W, Grégoire C, et al., Crystal structure, catalytic mechanism, and mitogenic properties of Trypanosoma cruzi proline racemase. Proc Natl Acad Sci U S A. 7 de fevereiro de 2006;103(6):1705–10.

Busse I, Preisfeld A. Unusually expanded SSU ribosomal DNA of primary osmotrophic euglenids: molecular evolution and phylogenetic inference. J Mol Evol. dezembro de 2002;55(6):757–67.

Camargo EP. Growth and Differentiation in Trypanosoma Cruzi. I. Origin of Metacyclic Trypanosomes in Liquid Media. Rev Inst Med Trop Sao Paulo. maio de 1964;6:93–100.

Campetella O, Henriksson J, Aslund L, Frasch AC, Pettersson U, Cazzulo JJ. The major cysteine proteinase (cruzipain) from Trypanosoma cruzi is encoded by multiple polymorphic tandemly organized genes located on different chromosomes. Mol Biochem Parasitol. fevereiro de 1992;50(2):225–34.

Campos RM, Nascimento M, Ferraz JC, Pereira MM, Rocha PO, Thompson GM, et al., Distinct mitochondrial HSP70 homologues conserved in various Leishmania species suggest novel biological functions. Mol Biochem Parasitol. agosto de 2008;160(2):157–62.

Capewell P, Cooper A, Clucas C, Weir W, Macleod A. A co-evolutionary arms race: trypanosomes shaping the human genome, humans shaping the trypanosome genome. Parasitology. fevereiro de 2015;142 Suppl 1:S108–19.

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. 9 de outubro de 2008;455(7214):757–63.

Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. abril de 2000;17(4):540–52.

Cavazzana M, Marcili A, Lima L, da Silva FM, Junqueira AC, Veludo HH, et al., Phylogeographical, ecological and biological patterns shown by nuclear (ssrRNA and gGAPDH) and mitochondrial (Cyt b) genes of trypanosomes of the subgenus Schizotrypanum parasitic in Brazilian bats. Int J Parasitol. 1 de março de 2010;40(3):345–55.

Cerqueira GC, Bartholomeu DC, DaRocha WD, Hou L, Freitas-Silva DM, Machado CR, et al., Sequence diversity and evolution of multigene families in Trypanosoma cruzi. Mol Biochem Parasitol. janeiro de 2008;157(1):65–72.

Chamond N, Cosson A, Coatnoan N, Minoprio P. Proline racemases are conserved mitogens: characterization of a Trypanosoma vivax proline racemase. Mol Biochem Parasitol. junho de 2009;165(2):170–9.

Chamond N, Goytia M, Coatnoan N, Barale J-C, Cosson A, Degrave WM, et al., Trypanosoma cruzi proline racemases are involved in parasite differentiation and infectivity. Mol Microbiol. outubro de 2005;58(1):46–60.

Chamond N, Grégoire C, Coatnoan N, Rougeot C, Freitas-Junior LH, da Silveira JF, et al., Biochemical characterization of proline racemases from the human protozoan parasite Trypanosoma cruzi and definition of putative protein signatures. J Biol Chem. 2 de maio de 2003;278(18):15484–94.

Coatnoan N, Berneman A, Chamond N, Minoprio P. Proline racemases: insights into Trypanosoma cruzi peptides containing D-proline. Mem Inst Oswaldo Cruz. julho de 2009;104 Suppl 1:295–300.

Conti P, Tamborini L, Pinto A, Blondel A, Minoprio P, Mozzarelli A, et al., Drug discovery targeting amino acid racemases. Chem Rev. 9 de novembro de 2011;111(11):6919–46.

Cooper ED. Horizontal gene transfer: accidental inheritance drives adaptation. Curr Biol CB. 16 de junho de 2014;24(12):R562–4.

Cornelis G, Vernochet C, Carradec Q, Souquere S, Mulot B, Catzeflis F, et al., Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. Proc Natl Acad Sci U S A. 3 de fevereiro de 2015;112(5):E487–96.

Cortez AP, Rodrigues AC, Garcia HA, Neves L, Batista JS, Bengaly Z, et al. Cathepsin L-like genes of Trypanosoma vivax from Africa and South America--characterization, relationships and diagnostic implications. Mol Cell Probes. fevereiro de 2009;23(1):44–51.

Corradi N, Slamovits CH. The intriguing nature of microsporidian genomes. Brief Funct Genomics. maio de 2011;10(3):115–24.

Coutinho L, Ferreira MA, Cosson A, Batista MM, Batista D da GJ, Minoprio P, et al., Inhibition of Trypanosoma cruzi proline racemase affects host-parasite interactions and the outcome of in vitro infection. Mem Inst Oswaldo Cruz. dezembro de 2009;104(8):1055–62.

D'Avila-Levy CM, Volotão ACC, Araújo FM, de Jesus JB, Motta MCM, Vermelho AB, et al., Bodo sp., a free-living flagellate, expresses divergent proteolytic activities from the closely related parasitic trypanosomatids. J Eukaryot Microbiol. outubro de 2009;56(5):454–8.

Dávila López M, Martínez Guerra JJ, Samuelsson T. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. PloS One. 2010;5(5):e10654.

De Greef C, Hamers R. The serum resistance-associated (SRA) gene of Trypanosoma brucei rhodesiense encodes a variant surface glycoprotein-like protein. Mol Biochem Parasitol. dezembro de 1994;68(2):277–84.

Deitsch KW, Moxon ER, Wellems TE. Shared themes of antigenic variation and virulence in bacterial, protozoal, and fungal infections. Microbiol Mol Biol Rev MMBR. setembro de 1997;61(3):281–93.

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

Dragon EA, Sias SR, Kato EA, Gabe JD. The genome of Trypanosoma cruzi contains a constitutively expressed, tandemly arranged multicopy gene homologous to a major heat shock protein. Mol Cell Biol. março de 1987;7(3):1271–5.

Dragovic Z, Broadley SA, Shomura Y, Bracher A, Hartl FU. Molecular chaperones of the Hsp110 family act as nucleotide exchange factors of Hsp70s. EMBO J. 7 de junho de 2006;25(11):2519–28.

Dvořáková N, Čepička I, Qablan MA, Gibson W, Blažek R, Široký P. Phylogeny and Morphological Variability of Trypanosomes from African Pelomedusid Turtles with Redescription of Trypanosoma mocambicum Pienaar, 1962. Protist. dezembro de 2015;166(6):599–608.

Eakin AE, Mills AA, Harth G, McKerrow JH, Craik CS. The sequence, organization, and expression of the major cysteine protease (cruzain) from Trypanosoma cruzi. J Biol Chem. 15 de abril de 1992;267(11):7411–20.

Easton DP, Kaneko Y, Subjeck JR. The hsp110 and Grp1 70 stress proteins: newly recognized relatives of the Hsp70s. Cell Stress Chaperones. outubro de 2000;5(4):276–90.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinforma Oxf Engl. 1 de outubro de 2010;26(19):2460–1.

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. 15 de julho de 2005a;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. 15 de julho de 2005b;309(5733):404–9.

Engman DM, Kirchhoff LV, Donelson JE. Molecular cloning of mtp70, a mitochondrial member of the hsp70 family. Mol Cell Biol. novembro de 1989;9(11):5163–8.

Estévez AM, Simpson L. Uridine insertion/deletion RNA editing in trypanosome mitochondria--a review. Gene. 29 de novembro de 1999;240(2):247–60.

Ettari R, Previti S, Tamborini L, Cullia G, Grasso S, Zappalà M. The Inhibition of Cysteine Proteases Rhodesain and TbCatB: a Valuable Approach to Treat Human African Trypanosomiasis. Mini Rev Med Chem. 9 de maio de 2016.

Feder ME, Hofmann GE. Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. Annu Rev Physiol. 1999;61:243–82.

Fermino BR, Paiva F, Soares P, Tavares LER, Viola LB, Ferreira RC, et al., Field and experimental evidence of a new caiman trypanosome species closely phylogenetically related to fish trypanosomes and transmitted by leeches. Int J Parasitol Parasites Wildl. dezembro de 2015;4(3):368–78.

Fermino BR, Viola LB, Paiva F, Garcia HA, de Paula CD, Botero-Arias R, et al., The phylogeography of trypanosomes from South American alligatorids and African crocodilids is consistent with the geological history of South American river basins and the transoceanic dispersal of Crocodylus at the Miocene. Parasit Vectors. 2013;6(1):313.

Ferreira RC, De Souza AA, Freitas RA, Campaner M, Takata CS, Barrett TV, et al., A phylogenetic lineage of closely related trypanosomes (Trypanosomatidae, Kinetoplastida) of anurans and sand flies (Psychodidae, Diptera) sharing the same ecotopes in brazilian amazonia. J Eukaryot Microbiol. setembro de 2008;55(5):427–35.

Figueiredo LM, Cross GAM, Janzen CJ. Epigenetic regulation in African trypanosomes: a new kid on the block. Nat Rev Microbiol. julho de 2009;7(7):504–13.

Fitzpatrick DA, Logue ME, Butler G. Evidence of recent interkingdom horizontal gene transfer between bacteria and Candida parapsilosis. BMC Evol Biol. 2008;8:181.

Flegontov P, Votýpka J, Skalický T, Logacheva MD, Penin AA, Tanifuji G, et al., Paratrypanosoma is a novel early-branching trypanosomatid. Curr Biol CB. 23 de setembro de 2013;23(18):1787–93.

Folgueira C, Requena JM. A postgenomic view of the heat shock proteins in kinetoplastids. FEMS Microbiol Rev. julho de 2007;31(4):359–77.

Fraga J, Montalvo AM, De Doncker S, Dujardin JC, Van der Auwera G. Phylogeny of Leishmania species based on the heat-shock protein 70 gene. Infect Genet Evol. março de 2010;10(2):238–45.

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 J Mol Epidemiol Evol Genet Infect Dis. agosto de 2013;18:229–37.

Fu X, Jiao W, Chang Z. Phylogenetic and biochemical studies reveal a potential evolutionary origin of small heat shock proteins of animals from bacterial class A. J Mol Evol. março de 2006;62(3):257–66.

Garcia HA, Kamyingkird K, Rodrigues AC, Jittapalapong S, Teixeira MM, Desquesnes M. High genetic diversity in field isolates of Trypanosoma theileri assessed by analysis of cathepsin L-like sequences disclosed multiple and new genotypes infecting cattle in Thailand. Vet Parasitol. 25 de agosto de 2011;180(3-4):363–7.

Garcia HA, Rodrigues AC, Martinkovic F, Minervino AHH, Campaner M, Nunes VLB, et al. Multilocus phylogeographical analysis of Trypanosoma (Megatrypanum) genotypes from sympatric cattle and water buffalo populations supports evolutionary host constraint and close phylogenetic relationships with genotypes found in other ruminants. Int J Parasitol. novembro de 2011b;41(13–14):1385–96.

Genevaux P, Georgopoulos C, Kelley WL. The Hsp70 chaperone machines of Escherichia coli: a paradigm for the repartition of chaperone functions. Mol Microbiol. novembro de 2007;66(4):840–57.

Germot A, Philippe H. Critical analysis of eukaryotic phylogeny: a case study based on the HSP70 family. J Eukaryot Microbiol. março de 1999;46(2):116–24.

Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, Ivens A, et al., Gene synteny and evolution of genome architecture in trypanosomatids. Mol Biochem Parasitol. abril de 2004;134(2):183–91.

Giambiagi-deMarval M, Souto-Padron T, Rondinelli E. Characterization and cellular distribution of heat-shock proteins HSP70 and HSP60 in Trypanosoma cruzi. Exp Parasitol. agosto de 1996;83(3):335–45.

Gibbs SP. The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae. Ann N Y Acad Sci. 1981;361:193–208.

Gilles HM. Protozoal diseases. London : New York: Arnold ; Oxford University Press; 1999.

Glass DJ, Polvere RI, Van der Ploeg LH. Conserved sequences and transcription of the hsp70 gene family in Trypanosoma brucei. Mol Cell Biol. dezembro de 1986;6(12):4657–66.

Goloubinoff P, De Los Rios P. The mechanism of Hsp70 chaperones: (entropic) pulling the models together. Trends Biochem Sci. agosto de 2007;32(8):372–80.

Graham SV, Terry S, Barry JD. A structural and transcription pattern for variant surface glycoprotein gene expression sites used in metacyclic stage Trypanosoma brucei. Mol Biochem Parasitol. 15 de outubro de 1999;103(2):141–54.

Guhl F, Vallejo GA. Trypanosoma (Herpetosoma) rangeli Tejera, 1920: an updated review. Mem Inst Oswaldo Cruz. junho de 2003;98(4):435–42.

Gupta RS, Golding GB. Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria, and eukaryotes. J Mol Evol. dezembro de 1993;37(6):573–82.

Gupta RS, Golding GB, Singh B. HSP70 phylogeny and the relationship between archaebacteria, eubacteria, and eukaryotes. J Mol Evol. novembro de 1994;39(5):537–40.

Hamilton PB, Adams ER, Malele I, Gibson WC. A novel, high-throughput technique for species identification reveals a new species of tsetse-transmitted trypanosome related to the Trypanosoma brucei subgenus, Trypanozoon. Infect Genet Evol. janeiro de 2008;8(1):26–33.

Hamilton PB, Adams ER, Njiokou F, Gibson WC, Cuny G, Herder S. Phylogenetic analysis reveals the presence of the Trypanosoma cruzi clade in African terrestrial mammals. Infect Genet Evol. janeiro de 2009;9(1):81–6.

Hamilton PB, Cruickshank C, Stevens JR, Teixeira MMG, Mathews F. Parasites reveal movement of bats between the New and Old Worlds. Mol Phylogenet Evol. maio de 2012a;63(2):521–6.

Hamilton PB, Gibson WC, Stevens JR. Patterns of co-evolution between trypanosomes and their hosts deduced from ribosomal RNA and protein-coding gene phylogenies. Mol Phylogenet Evol. julho de 2007;44(1):15–25.

Hamilton PB, Stevens JR, Gaunt MW, Gidley J, Gibson WC. Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA. Int J Parasitol. novembro de 2004;34(12):1393–404.

Hamilton PB, Stevens JR, Gidley J, Holz P, Gibson WC. A new lineage of trypanosomes from Australian vertebrates and terrestrial bloodsucking leeches (Haemadipsidae). Int J Parasitol. 1 de abril de 2005;35(4):431–43.

Hamilton PB, Teixeira MMG, Stevens JR. The evolution of Trypanosoma cruzi: the “bat seeding” hypothesis. Trends Parasitol. abril de 2012b;28(4):136–41.

Hannaert V, Opperdoes FR, Michels PA. Comparison and evolutionary analysis of the glycosomal glyceraldehyde-3-phosphate dehydrogenase from different Kinetoplastida. J Mol Evol. dezembro de 1998;47(6):728–38.

Hannaert V, Saavedra E, Duffieux F, Szikora J-P, Rigden DJ, Michels PAM, et al., Plant-like traits associated with metabolism of Trypanosoma parasites. Proc Natl Acad Sci U S A. 4 de fevereiro de 2003;100(3):1067–71.

Hartl FU, Hayer-Hartl M. Molecular chaperones in the cytosol: from nascent chain to folded protein. Science. 8 de março de 2002;295(5561):1852–8.

Haslbeck M, Franzmann T, Weinfurtner D, Buchner J. Some like it hot: the structure and function of small heat-shock proteins. Nat Struct Mol Biol. outubro de 2005;12(10):842–6.

Hertz-Fowler C, Figueiredo LM, Quail MA, Becker M, Jackson A, Bason N, et al., Telomeric expression sites are highly conserved in Trypanosoma brucei. PloS One. 2008;3(10):e3527.

von der Heyden S, Cavalier-Smith T. Culturing and environmental DNA sequencing uncover hidden kinetoplastid biodiversity and a major marine clade within ancestrally freshwater Neobodo designis. Int J Syst Evol Microbiol. novembro de 2005;55(Pt 6):2605–21.

Hirtreiter AM, Calloni G, Forner F, Scheibe B, Puype M, Vandekerckhove J, et al., Differential substrate specificity of group I and group II chaperonins in the archaeon Methanosarcina mazei. Mol Microbiol. dezembro de 2009;74(5):1152–68.

Hirt RP, Alsmark C, Embley TM. Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites. Curr Opin Microbiol. fevereiro de 2015;23:155–62.

Hoare CA. Morphological and taxonomic studies on mammalian trypanosomes. X. Revision of the systematics. J Protozool. maio de 1964;11:200–7.

Hoare CA. The trypanosomes of mammals: a zoological monograph. Oxford,: Blackwell Scientific Publications; 1972.

Horn D, McCulloch R. Molecular mechanisms underlying the control of antigenic variation in African trypanosomes. Curr Opin Microbiol. dezembro de 2010;13(6):700–5.

Hrdá Š, Fousek J, Szabová J, Hampl V, Hampl V, Vlček Č. The plastid genome of Eutreptiella provides a window into the process of secondary endosymbiosis of plastid in euglenids. PloS One. 2012;7(3):e33746.

Hubel A, Krobitsch S, Horauf A, Clos J. Leishmania major Hsp100 is required chiefly in the mammalian stage of the parasite. Mol Cell Biol. outubro de 1997;17(10):5987–95.

Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 1 de dezembro de 2012;61(6):1061–7.

Isobe T, Holmes EC, Rudenko G. The transferrin receptor genes of Trypanosoma equiperdum are less diverse in their transferrin binding site than those of the broad-host range Trypanosoma brucei. J Mol Evol. abril de 2003;56(4):377–86.

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

Jackson AP. Tandem gene arrays in Trypanosoma brucei: comparative phylogenomic analysis of duplicate sequence variation. BMC Evol Biol. 2007;7:54.

Jackson AP. Genome evolution in trypanosomatid parasites. Parasitology. fevereiro de 2015a;142 Suppl 1:S40–56.

Jackson AP. Preface. The evolution of parasite genomes and the origins of parasitism. Parasitology. fevereiro de 2015b;142 Suppl 1:S1–5.

Jackson AP, Allison HC, Barry JD, Field MC, Hertz-Fowler C, Berriman M. A cell-surface phylome for African trypanosomes. PLoS Negl Trop Dis. 2013;7(3):e2121.

Jackson AP, Berry A, Aslett M, Allison HC, Burton P, Vavrova-Anderson J, et al., Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. Proc Natl Acad Sci U S A. 28 de fevereiro de 2012;109(9):3416–21.

Jackson AP, Otto TD, Aslett M, Armstrong SD, Bringaud F, Schlacht A, et al., Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. Curr Biol CB. 25 de janeiro de 2016;26(2):161–72.

Jefferson T, McShan D, Warfield J, Ogungbe IV. Screening and Identification of Inhibitors of Trypanosoma brucei Cathepsin L with Antitrypanosomal Activity. Chem Biol Drug Des. janeiro de 2016;87(1):154–8.

Jensen AT, Curtis J, Montgomery J, Handman E, Theander TG. Molecular and immunological characterisation of the glucose regulated protein 78 of Leishmania donovani(1). Biochim Biophys Acta. 10 de setembro de 2001;1549(1):73–87.

Kabani M. Structural and functional diversity among eukaryotic Hsp70 nucleotide exchange factors. Protein Pept Lett.

2009;16(6):623–60.

Kabani M, Martineau CN. Multiple hsp70 isoforms in the eukaryotic cytosol: mere redundancy or functional specificity? Curr Genomics. 2008;9(5):338–248.

Karnkowska A, Bennett MS, Watza D, Kim JI, Zakryś B, Triemer RE. Phylogenetic Relationships and Morphological Character Evolution of Photosynthetic Euglenids (Excavata) Inferred from Taxon-rich Analyses of Five Genes. J Eukaryot Microbiol. junho de 2015;62(3):362–73.

Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. agosto de 2008;9(8):605–18.

Kelly S, Ivens A, Manna PT, Gibson W, Field MC. A draft genome for the African crocodilian trypanosome Trypanosoma grayi. Sci Data. 2014;1:140024.

Kent ML, Elston RA, Nerad TA, Sawyer TK. An Isonema-like flagellate (Protozoa: Mastigophora) infection in larval geoduck clams, Panope abrupta. J Invertebr Pathol. novembro de 1987;50(3):221–9.

Klein CC, Alves JMP, Serrano MG, Buck GA, Vasconcelos ATR, Sagot M-F, et al., Biosynthesis of vitamins and cofactors in bacterium-harbouring trypanosomatids depends on the symbiotic association as revealed by genomic analyses. PloS One. 2013;8(11):e79786.

Korený L, Lukes J, Oborník M. Evolution of the haem synthetic pathway in kinetoplastid flagellates: an essential pathway that is not essential after all? Int J Parasitol. fevereiro de 2010;40(2):149–56.

Kostygov AY, Grybchuk-Ieremenko A, Malysheva MN, Frolov AO, Yurchenko V. Molecular revision of the genus Wallaceina. Protist. setembro de 2014;165(5):594–604.

Krauth-Siegel RL, Comini MA. Redox control in trypanosomatids, parasitic protozoa with trypanothione-based thiol metabolism. Biochim Biophys Acta. novembro de 2008;1780(11):1236–48.

Krobitsch S, Clos J. A novel role for 100 kD heat shock proteins in the parasite Leishmania donovani. Cell Stress Chaperones. setembro de 1999;4(3):191–8.

Lara E, Moreira D, Vereshchaka A, López-García P. Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. Environ Microbiol. janeiro de 2009;11(1):47–55.

Larreta R, Soto M, Alonso C, Requena JM. Leishmania infantum: gene cloning of the GRP94 homologue, its expression as recombinant protein, and analysis of antigenicity. Exp Parasitol. outubro de 2000;96(2):108–15.

Lee MG, Atkinson BL, Giannini SH, Van der Ploeg LH. Structure and expression of the hsp 70 gene family of Leishmania major. Nucleic Acids Res. 25 de outubro de 1988;16(20):9567–85.

Lee MG, Polvere RI, Van der Ploeg LH. Evidence for segmental gene conversion between a cognate hsp 70 gene and the temperature-sensitively transcribed hsp70 genes of Trypanosoma brucei. Mol Biochem Parasitol. junho de 1990;41(2):213–20.

Leonard G, Soanes DM, Stevens JR. Resolving the question of trypanosome monophyly: a comparative genomics approach using whole genome data sets with low taxon sampling. Infect Genet Evol. julho de 2011;11(5):955–9.

Lima L, Espinosa-Álvarez O, Hamilton PB, Neves L, Takata CSA, Campaner M, et al., Trypanosoma livingstonei: a new species from African bats supports the bat seeding hypothesis for the Trypanosoma cruzi clade. Parasit Vectors. 2013;6(1):221.

Lima L, Espinosa-Álvarez O, Pinto CM, Cavazzana M, Pavan AC, Carranza JC, et al., New insights into the evolution of the Trypanosoma cruzi clade provided by a new trypanosome species tightly linked to Neotropical Pteronotus bats and related to an Australian lineage of trypanosomes. Parasit Vectors. 2015;8:657.

Lima L, Ortiz PA, da Silva FM, Alves JMP, Serrano MG, Cortez AP, et al., Repertoire, genealogy and genomic organization of cruzipain and homologous genes in Trypanosoma cruzi, T. cruzi-like and other trypanosome species. PloS One. 2012a;7(6):e38385.

Lima L, Silva FM da, Neves L, Attias M, Takata CSA, Campaner M, et al., Evolutionary insights from bat trypanosomes: morphological, developmental and phylogenetic evidence of a new species, Trypanosoma (Schizotrypanum) erneyi sp. nov., in African bats closely related to Trypanosoma (Schizotrypanum) cruzi and allied species. Protist. novembro de 2012b;163(6):856–72.

Lindquist S, Craig EA. The heat-shock proteins. Annu Rev Genet. 1988;22:631–77.

Linton EW, Hittner D, Lewandowski C, Auld T, Triemer RE. A molecular study of euglenoid phylogeny using small subunit rDNA. J Eukaryot Microbiol. abril de 1999;46(2):217–23.

Lira CBB, Giardini MA, Neto JLS, Conte FF, Cano MIN. Telomere biology of trypanosomatids: beginning to answer some questions. Trends Parasitol. agosto de 2007;23(8):357–62.

López-García P, Philippe H, Gail F, Moreira D. Autochthonous eukaryotic diversity in hydrothermal sediment and experimental microcolonizers at the Mid-Atlantic Ridge. Proc Natl Acad Sci U S A. 21 de janeiro de 2003;100(2):697–702.

Lukeš J, Kuchta R, Scholz T, Pomajbíková K. (Self-) infections with parasites: re-interpretations for the present. Trends Parasitol. agosto de 2014;30(8):377–85.

Lun Z-R, Lai D-H, Li F-J, Lukes J, Ayala FJ. Trypanosoma brucei: two steps to spread out from Africa. Trends Parasitol. setembro de 2010;26(9):424–7.

Lv Z, Wu Z, Zhang L, Ji P, Cai Y, Luo S, et al., Genome mining offers a new starting point for parasitology research. Parasitol Res. fevereiro de 2015;114(2):399–409.

Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 10 de novembro de 2000;290(5494):1151–5.

Maia Da Silva F, Junqueira AC, Campaner M, Rodrigues AC, Crisante G, Ramirez LE, et al., Comparative phylogeography of Trypanosoma rangeli and Rhodnius (Hemiptera: Reduviidae) supports a long coexistence of parasite lineages and their sympatric vectors. Mol Ecol. agosto de 2007;16(16):3361–73.

Maia da Silva F, Marcili A, Ortiz PA, Epiphanio S, Campaner M, Catao-Dias JL, et al., Phylogenetic, morphological and behavioural analyses support host switching of Trypanosoma (Herpetosoma) lewisi from domestic rats to primates. Infect Genet Evol. maio de 2010;10(4):522–9.

Marche S, Roth C, Manohar SK, Dollet M, Baltz T. RNA virus-like particles in pathogenic plant trypanosomatids. Mol Biochem Parasitol. fevereiro de 1993;57(2):261–7.

Marcili A, Lima L, Cavazzana M, Junqueira AC, Veludo HH, Maia Da Silva F, et al., A new genotype of Trypanosoma cruzi associated with bats evidenced by phylogenetic analyses using SSU rDNA, cytochrome b and Histone H2B genes and genotyping based on ITS1 rDNA. Parasitology. maio de 2009;136(6):641–55.

Maresca B, Carratu L. The biology of the heat shock response in parasites. Parasitol Today. agosto de 1992;8(8):260–6.

Marr JJ, Berens RL, Nelson DJ. Purine metabolism in Leishmania donovani and Leishmania braziliensis. Biochim Biophys Acta. 1 de dezembro de 1978;544(2):360–71.

Maslov DA, Lukes J, Jirku M, Simpson L. Phylogeny of trypanosomes as inferred from the small and large subunit rRNAs: implications for the evolution of parasitism in the trypanosomatid protozoa. Mol Biochem Parasitol. janeiro de 1996;75(2):197–205.

Maslov DA, Votýpka J, Yurchenko V, Lukeš J. Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. Trends Parasitol. janeiro de 2013;29(1):43–52.

Merzlyak E, Yurchenko V, Kolesnikov AA, Alexandrov K, Podlipaev SA, Maslov DA. Diversity and phylogeny of insect trypanosomatids based on small subunit rRNA genes: polyphyly of Leptomonas and Blastocrithidia. J Eukaryot Microbiol. abril de 2001;48(2):161–9.

de Miguel N, Braun N, Bepperling A, Kriehuber T, Kastenmuller A, Buchner J, et al., Structural and functional diversity in the family of small heat shock proteins from the parasite Toxoplasma gondii. Biochim Biophys Acta. novembro de

2009;1793(11):1738–48.

Molyneux DH. Virus-like particles in Leishmania parasites. Nature. 7 de junho de 1974;249(457):588–9.

Molyneux DH. Trypanosomes of bats. Parasit Protozoa. Lodon: Academic Press; 1991. p. 195–223.

Montalvo-Alvarez AM, Folgueira C, Carrion J, Monzote-Fidalgo L, Canavate C, Requena JM. The Leishmania HSP20 is antigenic during natural infections, but, as DNA vaccine, it does not protect BALB/c mice against experimental L. amazonensis infection. J Biomed Biotechnol. 2008;2008:695432.

Montalvo AM, Fraga J, Rodríguez O, Blanco O, Llanos-Cuentas A, García AL, et al., [Detection of Leishmania spp. based on the gene encoding HSP20]. Rev Peru Med Exp Salud Pública. dezembro de 2014;31(4):635–43.

Moreira D, von der Heyden S, Bass D, López-García P, Chao E, Cavalier-Smith T. Global eukaryote phylogeny: Combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata. Mol Phylogenet Evol. julho de 2007;44(1):255–66.

Moreira D, Lopez-Garcia P, Rodriguez-Valera F. New insights into the phylogenetic position of diplonemids: G+C content bias, differences of evolutionary rate and a new environmental sequence. Int J Syst Evol Microbiol. novembro de 2001;51(Pt 6):2211–9.

Moreira D, Lopez-Garcia P, Vickerman K. An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the class Kinetoplastea. Int J Syst Evol Microbiol. setembro de 2004;54(Pt 5):1861–75.

Morrison LJ, MacLeod A. African trypanosomiasis. Parasite Immunol. agosto de 2011;33(8):421–2.

Motta MCM, de Souza W, Thiry M. Immunocytochemical detection of DNA and RNA in endosymbiont-bearing trypanosomatids. FEMS Microbiol Lett. 11 de abril de 2003;221(1):17–23.

Mottram JC, Murphy WJ, Agabian N. A transcriptional analysis of the Trypanosoma brucei hsp83 gene cluster. Mol Biochem Parasitol. novembro de 1989;37(1):115–27.

Myler PJ, Audleman L, deVos T, Hixson G, Kiser P, Lemley C, et al., Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes. Proc Natl Acad Sci U S A. 16 de março de 1999;96(6):2902–6.

Naderer T, McConville MJ. Intracellular growth and pathogenesis of Leishmania parasites. Essays Biochem. 24 de outubro de 2011;51:81–95.

Nakayima J, Nakao R, Alhassan A, Hayashida K, Namangala B, Mahama C, et al. Genetic diversity among Trypanosoma (Duttonella) vivax strains from Zambia and Ghana, based on cathepsin L-like gene. Parasite Paris Fr. 2013;20:24.

Nisbet B. Nutrition and feeding strategies in protozoa. London : Croom Helm; 1984.

Oliveira P, Lima FM, Cruz MC, Ferreira RC, Sanchez-Flores A, Cordero EM, et al., Trypanosoma cruzi: Genome characterization of phosphatidylinositol kinase gene family (PIK and PIK-related) and identification of a novel PIK gene. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. julho de 2014;25:157–65.

Opperdoes FR, Michels PAM. Horizontal gene transfer in trypanosomatids. Trends Parasitol. outubro de 2007;23(10):470–6.

Ortiz PA, Maia da Silva F, Cortez AP, Lima L, Campaner M, Pral EMF, et al., Genes of cathepsin L-like proteases in Trypanosoma rangeli isolates: markers for diagnosis, genotyping and phylogenetic relationships. Acta Trop. dezembro de 2009;112(3):249–59.

Ouellette M, Drummelsmith J, El-Fadili A, Kündig C, Richard D, Roy G. Pterin transport and metabolism in Leishmania and related trypanosomatid parasites. Int J Parasitol. abril de 2002;32(4):385–98.

Paba J, Santana JM, Teixeira ARL, Fontes W, Sousa MV, Ricart CAO. Proteomic analysis of the human pathogen Trypanosoma cruzi. Proteomics. abril de 2004;4(4):1052–9.

Paparini A, Macgregor J, Irwin PJ, Warren K, Ryan UM. Novel genotypes of Trypanosoma binneyi from wild platypuses

(Ornithorhynchus anatinus) and identification of a leech as a potential vector. Exp Parasitol. outubro de 2014;145:42–50.

Pérez-Morales D, Espinoza B. The role of small heat shock proteins in parasites. Cell Stress Chaperones. setembro de 2015;20(5):767–80.

Perez-Morales D, Ostoa-Saloma P, Espinoza B. Trypanosoma cruzi SHSP16: Characterization of an alpha-crystallin small heat shock protein. Exp Parasitol. outubro de 2009;123(2):182–9.

Pratt WB, Toft DO. Regulation of signaling protein function and trafficking by the hsp90/hsp70-based chaperone machinery. Exp Biol Med Maywood NJ. fevereiro de 2003;228(2):111–33.

Preisfeld A, Busse I, Klingberg M, Talke S, Ruppel HG. Phylogenetic position and inter-relationships of the osmotrophic euglenids based on SSU rDNA data, with emphasis on the Rhabdomonadales (Euglenozoa). Int J Syst Evol Microbiol. maio de 2001;51(Pt 3):751–8.

Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PloS One. 2010;5(3):e9490.

Pusnik M, Charrière F, Mäser P, Waller RF, Dagley MJ, Lithgow T, et al., The single mitochondrial porin of Trypanosoma brucei is the main metabolite transporter in the outer mitochondrial membrane. Mol Biol Evol. março de 2009;26(3):671–80.

Raina P, Kaur S. Knockdown of LdMC1 and Hsp70 by antisense oligonucleotides causes cell-cycle defects and programmed cell death in Leishmania donovani. Mol Cell Biochem [Internet]. 31 de julho de 2011; Recuperado de: http://www.ncbi.nlm.nih.gov/pubmed/21805355

Reina-San-Martín B, Degrave W, Rougeot C, Cosson A, Chamond N, Cordeiro-Da-Silva A, et al., A B-cell mitogen from a pathogenic trypanosome is a eukaryotic proline racemase. Nat Med. agosto de 2000;6(8):890–7.

Requena JM, Lopez MC, Jimenez-Ruiz A, de la Torre JC, Alonso C. A head-to-tail tandem organization of hsp70 genes in Trypanosoma cruzi. Nucleic Acids Res. 25 de fevereiro de 1988;16(4):1393–406.

Richter K, Buchner J. Hsp90: chaperoning signal transduction. J Cell Physiol. setembro de 2001;188(3):281–90.

Rodrigues AC, Garcia HA, Ortiz PA, Cortez AP, Martinkovic F, Paiva F, et al., Cysteine proteases of Trypanosoma (Megatrypanum) theileri: cathepsin L-like gene sequences as targets for phylogenetic analysis, genotyping diagnosis. Parasitol Int. setembro de 2010;59(3):318–25.

Rodrigues AC, Paiva F, Campaner M, Stevens JR, Noyes HA, Teixeira MM. Phylogeny of Trypanosoma (Megatrypanum) theileri and related trypanosomes reveals lineages of isolates associated with artiodactyl hosts diverging on SSU and ITS ribosomal sequences. Parasitology. fevereiro de 2006;132(Pt 2):215–24.

Rodrigues CMF, Batista JS, Lima JM, Freitas FJC, Barros IO, Garcia HA, et al. Field and experimental symptomless infections support wandering donkeys as healthy carriers of Trypanosoma vivax in the Brazilian Semiarid, a region of outbreaks of high mortality in cattle and sheep. Parasit Vectors. 2015;8:564.

Rudenko G. Epigenetics and transcriptional control in African trypanosomes. Essays Biochem. 20 de setembro de 2010;48(1):201–19.

Ruszczyk A, Forlenza M, Savelkoul HFJ, Wiegertjes GF. Molecular cloning and functional characterisation of a cathepsin L-like proteinase from the fish kinetoplastid parasite Trypanosoma carassii. Fish Shellfish Immunol. fevereiro de 2008;24(2):205–14.

Sajid M, McKerrow JH. Cysteine proteases of parasitic organisms. Mol Biochem Parasitol. março de 2002;120(1):1–21.

Sakanari JA, Nadler SA, Chan VJ, Engel JC, Leptak C, Bouvier J. Leishmania major: comparison of the cathepsin L- and B-like cysteine protease genes with those of other trypanosomatids. Exp Parasitol. janeiro de 1997;85(1):63–76.

Schell D, Borowy NK, Overath P. Transferrin is a growth factor for the bloodstream form of Trypanosoma brucei. Parasitol Res. 1991;77(7):558–60.

Schwarz RS, Bauchan GR, Murphy CA, Ravoet J, de Graaf DC, Evans JD. Characterization of Two Species of

Trypanosomatidae from the Honey Bee Apis mellifera: Crithidia mellificae Langridge and McGhee, and Lotmaria passim n. gen., n. sp. J Eukaryot Microbiol. outubro de 2015;62(5):567–83.

Shapira M, Pedraza G. Sequence analysis and transcriptional activation of heat shock protein 83 of Leishmania mexicana amazonensis. Mol Biochem Parasitol. outubro de 1990;42(2):247–55.

Sharma D, Masison DC. Hsp70 structure, function, regulation and influence on yeast prions. Protein Pept Lett. 2009;16(6):571–81.

Shorter J, Lindquist S. Hsp104, Hsp70 and Hsp40 interplay regulates formation, growth and elimination of Sup35 prions. EMBO J. 22 de outubro de 2008;27(20):2712–24.

Silva DCF, Silva RC, Ferreira RC, Briones MRS. Examining marginal sequence similarities between bacterial type III secretion system components and Trypanosoma cruzi surface proteins: horizontal gene transfer or convergent evolution? Front Genet. 2013;4:143.

Simpson AG, Lukes J, Roger AJ. The evolutionary history of kinetoplastids and their kinetoplasts. Mol Biol Evol. dezembro de 2002;19(12):2071–83.

Simpson AG, Roger AJ. Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. Mol Phylogenet Evol. janeiro de 2004;30(1):201–12.

Simpson AG, Stevens JR, Lukes J. The evolution and diversity of kinetoplastid flagellates. Trends Parasitol. abril de 2006;22(4):168–74.

Sogin ML, Elwood HJ, Gunderson JH. Evolutionary diversity of eukaryotic small-subunit rRNA genes. Proc Natl Acad Sci U S A. março de 1986;83(5):1383–7.

Stadtman TC, Elliott P. Studies on the enzymic reduction of amino acids. II. Purification and properties of D-proline reductase and a proline racemase from Clostridium sticklandii. J Biol Chem. outubro de 1957;228(2):983–97.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinforma Oxf Engl. 1 de maio de 2014;30(9):1312–3.

Stevens JR. Kinetoplastid phylogenetics, with special reference to the evolution of parasitic trypanosomes. Parasite Paris Fr. setembro de 2008;15(3):226–32.

Stevens JR, Gibson W. The molecular evolution of trypanosomes. Parasitol Today. novembro de 1999;15(11):432–7.

Stevens JR, Noyes HA, Dover GA, Gibson WC. The ancient and divergent origins of the human pathogenic trypanosomes, Trypanosoma brucei and T. cruzi. Parasitology. janeiro de 1999;118 (Pt 1):107–16.

Stevens JR, Noyes HA, Schofield CJ, Gibson W. The molecular evolution of Trypanosomatidae. Adv Parasitol. 2001;48:1–56.

Steverding D. The transferrin receptor of Trypanosoma brucei. Parasitol Int. janeiro de 2000;48(3):191–8.

Steverding D. The significance of transferrin receptor variation in Trypanosoma brucei. Trends Parasitol. março de 2003;19(3):125–7.

Stoco PH, Wagner G, Talavera-Lopez C, Gerber A, Zaha A, Thompson CE, et al., Genome of the avirulent human-infective trypanosome--Trypanosoma rangeli. PLoS Negl Trop Dis. setembro de 2014;8(9):e3176.

Stöver BC, Müller KF. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics. 2010;11:7.

Sturm NR, Maslov DA, Grisard EC, Campbell DA. Diplonema spp. possess spliced leader RNA genes similar to the Kinetoplastida. J Eukaryot Microbiol. junho de 2001;48(3):325–31.

Sulaiman IM, Morgan UM, Thompson RC, Lal AA, Xiao L. Phylogenetic relationships of Cryptosporidium parasites based on the 70-kilodalton heat shock protein (HSP70) gene. Appl Env Microbiol. junho de 2000;66(6):2385–91.

Svobodova M, Zidkova L, Cepicka I, Obornik M, Lukes J, Votypka J. Sergeia podlipaevi gen. nov., sp. nov. (Trypanosomatidae, Kinetoplastida), a parasite of biting midges (Ceratopogonidae, Diptera). Int J Syst Evol Microbiol. fevereiro de 2007;57(Pt 2):423–32.

Teixeira MM, Borghesan TC, Ferreira RC, Santos MA, Takata CS, Campaner M, et al., Phylogenetic validation of the genera Angomonas and Strigomonas of trypanosomatids harboring bacterial endosymbionts with the description of new species of trypanosomatids and of proteobacterial symbionts. Protist. julho de 2011;162(3):503–24.

Teixeira MMG. Trypanosomatidae diversidade, filogênia e taxonomia. 2010.

Thompson MD, Zhang L, Hong L, Hallick RB. Two new group-II twintrons in the Euglena gracilis chloroplast are absent in basally branching Euglena species. Curr Genet. janeiro de 1997;31(1):89–95.

Thompson RJ, Bouwer HG, Portnoy DA, Frankel FR. Pathogenicity and immunogenicity of a Listeria monocytogenes strain that requires D-alanine for growth. Infect Immun. agosto de 1998;66(8):3552–61.

Tibbetts RS, Jensen JL, Olson CL, Wang FD, Engman DM. The DnaJ family of protein chaperones in Trypanosoma cruzi. Mol Biochem Parasitol. 15 de março de 1998;91(2):319–26.

Tibbetts RS, Kim IY, Olson CL, Barthel LM, Sullivan MA, Winquist AG, et al., Molecular cloning and characterization of the 78-kilodalton glucose-regulated protein of Trypanosoma cruzi. Infect Immun. junho de 1994;62(6):2499–507.

Truc P, Büscher P, Cuny G, Gonzatti MI, Jannin J, Joshi P, et al., Atypical human infections by animal trypanosomes. PLoS Negl Trop Dis. 2013;7(9):e2256.

Truc P, Nzoumbou-Boko R, Desquesnes M, Semballa S, Vincendeau P. [Atypical human trypanosomoses]. Médecine Santé Trop. setembro de 2014;24(3):249–52.

Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL, et al., The genomes of four tapeworm species reveal adaptations to parasitism. Nature. 4 de abril de 2013;496(7443):57–63.

Turmel M, Gagnon M-C, O'Kelly CJ, Otis C, Lemieux C. The chloroplast genomes of the green algae Pyramimonas, Monomastix, and Pycnococcus shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. Mol Biol Evol. março de 2009;26(3):631–48.

Tutar L, Tutar Y. Heat shock proteins; an overview. Curr Pharm Biotechnol. fevereiro de 2010;11(2):216–22.

Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RHY, Aerts A, et al., Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science. 1 de setembro de 2006;313(5791):1261–6.

Vickerman K. The evolutionary expansion of the trypanosomatid flagellates. Int J Parasitol. dezembro de 1994;24(8):1317–31.

Viola LB, Attias M, Takata CS, Campaner M, De Souza W, Camargo EP, et al., Phylogenetic analyses based on small subunit rRNA and glycosomal glyceraldehyde-3-phosphate dehydrogenase genes and ultrastructural characterization of two snake Trypanosomes: Trypanosoma serpentis n. sp. from Pseudoboa nigra and Trypanosoma cascavelli from Crotalus durissus terrificus. J Eukaryot Microbiol. novembro de 2009;56(6):594–602.

Visser WF, Verhoeven-Duif NM, de Koning TJ. Identification of a human trans-3-hydroxy-L-proline dehydratase, the first characterized member of a novel family of proline racemase-like enzymes. J Biol Chem. 22 de junho de 2012;287(26):21654–62.

Votýpka J, Kostygov AY, Kraeva N, Grybchuk-Ieremenko A, Tesařová M, Grybchuk D, et al., Kentomonas gen. n., a new genus of endosymbiont-containing trypanosomatids of Strigomonadinae subfam. n. Protist. dezembro de 2014;165(6):825–38.

Votýpka J, Oborník M, Volf P, Svobodová M, Lukes J. Trypanosoma avium of raptors (Falconiformes): phylogeny and identification of vectors. Parasitology. setembro de 2002;125(Pt 3):253–63.

Votýpka J, Suková E, Kraeva N, Ishemgulova A, Duží I, Lukeš J, et al., Diversity of trypanosomatids (Kinetoplastea:

Trypanosomatidae) parasitizing fleas (Insecta: Siphonaptera) and description of a new genus Blechomonas gen. n. Protist. novembro de 2013;164(6):763–81.

Walsh P, Bursać D, Law YC, Cyr D, Lithgow T. The J-protein family: modulating protein assembly, disassembly and translocation. EMBO Rep. junho de 2004;5(6):567–71.

Whitesell L, Lindquist SL. HSP90 and the chaperoning of cancer. Nat Rev Cancer. outubro de 2005;5(10):761–72.

Wiesgigl M, Clos J. Heat shock protein 90 homeostasis controls stage differentiation in Leishmania donovani. Mol Biol Cell. novembro de 2001;12(11):3307–16.

Wijayawardena BK, Minchella DJ, DeWoody JA. Hosts, parasites, and horizontal gene transfer. Trends Parasitol. julho de 2013;29(7):329–38.

Witola WH, Sarataphan N, Inoue N, Ohashi K, Onuma M. Genetic variability in ESAG6 genes among Trypanosoma evansi isolates and in comparison to other Trypanozoon members. Acta Trop. janeiro de 2005;93(1):63–73.

Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, et al., The genome of Cryptosporidium hominis. Nature. 28 de outubro de 2004;431(7012):1107–12.

Yamasaki M, Tajima M, Lee KW, Jeong JR, Yamato O, Maede Y. Molecular cloning and phylogenetic analysis of Babesia gibsoni heat shock protein 70. Vet Parasitol. 11 de dezembro de 2002;110(1-2):123–9.

Yoshimura T, Esak N. Amino acid racemases: functions and mechanisms. J Biosci Bioeng. 2003;96(2):103–9.

Young R, Taylor JE, Kurioka A, Becker M, Louis EJ, Rudenko G. Isolation and analysis of the genetic diversity of repertoires of VSG expression site containing telomeres from Trypanosoma brucei gambiense, T. b. brucei and T. equiperdum. BMC Genomics. 2008;9:385.

Yubuki N, Edgcomb VP, Bernhard JM, Leander BS. Ultrastructure and molecular phylogeny of Calkinsia aureus: cellular identity of a novel clade of deep-sea euglenozoans with epibiotic bacteria. BMC Microbiol. 2009;9:16.

Zhang J. Evolution by gene duplication: an update. Trends Ecol Evol. 1 de junho de 2003;18(6):292–8.

Zylicz M, Wawrzynow A. Insights into the function of Hsp70 chaperones. IUBMB Life. maio de 2001;51(5):283–7.

Infection, Genetics and Evolution 23 (2014) 20–31

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Congopain genes diverged to become specific to Savannah, Forest and Kilifi subgroups of *Trypanosoma congolense*, and are valuable for diagnosis, genotyping and phylogenetic inferences

Adriana C. Rodrigues [a,1], Paola A. Ortiz [a,1], André G. Costa-Martins [a], Luis Neves [b,c], Herakles A. Garcia [a,d], João M.P. Alves [a], Erney P. Camargo [a], Silvia C. Alfieri [a], Wendy Gibson [e], Marta M.G. Teixeira [a,*]

[a] Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, SP, Brazil
[b] Biotechnology Centre, Eduardo Mondlane University, Maputo, Mozambique
[c] Faculty of Veterinary Science, University of Pretoria, South Africa
[d] Department of Veterinary Pathology, Faculty of Veterinary Sciences, Central University of Venezuela, Maracay, Venezuela
[e] School of Biological Sciences, University of Bristol, Bristol BS8 1UG, United Kingdom

ARTICLE INFO

Article history:
Received 3 September 2013
Received in revised form 10 January 2014
Accepted 11 January 2014
Available online 27 January 2014

Keywords:
Trypanosoma congolense
Congopain
Savannah
Forest
Kilifi
Genotyping

ABSTRACT

Trypanosoma congolense is the most important agent of nagana, a wasting livestock trypanosomosis in sub-Saharan Africa. This species is a complex of three subgroups (Savannah, Forest and Kilifi) that differ in virulence, pathogenicity, drug resistance, vectors, and geographical distribution. Congopain, the major Cathepsin L-like cysteine protease (CP2) of *T. congolense*, has been extensively investigated as a pathogenic factor and target for drugs and vaccines, but knowledge about this enzyme is mostly restricted to the reference strain IL3000, which belongs to the Savannah subgroup. In this work we compared sequences of congopain genes from IL3000 genome database and isolates of the three subgroups of *T. congolense*. Results demonstrated that the congopain genes diverged into three subclades consistent with the three subgroups within *T. congolense*. Laboratory and field isolates of Savannah exhibited a highly polymorphic repertoire both inter- and intra-isolates: sequences sharing the archetypical catalytic triad clustered into SAV1–SAV3 groups, whereas polymorphic sequences that, in general, exhibited unusual catalytic triad (variants) assigned to SAV4 or not assigned to any group. Congopain homologous genes from Forest and Kilifi isolates showed, respectively, moderate and limited diversity. In the phylogenetic tree based on congopain and homologues, Savannah was closer to Forest than to Kilifi. All *T. congolense* subgroup nested into a single clade, which together with the sister clade formed by homologues from *Trypanosoma simiae* and *Trypanosoma godfreyi* formed a clade supporting the subgenus *Nannomonas*. A single PCR targeting congopain sequences was developed for the diagnosis of *T. congolense* isolates of the three subgroups. Our findings demonstrated that congopain genes are valuable targets for the diagnosis, genotyping, and phylogenetic and taxonomic inferences among *T. congolense* isolates and other members of the subgenus *Nannomonas*.

1. Introduction

Livestock trypanosomosis (nagana) is a chronic wasting disease that poses a major constraint to livestock productivity in sub-Saharan Africa. The causative agents are tsetse-borne trypanosomes, of which *Trypanosoma congolense* is the most prevalent and widespread. *T. congolense* comprises three morphologically indistinguishable but genetically recognisable subgroups—Savannah, Forest, and Kilifi (Gibson, 2002, 2007)—which vary in virulence, pathogenicity, and geographical distribution. The Savannah and Forest subgroups were originally evidenced by isoenzymes (Young and Godfrey, 1983; Gashumba et al., 1988), whereas RFLP and karyotyping disclosed a further subgroup: Kenya Coast or Kilifi (Knowles et al., 1988; Majiwa et al., 1985, 1986). Methods based on repetitive DNA sequences were developed to identify these three subgroups (Gibson et al., 1988; Masiga et al., 1992), all further corroborated by other molecular markers (Gibson et al., 2001; Desquesnes et al., 2001; Gibson, 2007, 2011; Adams et al., 2010).

The Savannah subgroup of *T. congolense* is the most widespread. Field investigations associated the Savannah subgroup with a

* Corresponding author. Tel.: +55 1130917429; fax: +55 1130917417.
E-mail address: mmgteix@icb.usp.br (M.M.G. Teixeira).
[1] Both authors contributed equally to this work.

http://dx.doi.org/10.1016/j.meegid.2014.01.012
1567-1348/

range of tsetse (*Glossina*) species (morsitans, palpalis and fusca groups) and a broad range of ungulates and carnivore hosts across the whole of sub-Saharan Africa. In contrast, *T. congolense* Forest appears to be largely restricted to the palpalis group of tsetse flies and, consequently, to riverine-forest biomes. It has been recorded in pigs, goats, cattle, and dogs in West and Central Africa, and also at low prevalence in parts of East Africa. *T. congolense* Kilifi was first isolated from Kenya but has since been widely reported throughout south-eastern Africa; it is associated with tsetse of the morsitans group and has been reported in cattle, sheep, and goats (Majiwa et al., 1993, Majiwa et al., 1985; Reifenberg et al., 1997; Knowles et al., 1988; Masiga et al., 1996; Njiru et al., 2004; Malele et al., 2011; Simo et al., 2012;Simo et al., 2013). Infections with a mixture of subgroups are frequent in ungulates and tsetse flies: co-infections with Savannah and Forest subgroups are common in West and Central Africa (Seck et al., 2010; Simo et al., 2012,2013), whereas Savannah and Kilifi subgroups mixed infections occur in East and South Africa (Mekata et al., 2008; Mamabolo et al., 2009). The three subgroups coexist in Zambia, Kenya, and Tanzania (Njiru et al., 2004; Mekata et al., 2008; Malele et al., 2011). Infection of susceptible zebuine cattle revealed Kilifi as non-pathogenic, Forest of low pathogenicity, and Savannah as the most virulent subgroup (Bengaly et al., 2002a,b). Isolates of Savannah differed markedly in virulence and drug resistance, even in the same location (Seck et al., 2010; Van den Bossche et al., 2011; Vitouley et al., 2011; Moti et al., 2012).

Phylogenetic analyses based on SSU rRNA and gGAPDH genes showed that the three subgroups of *T. congolense* clustered together, forming a clade within a monophyletic assemblage corresponding to the subgenus *Nannomonas* that also includes *Trypanosoma simiae* and *Trypanosoma godfreyi* (Hamilton et al., 2004). However, these genes were unable to resolve the relationships among the subgroups of *T. congolense*. Previous studies have suggested a closer relationship between Savannah and Forest than between these subgroups and Kilifi. *T. congolense* Savannah and Forest share lengths of the major satellite DNA repeat, kDNA minicircles, and mini-exon gene repeats and also polymorphisms in the beta tubulin and rRNA genes, whereas these markers are significantly different in Kilifi (Garside and Gibson, 1995). Sequence analysis of the gene coding a major surface glycoprotein, glutamate- and alanine-rich protein (GARP), demonstrated a similar relationship among the subgroups of *T. congolense* (Asbeck et al., 2000).

The Cathepsin L (CATL)-like cysteine proteases (CPs) have been extensively studied in trypanosomes due to their important roles in pathogenicity, virulence, cell differentiation, and immune evasion. These CPs belong to the papain family (clan CA, family C1) that typically consists of a signal peptide, pro-peptide, catalytic domain (cd), and a C-terminal extension of variable size unique to kinetoplastid CATL (Sajid and Mckerrow, 2002; Atkinson et al., 2009; Alvarez et al., 2012). The two main CATL-like CPs (CP1 and CP2) characterized in *T. congolense* can be distinguished by polymorphisms in the cds, which result in functional differences. CP2, usually referred to as congopain, is the major CP of *T. congolense* (Fish et al., 1995; Jaye et al., 1993; Authié et al., 1992, 1994, 2001; Boulangé et al., 2001). Congopain is an important antigen in the development of vaccines and target for chemotherapy (Authié et al., 1992, 1994, 2001; Boulangé et al., 2001; Huson et al., 2009; Katerega et al., 2012; Lalmanach et al., 2002). In addition, an unusual CP identified in *T. congolense* differs from CP1 and CP2 by a serine replacing cysteine in the catalytic triad (Downey and Donelson, 1999; Pillay et al., 2010); in addition, other variants have been reported among the CP enzymes of this species (Kakundi, 2008).

Our studies of other trypanosome species demonstrated that CATL-like genes are useful markers for diagnosis, genotyping, and phylogenetic reconstruction at species and genotype levels in *Trypanosoma vivax*, *Trypanosoma theileri*, *Trypanosoma rangeli*, and *Trypanosoma cruzi* and allied species (Cortez et al., 2009; Garcia et al., 2011a,b; Lima et al., 2012; Ortiz et al., 2009; Rodrigues et al., 2010). While congopain have been very well characterised in the laboratory strain IL3000 of *T. congolense* Savannah, there is an absence of data on CPs from Forest and Kilifi subgroups. In this study, we characterised the catalytic domains of genes encoding congopain from *T. congolense* isolates of Savannah, Forest and Kilifi subgroups, including samples from west, central, and east Africa. Our main goals were: (a) to compare the congopain genetic repertoires of the three subgroups; (b) to infer the genealogy and to evaluate the suitability of congopain sequences for diagnosis, genotyping, and population structure analysis; (c) to infer the phylogenetic relationships among isolates from the three subgroups of *T. congolense*, *T. simiae* and *T. godfreyi* (subgenus *Nannomonas*) based on congopain and homologous genes.

2. Materials and methods

2.1. Trypanosomes and PCR amplification of the catalytic domains of congopain genes

T. congolense clone IL3000 represents the Savannah subgroup and was selected for the genome project. This clone, obtained in 1966, has been maintained for decades by successive passages in mice (Gibson, 2012). Additional *T. congolense* isolates characterised in this study were from different hosts and geographic origins (Table 1). DNA templates were prepared from *T. congolense* Savannah (WG81, Gam2, IL1180, TREU1457), Forest (ANR3, Cam22), and Kilifi (WG5, WG84) laboratory stocks. Field-collected blood samples were obtained from infected cattle, water buffalo, and goats from Mozambique collected from endemic settlings in the provinces of Maputo (isolates Ma.ca01, Ma.ca06, Ma.bu03, Ma.bu05, Ma.bu04), Sofala (So.go01) and Tete (isolates Te.ca09, Te.ca016, Te.ca018) in the Southern, Central, and Northern regions, respectively (Table 2, Fig. 1). CATL homologous sequences from *T. simiae* (Ken14) and *T. godfreyi* (Ken7) were determined in this study, and included in the phylogenetic analysis of *Trypanosoma*. The DNA from *T. congolense* isolates, *T. simiae* (Ken14) and *T. godfreyi* (Ken7) were used for the PCR-amplification of partial sequences (477 bp) corresponding to the cds of CP (Fig. 2) using primers and conditions previously described for the amplification of CATL-like encoding genes from several trypanosome species (Lima et al., 1994, 2012; Cortez et al., 2009; Garcia et al., 2011a,b; Rodrigues et al., 2010; Ortiz et al., 2009). From 6 to 10 cloned cdCP sequences were determined for each isolate included in this study.

2.2. Genealogies of CATL-like nucleotide and deduced amino acid sequences

The genealogies of sequences encoding CATL-like enzymes were inferred by analyses of either nucleotide or deduced amino acid sequence data sets. The present study benefited from congopain sequences retrieved from the on-going genome project of *T. congolense* IL3000 (http://www.genedb.org). In addition to sequences determined in this study and those retrieved from the genome project, we included in the alignments: (a) the prototype sequences available from Genbank of CP1 (Z25813), CP2 (L25130), and the variant CP2-like from the strain TRUM 183 (AF139913); (b) CATL-like genes determined in this study from *Trypanosoma simiae* and *Trypanosoma godfreyi*; (c) CATL-like genes from *Trypanosoma b. brucei*, *Trypanosoma b. rhodesiense*, *Trypanosoma b. gambiense*, *Trypanosoma evansi*, *Trypanosoma equiperdum*, *Trypanosoma vivax*, *Trypanosoma theileri*, *Trypanosoma rangeli*, and *Trypanosoma carassii*, all from previous studies (Table 2).

Table 1
Geographic and host origin of *T. congolense* isolates, CATL-like sequences of *T. congolense*, and homologous sequences from other trypanosome species.

Trypanosoma	Subgroup	Host origin	Geographic origin	GenBank and GeneDB accession number	CATL-like sequence
T. congolense laboratory isolates					
Cam22	Forest	Goat	Cameroon	KF414001–KF414024	Forest
ANR3	Forest	Tsetse	The Gambia	KF414025–KF414036	Forest
WG5	Kilifi	Goat	Kenya	KF413898–KF413922	Kilifi
WG84	Kilifi	Sheep	Kenya	KF413923–KF413933	Kilifi
WG81	Savannah	–	Kenya	KF413934–KF413940	SAV1, SAV2, SAV3
Gam2	Savannah	Cow	The Gambia	KF413941–KF413948	SAV2, SAV3
IL1180	Savannah	Lion	Tanzânia	KF413977–KF413983	SAV1, SAV2
TREU1457	Savannah	Cow	Nigeria	KF413949–KF413951	SAV1, SAV2
TRUM 183[a]	Savannah	–	–	AF139913	SAV1
CP2 (archetype)	Savannah	–	–	L25130	SAV1
IL3000 (CP1 archetype)	Savannah	Cow	Kenya	Z25813	CP1
IL3000[b]	Savannah	Cow	Kenya	TcIL3000: 0.28270, 0.53250, 0.12120, 0.11840, 0.44210, 0.17860, 0.25670	CP1 SAV1
				TcIL3000.0.26770	SAV2
				TcIL3000.0.31720	SAV3
				TcIL3000.0.60020	SAV4
				TcIL3000: 0.26780, 0.49190, 0.47610, 0.18880	SAVna[c]
				TcIL3000: 0.28390, 0.55820, 0.55830, 0.55780, 0.37240, 0.31730	
Field isolates					
Ma.ca01	Savannah	Cattle	Mozambique Maputo	KF413965–KF413976	SAV1, SAV3
Ma.ca06	Savannah	Cattle	Mozambique Maputo	KF413984–KF413986	SAV1
Ma.bu03	Savannah	Buffalo	Mozambique Maputo	KF414051–KF414053	Savannah[d]
Ma.bu04	Savannah	Buffalo	Mozambique Maputo	KF414048–KF414048	Savannah[d]
Ma.bu05	Savannah	Buffalo	Mozambique Maputo	KF413987–KF413994	SAV1, SAV2, SAV3
Te.ca09	Savannah	Cow	Mozambique Tete	KF413957–KF413960	SAV1
Te.ca016	Savannah	Cow	Mozambique Tete	KF413952–KF413956	SAV1
Te.ca018	Savannah	Cow	Mozambique Tete	KF413961–KF413964	SAV1, SAV3
So.go01	Savannah	Goat	Mozambique Sofala	KF413995–KF414000, KF414054	SAV1, SAV4
MzGlo92	Savannah	Tsetse	Mozambique Sofala	KF414041–KF414042	SAV1, SAV3
MzGlo93	Savannah	Tsetse	Mozambique Sofala	KF414043–KF414047	Savannah[d]
Other species					
T. vivax	Y486	Cow	Nigeria	Tviv534d01.q1k7, Tviv290f05.q1k11[e]	
T. vivax	TviMzNy	Nyala	Mozambique	EU753814	
T. b. brucei	427	Sheep	Uganda	EU753820	
T. b. brucei	Star	–	–	X16465	
T. b. gambiense	TB26	Pig	Congo	EU753821	
T. b. rhodesiense	AntTat1.12	Human	–	EU753822	
T. equiperdum	Botat1.	Horse	–	EU753819	
T. evansi	Ted2	Dog	Brazil	EU753818	
T. simiae	Ken14	Tsetse	The Gambia	KF414037–KF414038	
T. godfreyi	Ken7	Tsetse	The Gambia	KF414039–KF414040	
T. rangeli	(AM80)	Human	Brazil	FJ997560	
T. theileri	Tthc12	Cow	Brazil	GU299366	
T. carassi	–	Fish	–	EF538803	

[a] CP sequence from *T. congolense* TRUM 183 retrieved from Genebank.
[b] *T. congolense* sequence data obtained from the Sanger Institute website at http://www.sanger.ac.uk/Projects/T_congolense/.
[c] SAVna, na = sequence not assigned to SAV1–4 groups.
[d] Diagnosed as *T. congolense* and genotyped into the Savannah subgroup by the method of TcoCATL-PCR followed by sequencing of the amplified fragments; the accession numbers of sequences determined in this work and deposited at the Genbank are KF413898–414047.
[e] *T. vivax* sequence data obtained from the Sanger Institute website at http://www.sanger.ac.uk/Projects?T_vivax.

The alignments created for this study included: (a) amino acid sequences comprising the pre-, pro-, and catalytic domains (cd) (Fig. 2) of 12 genes encoding CP2, and 7 encoding CP1, all from the IL3000 genome; (b) 72 cd amino acid sequences (477 bp) from Savannah (5 isolates), including sequences from IL3000 (13), livestock blood samples and one from tsetse fly from Mozambique (CP1 sequences were excluded from this alignment); (c) amino acid sequences of cdCP2 from 7 isolates of Savannah, Forest and Kilifi subgroups, and homologous genes from *T. simiae*, *T. godfreyi*, and other species of *Trypanosoma*; (d) sequences of 211 bp fragments from cdCP genes obtained by TcoCATL-PCR (see below). Genealogies were inferred with the neighbour-joining (NJ) algorithm in Mega 5 software, treating gaps as deletions, and maximum-likelihood (ML) analysis was carried out using RAxML v.7.0 as described in our previous studies (Garcia et al., 2011b; Lima et al., 2012; Rodrigues et al., 2010). Phylogenetic trees were constructed using the Neighbor-Net method with Kimura 2 parameters implemented in Splits Tree4 V4.10 (Huson and Bryant, 2006). Internode support was estimated by performing 100 bootstrap replicates using the same parameters optimised for tree inferences. To provide a visual representation of the distance matrix, we used the multidimensional scaling (MDS) plot with two dimensions (2D). The MDS statistical analysis and graphing were performed using the Bios2mds package of the R language and

Table 2
Biogeographical characteristics and congopain repertoires of *T. congolense* isolates from Savannah, Forest and Kilifi subgropus.

T. congolense subgroup	Distribution	Host range	Tsetse transmission	Pathogenicity[a]	Drug susceptibility[b]	CP2 repertoire
Savannah	Tropical Africa	Numerous species of ungulates and other mammals	Morsitans, palpalis and fusca groups	High	Susceptible resistant	SAV1–SAV4, SAVna[c]
Forest	West and Central Africa	Pigs, goats, cattle, dogs	Palpalis group	Low	Susceptible	Forest CP2
Kilifi	Southeast Africa	Cattle, sheep, goats	Morsitans group	Non-pathogenic	Not-determined	Kilifi CP2

[a] Differential pathogenicity among *T. congolense* subgroups, and variable virulence within Savannah, revealed by studies in mice and cattle (Bengaly et al., 2002a,b; Masumu et al., 2006).
[b] Resistance to diminazene has been reported for Savannah in field infected cattle and mice (Van den Bossche et al., 2011; Moti et al., 2012).
[c] SAVna, na = sequence not assigned to SAV1–4 groups.

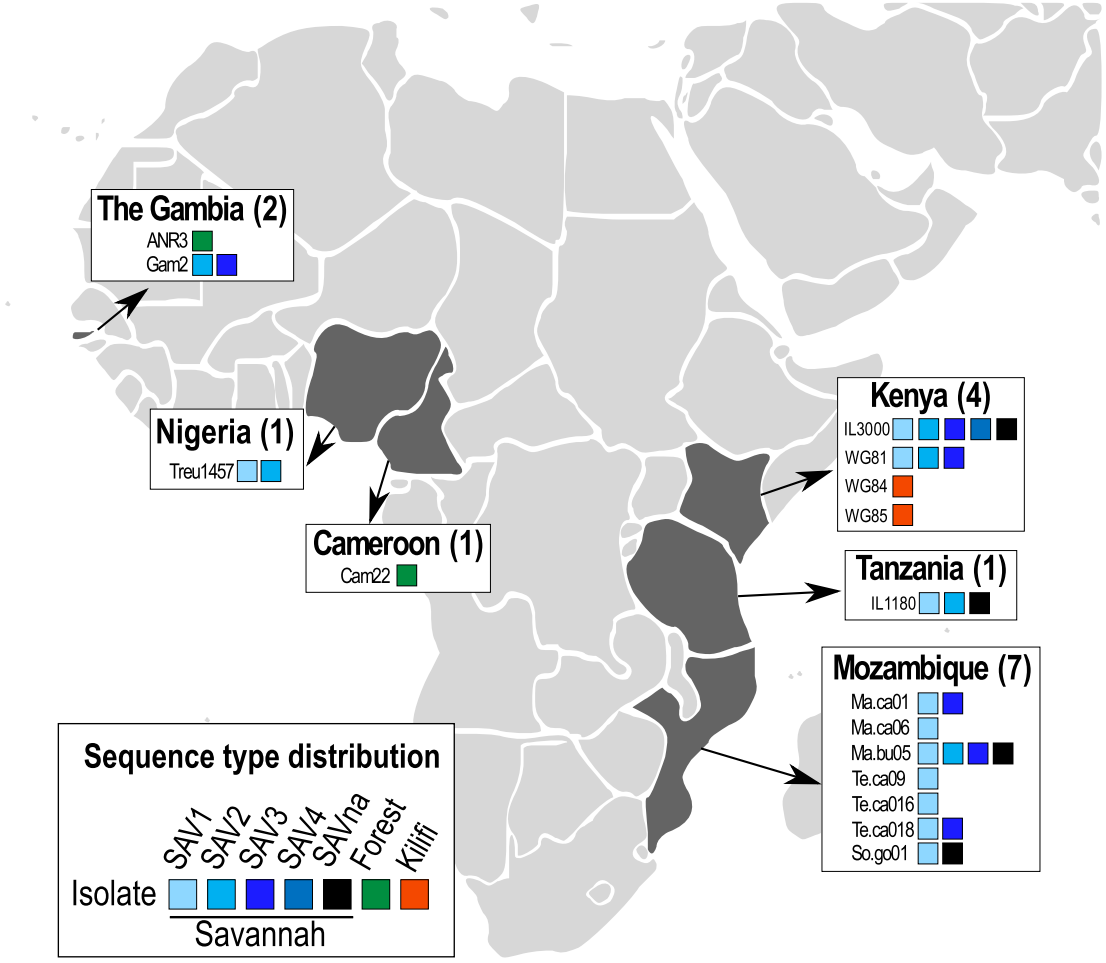

Fig. 1. Geographical origin of the *T. congolense* Savannah, Forest, and Kilifi subgroups characterised in this study and the respective genetic repertoire of CP2 sequences. Seven groups of CATL-like genes (indicated by different colours) were defined within *T. congolense* by the genealogy of the catalytic domain sequences and specific amino acid signatures: SAV1–SAV4 plus SAVna (=sequences not assigned to any group) for Savannah, and only one group of sequences each for the Forest and Kilifi subgroups. The number of isolates from each country is indicated within parentheses.

environment for statistical computing (Pelé et al., 2012). To find conserved motifs, "Multiple EM for Motif Elicitation" (MEME) version 3.5.4 (Bailey et al., 2009) was used. The parameters used for the analysis were number of repetitions – any, maximum number of motifs – 50, and optimum width of motif $\geqslant 3$ and $\leqslant 5$.

2.3. Codon usage and recombination analyses

The ratio of non-synonymous to synonymous (dN/dS) amino acid changes was calculated according to Yang and Nielsen (2000) using PAML v.4.2 software to infer relative selection pressures (Yang, 2007). A positive value for this test indicates an over-abundance of nonsynonymous substitutions, and in this case, the probability of rejecting the null hypothesis of neutral evolution (*P*-value) is calculated. The existence of putative recombination events in the genes encoding congopains was investigated using the RDP3 programme (Martin et al., 2010). All eight methods available were employed, and recombination events were considered valid if detected by at least four methods, with a minimum significance *P*-value of 0.05.

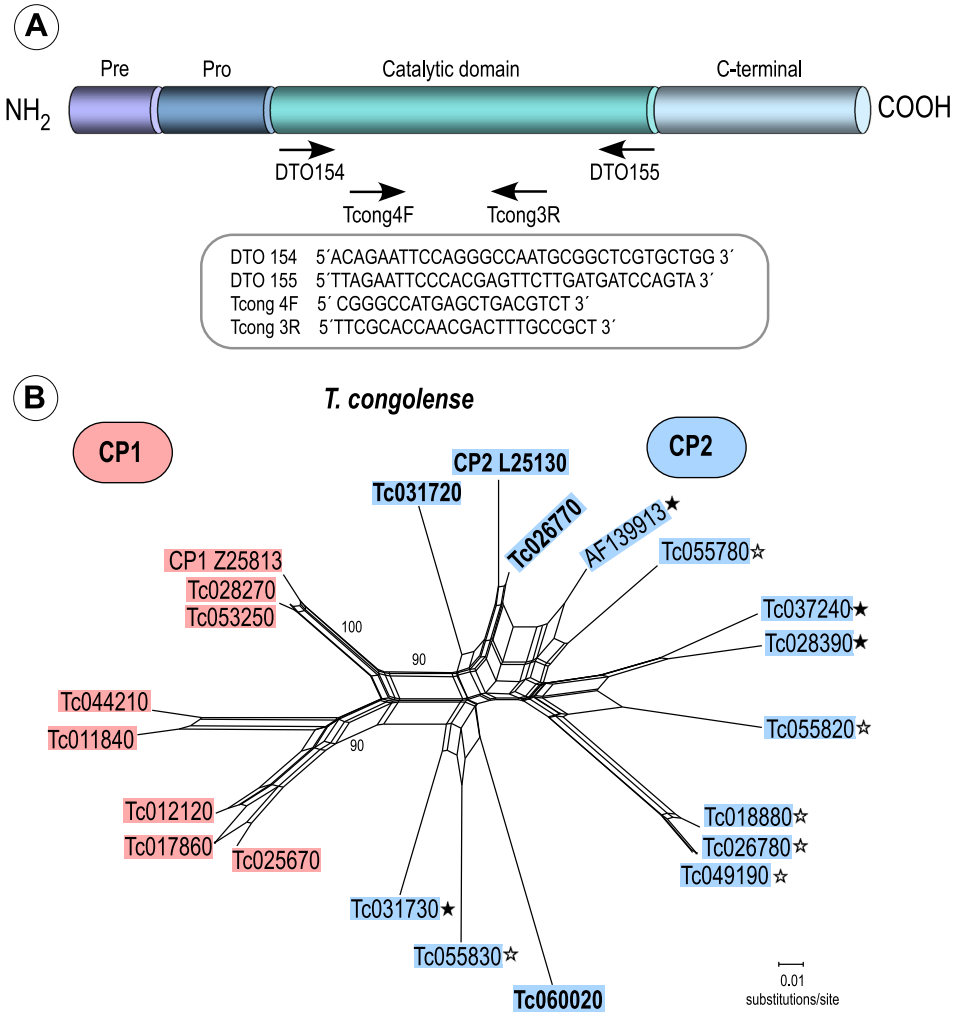

Fig. 2. (A) Schematic representation of the congopain gene indicating the protein domains and the primers employed for PCR amplification of the catalytic domains (primers DTO 154 and DTO 155) and the *T. congolense*-specific fragment of 211 bp (primers Tcong4F and Tcong 3R). (B) Network genealogy of CP predicted amino acid sequences (pre-, pro-, and catalytic domain) from the genome database of *T. congolense* IL3000 and from GenBank, inferred using the Neighbour-Net method with the K2P parameter and 1000 bootstrap replicates. The archetypical CP1 and CP2 (in bold) and variant sequences: the CP2-like and new CP2-like sequences are indicated by filled and unfilled stars, respectively.

2.4. Standardisation of PCR targeting CATL-like sequences for the diagnosis of T. congolense

An alignment including nucleotide sequences of cdCATL genes from *T. congolense* Savannah, Forest, Kilifi and homologues from closely related *T. simiae* and *T. godfreyi* and other trypanosome species was used to design the *T. congolense*-specific primers Tco4F and Tco3R (Fig. 2A). A PCR assay, designated TcoCATL-PCR, was developed for the amplification of a 211 bp DNA sequence specific for *T. congolense* using the following PCR conditions: 35 cycles of 94 °C (1 min), 63 °C (1 min) and 72 °C (1 min), with a final extension of 10 min at 72 °C.

To assess the species-specificity, TcoCATL-PCR was tested using DNA from *T. congolense* isolates of all subgroups (Table 1) and *T. vivax* from Brazil (TviBrMi, TviBrCa), West (Y486), and East (TviMzNy) Africa; *T. b. brucei* (427, 8195, AnTat1.1), *T. b. gambiense* group 2(TB26), *T. b. rhodesiense* (AnTat 1.12), *T. equiperdum* (BoTat1.1), *T. evansi* from Brazil (Ted1, Tec2,Teh1) and Africa (TeET); and *T. theileri* from cattle (TthATCC, Tthc3, Tthc17) and water buffalo (Tthb4, Tthb6). *T. simiae* (Ban2, Ken14) and *T. godfreyi* (Ken7) were also tested. DNA from field collected blood samples preserved in filter paper or in ethanol was obtained as previously described (Rodrigues et al., 2010; Garcia et al., 2011a), and tested using the method of TcoCATL-PCR. PCR-amplified DNA fragments were separated in 2% agarose gels, and stained with ethidium bromide.

3. Results

3.1. Diversity of CATL-like genes in T. congolense IL3000 genome database

To investigate the repertoire of all potential CATL-like genes encoded by *T. congolense*, we performed a BLAST search for proteins with high sequence similarity to the archetypes of CP2 (GenBank accession number L25130) and CP1 (GenBank Z25813) in the *T. congolense* IL3000 genome database available from http://www.genedb.org. All sequences sharing high similarity with CP2 (14) or CP1 (8 sequences) and containing the catalytic domain were downloaded, aligned and employed for phylogenetic analysis (Table 1, Fig. 1). Our analysis of the catalytic triads disclosed: (a) three sequences exhibiting the archetypical catalytic triad (CHN) of CP2; (b) four sequences showing SYN or SHN triad that, unexpectedly, encoded congopain-like enzymes active against classical CP substrates but differing slightly from one another and also from CP2 in substrate preferences (Pillay et al., 2010); (c) 7 sequences

showing SSN and PHN triads, designated herein as "new congopain-like" because although sequences with a serine or tyrosine replacing histidine have been reported (Kakundi, 2008; Pillay et al., 2010), the enzymatic activities of these variants have not investigated to date.

The amino acid sequences from the cds of CATL-like genes retrieved from *T. congolense* IL3000 genome database showed relevant polymorphisms (~18%) between the archetypical and variant CP sequences. The network genealogy of the 14 cd sequences found in the IL3000 genome confirmed and improved the divergent repertoire previously demonstrated by analysing CP sequences from a cosmid library of this strain, or PCR-amplified using degenerate primers (Kakundi, 2008; Pillay et al., 2010). Unfortunately, the sequences from CPs determined by Kakundi (2008), even those encoding congopain-like enzymes characterized by Pillay et al. (2010), are not available from public databases.

3.2. Repertoire and genealogy of the catalytic domains of CATL-like sequences in T. congolense isolates of Savannah, Forest, and Kilifi subgroups

We determined 100 sequences (477 bp) of the CATL-like catalytic domains from isolates of Savannah (WG81, Gam2, IL1180 and TREU1457), Forest (Cam22 and ANR3), and Kilifi (WG5 and WG84) (Fig. 1A). We also determined 44 sequences from blood samples of naturally infected cattle (5), water buffalo (1) and goats (1), plus one sample from a tsetse fly, all from Mozambique. Congopain analyses included several cloned sequences from each isolate, ranging from 10 sequences, for most Savannah laboratory isolates, to ≈30 for some isolates from Forest and Kilifi subgroups. Excluding the identical sequences, we generated alignments containing 70 different nucleotide sequences or 57 different predicted amino acid sequences (36 from Savannah, 9 from Kilifi, and 12 from Forest isolates), and the IL3000 genome sequences. The genealogy branching pattern not only segregated the sequences according to the *T. congolense* subgroups but also supported 4 subclades of sequences within the Savannah subgroup.

We selected sequences representative of the genetic diversity to illustrate the repertoire of CATL-like genes in *T. congolense* (Fig. 3A and B). Sequences from Forest and Kilifi subgroups always exhibited the typical catalytic triad and, respectively, QQLD or QQLN residues, preferentially, in the S2′ subsite. In contrast, sequences from the Savannah isolates diverged highly in both the catalytic triad and S2′ subsite (Fig. 4 and Fig S1 available as Supplementary online documentation).

To more specifically evaluate the genetic repertoire of congopain and homologues in *T. congolense* Savannah, we compared the 36 new sequences (from 4 laboratory and 7 field isolates) determined in this study with 13 sequences from the IL3000 genome. The analysis revealed a high level of polymorphism among and within the isolates. In the network genealogy, CATL-like sequences were subdivided into 4 subclades (SAV1 to SAV4 subgroups). Most sequences were homologous to CP2 and assigned to SAV1 (21 sequences), SAV2 (10) or SAV3 (7). Sequences of SAV1 predominated and, despite preferentially showing typical triad (CHN), this group included the first variant (SHN) reported in *T. congolense* (Downey and Donelson, 1999) in the strain TRUM 183. The group SAV4 consisted of sequences from the IL3000 genome with the SSN variant triad previously described by Pillay et al. (2010) (Fig. 3A and B). Each group, SAV1–4, exhibited specific amino acid signatures characterized by two or three motifs, including the S2 and S2′ subsite regions (Fig. 4). Nine sequences (6 from IL3000 and 3 from field isolates) did not share these signatures and did not cluster into the groups SAV1–4; these sequences were provisionally denominated as "SAV not assigned" (SAVna). From field samples, a single sample of Savannah (So.go01) showed a SAVna sequence exhibiting a unique variant catalytic triad (CSN) (Figs. 3 and 4; Fig. 1 Supplementary material).

To quantify the overall divergence of the whole repertoire of CP amino acid sequences from the three subgroups of *T. congolense*, we calculated the mean divergences and the number of polymorphic sites (PS) in amino acid (95 PS) or nucleotide (222 PS) sequences. The amino acid sequences among the Savannah isolates were highly divergent (≈14% internal divergence and 87 PS) compared to Forest (≈4.0% internal divergence and 32 PS) or Kilifi (≈1.3% internal divergence and 11 PS) isolates. However, the mean divergences of amino acid sequences between the three subgroups were comparable (12–16%).

Only one copy of CP1 was obtained from three isolates of the Savannah subgroup (Ma.ca06, Ma.bu05, and Te.ca16), and no CP1 homologues were found among 82 sequences determined in this study for Kilifi and Forest isolates (Fig. 3B).

3.3. Molecular evolution of CP repertoires

To evaluate the role of positive selection in the molecular evolution of the CP repertoires of *T. congolense* Savannah, Forest and Kilifi, we analysed the ratio of non-synonymous to synonymous substitutions (dN–dS), considering each different nucleotide triplet; the analysis included 72 CP2 sequences and a total of 159 amino acids corresponding to 477 nucleotide positions in the final dataset. Analysis of the overall inferred substitutions was performed for all subgroups. The number of non-synonymous substitutions was always higher than synonymous substitutions, regardless of the subgroup analysed. A large proportion of the substitutions were within the subgroup Savannah, and no significant difference was observed when variant CP sequences were excluded from the analysis. A specific analysis of codons revealed a high degree of conservation for the S2 and S1 subsites, with more polymorphisms for the S2′ subsite when sequences from each subgroup were compared separately. Despite non-synonymous substitutions in all subsites, non-significant *P*-values were estimated within Kilifi (0–0.45), Forest (0.45–0.70), and Savannah (0.08–0.70) subgroups. However, a comparison between the subgroups gave significant *P*-values for S1 and S2′ subsites varying from 0.007 to 0.042. In the dN–dS test, *P*-values < 0.05 suggest rejection of the null hypothesis of neutral evolution. Therefore, these results suggest that CP genes are not subject to highly constrained evolution and diverged to constitute subgroup-specific subclades within *T. congolense*. In contrast to the highly polymorphic repertoires among and within the isolates of the Savannah subgroup, our findings suggest that congopain genes are moderately divergent in Forest, and more homogeneous in the Kilifi isolates.

The occurrence of recombination events in CP genes was evaluated by eight detection methods using the RDP3 program; recombination events were considered valid if detected by at least four methods, with a minimum significance *P*-value of 0.05. Six recombination methods indicated that one sequence (Ma.bu05.c6 from a buffalo isolate from Mozambique), which nested into CP1, was a product of recombination between CP1 and CP2-like sequence (*P*-value ranging from 2.5×10^{-9} to 3.0×10^{-2}). In addition, at least one sequence from each of the SAV1 and SAV2 groups also appears to be derived from recombination events (indicated by 4 methods) (supplementary Fig. 2). Although preliminary, these findings suggest that recombination may be an important process in generating the diverse repertoire of CP sequences within the Savannah subgroup, providing additional insights to the mating capability of this subgroup (Morrison et al., 2009).

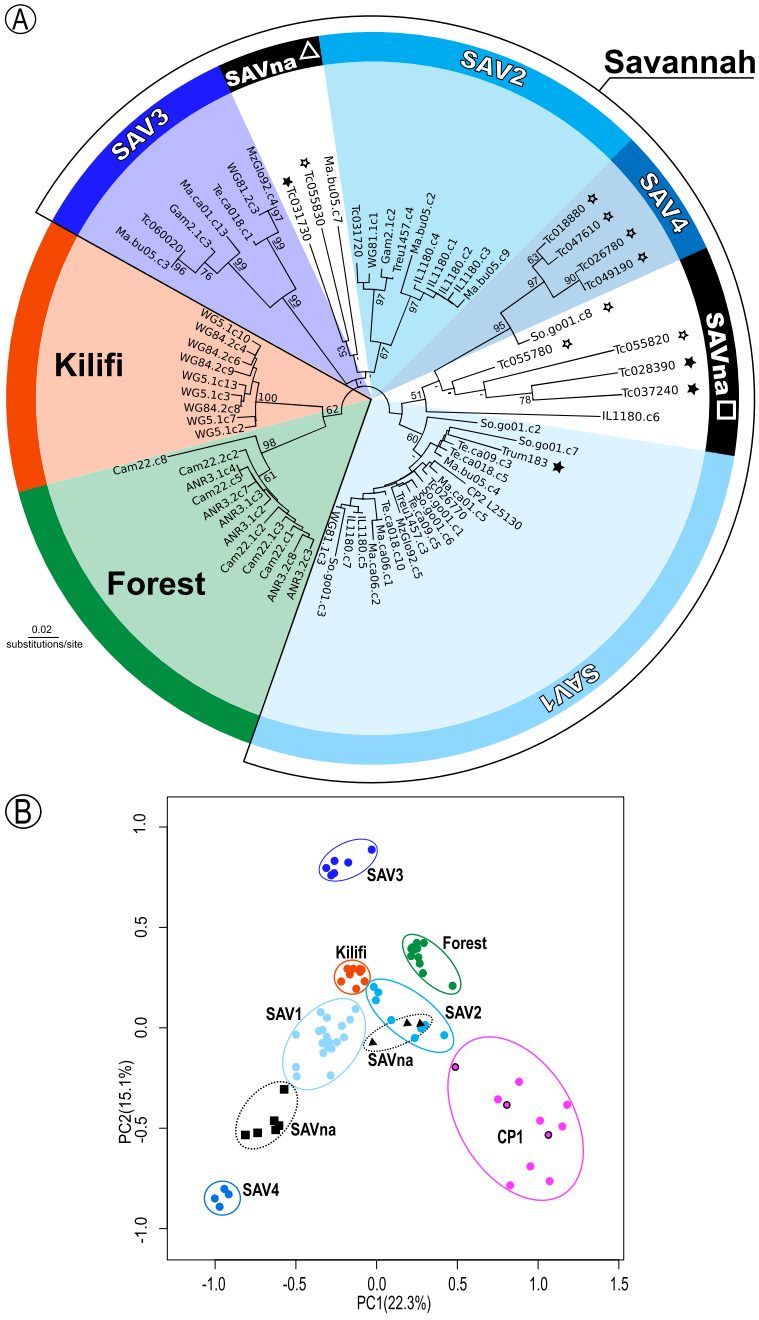

Fig. 3. (A) Genealogy of *T. congolense* catalytic domain of CATL-like inferred with 72 amino acid sequences from isolates representative of the Savannah, Forest, and Kilifi subgroups, including 36 sequences from livestock blood samples and tsetse flies determined in this study plus 13 sequences from the IL3000 genome. The four groups of sequences within the Savannah subgroup are indicated by different colours (SAV1–SAV4), and SAVna (Savannah not assigned) sequences are indicated by triangles and squares; sequences exhibiting variant catalytic triads corresponding to CP2-like (SAV4) and new CP2-like (SAVna) sequences are indicated by filled and unfilled stars, respectively. (B) The 2D sequence space of CP1 (pink circles with black stroke correspond to sequences determined in this study) and CP2 sequences were defined by the first two components (PC1 and PC2) of multidimensional scaling (MDS) plot constructed using the pairwise alignment of the 72 sequences, and the K-means method to define the groups of sequences.

3.4. Relationships of T. congolense and other trypanosome species in phylogenetic trees based on CATL-like homologous genes

We inferred phylogenies based on catalytic domains of congopain genes from the three subgroups of *T. congolense* and homologues from *T. simiae*, *T. godfreyi*, and several other *Trypanosoma* species. Only sequences sharing the catalytic triad and active sites with the archetype congopain (CP2) were included in the analysis. The results revealed an heterogeneous clade comprising all and exclusively the *T. congolense* sequences, with a sister clade formed by sequences from *T. simiae* and *T. godfreyi*, which are more closely related than those from the isolates of distinct subgroups of *T. congolense*. Together, these three species formed a monophyletic assemblage that strongly validated the subgenus *Nannomonas*. The phylogenetic positioning of this subgenus closer to the clade *T. brucei* than to *T. vivax* is in agreement with previous phylogenetic trees of *Trypanosoma* inferred with SSU rRNA and gGAPDH genes (Hamilton et al., 2004).

The congopain homologous genes from isolates of the same *T. congolense* subgroup always clustered together, forming three

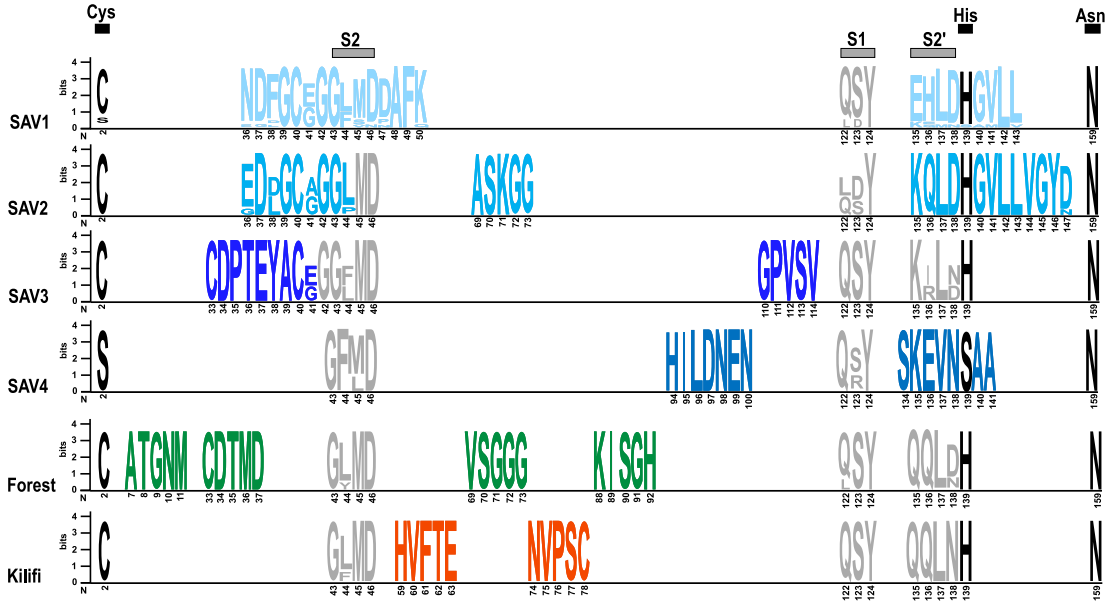

Fig. 4. Representation of signature residues within the catalytic domains of CATL-like of *T. congolense* defining the groups of sequences found within Savannah (SAV1–SAV4), Forest, and Kilifi subgroups. The amino acid motifs unique to each group of sequences were identified by analysing 155 amino acid sequences and used to design the logos. The CP2 catalytic triad, cysteine (Cys), histidine (His) and asparagine (Asn), and the subsites S2, S1, and S2′ are indicated in black and grey, respectively. The numbers indicate the position in the CP catalytic domain amino acid sequence.

well-supported subclades corresponding to Savannah, Forest, and Kilifi. In the best resolved phylogenetic tree, obtained using archetypical congopain genes, Savannah was more related to Forest than to Kilifi (Fig. 5). However, in the network genealogy inferred with all sequences including those from variant CPs (Fig. 3A), Forest appears to be more related to Kilifi than to Savannah subgroup. Most likely, inconsistencies in the phylogenetic relationships among the *T. congolense* subgroups observed when applying different inference methods (data not shown) are due to their comparable genetic distances between the nucleotide/amino acid sequences: Savannah and Forest (14%/12%), Savannah and Kilifi (14%/12%), and Forest and Kilifi (12%/12%).

The comparison of cd sequences showed large divergences separating congopain genes and all other CATL-like homologous genes, even those from the closest phylogenetically related *T. simiae* and *T. godfreyi* (26%). Sequences from *T. congolense* also largely diverged from the highly homogeneous CATL-like genes from *T. brucei* ssp. and *T. evansi* of the subgenus *Trypanozoon* (30% sequence divergence), and from the moderately variable CATL-like sequences from *T. vivax* isolates (32% sequence divergence).

3.5. *T. congolense* diagnosis, genotyping, and preliminary epidemiological study targeting CATL-like sequences

To standardise the PCR assay targeting cdCATL-like sequences (TcoCATL-PCR), the primers Tco3F and Tco4R were designed to be complementary to sequences conserved in all congopain sequences from *T. congolense* of the three subgroups, and non-complementary to sequences from any other pathogenic African trypanosomes and *T. theileri* (non-pathogenic species of ruminants). This method amplified a *T. congolense*-specific DNA fragment of ≈211 bp for isolates of all subgroups (Fig. 6A). No amplified products were detected using DNA from *T. simiae*, *T. godfreyi*, *T. vivax* (South American and West and East African genotypes), *T. brucei* ssp., *T. evansi*, *T. equiperdum*, and *T. theileri*. Negative results were obtained using DNA templates from the hemoprotozoans *Babesia bovis*, *Babesia bigemina*, and *Anaplasma* sp. (Fig. 6A).

The suitability of TcoCATL-PCR for epidemiological studies was evaluated using blood samples preserved in ethanol or spotted on filter paper from cattle (97), goats (28) and water buffalo (6), and the gut contents of tsetse flies (16), all from Mozambique (Fig. 6B). The TcoCATL-PCR was able to detect *T. congolense* in cattle blood sample that tested negative by microhaematocrit, and in tsetse samples that exhibited predominantly *T. simiae* and *T. godfreyi* mixed with very low amounts of *T. congolense* previously identified using the highly sensitive method of FFLB (Hamilton et al., 2008; Garcia et al., in preparation). Confirmation of PCR-amplified DNA bands was performed by sequencing randomly selected DNA fragments, and the results revealed exclusively isolates of the subgroup Savannah in livestock (Fig. 6C and D).

The sequences from the PCR-amplified DNA fragments (211 bp) were aligned with the corresponding sequences of congopain catalytic domains determined from the subgroups Savannah, Forest and Kilifi to evaluate their suitability for *T. congolense* genotyping. Small (211 bp) (Fig. 6C) and large (477 bp) (Fig. 3) congopain sequences resulted in similar groups corroborating the value of the TcoCATL-PCR-amplified sequences for *T. congolense* genotyping. Short sequences were sufficient for the identification of all subgroups by assessing the polymorphic sites (Fig. 6D), and the genealogy pattern (Fig. 6C).

The first epidemiological survey using TcoCATL-PCR followed by sequence analysis of the selected amplified DNA revealed that ≈27.5% of the livestock from Mozambique was infected with *T. congolense* Savannah (Fig. 6C). Our in progress study based on FFLB barcoding (Hamilton et al., 2008) has revealed high prevalence of both Savannah and Kilifi in tsetse flies from Mozambique (Garcia et al., in preparation).

4. Discussion

The present study examined the repertoire of CATL-like genes of isolates from the three subgroups of *T. congolense* (Savannah, Kilifi and Forest). This is the most comprehensive comparative study using protein coding genes from isolates of these three subgroups,

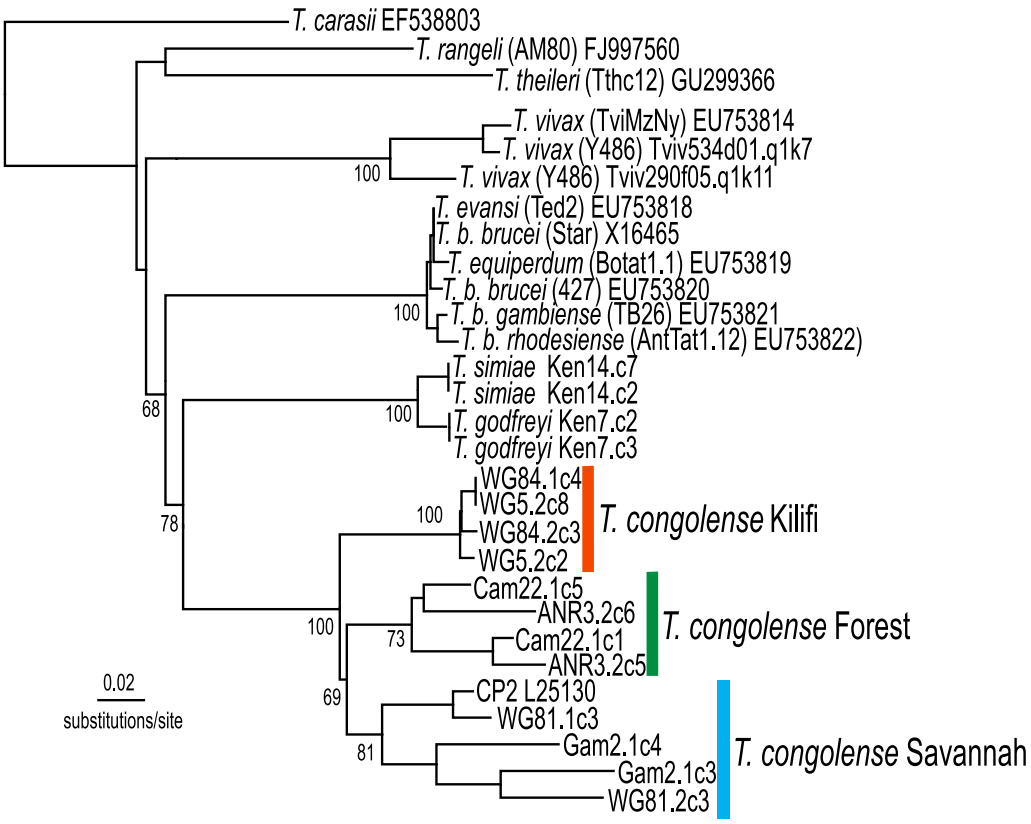

Fig. 5. *Trypanosoma* phylogenetic tree (neighbour joining tree) using congopain (CP2) amino acid sequences of *T. congolense* Savannah, Forest and Kilifi, and CATL-like homologous sequences from other trypanosome species. The numbers at the nodes are bootstrap support values from 500 replicates.

and it was carried out with the aim of investigating the genetic repertoire of congopain-encoding genes (catalytic domains), and to assess the suitability of these sequences for diagnosis, genotyping, and phylogenetic inferences.

The analysis of CP sequences demonstrated significant variability among *T. congolense* Savannah, Forest and Kilifi subgroups, with extensive polymorphism within Savannah, moderate polymorphism within Forest, and relative homogeneity within Kilifi. From subgroup Savannah, we evaluated 9 laboratory isolates plus 7 field samples from cow, buffalo, goat and sheep, and tsetse; the isolates were collected in sites differing in ecological traits and separated by large geographical distances. The two Forest isolates were obtained from tsetse and goat from The Gambia and Cameroon, respectively, while the two Kilifi isolates were from the same farm at Matuga, Kenya. Therefore, the high diversity of Savannah may reflect sampling from wider geographic and host ranges, compared with limited sampling from the other groups. Nevertheless, in contrast to isolates from Forest and Kilifi, all Savannah isolates showed very polymorphic sequences, even when derived from animals living in sympatry.

In general, the CP sequences were conserved in regions involved in both substrate specificity and enzymatic activity regardless of subgroup affiliation. However, several sequences from isolates of the subgroup Savannah, mostly from the IL3000 genome, exhibited a polymorphic S2′ subsite and unusual catalytic triads, corroborating previous reports of variant triads and the expression of congopain-like enzymes (Kakundi, 2008; Pillay et al., 2010).

The *T. congolense* IL3000 strain was selected for the genome project, and has been the subject of many studies regarding drugs and vaccines. Therefore, it is important to determine whether the remarkably diverse genetic repertoires of both congopain and congopain-like enzymes found in this strain are common to other Savannah strains, and to strains of Forest and Kilifi subgroups. With this aim, using degenerate primers designed specifically for PCR amplification of variant CPs, Kakundi (2008) obtained sequences varying in the catalytic triad, in three other Savannah strains and in one Forest strain; a single sequence was reported from each strain. The primers we have employed in this study, regardless of the polymorphisms at the primer DTO154 annealing region, allowed for the amplification of highly polymorphic sequences from Savannah isolates (Sav1-4 and SAVna sequences), from isolates of the three subgroups of *T. congolense*, and also of CATL-like from all trypanosome species examined to date (Cortez et al., 2009; Garcia et al., 2011a,b; Lima et al., 2012; Ortiz et al., 2009; Rodrigues et al., 2010). However, we cannot rule out the possibility that the primers employed for PCR amplification of CATL-like genes could have hampered the amplification of variant genes. Therefore, further studies are still necessary for a better appraisal of the congopain repertoire within *T. congolense*.

Phylogenetic analysis demonstrates that CP genes have diverged in specific subgroups, with a highly heterogeneous genetic repertoire among and within the isolates of the subgroup Savannah. The positive selection shaping the subgroup-specific and intra-Savannah genetic diversity suggests that CP2-encoding genes are not subject to highly constrained evolution among subgroups or within the subgroup Savannah. This process may have prevented extensive homogenisation, allowing for the emergence of subgroup-specific and highly divergent CP2 and CP2-like genes within Savannah. Differences in the ability to recombine may account for the higher diversity within Savannah. Microsatellite analyses suggested high variability, most likely resulting from mating in Savannah (Morrison et al., 2009), and low genetic variability and predominant clonal reproduction in Forest (Simo et al., 2013). Compared with data from *T. congolense*, there is a limited

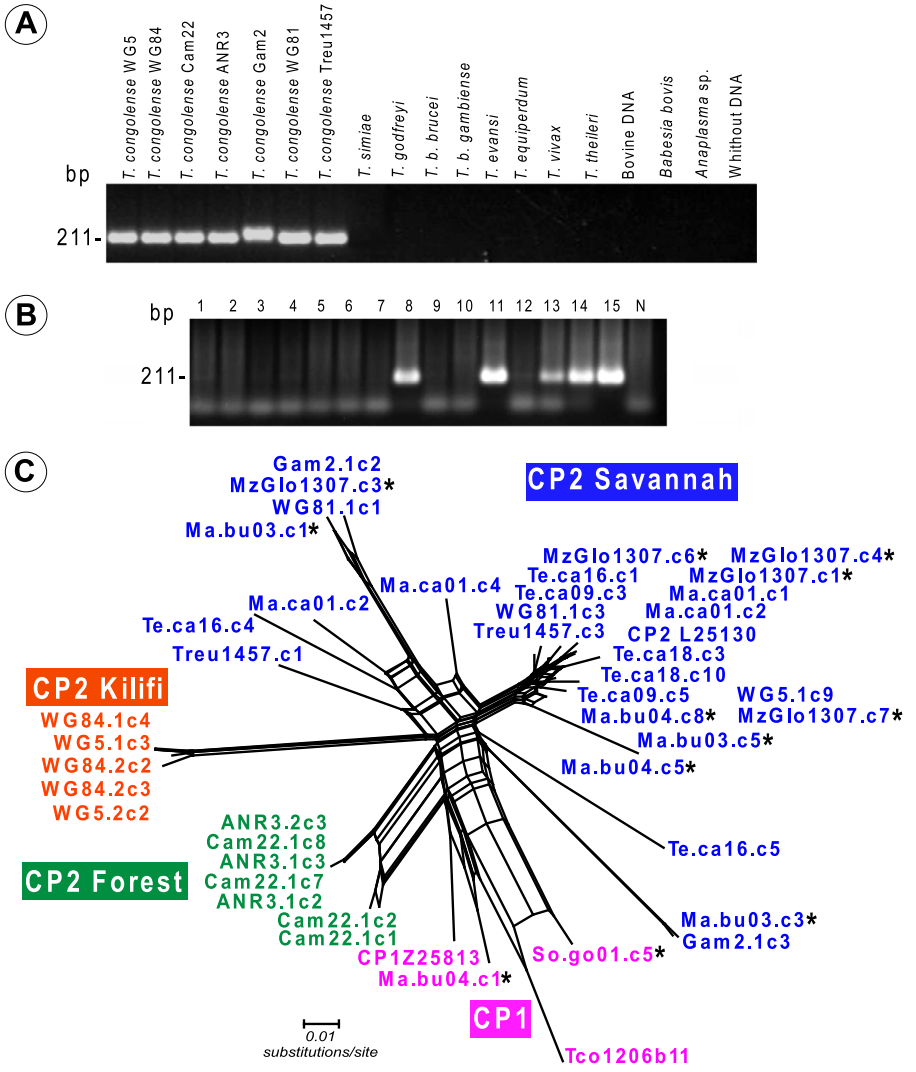

Fig. 6. Agarose gels stained with ethidium bromide (EtBr) showing DNA fragments amplified by TcoCATL-PCR: (A) specificity analysis using the DNA of isolates from the subgroups Savannah, Forest, and Kilifi, and other trypanosome species; (B) illustrative results from the evaluation of the suitability of the TcoCATL-PCR using crude DNA preparations from field-collected blood samples from sheep (1–5), buffalo (6–9), and cattle (11–14); positive (15) and negative (N) controls; (C) network genealogy of sequences amplified by TcoCATL-PCR (*) aligned with the corresponding region from *T. congolense* of all subgroups; analysis performed with the Neighbour-Net method with the K2P parameter and 1000 bootstrap replicates. (D) Selected region of aligned sequences illustrating the polymorphic sites useful for the genotyping of *T. congolense* subgroups.

polymorphism in CATL-like genes from *T. brucei* ssp. and *T. evansi* and moderate diversity in *T. vivax* (Cortez et al., 2009). Interestingly, analysis of Cathepsin B genes from *T. congolense* IL3000 revealed 13 gene copies with unusual polymorphisms in contrast to the single-copy gene from other trypanosome species (Mendoza-Palomares et al., 2008).

The results from this study provided new insights into the diversity of *T. congolense* CATL-like enzymes. Differences regarding the development in tsetse fly of different species and in experimental (livestock and mice) and field-infected animals have been observed among *T. congolense* Savannah, Forest and Kilifi isolates, and also within Savannah isolates (Bengaly et al., 2002a,b; Masumu et al., 2006; Seck et al., 2010; Van den Bossche et al., 2011; Vitouley et al., 2011; Moti et al., 2012). Evidence from this study showing that highly virulent (Savannah), moderate (Forest) and non-virulent (Kilifi) isolates differ in their CP2 gene repertoires deserve to be better investigated regarding the association of virulence with distinct enzymes.

Findings from this study demonstrated for the first time that congopain genes are valuable markers for genotyping and phylogenetic inferences in *T. congolense* Savannah, Forest and Kilifi. Inferred phylogenetic trees based on CATL-like genes were similar to those based on SSU rRNA and gGAPDH genes (Hamilton et al., 2004) clustering *T. congolense* together with *T. simiae* and *T. godfreyi* in the clade corresponding to subgenus *Nannomonas*. The assemblages comprising all and exclusively sequences from *T. congolense* were formed by three well-supported subclades corresponding to the three known subgroups: Savannah, Forest, and Kilifi. The best resolved phylogenetic analysis showed that Savannah and Forest isolates were more closely related and distant from isolates of the subgroup Kilifi, consistent with the results based on GARP (glutamate- and alanine-rich protein) gene sequences (Asbeck et al., 2000). Sequences of ribosomal RNA genes also supported a closer relationship between Savannah and Forest (Auty et al., 2012). We provide additional genetic evidence based on congopain genes corroborating that the three subgroups of *T. congolense* diverged enough to be separated into phylogenetically supported species. The genetic distances separating the subgroups of *T. congolense* are larger than the divergences between *T. simiae* and *T. godfreyi*. However, the characterization of more samples of Forest and Kilifi subgroups are required for a better appraisal of diversity, recombination, and taxonomic status of subgroups within *T. congolense* throughout sub-Saharan Africa (Gibson, 2007).

We developed a *T. congolense*-specific PCR assay targeting CATL-like sequences using crude DNA templates from field-collected blood samples preserved in ethanol at room temperature and, hence, amenable to epidemiological studies. The diagnostic PCR assay generated the same-sized fragments for the three subgroups of *T. congolense*, which can be genotyped by sequencing the small PCR-amplified fragment. The first method developed for this purpose, based on repetitive DNA, required three independent PCR reactions, one for each subgroup (Masiga et al., 1992), whereas a single PCR based on ITS1 rDNA can distinguish Savannah (700 bp), Forest (710 bp) and Kilifi (620 bp) (Desquesnes et al., 2001). However, epidemiological studies have required further sequencing of amplified ITS rDNA to distinguish between Savannah and Forest, which are often found as mixed infections (Malele et al., 2011; Auty et al., 2012).

5. Conclusion

We showed for the first time *T. congolense* Savannah, Forest and Kilifi specific repertoires of genes encoding congopain enzymes. The knowledge of genetic repertoires of CP enzymes and, specifically, of congopain, are valuable for studies about the roles of these enzymes in pathogenicity and virulence, and in the design of targets for the development of polyvalent vaccines and enzyme inhibitors useful as drugs against infections caused by isolates of the three subgroups of *T. congolense*. The method of PCR developed in this study can be helpful to improve the diagnosis and genotyping of *T. congolense* subgroups using crude DNA preparations from field-samples from livestock, wild reservoirs and tsetse flies.

Acknowledgements

We are indebted to the several students, colleagues, and local people who helped with the fieldwork in Mozambique. We are grateful to The Wellcome Trust for making available sequences from the genome of *T. congolense*. Sequence data were obtained from the Sanger Institute at the website of Genedb, http://www.genedb.org. This work was supported by the Brazilian agency CNPq within the PROAFRICA and UNIVERSAL programs. Adriana C. Rodrigues is a postdoctoral fellow of PNPD-CAPES, and Paola A. Ortiz and André G. Martins are recipients of scholarships from PROTAX-CNPq.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2014.01.012.

References

Adams, E.R., Hamilton, P.B., Gibson, W.C., 2010. African trypanosomes: celebrating diversity. Trends Parasitol. 26, 324–328.
Alvarez, V.E., Niemirowicz, G.T., Cazzulo, J.J., 2012. The peptidases of *Trypanosoma cruzi*: digestive enzymes, virulence factors, and mediators of autophagy and programmed cell death. Biochim. Biophys. Acta 1824, 195–206.
Asbeck, K., Ruepp, S., Roditi, I., Gibson, W., 2000. GARP is highly conserved among *Trypanosoma congolense* Savannah, Forest and Kilifi subgroups. Mol. Biochem. Parasitol. 106, 303–306.
Atkinson, H.J., Babbitt, P.C., Sajid, M., 2009. The global cysteine peptidase landscape in parasites. Trends Parasitol. 25, 573–581.
Authié, E., 1994. Trypanosomiasis and trypanotolerance in cattle: a role for congopain? Parasitol. Today 10, 360–364.
Authié, E., Boulangé, A., Muteti, D., Lalmanach, G., Gauthier, F., Musoke, A.J., 2001. Immunisation of cattle with cysteine proteinases of *Trypanosoma congolense*: targetting the disease rather than the parasite. Int. J. Parasitol. 31, 1429–1433.
Authié, E., Muteti, D.K., Mbawa, Z.R., Lonsdale-Eccles, J.D., Webster, P., Wells, C.W., 1992. Identification of a 33-kilodalton immunodominant antigen of *Trypanosoma congolense* as a cysteine protease. Mol. Biochem. Parasitol. 56, 103–116.
Auty, H., Anderson, N.E., Picozzi, K., Lembo, T., Mubanga, J., Hoare, R., Fyumagwa, R.D., Mable, B., Hamill, L., Cleaveland, S., Welburn, S.C., 2012. Trypanosome diversity in wildlife species from the serengeti and Luangwa Valley ecosystems. PLoS Negl. Trop. Dis. 6, 1828.
Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S., 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37, 202–208.
Bengaly, Z., Sidibe, I., Boly, H., Sawadogo, L., Desquesnes, M., 2002a. Comparative pathogenicity of three genetically distinct *Trypanosoma congolense*-types in inbred Balb/c mice. Vet. Parasitol. 105, 111–118.
Bengaly, Z., Sidibe, I., Ganaba, R., Desquesnes, M., Boly, H., Sawadogo, L., 2002b. Comparative pathogenicity of three genetically distinct types of *Trypanosoma congolense* in cattle: clinical observations and haematological changes. Vet. Parasitol. 108, 1–19.
Boulangé, A., Serveau, C., Brillard, M., Minet, C., Gauthier, F., Diallo, A., Lalmanach, G., Authié, E., 2001. Functional expression of the catalytic domains of two cysteine proteinases from *Trypanosoma congolense*. Int. J. Parasitol. 31, 1435–1440.
Cortez, A.P., Rodrigues, A.C., Garcia, H.A., Neves, L., Batista, J.S., Bengaly, Z., Paiva, F., Teixeira, M.M.G., 2009. Cathepsin L-like genes of *Trypanosoma vivax* from Africa and South America – characterization, relationships and diagnostic implications. Mol. Cell. Probes 23, 44–51.
Desquesnes, M., McLaughlin, G., Zoungrana, A., Dávila, A.M., 2001. Detection and identification of *Trypanosoma* of African livestock through a single PCR based on internal transcribed spacer 1 of rDNA. Int. J. Parasitol. 31, 610–614.
Downey, N., Donelson, J.E., 1999. Expression of foreign proteins in *Trypanosoma congolense*. Mol. Biochem. Parasitol. 104, 39–53.
Fish, W.R., Nkhungulu, Z.M., Muriuki, C.W., Ndegwa, D.M., Lonsdale-Eccles, J.D., Steyaert, J., 1995. Primary structure and partial characterization of a life-cycle-

regulated cysteine protease from *Trypanosoma (Nannomonas) congolense*. Gene 161, 125–128.
Garcia, H.A., Kamyingkird, K., Rodrigues, A.C., Jittapalapong, S., Teixeira, M.M.G., Desquesnes, M., 2011a. High genetic diversity in field isolates of *Trypanosoma theileri* assessed by analysis of cathepsin L-like sequences disclosed multiple and new genotypes infecting cattle in Thailand. Vet. Parasitol. 180, 363–367.
Garcia, H.A., Rodrigues, A.C., Martinkovic, F., Minervino, A.H., Campaner, M., Nunes, V.L., Paiva, F., Hamilton, P.B., Teixeira, M.M.G., 2011b. Multilocus phylogeographical analysis of *Trypanosoma* (*Megatrypanum*) genotypes from sympatric cattle and water buffalo populations supports evolutionary host constraint and close phylogenetic relationships with genotypes found in other ruminants. Int. J. Parasitol. 41, 1385–1396.
Garside, L.H., Gibson, W.C., 1995. Molecular characterization of trypanosome species and subgroups within subgenus *Nannomonas*. Parasitology 111, 301–312.
Gashumba, J.K., Baker, R.D., Godfrey, D.G., 1988. *Trypanosoma congolense*: the distribution of enzymic variants in east and West Africa. Parasitology 96, 475–486.
Gibson, W., 2002. Epidemiology and diagnosis of African trypanosomiasis using DNA probes. Trans. R. Soc. Trop. Med. Hyg. 96, 141–143.
Gibson, W., 2007. Resolution of the species problem in African trypanosomes. Int. J. Parasitol. 37, 829–838.
Gibson, W., 2012. The origins of the trypanosome genome strains *Trypanosoma brucei brucei TREU 927, T. b. gambiense* DAL 972, *T. vivax* Y486 and *T. congolense* IL3000. Parasit. Vectors 5, 1.
Gibson, W.C., Dukes, P., Gashumba, J.K., 1988. Species-specific DNA probes for the identification of African trypanosomes in tsetse flies. Parasitology 97, 63–73.
Gibson, W.C., Stevens, J.R., Mwendia, C.M., Ngotho, J.N., Ndung'u, J.M., 2001. Unravelling the phylogenetic relationships of African trypanosomes of suids. Parasitology 122, 625–631.
Hamilton, P.B., Adams, E.R., Malele, I.I., Gibson, W.C., 2008. A novel, high-throughput technique for species identification reveals a new species of tsetse-transmitted trypanosome related to the *Trypanosoma brucei* subgenus *Trypanozoon*. Infect. Genet. Evol. 8, 26–33.
Hamilton, P.B., Stevens, J.R., Gaunt, M.W., Gidley, J., Gibson, W.C., 2004. Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA. Int. J. Parasitol. 34, 1393–1404.
Huson, L.E., Authié, E., Boulangé, A.F., Goldring, J.P., Coetzer, T.H., 2009. Modulation of the immunogenicity of the *Trypanosoma congolense* cysteine protease, congopain, through complexation with alpha(2)-macroglobulin. Vet. Res. 40, 52.
Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.
Jaye, A.B., Nantulya, V.M., Majiwa, P.A., Urakawa, T., Masake, R.A., Wells, C.W., ole-MoiYoi, O.K., 1993. EMBL/GenBank/DDBJ databases.
Kateregga, J., Lubega, G.W., Lindblad, E.B., Authié, E., Coetzer, T.H., Boulangé, A.F., 2012. Effect of adjuvants on the humoral immune response to congopain in mice and cattle. Vet. Res. 8, 63.
Kakundi, E. M., 2008. Molecular Analysis of the Congopain Gene Family, School of Biochemistry, Genetics, Microbiology and Plant Pathology, University of Kwa Zulu-Natal, Pietermaritzburg (MSc Dissertation).
Knowles, G., Betschart, B., Kukla, B.A., Scott, J.R., Majiwa, P.A., 1988. Genetically discrete populations of *Trypanosoma congolense* from livestock on the Kenyan coast. Parasitology 96, 461–474.
Lalmanach, G., Boulangé, A., Serveau, C., Lecaille, F., Scharfstein, J., Gauthier, F., Authié, E., 2002. Congopain from *Trypanosoma congolense*: drug target and vaccine candidate. Biol. Chem. 383, 739–749.
Lima, A.P., Tessier, D.C., Thomas, D.Y., Scharfstein, J., Storer, A.C., Vernet, T., 1994. Identification of new cysteine protease gene isoforms in *Trypanosoma cruzi*. Mol. Biochem. Parasitol. 67, 333–338.
Lima, L., Ortiz, P.A., da Silva, F.M., Alves, J.M.P., Serrano, M.G., Cortez, A.P., Alfieri, S.C., Buck, G.A., Teixeira, M.M.G., 2012. Repertoire, genealogy and genomic organization of cruzipain and homologous genes in *Trypanosoma cruzi*, *T. cruzi*-like and other trypanosome species. PLoS One 7, e38385.
Majiwa, P.A., Hamers, R., Van Meirvenne, N., Matthyssens, G., 1986. Evidence for genetic diversity in *Trypanosoma (Nannomonas) congolense*. Parasitology 93, 291–304.
Majiwa, P.A., Maina, M., Waitumbi, J.N., Mihok, S., Zweygarth, E., 1993. *Trypanosoma (Nannomonas) congolense*: molecular characterization of a new genotype from Tsavo, Kenya. Parasitology 106, 151–162.
Majiwa, P.A., Masake, R.A., Nantulya, V.M., Hamers, R., Matthyssens, G., 1985. *Trypanosoma (Nannomonas) congolense*: identification of two karyotypic groups. EMBO J. 4, 3307–3313.
Malele, I.I., Magwisha, H.B., Nyingilili, H.S., Mamiro, K.A., Rukambile, E.J., Daffa, J.W., Lyaruu, E.A., Kapange, L.A., Kasilagila, G.K., Lwitiko, N.K., Msami, H.M., Kimbita, E.N., 2011. Multiple *Trypanosoma* infections are common amongst *Glossina* species in the new farming areas of Rufiji district, Tanzania. Parasit. Vectors 4, 217.
Mamabolo, M.V., Ntantiso, L., Latif, A., Majiwa, P.A.O., 2009. Natural infection of cattle and tsetse flies in South Africa with two genotypic groups of *Trypanosoma congolense*. Parasitology 136, 425–431.
Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., Lefeuvre, P., 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26, 2462–2463.
Masiga, D.K., McNamara, J.J., Laveissière, C., Truc, P., Gibson, W.C., 1996. A high prevalence of mixed trypanosome infections in tsetse flies in Sinfra, Côte d'Ivoire, detected by DNA amplification. Parasitology 112, 75–80.
Masiga, D.K., Smyth, A.J., Hayes, P., Bromidge, T.J., Gibson, W.C., 1992. Sensitive detection of trypanosomes in tsetse flies by DNA amplification. Int. J. Parasitol. 22, 909–918.
Masumu, J., Marcotty, T., Geysen, D., Geerts, S., Vercruysse, J., Dorny, P., den Bossche, P.V., 2006. Comparison of the virulence of *Trypanosoma congolense* strains isolated from cattle in a trypanosomiasis endemic area of eastern Zambia. Int. J. Parasitol. 36, 497–501.
Mekata, H., Konnai, S., Simuunza, M., Chembensofu, M., Kano, R., Witola, W.H., Tembo, M.E., Chitambo, H., Inoue, N., Onuma, M., Ohashi, K., 2008. Prevalence and source of trypanosome infections in field-captured vector flies (*Glossina pallidipes*) in southeastern Zambia. J. Vet. Med. Sci. 70, 923–928.
Mendoza-Palomares, C., Biteau, N., Giroud, C., Coustou, V., Coetzer, T., Authié, E., Boulangé, A., Baltz, T., 2008. Molecular and biochemical characterization of a cathepsin B-like protease family unique to *Trypanosoma congolense*. Eukaryot. Cell 7, 684–697.
Morrison, L.J., Tweedie, A., Black, A., Pinchbeck, G.L., Christley, R.M., Schoenefeld, A., Hertz-Fowler, C., MacLeod, A., Turner, C.M.R., Tait, A., 2009. Discovery of mating in the major African livestock pathogen *Trypanosoma congolense*. PLoS One 4, e5564.
Moti, Y., Fikru, R., Van Den Abbeele, J., Büscher, P., Van den Bossche, P., Duchateau, L., Delespaux, V., 2012. Ghibe river basin in Ethiopia: present situation of trypanocidal drug resistance in *Trypanosoma congolense* using tests in mice and PCR-RFLP. Vet. Parasitol. 189, 197–203.
Njiru, Z.K., Makumi, J.N., Okoth, S., Ndungu, J.M., Gibson, W.C., 2004. Identification of trypanosomes in *Glossina pallidipes* and *G. longipennis* in Kenya. Infect. Genet. Evol. 4, 29–35.
Ortiz, P.A., Maia da Silva, F., Cortez, A.P., Lima, L., Campaner, M., Pral, E.M.F., Alfieri, S.C., Teixeira, M.M.G., 2009. Genes of cathepsin L-like proteases in *Trypanosoma rangeli* isolates: markers for diagnosis, genotyping and phylogenetic relationships. Acta Trop. 112, 249–259.
Pelé, J., Bécu, J.-M., Abdi, H., Chabbert, M., 2012. Bios2mds: an R package for comparing orthologous protein families by metric multidimensional scaling. BMC Bioinformatics 13, 133.
Pillay, D., Boulangé, A.F., Coetzer, T.H.T., 2010. Expression, purification and characterisation of two variant cysteine peptidases from *Trypanosoma congolense* with active site substitutions. Protein Expr. Purif. 74, 264–271.
Reifenberg, J.M., Cuisance, D., Frezil, J.L., Cuny, G., Duvallet, G., 1997. Comparison of the susceptibility of different *Glossina* species to simple and mixed infections with *Trypanosoma (Nannomonas) congolense* savannah and riverine forest types. Med. Vet. Entomol. 11, 246–252.
Rodrigues, A.C., Garcia, H.A., Ortiz, P.A., Cortez, A.P., Martinkovic, F., Paiva, F., Batista, J.S., Minervino, A.H., Campaner, M., Pral, E.M., Alfieri, S.C., Teixeira, M.M.G., 2010. Cysteine proteases of *Trypanosoma (Megatrypanum) theileri*: cathepsin L-like gene sequences as targets for phylogenetic analysis, genotyping diagnosis. Parasitol. Int. 59, 318–325.
Sajid, M., McKerrow, J.H., 2002. Cysteine proteases of parasitic organisms. Mol. Biochem. Parasitol. 2002 (120), 1–21.
Seck, M.T., Bouyer, J., Sall, B., Bengaly, Z., Vreysen, M.J.B., 2010. The prevalence of African animal trypanosomoses and tsetse presence in Western Senegal. Parasite 17, 257–265.
Simo, G., Silatsa, B., Flobert, N., Lutumba, P., Mansinsa, P., Madinga, J., Manzambi, E., De Deken, R., Asonganyi, T., 2012. Identification of different trypanosome species in the mid-guts of tsetse flies of the Malanga (Kimpese) sleeping sickness focus of the Democratic Republic of Congo. Parasit. Vectors 5, 201.
Simo, G., Sobgwi, P.F., Njitchouang, G.R., Njiokou, F., Kuiate, J.R., Cuny, G., Asonganyi, T., 2013. Identification and genetic characterization of *Trypanosoma congolense* in domestic animals of Fontem in the South-West region of Cameroon. Infect. Genet. Evol. 18, 66–73.
Van den Bossche, P., Chitanga, S., Masumu, J., Marcotty, T., Delespaux, V., 2011. Virulence in *Trypanosoma congolense* Savannah subgroup. A comparison between strains and transmission cycles. Parasit. Immunol. 33, 456–460.
Vitouley, H.S., Mungube, E.O., Allegye-Cudjoe, E., Diall, O., Bocoum, Z., Diarra, B., Randolph, T.F., Bauer, B., Clausen, P.-H., Geysen, D., Sidibe, I., Bengaly, Z., Van den Bossche, P., Delespaux, V., 2011. Improved PCR-RFLP for the detection of diminazene resistance in *Trypanosoma congolense* under field conditions using filter papers for sample storage. PLoS Negl. Trop. Dis. 5, e1223.
Yang, Z., 2007. PAML4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 8, 1586–1591.
Yang, Z., Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17, 32–43.
Young, C.J., Godfrey, D.G., 1983. Enzyme polymorphism and the distribution of *Trypanosoma congolense* isolates. Ann. Trop. Med. Parasitol. 77, 467–481.

Caballero et al. Parasites & Vectors (2015) 8:222
DOI 10.1186/s13071-015-0829-y

Parasites & Vectors

RESEARCH Open Access

Phylogenetic and syntenic data support a single horizontal transference to a *Trypanosoma* ancestor of a prokaryotic proline racemase implicated in parasite evasion from host defences

Zuleima C Caballero[1,2†], Andre G Costa-Martins[1†], Robson C Ferreira[1], João M P Alves[1], Myrna G Serrano[3], Erney P Camargo[1], Gregory A Buck[3], Paola Minoprio[4] and Marta M G Teixeira[1*]

Abstract

Background: Proline racemase (PRAC) enzymes of *Trypanosoma cruzi* (*Tc*PRAC), the agent of Chagas disease, and *Trypanosoma vivax* (*Tv*PRAC), the agent of livestock trypanosomosis, have been implicated in the B-cells polyclonal activation contributing to immunosuppression and the evasion of host defences. The similarity to prokaryotic PRAC and the absence in *Trypanosoma brucei* and *Trypanosoma congolense* have raised many questions about the origin, evolution, and functions of trypanosome PRAC (*Try*PRAC) enzymes.

Findings: We identified *Try*PRAC homologs as single copy genes per haploid genome in 12 of 15 *Trypanosoma* species, including *T. cruzi* and *T. cruzi marinkellei*, *T. dionisii, T. erneyi, T. rangeli*, *T. conorhini* and *T. lewisi,* all parasites of mammals. Polymorphisms in *Tc*PRAC genes matched *T. cruzi* genotypes: TcI-TcIV and Tcbat have unique genes, while the hybrids TcV and TcVI contain *Tc*PRACA and *Tc*PRACB from parental TcII and TcIII, respectively. PRAC homologs were identified in trypanosomes from anurans, snakes, crocodiles, lizards, and birds. Most trypanosomes have intact PRAC genes. *T. rangeli* possesses only pseudogenes, maybe in the process of being lost. *T. brucei, T. congolense* and their allied species, except the more distantly related *T. vivax*, have completely lost PRAC genes.

Conclusions: The genealogy of *Try*PRAC homologs supports an evolutionary history congruent with the *Trypanosoma* phylogeny. This finding, together with the synteny of PRAC loci, the relationships with prokaryotic PRAC inferred by taxon-rich phylogenetic analysis, and the absence in trypanosomatids of any other genera or in bodonids or euglenids suggest that a common ancestor of *Trypanosoma* gained PRAC gene by a single and ancient horizontal gene transfer (HGT) from a Firmicutes bacterium more closely related to *Gemella* and other species of Bacilli than to *Clostridium* as previously suggested. Our broad phylogenetic study allowed investigation of *Try*PRAC evolution over long and short timescales. *Try*PRAC genes diverged to become species-specific and genotype-specific for *T. cruzi* and *T. rangeli,* with resulting genealogies congruent with those obtained using vertically inherited genes. The inventory of *Try*PRAC genes described here is the first step toward the understanding of the roles of PRAC enzymes in trypanosomes differing in life cycles, virulence, and infection and immune evasion strategies.

Keywords: Proline racemase, *Trypanosoma cruzi*, *Trypanosoma vivax*, *Trypanosoma rangeli*, Horizontal gene transfer, Gene loss, Kinetoplastid evolution, Phylogeny, Synteny, Genotyping

* Correspondence: mmgteix@icb.usp.br
†Equal contributors
[1]Departamento de Parasitologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, SP 05508-900, Brazil
Full list of author information is available at the end of the article

Background

The kinetoplastids (Euglenozoa: Kinetoplastea) are composed of bodonids, which include free-living and parasitic species in aquatic environments, and their descendants, the obligate parasitic trypanosomatids. These include parasites of insects and plants as well as *Trypanosoma* and *Leishmania,* which alternate between invertebrate and vertebrate hosts, including humans [1-3]. Although the trypanosomes are parasites of all vertebrate classes, they are highly divergent in their host ranges. Some species have a single vertebrate host, while others can infect diverse genera and orders. Most trypanosomes are thought to develop exclusively in the bloodstream, but a few species can also live in extra-vascular (*T. brucei* ssp. and *T. vivax*) and intracellular (*T. cruzi* and allied species) compartments. In vectors such as flies, hemipterans, mosquitoes, fleas or leeches, most trypanosomes develop in the gut, although some can invade the haemolymph and multiply and differentiate to infective forms in the salivary glands of their vectors [1,3-5].

Morphological and functional diversification has given rise to trypanosomatids differing in life cycles, vertebrate hosts, and vectors. Parasite adaptations to the variable host and vector environments have resulted in the development of diverse physiological processes and unique mechanisms to evade the host defences. Characterisation of molecules essential to metabolism and host interactions is fundamental to the elucidation of the emergence of pathogenicity and the diverse evolutionary strategies used by trypanosomes to infect and survive within a wide range of vertebrates and invertebrates.

Polyclonal lymphocyte-B activation is one of the major immunological disorders observed during microbial infections and is among the main strategies used by *T. cruzi* to evade the host specific immune response, ensuring its survival in vertebrate hosts [6-11]. This process can be triggered by proline racemase (PRAC) enzymes released by *T. cruzi*, which are implicated in the virulence of the parasite and induce indiscriminate activation of B-cells producing high levels of non-specific antibodies that contribute to immunosuppression and, consequently, to parasite immune evasion and persistence in the host. Immunological and biochemical studies confirmed that both *Tc*PRAC (*T. cruzi*) and *Tv*PRAC (*T. vivax*) exhibited proline racemase activity and B-cell mitogenicity, inducing polyclonal activation, delayed specific immune responses to parasite antigens favouring parasite immune evasion, and concomitant increase of parasitemias in the early phase of infection [7-12].

Amino-acid racemases are enzymes that catalyse the interconversion of free L- and D-amino acids. D-amino acids released by bacteria are key factors for the cell wall remodelling essential for adaptation to environmental challenges. Alanine- and glutamate-racemases are necessary for the synthesis of bacterial cell wall (peptidoglycan), which provide protection against proteolysis and host immune defences [13-15]. The enzymatic activity of Proline-racemase (PRAC) enzymes has been identified in a restricted group of bacteria, acting as a virulence factor in the highly pathogenic *Clostridium difficile* and *Pseudomonas aeruginosa* [14-16]. The first PRAC enzyme was isolated from *Clostridium sticklandii* in 1968 [15]. The first eukaryotic PRAC was reported in 2000 in *T. cruzi* [11]. In 2009, a PRAC was reported in *T. vivax* [8]. It is now recognised that the PRAC-like gene family is widely distributed throughout prokaryotes but scarce in eukaryotes, which according to phylogenetic analyses have acquired distinct bacterial PRAC-like genes by independent horizontal gene transfer (HGT) events. The repertoires and roles of PRAC-like genes in eukaryotes other than trypanosomes, including fungi, humans and other animals, are just beginning to be appreciated [16,17].

The search for PRAC genes in the genome of the *T. cruzi* CL Brener strain revealed two genes encoding two enzyme isoforms essential for viability and differentially expressed during parasite development: *Tc*PRACA (secreted by metacyclic and bloodstream trypomastigotes) and *Tc*PRACB (intracellular protein of epimastigotes). The two enzymes share 96% amino-acid identity but differ in kinetic properties relevant to catalytic activities [9,10]. Evidence provided by inhibitors of *Tc*PRAC support its suitability as a target for chemotherapy against Chagas disease [18,19]. However, to date neither the genetic nor the enzymatic diversity of *Tc*PRAC was investigated for any other strain of *T. cruzi* besides CL Brener. All protein candidates for drug design should consider the diversity within *T. cruzi* [20].

T. cruzi is a complex of genetically diverse isolates distributed in seven intraspecific subdivisions: the DTUs (Discrete Typing Units) TcI-TcVI and Tcbat. The heterogeneity of *T. cruzi* isolates has been implicated in different clinical forms of the disease. Chagas disease pathology ranges from subclinical infection to severe cardiac and digestive syndromes. However, attempts to associate *T. cruzi* genotypes with clinical forms, degrees and types of host-cell invaded, virulence and metacyclogenesis suggested some degrees of association, but involved several factors from hosts and parasites that are not well understood [21].

As mentioned above, *Tc*PRAC enzymes contribute to delays in the effective host immune response by non-specifically activating B-lymphocytes, thus enhancing the ability of the parasite to avoid immune clearance [8,9,11]. Treatment of macrophages with recombinant *Tc*PRAC induces the secretion of a soluble factor that promotes B-cell proliferation. *Tc*PRAC also activates the production of a cytokine known to enhance host susceptibility to *T. cruzi,* thus enhancing parasite virulence [22,23]. Over-expression of *Tc*PRAC genes increased the

differentiation of non-infective epimastigotes into infective metacyclic trypomastigotes, suggesting that the enzyme may regulate intracellular metabolic pathways of L-proline internalised from the vector gut. The inhibition of *Tc*PRAC significantly reduced the invasion of cells, and the intracellular differentiation of *T. cruzi* [9,10]. L-proline is one major source of energy for *T. cruzi* not only in the vectors but also during host-cell invasion and, in addition, improves parasite protection against oxidative stress [24,25]. *Tc*PRAC may also participate in the addition of D-amino acids to peptides, generating less immunogenic parasites, and maybe providing resistance against host proteolytic mechanisms as described for bacterial cell walls [10,12-14,16,22].

To date, the only trypanosome other than *T. cruzi* in which a PRAC homolog was reported is *T. vivax*. Similar to *Tc*PRAC, *Tv*PRAC displays racemase enzymatic activity, and induces polyclonal activation (mitogenic activity) in B-cells [8]. *T. vivax* evades the host immune system due to VSG expression, multiplying extracellularly in the bloodstream, and invading and multiplying in tissue spaces and the CNS similarly to the *T. brucei* ssp. agents of Sleeping Sickness [26,27].

Throughout their evolutionary history, trypanosomes have relied on various strategies to infect their hosts, obtain energy from sources available in vectors (gut and haemolymph) and vertebrate hosts (blood, intravascular and intracellular compartments), evade host defences, and develop virulence factors that play different roles according to the trypanosome species. PRAC enzymes have been implicated in these processes in *T. cruzi* and *T. vivax* [7-12]. The discovery of PRAC enzymes in *T. cruzi* and *T. vivax*, which are species separated by large genetic distances, and the absence of homologs in the genomes of *T. brucei* and *T. congolense,* which together with *T. vivax* form the clade *T. brucei* exclusive of African pathogenic trypanosomes [3,28] and the lack in *Leishmania* spp. [8], suggest that PRAC have a complex evolutionary history in the Trypanosomatidae family.

The acquisition by HGT of a large number of foreign genes from viruses, bacteria, eukaryotes and even vertebrate hosts and vectors can change genetic and metabolic repertoires, and has played important roles in the evolution of trypanosomatids and other protistan parasites. HGT has been an important evolutionary force in the adaptation of trypanosomatids to parasitism and to specialised niches within hosts, largely contributing to amino acid and carbohydrate metabolic pathways. In addition, an increasing number of putative proteins of unknown function gained from bacterium have been identified in the *T. brucei*, *T. cruzi* and *Leishmania* spp. genomes [29-35]. Recent studies have characterised putative HGT contributing to host infection, cell invasion, virulence, and pathogenesis of trypanosomatids. It was suggested that *T. cruzi* acquired genes for the calcium mobilisation necessary for host-cell invasion via ancient HGT from *Salmonella* [36]. The analysis of the phosphatidylinositol kinase gene family revealed a novel gene of *T. cruzi*, *T. brucei*, *T. congolense*, *T. vivax,* and *Leishmania* spp. that may have been acquired from a virus through HGT [37].

To achieve a better understanding of the origin, the possible bacterial donors and the evolution of *Try*PRAC genes, we searched for *Tc*PRAC homologs in the genomes of *T. cruzi* representing the whole range of intra-specific diversity (DTUs TcI-TcVI and Tcbat), other trypanosomes of mammals (*T. c. marinkellei, T. dionisii, T. erneyi*, *T. rangeli, T. conorhini* and *T. lewisi),* trypanosomes of snakes, crocodiles, toads, lizards and birds, trypanosomatids of several other genera, and bodonids and euglenids. Here, we describe PRAC repertoires of species and genotypes of trypanosomes, taxon-rich phylogeny of eukaryotic and prokaryotic PRAC homologs, GC contents, selection pressures on the evolution of *Try*PRAC, and genome synteny analyses. Together, the results allowed us to hypothesise about the origin, and number and timing of PRAC transference that gave rise to *Try*PRAC genes.

Methods

Trypanosome genomes used for searches of *Tc*PRAC homologous genes

Searches for *Tc*PRAC homologs were performed by BLAST against draft and annotated genomes of trypanosomatids freely available in TriTrypDB, geneDB and NCBI data banks. Sequences of *Tc*PRAC and *Tv*PRAC [10] were used as queries; full-length sequences and specific motifs from PRAC-like gene family were used as baits for the genome analyses of *T. cruzi* CL Brener, Silvio X10 plus other strains of *T. cruzi* (Table 1) sequenced by the Kinetoplastid Genome Sequencing and Analysis Consortium NIH/NHGRI/NIAID, *T. brucei* ssp., *T. evansi*, *T. congolense*, *T. vivax*, *T. c. marinkellei, T. grayi,* and species of *Leishmania* (Table 1). In addition, we examined the freely available genomes from *Crithidia acanthocephali, Angomonas desouzai*, *Angomonas deanei*, *Strigomonas culicis*, *Strigomonas oncopelti* and *Herpetomonas muscarum,* all generated in our laboratories, plus genomes from *Phytomonas* sp., *Crithidia fasciculata* and *Endotrypanum schaudinni* (Table 1).

We also searched for PRAC genes in genomes that have been generated in our laboratories for a large number of euglenozoans within the ATOL (Assembling the Tree of Life, NSF-USA) and TCC-USP (Brazil) projects aiming highly comprehensive phylogenomic inferences. The following ongoing genomes were analyzed: *T. cruzi* (G and Tcbat), *T. cruzi marinkellei* (TCC344), *T. dionisii* (TCC211), *T. erneyi* (TCC1946), *T. rangeli* (AM80), *T. lewisi* (TCC34), *T. conorhini* (TCC025E), *T. serpentis*

Table 1 Trypanosomes, other trypanosomatids and free living kinetoplastids and euglenids examined in this study, and respective sequences of *Tc*PRAC homologous genes

Species isolate (genotype)	Host species	Data bank assessed	*Try*PRACAccess number Genome/GenBank
Trypanosomes			
Trypanosoma cruzi			
Sylvio X10.6 (TcI)	*Homo sapiens*	TritrypDB &	TCSYLVIO_010607
JR cl4 (TcI)	*Homo sapiens*	Genome draft (WU) &	KP001304
G (TcI)	*Didelphis marsupialis*	Genome draft (ATOL) #	KP001302
Esmeraldo (TcII)	*Homo sapiens*	Genome draft (WU) &	KP001301
M6241 cl6 (TcIII)	*Homo sapiens*	Genome draft (WU) &	KP001305
Can III (TcIV)	*Homo sapiens*	Genome draft (WU) &	KP001298
CLBrener –Esm (TcVI)	*Triatoma infestans*	TritrypDB &	TcCLB.506795.80
CLBrener-Non-Esm (TcVI)	*Triatoma infestans*	TritrypDB &	TcCLB.509935.29
Tula cl2 (TcVI)	*Homo sapiens*	Genome draft (WU) &	KP001312
1994 (Tcbat)	*Myotis levis* (bat)	Genome draft (USP)	KP001313
Trypanosoma cruzi marinkellei B7	*Phyllostomus discolor* (bat)	TritrypDB &	Tc_MARK_8728
TCC344	*Carollia perspicillata* (bat)	Genome draft (ATOL) #	KP001314
Trypanosoma dionisii TCC211	*Eptesicus brasiliensis* (bat)	Genome draft (ATOL) #	KP001263
Trypanosoma erneyi TCC1946	*Mops condylurus* (bat)	Genome draft (ATOL) #	KP001315
Trypanosoma rangeli AM80	*Homo sapiens*	Genome draft (ATOL) #	KP001264
Trypanosoma conorhini TCC025	*Rattus rattus*	Genome draft (ATOL) #	KP001316
Trypanosoma lewisi TCC034	*Rattus rattus*	Genome draft (ATOL) #	KP001317
Trypanosoma vivax Y486	*Bos taurus*	TritrypDB &	TvY486_0703770
Trypanosoma b. brucei TREU927	*Glossina pallidipes*	TritrypDB &	
Trypanosoma b. gambiense DAL972	*Homo sapiens*	TritrypDB &	
Trypanosoma congolense IL3000	*Bos* sp.	TritrypDB &	
Trypanosoma serpentis TCC1052	*Pseudoboa nigra* (snake)	Genome draft (USP) #	KP001318
Trypanosoma grayi ANR4	*Glossina palpalis*	TritrypDB &	Tgr.146.1080
Trypanosoma sp. TCC339	*Rhinella marina* (toad)	Genome draft (USP) #	KP001319
Trypanosoma sp. TCC1825	*Ramphocelus nigrogularis* (bird)	Genome draft (USP) #	KP001320
Trypanosoma sp. TCC878	*Mabuya frenata* (lyzard)	Genome draft (USP) #	KP001321
Other Trypanosomatids			
Crithidia fasciculata	*Anopheles quadrimaculatus*	TritrypDB &	
Crithidia acanthocephali	*Acanthocephala femorata* (fly)	(ATOL) & GenBank AUXI01000000	
Leptomonas costaricensis	*Ricolla simillima* (Hemiptera)	(ATOL) #	
Leishmania major Friedlin	*Homo sapiens*	TritrypDB &	
Leishmania tarentolae	*Tarentola mauritanica* (lyzard)	TritrypDB &	
Endotrypanum schaudinni	*Choloepus hoffmani* (sloth)	Genome draft (ATOL) #	
Angomonas desouzai	*Ornidia obesa* (fly)	(ATOL) & GenBank AUXL01000000	
Angomonas deanei	*Zelus leucogrammus* (Hemiptera)	(ATOL) & GenBank AUXM01000000	
Strigomonas culicis	*Aedes vexans*	(ATOL) & GenBank AUXH01000000	
Strigomonas oncopelti	*Oncopeltus sp.* (Hemiptera)	(ATOL) & GenBank AUXK01000000	
Herpetomonas muscarum	*Musca domestica*	(ATOL) & GenBank AUXJ01000000	
Phytomonas sp.	*Jatropha macrantha* (plant)	Genome draft (ATOL) #	

Table 1 Trypanosomes, other trypanosomatids and free living kinetoplastids and euglenids examined in this study, and respective sequences of *Tc*PRAC homologous genes *(Continued)*

Free living euglenozoans	
Bodo sp. ATCC 50149	Genome draft (ATOL) #
Parabodo caudatus ATCC 30905	Genome draft (ATOL) #
Discoplastis spatirhyncha SAG1224.42	Genome draft (ATOL) #
Eutreptia viridis SAG 1226-1c	Genome draft (ATOL) #

TCC: Trypanosomatid Culture Collection of the University of São Paulo, SP, Brazil.
TritrypDB (http://tritrypdb.org).
WU: Washington University (USA) - Kinetoplastid Genome Sequencing and Analysis Consortium (NIH/NHGRI/NIAID).
ATOL: Assembling the Tree of Life (NSF-USA);
USP: Department of Parasitology, University of São Paulo, USP.
& publicly available genomes; # access to these ongoing genomes can be obtained by contacting the corresponding author.

(TCC1052), *Trypanosoma* sp. of toad (TCC339), *Trypanosoma* sp. of lizard (TCC878) and *Trypanosoma* sp. of bird (TCC1825). PRAC genes were also searched in draft genomes of bodonids (*Bodo* sp. and *Parabodo caudatus*), and euglenids (*Euglena gracilis*, *Eutrepia viridis*, *Discoplastis spathirhyncha*) (Table 1). The trypanosomatids employed for genome sequencing are cryopreserved at Trypanosomatid Culture Collection of the University of São Paulo (TCC-USP). PRAC sequences retrieved from the genomes were all deposited in GenBank (Table 1).

The draft genomes generated in our laboratories were sequenced using standard pyrosequencing shotgun methodology according to Roche 454 protocols and assembled by Roche's Newbler software (version 2.3) as previously described [32,33]. The ongoing genomes from trypanosomes of toad, snake, lizard and bird were obtained using the MiSeq Illumina plataform (mate-pair reads), and assembled using Newbler (version 2.9) as described [32]. Access to the unpublished draft and ongoing genomes analyzed in this paper can be obtained by contacting the corresponding author.

Essential motifs and residues, alignments, and phylogenetic analyses of PRAC sequences

Predicted amino-acid sequences from PRAC genes identified in the trypanosome genomes were evaluated regarding motifs essential for racemase activity to identify putative PRAC homologous enzymes, thus ensuring the selection of genes encoding racemases, and excluding closely related PRAC-like genes such as those coding for epimerases. Previous studies on *T. cruzi* PRAC enzymes and bacterial PRACs demonstrated that catalytic cysteines (Cys130 and Cys300), active site (SPCGT) and essential motifs (MCGH and MIII) are not sufficiently stringent to discriminate between PRAC-like enzymes such as hydroxyproline-2 epimerase (HyPRE) and racemase. Therefore, the residues R1, R2, and R3, which are involved in substrate specificity, were used to distinguish between PRAC from HyPRE enzymes *in silico* [7,8,10,16]. These features were examined to select the genes encoding putative homologous PRAC enzymes in trypanosome genomes.

Amino-acid and nucleotide sequences of whole *Tc*PRAC-homologous genes (~1062 bp) from the various trypanosome species were obtained from genome data banks, aligned using Clustal X v2.0 and manually adjusted. In addition, partial (~1015 bp) PRAC nucleotide sequences obtained by PCR-sequencing were used for polymorphism analysis within *T. cruzi* by comparing sequences from TcI-TcVI and Tcbat isolates. An alignment was created with partial PRAC amino acid sequences from *T. rangeli* isolates of lineages A-E and *T. conorhini* as outgroup.

Maximum-likelihood (ML) and maximum-parsimony (MP) analyses were performed respectively with RAxML v7.2.8 and PAUP*v4b10 based on nucleotide and amino acid alignments. The MP tree search and bootstrap analysis were done using 500 replicates of random addition sequence swapped using TBR. The ML analysis employed GTRGAMMAI with 500 maximum parsimony starting trees. Model parameters of ML analysis were estimated over the tree search and bootstrap support was estimated with 1000 replicates in RAxML using maximum parsimony as starting trees and optimized in the best tree as previously described [38-43].

To compare the highly conserved *Tc*PRAC genes from all DTUs, a network genealogy was inferred using nucleotide sequences and the neighbor-net method with Kimura's 2-parameter model implemented in SplitsTree4 V4.10 as described previously [28,44]. Internode support was estimated by performing 100 bootstrap replicates using the same parameters optimized for network inferences.

PCR amplification and sequencing of PRAC gene sequences from *T. cruzi* and *T. rangeli* isolates

PCR amplification of partial *Tc*PRAC sequences (~1015 bp comprising all essential motifs and residues of *Tc*PRAC enzyme) from large number of isolates from *T. cruzi* and *T. rangeli* (Additional file 1) was conducted as previously described [8] using the primers PRAC1 (5'-CTTCCCATGG GGCAGGAAAAGCTTCTG-3') and PRAC2 (5'-CTGA GCTCGACCAGATCTATCTGC-3'). The PCR-amplified products were cloned, and 3–5 clones from each isolate

were sequenced, whereas ~10 clones were sequenced from each of the hybrid isolates. The PRAC sequences representing the genetic diversity within *T. cruzi* and *T. rangeli* were deposited in GenBank and the access numbers are listed in Additional file 2.

Phylogenetic analysis based on gGAPDH gene sequences

Phylogenetic tree of Euglenozoa species based on gGAPDH gene sequences was inferred by ML and MP as described above for PRAC genes. The alignment created for this analysis was done using for guidance a comprehensive alignment of kinetoplastid gGAPDH genes [3] and included sequences from 32 trypanosome species, non-trypanosome trypanosomatids of seven genera, and five bodonids and euglenids as outgroups. Bootstrap support was estimated with 100 pseudoreplicates in RAxML using GTRGAMMA. The Genbank access numbers of all gGAPDH genes included in the phylogenetic trees are listed in Additional file 1.

Trypanosomes lacking PRAC genes as determined by genome search and/or PCR amplification

The absence of PRAC genes in the genomes of *T. brucei* ssp. and *T. evansi* was confirmed by negative results in PCR tests of additional isolates of each species. Besides the lack of PRAC homologs in the genome of *T. congolense* IL3000 (subgroup Savannah), results were also negative for all other members of *Nannomonas* tested: *T. congolense* Cam22 (Forest), WG5 (Kilifi) and TREU1475 (Savannah), *T. simiae* and *T. godfreyi* [28]. DNA samples from these trypanosomes were kindly provided by Wendy Gibson, Bristol University, UK.

Horizontal-gene-transfer analysis

The horizontal-gene-transfer (HGT) analysis includes a comprehensive dataset of 2,530 PRAC-like protein sequences from prokaryotes and eukaryotes in the non redundant (NR) database. A BLASTp search was performed with a maximum-expected-value threshold of 1e-20, using the *Tc*PRACA and *Tc*PRACB sequences as queries. The retrieved sequences were checked for PRAC-like domains using the Batch search tool in the Conserved Domain Database (http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi). Multiple-sequence alignment was performed using MUSCLE v3.8, and edited using Gblocks v0.91b [45] to eliminate poorly aligned positions. The final phylogenetic tree was obtained by ML analysis with 2,530 sequences under the WAG substitution model with gamma-distributed heterogeneity rate categories, and estimated empirical residue frequencies (model PROTGAMMAWAG) as implemented in RAxMLv7.2.8. One hundred different best tree searches were performed, and the tree with best likelihood found was kept. RAxML rapid bootstrap was performed with 100 pseudoreplicates. The tree was also visualized using Dendroscope v3.2.4 [46] with further cosmetic adjustments done using the Inkscape vector image editor (http://inkscape.org). To better resolve and visualize the putative HGT donor lineages, a ML analysis was performed using a subset of 303 PRAC-like sequences from NR database and 39 *Try*PRAC aligned with the nearest neighbor taxons identified in the analysis using 2,530 sequences. The tree search and bootstrap were conducted using the same parameters for both datasets.

Genomic organization, GC content and codon pressure analyses of trypanosome PRAC genes

The comparison of PRAC genomic organization in the analyzed genomes was performed with the bl2seq BLASTX algorithm using the flanking downstream and upstream regions (~10,000 bp) previously reported for *T. cruzi* and *T. vivax* [8,9] in all trypanosome genomes investigated in this study. Codon-selection analysis was performed using the HyPhy v2.2 package [47] with a threshold p-value < 0.05. The GC content comparison between *Try*PRAC homologs and both flanking genes and whole genomes were conducted using the mfsizes v. 1.8.3 software (http://sourceforge.net/projects/mfsizes/).

Results and discussion

Analysis of kinetoplastid and euglenid genomes shows *Tc*PRAC homologs exclusively in *Trypanosoma*

We searched for *Tc*PRAC homologs in the genomes of trypanosomatids, bodonids and euglenids using *Tc*PRACA and *Tc*PRACB sequences as queries. Homologs were identified in *T. cruzi*, *T. c. marinkellei*, *T. dionisii*, *T. erneyi*, *T. rangeli*, *T. conorhini* and *T. lewisi*. In addition to these mammalian parasites, *Tc*PRAC homologs were found in trypanosomes from snake (*T. serpentis*) [42], crocodile [43,48], lizard (TCC878), bird (TCC1825) and toad (TCC339) [41] (Table 1, Figure 1). All these species exhibit a single copy of a *Try*PRAC homolog per haploid genome, and no other PRAC-like gene was identified in the kinetoplastid genomes.

Previous studies reported the absence of PRAC homologs in the genomes of *T. b. brucei* and *T. b. gambiense,* while only a remnant of a PRAC gene was suggested for *T. congolense* [8-10]. Thus, it was not surprising that *T. evansi*, which is highly closely related to *T. brucei* ssp. and *T. equiperdum*, all forming the subgenus *Trypanozoon*, also lacked PRAC. In addition, attempts to detect even small fragments of PRAC genes by PCR amplification failed for all investigated species of the subgenus *Nannomonas*: *T. congolense* of subgroups Forest, Kilifi and Savannah, *T. simiae* and *T. godfreyi.* We were also unable to detect any PRAC fragment in the recently assembled genome of *T. congolense* (Savannah). In the phylogenetic trees, the species of *Trypanozoon* and *Nannomonas* formed

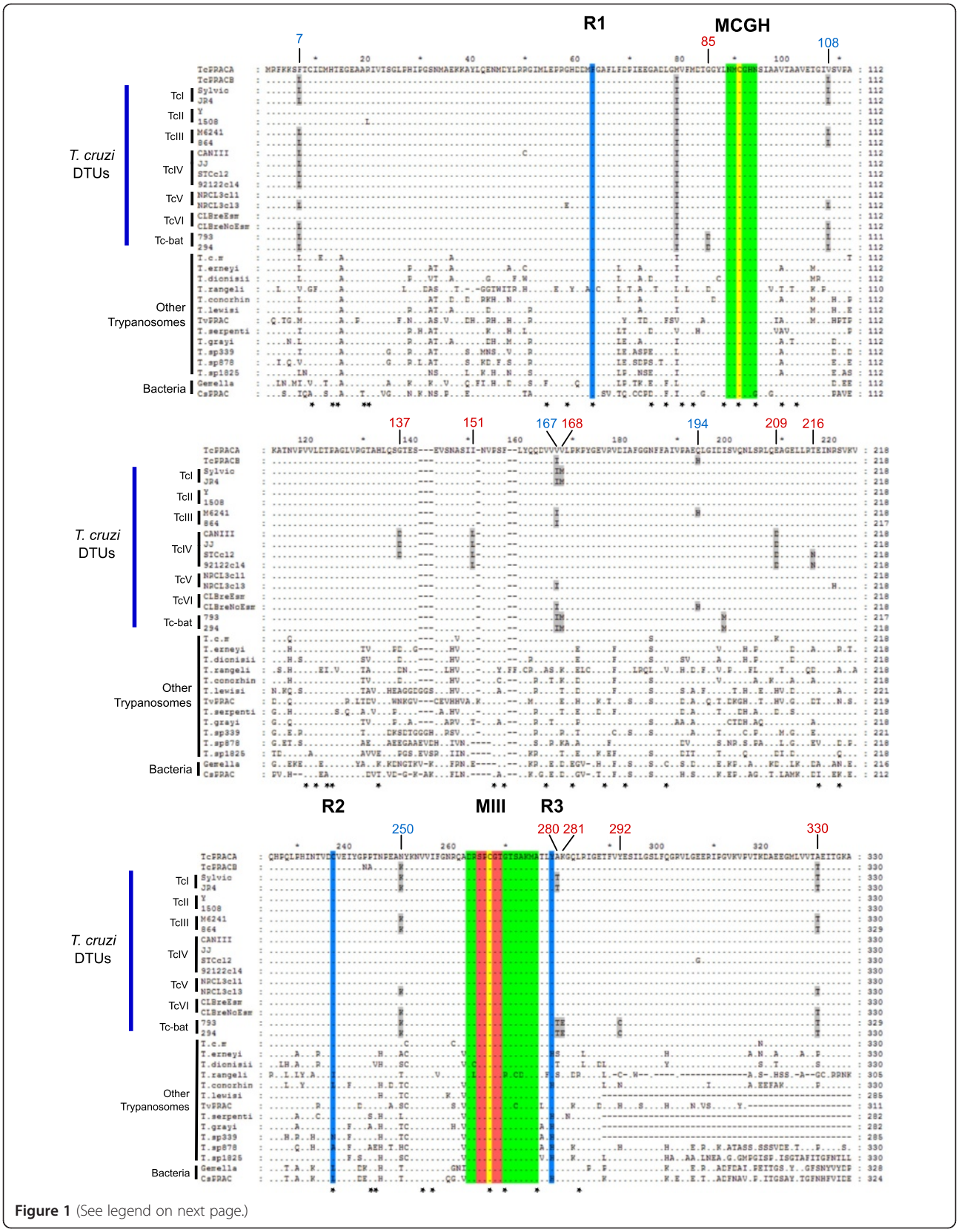

Figure 1 (See legend on next page.)

(See figure on previous page.)
Figure 1 Alignment of predicted amino acid sequences of proline racemase (PRAC) homologous genes from *T. cruzi* (DTUs TcI-TcVI and Tcbat), *T. cruzi marinkellei* (*T. c. m*), *T. erneyi*, *T. dionisii*, *T. rangeli*, *T. conorhini*, *T. lewisi*, *Tv*PRAC - *T. vivax*, *T. serpentis*, *T. grayi*, *T. sp*. from toad (TCC339), *T.* sp. from lizard (TCC878), and *T.* sp. from bird (TCC1825) and PRAC from *Gemella haemolysans and Clostridium difficile* (CsPRAC). Essential motifs (MCGH and MIII) are in green, and the active site (SPCGT) in red. R1, R2 and R3 are residues involved in substrate specificity. Cys91 and Cys267 are the catalytic cysteines. Blue numbers indicate differences between *Tc*PRACA and *Tc*PRACB, and red numbers indicate substitutions found in newly identified *Try*PRAC homologs. Black stars indicate negatively selected amino acid residues.

a monophyletic assemblage within the main *T. brucei* clade in which *T. vivax* has a basal position [3].

Examination of non-trypanosome Trypanosomatidae species revealed that PRAC genes, and even other genes of the PRAC-like family, are absent not only from the *L. major* genome as previously reported [8-10,31], but also from other *Leishmania* species. Moreover, the genomes of monoxenous parasites of insects of the genera *Crithidia, Leptomonas*, *Angomonas*, *Strigomonas* and *Herpetomonas* and the plant parasites of the genus *Phytomonas* all lacked PRAC. Regarding other kinetoplastids, our searches did not reveal any putative PRAC-like genes in *Bodo* sp. and *Parabodo caudatus.* In addition, we did not identify PRAC-like genes in the genomes of the basal species within Euglenozoa: *Euglena gracilis*, *Eutrepia viridis* and *Discoplastis spathirhyncha* (Figure 2B).

In conclusion, PRAC homologs are widespread in trypanosomes but absent from the genomes of non-trypanosome trypanosomatids of all genera investigated, and in the bodonids and euglenids to date examined. PRAC homologs were identified in species of the main clades of the *Trypanosoma* phylogenetic tree, including the basal Aquatic clade [3,41], but in the *T. brucei* clade only *T. vivax* carries a homologous PRAC gene. Results strongly suggest that a prokaryotic PRAC was introduced at the root of *Trypanosoma* and fixed in the genome of a common trypanosome ancestor.

Molecular characterisation of new putative PRAC-like enzymes of trypanosomes

The catalytic mechanism of *Tc*PRAC is essentially identical to that of the prokaryotic PRAC enzymes. The activity of PRAC enzymes depends mainly on two cysteine residues that transfer protons to the chiral carbon (C^{α}) of L-proline/D-proline enantiomers, resulting in the stereoinversion of its configuration [7,10,16]. PRAC-like genes include diverse racemase-like genes that exhibit strong sequence similarity to proline racemases. A few eukaryotic PRAC-like enzymes have all the residues critical for racemase activity. Most of these enzymes function as proline epimerases, which are common in prokaryotes or as proline dehydratases as reported in humans [16,17].

Alignment of *Try*PRAC homologs from 13 trypanosome species, using *Tc*PRACs and prokaryotic PRAC sequences for guidance, revealed the two cysteine residues, the active site SPCGT, and the MCGH motif in all sequences. Only the MIII motif had relevant polymorphism. At the residues RI, R2, and R3, which are involved in substrate specificity, all *Try*PRAC homologs have a conserved R1 but a variable R2 and R3. All species of the subgenus *Schizotrypanum* (*T. cruzi, T. c. marinkellei*, *T. dionisii* and *T. erneyi*) share identical motifs, and highly conserved essential residues. Within this clade, only *T. erneyi* showed one non-synonymous substitution at R3. In contrast, more distantly related trypanosomes showed synonymous and non-synonymous substitutions at R2 (*T. rangeli, T. conorhini*, *T.* sp. from toad and *T.* sp. from lizard) and R3 (*T. conorhini*, *T. serpentis, T. grayi* and *T.* sp. from toad) (Figure 1). The implications of these substitutions in substrate specificity merit further investigation.

All putative *Try*PRAC sequences lacked a signal peptide, suggesting that the encoded enzymes are intracellular, and can be released through the flagellar pocket and/or parasite death [8,10,18]. Despite highly conserved catalytic domains, we identified at least one novel PRAC homolog for each species of trypanosome. *In silico* analysis suggested that most trypanosomes can express PRAC proteins with racemase activities (Figure 1). Homologous PRAC of *T. vivax* differed in several residues when compared to those from *T. cruzi*. Although the PRAC genes from trypanosomes of non-mammalian hosts such as snakes (*T. serpentis*), crocodiles (*T. grayi*), toads, lizards and birds differed in several residues when compared to homologs of *Schizotrypanum* species, all sequences could be aligned with confidence with both *Tc*PRAC and prokaryotic PRAC genes (Figure 1).

Phylogenetic relationships of PRAC homologs from 13 trypanosome species agree with the currently recognised phylogeny of *Trypanosoma*

T. cruzi is highly closely related to all other species of the subgenus *Schizotrypanum* (*T. c. marinkellei*, *T. dionisii* and *T. erneyi*), which are all called *T. cruzi*-like because they share morphology of blood and culture forms, although they differ in hosts, vectors and pathogenicity. Development as amastigotes and differentiation into trypomastigotes within mammalian cells *in vitro* is a unique feature of *Schizotrypanum*, whereas *in vivo,* only

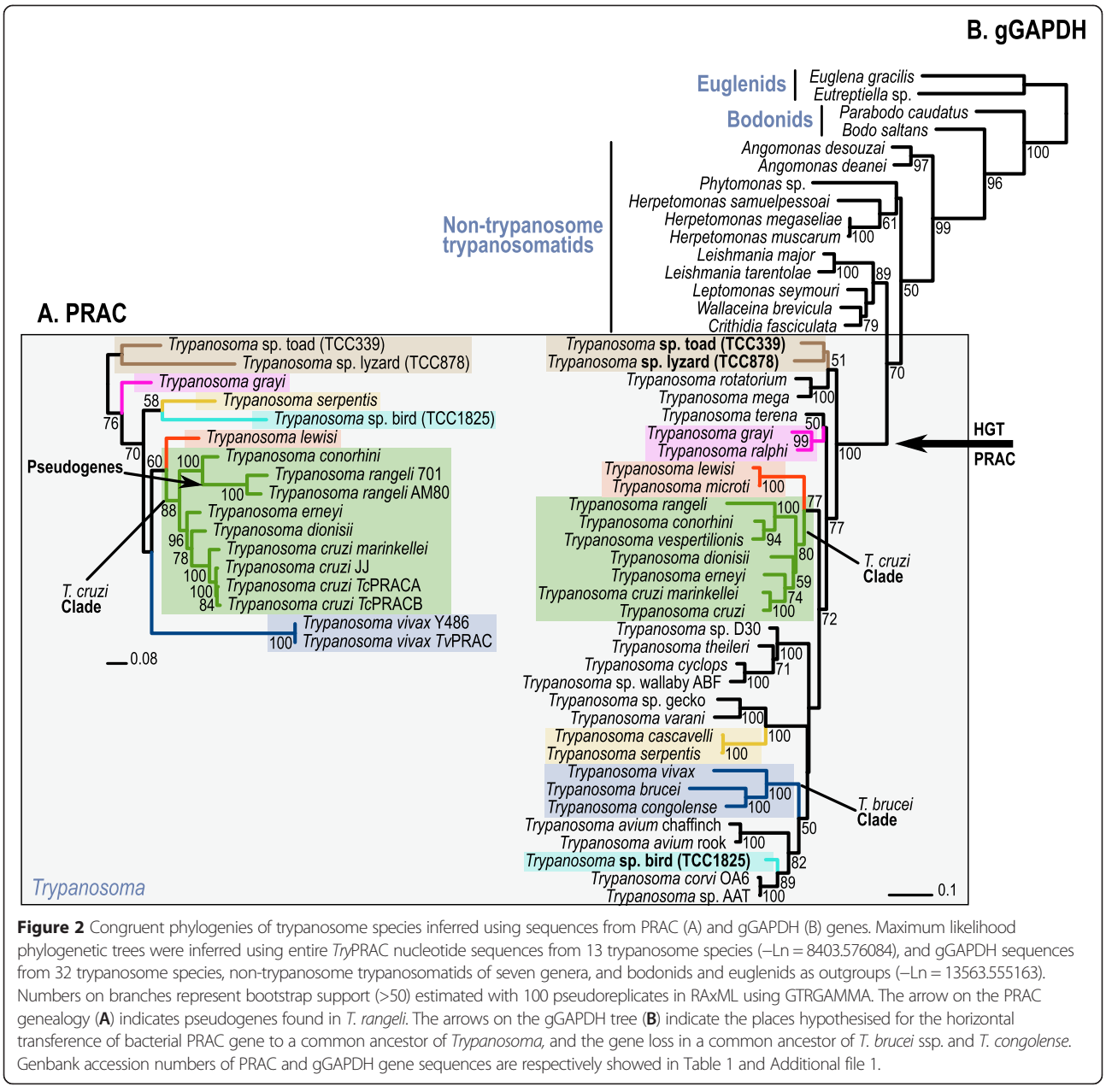

Figure 2 Congruent phylogenies of trypanosome species inferred using sequences from PRAC (A) and gGAPDH (B) genes. Maximum likelihood phylogenetic trees were inferred using entire *Try*PRAC nucleotide sequences from 13 trypanosome species (–Ln = 8403.576084), and gGAPDH sequences from 32 trypanosome species, non-trypanosome trypanosomatids of seven genera, and bodonids and euglenids as outgroups (–Ln = 13563.555163). Numbers on branches represent bootstrap support (>50) estimated with 100 pseudoreplicates in RAxML using GTRGAMMA. The arrow on the PRAC genealogy (**A**) indicates pseudogenes found in *T. rangeli*. The arrows on the gGAPDH tree (**B**) indicate the places hypothesised for the horizontal transference of bacterial PRAC gene to a common ancestor of *Trypanosoma,* and the gene loss in a common ancestor of *T. brucei* ssp. and *T. congolense*. Genbank accession numbers of PRAC and gGAPDH gene sequences are respectively showed in Table 1 and Additional file 1.

T. cruzi infects mammals other than bats. Nevertheless, as in *T. cruzi* infection, nests of amastigotes in cardiac cells can be found in bats infected with *T. cruzi*-like species. *T. cruzi* is transmitted by triatomines, while cimicids are vectors of *T. dionisii.* These trypanosomes share development restricted to the vector guts [5,38-40,44].

For phylogenetic inferences within *Schizotrypanum,* we compared isolates of all species mentioned above. According to strongly supported branching patterns on both PRAC (Figure 2A) and gGAPDH (Figure 2B) phylogenetic trees, all species clustered tightly, forming a monophyletic assemblage of trypanosomes. The relationships among the *Schizotrypanum* species and the DTUs of *T. cruzi* were inferred using entire PRAC amino acid sequences. Results corroborated the clustering of sequences according to species (Additional file 3). In agreement with previous analysis of several other genes [49] such as gGAPDH and cathepsin L-like genes, *Try*PRAC genealogy and nucleotide sequence divergences confirmed *T. c. marinkellei* as the closest relative of *T. cruzi* (~7.5 of *Try*PRAC sequence divergence between the two species). This species was followed by *T. erneyi* from African bats (~14%) and *T. dionisii* (14.5%) from Old World bats. Large genetic distances separated *T. cruzi* and *T. rangeli* (~30% PRAC sequence divergence), and *T. cruzi* and *T. vivax* (~38%) PRAC

sequences. Compared to divergence among *Try*PRAC sequences, the species of the subgenus *Schizotrypanum* were separated by much smaller gGAPDH sequence divergences (~8.0% between *T. cruzi* and *T. dionisii*), whereas ~15% and ~17.5% of gGAPDH sequence divergence separated *T. cruzi* from *T. rangeli* and *T. vivax*, respectively. Divergences of gGAPDH and the more conserved SSU rRNA genes, which are the traditional genes employed for phylogenetic inferences of the Trypanosomatidae, were previously reported for the trypanosome species included in PRAC phylogeny [5,39,40].

In the bat-seeding hypothesis for the origin of the *T. cruzi* clade [39], a scenario has been proposed in which ancestral trypanosomes of bats evolved exclusively in Chiroptera, giving rise to the bat-restricted species [38-40], or evolved through multiple independent host jumps, giving rise to species infecting other mammals (such as rats, civets and monkeys) in the Old World, and to the generalists *T. cruzi* and *T. rangeli,* which are species infective to bats plus a broad range of other mammals, including human and non-human primates in the New World [3,5,38-40,50-53]. Also in agreement with previous studies, in the PRAC genealogy *T. rangeli* and *T. conorhini* formed the sister group of *Schizotrypanum,* together constituting the clade *T. cruzi* that also harbours other trypanosomes, mostly from bats [39,40]. *T. lewisi,* the basal species of the clade *T. cruzi*, is a non-pathogenic and cosmopolitan trypanosome of domestic rats transmitted by fleas, which can opportunistically infect immune-compromised human and non-human primates [54] (Figures 2A, 2B).

The toad and lizard trypanosomes included in PRAC genealogy represented the basal branches of the *Trypanosoma* gGAPDH phylogenetic trees showing PRAC genes highly divergent from *Tc*PRAC (Figure 2A). Previous [3,41-43,48] and herein inferred phylogenies based on gGAPDH genes (Figure 2B) demonstrated that anuran trypanosomes nested into the so-called "Aquatic clade", which also includes trypanosomes transmitted by aquatic leeches of fishes, turtles and platypus, besides a lizard trypanosome of unknown vector. The aquatic clade was strongly supported as the most basal of *Trypanosoma* [3]. Also concordant with previous phylogenies, PRAC genes of trypanosomes from snakes, birds and crocodiles, which are transmitted by insects, all clustered into the "terrestrial" clade that also includes PRAC sequences from all trypanosomes of mammals (Figure 2A) in agreement with previous SSU rRNA and gGAPDH phylogenies [3,41-43,48].

Repertoires and phylogenetic relationships among *T. cruzi* PRAC homologs of all DTUs and Tcbat

Comparison of whole *Try*PRAC amino acid sequences revealed relevant polymorphisms (~3.0% sequence divergence) within *T. cruzi*. Aiming an intra-specific analysis of *Tc*PRACs, we compared entire amino acid sequences from *T. cruzi* Sylvio X10.6, JRcl4 and G (TcI), Esmeraldo cl3 (TcII), M6241cl6 (TcIII), CANIII (TcIV), CL Brener (TcVI), Tula (TcVI), and Tcbat. Unlike all other isolates, which carried a single PRAC gene, *T. cruzi* CL Brener exhibited *Tc*PRACA and *Tc*PRACB [10], found in this work in the Esmeraldo-like and non-Esmeraldo-like haplotypes, respectively.

We evaluated all signatures defined for *Tc*PRAC activity, and polymorphisms used to differentiate between *Tc*PRACA and *Tc*PRACB. A leucine at position seven (typical of *Tc*PRACB) was found in TcI, TcIII, TcIV and Tcbat. The phenylalanine at this position that had been reported to be specific to *Tc*PRACA was found in Y and TCC1508 (TcII), but not in Esmeraldo (TcII). *Tc*PRACs from all DTUs have isoleucine at position 79 (like *Tc*PRACB), while methionine at this position in *Tc*PRACA was exclusive to CL Brener. Like *Tc*PRACA, TcII and TcIV had valine at positions 108 and 167 and asparagine at position 250. At these positions, TcI, TcIII and Tcbat had leucine, isoleucine and lysine, respectively, like *Tc*PRACB. Like the hybrid CL Brener (TcVI), both *Tc*PRACA and *Tc*PRACB were identified in the hybrid NRCL3 (TcV). New polymorphic residues evidenced novel *Tc*PRAC homologs defining TcI-TcIV-specific profiles while TcV and TcVI can be identified by the presence of both *Tc*PRACA and *Tc*PRACB. Polymorphic amino acids defining each *T. cruzi* DTU are showed in Additional file 3.

Due to the high sequence conservation throughout the *Tc*PRAC genes from *T. cruzi* of some DTUs, phylogenetic analyses based on amino acid sequences were unable to clearly resolve the closely related DTUs (Additional file 3 shows the network of *Tc*PRAC amino acid sequences). To assess the relationships within *T. cruzi* using the conserved *Tc*PRAC genes (63 polymorphic sites), we constructed a network using partial nucleotide sequences obtained by PCR-sequencing from 68 isolates previously genotyped [21,44,49]. The network clearly evidenced subclades corresponding to each TcI, Tcbat, TcII, TcIII, and TcIV DTUs. Sequences from TcV and TcVI clustered with TcII or TcIII, in agreement with their hybrid origin, forming a reticulate pattern in the network. The network confirmed TcI closest to Tcbat and TcII more related to TcIV and, in addition, corroborated the heterogeneity intra-DTUs TcI, TcIII, and TcIV (Figure 3).

Most previous phylogenetic analyses within *T. cruzi* lacked isolates of all DTUs (especially TcIV and Tcbat) and/or *T. cruzi*-like outgroups and, then, were insufficient to resolve intra-*T. cruzi* phylogenetic relationships. Here, *Tc*PRAC genealogy using *T. cruzi*-like species agreed with the relationships among the DTUs (including Tcbat) previously inferred using cruzipain, SSU rRNA, and cytochrome b sequences [44,49]. Diverse

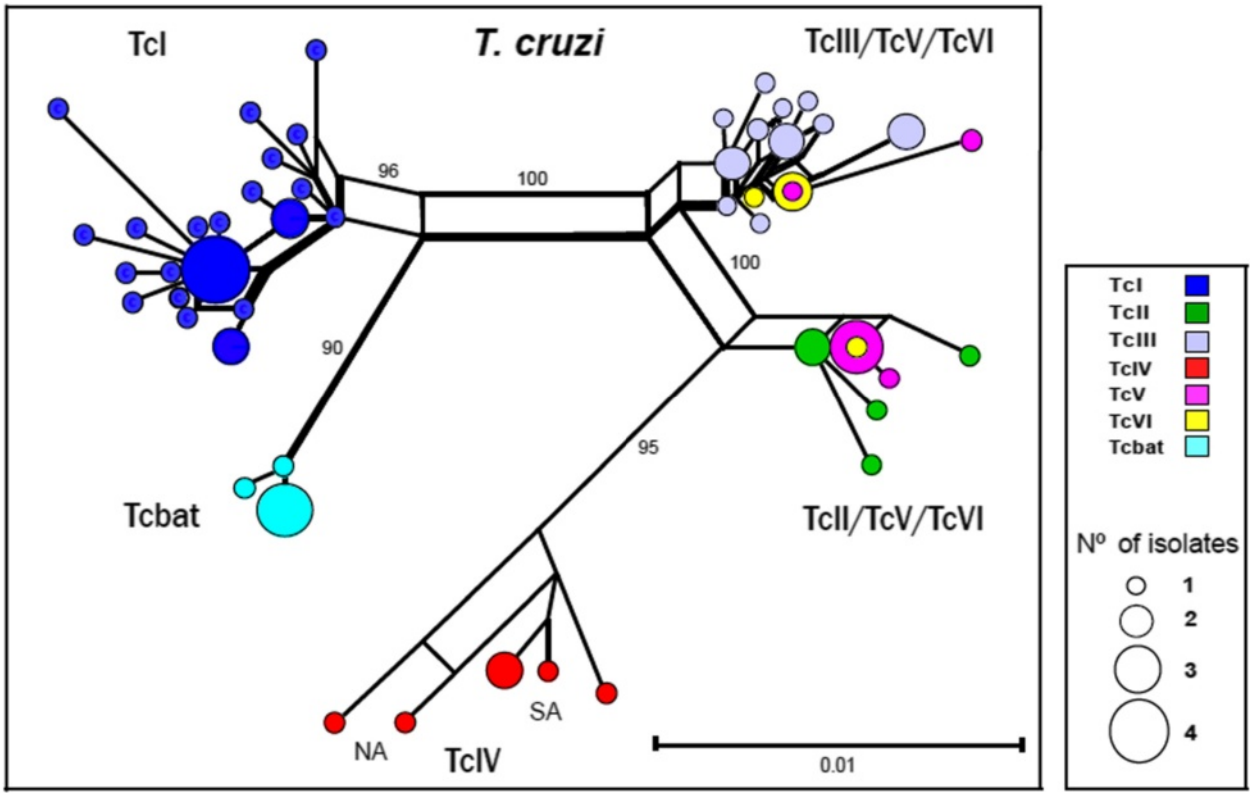

Figure 3 Network genealogy of partial PRAC nucleotide sequences from T. cruzi isolates showing the clustering of sequences according to DTUs, and sequences from the hybrids TcV and TcV positioned within both TcII and TcIII clades. The network branching pattern indicated high homogeneity of PRAC genes from isolates of Tcbat, TcII and TcIII contrasting with the large heterogeneity in TcI and TcIV isolates. Each colour represents one DTU, and the size of the circles indicates the number of isolates in the network. Genbank accession numbers of *Tc*PRAC genes are listed in Additional file 2, and the DTU-specific profiles of polymorphic sites is showed in Additional file 3.

genes have been employed as markers for *T. cruzi* genotyping [21]. We demonstrated that polymorphisms of *Tc*PRAC sequences allow the genotyping of all DTUs including hybrid genotypes, and are also valuable to infer inter-DTU relationships. To our knowledge, *Tc*PRAC is the first horizontally transferred gene (non-mitochondrial) characterized with these purposes.

Taking into account that *Tc*PRAC-A and TcPRAC-B participate in *T. cruzi* development in vertebrates and vectors and have been incriminated as host defence factors [9-12,19,22,55], it is tempting to speculate whether DTU-specific *Tc*PRAC enzymes can contribute to differential degrees of metacyclogenesis, parasitemias and virulence. *T. cruzi* of different DTUs interacts differently with the host, induces distinct immune responses and infections ranging from highly lethal to virtually asymptomatic, contributing to variable clinical forms of Chagas disease. Strains of TcI, the most widespread DTU in Latin American sylvatic cycles, exhibit high levels of metacyclogenesis. TcI is responsible for human outbreaks of oral infection and severe cardiomyopathies in Central America and Northern South America. Isolates of TcI are highly diverse genetically and in terms of virulence to mice. Although some isolates induced very low parasitemia and no mortality, it was suggested that Col strain (TcI) can evade the host immune response remaining unnoticed by mononuclear cells allowing rapid multiplication during acute infection [21,50-53,56-60]. Future studies are required to evaluate the roles played by TcI-specific *Tc*PRAC in low and highly virulent strains. TcII (PRAC-A), TcV and TcVI (PRAC-A and PRAC-B) are virulent to mice, induce high parasitemias and mortality, and have been associated with both cardiac and digestive forms in humans in Southern Cone countries [21,61]. TcIII strains (PRAC-B), found in Brazil and neighbouring countries, can induce important parasitemia and pathology in mice [57,62]. TcIV (unique *Tc*PRAC) is sylvatic, orally infects humans in Brazil and Venezuela, and induces low or moderate parasitemia and mortality in mice [21,53,60]. Tcbat, found in South and Central American bats and, apparently, able to infect humans [63], is not virulent to mice inducing extremely low parasitemias and no mortality, and is unable to develop in the commonest triatomine vectors of TcI-TcVI [50].

Trypanosoma rangeli pseudogenes diverged to be lineage-specific and more closely related to *T. conorhini* than to *T. cruzi* homologous PRAC

T. rangeli is a non-pathogenic parasite of humans and domestic and wild animals in Central and South America. This species is thought to be restricted to the bloodstream and survives host defences for months or years by unknown mechanisms. *T. cruzi* and *T. rangeli* are the only agents of human trypanosomosis in the Americas, sharing mammalian hosts and vectors in overlapping areas. *T. rangeli* overcomes the defences of the vector (*Rhodnius* spp.), multiplying in the gut and invading the haemolymph, where the parasites multiply outside and inside of haemocytes before reaching the salivary glands where metacyclogenesis takes place. This species differs from *T. cruzi*, which develops exclusively in the triatomine gut, and from *T. brucei*, which reaches the salivary glands of the vector (tsetse flies) from the proboscid [4,64,65].

Closer phylogenetic relationships of *T. rangeli* to *T. cruzi* than to *T. brucei* were strongly supported by comprehensive phylogenetic analysis based on diverse genes [3,5,38-40]. In addition, phylogenies based on PRAC (Figure 2A), gGAPDH (Figure 2B) and SSU rRNA genes have supporting *T. rangeli* more closely related to *T. conorhini* than to *T. cruzi* [3,38,40]. *T. conorhini* is a tropicopolitan species common in rats and transmitted by the also tropicopolitan *Triatoma rubrofasciata* [66,67]. This species shares features with both *T. cruzi* (development restricted to the gut of its triatomine vector) and *T. rangeli* (lack of both intracellular stages and pathogenicity to vertebrates). We are currently comparing the genomes of *T. conorhini, T. rangeli* and *T. cruzi* to better understand their relationships.

In contrast to predicted PRAC proteins in most trypanosomes including *T. conorhini*, which are compatible with the expression of racemases, all *T. rangeli* PRAC sequences were found disrupted by internal stop codons resulting in pseudogenes. This finding was confirmed in the PRAC found in the genome of *T. rangeli* AM80 (human isolate of basal lineage TrB from the Amazon region) and sequences from several isolates of all lineages determined by PCR-sequencing. Additional file 4 shows the alignment of *T. rangeli* PRAC pseudogenes.

We compared *T. rangeli* PRAC pseudogenes from 17 isolates of lineages TrA-TrE, all previously genotyped using other markers [49,50]. Previous phylogeographical studies suggest evolution within *T. rangeli* shaped by the coexistence of parasites with sympatric species of *Rhodnius* [4,51,52,64,65]. In this and previous studies based on SL, SSU rRNA, ITS rDNA, gGAPDH, and CATL sequences, TrB was always placed as the basal lineage of *T. rangeli*, whereas the relationships among the closely related TrA, C, D and E were far from resolved (Figure 4). In addition, an increasing genetic diversity within *T. rangeli* has recently been revealed [51,52,63,64]. Phylogenetic studies of all lineages using multilocus approaches are essential to better resolve the complex relationships among the lineages and to hypothesise about the evolutionary history of *T. rangeli*.

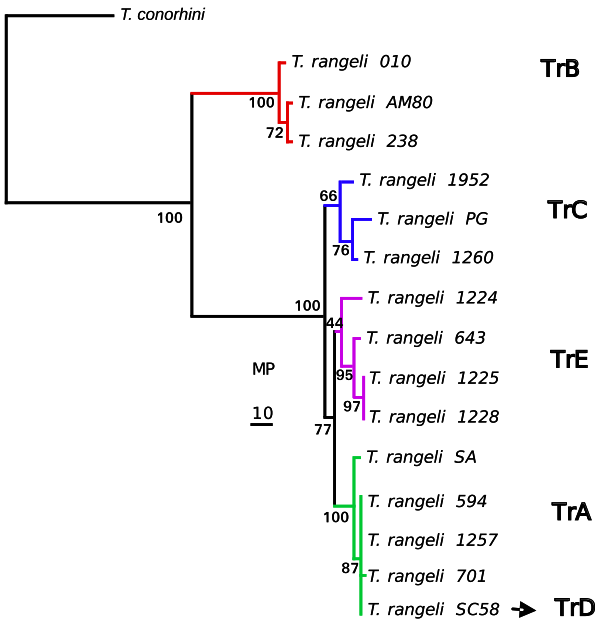

Figure 4 Phylogenetic inferences (MP) based on PRAC pseudogenes from 17 T. rangeli isolates using T. conorhini as outgroup. The branching pattern showing clades corresponding to the phylogenetic lineages TrA-TrE indicates a close phylogenetic relationship among the lineages TrA, C, D, E with TrB from Amazonia as the basal lineage of *T. rangeli*. Genbank accession numbers of PRAC genes included in this analysis are in
Additional file 2. The alignment of *T. rangeli* pseudogenes is showed in Additional file 4.

Phylogeny and pattern of presence/absence of PRAC homologs support a single HGT from a bacterium to an ancestor of *Trypanosoma*

The identification of bacterial PRAC homologs in *T. cruzi* and *T. vivax* and their absence from *T. brucei* and *T. congolense* [7-11] suggested a complex evolution of PRAC genes in trypanosomes. Here, a broad taxon sampling comprising 15 trypanosome species, trypanosomatids of 9 additional genera and five free-living ancestors of bodonids and euglenids provides relevant insights into this process at long and short timescales. The relative timing of the HGT event was investigated by searching the presence/absence of PRAC-like genes in the increased and broad taxon sampling. No PRAC-like genes were found in trypanosomes besides those encoding putative proline racemase enzymes, such as highly similar genes coding for epimerase and dehydratase found in prokaryotes and other eukaryotes such as fungi and metazoans. Although PRAC-like genes detected in fungi

and trypanosomes were all likely of prokaryotic origin, previous phylogenetic analyses revealed a polyphyletic pattern, indicating that they originated from different bacterial sources through independent HGT events [17,18]. Here, no PRAC-like genes were detected in the genomes of euglenids, bodonids or non-trypanosome trypanosomatids.

After identifying in our taxon-rich phylogenetic analysis of 2,530 eukaryotic (including trypanosome) and prokaryotic PRAC-like genes (Figure 5A) the general vicinity of *Try*PRAC sequences (the names and grouping of all organisms can be found in the rectangular phylogram in Additional file 5), our targeted analysis of 342 genes showed all *Try*PRAC homologs in a strongly supported clade exclusive of trypanosome sequences, evidencing their common ancestry (Figure 5B). The complete list of putative donor lineages currently available on NCBI NR protein database included in the phylogenetic analysis is presented in Additional file 6.

Considering the presence/absence of PRAC genes and the congruence of all species in phylogenies based on PRAC and gGAPDH genes (Figure 2), we hypothesised gains and losses of PRAC genes during trypanosome evolution. In the most parsimonious evolutionary scenario, one prokaryotic PRAC gene was transferred through a single HGT to the root of the genus *Trypanosoma* in a common ancestor of this genus, which gave origin to all extant trypanosome species with respective *Try*PRAC homologs. The absence in *T. brucei* and *T. congolense* suggested the loss of PRAC gene by a common ancestor of the subclade comprising these species (Figure 2B). In an unlikely, less parsimonious scenario,

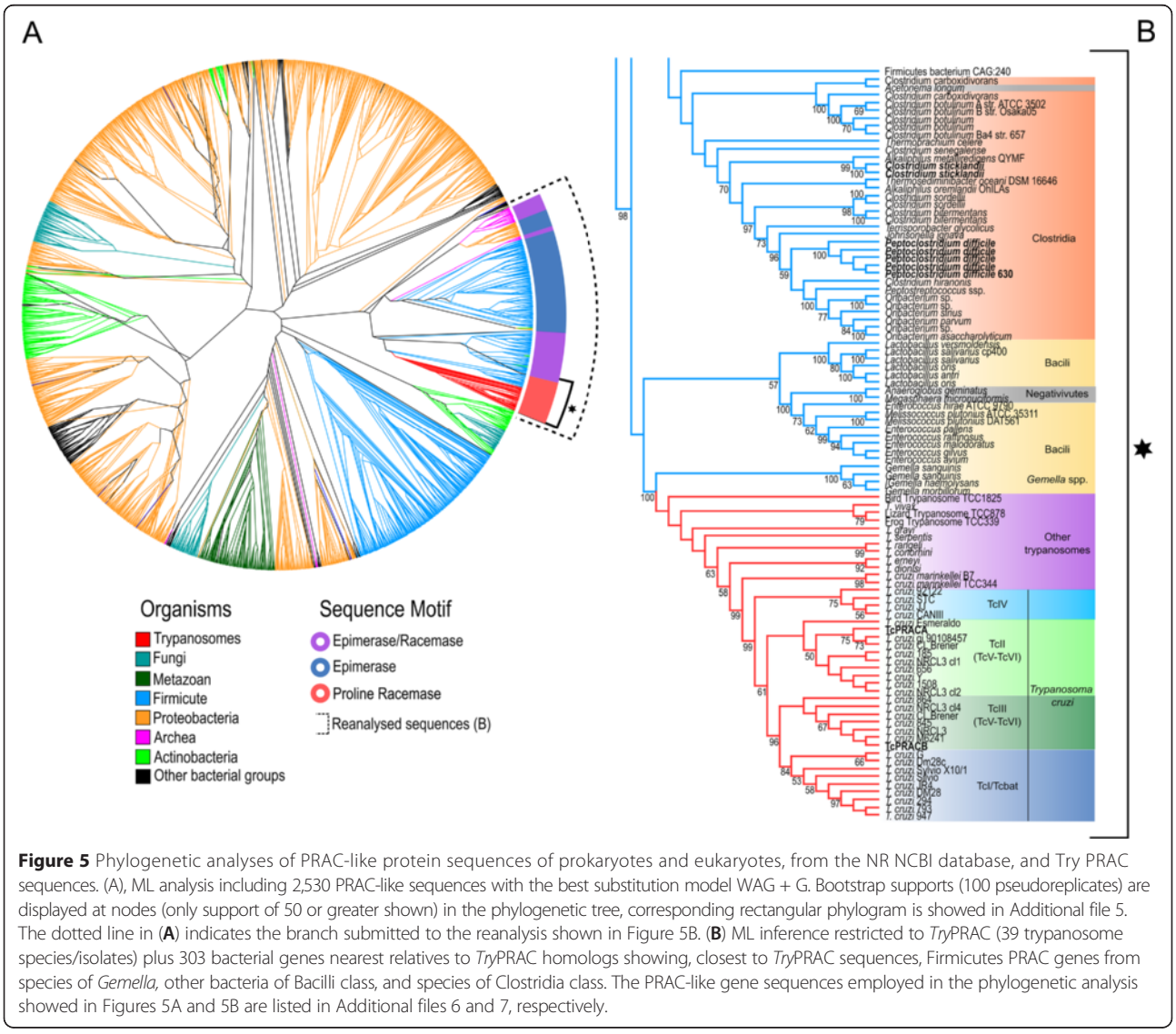

Figure 5 Phylogenetic analyses of PRAC-like protein sequences of prokaryotes and eukaryotes, from the NR NCBI database, and Try PRAC sequences. (A), ML analysis including 2,530 PRAC-like sequences with the best substitution model WAG + G. Bootstrap supports (100 pseudoreplicates) are displayed at nodes (only support of 50 or greater shown) in the phylogenetic tree, corresponding rectangular phylogram is showed in Additional file 5. The dotted line in (**A**) indicates the branch submitted to the reanalysis shown in Figure 5B. (**B**) ML inference restricted to *Try*PRAC (39 trypanosome species/isolates) plus 303 bacterial genes nearest relatives to *Try*PRAC homologs showing, closest to *Try*PRAC sequences, Firmicutes PRAC genes from species of *Gemella,* other bacteria of Bacilli class, and species of Clostridia class. The PRAC-like gene sequences employed in the phylogenetic analysis showed in Figures 5A and 5B are listed in Additional files 6 and 7, respectively.

PRAC homologs were present in all trypanosomatids or even in more basal euglenozoans, and were lost by most lineages being retained exclusively in the genus *Trypanosoma*.

Together, phylogenetic analysis and the pattern of absence/presence of PRAC-like genes strongly support our HGT hypothesis. However, it is important to consider other alternative scenarios, such as insufficient taxon sampling (donor lineages not represented in the analysis) and convergent evolution, which is an important cause of misidentified orthologs [68,69]. However, convergent evolution is in general restricted to functional domains and not across the protein length and, then, was not suggested by the alignment of *Try*PRAC with PRAC homologs from *Gemella haemolysans* and *Clostridum difficile* (Figure 1).

Taxonomy of *Try*PRAC donor prokaryotic lineages

In the most robust approach to assess taxonomic affiliation of putative HGT donors, all the most likely candidates selected by BLAST searches (useful for a primary screen of potential donors) should be submitted to deep phylogenetic analyses [17,18,31-35,68-71]. Here, the results obtained using this approach corroborate previous phylogenetic studies of prokaryotic and eukaryotic PRAC-like genes, suggesting prokaryotic donors for *Tc*PRAC and *Tv*PRAC [8-10]. In addition, we provided new insights into the origin and evolutionary history of *Try*PRAC homologs. The finding of all *Try*PRAC homologs clustering tightly together in a monophyletic assemblage, distant from any PRAC-like genes of other eukaryotes and within a large clade of prokaryotic PRAC-like genes, corroborated a single bacterium as donor lineage (Figures 5A and 5B) [Additional file 5].

To better resolve the phylogenetic relationships and visualise the most likely *Try*PRAC donor lineages, 303 prokaryotic sequences adjacent to *Try*PRACs were employed for further ML analysis (Figure 5B). The full list of taxa included in this analysis is shown in Additional file 7. The results suggest that the donor was a bacterium related to species of the Bacilli class of the orders Lactobacillales (bacteria that live in soil, water, plants and animals) and Negativicutes (anaerobes that live in rivers, lakes and animal guts) [72], including species of the genera *Gemella, Enterococcus*, *Lactobacillus* and *Melissococus* (Figure 5B). After these species, PRAC homologs from species of *Clostridium*, *Peptoclostridium* and *Oribacterium* of the Clostridia class were the most closely related to *Try*PRAC genes. In previous studies [8,10], the closest relatives of *Tc*PRAC were *Clostridium difficile* (reclassified as *Peptoclostridium*) and *Clostridium sticklandii*, both Firmicutes of the Clostridia class [72]. In this analysis, PRAC homologs from *Gemella haemolysans*, *G. morbillorum* and *G. sanguinis,* all exhibiting typical residues of PRAC racemase enzymes, were the nearest relatives to *Try*PRACs (~57% of identity) (Figure 5B). The species of *Gemella* are oral and gastrointestinal commensals of animals including humans that, as opportunistic pathogens, cause severe pulmonary, cardiac and cerebral infections [73]. Interestingly, *G. haemolysans* and *G. morbillorum* are highly prevalent among the bacterial fauna of haematophagous dipterans of Culicidae [74], which can transmit trypanosomes among anurans [41].

Synteny analysis revealed highly conserved gene order around *Try*PRAC homologs

Our findings demonstrated that *Try*PRAC homologs are ubiquitous in the genus *Trypanosoma* (Figure 2A). Previous studies revealed high conservation of gene segments containing the PRAC locus from *T. cruzi* and *T. vivax* [8,10]. To verify the genome organisation of PRAC genes in the different trypanosome species, we performed BLAST searches in genome databases for orthologous genes in PRAC loci. The results showed a syntenic block shared by all species (Figure 6). The adjacent regions of PRAC genes exhibited high synteny in *T. cruzi*, *T. c. marinkellei, T. erneyi*, *T. dionisii, T. vivax* and *T. grayi,* with at least 8 genes arranged in the same order: K39 kinesin, WD domain-containing protein, cold shock domain-containing protein, PRAC, hypothetical protein, zinc-finger protein, poly (A) polymerase, carbohydrate kinase and phosphatidyl serine. Syntenic organisation strongly supports orthology, with a single insertion of a prokaryotic PRAC gene between the cold-shock and

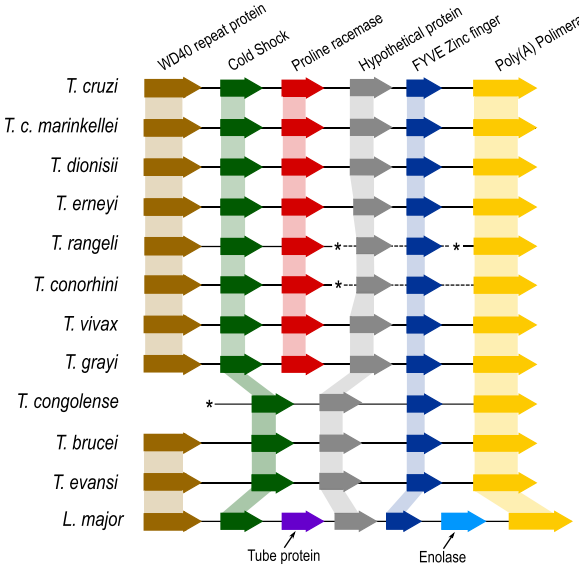

Figure 6 Synteny analysis showing the genomic organisation of orthologous genes arranged on Try PRAC loci in the genomes of several trypanosome species. The aligned genome segments showed total synteny in all trypanosomes exhibiting PRAC genes and the loss exclusively of PRAC gene in the corresponding *T. congolense*, *T. b. brucei* and *T. evansi* loci. Only partial synteny was observed in the *L. major* locus, with an additional putative enolase gene and the replacement of PRAC by a Tubby family gene.

hypothetical protein genes. Although *T. congolense* and *T. b. brucei* lack PRAC genes, the synteny of this genome segment was retained in these species as well as in this locus of the *T. evansi* genome. These species nested into a single clade with *T. vivax* as the basal species. The synteny in PRAC loci corroborated the loss of PRAC by a common ancestor of these species after the divergence of *T. vivax* (Figure 2B). Figure 6 shows the conserved gene order of five orthologs flanking the gene PRAC of trypanosomes.

Synteny analysis requires larger contigs still not available in some ongoing genome drafts. Consequently, synteny was confirmed partially for *T. rangeli* and *T. conorhini* (Figure 6). Only the two upstream genes were found in *T. serpentis,* and no genes flanking PRAC genes could be confidently located in the genomes of the non-mammalian trypanosomes. The PRAC locus showed only partially syntenic orthologs in the available genomes of non-trypanosomes trypanosomatids. In *Leishmania major* and *Crithidia* spp., which are phylogenetically related organisms [1,2], the position occupied by the PRAC gene was taken by a Tubby superfamily gene. In addition, these species also exhibited a putative enolase gene absent in this segment of trypanosome genomes (Figure 6). The insect trypanosomatids *Angomonas desouzai* and *Strigomonas culicis* lack both PRAC and Tubby family genes, while conserving the other flanking genes (data not shown). Therefore, corroborating the high plasticity of trypanosomatid genomes, the gain and loss of the PRAC gene in trypanosome genomes was not the only event that occurred in this genome segment during the evolution of the trypanosomatids.

PRAC homologs from trypanosomes are largely under the influence of purifying selection

One common characteristic of HGT is that the horizontally acquired bacterial gene and host vertically inherited genes can differ in GC content and codon usage, depending on the relative timing of the HGT event [68-71]. Despite considerable GC content divergences among the genomes of different trypanosome species, the averages of the GC contents of whole genomes were comparable to those from *Try*PRAC and flanking genes, such as in the genomes of *T. cruzi* G (~49, 48 and 48%, respectively, for PRAC, PRAC loci and whole genome), *T. dionisii* (~52, 55, 47%), *T. vivax* (~49, 49, 46%) and *T. grayi* (~63, 55, 54%). Results indicating that the PRAC gene acquired by HGT has been strongly adapted to the codon usage of the host genes are consistent with ancient acquisition of the bacterial PRAC gene by one ancestor of *Trypanosoma*.

Protein sequences can continuously evolve under the effect of evolutionary pressure that arises as a consequence of the host-parasite interactions including host immune defences. To examine positive or negative selection pressures on the evolution of *Try*PRAC, we calculated the dN/dS ratio for the putative *Try*PRAC homologs. Finding dN/dS ratios below one, indicative of negative or purifying selection, suggested that positive selection is not the driving evolutionary force shaping the *Try*PRAC repertoire. In addition, codon selection was specifically investigated on *Try*PRAC motifs and residues essential for enzyme activity. The results indicated 40 negatively selected codons, 7 of which have known relevance to PRAC activity in *T. cruzi* and *T. vivax* (Figure 1), and no positively selected codon. PRAC homologs, despite evolving to be species and genotype specific, are largely under the influence of purifying selection. The fixation and evolution of PRAC genes under strong constraint in several trypanosomes suggests that PRAC activity should be advantageous to these parasites.

The genetic polymorphism analysis should be one of the first steps in the selection of promising vaccine and drug candidates [20]. A target sharing high conservation of all essential motifs in isolates representing the *T. cruzi* genetic repertoire, as evidenced previously for cruzipain [44] and herein demonstrated for *Tc*PRAC enzymes are good candidates for a multivalent drug against Chagas disease.

Conclusions

T. cruzi and *T. vivax* PRAC enzymes are potent host B-cell mitogens that delay specific immune defences through the generation of non-specific B-cell proliferation, allowing parasite evasion and disease progression. These enzymes have also been linked to metabolism and parasite multiplication and differentiation. We identified *Try*PRAC homologs in the genomes of 12 trypanosome species, including newly sequenced genomes from trypanosomes of mammals, birds, snakes, lizards, crocodiles and toads. *Try*PRAC homologs were identified in most trypanosomes, including pathogenic and non-pathogenic species with different life cycles in vertebrates and vectors. *T. brucei* ssp., *T. evansi, T. congolense, T. simiae* and *T. godfreyi*, which are all pathogenic for mammals, were so far the only trypanosomes that lost PRAC genes.

The *Try*PRAC genealogy is congruent with the recognised relationships within *Trypanosoma*, with genes evolving to become species-specific and genotype-specific. A taxon-rich phylogenetic analysis strongly supports a bacterial origin for these genes. The presence of *Tc*PRAC homologs in trypanosomes of the Aquatic clade (basal in the phylogeny of *Trypanosoma*) and the absence of any PRAC-like gene in other trypanosomatid genera, bodonids and euglenids, together with the high synteny of PRAC gene neighbourhood allowed us to hypothesise that a common ancestor of *Trypanosoma* gained the bacterial gene through a single HGT. Our analysis supports a Firmicutes bacterium as the donor lineage, and suggested that the closest

relatives of *Try*PRACs are not in the genus *Clostridium* of Clostridia class as previously hypothesized, but more likely in the genus *Gemella* of the Bacilli class. However, this conclusion can change as more and more genome sequences become available in this important bacterial group. The results revealed unique PRAC homologs for each species as well as for each *T. cruzi* DTU and *T. rangeli* lineages. According to *in silico* analysis, all newly identified putative *Try*PRAC genes likely express functional racemases except *T. rangeli*, which has only pseudogenes. Together, our results suggest evolutionarily driven rearrangements on *Try*PRAC loci resulting in the fixation of intact PRAC genes in most trypanosomes, complete loss by the subclade *T. brucei-T. congolense,* and PRAC genes, apparently, in process of being lost in *T. rangeli.*

An understanding of the repertoire and evolutionary history of genes encoding *Try*PRAC homologs in a range of trypanosome species and genotypes can help understand the potential role of PRAC enzymes in host-trypanosome interactions. Further analyses are required to evaluate the expression and any involvement of novel putative *Try*PRAC enzymes in the life cycles, infection strategies, pathogenicity, virulence, and host immune evasion of the various trypanosome species.

Additional files

Additional file 1: Table containing the isolates of *T. cruzi* and *T. rangeli* and respective host species, geographic origin and lineages, and Genbank accession numbers of their respective PRAC gene sequences.

Additional file 2: Genbank accession numbers of gGAPDH gene sequences included in the Figure 2B.

Additional file 3: (A) Network genealogy of entire PRAC amino-acid sequences from *T. cruzi* isolates of different DTUs. Colors represent the DTUs, and the size of circles indicates the numbers of isolates. **(B)** Polymorphic amino acids detected on the alignment comprising 68 partial PRAC sequences from *T. cruzi* isolates of TcI-TcVI DTUs and Tcbat.

Additional file 4: Alignment of predicted amino acid sequences of PRAC pseudogenes from *T. rangeli* isolates of all phylogenetic lineages (TrA-TrE). Essential motifs (MCGH and MIII) are in green, active site in red (SPCGT), R1, R2 and R3 in blue and Cys91/267 in yellow. Red asterisks indicate stop codons.

Additional file 5: Maximum likelihood phylogeny of 2,530 PRAC-like protein sequences (Figure 5A) displayed as rectangular phylogram.

Additional file 6: Genbank acession numbers of 2,530 PRAC-like family genes from prokaryotes and eukaryotes retrieved from full NCBI NR database and included in the Figure 5A.

Additional file 7: Genbank accession numbers of 303 prokaryotic PRAC-like genes closest related to *Try*PRAC genes retrieved from full NCBI NR database and included in the Figure 5B.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
ZEC, AGCM and RCF performed the sequencing, genome and phylogenetic analyses and participated in the manuscript preparation. JMPA, MGS and GB were responsible for genome sequencing of draft genomes within the ATOL project. PM, JMPA, GB and EPC contributed with data interpretation, discussion and manuscript preparation. MMGT designed and coordinated the study and drafted the manuscript. All authors contributed with discussions and have read, revised and approved the final version of the manuscript.

Acknowledgements
We are grateful to The Wellcome Trust (TriTrypDB) and to the Kinetoplastid Genome Sequencing and Analysis Consortium NIH/NHGRI/NIAID (project ID 59941) of the Genome Institute at Washington University School of Medicine (USA) for making genomes of trypanosomatids freely available. Additional draft and ongoing genomes were obtained within ATOL (Assembling the tree of Life - Phylum Euglenozoa) program sponsored by the National Science Foundation, USA (PI Gregory Buck: DEB-0830056), and by projects supported by CNPq, CAPES and USP to MMGT. ZEC is recipient of a PhD fellowship from SENACYT (Panamá). AGCM, RCF and JMPA are recipients of fellowships from Brazilian agencies CNPq, CAPES and FAPESP (grant #2013/14622-3, São Paulo Research Foundation), respectively. We thank Marta Campaner for parasite cultures, and Carmen SA Takata and Tania E. Matsumoto for genes and genome sequencing (Illumina platform). We are grateful to Dr Carmenza Spadafora for valuable support given to this project in INDICASAT-AIP, Panama.

Author details
[1]Departamento de Parasitologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, SP 05508-900, Brazil. [2]Instituto de Investigaciones Científicas y Servicios de Alta Tecnología-AIP (INDICASAT-AIP), Ciudad del Saber, Clayon, Panamá. [3]Department of Microbiology and Immunology, Virginia Commonwealth University, Virginia, USA. [4]Département Infection et Epidemiologie, Institut Pasteur, Laboratoire des Processus Infectieux à Trypanosomatidés, Paris, France.

Received: 25 October 2014 Accepted: 25 March 2015
Published online: 12 April 2015

References
1. Lukes J, Skalicky T, Tyc J, Votypka J, Yurchenko V. Evolution of parasitism in kinetoplastid flagellates. Mol Biochem Parasitol. 2014;195:115–22.
2. Maslov DA, Votypka J, Yurchenko V, Lukes J. Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. Trends Parasitol. 2013;29:43–52.
3. Hamilton PB, Gibson WC, Stevens JR. Patterns of co-evolution between trypanosomes and their hosts deduced from ribosomal RNA and protein-coding gene phylogenies. Mol Phylogenet Evol. 2007;44:15–25.
4. Vallejo GA, Guhl F, Schaub GA. Triatominae-*Trypanosoma cruzi*/*T. rangeli*: Vector-parasite interactions. Acta Trop. 2009;110:137–47.
5. Cavazzana Jr M, Marcili A, Lima L, da Silva FM, Junqueira AC, Veludo HH, et al. Phylogeographical, ecological and biological patterns shown by nuclear (ssrRNA and gGAPDH) and mitochondrial (Cyt b) genes of trypanosomes of the subgenus *Schizotrypanum* parasitic in Brazilian bats. Int J Parasitol. 2010;40:345–55.
6. Minoprio P. Parasite polyclonal activators: new targets for vaccination approaches? Int J Parasitol. 2001;31:588–91.
7. Buschiazzo A, Goytia M, Schaeffer F, Degrave W, Shepard W, Gregoire C, et al. Crystal structure, catalytic mechanism, and mitogenic properties of *Trypanosoma cruzi* proline racemase. Proc Natl Acad Sci U S A. 2006;103:1705–10.
8. Chamond N, Cosson A, Coatnoan N, Minoprio P. Proline racemases are conserved mitogens: characterization of a *Trypanosoma vivax* proline racemase. Mol Biochem Parasitol. 2009;165:170–9.
9. Chamond N, Goytia M, Coatnoan N, Barale JC, Cosson A, Degrave WM, et al. *Trypanosoma cruzi* proline racemases are involved in parasite differentiation and infectivity. Mol Microbiol. 2005;58:46–60.
10. Chamond N, Gregoire C, Coatnoan N, Rougeot C, Freitas-Junior LH, da Silveira JF, et al. Biochemical characterization of proline racemases from the human protozoan parasite *Trypanosoma cruzi* and definition of putative protein signatures. J Biol Chem. 2003;278:15484–94.

11. Reina-San-Martin B, Degrave W, Rougeot C, Cosson A, Chamond N, Cordeiro-Da-Silva A, et al. A B-cell mitogen from a pathogenic trypanosome is a eukaryotic proline racemase. Nat Med. 2000;6:890–7.
12. Coatnoan N, Berneman A, Chamond N, Minoprio P. Proline racemases: insights into *Trypanosoma cruzi* peptides containing D-proline. Mem Inst Oswaldo Cruz. 2009;104 Suppl 1:295–300.
13. Radkov AD, Moe LA. Bacterial synthesis of D-amino acids. Appl Microbiol Biotechnol. 2014;98:5363–74.
14. Horcajo P, Pedro MA, Cava F. Peptidoglycan plasticity in bacteria: stress-induced peptidoglycan editing by noncanonical D-amino acids. Microb Drug Resist. 2012;18:306–13.
15. Cardinale GJ, Abeles RH. Purification and mechanism of action of proline racemase. Biochemistry. 1968;7:3970–8.
16. Goytia M, Chamond N, Cosson A, Coatnoan N, Hermant D, Berneman A, et al. Molecular and structural discrimination of proline racemase and hydroxyproline-2-epimerase from nosocomial and bacterial pathogens. PLoS One. 2007;2:e885.
17. Fitzpatrick DA, Logue ME, Butler G. Evidence of recent interkingdom horizontal gene transfer between bacteria and *Candida parapsilosis*. BMC Evol Biol. 2008;8:181.
18. Visser WF, Verhoeven-Duif NM, de Koning TJ. Identification of a human trans-3-hydroxy-L-proline dehydratase, the first characterized member of a novel family of proline racemase-like enzymes. J Biol Chem. 2012;287:21654–62.
19. Coutinho L, Ferreira MA, Cosson A, Batista MM, Batista Dda G, Minoprio P, et al. Inhibition of *Trypanosoma cruzi* proline racemase affects host-parasite interactions and the outcome of *in vitro* infection. Mem Inst Oswaldo Cruz. 2009;104:1055–62.
20. Zingales B, Miles MA, Moraes CB, Luquetti A, Guhl F, Schijman AG, et al. Drug discovery for Chagas disease should consider *Trypanosoma cruzi* strain diversity. Mem Inst Oswaldo Cruz. 2014;109:828–33.
21. Zingales B, Miles MA, Campbell DA, Tibayrenc M, Macedo AM, Teixeira MMG, et al. The revised *Trypanosoma cruzi* subspecific nomenclature: rationale, epidemiological relevance and research applications. Infect Genet Evol. 2012;12:240–53.
22. Spera JM, Herrmann CK, Roset MS, Comerci DJ, Ugalde JE. A *Brucella* virulence factor targets macrophages to trigger B-cell proliferation. J Biol Chem. 2013;288:20208–16.
23. Bryan MA, Norris KA. Genetic immunization converts the *Trypanosoma cruzi* B-Cell mitogen proline racemase to an effective immunogen. Infect Immun. 2010;78:810–22.
24. Martins RM, Covarrubias C, Rojas RG, Silber AM, Yoshida N. Use of L-proline and ATP production by *Trypanosoma cruzi* metacyclic forms as requirements for host cell invasion. Infect Immun. 2009;77:3023–32.
25. Saye M, Miranda MR, Di Girolamo F, De Los Milagros Camara M, Pereira CA. Proline modulates the *Trypanosoma cruzi* resistance to reactive oxygen species and drugs through a novel D, L-proline transporter. PLoS One. 2014;9:e92028.
26. Batista JS, Rodrigues CM, Garcia HA, Bezerra FS, Olinda RG, Teixeira MMG, et al. Association of *Trypanosoma vivax* in extracellular sites with central nervous system lesions and changes in cerebrospinal fluid in experimentally infected goats. Vet Res. 2011;42:63.
27. D'Archivio S, Cosson A, Medina M, Lang T, Minoprio P, Goyard S. Non-invasive *in vivo* study of the *Trypanosoma vivax* infectious process consolidates the brain commitment in late infections. PLoS Negl Trop Dis. 2013;7:e1976.
28. Rodrigues AC, Ortiz PA, Costa-Martins AG, Neves L, Garcia HA, Alves JM, et al. Congopain genes diverged to become specific to Savannah, Forest and Kilifi subgroups of *Trypanosoma congolense*, and are valuable for diagnosis, genotyping and phylogenetic inferences. Infect Genet Evol. 2014;23:20–31.
29. Opperdoes FR, Michels PA. Horizontal gene transfer in trypanosomatids. Trends Parasitol. 2007;23:470–6.
30. Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. 2008;9:605–18.
31. 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.
32. Alves JM, Klein CC, da Silva FM, Costa-Martins AG, Serrano MG, Buck GA, et al. Endosymbiosis in trypanosomatids: the genomic cooperation between bacterium and host in the synthesis of essential amino acids is heavily influenced by multiple horizontal gene transfers. BMC Evol Biol. 2013;13:190.
33. Alves JM, Serrano MG, Maia Da Silva F, Voegtly LJ, Matveyev AV, Teixeira MMG, et al. Genome evolution and phylogenomic analysis of Candidatus Kinetoplastibacterium, the betaproteobacterial endosymbionts of *Strigomonas* and *Angomonas*. Genome Biol Evol. 2013;5:338–50.
34. Klein CC, Alves JM, Serrano MG, Buck GA, Vasconcelos AT, Sagot MF, et al. Biosynthesis of vitamins and cofactors in bacterium-harbouring trypanosomatids depends on the symbiotic association as revealed by genomic analyses. PLoS One. 2013;8:e79786.
35. Hirt RP, Alsmark C, Embley TM. Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites. Curr Opin Microbiol. 2014;23:155–62.
36. Silva DC, Silva RC, Ferreira RC, Briones MR. Examining marginal sequence similarities between bacterial type III secretion system components and *Trypanosoma cruzi* surface proteins: horizontal gene transfer or convergent evolution? Front Genet. 2013;4:143.
37. Oliveira P, Lima FM, Cruz MC, Ferreira RC, Sanchez-Flores A, Cordero EM, et al. *Trypanosoma cruzi*: Genome characterization of phosphatidylinositol kinase gene family (PIK and PIK-related) and identification of a novel PIK gene. Infect Genet Evol. 2014;25:157–65.
38. Lima L, Silva FM, Neves L, Attias M, Takata CS, Campaner M, et al. Evolutionary insights from bat trypanosomes: morphological, developmental and phylogenetic evidence of a new species, *Trypanosoma* (*Schizotrypanum*) *erneyi* sp. nov., in African bats closely related to *Trypanosoma* (*Schizotrypanum*) *cruzi* and allied species. Protist. 2012;163:856–72.
39. Hamilton PB, Teixeira MMG, Stevens JR. The evolution of *Trypanosoma cruzi*: the 'bat seeding' hypothesis. Trends Parasitol. 2012;28:136–41.
40. Lima L, Espinosa-Alvarez O, Hamilton PB, Neves L, Takata CS, Campaner M, et al. *Trypanosoma livingstonei*: a new species from African bats supports the bat seeding hypothesis for the *Trypanosoma cruzi* clade. Parasit Vectors. 2013;6:221.
41. Ferreira RC, De Souza AA, Freitas RA, Campaner M, Takata CS, Barrett TV, et al. A phylogenetic lineage of closely related trypanosomes (Trypanosomatidae, Kinetoplastida) of anurans and sand flies (Psychodidae, Diptera) sharing the same ecotopes in brazilian amazonia. J Eukaryot Microbiol. 2008;55:427–35.
42. Viola LB, Attias M, Takata CS, Campaner M, de Souza W, Camargo EP, et al. Phylogenetic analyses based on small subunit rRNA and glycosomal glyceraldehyde-3-phosphate dehydrogenase genes and ultrastructural characterization of two snake Trypanosomes: *Trypanosoma serpentis* n. sp. from *Pseudoboa nigra* and *Trypanosoma cascavelli* from *Crotalus durissus terrificus*. J Eukaryot Microbiol. 2009;56:594–602.
43. Fermino BR, Viola LB, Paiva F, Garcia HA, de Paula CD, Botero-Arias R, et al. The phylogeography of trypanosomes from South American alligatorids and African crocodilids is consistent with the geological history of South American river basins and the transoceanic dispersal of *Crocodylus* at the Miocene. Parasit Vectors. 2013;6:313.
44. Lima L, Ortiz PA, da Silva FM, Alves JM, Serrano MG, Cortez AP, et al. Repertoire, genealogy and genomic organization of cruzipain and homologous genes in *Trypanosoma cruzi,* T. cruzi-like and other trypanosome species. PLoS One. 2012;7:e38385.
45. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52.
46. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 2012;61:1061–7.
47. Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005;21:676–9.
48. Kelly S, Ivens A, Manna PT, Gibson W, Field MC. A draft genome for the African crocodilian trypanosome *Trypanosoma grayi*. Sci Data. 2014;1:140024.
49. Flores-Lopez CA, Machado CA. Analyses of 32 loci clarify phylogenetic relationships among *Trypanosoma cruzi* lineages and support a single hybridization prior to human contact. PLoS Negl Trop Dis. 2011;5:e1272.
50. Marcili A, Lima L, Cavazzana M, Junqueira AC, Veludo HH, Maia Da Silva F, et al. A new genotype of *Trypanosoma cruzi* associated with bats evidenced by phylogenetic analyses using SSU rDNA, cytochrome b and Histone H2B genes and genotyping based on ITS1 rDNA. Parasitology. 2009;136:641–55.
51. Maia Da Silva F, Junqueira AC, Campaner M, Rodrigues AC, Crisante G, Ramirez LE, et al. Comparative phylogeography of *Trypanosoma rangeli* and

Rhodnius (Hemiptera: Reduviidae) supports a long coexistence of parasite lineages and their sympatric vectors. Mol Ecol. 2007;16:3361–73.
52. Maia Da Silva F, Marcili A, Lima L, Cavazzana Jr M, Ortiz PA, Campaner M, et al. *Trypanosoma rangeli* isolates of bats from Central Brazil: genotyping and phylogenetic analysis enable description of a new lineage using spliced-leader gene sequences. Acta Trop. 2009;109:199–207.
53. Marcili A, Valente VC, Valente SA, Junqueira AC, da Silva FM, Pinto AY, et al. *Trypanosoma cruzi* in Brazilian Amazonia: Lineages TCI and TCIIa in wild primates, *Rhodnius* spp. and in humans with Chagas disease associated with oral transmission. Int J Parasitol. 2009;39:615–23.
54. Truc P, Büscher P, Cuny G, Gonzatti MI, Jannin J, Joshi P, et al. Atypical human infections by animal trypanosomes. PLoS Negl Trop Dis. 2013;7:e2256.
55. Berneman A, Montout L, Goyard S, Chamond N, Cosson A, d'Archivio S, et al. Combined approaches for drug design points the way to novel proline racemase inhibitor candidates to fight Chagas' disease. PLoS One. 2013;8:e60955.
56. Duz ALC, Vieira PMA, Roatt BM, Aguiar-Soares RDO, Cardoso JMO, Oliveira FCB, et al. The TcI and TcII *Trypanosoma cruzi* experimental infections induce distinct immune responses and cardiac fibrosis in dogs. Mem Inst Oswaldo Cruz. 2014;109:1005–13.
57. Segovia M, Carrasco HJ, Martinez CE, Messenger LA, Nessi A, Londono JC, et al. Molecular epidemiologic source tracking of orally transmitted Chagas disease, Venezuela. Emerg Infect Dis. 2013;19:1098–101.
58. Teixeira MMG, da Silva FM, Marcili A, Umezawa ES, Shikanai-Yasuda MA, Cunha-Neto E, et al. Short communication: *Trypanosoma cruzi* lineage I in endomyocardial biopsy from a north-eastern Brazilian patient at end-stage chronic Chagasic cardiomyopathy. Trop Med Int Health. 2006;11:294–8.
59. Samudio F, Ortega-Barria E, Saldana A, Calzada J. Predominance of *Trypanosoma cruzi* I among Panamanian sylvatic isolates. Acta Trop. 2007;101:178–81.
60. Monteiro WM, Margioto Teston AP, Gruendling AP, dos Reis D, Gomes ML, de Araujo SM, et al. *Trypanosoma cruzi* I and IV stocks from Brazilian Amazon are divergent in terms of biological and medical properties in mice. PLoS Negl Trop Dis. 2013;7:e2069.
61. Ragone PG, Perez Brandan C, Padilla AM, Monje Rumi M, Lauthier JJ, Alberti D'Amato AM, et al. Biological behavior of different *Trypanosoma cruzi* isolates circulating in an endemic area for Chagas disease in the Gran Chaco region of Argentina. Acta Trop. 2012;123:196–201.
62. Llewellyn MS, Lewis MD, Acosta N, Yeo M, Carrasco HJ, Segovia M, et al. *Trypanosoma cruzi* IIc: phylogenetic and phylogeographic insights from sequence and microsatellite analysis and potential impact on emergent Chagas disease. PLoS Negl Trop Dis. 2009;3:e510.
63. Guhl F, Auderheide A, Ramirez JD. From ancient to contemporary molecular eco-epidemiology of Chagas disease in the Americas. Int J Parasitol. 2014;44:605–12.
64. Ortiz PA, Maia Da Silva F, Cortez AP, Lima L, Campaner M, Pral EM, et al. Genes of cathepsin L-like proteases in *Trypanosoma rangeli* isolates: markers for diagnosis, genotyping and phylogenetic relationships. Acta Trop. 2009;112:249–59.
65. Urrea DA, Guhl F, Herrera CP, Falla A, Carranza JC, Cuba-Cuba C, et al. Sequence analysis of the spliced-leader intergenic region (SL-IR) and random amplified polymorphic DNA (RAPD) of *Trypanosoma rangeli* strains isolated from *Rhodnius ecuadoriensis*, *R. colombiensis*, *R. pallescens* and *R. prolixus* suggests a degree of co-evolution between parasites and vectors. Acta Trop. 2011;120:59–66.
66. Deane LM, Deane MP, Lourenco-de-Oliveira R. Are Asian monkeys the original mammalian hosts of *Trypanosoma conorhini*? Mem Inst Oswaldo Cruz. 1986;81:127–9.
67. Deane MP, Deane LM. Studies on the life cycle of *Trypanosoma conorrhini*. "In vitro" development and multiplication of the bloodstream trypanosomes. Rev Inst Med Trop Sao Paulo. 1961;3:149–60.
68. Beiko RG, Ragan MA. Detecting lateral genetic transfer: a phylogenetic approach. Methods Mol Biol. 2008;452:457–69.
69. Beiko RG, Ragan MA. Untangling hybrid phylogenetic signals: horizontal gene transfer and artifacts of phylogenetic reconstruction. Methods Mol Biol. 2009;532:241–56.
70. Wisecaver JH, Hackett JD. The impact of automated filtering of BLAST-determined homologs in the phylogenetic detection of horizontal gene transfer from a transcriptome assembly. Mol Phylogenet Evol. 2014;71:184–92.
71. Strese A, Backlund A, Alsmark C. A recently transferred cluster of bacterial genes in *Trichomonas vaginalis* - lateral gene transfer and the fate of acquired genes. BMC Evol Biol. 2014;14:119.
72. Yutin N, Galperin MY. A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. Environ Microbiol Rep. 2013;15:2631–41.
73. Galen BT, Banach DB, Gitman MR, Trow TK. Meningoencephalitis due to *Gemella haemolysans*. J Med Microbiol. 2014;63:138–9.
74. Alves WCL, Gorayeb IS, Loureiro CB. Isolated bacteria from hematophagous Culicidae (Diptera:Nematocera) in Belém, Pará State, Brazil. Rev Pan-Amaz Saude. 2010;1:131–41.

Alves et al. BMC Evolutionary Biology 2013, 13:190
http://www.biomedcentral.com/1471-2148/13/190

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Endosymbiosis in trypanosomatids: the genomic cooperation between bacterium and host in the synthesis of essential amino acids is heavily influenced by multiple horizontal gene transfers

João MP Alves[1,5*†], Cecilia C Klein[2,3,4†], Flávia Maia da Silva[5], André G Costa-Martins[5], Myrna G Serrano[1], Gregory A Buck[1], Ana Tereza R Vasconcelos[4], Marie-France Sagot[2,3], Marta MG Teixeira[5], Maria Cristina M Motta[6] and Erney P Camargo[5]

Abstract

Background: Trypanosomatids of the genera *Angomonas* and *Strigomonas* live in a mutualistic association characterized by extensive metabolic cooperation with obligate endosymbiotic Betaproteobacteria. However, the role played by the symbiont has been more guessed by indirect means than evidenced. Symbiont-harboring trypanosomatids, in contrast to their counterparts lacking symbionts, exhibit lower nutritional requirements and are autotrophic for essential amino acids. To evidence the symbiont's contributions to this autotrophy, entire genomes of symbionts and trypanosomatids with and without symbionts were sequenced here.

Results: Analyses of the essential amino acid pathways revealed that most biosynthetic routes are in the symbiont genome. By contrast, the host trypanosomatid genome contains fewer genes, about half of which originated from different bacterial groups, perhaps only one of which (ornithine cyclodeaminase, EC:4.3.1.12) derived from the symbiont. Nutritional, enzymatic, and genomic data were jointly analyzed to construct an integrated view of essential amino acid metabolism in symbiont-harboring trypanosomatids. This comprehensive analysis showed perfect concordance among all these data, and revealed that the symbiont contains genes for enzymes that complete essential biosynthetic routes for the host amino acid production, thus explaining the low requirement for these elements in symbiont-harboring trypanosomatids. Phylogenetic analyses show that the cooperation between symbionts and their hosts is complemented by multiple horizontal gene transfers, from bacterial lineages to trypanosomatids, that occurred several times in the course of their evolution. Transfers occur preferentially in parts of the pathways that are missing from other eukaryotes.

Conclusion: We have herein uncovered the genetic and evolutionary bases of essential amino acid biosynthesis in several trypanosomatids with and without endosymbionts, explaining and complementing decades of experimental results. We uncovered the remarkable plasticity in essential amino acid biosynthesis pathway evolution in these protozoans, demonstrating heavy influence of horizontal gene transfer events, from Bacteria to trypanosomatid nuclei, in the evolution of these pathways.

Keywords: Endosymbiosis, Trypanosomatids, Amino acid biosynthesis, Phylogeny, Genomic analyses, Metabolic pathway evolution, Proteobacteria

* Correspondence: alvesjmp@gmail.com
†Equal contributors
[1]Virginia Commonwealth University, Richmond, VA, USA
[5]Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil
Full list of author information is available at the end of the article

Background

Many protozoan and metazoan cells harbor vertically inherited endosymbionts in their cytoplasm. Prominent among them are the associations between Alphaproteobacteria and leguminous root cells, as well as Gammaproteobacteria and cells lining the digestive tube of insects. Comprehensive reviews have covered most aspects of these ancient mutualistic relationships, including metabolism, genetics, and evolutionary history of the endosymbiont/host cell associations [1-7]. Much less is known about the relationship between protists and their bacterial endosymbionts [8-10], including the symbiosis between trypanosomatids and Betaproteobacteria, herein examined [11-14].

The Trypanosomatidae (Euglenozoa, Kinetoplastea) are well studied mainly because species of the genera *Trypanosoma* and *Leishmania* are pathogenic in humans and domestic animals [15]. However, despite their importance, these pathogens are a minority within the family, and most species are non-pathogenic commensals in the digestive tube of insects [16-18]. Usually, trypanosomatids are nutritionally fastidious and require very rich and complex culture media, however a very small group of these protozoa can be cultivated in very simple and defined media [19-23]. This reduced group of insect trypanosomatids carries cytoplasmic endosymbionts and is known as symbiont-harboring trypanosomatids, to distinguish them from regular insect trypanosomatids naturally lacking symbionts. Symbiont-harboring trypanosomatids belong to the genera *Strigomonas* and *Angomonas* [24], and their lesser nutritional requirements indicate that they have enhanced biosynthetic capabilities. In a few cases, it has been shown that the symbiotic bacterium contains enzymes involved in host biosynthetic pathways, but in most cases the metabolic contribution of the endosymbiont has been inferred from nutritional data rather than genetically demonstrated [12,14].

Each symbiont-harboring trypanosomatid carries just one symbiont in its cytoplasm, which divides synchronously with other host cell structures and is vertically transmitted [14]. The endosymbionts' original association with an ancestral trypanosomatid is thought to have occurred sometime in the Cretaceous [13]. This long partnership has led to considerable changes on the endosymbiont genomes including gene loss, with clear preferential retention of genes involved in metabolic collaboration with the host, and consequent genomic size reduction [25,26], as seen in other obligatory symbiotic associations [1,2,7,27-29].

Extensive comparative studies between symbiont-harboring trypanosomatids (wild and cured strains, obtained after antibiotic treatment) and regular trypanosomatids have permitted inferences about the symbiont dependence and contribution in the overall metabolism, in particular phospholipid [30-32] and amino acid [33-41] production of the host cell. Previous comparative studies on these organisms, often involving trace experiments using radioactive compounds, reported the requirement, substitution, and sparing of amino acids in culture media [22,33,42-55]. Nutritional data revealed that, as for most animals, including humans, the amino acids lysine, histidine, threonine, isoleucine, leucine, methionine, cysteine, tryptophan, valine, phenylalanine, tyrosine, and arginine/citrulline are essential for regular trypanosomatids. However, similar analyses showed that symbiont-harboring trypanosomatids require only methionine or tyrosine in culture media, suggesting that they possess the necessary enzymatic equipment to synthesize most amino acids [20,22,23,34]. Unfortunately, besides the symbiont-harboring trypanosomatids, most of these studies were performed only on *Crithidia fasciculata*, largely ignoring other trypanosomatids. Of the hundreds of enzymes known to be involved in the synthesis of essential amino acids in other organisms, only a few, i.e., diaminopimelic decarboxylase, threonine deaminase, ornithine carbamoyl transferase, argininosuccinate lyase, citrulline hydrolase, ornithine acetyl transferase, acetyl ornithinase, and arginase have been identified and characterized in trypanosomatids [21,33,35-41,46,56]. Thus, in contrast to the advanced state of knowledge of genes involved in amino acid biosynthesis in many microorganisms [57], the potential for amino acid synthesis in trypanosomatids remains largely unknown. In symbiont-harboring trypanosomatids, nutritional inferences provided little information about the effective participation of the symbiotic bacterium in the various metabolic pathways of the host protozoan. This contrasts with the advancement of knowledge about the presence/absence of genes for complete pathways for amino acid synthesis in many microorganisms.

Herein, we have identified the genes involved in the biosynthetic pathways of the essential amino acids in the genomes of symbiont-harboring and regular trypanosomatids of different genera (see Methods), through the characterization of each gene by similarity searches and protein domain analyses. We apply extensive phylogenetic inferences to determine the most likely origins of these genes, as it has been previously shown that other important metabolic enzymes in trypanosomatids have been transferred from bacteria, other than the present symbiont [58]. Although detection of a gene with a presumed function does not definitely prove its activity, the association of its presence with complementary nutritional and biochemical data supports the conclusion that it functions as predicted. In the present work, we establish the clear and defined contribution of endosymbionts to the amino acid metabolism of their trypanosomatid hosts, which is related to high amounts of lateral transfer of genes from diverse bacterial lineages to trypanosomatid genomes.

Results and discussion

In this work, the presence or absence of a given gene for a particular enzyme was verified in the genomes of endosymbionts, symbiont-harboring and regular trypanosomatids and then compared with the available nutritional and enzymatic data on essential amino acid biosynthesis in insect trypanosomatids. Extensive phylogenetic analyses were also performed on most of the identified trypanosomatid genes, in addition to some symbiont genes of interest. Data are mostly limited to the regular and symbiont-harboring trypanosomatid and endosymbiont genomes that have been sequenced here. Although the genomes of all available symbiont-harboring trypanosomatids and endosymbionts have been examined, only a very limited sample of regular trypanosomatid genomes (*H. muscarum* and *C. acanthocephali*) was included in these analyses, precluding generalizations about trypanosomatids as a whole. Data on the genomes of leishmaniae and trypanosomes available in KEGG were also used for comparison, but a wider sampling of genomes from more diverse groups of Trypanosomatidae and other, more distant Kinetoplastida will be necessary to enable more generalizing conclusions on the evolution of essential amino acid synthesis pathways in these organisms.

Given the incomplete nature of the trypanosomatid genomes sequenced here and the possibility of contaminant sequences, we have taken extensive precautions before including each gene in our analyses (see Methods). Our genomic context analyses of the genes identified as horizontally transferred show (Additional file 1) that genes used in this work occurred, with one exception, in long contigs presenting the typical trypanosomatid architecture of long stretches of genes in the same orientation. Moreover, all these genes overwhelmingly matched those from previously sequenced trypanosomatids. The one exception is a gene (2.7.1.100, see below) that occurs only in the two regular trypanosomatids sequenced here, and whose sequences are isolated in short contigs. As described below, they form a monophyletic group in the phylogeny. GC percent (Additional file 1) and sequencing coverage (Additional file 2) analyses also show that all genes identified in this work present statistics typical of other genes from these organisms. In short, these data show that the trypanosomatid genes employed here are highly unlikely to be contaminants.

Pathways of amino acid synthesis

Lysine

Lysine, as well as methionine and threonine, are essential amino acids generated from aspartate, a non-essential amino acid, which is synthesized from oxaloacetate that is produced in the Krebs cycle. There are two main routes for the biosynthesis of lysine: the diaminopimelate (DAP) and the aminoadipate (AA) pathways. The former is largely confined to bacteria, algae, some fungi, and plants, whereas the latter is described in fungi and euglenids [59-63].

Early nutritional studies [46] showed that lysine is essential for the growth of regular trypanosomatids, but could be efficiently replaced by DAP. In accordance, radioactive tracer and enzymatic experiments revealed that DAP is readily incorporated as lysine into proteins. Moreover, DAP-decarboxylase (EC:4.1.1.20), the enzyme that converts DAP into lysine, was detected in cell homogenates of *C. fasciculata* [46]. Nevertheless, either lysine or DAP were always necessary for growth of these flagellates in defined medium, indicating that the lysine pathway was somehow incomplete. In contrast, symbiont-harboring trypanosomatids required neither lysine nor DAP to grow in defined media [19-23]. Interestingly, the genes encoding the nine enzymes of the bacterial-type DAP pathway, leading from aspartate to lysine, were identified in the genomes of all endosymbionts (Figure 1). In contrast, only the final gene of the DAP pathway was found in the genomes of the symbiont-harboring trypanosomatids, and the final two found in one regular trypanosomatid examined (*H. muscarum*), which explains why DAP could substitute for lysine in growth media of some regular trypanosomatids. There are no genes for lysine biosynthesis annotated in the leishmaniae and trypanosomes present in KEGG. It is worth mentioning that, with respect to the alternative AA pathway, we were unable to find any genes for the synthesis of lysine in any of the endosymbiont, symbiont-harboring or regular trypanosomatid genomes analyzed.

In summary, our findings using comparative genomics are in concordance with the data from previous nutritional and enzymatic studies, showing that only symbiont-harboring trypanosomatids, and not regular ones, are autotrophic for lysine and that this autonomy is provided by the DAP pathway present in their symbionts. The presence of DAP-decarboxylase in symbiont-harboring trypanosomatids may suggest that although the symbiont contains the great majority of genes for the lysine production, the host protozoan somehow controls the production of this essential amino acid.

Methionine and cysteine

Methionine is included in all defined media designed for the growth of trypanosomatids with or without symbionts [20,22,43], suggesting that these protozoans are incapable of methionine synthesis. However, experimental evidence has shown that homocysteine and/or cystathionine could substitute for methionine in culture media for trypanosomatids [43,45,64].

Our analyses suggest that regular trypanosomatids and symbiont-harboring trypanosomatids have the necessary genes to produce cystathionine, homocysteine, and methionine from homoserine (Figure 2), whereas

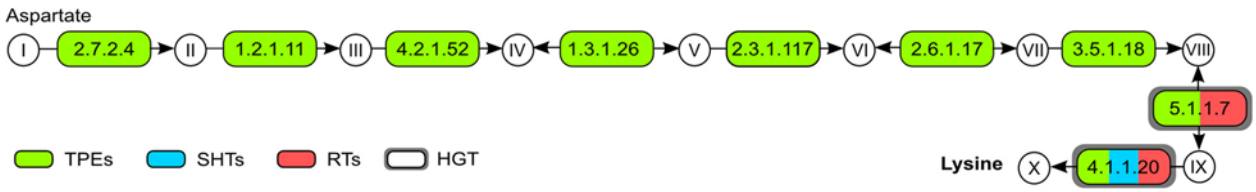

Figure 1 DAP pathway for lysine biosynthesis. Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – **I**: L-aspartate; **II**: 4-aspartyl-phosphate; **III**: aspartate 4-semialdehyde; **IV**: 2,3-dihydrodipicolinate; **V**: 2,3,4,5-tetrahydrodipicolinate; **VI**: N-succinyl-L-2-amino-6-oxopimelate; **VII**: N-succinyl-LL-2,6-diaminopimelate; **VIII**: LL-2,6-diaminopimelate; **IX**: meso-2,6-diaminopimelate; **X**: lysine. Enzymes – 2.7.2.4: aspartate kinase; 1.2.1.11: aspartate-semialdehyde dehydrogenase; 4.2.1.52: dihydrodipicolinate synthase; 1.3.1.26: dihydrodipicolinate reductase; 2.3.1.117: tetrahydrodipicolinate succinyltransferase; 2.6.1.17: succinyldiaminopimelate transaminase; 3.5.1.18: succinyldiaminopimelate desuccinylase; 5.1.1.7: diaminopimelate epimerase; 4.1.1.20: diaminopimelate decarboxylase. SHT: symbiont-harboring trypanosomatid; RT: regular trypanosomatid; TPE: trypanosomatid proteobacterial endosymbiont.

the endosymbiont genomes have no gene for the enzymes involved in the synthesis of methionine from homoserine. However, homoserine is produced from aspartate semialdehyde through the mediation of homocysteine methyltransferase (EC:1.1.1.3), which is universally present in the genomes of all the endosymbionts, symbiont-harboring and regular trypanosomatids examined.

With respect to cysteine synthesis, it has been shown that the incubation of cell homogenates of *C. fasciculata* with ^{35}S-methionine produced radioactive adenosyl-methionine (SAM), adenosyl-homocysteine (SAH), homocysteine, cystathionine, and cysteine [45]. Thus, this trypanosomatid is fully equipped to methylate methionine to produce homocysteine and, thereon, to convert homocysteine into cysteine through the trans-sulfuration pathway. However, with respect to the cystathionine/cysteine interconversion, there is some ambiguity concerning the presence or absence of cystathionine gamma-lyase (EC:4.4.1.1) in regular trypanosomatids. Many sulfhydrolases have a domain composition very similar to that of EC:4.4.1.1, which makes a definitive *in silico* function assignment to any of them difficult. Specifically, the enzymes cystathionine gamma-synthase (EC:2.5.1.48) and O-acetylhomoserine aminocarboxypropyltransferase (EC:2.5.1.49), and the two versions of cystathionine beta-lyase (EC:4.4.1.8) are possible candidates to mediate the trans-sulfuration step attributed to EC:4.4.1.1, but further research is required to establish which of these enzymes, if any, performs that reaction. We also found that, in addition to the standard pathway for methionine/cysteine synthesis (Figure 2, compounds III-X), all symbiont-harboring and regular trypanosomatids examined had the genes to produce cysteine from serine in a simple two-step reaction, with acetylserine as an intermediate (Figure 2, I-III).

In summary, if regular and symbiont-harboring trypanosomatids are capable of interconverting methionine and cysteine, as shown for *C. fasciculata* [43], none of these two amino acids can be considered essential for trypanosomatids as the presence of one renders the other unnecessary. In that case, both can be synthesized by trypanosomatids, without any participation of their symbionts, except in the optional production of aspartate semialdehyde and homoserine. However, the expression of these genes remains to be confirmed.

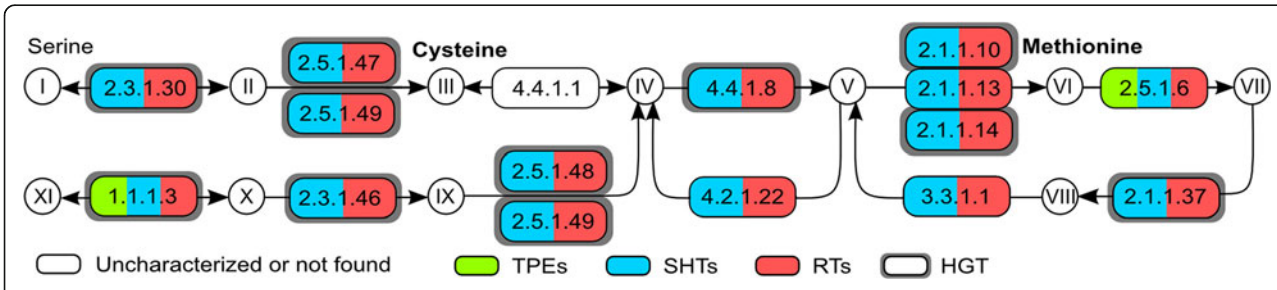

Figure 2 Cysteine and methionine synthesis and interconversion pathway. Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – **I**: L-serine; **II**: O-acetyl-serine; **III**: cysteine; **IV**: cystathionine; **V**: homocysteine; **VI**: methionine; **VII**: S-adenosyl-methionine; **VIII**: S-adenosyl-homocysteine; **IX**: succinyl-homoserine; **X**: homoserine; **XI**: aspartate 4-semialdehyde. Enzymes – 2.3.1.30: serine O-acetyltransferase; 2.5.1.47: cysteine synthase; 4.4.1.1: cystathionine gamma-lyase; 4.4.1.8: cystathionine beta-lyase; 2.1.1.x: 2.1.1.10, homocysteine S-methyltransferase, 2.1.1.13, 5-methyltetrahydrofolate–homocysteine methyltransferase, 2.1.1.14, 5-methyltetrahydropteroyltriglutamate–homocysteine methyltransferase; 2.5.1.6: S-adenosyl-methionine synthetase; 2.1.1.37: DNA (cytosine-5-)-methyltransferase; 3.3.1.1: adenosylhomocysteinase; 4.2.1.22: cystathionine beta-synthase; 2.5.1.48: cystathionine gamma-synthase; 2.3.1.46: homoserine O-succinyltransferase; 1.1.1.3: homoserine dehydrogenase. Abbreviations as for Figure 1.

Threonine

In trypanosomatids, initial investigations about the nutritional requirements for threonine were controversial. Most results suggested that this amino acid is essential [43,45,48,64-66], but other studies considered the addition of threonine to the growth media of regular trypanosomatids unnecessary [33]. Our genomic analysis favors the latter observations.

Threonine, one of the precursors of isoleucine, can be produced by different biosynthetic pathways. We have examined two of these possible routes, one starting from glycine and the other from aspartate, as presented in Figure 3. The conversion of glycine plus acetoaldehyde into threonine is mediated by threonine aldolase (EC:4.1.2.5). The gene for this enzyme is absent from endosymbionts but present in the genomes of symbiont-harboring trypanosomatids and *C. acanthocephali*, but not *Herpetomonas*. It is also absent from the genomes of trypanosomes but present in the genome of *Leishmania major* (KEGG data).

The pathway from aspartate utilizes the first two enzymes (EC:2.7.2.4 and EC:1.2.1.11) of the DAP pathway from lysine synthesis for the production of aspartate semialdehyde. These genes are present exclusively in the symbiont genomes. Aspartate semialdehyde is then sequentially converted into homoserine, phosphohomoserine, and threonine. The gene encoding homoserine dehydrogenase (EC:1.1.1.3) is universally present in the genomes of the endosymbionts, symbiont-harboring and regular trypanosomatids. It is also present in the genomes of *T. cruzi* and *Leishmania* spp. In contrast, the genes for the enzymes leading from homoserine to threonine via phosphohomoserine (EC:2.7.1.39 and EC:4.2.3.1) are present in the genomes of all insect trypanosomatids (including symbiont-harboring ones), of *Trypanosoma* spp., and *Leishmania* spp., but totally absent from the endosymbiont genomes.

Thus, the genetic constitution of regular trypanosomatids is consistent with earlier nutritional data showing the insect trypanosomatids, with or without symbionts, to be autotrophic for threonine. This observation suggests that endosymbionts are able to enhance the host cell threonine synthesis by producing the metabolic precursor aspartate semialdehyde that is also involved in other metabolic pathways.

The overall genomic and enzymatic picture is in apparent contradiction with early nutritional findings showing that threonine promoted the growth of trypanosomatids in culture [67]. This contradiction might find its basis in the fact that endogenously produced threonine is required by many metabolic processes, such that supplementation of the culture media could enhance the growth of the trypanosomatids.

Isoleucine, valine, and leucine

Isoleucine, valine, and leucine are considered essential nutrients for the growth of all trypanosomatids, except symbiont-harboring ones. The canonic pathway for the synthesis of isoleucine is depicted in Figure 4. Oxobutanoate (alpha-ketoglutaric acid) is the starting point of the pathway, and can be produced in two ways: from threonine (Figure 4, compounds II-III) or from pyruvate (Figure 4, compounds I, IX). The conversion of threonine into oxobutanoate is mediated by threonine deaminase (EC:4.3.1.19). The specific activity of this enzyme was higher in symbiont-enriched subcellular fractions of symbiont-harboring trypanosomatid homogenates than in any other cell fraction or in the cytosol, suggesting that this enzyme was located in the symbiont [33]. However, genes for EC:4.3.1.19 are present in the genomes of endosymbionts, as well as those of symbiont-harboring and regular trypanosomatids (except *Leishmania* and *Trypanosoma*), contrasting with enzymatic determinations showing the absence of enzyme activity in regular trypanosomatids [33]. Since the presence of the gene does not guarantee the functionality of the enzyme for that specific reaction, the issue remains to be experimentally verified. The next enzymatic step, the transference of the acetaldehyde from pyruvate to oxobutanoate, is mediated by the enzyme acetolactate synthase (EC:2.2.1.6), which is present exclusively in the genomes of endosymbionts. Also present only in symbionts are the genes for the next four enzymes of the pathway, which are common for valine and isoleucine synthesis. However, the gene for a branched-chain amino acid transaminase (EC:2.6.1.42), mediating the last step in the synthesis of isoleucine,

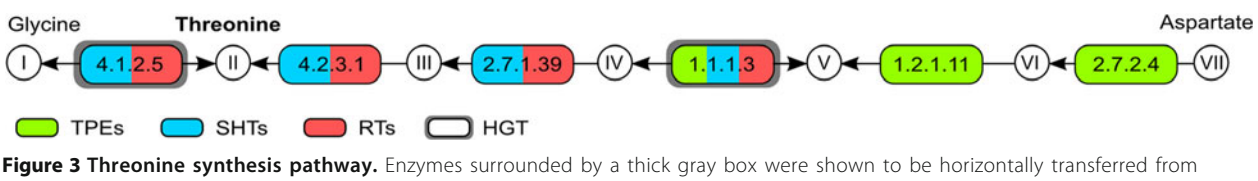

Figure 3 Threonine synthesis pathway. Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – **I**: glycine; **II**: threonine; **III**: phosphohomoserine; **IV**: homoserine; **V**: aspartate 4-semialdehyde; **VI**: 4-aspartyl-phosphate; **VII**: L-aspartate. Enzymes – 4.1.2.5: threonine aldolase; 2.7.2.4: aspartate kinase; 1.2.1.11: aspartate-semialdehyde dehydrogenase; 1.1.1.3: homoserine dehydrogenase; 2.7.1.39: homoserine kinase; 4.2.3.1: threonine synthase. Abbreviations as for Figure 1.

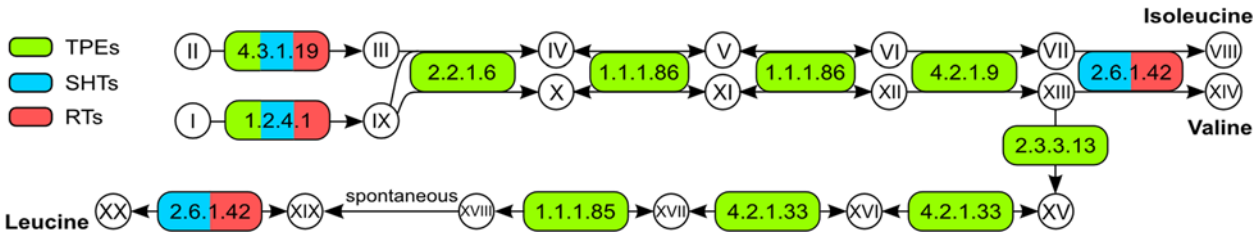

Figure 4 Isoleucine, valine, and leucine synthesis pathway. Metabolites – **I**: pyruvate; **II**: threonine; **III**: 2-oxobutanoate; **IV**: (S)-2-aceto-2-hydroxybutanoate; **V**: (R)-3-hydroxy-3-methyl-2-oxopentanoate; **VI**: (R)-2,3-dihydroxy-3-methylpentanoate; **VII**: (S)-3-methyl-2-oxopentanoate; **VIII**: isoleucine; **IX**: 2-(alpha-hydroxyethyl) thiamine diphosphate; **X**: (S)-2-acetolactate; **XI**: 3-hydroxy-3-methyl-2-oxobutanoate; **XII**: (R)-2,3-dihydroxy-3-methylbutanoate; **XIII**: 2-oxoisovalerate; **XIV**: valine; **XV**: (2S)-2-isopropylmalate; **XVI**: 2-isopropylmaleate; **XVII**: (2R,3S)-3-isopropylmalate; **XVIII**: (2S)-2-isopropyl-3-oxosuccinate; **XIX**: 4-methyl-2-oxopentanoate; **XX**: leucine. Enzymes – 1.2.4.1: pyruvate dehydrogenase E1 component subunit alpha; 4.3.1.19: threonine ammonia-lyase; 2.2.1.6: acetolactate synthase small and large subunits; 1.1.1.86: ketol-acid reductoisomerase; 4.2.1.9: dihydroxy-acid dehydratase; 2.6.1.42: branched-chain amino acid transaminase; 2.3.3.13: 2-isopropylmalate synthase; 4.2.1.33: 3-isopropylmalate dehydratase small and large subunits; 1.1.1.85: 3-isopropylmalate dehydrogenase. Abbreviations as for Figure 1.

valine, and leucine, is present in the genomes of symbiont-harboring and regular trypanosomatids, but not endosymbionts.

The first step of the valine pathway is the conversion of pyruvate into hydroxymethyl ThPP, mediated by an enzyme of the pyruvate dehydrogenase complex (EC:1.2.4.1) whose gene is present in the genomes of endosymbionts and symbiont-harboring and regular trypanosomatids. The next reaction, leading to acetolactate, is mediated by acetolactate synthase (EC:2.2.1.6), whose gene is present exclusively in the genomes of the endosymbionts. The reactions that follow from acetoacetate into valine involve the same endosymbiont genes from isoleucine synthesis.

Synthesis of leucine uses oxoisovalerate, an intermediate metabolite of the valine pathway that is converted into isopropylmalate by 2-isopropylmalate synthase (EC:2.3.3.13), encoded by a gene present only in the endosymbionts – as are the genes for the enzymes catalyzing the next three steps for leucine biosynthesis. The presence of the gene for this branched-chain amino acid transaminase (EC:2.6.1.42) in the genomes of regular trypanosomatids explains the earlier finding that oxopentanoate and oxoisovalerate, the immediate precursors of isoleucine, valine, and leucine could substitute for these amino acids when added to regular trypanosomatid synthetic culture media [43]. Interestingly, this gene is present in all symbiont-harboring and regular trypanosomatid genomes examined, but absent from endosymbiont genomes (Figure 4). It is also present in the genomes of *T. brucei* and the leishmaniae available from KEGG. In addition to isoleucine, valine, and leucine biosynthesis, this enzyme also participates in the degradation of these amino acids for their use in other metabolic processes in the cell, which might explain the presence of this enzyme as the only representative of the pathway in all regular trypanosomatids examined.

A coupled biosynthetic pathway of the branched-chain amino acids was also described for the symbiotic bacterium *Buchnera* and its aphid host, where the symbiont has the capability to synthesize the carbon skeleton of these amino acids but lacks the genes for the terminal transaminase reactions [68,69]. The aphid possesses genes hypothesized to accomplish these missing steps, even if orthologs of those are found in other insects and carry out different functions [70]. The branched-chain amino acid transaminase (EC:2.6.1.42) encoded by an aphid gene was shown to be up-regulated in the bacteriocytes, supporting the cooperation of *Buchnera* and its host in the synthesis of essential amino acids [71]. Since this transamination involves the incorporation of amino-N and the aphid diet is low in nitrogen, the host mediation of this step would be a way of maintaining a balanced profile of amino acids through transamination between those that are over abundant and those that are rare [71,72].

In summary, the presence in endosymbionts of most genes involved in isoleucine, valine and leucine synthesis explains why symbiont-harboring trypanosomatids, but not regular ones, are autotroph for these essential amino acids. However, it is worth noting that the presence of the branched-chain amino acid transaminase in trypanosomatids indicates that the host might control amino acid production according to their necessity and the nutrient availability in the medium.

Phenylalanine, tyrosine, and tryptophan

There are no enzymatic data concerning the synthesis of phenylalanine, tryptophan, and tyrosine in trypanosomatids. However, it is well known that these amino acids are essential in defined culture media designed for regular trypanosomatids, but not for symbiont-harboring ones [20,22,43,44]. The biosynthetic routes for these three amino acids use chorismate, which is produced from phosphoenolpyruvate (PEP) via the shikimate pathway, as a common substrate. The genomes of all endosymbionts contain the genes for this route, while the genomes of

symbiont-harboring and regular trypanosomatids do not (Figure 5).

The genes for the enzymes converting chorismate into prephenate and for transforming this compound into phenylalanine and tyrosine are present in all endosymbiont genomes. Symbiont-harboring and regular trypanosomatid genomes also have the genes for the last step in the synthesis of phenylalanine and tyrosine, but it is not known whether all of these enzymes are functional. The gene for phenylalanine-4-hydroxylase (EC:1.14.16.1), which converts phenylalanine into tyrosine, is present in symbiont-harboring and regular trypanosomatids, including the leishmaniae, but not in endosymbionts. Similarly, this enzyme is present only in the aphid concerning the metabolic partnership between *Buchnera* and its insect host. Furthermore, the gene encoding this enzyme is up-regulated in bacteriocytes, thus enhancing the production and interconversion of such amino acids [71]. On the other hand, endosymbionts have an additional route for the synthesis of phenylalanine from prephenate, involving enzymes aromatic-amino-acid aminotransferase (EC:2.6.1.57) and prephenate dehydratase (EC:4.2.1.51), whose genes are absent in symbiont-harboring and regular trypanosomatid genomes.

The case of the last enzyme of the tryptophan pathway is rather interesting. Tryptophan synthase (EC:4.2.1.20) possesses two subunits. This bi-enzyme complex (a tetramer of two alpha and two beta subunits) channels the product of the alpha subunit (indole) to the beta subunit, which condenses indole and serine into tryptophan [73]. Both subunits are present in the endosymbionts, whereas the genomes of symbiont-harboring trypanosomatids and *H. muscarum* have only the beta subunit. None of the other trypanosomatid genomes examined presented either subunit of tryptophan synthase.

In summary, the endosymbionts have all the genes for the different routes leading from chorismate to tryptophan, tyrosine, and phenylalanine, which are absent from symbiont-harboring and regular trypanosomatid genomes. This obviously prevents regular trypanosomatids from synthesizing any of these three amino acids and growing without supplementation. It is worth observing that the presence of phenylalanine hydroxylase, which converts phenylalanine into tyrosine, in trypanosomatids but not in endosymbionts indicates that the host might control tyrosine production.

Histidine

Histidine is derived from three precursors: the ATP purine ring furnishes a nitrogen and a carbon, the glutamine contributes with the second ring nitrogen, while PRPP donates five carbons. Histidine is a truly essential amino acid for most trypanosomatids, as corroborated by its obligatory presence in every synthetic media so far devised for regular

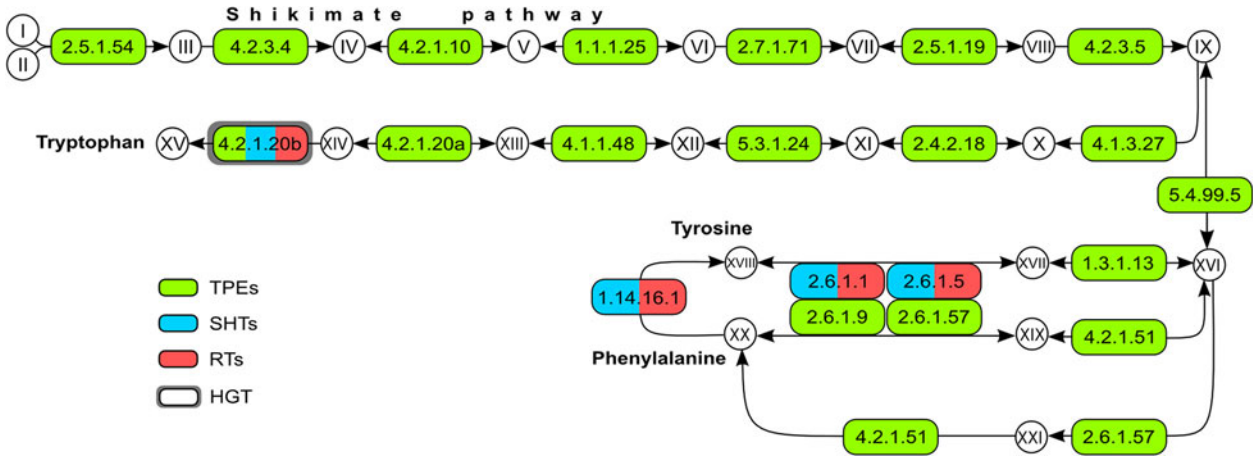

Figure 5 Phenylalanine, tyrosine, and tryptophan synthesis pathway. Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – **I**: D-erythrose 4-phosphate; **II**: phosphoenolpyruvate; **III**: 7-phosphate-2-dehydro-3-deoxy-D-arabinoheptonate; **IV**: 3-dehydroquinate; **V**: 3-dehydroshikimate; **VI**: shikimate; **VII**: shikimate 3-phosphate; **VIII**: 5-O-(1-carboxyvinyl)-3-phosphoshikimate; **IX**: chorismate; **X**: anthranilate; **XI**: N-(5-Phospho-D-ribosyl)anthranilate; XII: 1-(2-carboxyphenylamino)-1-deoxy-D-ribulose 5-phosphate; **XIII**: indoleglycerol phosphate; **XIV**: indole; **XV**: tryptophan; **XVI**: prephenate; **XVII**: arogenate; **XVIII**: phenylpyruvate; **XIX**: 4-hydroxyphenylpyruvate; **XX**: tyrosine; **XXI**: phenylalanine. Enzymes – 2.5.1.54: 3-deoxy-7-phosphoheptulonate synthase; 4.2.3.4: 3-dehydroquinate synthase; 4.2.1.10: 3-dehydroquinate dehydratase I; 1.1.1.25: shikimate dehydrogenase; 2.7.1.71: shikimate kinase; 2.5.1.19: 3-phosphoshikimate 1-carboxyvinyltransferase; 4.2.3.5: chorismate synthase; 4.1.3.27: anthranilate synthase; 2.4.2.18: anthranilate phosphoribosyltransferase; 5.3.1.24: phosphoribosylanthranilate isomerase; 4.1.1.48: indoleglycerol phosphate synthetase; 4.2.1.20a/b: tryptophan synthase alpha (a) and beta (b) subunits; 4.2.1.51/5.4.99.5: bifunctional prephenate dehydratase/chorismate mutase; 2.6.1.57: aromatic amino acid aminotransferase; 1.3.1.13: prephenate dehydrogenase (NADP+); 2.6.1.1: aspartate aminotransferase; 2.6.1.5: tyrosine aminotransferase; 2.6.1.9: histidinol-phosphate aminotransferase; 1.14.16.1: phenylalanine-4-hydroxylase. Abbreviations as for Figure 1.

trypanosomatid growth [22,43,44]. Accordingly, symbiont-harboring and regular trypanosomatid genomes do not seem to carry a single gene for histidine synthesis (Figure 6). All genes for the enzymes that participate in its biosynthesis, except the gene for histidinol-phosphate phosphatase (HPP, EC:3.1.3.15), which converts histidinol phosphate into histidinol, are present in the endosymbiont genomes. Since symbiont-harboring trypanosomatids do not require histidine, it is presumed that the absent EC:3.1.3.15 is replaced by an equivalent enzyme yet to be characterized (see Other observation on amino acid pathway peculiarities).

Arginine and ornithine

Organisms autotrophic for ornithine use the glutamate pathway [74] for its synthesis via acetylated compounds as represented in Figure 7 (I-VI). All genes for this pathway are present in the genomes of endosymbionts. The last step in the synthesis of ornithine can also be performed by the enzymes aminoacylase (EC:3.5.1.14) or acetylornithine deacetylase (EC:3.5.1.16), which convert acetylornithine into ornithine and are present in the genomes of symbiont-harboring and regular trypanosomatids, but not endosymbionts.

As represented in Figure 7, organisms lacking the glutamate pathway for the synthesis of ornithine can nevertheless produce it by different routes utilizing either citrulline or arginine [37,39,54]. Ornithine can be produced from the hydrolysis of citrulline mediated by citrulline hydrolase (EC:3.5.1.20). This activity is present in cell homogenates of all trypanosomatids, except the leishmaniae and trypanosomes, but the corresponding gene has not yet been identified to date in any organism, making it impossible to perform similarity searches. Ornithine can also be produced from arginine by means of arginase (EC:3.5.3.1), which splits arginine into ornithine and urea. The gene for arginase is present in the genomes of symbiont-harboring trypanosomatids and some regular trypanosomatids (*Leishmania* and *C. acanthocephali*), but not in the genomes of endosymbionts or *H. muscarum* – although a fragment was found in the later (see HGT and arginine and ornithine biosynthesis).

Arginine can be synthesized from ornithine through a recognized universal enzymatic pathway [74], the first step of which is the conversion of ornithine and carbamoyl phosphate into citrulline mediated by OCT (ornithine carbamoyl transferase, EC:2.1.3.3). The gene for OCT was found in the genomes of all endosymbionts and also in *Herpetomonas*, but was absent from other regular, as well as symbiont-harboring, trypanosomatid genomes examined. These findings confirm earlier immunocytochemical ultrastructural experiments showing the presence of OCT in the symbiont of *Angomonas deanei* [36]. The absence of the OCT gene renders most trypanosomatids unable to make citrulline from ornithine [75]. However, the genes for the remaining enzymes leading from citrulline into arginine are all present in the genomes of all regular and symbiont-harboring trypanosomatids, but absent from the endosymbiont genomes. These data are in full accordance with earlier enzymatic determinations for argininosuccinate synthase (EC:6.3.4.5), argininosuccinate lyase (EC:4.3.2.1), and arginase (EC:3.5.3.1) in cell homogenates of trypanosomatids [38,39,56].

Taking all these data together, we can conclude that regular trypanosomatids require exogenous sources of arginine or citrulline in their culture medium to produce ornithine. This is related to the fact that regular trypanosomatids lack the glutamate pathway for ornithine synthesis. Furthermore, ornithine cannot substitute for arginine or citrulline because most regular trypanosomatids lack OCT. Conversely, symbiont-harboring trypanosomatids are autotrophic for ornithine. This is due to the fact that, although the symbiont lacks most genes for ornithine production, it contains sequences for key enzymes such as those for the glutamate route and OCT, which converts ornithine into citrulline thus completing the urea cycle.

Polyamines

As shown in Figure 7, putrescine, a polyamine associated with cell proliferation, can be produced from ornithine in a one-step reaction mediated by ODC (ornithine decarboxylase, EC:4.1.1.17), whose gene is present in the genomes from the

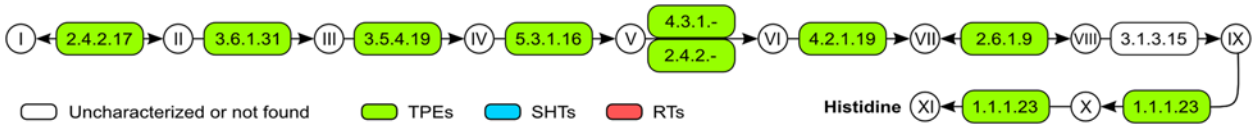

Figure 6 Histidine synthesis pathway. Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – **I**: 5-phosphoribosyl diphosphate (PRPP); **II**: phosphoribosyl-ATP; **III**: phosphoribosyl-AMP; **IV**: phosphoribosyl-formimino-AICAR phosphate; **V**: phosphoribulosyl-formimino-AICAR phosphate; **VI**: imidazole-glycerol 3-phosphate; **VII**: imidazole-acetol phosphate; **VIII**: histidinol phosphate; **IX**: histidinol; **X**: histidinal; **XI**: histidine. Enzymes – 2.4.2.17: ATP phosphoribosyltransferase; 3.6.1.31: phosphoribosyl-ATP pyrophosphohydrolase; 3.5.4.19: phosphoribosyl-AMP cyclohydrolase; 5.3.1.16: phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase; 4.1.3.-: cyclase HisF; 2.4.2.-: glutamine amidotransferase; 4.2.1.19: imidazole-glycerol phosphate dehydratase; 2.6.1.9: histidinol phosphate aminotransferase; 3.1.3.15: histidinol phosphatase; 1.1.1.23: histidinol dehydrogenase. Abbreviations as for Figure 1.

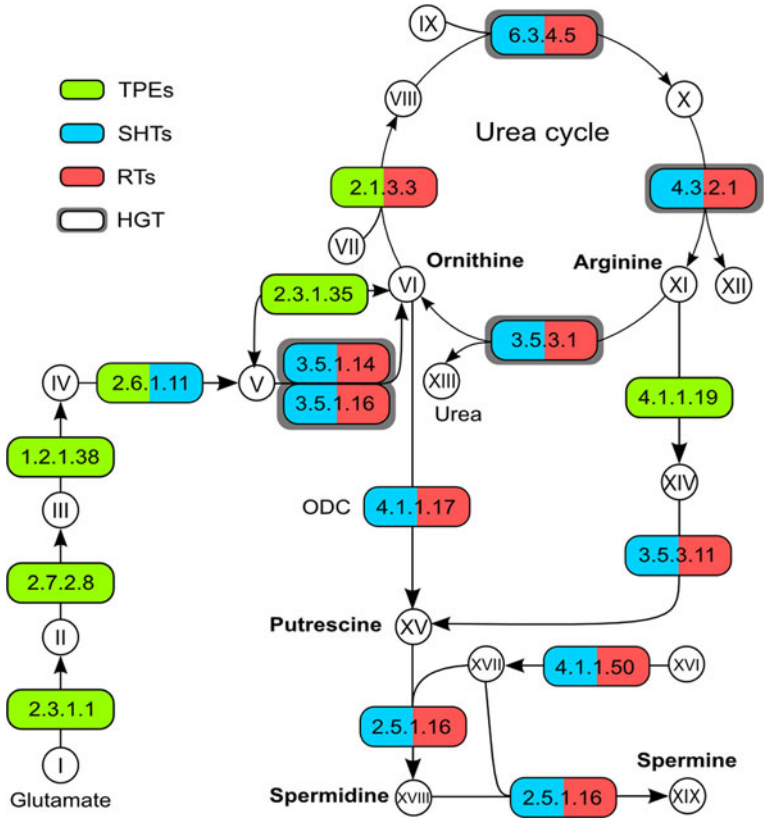

Figure 7 Arginine, ornithine, and polyamine synthesis pathway. Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – **I**: glutamate; **II**: N-acetylglutamate; **III**: N-acetylglutamyl-phosphate; **IV**: N-acetyl-glutamate semialdehyde; **V**: N-acetylornithine; **VI**: ornithine; **VII**: carbamoyl-phosphate; **VIII**: citruline; **IX**: aspartate; **X**: arginino succinate; **XI**: arginine; **XII**: fumarate; **XIII**: urea; **XIV**: agmatine; **XV**: putrescine; **XVI**: S-adenosylmethionine; **XVII**: S-adenosylmethioninamine; **XVIII**: spermidine; **XIX**: spermine. Enzymes – 2.1.3.3: ornithine carbamoyltransferase; 6.3.4.5: argininosuccinate synthase; 4.3.2.1: argininosuccinate lyase; 3.5.3.1: arginase; 4.1.1.17: ornithine decarboxylase; 3.5.1.14: aminoacylase; 3.5.1.16: acetylornithine deacetylase; 2.3.1.35: glutamate N-acetyltransferase; 2.6.1.11: acetylornithine aminotransferase; 1.2.1.38: N-acetyl-gamma-glutamyl-phosphate reductase; 2.7.2.8: acetylglutamate kinase; 2.3.1.1: amino-acid N-acetyltransferase; 3.5.3.11: agmatinase; 4.1.1.19: arginine decarboxylase; 4.1.1.50: adenosylmethionine decarboxylase; 2.5.1.16: spermidine synthase. Abbreviations as for Figure 1.

genus *Angomonas* and regular trypanosomatids, but not in endosymbionts or *Strigomonas*. Interestingly, it was proposed that the symbiont can enhance the ODC activity of *A. deanei* by producing protein factors that increase the production of polyamines in the host trypanosomatid [76]. Such high ODC activity may be directly connected to the lowest generation time described for trypanosomatids that is equivalent to 6 hours [13]. Putrescine could also be produced from agmatine since the genomes of regular and symbiont-harboring trypanosomatids have the gene for agmatinase (EC:3.5.3.11), converting agmatine into putrescine. However, the gene for the enzyme arginine decarboxylase (EC:4.1.1.19), which synthesizes agmatine, is present solely in the genomes of endosymbionts, thus completing the biosynthetic route for this polyamine, via agmatinase, in symbiont-harboring trypanosomatids. Putrescine is then converted to spermidine and spermine by enzymes S-adenosylmethionine decarboxylase (EC:4.1.1.50) and spermidine synthase (EC:2.5.1.16). The genes for these enzymes are present in the regular and symbiont-harboring trypanosomatids, but not in endosymbionts (Figure 7). Enzyme EC:2.5.1.16, converting S-adenosylmethioninamine and putrescine into S-methyl-5'-thioadenosine and spermidine, also participates in a reaction from the methionine salvage pathway. This pathway is present, complete in all symbiont-harboring and regular trypanosomatids examined (Additional file 3), although there are questions regarding the step catalyzed by acireductone synthase (EC:3.1.3.77, see HGT and methionine and cysteine biosynthesis).

Phylogenetic analyses

Our data on the phylogeny of the genes for essential amino acid biosynthesis have clearly shown that the genes present in the symbionts are of betaproteobacterial origin (for an illustrative example, see Figure 8), as shown before for the genes of heme synthesis [58] and many others

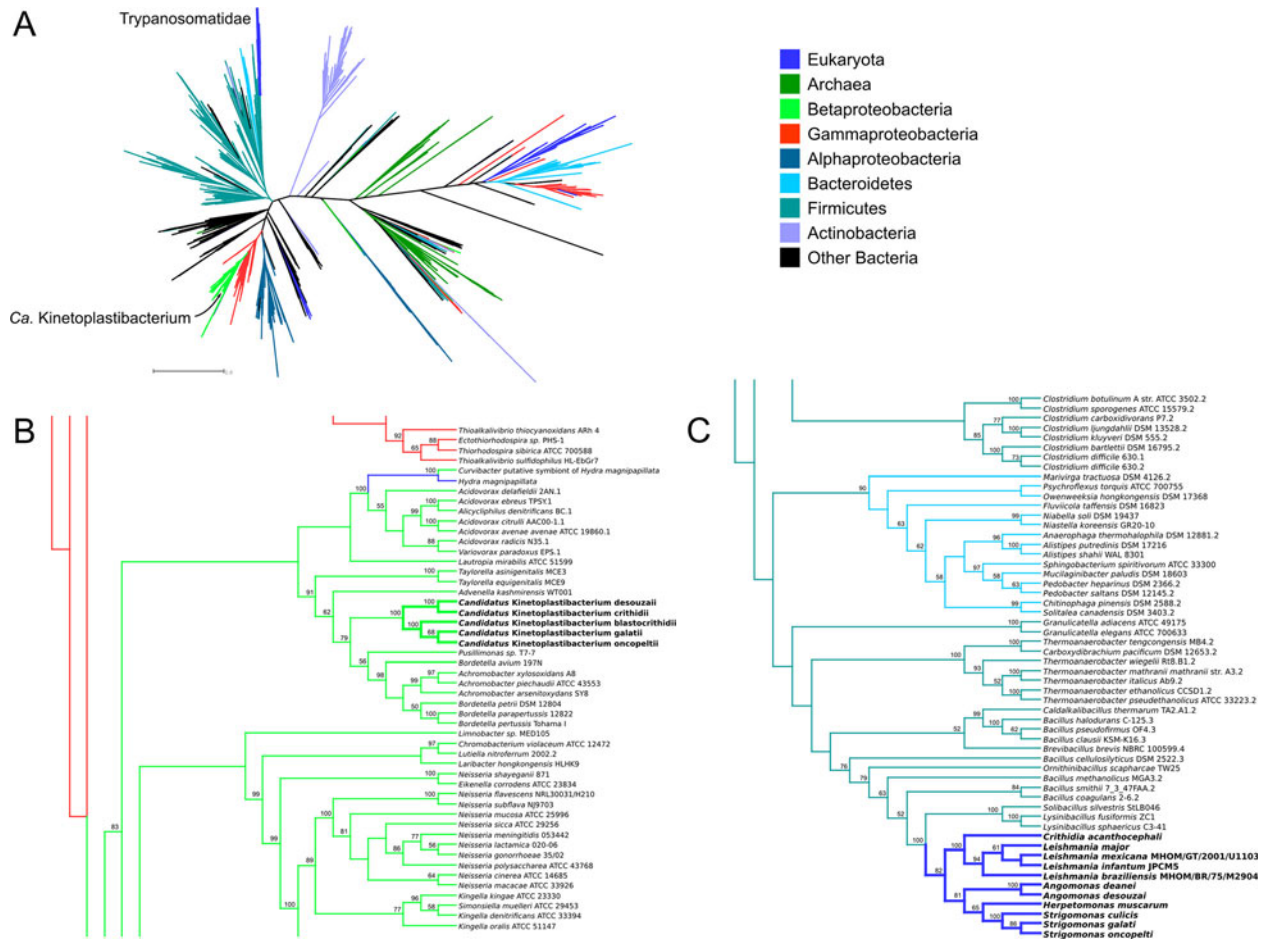

Figure 8 Maximum likelihood phylogenetic tree of homoserine dehydrogenase (EC:1.1.1.3). A – overall tree, colored according to taxonomic affiliation of each taxon, as per the legend on the right; distance bar only applies to panel **A**. **B** – details of the region of the tree where the *Ca.* Kinetoplastibacterium spp. are placed. **C** – details of the region of the tree where the Trypanosomatidae are placed. Values on nodes represent bootstrap support (only 50 or greater shown). Panels **B** and **C** are meant to only represent the branching patterns and do not portray estimated distances between sequences. Abbreviations as for Figure 1.

across the endosymbiont genomes [25]. The symbiont-harboring and regular trypanosomatid genomes, on the other hand, present a rather different situation. Thus, 18 of the 39 genes required for the biosynthesis of essential amino acids exhibited at least some phylogenetic evidence of having been horizontally transferred from a bacterial group to a trypanosomatid group, with three other genes presenting undetermined affiliation (see Additional file 2 for a summary of the phylogenetic analyses results). As detailed below, horizontal gene transfer (HGT) events seem to have originated from a few different bacterial taxa, although in some cases the exact relationship was not completely clear. Also, while some transfers are common to all trypanosomatid groups examined, others were found to be specific to certain subgroups. This could be due to multiple HGT events from associated bacteria at different points of the family's evolutionary history or, alternatively, to HGT events that occurred in the common ancestor of all trypanosomatids, whose corresponding genes were later differentially lost in certain taxa. Given the low number of genomes currently known in the family, it is difficult to assign greater probability to either scenario.

Regarding the taxonomic affiliation of the putative origin of these HGT events, it is possible to notice a preponderance of bacteria from a few phyla with three or more genes transferred, i.e. Firmicutes, Bacteroidetes, and Gammaproteobacteria, plus a few other phyla with two or less genes represented, like Actinobacteria, Betaproteobacteria, Acidobacteria, and Alphaproteobacteria. In a few other cases, the trypanosomatid genes grouped inside diverse bacterial phyla, in which case the assignment of a definite originating phylum was not possible. However, given the sometimes high rate of HGT in prokaryotic groups, it is difficult to assess with confidence the correct number of putative HGT events from Bacteria to Trypanosomatidae. It is possible that some of the genes

that seem to have originated from different phyla could actually have come from one bacterial line that was itself the recipient of one or more previous HGT events from other bacteria.

Analysis of all generated phylogenetic inferences has uncovered a clear pattern for the HGT events, which were shown to be concentrated preferentially in pathways or enzymatic steps that are usually reported to be absent in eukaryotes, particularly animals and fungi. Thus, the HGT events identified in this study involve pathways for the synthesis of lysine, cysteine, methionine, threonine, tryptophan, ornithine, and arginine (Figures 1, 2, 3, 5, and 7) and also the synthesis of a few non-essential amino acids such as glycine, serine, and proline. The detailed analysis of these events in different genes and pathways follows.

HGT of homoserine dehydrogenase

Some enzymes are common to a number of pathways involving key precursors to many compounds. Homoserine dehydrogenase (EC:1.1.1.3), for example, participates in the aspartate semialdehyde pathway for the synthesis of lysine, cysteine, methionine, and threonine (Figures 1, 2, and 3). The gene for EC:1.1.1.3 present in symbiont-harboring and regular trypanosomatid genomes seems to have been transferred from a member of the Firmicutes, clustering most closely with *Solibacillus silvestris*, *Lysinibacillus fusiformis*, and *L. sphaericus* with bootstrap support value (BSV) of 100 (Figure 8). On the other hand, the endosymbiont ortholog groups deep within the Betaproteobacteria, more specifically in the Alcaligenaceae family, as expected in the case of no HGT of this gene into the endosymbiont genomes.

HGT and lysine biosynthesis

The two genes of the lysine pathway (Figure 1) that were found in trypanosomatid genomes presented evidence of HGT. *H. muscarum* was the only trypanosomatid analyzed containing the next to last gene, for diaminopimelate epimerase (EC:5.1.1.7), which clusters strongly with the phylum Bacteroidetes, with BSV of 99 (Additional file 4). The last gene, for diaminopimelate decarboxylase (EC:4.1.1.20), was present in the symbiont-harboring and regular trypanosomatids. In the phylogeny, this particular gene has Actinobacteria as sister group (BSV of 79), although also grouping with a few other eukaryotic genera, most closely *Dictyostelium*, *Polysphondylium*, and *Capsaspora*, with BSV of 65 (Additional file 5). There are, overall, very few Eukaryota in the tree for 4.1.1.20, making it hard to reach a definite conclusion on the direction of transfer for this gene, since other eukaryotes are also present basally to this substantially large group of Actinobacteria plus Trypanosomatidae, with the high BV of 98.

Using the *C. acanthocephali* gene for EC:4.1.1.20 in a manual search against the *L. major* genome has shown a small fragment with significant similarity (57% identity and 67% similarity, from amino acid 177 to 227), but containing stop codons. Search against predicted *L. major* proteins yielded no results. These sequence remains suggest that *Leishmania* could have lost DAP-decarboxylase in a relatively recent past.

HGT and methionine and cysteine biosynthesis

The pathways for cysteine and methionine synthesis (Figure 2) present the highest number of HGT events identified among the pathways studied here. The gene for the enzyme EC:2.3.1.30, necessary for the conversion of serine to cysteine, seems to have been transferred from Bacteria to the genomes of host trypanosomatids. EC:2.3.1.30 of symbiont-harboring and regular trypanosomatids grouped inside a large cluster of diverse Bacteria (predominantly Bacteroidetes and Betaproteobacteria), with BSV of 80 (Additional file 6). An even deeper branch, which separates the subtree containing the trypanosomatids from the rest of the tree, has BSV of 97. The evolutionary history of the other enzyme with the same functionality, EC:2.5.1.47, is unclear and can not be considered a case of HGT given current results. Its gene is present in symbiont-harboring and regular trypanosomatids (including one sequence from *T. cruzi* CL Brener) and clusters as a sister group of Actinobacteria, although with low BSV (Additional file 7). Although there are many other eukaryotes in the tree, they are not particularly close to the subtree containing the Trypanosomatidae. Interestingly, one *Entamoeba dispar* sequence is a sister group to the Trypanosomatidae, although with low BSV, raising the possibility of eukaryote-to-eukaryote HGT, as previously reported (reviewed in [77]).

The gene for EC:2.3.1.46, the first in the pathway converting homoserine to cystathionine, is present in all symbiont-harboring trypanosomatids and *Herpetomonas*, but in no other regular trypanosomatid examined. This trypanosomatid gene groups within Bacteroidetes, with BSV of 53 and, in a deeper branch, BSV of 89, still clustering with Bacteroidetes only (Additional file 8).

The gene for EC:2.1.1.37, responsible for the first step in the conversion of S-adenosylmethionine into homocysteine, is present in all symbiont-harboring and regular trypanosomatids, although the sequence is still partial in the genome sequences of the *Angomonas* species. Almost all organisms in the tree are Bacteria of several different phyla (Additional file 9), with the few Eukaryota present forming a weakly supported clade. KEGG shows that many Eukaryota do possess a gene for enzyme EC:2.1.1.37, but their sequences are very different from that present in the trypanosomatids (and other eukaryotes) studied here. This therefore suggests a bacterial origin for the EC:2.1.1.37

from the Eukaryota in our tree, although the specific donor group cannot be currently determined with confidence. It is interesting to note that, besides the Trypanosomatidae, the clade of eukaryotes is composed of Stramenopiles and green algae (both groups that have, or once had, plastids), with a Cyanobacteria close to the base of the group. Although the BSV of 54 does not allow strong conclusions regarding this group, it is interesting to speculate about the possibility of eukaryote-to-eukaryote gene transfer, as previously seen (reviewed in [77]), after the acquisition of this gene from a so-far unidentified bacterium.

The genes for EC:2.5.1.48, EC:2.5.1.49, and EC:4.4.1.8 (two versions) are quite similar in sequence and domain composition. Therefore, similarity searches with any one of these genes also retrieves the other three. In spite of the similarities, these genes are found in rather different phyletic and phylogenetic patterns on the trypanosomatids (Additional file 10). EC:2.5.1.48 is present in all symbiont-harboring and regular trypanosomatids examined, plus *Trypanosoma* sp. and a few other Eukaryota (mostly Apicomplexa and Stramenopiles), all within a group of Acidobacteria (BSV of 94). The gene for EC:2.5.1.49 is present in the symbiont-harboring trypanosomatids and *Herpetomonas*, but in none of the other regular trypanosomatids examined. This trypanosomatid gene also clusters with diverse groups of Bacteria, although low BSV makes it hard to confidently identify its most likely nearest neighbor, and it is not possible to conclude with reasonable certainty that this gene is derived from HGT. The gene for EC:4.4.1.8 occurs, in symbiont-harboring and regular trypanosomatids, as two orthologs presenting very different evolutionary histories. One of the orthologs clusters with eukaryotes, with BSV of 95, while the other seems to be of bacterial descent, grouping mostly with Alphaproteobacteria of the Rhizobiales order, with BSV of 99.

The presence of two genes identified as EC:4.4.1.8 raises the possibility of them performing different enzymatic reactions. Given the overall domain composition similarities of several of the genes of the methionine and cysteine synthesis pathways, it is possible that one of the enzymes identified as EC:4.4.1.8 is actually the enzyme EC:4.4.1.1, for which no gene has been found in our searches of the Trypanosomatidae genomes, as detailed above (*Methionine and cysteine*).

Genes for two of the enzymes for the last step in the methionine synthesis, EC:2.1.1.10 and EC:2.1.1.14 (Additional files 11 and 12), are present in all regular and symbiont-harboring trypanosomatids (except for *Herpetomonas*, which lacks the latter). EC:2.1.1.14 appears to be of bacterial origin, grouping within the Gammaproteobacteria with moderate (74) bootstrap support. While EC:2.1.1.10 also groups near Gammaproteobacteria, BSV is low and this gene cannot be considered a case of HGT given current data.

As seen above, most genes in the *de novo* methionine synthesis pathway seem to have originated in one or more HGT events. Enzymes from the methionine salvage pathway (Additional file 3), on the other hand, are notably different. Of these, only S-methyl-5-thioribose kinase (EC:2.7.1.100), found in *C. acanthocephali* and *Herpetomonas* but not in the endosymbionts and symbiont-harboring trypanosomatids, seems to have originated in a bacterial group (Additional file 13). These two organisms' enzymes group deep within the Gammaproteobacteria, with BSV of 97.

Enzyme acireductone synthase (EC:3.1.3.77) presents an intriguing case, being the only methionine salvage pathway enzyme absent from the symbiont-harboring trypanosomatid genomes. This enzyme is of eukaryotic origin (not shown), and present in both *H. muscarum* and *C. acanthocephali*, but was not found in any other of the regular trypanosomatids available from KEGG. Interestingly, KEGG data for *Trypanosoma brucei* also shows the two enzymes preceding EC:3.1.3.77 as missing, which raises the question of whether this important pathway is in the process of being lost in trypanosomatids. If that is not the case, and given that all other enzymes from the pathway are present, the Trypanosomatidae must have a different enzyme (or enzymes) to perform the required reactions.

HGT and threonine biosynthesis

The gene for the enzyme that interconverts glycine and threonine (Figure 3), EC:4.1.2.5, was identified in all symbiont-harboring and regular trypanosomatids (except *Herpetomonas*), but the evolutionary histories of symbiont-harboring and regular trypanosomatid genes are very different (Additional file 14). The gene found in the regular trypanosomatids *Leishmania* sp. and *C. acanthocephali* groups deep within the Firmicutes, most closely *Clostridium*, with BSV of 63. The symbiont-harboring trypanosomatid genes, on the other hand, cluster as the most basal clade of one of the two large assemblages of eukaryotes present in this phylogeny, although all BSV are low and there is a large group of Bacteria from diverse phyla between the symbiont-harboring trypanosomatids plus a few other eukaryotic groups and the other eukaryotes in this part of the tree. It is therefore difficult to conclude whether the symbiont-harboring trypanosomatid gene is of bacterial or eukaryotic origin.

HGT and tryptophan biosynthesis

Tryptophan synthase beta subunit (EC:4.2.1.20), present in the symbiont-harboring trypanosomatids and *Herpetomonas*, is the last enzyme of the tryptophan biosynthesis pathway, and the only one present in trypanosomatids for this pathway. Its gene groups robustly (BSV of 97) with the

Bacteroidetes phylum (Additional file 15). It is also highly similar (around 80% identity and 90% similarity) to the corresponding genes of this phylum, suggesting either a very recent transfer or high sequence conservation. Given that the protein alignment of the orthologs (not shown) presents a maximum patristic distance value of 84.04% and a median of 47.22%, it is therefore likely that the transfer of EC:4.2.1.20 to the Trypanosomatidae is relatively recent.

HGT and arginine and ornithine biosynthesis

The arginine and ornithine synthesis pathway has been influenced by HGT events in a few key steps. As discussed above, one of the entry points for the urea cycle is through ornithine synthesized from glutamate. The last step, converting N-acetylornithine to ornithine, can be performed by either EC:3.5.1.14 or EC:3.5.1.16 (Figure 7). We have found that the genes for both enzymes, present in all symbiont-harboring and regular trypanosomatid genomes, originated from HGT events. All gene copies for EC:3.5.1.14 group as one clade with a gammaproteobacterium (BSV of 98), and with Bacteria of different phyla (predominantly Firmicutes) as nearest sister group, although with low BSV (Additional file 16). The few other eukaryotic groups present in the tree are very distant from the trypanosomatid group. The multiple copies of the gene for EC:3.5.1.16 in symbiont-harboring and regular trypanosomatids group together in a monophyletic clade (Additional file 17), which clusters within a large group of mostly Betaproteobacteria with BSV of 80, including the Alcaligenaceae, the family to which the endosymbionts belong. However, it seems highly unlikely that this sequence has been transferred from the endosymbiont genomes to their hosts genomes because the nuclear sequences are firmly removed from the Alcaligenaceae, and many regular trypanosomatids (including *Trypanosoma* spp.) also present this gene in the same part of the tree.

The only trypanosomatid analyzed which presented ornithine carbamoyl transferase (OCT, EC:2.1.3.3) was *Herpetomonas muscarum*. Our phylogenetic analysis of this gene indicates that it is of eukaryotic origin (not shown). The symbiont-harboring trypanosomatids utilize the OCT provided by their endosymbionts, and their OCT genes group firmly inside the Alcaligenaceae family, next to *Taylorella* and *Advenella*, as expected.

The genes for EC:6.3.4.5 and EC:4.3.2.1 present similar evolutionary patterns: both are absent from endosymbiont genomes and present in all symbiont-harboring and regular trypanosomatid genomes – the only exception being the lack of the latter in *Leishmania* spp. The Trypanosomatid genes form monophyletic groups in their respective trees, grouping within Firmicutes in both cases (Additional files 18 and 19). BSV is higher (82) in the tree of EC:4.3.2.1 than in that of EC:6.3.4.5 (69). In both cases, support falls for deeper branches in the trees. Although the host genomic sequences are still incomplete and in varying degrees of contiguity, it is interesting to note that the genes for EC:6.3.4.5 and EC:4.3.2.1 are present in tandem in one contig in all symbiont-harboring trypanosomatids (Additional file 1). The flanking genes are eukaryotic: terbinafine resistance locus protein and a multidrug resistance ABC transporter. As seen in the genome browser at TriTrypDB (http://tritrypdb.org), *Leishmania* spp. have most of these same genes, although in a slightly different order (EC:6.3.4.5 occurring after the two eukaryotic genes instead of between them) and lacking EC:4.3.2.1. *L. braziliensis* seems to be in the process of additionally losing EC:6.3.4.5, which is annotated as a pseudogene. These phylogenetic and genomic data strongly suggest that EC:4.3.2.1 and EC:6.3.4.5 have been transferred together from a Firmicutes bacterium to the common ancestor of the symbiont-harboring and regular trypanosomatids studied, and that these transferred genes have been or are being lost from *Leishmania* at least.

The final enzyme in the urea cycle, arginase (EC:3.5.3.1), is present in all symbiont-harboring and regular trypanosomatids examined here. However, the sequence from *Herpetomonas* presents a partial arginase domain; while the protein sequence length is as expected, the domain match starts only after 70 amino acids. We speculate that this divergence could be responsible for the lack of arginase activity previously seen in *Herpetomonas*. Differently from most other enzymes in this work, there are different evolutionary histories for the arginase genes: all trypanosomatid genes but that from *Herpetomonas* cluster together with very high bootstrap support of 98, within Eukaryota (Additional file 20). The sequence from *Herpetomonas* on the other hand is the sister group (BSV of 79) of a large assemblage of Bacteria from several different phyla, but predominantly Deltaproteobacteria, Firmicutes, Actinobacteria, and Cyanobacteria. It is therefore clear that *Herpetomonas* must have acquired a different arginase than that present in the other trypanosomatids studied, which possess eukaryotic genes. Furthermore, this gene seems to be undergoing a process of decay, given its lack of significant similarity to the known arginase domain in a significant portion of the protein.

HGT in other pathways: possible symbiont to host transfer

Ornithine cyclodeaminase (EC:4.3.1.12) converts ornithine directly into proline, a non-essential amino acid. In our analyses, we have found that the gene for EC:4.3.1.12 of symbiont-harboring trypanosomatid genomes is very similar to those from Betaproteobacteria of the Alcaligenaceae family, to which the endosymbionts belong. Regular trypanosomatid and endosymbiont genomes do not contain the gene for this enzyme. Accordingly, the phylogeny shows the symbiont-harboring trypanosomatid

gene grouping close to several Alcaligenaceae, although the clade is not monophyletic and presents BSV of 63 (Additional file 21). This grouping, together with the gene presence in symbiont-harboring trypanosomatid genomes only, poses the possibility that EC:4.3.1.12 has been transferred from the ancestral endosymbiont to the corresponding host, before the radiation of symbiont-harboring trypanosomatids into the two genera and five species analyzed here.

Other observation on amino acid pathway peculiarities

Some interesting peculiarities of specific genes from a few pathways deserve to be discussed. Interestingly, the gene for branched-chain-amino-acid transaminase (EC:2.6.1.42), the last step in the synthesis of isoleucine, valine, and leucine (Figure 4), was identified in all bacteria of the Alcaligenaceae family present in KEGG, except for the endosymbionts' closest relatives, *Taylorella* spp. (parasitic) and *Advenella kashmirensis* (free-living), which also lack the gene. The question is raised then of whether the common ancestor of *Taylorella* and the endosymbionts, which are sister groups [25], had already lost the gene. Another possibility is that independent loses occurred in endosymbionts, *Taylorella*, and *Advenella*. Considering that the rest of the pathway is present in these organisms and that the free-living *Advenella* would need the last gene to complete synthesis of these amino acids, it is reasonable to speculate that their EC:2.6.1.42 is novel or at least very different and thus could not be identified by similarity searches.

As mentioned above, the histidine pathway biosynthesis is performed by the endosymbionts and all enzymes, with the exception of histidinol-phosphate phosphatase (HPP, EC:3.1.3.15), have been identified. This is also the only enzyme of this pathway missing in other Betaproteobacteria available in KEGG. Recently, it was reported that such a gap in the histidine biosynthesis pathway in other organisms was completed by novel HPP families [78,79]. Our searches for the novel *C. glutamicum* HPP (cg0910, an inositol monophosphatase-like gene) have identified two possible candidate genes in the endosymbionts (BCUE_0333 and BCUE_0385, in *C.* K. blastocrithidii). As in *Corynebacterium*, neither of these genes is in the same operon as the known histidine synthesis genes. Given the absence of any other inositol phosphate metabolism genes in the endosymbiont genomes, except for these two IMPases, it is reasonable to hypothesize that at least one of the two aforementioned candidates could be the HPP.

Conclusions

In the present paper, we have put together nutritional, biochemical, and genomic data in order to describe how the metabolic co-evolution between the symbiont and the host trypanosomatid is reflected in amino acid production (Figure 9). In fact, amino acid biosynthetic pathways in symbiont-harboring trypanosomatids are frequently chimeras of host and endosymbiont encoded enzymes, with predominance of the latter in the synthesis of essential amino acids. After a careful analysis of different routes, it becomes clear that the symbiotic bacterium completes and/or potentiates most pathways of the host protozoa that are involved in amino acid production, as previously seen in other systems [7].

Sometimes, as in the lysine and histidine synthesis, the symbionts contain all genes for enzymes that compose the metabolic route. By contrast, in the cysteine and methionine pathways the bacterium lacks most genes involved in amino acid interconversion, which are present in host trypanosomatids. Interestingly, the last step of some metabolic routes such as those for lysine and tryptophan, contains two genes; one in the host genome, the other in the endosymbiont genome. This phenomenon has also been observed in the synthesis of heme [58,80], but the reasons for this peculiarity remain obscure. However, we have to consider the possibility that HGT events preceded the colonization of symbiont-harboring trypanosomatids by their endosymbionts, and that the genes present in the host genomes are just relics of previous HGT event(s). Alternatively, these genes could have been recruited to perform functions, as the control of amino acid production by the host trypanosomatid. This same strategy can be considered in isoleucine, valine, and leucine production, but in this case endosymbionts lack the enzyme for the last step, the branched-chain amino acid transaminase (EC:2.6.1.42).

A clear example of the integration of earlier nutritional and enzymatic data with the present gene screening is the synthesis of arginine and ornithine in trypanosomatids. Differently from other members of the family, the urea cycle is complete in symbiont-harboring trypanosomatids by the presence of the OCT gene (EC:2.1.3.3) in symbionts, making these protozoa entirely autotrophic for ornithine, citrulline, and arginine, as previously known from nutritional data [19,22,44,52]. Symbiont-bearing trypanosomatids contain genes for all enzymes leading from glutamate to arginine. The corresponding genes are located partly in the genomes of their endosymbionts and partly in the protozoan nucleus; in this last case, genes are of bacterial origin, resulting from HGT and including at least one transfer of two genes at once (EC:4.3.2.1 and EC:6.3.4.5), as demonstrated in our phylogenies. Furthermore, endosymbionts also contain most genes for the glutamate pathway, thus enhancing synthesis of ornithine, that once decarboxylated generates polyamine, which is related to cell proliferation and to the low generation time displayed by symbiont-harboring trypanosomatids. Results in this study confirm previous findings [25,58] showing the

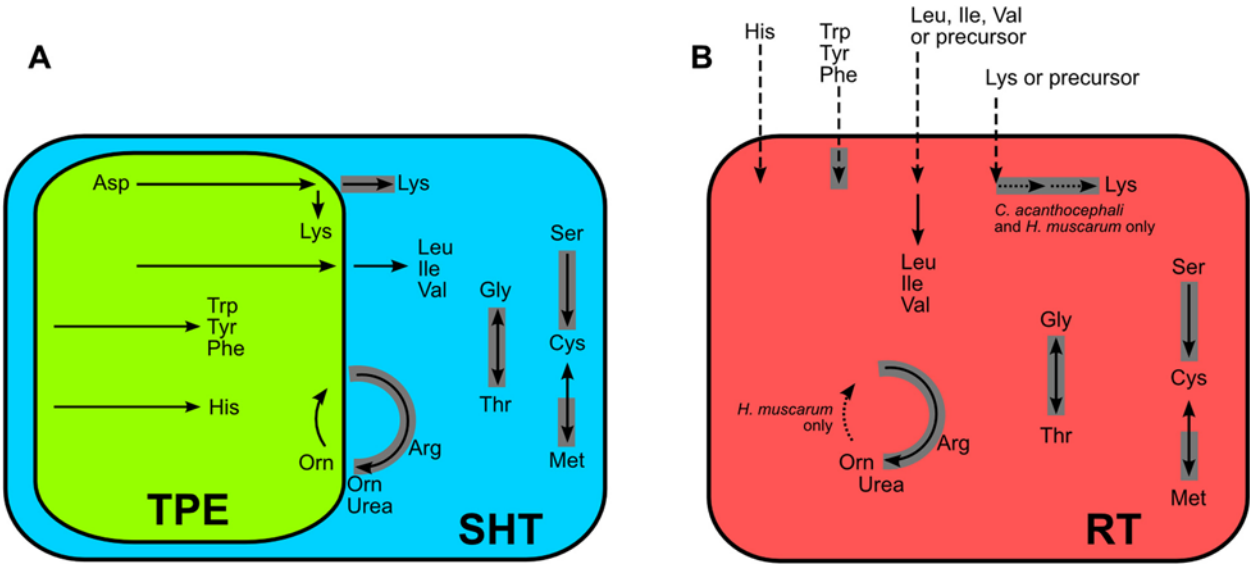

Figure 9 Overview of the biosynthetic pathways of essential amino acids in trypanosomatids. Dashed arrows: metabolite import; dotted arrows: reaction present in only some of the organisms analyzed; solid arrows: other reactions (a single arrow can summarize multiple steps); arrows surrounded by a gray box: enzymes possibly acquired through horizontal transfer from Bacteria to trypanosomatids (see main text). **A**. Contribution of symbiont-harboring trypanosomatids (SHT) and endosymbionts (TPE) based on the analysis of gene content in the genomes of *A. deanei*, *A. desouzai*, *S. culicis*, *S. oncopelti*, *S. galati*, and respective endosymbionts. **B**. Biochemical capability of trypanosomatids without symbionts (RT), based on the analysis of genomic data from *H. muscarum*, *C. acanthocephali*, and *L. major*.

betaproteobacterial origin of the genes of endosymbionts. The nuclear genes, on the other hand, present a much more convoluted evolutionary picture, with probably numerous ancient HGT events shaping the amino acid metabolism in trypanosomatids. A few pathways in particular have been heavily affected, i.e. methionine/cysteine and arginine/ornithine synthesis. Transferred genes originated preferentially from three bacterial phyla, namely Firmicutes, Bacteroidetes, and Gammaproteobacteria, although possible transfers from other phyla of Bacteria have also been uncovered. Especially interesting was the finding of a gene, coding for ornithine cyclodeaminase (EC:4.3.1.12), which closely groups with the Alcaligenaceae family of the Betaprotebacteria and that is likely to have been transferred from the endosymbiont to the host genome. Accordingly, it is present only in symbiont-harboring trypanosomatid nuclear genomes and not in any of the currently sequenced regular trypanosomatid genomes. During review of this work, a very recent report of a similar situation of multiple lineages contributing to the metabolism in the symbiosis of mealybugs, involving the three interacting partners and genes acquired through HGT from other bacterial sources (mainly Alphaproteobacteria, but also Gammaproteobacteria and Bacteroidetes) to the insect host, has been published [81]. This suggests that this phenomenon could be widespread and of great importance in genomic and metabolic evolution.

Having been detected in about half of the genes analyzed in this work, HGT events seem to have been fundamental in the genomic evolution of the Trypanosomatidae analyzed, and further phylogenetic studies of the whole host genomes should show the complete extent of this process and which additional pathways could be affected. Synthesis of vitamins (Klein et al., personal communication), heme, and amino acids have already been shown to benefit from bacterial-to-trypanosomatid HGT; many other processes in Trypanosomatidae metabolism might also be subjected to this evolutionary process.

Methods

Organisms and growth conditions

The symbiont-harboring trypanosomatid species genomes sequenced here were: *Strigomonas oncopelti* TCC290E, *S. culicis* TCC012E, *S. galati* TCC219, *Angomonas deanei* TCC036E, and *A. desouzai* TCC079E. These symbiont-harboring trypanosomatids harbor, respectively, the symbionts: *Candidatus* Kinetoplastibacterium oncopeltii, *Ca*. K. blastocrithidii, *Ca*. K. galatii, *Ca*. K. crithidii and *Ca*. K. desouzaii [24], and were previously sequenced [25]. In addition, we have also sequenced the genomes of two regular trypanosomatids, i.e. *Herpetomonas muscarum* TCC001E and *Crithidia acanthocephali* TCC037E. These organisms are cryopreserved at the Trypanosomatid Culture Collection of the University of São Paulo, TCC-USP. Symbiont-harboring trypanosomatids were grown in

Graces' medium (Gibco). Regular trypanosomatids were grown in LIT media [82].

DNA extraction and sequencing

Total genomic DNA was extracted by the phenol-chloroform method [83]. We applied kDNA depletion methods to minimize the presence of this type of molecule, as previously described [58], which result in less than about 5% of remaining kDNA in the sample. After kDNA depletion, about 5 μg of DNA were submitted to each Roche 454 shotgun sequencing run, according to the manufacturer's protocols. Different genomes have so far been sequenced to different levels of draft quality, with estimated coverages of 15X to 23X (considering a genome of ~30 Mbp). Sequences were assembled using the Newbler assembler version 2.3, provided by Roche. Resulting assemblies are available from GenBank, as detailed in "Availability of Supporting Data" below. Endosymbiont genomes were finished to a closed circle as previously described [25].

Gene discovery and annotation

Endosymbiont genes were used as previously published [25]. In an initial scan of the genome, trypanosomatid genes were discovered and mapped to metabolic pathways using ASGARD [84], employing as reference the UniRef100 [85] and the Kyoto Encyclopedia of Genes and Genomes, KEGG [86] databases. The identified segments of DNA were then extracted from the genome and manually curated for completion and proper location of start and stop codons by using the GBrowse genome browser [87]. Putative sequence functions were confirmed by domain searches against NCBI's Conserved Domain Database [88]. Genes and annotations from other trypanosomatids were used when needed and as available at KEGG. All trypanosomatid genes characterized in this study have been submitted to NCBI's GenBank and accession numbers are available from Additional file 22. All endosymbiont genes analyzed here have been previously sequenced [25]; gene identifiers are available from Additional file 23.

Due to the incomplete nature of our trypanosomatid assemblies, a set of criteria were used to avoid including contaminant sequences in our analyses. A gene was accepted as legitimate only when satisfying at least two of the following: genomic context compatible with a trypanosomatid gene (i.e. long stretches of genes in the same orientation in the contig, most neighboring genes similar to genes from other, previously sequenced trypanosomatids); sequencing coverage in the gene similar to, or higher than, that of the gene and genome averages (since contaminants that are difficult to detect will almost always be in small contigs of low coverage); GC percent content consistent with that of the neighboring genes, and of the overall genome; and phylogenetic congruence (i.e. whether genes from more than one trypanosomatid formed monophyletic assemblages). Genomic context and GC content graphs were drawn by GBrowse [87] and graphically edited for better use of space.

Phylogenetic analyses

For phylogenetic analysis of each enzyme characterized in this work, corresponding putative orthologous genes from all domains of life were collected from the public databases by BLAST search (E-value cutoff of 1e-10, maximum of 10,000 matches accepted) against the full NCBI NR protein database, collecting sequences from as widespread taxonomic groups as possible and keeping one from each species (except for alignments with more than ~1,500 sequences, in which case one organism per genus was kept). Only sequences that were complete and aligned along at least 75% of the length of the query were selected. All analyses were performed at the protein sequence level. Sequences were aligned by Muscle v. 3.8.31 [89]. Phylogenetic inferences were performed by the maximum likelihood method, using RAxML v. 7.2.8 [90] and employing the WAG amino acid substitution model [91], with four gamma-distributed substitution rate heterogeneity categories and empirically determined residue frequencies (model PROTGAMMAWAGF). Each alignment was submitted to bootstrap analysis with 100 pseudo-replicates. Trees were initially drawn and formatted using TreeGraph2 [92] and Dendroscope [93], with subsequent cosmetic adjustments performed with the Inkscape vector image editor (http://inkscape.org). Phylogenetic conclusions have been displayed as strong in the summary table for phylogenetic results (Additional file 2) if BSV was 80 or greater, and moderate if BSV was between 50 and 80 – with one exception, EC:2.1.1.37, described in the results. Resulting phylogenetic trees are available from TreeBase (accession number S14564), as detailed in "Availability of Supporting Data" below.

Availability of supporting data

The data sets supporting the results of this article are available in the GenBank and TreeBase repositories, under accession numbers AUXH00000000, AUXI00000000, AUXJ00000000, AUXK00000000, AUXL00000000, AUXM 00000000, and AUXN00000000 (genome sequences of *S. culicis*, *C. acanthocephali*, *H. muscarum*, *S. oncopelti*, *A. desouzai*, *A. deanei*, and *S. galati*, respectively), and S14564, for the sequence alignments and phylogenetic trees (http://purl.org/phylo/treebase/phylows/study/TB2:S14564).

Additional files

Additional file 1: Genomic context and GC content for candidate HGT genes in the Trypanosomatidae analyzed in this work. Arrows show TBLASTN alignments of the genome against UniRef100 and KEGG proteins. Alignment orientation is displayed in blue or red, except for the

alignment for the gene currently in focus, which is colored black. Coordinates are in kilobases.

Additional file 2: Summary of phylogenetic and genome coverage analyses of the candidate HGT genes in the Trypanosomatidae analyzed in this work and a few other genes of interest.

Additional file 3: Methionine salvage pathway. Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – I: methionine; II: S-adenosylmethionine; III: S-adenosylmethioninamine; IV: S-methyl-5-thioadenosine; V: S-methyl-5-thioribose; VI: S-methyl-5-thioribose 1-phosphate; VII: S-methyl-5-thioribulose 1-phosphate; VIII: 2,3-diketomethylthiopentyl-1-phosphate; IX: 2-hydroxy-3-keto-5-methylthiopenctenyl-1-phosphate; X: 1,2-dihydroxy3-keto-5-methylthiopentene; XI: 4-methylthio-2-oxobutanoate. Enzymes – 2.5.1.6: methionine adenosyltransferase; 4.1.1.50: adenosylmethionine decarboxylase; 2.5.1.16: spermidine synthase; 3.2.2.9: adenosylhomocysteine nucleosidase; 3.2.2.16: methylthioadenosine nucleosidase; 2.7.1.100: S-methyl-5-thioribose kinase; 2.4.2.28: S-methyl-5'-thioadenosine phosphorylase; 5.3.1.23: S-methyl-5-thioribose-1-phosphate isomerase; 4.2.1.109: methylthioribulose 1-phosphate dehydratase; 3.1.3.77: acireductone synthase; 1.13.11.54: acireductone dioxygenase; 2.6.1.5: tyrosine transaminase; 2.6.1.57: aromatic-amino-acid transaminase. SHT: symbiont-harboring trypanosomatid; RT: regular trypanosomatid; TPE: trypanosomatid proteobacterial endosymbiont.

Additional file 4: Maximum likelihood phylogeny of diaminopimelate epimerase (EC:5.1.1.7). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 5: Maximum likelihood phylogeny of diaminopimelate decarboxylase (EC:4.1.1.20). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 6: Maximum likelihood phylogeny of serine O-acetyltransferase (EC:2.3.1.30). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 7: Maximum likelihood phylogeny of cysteine synthase (EC:2.5.1.47). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 8: Maximum likelihood phylogeny of homoserine O-succinyltransferase (EC:2.3.1.46). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 9: Maximum likelihood phylogeny of DNA (cytosine-5-)-methyltransferase (EC:2.1.1.37). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 10: Maximum likelihood phylogeny of cystathionine gamma-synthase, O-acetylhomoserine aminocarboxypropyltransferase, and cystathionine beta-lyase (EC:2.5.1.48, EC:2.5.1.49, and EC:4.4.1.8). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 11: Maximum likelihood phylogeny of homocysteine S-methyltransferase (EC:2.1.1.10). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 12: Maximum likelihood phylogeny of 5-methyltetrahydropteroyltriglutamate-homocysteine S-methyltransferase (EC:2.1.1.14). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 13: Maximum likelihood phylogeny of S-methyl-5-thioribose kinase (EC:2.7.1.100). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 14: Maximum likelihood phylogeny of L-threonine aldolase (EC:4.1.2.5). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 15: Maximum likelihood phylogeny of (EC:4.2.1.20). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 16: Maximum likelihood phylogeny of aminoacylase (EC:3.5.1.14). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 17: Maximum likelihood phylogeny of acetylornithine deacetylase (EC:3.5.1.16). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 18: Maximum likelihood phylogeny of argininosuccinate synthase (EC:6.3.4.5). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 19: Maximum likelihood phylogeny of argininosuccinate lyase (EC:4.3.2.1). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 20: Maximum likelihood phylogeny of arginase (EC:3.5.3.1). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 21: Maximum likelihood phylogeny of ornithine cyclodeaminase (EC:4.3.1.12). Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

Additional file 22: Genbank accession numbers for Trypanosomatidae genes characterized in this study.

Additional file 23: GenBank locus tags for the *Ca.* Kinetoplastibacterium genes analyzed in this study.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JMPA generated and analyzed metabolic pathway data, performed phylogenetic analyses, and wrote the manuscript. CCK generated and analyzed metabolic pathway data, and edited the manuscript. FMS and AGCM analyzed metabolic pathway data. MGS and GAB generated and analyzed genomic data. ATRV, MFS, and MMGT, participated in study design and data interpretation. MCM and EPC conceived the study and wrote the manuscript. All authors read and approved the manuscript.

Acknowledgments
We would like to thank Marta Campaner and Carmen C. Takata (USP), Vladimir Lee, Andrey Matveyev, and Yingping Wang (VCU) for technical support, and Carlisle G. Childress Jr. and J. Michael Davis (VCU Center for High-Performance Computing). Sequencing was performed in the Nucleic Acids Research Facilities, and analyses were performed in the Bioinformatics Computational Core Lab and the Center for High Performance Computing at VCU. The research leading to these results was funded by: the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement n° [247073]10; the French project ANR MIRI BLAN08-1335497; FAPERJ grant coordinated by Dr. Cristina Motta and the FAPERJ-INRIA project RAMPA; by the Laboratoire International Associé (LIA) LIRIO co-coordinated by Ana Tereza R. de Vasconcelos (Labinfo, LNCC, Brazil) and Marie-France Sagot (LBBE, UCBL-CNRS-INRIA, France); the National Science Foundation [USA, grant number NSF DEB-0830056 to Gregory Buck]; and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brazil) to Cristina Motta, Marta M.G. Teixeira and Erney P. Camargo.

Author details
[1]Virginia Commonwealth University, Richmond, VA, USA. [2]BAMBOO Team, INRIA Grenoble-Rhône-Alpes, Villeurbanne, France. [3]Laboratoire Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558, Villeurbanne, France. [4]Laboratório Nacional de Computação Científica, Petrópolis, Rio de Janeiro, Brazil. [5]Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil. [6]Laboratório de Ultraestrutura Celular Hertha Meyer. Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.

Received: 12 June 2013 Accepted: 6 September 2013
Published: 9 September 2013

References

1. Baumann P, Moran NA, Baumann L: **The evolution and genetics of aphid endosymbionts.** *BioScience* 1997, **47:**12–20.
2. Wernegreen JJ: **Genome evolution in bacterial endosymbionts of insects.** *Nat Rev Genet* 2002, **3:**850–861.
3. Wernegreen JJ: **Endosymbiosis: lessons in conflict resolution.** *PLoS Biol* 2004, **2:**E68.
4. Moran NA: **Symbiosis.** *Curr Biol* 2006, **16:**R866–R871.
5. Moran NA, McCutcheon JP, Nakabachi A: **Genomics and evolution of heritable bacterial symbionts.** *Annu Rev Genet* 2008, **42:**165–190.
6. Wernegreen JJ: **Strategies of genomic integration within insect-bacterial mutualisms.** *Biol Bull* 2012, **223:**112–122.
7. McCutcheon JP, von Dohlen CD: **An interdependent metabolic patchwork in the nested symbiosis of mealybugs.** *Curr Biol* 2011, **21:**1366–1372.
8. Horn M, Wagner M: **Bacterial endosymbionts of free-living amoebae.** *J Eukaryot Microbiol* 2004, **51:**509–514.
9. Heinz E, Kolarov I, Kästner C, Toenshoff ER, Wagner M, Horn M: **An *Acanthamoeba* sp. containing two phylogenetically different bacterial endosymbionts.** *Environ Microbiol* 2007, **9:**1604–1609.
10. Nowack ECM, Melkonian M: **Endosymbiotic associations within protists.** *Philos Trans R Soc Lond B Biol Sci* 2010, **365:**699–712.
11. Chang KP, Chang CS, Sassa S: **Heme biosynthesis in bacterium-protozoon symbioses: enzymic defects in host hemoflagellates and complemental role of their intracellular symbiotes.** *Proc Natl Acad Sci USA* 1975, **72:**2979–2983.
12. Roitman I, Camargo EP: **Endosymbionts of trypanosomatidae.** *Parasitol Today (Regul. Ed.)* 1985, **1:**143–144.
13. Du Y, Maslov DA, Chang KP: **Monophyletic origin of beta-division proteobacterial endosymbionts and their coevolution with insect trypanosomatid protozoa *Blastocrithidia culicis* and *Crithidia* spp.** *Proc Natl Acad Sci USA* 1994, **91:**8437–8441.
14. Motta MCM, Catta-Preta CMC, Schenkman S, De Azevedo Martins AC, Miranda K, De Souza W, Elias MC: **The bacterium endosymbiont of *Crithidia deanei* undergoes coordinated division with the host cell nucleus.** *PLoS ONE* 2010, **5:**e12415.
15. Hoare CA: **Herpetosoma from man and other mammals.** In *The Trypanosomes of Mammals: A Zoological Monograph.* Oxford: Blackwell Scientific Publications; 1972:288–314.
16. Wenyon CM: *Protozoology - A Manual for Medical Men, Veterinarians and Zoologists.* London: Bailliere, Tindall and Cox; 1926:1.
17. Wallace FG: **The trypanosomatid parasites of insects and arachnids.** *Exp Parasitol* 1966, **18:**124–193.
18. Vickerman K: **The evolutionary expansion of the trypanosomatid flagellates.** *Int J Parasitol* 1994, **24:**1317–1331.
19. Newton BA: **A synthetic growth medium for the trypanosomid flagellate *Strigomonas (Herpetomonas) oncopelti*.** *Nature* 1956, **177:**279–280.
20. Newton BS: **Nutritional requirements and biosynthetic capabilities of the parasitic flagellate *Strigomonas oncopelti*.** *J Gen Microbiol* 1957, **17:**708–717.
21. Kidder GW, Davis JS, Cousens K: **Citrulline utilization in *Crithidia*.** *Biochem Biophys Res Commun* 1966, **24:**365–369.
22. Mundim MH, Roitman I, Hermans MA, Kitajima EW: **Simple nutrition of *Crithidia deanei*, a reduviid trypanosomatid with an endosymbiont.** *J Protozool* 1974, **21:**518–521.
23. Menezes MCND, Roitman I: **Nutritional requirements of *Blastocrithidia culicis*, a trypanosomatid with an endosymbiont.** *J Eukaryotic Microbiol* 1991, **38:**122–123.
24. Teixeira MMG, Borghesan TC, Ferreira RC, Santos MA, Takata CSA, Campaner M, Nunes VLB, Milder RV, de Souza W, Camargo EP: **Phylogenetic validation of the genera *Angomonas* and *Strigomonas* of trypanosomatids harboring bacterial endosymbionts with the description of New species of trypanosomatids and of proteobacterial symbionts.** *Protist* 2011, **162:**503–524.
25. Alves JMP, Serrano MG, Silva FM D, Voegtly LJ, Matveyev AV, Teixeira MMG, Camargo EP, Buck GA: **Genome evolution and phylogenomic analysis of *candidatus* kinetoplastibacterium, the betaproteobacterial endosymbionts of *Strigomonas* and *Angomonas*.** *Genome Biol Evol* 2013, **5:**338–350.
26. Motta MCM, Martins AC De A, De Souza SS, Catta-Preta CMC, Silva R, Klein CC, De Almeida LGP, De Lima Cunha O, Ciapina LP, Brocchi M, Colabardini AC, De Araujo Lima B, Machado CR, De Almeida Soares CM, Probst CM, De Menezes CBA, Thompson CE, Bartholomeu DC, Gradia DF, Pavoni DP, Grisard EC, Fantinatti-Garboggini F, Marchini FK, Rodrigues-Luiz GF, Wagner G, Goldman GH, Fietto JLR, Elias MC, Goldman MHS, Sagot M-F, Pereira M, Stoco PH, De Mendonça-Neto RP, Teixeira SMR, Maciel TEF, De Oliveira Mendes TA, Urményi TP, De Souza W, Schenkman S, De Vasconcelos ATR: **Predicting the proteins of *Angomonas deanei*, *Strigomonas culicis* and their respective endosymbionts reveals New aspects of the trypanosomatidae family.** *PLoS ONE* 2013, **8:**e60209.
27. Andersson SG, Kurland CG: **Reductive evolution of resident genomes.** *Trends Microbiol* 1998, **6:**263–268.
28. Itoh T, Martin W, Nei M: **Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts.** *Proc Natl Acad Sci USA* 2002, **99:**12944–12948.
29. Gómez-Valero L, Silva FJ, Christophe Simon J, Latorre A: **Genome reduction of the aphid endosymbiont *Buchnera aphidicola* in a recent evolutionary time scale.** *Gene* 2007, **389:**87–95.
30. Palmié-Peixoto IV, Rocha MR, Urbina JA, de Souza W, Einicker-Lamas M, Motta MCM: **Effects of sterol biosynthesis inhibitors on endosymbiont-bearing trypanosomatids.** *FEMS Microbiol Lett* 2006, **255:**33–42.
31. De Azevedo-Martins AC, Frossard ML, de Souza W, Einicker-Lamas M, Motta MCM: **Phosphatidylcholine synthesis in *Crithidia deanei*: the influence of the endosymbiont.** *FEMS Microbiol Lett* 2007, **275:**229–236.
32. De Freitas-Junior PRG, Catta-Preta CMC, da Silva Andrade L, Cavalcanti DP, De Souza W, Einicker-Lamas M, Motta MCM: **Effects of miltefosine on the proliferation, ultrastructure, and phospholipid composition of *Angomonas deanei*, a trypanosomatid protozoan that harbors a symbiotic bacterium.** *FEMS Microbiol Lett* 2012, **333:**129–137.
33. Alfieri SC, Camargo EP: **Trypanosomatidae: isoleucine requirement and threonine deaminase in species with and without endosymbionts.** *Exp Parasitol* 1982, **53:**371–380.
34. Chang KP, Trager W: **Nutritional significance of symbiotic bacteria in two species of hemoflagellates.** *Science* 1974, **183:**531–532.
35. Fair DS, Krassner SM: **Alanine aminotransferase and aspartate aminotransferase in *Leishmania tarentolae*.** *J Protozool* 1971, **18:**441–444.
36. Camargo EP, Freymuller E: **Endosymbiont as supplier of ornithine carbamoyltransferase in a trypanosomatid.** *Nature* 1977, **270:**52–53.
37. Figueiredo EN, Yoshida N, Roitman C, Camargo EP: **Enzymes of the ornithine-arginine metabolism of trypanosomatids of the genus *Crithidia*.** *J Eukaryotic Microbiol* 1978, **25:**546–549.

38. Yoshida N, Jankevicius JV, Roitman I, Camargo EP: **Enzymes of the ornithine-arginine metabolism of trypanosomatids of the genus *Herpetomonas*.** *J Eukaryotic Microbiol* 1978, **25:**550–555.
39. Camargo EP, Coelho JA, Moraes G, Figueiredo EN: **Trypanosoma spp., leishmania spp. and leptomonas spp.: enzymes of ornithine-arginine metabolism.** *Exp Parasitol* 1978, **46:**141–144.
40. Galinari S, Camargo EP: **Trypanosomatid protozoa: survey of acetylornithinase and ornithine acetyltransferase.** *Exp Parasitol* 1978, **46:**277–282.
41. Galinari S, Camargo EP: **Urea cycle enzymes in wild and aposymbiotic strains of *Blastocrithidia culicis*.** *J Parasitology* 1979, **65:**88.
42. Cowperthwaite J, Weber MM, Packer L, Hutner SH: **Nutrition of *Herpetomonas (Strigomonas) culicidarum*.** *Ann N Y Acad Sci* 1953, **56:**972–981.
43. Kidder GW, Dutta BN: **The growth and nutrition of *Crithidia fasciculata*.** *J Gen Microbiol* 1958, **18:**621–638.
44. Guttman HN: **First defined media for *Leptomonas* spp. from insects.** *J Protozool* 1966, **13:**390–392.
45. Guttman HN: **Patterns of methionine and lysine biosynthesis in the trypanosomatidae during growth.** *Journal of Eukaryotic Microbiology* 1967, **14:**267–271.
46. Gutteridge WE: **Presence and properties of diaminopimelic acid decarboxylases in the genus *Crithidia*.** *Biochim Biophys Acta* 1969, **184:**366–373.
47. Krassner SM, Flory B: **Essential amino acids in the culture of *Leishmania tarentolae*.** *J Parasitol* 1971, **57:**917–920.
48. Kidder GW, Dewey VC: **Methionine or folate and phosphoenolpyruvate in the biosynthesis of threonine in *Crithidia fasciculata*.** *J Protozool* 1972, **19:**93–98.
49. Cross GAM, Klein RA, Baker JR: ***Trypanosoma cruzi*: growth, amino acid utilization and drug action in a defined medium.** *Ann Trop Med Parasitology* 1975, **69:**513–514.
50. Anderson SJ, Krassner SM: **Axenic culture of *Trypanosoma cruzi* in a chemically defined medium.** *J Parasitol* 1975, **61:**144–145.
51. Cross GA, Klein RA, Linstead DJ: **Utilization of amino acids by *Trypanosoma brucei* in culture: L-threonine as a precursor for acetate.** *Parasitology* 1975, **71:**311–326.
52. Mundim MH, Roitman I: **Extra nutritional requirements of artificially aposymbiotic *Crithidia deanei*.** *J Eukaryotic Microbiol* 1977, **24:**329–331.
53. Roitman I, Mundim MH, Azevedo HP, Kitajima EW: **Growth of *Crithidia* at high temperature: *Crithidia hutneri* sp. n. and *Crithidia luciliae thermophila* s. sp. n.** *Journal of Eukaryotic Microbiology* 1977, **24:**553–556.
54. Yoshida N, Camargo EP: **Ureotelism and ammonotelism in trypanosomatids.** *J Bacteriol* 1978, **136:**1184–1186.
55. Hutner SH, Bacchi CJ, Baker H: **Nutrition of the kinetoplastida.** In *Biology of the Kinetoplastida. Vol. 2*, Volume 2. Edited by Lumsden WHR, Evans DA. London & New York: Academic; 1979:645–691.
56. Camargo EP, Silva S, Roitman I, Souza W, Jankevicius JV, Dollet M: **Enzymes of ornithine-arginine metabolism in trypanosomatids of the genus *phytomonas*.** *J Eukaryotic Microbiology* 1987, **34:**439–441.
57. Bono H, Ogata H, Goto S, Kanehisa M: **Reconstruction of amino acid biosynthesis pathways from the complete genome sequence.** *Genome Res* 1998, **8:**203–210.
58. Alves JMP, Voegtly L, Matveyev AV, Lara AM, da Silva FM, Serrano MG, Buck GA, Teixeira MMG, Camargo EP: **Identification and phylogenetic analysis of heme synthesis genes in trypanosomatids and their bacterial endosymbionts.** *PLoS ONE* 2011, **6:**e23518.
59. Bhattacharjee JK: **alpha-Aminoadipate pathway for the biosynthesis of lysine in lower eukaryotes.** *Crit Rev Microbiol* 1985, **12:**131–151.
60. Nishida H: **Distribution of genes for lysine biosynthesis through the aminoadipate pathway among prokaryotic genomes.** *Bioinformatics* 2001, **17:**189–191.
61. Velasco AM, Leguina JI, Lazcano A: **Molecular evolution of the lysine biosynthetic pathways.** *J Mol Evol* 2002, **55:**445–459.
62. Hudson AO, Bless C, Macedo P, Chatterjee SP, Singh BK, Gilvarg C, Leustek T: **Biosynthesis of lysine in plants: evidence for a variant of the known bacterial pathways.** *Biochim Biophys Acta* 2005, **1721:**27–36.
63. Torruella G, Suga H, Riutort M, Peretó J, Ruiz-Trillo I: **The evolutionary history of lysine biosynthesis pathways within eukaryotes.** *J Mol Evol* 2009, **69:**240–248.
64. Hutner SH, Provasoli L: **Comparative physiology: nutrition.** *Annu Rev Physiol* 1965, **27:**19–50.
65. Nathan HA, Cowperthwaite J: **Use of the trypanosomid flagellate, *Crithidia fasciculata*, for evaluating antimalarials.** *Proc Soc Exp Biol Med* 1954, **85:**117–119.
66. Janakidevi K, Dewey VC, Kidder GW: **Serotonin in protozoa.** *Arch Biochem Biophys* 1966, **113:**758–759.
67. Hutner SH, Bacchi CJ, Shapiro A, Baker H: **Protozoa as tools for nutrition research.** *Nutr Rev* 1980, **38:**361–364.
68. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H: **Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS.** *Nature* 2000, **407:**81–86.
69. Macdonald SJ, Lin GG, Russell CW, Thomas GH, Douglas AE: **The central role of the host cell in symbiotic nitrogen metabolism.** *Proc Biol Sci* 2012, **279:**2965–2973.
70. Wilson ACC, Ashton PD, Calevro F, Charles H, Colella S, Febvay G, Jander G, Kushlan PF, Macdonald SJ, Schwartz JF, Thomas GH, Douglas AE: **Genomic insight into the amino acid relations of the pea aphid, *Acyrthosiphon pisum*, with its symbiotic bacterium *Buchnera aphidicola*.** *Insect Mol Biol* 2010, **19**(Suppl 2):249–258.
71. Hansen AK, Moran NA: **Aphid genome expression reveals host-symbiont cooperation in the production of amino acids.** *Proc Natl Acad Sci USA* 2011, **108:**2849–2854.
72. Sandström J, Moran N: **How nutritionally imbalanced is phloem sap for aphids?** *Entomol Exp Appl* 1999, **91:**203–210.
73. Dunn MF, Niks D, Ngo H, Barends TRM, Schlichting I: **Tryptophan synthase: the workings of a channeling nanomachine.** *Trends Biochem Sci* 2008, **33:**254–264.
74. Meister A: *Biochemistry of the amino acids*. New York: Academic Press Inc.; 1965.
75. Beutin L, Eisen H: **Regulation of enzymes involved in ornithine/arginine metabolism in the parasitic trypanosomatid *Herpetomonas samuelpessoai*.** *Mol Gen Genet* 1983, **190:**278–283.
76. Frossard ML, Seabra SH, DaMatta RA, de Souza W, de Mello FG, Machado Motta MC: **An endosymbiont positively modulates ornithine decarboxylase in host trypanosomatids.** *Biochem Biophys Res Commun* 2006, **343:**443–449.
77. Andersson JO: **Horizontal gene transfer between microbial eukaryotes.** *Methods Mol Biol* 2009, **532:**473–487.
78. Mormann S, Lömker A, Rückert C, Gaigalat L, Tauch A, Pühler A, Kalinowski J: **Random mutagenesis in *Corynebacterium glutamicum* ATCC 13032 using an IS6100-based transposon vector identified the last unknown gene in the histidine biosynthesis pathway.** *BMC Genomics* 2006, **7:**205.
79. Petersen LN, Marineo S, Mandalà S, Davids F, Sewell BT, Ingle RA: **The missing link in plant histidine biosynthesis: *Arabidopsis* myoinositol monophosphatase-like2 encodes a functional histidinol-phosphate phosphatase.** *Plant Physiol* 2010, **152:**1186–1196.
80. Korený L, Lukes J, Oborník M: **Evolution of the haem synthetic pathway in kinetoplastid flagellates: an essential pathway that is not essential after all?** *Int J Parasitol* 2010, **40:**149–156.
81. Husnik F, Nikoh N, Koga R, Ross L, Duncan RP, Fujie M, Tanaka M, Satoh N, Bachtrog D, Wilson ACC, von Dohlen CD, Fukatsu T, McCutcheon JP: **Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis.** *Cell* 2013, **153:**1567–1578.
82. Camargo EP: **Growth and differentiation in *Trypanosoma cruzi*. I. Origin of metacyclic trypanosomes in liquid media.** *Rev. Inst. Med. Trop. Sao Paulo* 1964, **6:**93–100.
83. Ozaki LS, Czeko YMT: **Genomic DNA cloning and related techniques.** In *Genes and Antigens of Parasites. A Laboratory Manual.* Edited by Morel CM. Rio de Janeiro: Fundação Oswaldo Cruz; 1984:165–185.
84. Alves JMP, Buck GA: **Automated system for gene annotation and metabolic pathway reconstruction using general sequence databases.** *Chem Biodivers* 2007, **4:**2593–2602.
85. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **Uniref: comprehensive and non-redundant uniprot reference clusters.** *Bioinformatics* 2007, **23:**1282–1288.
86. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 1999, **27:**29–34.
87. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12:**1599–1610.
88. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH: **CDD: a conserved domain database for the functional annotation of proteins.** *Nucleic Acids Res* 2011, **39:**D225–D229.

89. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5**:113.
90. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**:2688–2690.
91. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.** *Mol Biol Evol* 2001, **18**:691–699.
92. Stöver BC, Müller KF: **TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses.** *BMC Bioinformatics* 2010, **11**:7.
93. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R: **Dendroscope: an interactive viewer for large phylogenetic trees.** *BMC Bioinformatics* 2007, **8**:460.

doi:10.1186/1471-2148-13-190
Cite this article as: Alves *et al.*: Endosymbiosis in trypanosomatids: the genomic cooperation between bacterium and host in the synthesis of essential amino acids is heavily influenced by multiple horizontal gene transfers. *BMC Evolutionary Biology* 2013 **13**:190.