Bioinformatic study of the metabolic dialog between a non-pathogenic trypanosomatid and its endosymbiont with evolutionary and functional goals

Cecilia Coimbra Klein

N^o d'ordre: 208-2013 Année 2013

Thèse

Présentée

devant l'Université Claude Bernard - Lyon 1

pour l'obtention

du Diplôme de Doctorat

(arrêté du 7 août 2006)

et soutenue publiquement le
12 Novembre 2013

par

Cecilia Coimbra Klein

Bioinformatic study of the metabolic dialog between a non-pathogenic trypanosomatid and its endosymbiont with evolutionary and functional goals

---

Directeur de thèse: Marie-France Sagot
Co-Directeur: Ana Tereza Ribeiro de Vasconcelos

Jury:	Siv Andersson,	Rapporteur
	Michael Barrett,	Rapporteur
	Frédéric Bringaud,	Examinateur
	Christian Gautier,	Examinateur
	Eduardo Rocha,	Rapporteur

UNIVERSITÉ CLAUDE BERNARD-LYON 1

Président de l'Université	**M. M. François-Noël GILLY**
Vice-Président du Conseil d'Administration	M. le Professeur Hamda BEN HADID
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur Philippe LALLE
Vice-Président du Conseil Scientifique	M. le Professeur Germain GILLET
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecin Lyon-Est - Claude Bernard	Directeur: M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur: Mme la Professeure C. BURILLON
UFR d'Ontologie	Directeur: D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur: Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de Réadaptation	Directeur: M. le Professeur Y. MATILLON
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur: M. le Professeur P. FARGE

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur: M. le Professeur S. De MARCHI
Département Biologie	Directeur: M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur: Mme. le Professeur H. PARROT
Département Génie Electrique et des Procédés	Directeur: M. N. SIAUVE
Département Informatique	Directeur: M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur: M. le Professeur A. GOLDMAN
Département Mécanique	Directeur: M. le Professeur H. BEN HADID
Département Physique	Directeur: M. le Professeur S. FLECK
Département Sciences de la Terre	Directeur: M. le Professeur I. DANIEL
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur: M. C. COLLIGNON
Observatoire de Lyon	Directeur: M. B. GUIDERDONI
Ecole Polytechnique Universitaire de Lyon 1	Directeur: M. P. FOURNIER
Ecole Supérieure de Chimie Physique Electronique	Directeur: M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur: M. C. VITON
Institut Universitaire de Formation des Maîtres	Directeur : M. A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Administrateur provisoire : M. N. LEBOISNE

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° [247073]10 SISYPHE.

Contents

Introduction **1**

1 Biological and methodological context **3**
1.1 Symbiosis and cellular evolution . . . 4
1.1.1 Types of symbiosis: a continuum . . . 4
1.1.2 Nutritional mutualists: intricate metabolic exchanges . . . 4
1.1.3 Transition from symbionts to organelles . . . 6
1.2 Trypanosomatids: symbiosis and metabolism . . . 9
1.2.1 Ecological aspects . . . 9
1.2.2 Morphological characteristics and special features . . . 11
1.2.3 Symbiont-harbouring trypanosomatids (SHTs) . . . 13
1.3 Metabolic networks . . . 21
1.3.1 Overview of metabolism . . . 21
1.3.2 Metabolic network reconstruction . . . 22
1.3.3 Modelling of metabolic networks . . . 26

2 Metabolic dialogue between a trypanosomatid and its symbiont **29**
2.1 Predicting the proteins of *Angomonas deanei* and *Strigomonas culicis* and of their respective endosymbionts reveals new aspects of the Trypanosomatidae family . . . 30
2.2 Biosynthetic pathways of amino acids and vitamins . . . 31
2.2.1 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 . . . 32
2.2.2 Biosynthesis of vitamins and cofactors in bacterium-harbouring trypanosomatids depends on the symbiotic association as revealed by genomic analyses . . . 54
2.3 Metabolic networks of host and symbiont . . . 81
2.3.1 Overview . . . 81
2.3.2 Reconstruction of the metabolic networks . . . 81
2.3.3 Metabolic network of *A. deanei* . . . 82
2.3.4 Metabolic reconstruction of the endosymbiont . . . 85
2.3.5 Potential metabolic exchanges between the host and its symbiont . . . 87
2.3.6 Perspectives . . . 89

3 Comparative analyses of metabolic networks **93**
3.1 Exploration of the core metabolism of symbiotic bacteria . . . 94
3.1.1 Background . . . 94

3.1.2 Methods . . . 95
3.1.3 Results . . . 100
3.1.4 Discussion . . . 110
3.1.5 Conclusions . . . 113
3.2 The extended core of metabolic networks . . . 115
3.2.1 Overview . . . 115
3.2.2 Dataset description . . . 115
3.2.3 Computation of the core/periphery reactions . . . 116
3.2.4 Results and discussion . . . 119
3.2.5 Conclusion and perspectives . . . 124

4 Exploring metabolomics data **125**
4.1 Overview . . . 125
4.2 Metabolic stories . . . 125
4.3 Yeast response to cadmium exposure . . . 126
4.4 Perspectives . . . 128

Conclusion and Perspectives **129**

Bibliography **131**

A Published article: Structural and dynamical analysis of biological networks **159**

B Published article: Predicting the Proteins of *Angomonas deanei*, *Strigomonas culicis* and Their Respective Endosymbionts Reveals New Aspects of the Trypanosomatidae Family **175**

C Additional material: 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 **197**

D Published article: Telling metabolic stories to explore metabolomics data: A case study on the Yeast response to cadmium exposure **223**

Introduction

This thesis had two main concerns: metabolism and symbiosis, the latter being explored through the prism of the first and seen as any persistent association between different species. Our work on metabolism spanned from an investigation of classically defined pathways to genome-scale metabolic models, and even reached comparative analyses of a large number of metabolic networks as well as one application of metabolomics. One model deeply investigated was the one of trypanosomatids that harbour one single bacterial symbiont whose interaction is characterised by intensive metabolic exchanges. The comparative analyses performed concerned however also a vaster set of bacterial organisms covering a wider spectrum of associations with their hosts and/or environment.

Symbiont-harbouring trypanosomatids (herein termed SHTs) represent an interesting model to study co-evolution and the evolution of the eukaryotic cell. This is due to the presence of one single betaproteobacterial endosymbiont (herein called TPEs for trypanosomatid proteobacterial endosymbionts) which divides synchronically with the host cell and is vertically transmitted. As bacterial mutualistic symbionts of insects, TPEs show similarities with organelles in terms of genome size and integration. Since the 1950s, nutritional data started to elucidate the diverse nutritional needs of trypanosomatids. It was later evidenced that the ones bearing symbionts had simpler nutritional requirements when compared to their counterparts without symbionts (called RTs for regular trypanosomatids). These studies, together with biochemical data already indicated the enhanced capability of SHTs to synthesise amino acids and hemin, demonstrating in a few cases that the enzyme catalysing the corresponding reactions is encoded by the bacterium. However, it is only the recent genome sequencing of these organisms that is allowing to investigate the presence and genome location of those genes, with all due caution in interpreting the results observed since their expression and functionality remain unknown. These analyses are part of this thesis as concerns the biosynthetic pathways of amino acids and vitamins and the reconstruction and exploration of their whole metabolic networks based on genomic data (for the annotation of the genomes, see: Motta *et al.* (2013), for the analyses on the synthesis of amino acids, see: Alves *et al.* (2013a), and for the synthesis of vitamins, see: Klein *et al.* (2013)). Investigation of the whole metabolic networks is ongoing and is mainly focused on the metabolic exchange between the host mitochondrial, glycosomal and cytoplasmic metabolism and the symbiont. In addition to the above, phylogenetic analyses of the host genes involved in those biosynthetic routes were also performed aiming to characterise potential horizontal gene transfers (HGT) from the symbiont to the host nucleus.

Comparative analyses of metabolic networks became possible only recently thanks to the availability of a large number of genome-scale metabolic models for many organisms, mostly bacteria. These metabolic reconstructions are based on genomic data. Their completeness is a

current limitation as some reactions remain to be discovered and will be missing in the model while some false positive reactions may be wrongly included in the network. The level of annotation of the available genomes is widely heterogeneous across species, making it crucial for any comparative analysis to carefully choose a set of species for which we can guarantee a good enough annotation, and to follow a same procedure for inferring the metabolic network from the annotated genomes. Here, we work at the level of whole metabolic networks for each organism and we analyse the common elements of the small molecule metabolism of diverse bacteria. Working with entire metabolic networks allows for a more systemic view contrasting with the more reductionistic approach based on the analysis of *a priori* selected pathways. Our first work concerning this part of the thesis, entitled *Exploration of the core metabolism of symbiotic bacteria* (Klein *et al.*, 2012b), focuses on symbiotic bacteria, exploring common and variable portions as well as the contribution of different lifestyle groups to the reduction of a common set of metabolic capabilities. The second one treats of *The extended core of metabolic networks*, and is an ongoing analysis of common metabolic capabilities shared by a set of species (not requiring omnipresence) using a new approach where common and group-specific reactions are split automatically. The corresponding method was developed in collaboration with statisticians and is based on the presence/absence of a reaction in an organism. In addition to that, we propose a second approach that relies on a neighbour relationship between reactions.

Finally, our work on metabolomics concerns what has been called *metabolic stories*. These represent possible scenarios explaining the flow of matter among the metabolites in a set of interest based on data from a metabolomics experiment. This approach was proposed, formally defined and modelled by members of the BAMBOO team and we have applied it to data on the response of yeast to cadmium exposure (Milreu *et al.*, 2014). This work can be used in future to better understand the response of an organism to the presence of another species with which it lives in close relation.

This thesis is organised in four main chapters as follows. Chapter 1 introduces some biological and methodological concepts important for the following chapters. Chapter 2 explores this intricate relationship of trypanosomatids and their symbionts focusing mainly on metabolic and evolutionary issues. Chapter 3 comprises the two comparative analyses of metabolic networks of bacteria while Chapter 4 presents the exploration of metabolomics data of yeast exposed to the toxic cadmium. We conclude by presenting the perspectives of this work. The chapters/sections of results are mostly composed of papers either already published, submitted or in preparation.

Chapter 1

Biological and methodological context

Contents

1.1 Symbiosis and cellular evolution 4
1.1.1 Types of symbiosis: a continuum 4
1.1.2 Nutritional mutualists: intricate metabolic exchanges 4
1.1.3 Transition from symbionts to organelles 6
1.2 Trypanosomatids: symbiosis and metabolism 9
1.2.1 Ecological aspects 9
1.2.2 Morphological characteristics and special features 11
1.2.3 Symbiont-harbouring trypanosomatids (SHTs) 13
1.3 Metabolic networks 21
1.3.1 Overview of metabolism 21
1.3.2 Metabolic network reconstruction 22
1.3.3 Modelling of metabolic networks 26

The goal of this chapter is to introduce the biological and methodological concepts that form the basis of this thesis. These mainly include symbiosis, *i.e.* a long-term relationship between two or more different species (Section 1.1) and metabolism (Section 1.3). Symbiosis is more deeply detailed and analysed herein in the case of the trypanosomatids that harbour a symbiotic bacterium in a nutritional mutualistic association (Section 1.2). As this is a long and intricate relationship, it constitutes an important model to study cellular evolution and the possible loss of identity of the symbiotic partners (Sections 1.1.3 and 1.2.3).

1.1 Symbiosis and cellular evolution

1.1.1 Types of symbiosis: a continuum

The term symbiosis (from the Greek living together) was introduced by Anton de Bary (1879) to refer to *any association between different species*, in persistent contact regardless of whether the association is beneficial or not to all participants. This is not however a widely adopted definition which, in many cases, is restricted to the interactions that are beneficial to all participating organisms. Assuming the first definition, we can further classify the wide diversity of symbioses according to the benefits or deficits for each partner which is assessed by comparing its performance (survival, growth, reproductive output, etc) in association or not (Douglas, 2010). When the interaction is advantageous to all, it is called mutualism; when it is beneficial to one while not harming the other species, it is called commensalism; and finally, when one is harmed in the benefit of the other, it is termed parasitism. It is not always an easy task to associate an interaction to one of these categories due to practical difficulties in accessing the changes in the organism's performance with and without the partner. Moreover, there are no precise limits among them as concerns the variability of real associations where environmental changes or other factors imply that benefit is not a fixed trait (Douglas, 2010). Transitions among these categories may happen in the early stages of host adaptation, or may even break down and reverse to autonomy; however increased host dependency and extreme genome reduction may prevent those events from happening (Toft et Andersson, 2010).

Microorganisms intensively interact with eukaryotic cells through symbiotic associations ranging from mutualism to parasitism. The adaptation to a host-associated lifestyle leads the organisms to use common strategies such as the acquisition of essential nutrients from the host cell in parasitic associations or intense metabolic exchanges in some mutualistic associations.

Many protozoan and metazoan cells harbour vertically inherited endosymbionts in their cytoplasm. Prominent examples of such are the associations between gammaproteobacteria and cells lining the digestive tube of insects. Various comprehensive reviews have covered most aspects of these ancient mutualistic relationships, including metabolism, genetics, and evolutionary history of the endosymbiont/host cell associations (Baumann *et al.*, 1997; Wernegreen, 2002, 2004; Moran, 2006; Moran *et al.*, 2008; Wernegreen, 2012; McCutcheon et von Dohlen, 2011). Much less is known about the relationship between protists and their bacterial endosymbionts (Horn et Wagner, 2004; Heinz *et al.*, 2007; Nowack et Melkonian, 2010), including the symbiosis between trypanosomatids and betaproteobacteria (Chang *et al.*, 1975; Roitman et Camargo, 1985; Du *et al.*, 1994a; Motta *et al.*, 2010), herein examined and introduced in Section 1.2.

1.1.2 Nutritional mutualists: intricate metabolic exchanges

The acquisition of metabolic capabilities through a mutualistic symbiosis with bacteria is widespread among eukaryotes. The sap-feeding insects are well studied examples of this (Zientz *et al.*, 2004; Moya *et al.*, 2008; Moran *et al.*, 2008). The great majority of these associations enables the synthesis of essential amino acids not available in the poor diet of the insect hosts such as with *Buchnera* and *Candidatus* Blochmannia (Baumann *et al.*, 1995, 1997; Tamas *et al.*, 2002; Shigenobu *et al.*, 2000; Gil *et al.*, 2003; Zientz *et al.*, 2004; Degnan *et al.*, 2005; Zientz *et al.*, 2006; Pérez-Brocal *et al.*, 2006; Feldhaar *et al.*, 2007; Williams et Wernegreen, 2010). In some cases, the bacterial symbionts are able to produce vitamins of the B complex and cofactors. Such is the case of the endosymbiont, *Wigglesworthia glossinidia*,

of the tsetse fly, and also of *Candidatus* Baumannia cicadellinicola, an endosymbiont of the sharpshooter (Akman *et al.*, 2002; Wu *et al.*, 2006; McCutcheon et Moran, 2007). The latter is in a dual bacterial symbiosis, where one partner (the bacteroidetes *Ca.* Sulcia muelleri) supplies most of the essential amino acids to the host whereas the other (the gammaproteobacterium *Ca.* B. cicadellinicola) provides vitamins and cofactors; this renders the sharpshooter less nutritionally exigent (Wu *et al.*, 2006). *Ca.* Sulcia muelleri has been shown to have other co-resident intracellular symbionts in association with different insect hosts, the alphaproteobaterium *Ca.* Hodgkinia cicadicola in the cicada, and the betaproteobacterium *Ca.* Zinderia insecticola in the spittlebug (McCutcheon *et al.*, 2009; McCutcheon et Moran, 2010). In all three dual symbioses involving *Ca.* Sulcia, this bacterium provides eight (seven in the latter case) essential amino acids to the host leaving to the co-symbiont the role of providing the remaining two (histidine and methionine, plus tryptophan in the case of the co-symbiosis with *Ca.* Zinderia) (Wu *et al.*, 2006; McCutcheon *et al.*, 2009; McCutcheon et Moran, 2010). This can get even more intricate in the nested tripartite symbiosis of mealybugs where the endosymbiont *Candidatus* Tremblaya princeps harbours *Candidatus* Moranella endobia, thus providing the example of a bacterial-bacterial endosymbiosis (von Dohlen *et al.*, 2001; McCutcheon et von Dohlen, 2011; Husnik *et al.*, 2013). Both bacterial partners contribute in a patchwork manner to a same pathway, that is mainly involved in the synthesis of essential amino acids (McCutcheon et von Dohlen, 2011; Husnik *et al.*, 2013).

One remarkable feature of these bacterial symbionts is their extreme genome reduction (Andersson et Kurland, 1998; Tamas *et al.*, 2001; Wernegreen, 2002; Gil *et al.*, 2002; Klasson et Andersson, 2004; McCutcheon, 2010; Moya *et al.*, 2008; McCutcheon et Moran, 2012). Ranging from approximately 0.11 to 14 Mb pairs in length (Husnik *et al.*, 2013; Bennett et Moran, 2013), the smallest bacterial genomes to date are the ones of obligate intracellular (*i.e.* host-restricted) symbionts. Genomic stasis is another striking characteristic, where symbiont genome pairs that diverged by 30 (*Ca.* Blochmannia) to 200 million years (*Ca.* Sulcia) show very stable gene content and order and no rearrangements or duplications, indicating that these endosymbionts are no longer sources of genetic diversification to their hosts (Tamas *et al.*, 2002; Silva *et al.*, 2003; Degnan *et al.*, 2005; McCutcheon *et al.*, 2009). In contrast, a genomic inversion was found in the genome of *Ca.* Tremblaya for which it remains unknown whether it is advantageous and/or a recent event (McCutcheon et von Dohlen, 2011). Other than that, these genomes are generally gene dense (from 73% coding density in *Ca.* Tremblaya to 97% in *Ca.* Carsonella) and AT rich with a few exceptions among the smallest genomes such as *Ca.* Hodgkinia and *Ca.* Tremblaya (McCutcheon et von Dohlen, 2011; McCutcheon et Moran, 2012). The availability of the genome sequences of an important number of host-dependent bacteria allows for a better picture of this process of co-evolution.

The major stages of a genome reduction during host adaptation were described by Toft et Andersson (2010) and McCutcheon et Moran (2012):

1. Stage 1: free-living and extracellular bacterium, *e.g. Escherichia coli* (few pseudogenes, ongoing gene acquisition by horizontal gene transfer (through plasmids, genomic islands and/or bacteriophages) and loss, interstrain recombination, rearrangement, functional divergence, etc);

2. Stage 2: recent host-restricted symbionts or pathogens (facultative intracellular), *e.g. Sodalis glossinidius* (many pseudogenes and mobile elements, large and small deletions and chromosome rearrangements);

3. Stage 3: long-term obligate symbionts or pathogens (obligate intracellular) such as *Buchnera aphidocola*, with genome size ranging from 400-700kb (few pseudogenes, no mobile elements and stable chromosomes);

4. Stage 4: tiny-genome symbionts (obligate intracellular mutualists) such as *Ca.* Tremblaya princeps, with genome size ranging from 140-250kb (ongoing gene loss);

5. Stage 5: organelles (gene loss, gene transfer to the host nuclear genome (genetic assimilation) or replacement by functions encoded by host nuclear genes).

The reasoning for this reductive genome evolution is the accumulation of slightly deleterious mutations (a process termed Muller's Hatchet) (Moran, 1996) due to asexual reproduction, small effective population size and bottlenecks during transmission of those intracellular bacteria (see reviews in Andersson et Kurland (1998); Moya *et al.* (2008); McCutcheon et Moran (2012)). The outcome is, therefore, gene loss and rapid evolution of the protein sequences. This loss is not random and the genes kept are involved in the core information processing (*i.e.* replication, transcription, translation) and in the metabolic routes important for interaction with the host (Klasson et Andersson, 2004; Gil *et al.*, 2004; Moran, 2007; Moran *et al.*, 2008; Moya *et al.*, 2008; McCutcheon et Moran, 2012).

Other interacting partners play an important role as a redundancy of gene functions from another source may relax the selection on some genes, which favours further genome reduction. Such is the case of the nested symbiosis of mealybugs where the betaproteobacterial endosymbiont *Ca.* Tremblaya princeps harbours the gammaproteobacterium *Ca.* Moranella endobia (McCutcheon et von Dohlen, 2011). The latter was possibly more recently acquired, and triggered once again a reductive genome process (even if it was already an extreme case of tiny genome) (Husnik *et al.*, 2013). When a second symbiotic partner joins and the association becomes stable, either both bacteria may be kept as in the above mentioned metabolic complementation of *Ca.* Sulcia muelleri and *Ca.* Baumannia cicadellinicola, or one bacterium may follow an extreme degenerative process ending in its extinction (Moya *et al.*, 2009). The process of a symbiont becoming an organelle is much less clear and will be explored in the next section.

1.1.3 Transition from symbionts to organelles

There are more open questions than answers in this section, however there are promising advances on this subject as more and more case studies of host-restricted symbionts and symbiont-derived organelles have been investigated lately. Even more important than that is the increasing attention given to the under-explored research linking these two topics (Keeling, 2011). Limitations to relate them may remain, since the only two well studied and widely accepted examples of symbiont-derived organelles are mitochondria and plastids which have evolved over a billion years ago and have diversified since then (Douglas, 2010). Overcoming this limitation strongly depends on finding intermediate examples in this long evolutionary path that could give some clues on the important traits of this process. Maybe the recent effort in gathering information in these two research fields will show that this bridge is not that long.

Similarities in terms of genome size and organismal integration of host-restricted symbionts and organelles (Douglas et Raven, 2003; Toft et Andersson, 2010; McCutcheon et Moran, 2012; Husnik *et al.*, 2013) instigate the search for common patterns. One feature of mitochondria and plastids that first draws attention is the gene transfer from the symbiont (alphaproteobacterial

and cyanobacterial ancestors, respectively) to the host nucleus (*i.e.* genetic assimilation) in the course of the evolutionary transition into an organelle (Douglas, 2010; Karlberg *et al.*, 2000; Kurland et Andersson, 2000). It is estimated that more that 90% of the protein-coding genes that act exclusively on these organelles are located in the nucleus, meaning that their proteome is ten times larger than their genome, and that their functioning largely depends on the products of those transferred genes (review in Timmis *et al.* (2004); Douglas (2010)). Gabaldón et Huynen (2007) suggested that the proto-mitochondrion has been hijacked by the eukaryotic host, taking control of its protein synthesis and metabolism. This transfer of symbiont genes to the host nucleus can be a route to compensate for the problem of genome decay of vertically transmitted symbionts, conferring an advantage to a small genome size in a competition among the symbionts in a cell (symbionts with fewer genes can have smaller genomes and replicate faster than those with larger genomes) (Douglas, 2010). In addition to the population dynamics of intracellular bacteria, this minimisation in the size of the genomes, as well as the specific targeting back to the organelle system, may be driven by selection at the host level, preventing the endosymbiont from reverting to autonomy or changing partners; it thus provides a way to stabilise this cooperative relationship (Douglas, 2010; Toft et Andersson, 2010).

Most of the genome sequencing has been done in the symbionts and almost never in their hosts, restricting the analyses of the host interplay and of the symbiont-host gene transfer. One such host genome available and analysed for this purpose is the one of the pea aphid that harbours the gammaproteobacterium *Buchnera aphidicola* (International Aphid Genomics Consortium, 2010; Nikoh *et al.*, 2010). This study revealed no functional symbiont-host gene transfer; conversely, a few genes originating from alphaproteobacteria (possibly from the genus *Wolbachia*) were identified and shown to be highly expressed in the bacteriocyte (a specialised host cell where reside most of the vertically transmitted mutualistic symbionts of animals) (Nikoh *et al.*, 2010). Moreover, the importation of host proteins into *Buchnera* cells was investigated by proteomics, yielding no evidence for a selective transfer (Poliakov *et al.*, 2011). In addition to the pea aphid, the genomes of the body louse and of its primary bacterial endosymbiont *Candidatus* Riesia pediculicola were sequenced, and genes of prokaryotic origin are apparently not present in the host genome, suggesting the absence of symbiont-host gene transfers (Kirkness *et al.*, 2010). In the tripartite nested mealybug symbiosis, multiple lineages seem to contribute to their metabolism, 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, however no symbiont-host gene transfer was found (Husnik *et al.*, 2013). Thus, differing from the symbiont-derived organelles, these findings indicate that symbiont-host gene transfer is not the process enabling survival of these small genome bacteria (McCutcheon et Moran, 2012). McCutcheon et Moran (2012) suggested that symbiont-host gene transfer and/or the importation of host proteins into the symbiont cell might occur in symbionts showing greater genome erosion than *Buchnera*. However, the work of Husnik *et al.* (2013) pointed to the absence of symbiont-host gene transfer in the symbiosis of mealybugs, where *Ca.* Tremblaya princeps has one of the smallest bacterial genomes so far reported. This leaves open the possibility of symbiont-host gene transfer in other tiny-genome symbionts without a nested symbiosis as well as of the import of host proteins into the symbiont, and whether these processes could still happen further on in the evolutionary path from a host-restricted symbiont towards an organelle.

Douglas (2010) explained that genetic assimilation is such a rare event due to its evolutionary difficulty. The first step is the persistent DNA transfer to the nucleus which is a hard condition in multicellular organisms as this may happen in somatic cells preventing its

continuation in the next generation. The case is different in *Wolbachia* which inhabits the reproductive cells (oocytes) and its DNA transfer to the host nucleus has been evidenced in both insects and nematodes (Nikoh *et al.*, 2008; Hotopp *et al.*, 2007). Moreover, nearly the entire *Wolbachia* genome was transferred to the nuclear genome of *Drosophila ananassae* (Hotopp *et al.*, 2007). As single cell eukaryotes, this task is facilitated in the case of protozoa (Douglas, 2010). The second step consists in targeting back to the organelle the protein coded by the transferred gene (Douglas, 2010). This implies a dedicated targeting system, meaning thus that the endosymbiont-turned-organelle depends strictly on its host to maintain its genetic information (Keeling et Archibald, 2008). A previous definition of a bacterial-derived organelle was given by Douglas et Raven (2003) as an intracellular derivative of a symbiotic bacterium with transfer from the symbiont to the nucleus of one or more genes whose product(s) is (are) targeted back to the organelle, accompanied by a loss of the bacterial identity. Such absolute dependency on its host prevents the organelle to switch to a host lineage with which it has not co-evolved (Douglas, 2010). Assuming an organelle is defined by an inescapable reliance on its host for genetic information maintenance, as mentioned before, host-restricted symbionts still keep the core genetic information processing genes (Klasson et Andersson, 2004; Gil *et al.*, 2004; Moran, 2007; Moran *et al.*, 2008; Moya *et al.*, 2008; McCutcheon et Moran, 2012). McCutcheon (2010) compared the gene content of insect symbiont and organelle genomes and found a clear difference in the retained activities even if they have a similar number of predicted genes in some cases, suggesting that the forces governing gene loss in these two groups are different. Most of these obligate symbionts retain more robust gene sets when compared to organelles, and these are considered complete enough to support autonomous life (McCutcheon et Moran, 2012). Striking once again is the nested symbiosis in mealybugs where *Ca.* Tremblaya has lost essential genes that were unprecedentedly reported and are involved in translation, such as both translational release factors and the complete set of functional aminoacyl-tRNA synthetases (McCutcheon, 2010; McCutcheon et von Dohlen, 2011). Since translation potentially still occurs in *Ca.* Tremblaya cells considering the presence of the ribosomal protein genes, the missing proteins could be supplied either by the host (through genetic assimilation of those genes and specific targeting back to *Ca.* Tremblaya) or by *Ca.* Moranella (passively through lysis or specific targeting back) (McCutcheon et von Dohlen, 2011; Keeling, 2011). Taking into account the findings of Husnik *et al.* (2013) showing no *Ca.* Tremblaya-host gene transfer, the second option depending on *Ca.* Moranella seems more plausible, also when one considers their patchwork metabolism. As a general trend until now, nutritional symbionts have a stronger bacterial identity than organelles (McCutcheon, 2010).

Keeling (2011) raised one more feature that is worth paying attention to when comparing endosymbionts and organelles: the degree of functional integration, which was exemplified by the tripartite symbiosis of mealybugs. In this system, both bacterial symbionts are needed to have the complete biosynthetic pathways of essential amino acids which requires the transport of intermediate metabolites between symbionts before the final product is ready and can be used by the three partners (McCutcheon et von Dohlen, 2011). Maybe unusual among symbionts, this feature is described among the different compartments in the eukaryotic cell such as the heme biosynthesis in many eukaryotes that spans the cytosol and the mitochondrion (Keeling, 2011). Apicomplexan parasites appear to have included one more layer, its non-photosynthetic plastid (review in Lim et McFadden (2010)). This partition of metabolic functions among semi-independent compartments might be expected to increase with the complexity of the system (Keeling, 2010; Keeling et Corradi, 2011).

Further investigation of host-restricted symbionts may give insights on the evolutionary

path of symbiont-derived organelles, on the loss of bacterial identity and on the limits between host-restricted symbionts and organelles.

1.2 Trypanosomatids: symbiosis and metabolism

1.2.1 Ecological aspects

Kinetoplastid flagellates: evolution of parasitism

Kinetoplastids are unicellular eukaryotes which contain a range of ubiquitous free-living species. Among them, there are the well studied pathogens of humans and domestic animals of the genera *Trypanosoma* and *Leishmania* causing Chagas disease, sleeping sickness and leishmaniasis (Hoare, 1972). They are part of the family Trypanosomatidae (Euglenozoa, Kinetoplastea) which includes obligate parasites of invertebrates, vertebrates and plants. Most species are non-pathogenic commensals in the digestive tube of insects (Wenyon, 1926; Wallace, 1966; Vickerman, 1994). The ancestral trypanosomatids were probably parasites of insects (insect-first model) and their closest relative is likely to be the free-living *Bodo saltans* (Simpson *et al.*, 2006; Deschamps *et al.*, 2011) (Figure 1.1). In addition to this group of parasites among the free-living kinetoplastids, there are other clades which indicate that parasitism evolved more than once in kinetoplastids (Simpson *et al.*, 2006). Assuming that there were no reversions to a free-living state, Simpson *et al.* (2006) suggest that there were at least four independent adoptions of obligate parasitism or commensalism (see Figure 1.1). The above mentioned insect-first model for the ancestry of trypanosomatids is currently the most accepted where these flagellates descended from parasites of blood-sucking insects that survived accidental transmission into a vertebrate host during feeding (Simpson *et al.*, 2006). Even if such transmission must have occurred often, the rare successful cases would presumably open a large niche to the parasite (Simpson *et al.*, 2006).

Life cycle and the interaction between the flagellate and its insect host

Trypanosomatids include both monoxenic insect parasites (termed insect trypanosomatids herein) and heteroxenic taxa that alternate between insects and vertebrates (or plants). The latter trypanosomatids reach a secondary host via an insect vector, such as vertebrates via blood-sucking insects and plants via phytophagous bugs (Wallace, 1966).

The focus of the present work will be on the monoxenic insect parasites which will be more detailed hereafter. On the other hand, the heteroxenics, such as *Trypanosoma* and *Leishmania* spp., will be used as comparative models when pertinent since they have been more investigated due to their medical interest.

There are few studies of the complex interaction between insect trypanosomatids and their hosts (Wallace, 1966; Schaub et Sehnitker, 1988; Schaub, 1988; Podlipaev, 2000; Nascimento *et al.*, 2010). In addition to those, some studies focused on the composition of the cell surface that are important for specific recognition and adherence between parasites and host cells such as glycoconjugates (D'Avila-Levy *et al.*, 2005; Nogueira de Melo *et al.*, 2006; D'Avila-Levy *et al.*, 2008; Pereira *et al.*, 2009). The trypanosomatids develop within different insect organs, including the midgut, hindgut, Malphigian tubes, salivary glands and haemocoel (Wallace, 1966). The monoxenic protozoan *Blastocrithidia triatomae* colonises the digestive tract of the insect *Triatoma infestans* which is also the host of pathogenic heteroxenic trypanosomatids such as *Trypanosoma cruzi* (Schaub et Sehnitker, 1988; Schaub, 1988). The monoxenic *Strigomonas culicis* interacts mainly with midgut cells of *Aedes aegypt* through its flagellum, which penetrates the microvilli preferentially near the tight junctions; the protozoan may reach the hemocel in cases of prolonged infections (Corrêa-da-silva *et al.*, 2006). Nascimento *et al.* (2010) showed that this same trypanosomatid is able to adhere and invade the salivary

Figure 1 page 169 from Simpson *et al.* (2006).

Figure 1.1: Evolutionary relationships among kinetoplastids. Extracted from Simpson *et al.* (2006).

glands of *A. aegypt*, reaching the acinar space where the saliva is stored. This suggests that vector transmission of monoxenic trypanosomatids to vertebrate host may occur in nature (Corrêa-da-silva *et al.*, 2006). Although these trypanosomatids are considered nonpathogenic to mammals, they have been reported to infect different vertebrate hosts (McGhee, 1957; Jansen *et al.*, 1988; Pacheco *et al.*, 1998; Morio *et al.*, 2008; Barreto-de-Souza *et al.*, 2008).

Diversity of insect hosts

The insect trypanosomatids are mostly found in Diptera and Hemiptera. However, this may be underestimated, since their presence has been investigated only in a minority of insect taxa from limited locations (Podlipaev, 2000). Several trypanosomatid species were found within one insect specimen and, conversely, the same parasite was identified in a wide range of insect hosts in a large geographical area (Podlipaev *et al.*, 2004). One such example is the promiscuous *Strigomonas culicis* isolated from mosquitoes such as *Aedes vexans*, *Culex pipens*, *Mansonia richardii* and *Anopheles maculipennis* (Wallace, 1966). This lack of specificity is not equivalent to stating that insect trypanosomatids do not have host preferences and it is conceivable that some species may be restricted to a particular host and/or place (Borghesan *et al.*, 2013). The *Herpetomonas* species showed a marked preference for dipterans (Borghesan

et al., 2013). This low level of host specificity, more characterised in the case of Hemipteran hosts, indicates that co-evolution of the partners is unlikely (Podlipaev, 2000). Moreover, these interactions may be more or less "occasional", in the sense that not only insects but also plants and other organisms could be involved, increasing the chances of the establishment of new host-parasite systems (Podlipaev, 2000).

1.2.2 Morphological characteristics and special features

Figure 1.2 shows a general view of a trypanosomatid (reviewed by de Souza (2002); de Souza et da Cunha-e Silva (2003)). The protozoan is surrounded by a typical plasma membrane, and its flagellum emerges from the basal body located in the anterior region of the cell and projects forward. The flagellum is responsible for motility and participates in the interaction with the host by adhering to the insect digestive tract. In addition to the typical flagellar structure, the protozoan exhibits a unique paraflagellar rod which is a highly elaborated network of filamentous structures connected to the axoneme. The nucleus is centrally located and the cytoplasm contains randomly distributed ribosomes and profiles of the endoplasmic reticulum. The Golgi complex is located in the anterior region, close to the flagellar pocket.

A. Figure 1 page 155 from de Souza et da Cunha-e Silva (2003).
B. Figure 1 page 252 from Docampo *et al.* (2005).

Figure 1.2: **A general view of a trypanosomatid. A.** A thin section of the trypanosomatid *Herpetomonas anglusteri* showing structures such as the flagellum (F), the nucleus (N), the kinetoplast (K), mitochondria (M), glycosomes (G) and the flagellar pocket (FP), examined by transmission electron microscopy. Bar = 0.5 μm. **Extracted from** de Souza et da Cunha-e Silva (2003). **B.** A schematic representation of longitudinal section of an epimastigote form of *T. cruzi*. **Extracted from** Docampo *et al.* (2005). **C.** The main cellular forms of trypanosomatids as defined by cell shape, flagellum presence and attachment (1), and position of the basal body (2), kinetoplast (3) and nucleus (4). A- Amastigote B- Epimastigote C- Trypomastigote D- Choanomastigote E- Promastigote F- Paramastigote G- Opistomastigote.

Other than these typical eukaryotic organelles, there are some specific to the trypanosomatids (reviewed by de Souza (2002) and by de Souza et da Cunha-e Silva (2003)). One is the kinetoplast, which corresponds to the extranuclear DNA that lies within the unique ramified mitochondrion, and is localised in front of the basal body that gives rise to the flagellum.

Another one is the glycosome, a specialised peroxisome that contains most of the enzymes involved in the glycolytic pathway (Opperdoes et Borst, 1977). In addition to those, there are acidocalcisomes, which were first described in trypanosomatids and have been characterised in most detail in this group of organisms, however they are actually conserved from bacteria to mammals (Vercesi *et al.*, 1994; de Souza et da Cunha-e Silva, 2003; Docampo *et al.*, 2005; Moreno et Docampo, 2009; Docampo *et al.*, 2010; Docampo et Moreno, 2011).

The kinetoplast

One of the most striking features of trypanosomatids is their mitochondrial DNA, termed kinetoplast DNA (kDNA). It is one of the largest organellar genomes and is a network of thousands of interlocked DNA rings of two types: thousands of minicircles (small DNA ring - 0.5 to 2.5 kbp) and dozens of maxicircles (large DNA ring - 20 to 40 kbp) (review in Jensen et Englund (2012)). While maxicircles encode rRNAs and some subunits of the mitochondrial bioenergetics machinery (including subunits of cytochrome oxidase, NADH dehydrogenase, and the ATP synthase), minicircles encode most of the guide RNAs that control the specificity for editing maxicircle transcripts. This editing consists in adding or removing uridylate residues from specific internal sites within the transcript to form functional messenger RNAs. This extensive editing requires a large number of different minicircles; moreover, the lack of even one class of those accounts for incomplete editing of the maxicircle transcripts and leads to the death of the parasite. It is a complex and energy consuming mitochondrial RNA editing (Simpson *et al.*, 2006). The synthesis of maxicircle mRNAs depends on a nuclear-encoded single subunit mtRNAP (mitochondrial RNA polymerase), and this process remains unknown in the case of the minicircles. Moreover, the kDNA contains no tRNA genes, which are transcribed in the nucleus and imported into the mitochondrion using a variety of targeting signals (review in Campbell *et al.* (2003)).

The glycosome

The correspondence between the trypanosomatid glycosome and the peroxisome was based on the presence of catalase and enzymes involved in the β-oxidation of lipids in some monoxenic species (*e.g. Crithidia fasciculata* and *Herpetomonas samueli*), and on the conservation of protein import processes and of the same kind of topogenic signals (Parsons *et al.*, 2001; Gualdrón-López *et al.*, 2012). Conversely, heteroxenic trypanosomatids such as *T. brucei*, *T. cruzi* and *Leishmania* had no significant catalase activity detected (de Souza, 2002). Other than the above mentioned metabolic routes, additional pathways are observed in the glycosomes while they occur in the cytosol of other cells; this is the case for most of the glycolytic pathway, carbon dioxide fixation, purine salvage and pyrimidine *de novo* biosynthesis (reviewed by Michels *et al.* (2000); Parsons *et al.* (2001); Hannaert *et al.* (2003); Michels *et al.* (2006); Gualdrón-López *et al.* (2012)). Glycosomes were suggested to have originate in a free-living common ancestor of kinetoplastids and Diplonemida (see Figure 1.1) (Gualdrón-López *et al.*, 2012). Since the glycosome has no genome, all its proteins are encoded by nuclear genes and post-translationally imported into the organelle (de Souza, 2002). The import machinery of the family of peroxisomes is unique (Lanyon-Hogg *et al.*, 2010). Differing from the mitochondrial and chloroplast protein import, the evidence that proteins lacking a peroxisomal targeting sequence (PTS) can be imported into this organelle when associated with a protein bearing a PTS suggests that proteins may be imported in a folded or oligomeric state (review in Parsons *et al.* (2001); Lanyon-Hogg *et al.* (2010); Rucktäschel *et al.* (2011); Theodoulou *et al.* (2013)).

As stated above, the presence of the glycolytic enzymes for the conversion of glucose into 3-phosphoglycerate (3PGA) distinguishes the glycosome from the peroxisome (Opperdoes et Borst, 1977). There is no net ATP synthesis within the glycosome and the 3PGA produced within the organelle is further metabolised into pyruvate generating ATP in the cytosol (Opperdoes et Borst, 1977). Glycosomal glycolytic enzymes show stage-specific changes in abundance (for cellular forms of trypanosomatids see Figure 1.2C and the topic below *The main cellular forms of trypanosomatids*); for example in the heteroxenic *T. brucei* the levels of those enzymes are much higher in bloodstream forms than procyclic forms. This is due to the energy generation in the first through glycolysis and in the latter through cytochrome-mediated respiration (review in Michels *et al.* (2000); Parsons *et al.* (2001); Hannaert *et al.* (2003); Michels *et al.* (2006); Gualdrón-López *et al.* (2012)). The metabolic compartmentation of the glycolytic pathway may be related to an increased metabolic flexibility, accounting for a more readily and efficient adaptation of the organism to different environmental conditions (Gualdrón-López *et al.*, 2012). Furthermore, Gualdrón-López *et al.* (2012) propose that glycosomes played a facilitating role in the multiple development of parasitism and its elaborated life cycles involving different hosts in the kinetoplastids.

The acidocalcisome

The acidocalcisomes are acidic organelles involved in the storage of cations and phosphorous, in the metabolism of pyrophosphate (PPi) and polyphosphate (polyP), in the regulation of the cytoplasmic concentration of calcium, in the maintenance of intracellular pH and in osmoregulation (recently reviewed by Docampo *et al.* (2010); Docampo et Moreno (2011)). Several enzymes and transporters were identified in acidocalcisomes of protists (Docampo et Moreno, 2011). Furthermore, polyP functions as an energy source to replace ATP; in cell membrane alterations that might be related to a channel for DNA import; in responses to nutritional limitations and environmental stresses; in cellular growth and virulence of pathogens (Kornberg *et al.*, 1999; Rao *et al.*, 2009). Moreover, in the acidocalcisomes of parasitic protozoa, reduced levels of polyP were found to be related to decreased virulence and ability to respond to osmotic or nutritional stresses (Lemercier *et al.*, 2004; Luo *et al.*, 2005; Docampo *et al.*, 2011).

The unusual features of trypanosomatids genomes

In addition to the mitochondrial genome, the nuclear genome of trypanosomatids presents some uncommon characteristics. Most of the protein-coding genes are arranged in giant polycistronic clusters such that tens-to-hundreds of functionally unrelated genes are co-transcribed (Campbell *et al.*, 2003; Martínez-calvillo *et al.*, 2004; Martínez-Calvillo *et al.*, 2010). These genomes are almost devoid of introns (El-Sayed *et al.*, 2005). The mRNA is cleaved into single gene transcripts that are *trans*-spliced to small spliced leader RNAs (Campbell *et al.*, 2003). These features are common in kinetoplastids and pre-date the adoption of parasitism (Simpson *et al.*, 2006).

The main cellular forms of trypanosomatids

Figure 1.2C shows the main cellular forms of trypanosomatids as defined by cell shape, flagellum presence and attachment, and position of the basal body, kinetoplast and nucleus (Docampo *et al.*, 2005). The monoxenics generally present one form as they inhabit only one host

whereas heteroxenics differentiate during their life cycle as they alternate hosts. One such example of monoxenic is the choanomastigote *Angomonas deanei* (Teixeira *et al.*, 2011).

1.2.3 Symbiont-harbouring trypanosomatids (SHTs)

The non-pathogenic, insect-exclusive parasites comprise the largest number of trypanosomatid species, and the digestive tube of dipterans and hemipterans represents their most common habitat. Cultures of insect trypanosomatids, also referred to as monoxenics, were first obtained in the 1920s. However, most designated species of these protozoa have not been cultivated and are only known from morphological descriptions recorded in drawings published from the end of the nineteenth century on (Noguchi et Tilden, 1926). The modest number of available cultures of insect trypanosomatids is in part due to the difficulties inherent to growing these organisms in artificial media. This is related to the fastidiousness of insect trypanosomatids, which require nutritionally very rich and complex media in order to grow (Lwoff, 1940; Cowperthwaite *et al.*, 1953; Guttman, 1966). The first defined medium for an insect trypanosomatid was published in 1958 (Kidder et Dutta, 1958), as an attempt to cultivate *Crithidia fasciculata*, a species isolated from mosquitoes. The identity of the flagellate, however, cannot be taken at face value because some confusion prevailed at the time (and even today) with respect to the authenticity of strains and species of insect trypanosomatids.

In most cases, cultivation of insect trypanosomatids required all essential amino acids, vitamins of the B-complex, para-aminobenzoate (pABA), inositol, and choline, in addition to purines, glucose, and salts (Guttman, 1966; Kidder et Dutta, 1958). Earlier, Newton (Newton, 1956, 1957) had described the much simpler nutritional requirements of *Strigomonas oncopelti*, which in addition to the B vitamins needed only methionine, adenine, glucose, and salts for its growth. Later, it was shown that *S. oncopelti* carries a symbiotic bacterium in its cytoplasm (Gill et Vogel, 1963), an observation soon extended to some other insect trypanosomatids (Table 1.1) (Mundim *et al.*, 1974; Faria e Silva *et al.*, 1991; Chang, 1975; Chang et Trager, 1974; Teixeira *et al.*, 2011). This reduced group of insect trypanosomatids carries cytoplasmic endosymbionts (referred to as TPEs for trypanosomatid proteobacterial endosymbionts) and is known as symbiont harbouring trypanosomatids (SHTs), to distinguish them from regular insect trypanosomatids naturally lacking symbionts (RTs). SHTs comprise six species that belong to the genera *Strigomonas* and *Angomonas*, and they form a monophyletic cluster split in two subclades, one for each genus (Figure 1.3) (Hollar *et al.*, 1998; Teixeira *et al.*, 2011).

SHTs	Previous names	Bacterial endosymbiont	Insect host
Angomonas deanei[1]	*C. deanei; H. roitmani*	*Ca.* Kinetoplastibacterium crithidii	H/D
Angomonas desouzai[2]	*C. desouzai*	*Ca.* K. desouzaii	D
Angomonas ambiguus[3]	-	*Ca.* K. crithidii	D
Strigomonas culicis[4]	*Blastocrithidia culicis*	*Ca.* K. blastocrithidii	H/D
Strigomonas oncopelti[5]	*Crithidia oncopelti*	*Ca.* K. oncopeltii	H
Strigomonas galati[6]	-	*Ca.* K. galatii	D

Table 1.1: **The six species of symbiont-harbouring trypanosomatids, respective symbionts and insect host origin.** H: Hemiptera; D: Diptera. [1]Mundim *et al.* (1974); Fiorini (1989); Faria e Silva *et al.* (1991); [2]Fiorini (1989); [3]Teixeira *et al.* (2011) [4]Novy *et al.* (1907); [5]Newton et Horne (1957); [6]Teixeira *et al.* (2011); .

A considerable amount of information has been gathered about the morphology and cell biology of the host/symbiont association (Motta *et al.*, 2010; Freymuller et Camargo, 1981;

Roitman et Camargo, 1985; Motta, 2010). From early on, it was suspected that the symbiont was responsible for the enhanced nutritional capabilities of the SHTs, a fact supported by the loss of these capabilities in strains cured of the symbiont (aposymbiotic strains) by chloramphenicol treatment (Guttman et Eisenman, 1965; Mundim et Roitman, 1977; Chang et Trager, 1974). Further nutritional studies have shown that, indeed, the requirements of the SHTs are minimal compared to those of RTs (Mundim *et al.*, 1974; de Menezes et Roitmanz, 1991).

Figure 1 page 507 from Teixeira *et al.* (2011).

Figure 1.3: **Phylogenetic tree of symbiont-harbouring trypanosomatids and representatives of distinct trypanosomatid genera inferred by maximum likelihood. Extracted from** Teixeira *et al.* (2011).

Mutualistic association and its origin

SHTs and TPEs establish an obligate mutualistic association, in which the symbiont is unable to survive and replicate without its host, whereas the aposymbiotic trypanosomatid loses its ability to colonise the insects (Fampa *et al.*, 2003; Motta, 2010). The original association of TPEs with an ancestral trypanosomatid is thought to have occurred 40-120 million years ago, based on the genetic distances of the bacterial SSU rRNA genes and on an evolutionary rate

of 0.01-0.02 per site per 50 million years (Du *et al.*, 1994a; Moran *et al.*, 1993). Phylogenetic and phylogenomic analyses indicate a common origin for all TPEs clustering within the betaproteobacteria from the Alcaligenaceae family (*Taylorella* genus as sister group) (Figure 1.4), thus suggesting that a single event gave rise to this symbiotic relationship (Du *et al.*, 1994a,b; Teixeira *et al.*, 2011; Alves *et al.*, 2013b). The clade of the TPEs is divided in two subclades, similar to the protozoan host tree, one for the symbionts of *Angomonas* hosts and the other for those of the *Strigomonas* hosts (Teixeira *et al.*, 2011; Alves *et al.*, 2013b). Teixeira *et al.* (2011) performed a congruence analysis between the ITS rDNA-based phylogenetic trees of TPEs and SHTs, which indicated perfect congruence at the genus level and partial at the species level. Assuming the common origin of all TPEs, the authors suggested an overall host-symbiont co-divergence and different rates of evolution for symbionts and hosts.

As concerns the acquisition of endosymbionts by protozoan hosts, Du *et al.* (1994a) suggested that this event might have occurred when the ancestral trypanosomatid still fed upon bacteria to recruit endosymbionts, which points to the free-living and still bacteriovore ancestors of trypanosomatids, *Bodo saltans* (for the ancestry of trypanosomatids see Section 1.2.1). SHTs are neither phagocytic nor susceptible to experimental infection by symbionts or other bacteria. Furthermore, these same authors proposed that a single event of acquisition of a bacterium by phagotrophy by a *Bodo*-like ancestor gave rise to the contemporary endosymbioses in SHTs. Limitations to this proposal were highlighted by the authors: either there were multiple losses of the symbionts from the RTs that interrupt the evolutionary descendence of the SHTs from *Bodo*; or SHTs descend from another still unidentified ancestral lineage. Supporting the *Bodo*-like ancestor, a cytoplasmic endosymbiont was described for the bacteriovore *B. saltans* (Brooker, 1971). Using light and electron microscopy, Brooker (1971) identified as many as 4 bacteria and suggested a total population much larger. Furthermore, mid-point constrictions were observed indicating independent division of the endosymbiont and the host cell (Brooker, 1971). In addition to *B. saltans*, for the first time in an heteroxenic trypanosomatid, an endosymbiont was described in all the stages of the life cycle of the fish *Trypanosoma cobitis* (Lewis et Ball, 1981). These microorganisms are of the gram-negative type and their division does not seem to be synchronised with that of the host (Lewis et Ball, 1981).

Bacterial endosymbiont

The bacterium is in close association with the host cell nucleus, surrounded by glycosomes, and it presents different shapes during the cell cycle of the protozoan host (for more details see Section 1.2.3) (Motta *et al.*, 1997a; Faria-e Silva *et al.*, 2000; Motta *et al.*, 2010). Moreover, it is enclosed by 2 unit membranes and a reduced peptidoglycan layer, possibly facilitating the intense metabolic exchanges with the host and playing important physiological roles in shape maintenance and bacterial division (Motta *et al.*, 1997b). As concerns prokaryote division, similar to the mitochondria of animal, fungi and higher plants but different from overall bacteria, TPEs do not form the FtsZ ring and lack the septum (Motta *et al.*, 2004; Margolin, 2005). Some symbionts with an extreme genome reduction have lost the *fts*Z gene which might have been transferred to the host nucleus (as for the plastids of plants and algae) or replaced by host-derived functions, such as the dynamin-like protein ring found in most mitochondria, plastids as well as in the chloroplast-like organelle of Aplicomplexa parasites, the apicoplast (see reviews in Margolin (2005); Vaishnava et Striepen (2006); Adams et Errington (2009); Bernander et Ettema (2010); McFadden (2011)). A fragment containing the *fts*Z gene transferred from *Wolbachia*, a bacterial endosymbiont, to the X chromosome of an insect host was

A. Figure 2 page 3083 from Du *et al.* (1994b).
B. Figure 2 page 342 from Teixeira *et al.* (2011).

Figure 1.4: **Phylogenetic and phylogenomic analyses of trypanosomatid proteobacterial endosymbionts (TPEs). A.** Phylogenetic analysis of 16S rRNA gene sequences of TPEs and other bacteria. **Extracted from** Du *et al.* (1994b). **B.** Maximum likelihood supermatrix phylogeny (233 concatenated orthologs) of TPEs and other bacteria. **Extracted from** Alves *et al.* (2013b).

reported, however it was proven to be non functional (Kondo *et al.*, 2002). On the other hand, some bacteria which are known to lack this gene, such as Chlamydiae, Planctomycetes, *Ureaplasma urealyticum* and *Mycoplasma mobile*, seem to be capable of independent cell division (McFadden, 2011). Moreover, the typical eukaryotic membrane phospholipid, phosphatydilcholine (PC), is present in the membranes of TPEs and part of PC or of a PC precursor is supplied by the host, indicating that this phospholipid is important for the establishment of the symbiosis in trypanosomatids (Palmié-Peixoto *et al.*, 2006; de Azevedo-Martins *et al.*, 2007). PC is not a frequent constituent of bacterial membranes, however it is found in symbiotic and pathogenic bacteria interacting with plants and animals, such as the nitrogen fixing symbionts of plants and the human pathogens, *Brucella abortus* and *Legionella pneumophila*, that require PC for full virulence (see review in Aktas *et al.* (2010)).

This long partnership has led to considerable changes in the genomes of TPEs including gene loss, with clear preferential retention of genes involved in metabolic collaboration with the host, and consequent genomic size reduction (Alves *et al.*, 2013b; Motta *et al.*, 2013), as seen in other obligatory symbiotic associations (Baumann *et al.*, 1997; Wernegreen, 2002; McCutcheon et von Dohlen, 2011; Andersson et Kurland, 1998; Itoh *et al.*, 2002; Gómez-

Valero *et al.*, 2007). Genomic stasis is also a common feature of host-restricted symbionts and is found in TPEs where the five sequenced genomes are highly syntenic despite million years of divergence (Alves *et al.*, 2013b). The loss of some but not all of the DNA recombination genes may be related to this process (Alves *et al.*, 2013b). Moreover, these genomes are AT rich with only about 30-32% CG content, similar to the obligate symbiont of the sharpshooter *Ca.* Baumannia cicadellinicola (Wu *et al.*, 2006; Alves *et al.*, 2013b).

Changes in the presence of symbionts

The presence of the symbiont results in morphological and physico-chemical alterations in the trypanosomatid host (see review in Motta (2010)). As concerns the first, such changes include the rearrangement of kinetoplast DNA fibers and reduced paraflagellar structure (Freymuller et Camargo, 1981; Gadelha *et al.*, 2005; Cavalcanti *et al.*, 2008). Contrary to the tightly packed kDNA fibers in RTs, these fibers display a looser arrangement in SHTs (Cavalcanti *et al.*, 2008). The paraflagellar rod is important for full motility and adhesion to the host epithelia; moreover the reduced nature of this structure does not result in problems for the SHTs which keep these capabilities (Gadelha *et al.*, 2005). In addition to that, the glycosomes, which are generally distributed throughout the cells, are concentrated around the symbiont in SHTs (Motta *et al.*, 1997a).

The composition of the cell surface, which is of key importance for the interaction with the insect host cell and for the cellular response to environmental stimuli, is different in SHTs and in their aposymbiotic counterparts (Motta, 2010; D'Avila-Levy *et al.*, 2005). For this reason, endosymbiont-bearing strains interact better with insect cells and guts when compared to the aposymbiotic counterparts; this is related to changes in polysaccharides, glycoprotein and carbohydrate composition (Dwyer et Chang, 1976; Fampa *et al.*, 2003; D'Avila-Levy *et al.*, 2005). The highly negative surface charge of the symbiont-free *A. deanei* is slightly reduced by the presence of the endosymbiont (Oda *et al.*, 1984). The altered profile of glycoconjugates, that are important for specific recognition between parasites and host cells, impairs the interaction of the aposymbiotic strains with the insect cells and guts (D'Avila-Levy *et al.*, 2005). Similarly, the protozoan *A. deanei* is considerably more prone to adhere to the explanted guts of *Aedes aegypti* than the aposymbiotic parasite due to the higher expression of surface gp63 molecules (glycosylphosphatidylinositol) which is influenced by the presence of the endosymbiont (D'Avila-Levy *et al.*, 2008).

Coordinated cell division

The single vertically transmitted betaproteobacterial symbiont divides synchronously with other host cell structures (Figure 1.5) (Motta *et al.*, 2010). The cell cycle of trypanosomatids involves a coordinated replication and segregation of the flagellum, the kinetoplast and the nucleus. This process is more complex in SHTs in which the endosymbiont lies down over the host nucleus and is the first structure to divide. It is followed by the segregation of the kinetoplast and the nucleus. The association of the bacterium and the nucleus is suggested to be related to the maintenance of precisely one symbiont per daughter cell. This restriction of a single symbiont per cell indicates that the host protozoan imposes tight control over the fission of the endosymbiont which may also be related to the genes responsible for the prokaryote cell division such as the above mentioned *fts*Z. Douglas (2010) indicated that the restriction of the habitat space, the growth and the proliferation of its partners are possible ways to exercise a control over the abundance and distribution of the partners, and this control is essential for the persistence of symbiosis. As in the TPEs, in *Ca.* Blochmannia most genes are single copy,

and it was suggested that its overall level of gene expression could be influenced by controlling the replication of the bacterium in certain life stages of the insect host (Stoll *et al.*, 2009). In SHTs, the control over the abundance is extremely strict allowing only a single bacterium during all the protozoan host life cycle (Motta *et al.*, 2010); therefore abundance is not the strategy to regulate the gene expression of this bacterium.

Figure 5 page 7 from Motta *et al.* (2010).

Figure 1.5: **Schematic representation that summarizes the morphological alterations during the *A. deanei* cell cycle.** Recently replicated protozoa present a single symbiotic bacterium in rod-shape format (A), the endosymbiont elongates and lies down over the host cell nucleus (B). The bacterium is the first structure to divide (C). After the symbiont duplication, the kinetoplast migrates to the posterior end of the host protozoan (arrow) and the new flagellum grows inside the flagellar pocket (D–E). Then, the kinetoplast segregates (F) and the nucleus divides (G). When the cytokinesis begins, the duplicated bacteria are seen in the posterior end of the protozoan, as well as the duplicated kinetoplasts, considering the nuclear position (G). As the cytokinesis advances, kinetoplasts return to the anterior cell end (arrows), while the symbiont remains in the posterior part of the cell body (H–I). The new flagellum only emerges from the flagellar pocket at the end of cytokinesis, when the flagellar pocket probably segregates (Fig. 5 H–J). The flagellar beat in opposite directions (arrows) generates a propelling force in a late dividing protozoan (Fig. 5 I). At the end of the division process each daughter cell contains a single copy structure, including the symbiotic bacterium (J). The symbiont remains as a single rod-shape bacterium for 1.0 h (A), whereas the constricted symbiont persists in this format for 3 h (B and C'). After the symbiont division, both bacteria are maintained in the host trypanosomatid for 2 h, before the generation of two new daughter cells (C–K). **Extracted from** Motta *et al.* (2010).

Metabolic exchanges

The mutualistic association of the host trypanosomatid and its endosymbiont confers the host protozoan less stringent nutritional requirements because of the endosymbiont supply of essential growth factors (Chang, 1975; Chang *et al.*, 1975; Newton, 1956, 1957; Mundim *et al.*, 1974). Extensive comparative studies between SHTs (wild and cured strains, obtained after antibiotic treatment) and RTs have permitted inferences about the symbiont dependence and contribution to the overall metabolism, in particular the phospholipid (Palmié-Peixoto *et al.*,

2006; de Azevedo-Martins *et al.*, 2007; de Freitas-Junior *et al.*, 2012) and amino acid (Alfieri et Camargo, 1982; Chang et Trager, 1974; Fair et Krassner, 1971; Camargo et Freymuller, 1977; Figueiredo *et al.*, 1978b; Yoshida *et al.*, 1978; Camargo *et al.*, 1987; Galinari et Camargo, 1978, 1979) production of the host cell. In a few cases, it has been shown that the symbiotic bacterium contains enzymes involved in the biosynthetic pathways of the host, but in most cases the metabolic contribution of the endosymbiont has been inferred from nutritional data rather than been genetically demonstrated (Newton, 1956, 1957; Mundim *et al.*, 1974; Chang, 1975; Chang *et al.*, 1975; Roitman et Camargo, 1985; de Menezes et Roitmanz, 1991; Motta, 2010). From these nutritional studies, it was suggested that SHTs require neither hemin nor the amino acids that are essential for the growth of RTs. Some biochemical studies indicated the complementarity of both partners for the synthesis of heme (Chang *et al.*, 1975; Salzman *et al.*, 1985). This was recently confirmed by the presence of the complete set of genes that code for enzymes of the heme pathway, which showed that those genes are unequally distributed between the host and the endosymbiont genomes, with most of them located in the bacterium (Alves *et al.*, 2011, 2013b).

In addition to hemin, the presence of the symbiont dispenses the with need for either citrulline or arginine to produce ornithine, which is due to the presence of the enzyme ornithine carbamoyl-transferase (OCT) in the bacterium completing the urea cycle (Figure 1.6) (Camargo et Freymuller, 1977; Figueiredo *et al.*, 1978a; Galinari et Camargo, 1978). In this cycle, the production of ornithine leads to the production of an intermediate used for the synthesis of pyrimidines, proline and the polyamine putrescine (Kidder *et al.*, 1966; Yoshida *et al.*, 1978). This intricate dialog includes not only the metabolic point of view, but also regulation and cell cycle, among others. One such example is the uptake of L-proline, which is not required for growth. In *A. deanei*, it seems to be upregulated by the presence of the symbiotic bacterium (Galvez Rojas *et al.*, 2008). The just mentioned polyamine, putrescine, can be produced from ornithine through the activity of the enzyme ornithine decarboxylase (ODC). The ODC activity is higher in *A. deanei* when compared to the aposymbiotic strain, increasing thus the polyamine metabolism (Frossard *et al.*, 2006). This augmentation could be related to the faster growth of SHTs than of symbiont-free strains, due to the involvement of polyamines in cell growth and proliferation (see review in Willert et Phillips (2012)).

Less is known about the contribution of the protozoan host to the metabolism of the bacterium. Besides the previously mentioned host supply of PC to the symbiont (de Azevedo-Martins *et al.*, 2007), the symbiont may obtain ATP through the activity of host glycosomes (Motta *et al.*, 1997a). The latter indicates important metabolic exchanges, especially as concerns energy metabolism, between the bacterium and the host glycosomes, which are known to be physically close, and probably also including the mitochondrion (Motta *et al.*, 1997a; Faria-e Silva *et al.*, 2000; Motta, 2010).

Symbiont identity and gene transfer

There are a few characteristics that make the symbiosis in trypanosomatids a singular model to study cellular evolution. One such feature is the presence of only one symbiotic bacterium that divides synchronically with the host cell, indicating that the protozoan imposes tight control over the endosymbiont division (Motta *et al.*, 2010). As previously mentioned, host control of cellular division is found in organelles and is important for the persistence of the interaction (Douglas, 2010). Finally, genetic assimilation, as described in symbiont-derived organelles, is facilitated in the case of single cell eukaryotes such as the trypanosomatids (for more details see Section 1.1.3) (Douglas, 2010); thus, investigation of the symbiont-host gene

Figure 6 page 145 from Motta (2010).

Figure 1.6: **Urea cycle in endosymbiont-containing (a) and endosymbiont-free (b) *Angomonas* species.** Notice that the enzyme ornithine carbamoyl-transferase (OCT) is only present in endosymbiont-bearing strains, closing the urea cycle in these protozoa. Conversely, the citrulline hydrolase is only found in endosymbiont-free species that need exogenous arginine or citrulline in culture medium, but not ornithine, which does not substitute for either aminoacids. **Extracted from** Motta (2010).

transfer and of the identity of the bacterium is of great interest.

1.3 Metabolic networks

1.3.1 Overview of metabolism

Metabolism is the whole network of chemical reactions occurring in a living organism. A reaction is the transformation of a set of compounds, called *substrates* into another set of compounds, called *products*. Classically, the analysis of the metabolism of an organism is performed by splitting the metabolic network into several *metabolic pathways* and by analysing each one independently. Conversely, the analysis of the whole metabolic network in a systemic way is enabled by high-throughput technologies, and allows investigating the functions of a complex system that cannot be understood by its components and which are termed emergent or systemic properties (Breitling *et al.*, 2008; Palsson, 2006).

Getting more deeply into the parts that compose the system, some basic concepts will be hereafter described relying on Cornish-Bowden (2004). Most of the chemical reactions taking place in living organisms do not happen spontaneously due to mild conditions inside a cell, they need catalysts which in this case are termed enzymes. The majority of them are made of protein, and are neither consumed nor produced but are necessary for the reaction to happen. A remarkable feature of an enzyme is the specificity to catalyse one or a set of reactions but not the remaining ones. This precision comes with a high cost and a complex structure, where the enzyme is generally about 50-100 times the combined volume of the molecules it acts on. In order to act on its substrates (generally two but there are enzymes that act on one, three or more), it needs a cavity to fit them (and not unwanted molecules) and regions that attract them such as charged groups and hydrophobic regions (Figure 1.7). Then, the catalytic groups can interact with the substrates leading to the proper transformation.

Depending on thermodynamic constraints, the transformation from substrate(s) (A) into product(s) (B) can be *reversible*, i.e. $A \leftrightarrow B$, or *irreversible*, i.e. $A \rightarrow B$. Moreover, there are coefficients to describe the balance of molecules from both sides of the chemical reaction, e.g. $2H_2 + O_2 \rightarrow 2H_2O$, which are called stoichiometric coefficients.

Figure 1.2 page 7 from Cornish-Bowden (2004).

Figure 1.7: **Required features for the activity of an enzyme. Extracted from** Cornish-Bowden (2004).

The unity of biochemistry is highlighted by Cornish-Bowden (2004) and it is interesting to mention it here. While the characteristics of one type of organism are most often considerably different from another, their biochemical components - carbohydrates, proteins and fats - use the same sort of chemistry to transform compounds into one another and to produce the basic building blocks from the substrates which they acquire from the environment. Even the sequence of reactions, i.e. pathways, such as for the conversion of glucose into energy, are

similar in quite different organisms. This is a key notion to keep in mind when considering the uneven distribution of biochemical data currently available where few model species are well-studied and constitute the main part of the metabolic databases, while inferring the metabolism of the vast majority of the organisms relies on a propagation of such knowledge. This fact will be further discussed below. This is however limiting in the sense that only a same core metabolism is known in a wide diversity of species whereas organism-specific information is missing (Breitling *et al.*, 2008). Furthermore, Breitling *et al.* (2008) raised the fact that condition-specific data are also missing and suggest promising experimental methodologies to focus on this unexplored metabolism. Thus, those gaps in biochemical knowledge, as well as the ones that remain in the core metabolism (i.e., reactions for which no enzyme catalysing them has been identified), impact the quality of the reconstructed networks (Palsson, 2006).

1.3.2 Metabolic network reconstruction

Diverse omics data, such as genomics, transcriptomics, proteomics and metabolomics, can be integrated into a biological network or used to analyse it. A metabolic network reconstruction relies mainly on genomic data for defining the set of reactions potentially occurring in a given organism. This issue has been extensively reviewed (Francke *et al.*, 2005; Pinney *et al.*, 2007; Lacroix *et al.*, 2008; Rocha *et al.*, 2008; Durot *et al.*, 2009; Feist *et al.*, 2009; Pitkänen *et al.*, 2010; Haggart *et al.*, 2011; Santos *et al.*, 2011; Chen *et al.*, 2012; Kim *et al.*, 2012; de Oliveira Dal'Molin et Nielsen, 2013) and a step-by-step procedure was recently described by Thiele et Palsson (2010). This process starts with a draft reconstruction based on the genome sequence of the target organism and is followed by the time-consuming refinement of this reconstruction depending on the intended use of the metabolic model. These topics are further detailed hereafter.

Metabolic databases

Depending on the chosen method for the reconstruction process, it may rely on the metabolic pathway information gathered in databases such as MetaCyc/BioCyc (Caspi *et al.*, 2012) and KEGG (Kanehisa *et al.*, 2012). These provide data on experimentally characterised metabolic pathways of primary and secondary metabolism for different organisms as well as associated experimental literature, compounds, enzymes and genes, thus gathering a catalog of the universe of metabolism, a global metabolic pathway map. In addition to that, information can be obtained from organism-specific databases such as EcoCyc (Keseler *et al.*, 2009), and from biochemical databases for enzymes or transporters, e.g. BRENDA (Schomburg *et al.*, 2013), ExplorEnz (McDonald *et al.*, 2009), Transport DB (Ren *et al.*, 2007), Transport Classification DB (Saier *et al.*, 2009).

From the genome to the chemical reactions

This process may be split in two parts: (i) functional annotation of metabolic genes to determine the potential enzymatic capabilities of the target organism, mainly based on sequence homology; and (ii) inference of the list of reactions enabled by the assigned enzymatic activities coded by its EC number (e.g. 1.1.1.1), which can be obtained from enzyme databases such as BRENDA (Lacroix *et al.*, 2008; Schomburg *et al.*, 2013). At this stage, the gene-protein-reaction (GPR) associations are initially established, and will be further refined to include cases where the relation between genes and reactions is not one-to-one, e.g. isozymes perform a same reaction or the enzyme complex requires more than one subunit coded by

different genes. The GPR associations will be modelled in the network as a boolean system, for which an example can be found in Figure 1.8.

Figure 5 page 99 from Thiele et Palsson (2010).

Figure 1.8: **Gene-protein-reaction (GPR) associations.** Examples of GPR associations in *Escherichia coli* and the boolean representation. **Extracted from** Thiele et Palsson (2010).

This draft network step can currently be accomplished by a variety of automatic reconstruction tools, listed and detailed in various reviews about this topic (Francke *et al.*, 2005; Pinney *et al.*, 2007; Lacroix *et al.*, 2008; Rocha *et al.*, 2008; Durot *et al.*, 2009; Feist *et al.*, 2009; Pitkänen *et al.*, 2010; Haggart *et al.*, 2011; Santos *et al.*, 2011; Chen *et al.*, 2012; Kim *et al.*, 2012; de Oliveira Dal'Molin et Nielsen, 2013). I will describe three of them focusing on the input and output of each one, observing that none of them spares the need for a subsequent time-consuming manual refinement of the network, although it may facilitate the task.

I start with the well established approach of PathoLogic from the Pathway tools software (Karp *et al.*, 2010), which is associated to the BioCyc database (Karp *et al.*, 2005) and uses an annotated genome sequence as input to build a model-organism database called a Pathway/Genome Database (PGDB) such as EcoCyc (Keseler *et al.*, 2009). It enables to infer operons, transport reactions and metabolic pathways, gives access to its own gap-filling approach called pathway hole-filler, performs consistency checks, and gives access to a cellular overview with an omics viewer, to an iterative editing platform to curate the data, as well as to various other analysis and visualisation tools (Green et Karp, 2004; Romero et Karp, 2004; Paley et Karp, 2006; Lee *et al.*, 2008; Dale *et al.*, 2010; Latendresse et Karp, 2011). From an annotated genome sequence, it uses the assigned EC number as well as the enzyme name which is compared to a dictionary where the enzymatic activity names are linked to potential reactions. This method was first designed to build a database of pathways and to add regulation and other information into the platform rather than to modelling for a constraint-based analysis (Santos *et al.*, 2011). On the other hand, it is a software that is constantly evolving since 1996 (Karp et Paley, 1996; Karp *et al.*, 2002), one example being the recent inclusion of flux balance analysis (FBA) and multiple gap-filling to Pathway tools (Latendresse *et al.*, 2012). The aim of this multiple gap-filling method is to accelerate the development of FBA models directly from Pathway/Genome Databases, using the new tool

called MetaFlux based on mixed integer linear programming. Its singularity is the capability to simultaneously suggest modifications to the four essential sets describing an FBA model: the set of reactions, of metabolites of the biomass reaction, of nutrients and of secretions.

The second automatic reconstruction approach that will be introduced here focuses on constraint-based modelling. Notebaart *et al.* (2006) presented the detailed Autograph method and applied it to reconstruct a genome-scale metabolic model of *Lactococcus lactis*; however no tool or algorithm was made available. The aim is to reuse as much as possible the available and well-curated genome-scale metabolic models. There are currently almost 100 of those models, mostly from bacteria, and the complete list can be found at `http://gcrg.ucsd.edu/InSilicoOrganisms/OtherOrganisms`. When using this approach, one takes advantage of the verification steps and the collected knowledge of a curated model that may not be found in databases, due to wrong or incomplete information or to new data which are not yet included in the databases. Examples of such include: reaction stoichiometries, gap-filling and extensive bioinformatic, experimental and bibliomic analyses which were performed to curate the model (Santos *et al.*, 2011). The pipeline is the following: (i) orthology search between the genes of the query and the reference genome (the latter is one that has a well-curated genome-scale metabolic model available); (ii) establishment of the link between orthologs and reactions; (iii) transfer of the reactions to the genes of the query genome (Notebaart *et al.*, 2006).

Finally, a more recent method called the Model Seed (Henry *et al.*, 2010) is a web-based pipeline for metabolic reconstruction that is based on a constraint-based modelling and allows to go further in the list of recommended refinements of a reconstruction (Thiele et Palsson, 2010) when compared to other methodologies. It requires only an assembled genome sequence as input and generates an SBML (Systems Biology Markup Language Hucka *et al.* (2003)) model. A singularity of this approach is the generation of a metabolic model containing already an organism-specific biomass reaction. The quite reduced list of manual curation steps that remains after applying this method includes experimental data collection, assignment of gene and reaction localisation, inference of intracellular transport reactions (mostly for eukaryotic models since only cytosol and extracellular compartments are included in the model), determination of biomass reaction coefficients and loading of the models into the Cobra toolbox (Henry *et al.*, 2010; Schellenberger *et al.*, 2011). This pipeline goes further in some decision-making steps during model optimisation, such as filling or not one gap, that are error-prone and should be verified.

Thus, the three approaches presented above for the draft level reconstruction of a metabolic network have important differences that should be taken into account when selecting one to be used. Pathway tools for instance allows for a step-by-step draft reconstruction process where the user interacts with the method solving ambiguities. On the other hand, Autograph and Model Seed might output a draft reconstruction that goes further in the manual refinements (presented in the next topic) while it passes additional decision-making steps that are subject to error and require verification during the manual refinement. None of the methods to date completely eliminates the need for a manual curation, if the aim is to have a high-quality model for simulations and predictions (Thiele et Palsson, 2010; Santos *et al.*, 2011). Moreover, the decision of which approach to adopt may depend on:

1. The size of the reconstruction model; for instance if one is working with a small bacterial metabolic network, it may be more efficient to perform a reaction-by-reaction reconstruction without a first draft reconstruction since the latter would require a manual inspection of the included reactions, as suggested by Thiele et Palsson (2010).

2. The level of annotation of the target genome, *i.e.* if it was carefully manually annotated, the chosen method should probably make use of it.

3. The availability of well-curated genome-scale metabolic models and organism-specific databases one could rely on, *i.e.* models or databases of organisms with similar taxonomic position and/or lifestyle of your target organism.

4. A choice to iteratively interact and control for each step of the draft reconstruction process or a verification/inclusion/exclusion step after an enhanced draft reconstruction.

Iterative refinement process

The draft level reconstruction of a metabolic network, which was presented in the previous topic, is encouraged to be followed by a manual refinement which can be performed as an iterative process as follows: refine the reconstruction, model it (see the next Section *Modelling of metabolic networks*) and evaluate it based on its topological and functional properties in a cyclic way, i.e. getting back to manual refinements, re-evaluating it and so on (Thiele et Palsson, 2010). During this process, one can check for correctness of the inclusion of each reaction and of its links with the rest of the network, as well as for gaps and inconsistencies (Thiele et Palsson, 2010; Santos *et al.*, 2011). In addition to that, the growth medium requirements and the biomass composition should be defined. In the reconstruction protocol, Thiele et Palsson (2010) recommended a curation process done in a pathway-by-pathway manner using the canonical pathways and leaving the peripheral pathways and reactions with no assigned pathway for later, e.g. starting with the central metabolism, followed by the biosynthesis of individual macromolecular building blocks such as amino acids, nucleotides, and lipids. Organism-specific data are important in different steps of the curation process and one such example is physiological data on growth conditions. The already mentioned reconstruction protocol includes 96 steps that can be found in Figure 1.9. This iterative process should be repeated until the phenotypic characteristics of the target organism are similar to model predictions and/or all experimental data for comparison is exhausted (Thiele et Palsson, 2010).

Figure 1 page 95 from Thiele et Palsson (2010).

Figure 1.9: **Overview of the procedure to iteratively reconstruct metabolic networks. Extracted from** Thiele et Palsson (2010).

1.3.3 Modelling of metabolic networks

The analysis of metabolic networks can be structural or dynamic and the modelling commonly ranges from graphs, constraint-based models to differential equations (Lacroix *et al.*, 2008). Further comparison of those analyses and models can be found in Santos *et al.* (2011); Haggart *et al.* (2011); Klein *et al.* (2012a). The latter presents our review on these topics and is included in the Appendix A. The first two models will be briefly described below.

Graph models

A metabolic network may be interpreted and built in various ways (Figure 1.10): nodes can be metabolites or reactions (respectively giving rise to the compound and the reaction graphs), and arcs (i.e. directed edges) can be reactions or shared metabolites. In both cases, the reconstruction may lead to a loss of fundamental information, e.g. in Figure 1.10 reaction R1 has two substrates (A and B) and two products (C and D), however, by looking at the corresponding compound graph one could imagine that the production of C only requires A, and by looking at the corresponding reaction graph we notice that the arc between R1 and R2 exists only because of the compound D regardless of the presence of E. These limitations ask for a full treatment of the complex reactions in a metabolic network (discussed in detail e.g. in Lacroix *et al.* (2008); Cottret et Jourdan (2010)): bipartite graphs and hypergraphs help to overcome these problems at the price of a higher algorithmic complexity. Hypergraphs are indeed generalisations of graphs and thus problems may become harder to solve (see Klamt *et al.* (2009) for some examples of hypergraphs applied to biological questions and the associated computational problems).

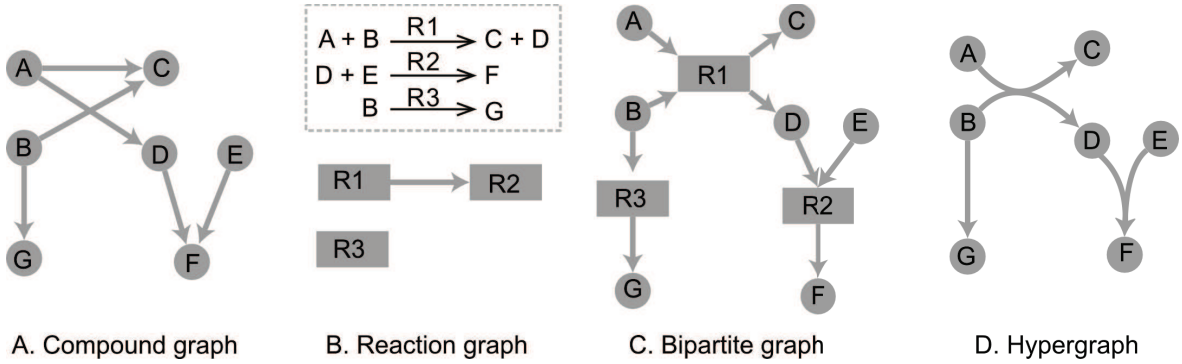


Figure 1.10: **Graph models to represent a metabolic network.** Given three biochemical reactions (R1, R2, R3), metabolic graphs are built with metabolites as round nodes and reactions as square nodes. The same system can be represented using different kinds of networks. **A.** Compound graph, where nodes are metabolites and there is an arc between a substrate and a product of a reaction; **B.** reaction graph, where nodes correspond to reactions and are connected when a product of one reaction is a substrate of the next one; **C.** bipartite graph: nodes are either compounds or reactions in which there is an arc between the substrate/reaction and reaction/product; **D.** hypergraph: nodes are compounds and a hyperarc links the substrate(s) to the product(s) of a reaction. **Extracted from** Klein *et al.* (2012a).

Structural analyses of metabolic networks modelled as a graph include the classical measures from graph theory such as centrality, degree distribution, diameter and average internode distance (Lacroix *et al.*, 2008; Klein *et al.*, 2012a). Besides, it allows also for the evaluation of the metabolic reconstruction. Such measures can be applied to other types of biological networks and may provide further insight into the structure and general characteristics of the

network. Biases in the network reconstructions or manipulation can strongly affect the results of the analysis, confounding the observed correlations (if any exist) between biological and topological properties (Coulomb *et al.*, 2005). Thus, depending on the structural analyses performed, some filtering of cofactors and ubiquitous compounds may be necessary to avoid meaningless biochemical paths and conclusions (Ma et Zeng, 2003). Such filters can be applied to the metabolic graph using MetExplore (Cottret *et al.*, 2010), which is a web-server that allows to build, curate and analyse genome-scale metabolic networks. Consequently, one needs to carefully interpret the topological measures obtained (for further discussion of this topic see Klein *et al.* (2012a)).

Constraint-based models

In this framework, the network is modelled as a stoichiometric matrix (Figure 1.11), where metabolites compose the lines whereas reactions represent the columns and the stoichiometric coefficients fill the cells of the matrix. The signs of those coefficients indicate whether the compound is consumed or produced (Covert et Palsson, 2003; Palsson, 2000, 2006). This approach was introduced by Covert et Palsson (2003) and made available in the Cobra toolbox (Schellenberger *et al.*, 2011) which allows for the reconstruction and analysis of constraint-based models.

Figure in Box 1 page 157 from Breitling *et al.* (2008).

Figure 1.11: **Example of a stoichiometric matrix (S) representing a pathway-map.** S is the mathematical representation of the pathway shown at the top of the figure. Reaction (r) 1 consumes one molecule each of red and green combined to form one molecule of red–green product. A metabolic network can be reconstructed from S, but the opposite is not necessarily the case. For example, in r4, two molecules of the purple compound are produced for every blue molecule that is consumed. Such information is not always included in the biochemical-pathway map. **Extracted from** Breitling *et al.* (2008).

The motivation here is to investigate the metabolic capabilities of an organism under specified growth conditions based on the distribution of mass fluxes through the reactions, imposing some constraints in order to reduce the space of feasible solutions. Such constraints include the mass-balance, also called steady-state, constraint where the concentration of an

internal compound is constant in time since it is balanced by its production and consumption rates, as well as the capacity, also termed thermodynamic, constraint which concerns any limitation imposed on the individual rates of the reactions. This framework allows for a quantitative structural analysis called Flux Balance Analysis (FBA). Since the solution space for such models is very large even under the constraints used, FBA seeks an optimal flux distribution with respect to a carefully chosen objective function using optimisation techniques. The assumption behind FBA is that metabolism maximises some objective, but there may exist many suboptimal flux distributions that help the organism during adaptation to specific environmental conditions.

One of the most appealing properties of constraint-based models is that they provide a way to explore the consequences of genetic manipulations on the whole metabolic network: one or more reactions can be eliminated (simulating knock-out mutants) (Pharkya *et al.*, 2003, 2004; Wunderlich et Mirny, 2006; Suthers *et al.*, 2009) or otherwise manipulated, and simulations can be run to see if and how the objective function can be improved with respect to the wild-type model (Trinh *et al.*, 2008). By coupling two levels of optimisation, it is possible to predict the best engineering strategy to have mutants that maximise some by-product of interest, such as ethanol (Trinh *et al.*, 2008) or lactate (Fong *et al.*, 2005), while growing. A recent survey on FBA and its applications can be found in Raman et Chandra (2009).

Chapter 2

Metabolic dialogue between a trypanosomatid and its symbiont

Contents

2.1 Predicting the proteins of *Angomonas deanei* and *Strigomonas culicis* and of their respective endosymbionts reveals new aspects of the Trypanosomatidae family 30

2.2 Biosynthetic pathways of amino acids and vitamins 31

2.2.1 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 32

2.2.2 Biosynthesis of vitamins and cofactors in bacterium-harbouring trypanosomatids depends on the symbiotic association as revealed by genomic analyses 54

2.3 Metabolic networks of host and symbiont 81

2.3.1 Overview 81

2.3.2 Reconstruction of the metabolic networks 81

2.3.3 Metabolic network of *A. deanei* 82

2.3.4 Metabolic reconstruction of the endosymbiont 85

2.3.5 Potential metabolic exchanges between the host and its symbiont . . 87

2.3.6 Perspectives 89

This chapter presents a sequence of studies aiming to characterise the intricate metabolic exchanges between trypanosomatids and their symbiotic bacterium based on genomic data. It starts with the genome sequencing of two symbiont-harbouring trypanosomatids: *Angomonas deanei* and *Strigomonas culicis* and their respective symbionts (Section 2.1). It is followed by the analyses of the biosynthetic pathways of essential amino acids and vitamins for which the bacterial symbionts are known to play an important role based on nutritional data (Section 2.2). The chapter ends with the ongoing genome-scale metabolic model of these two pairs of host and symbiont (Section 2.3).

2.1 Predicting the proteins of *Angomonas deanei* and *Strigomonas culicis* and of their respective endosymbionts reveals new aspects of the Trypanosomatidae family

My main contribution to the following manuscript – Motta *et al.* (2013) Predicting the proteins of *Angomonas deanei*, *Strigomonas culicis* and their respective endosymbionts reveals new aspects of the trypanosomatidae family. *PLoS One.* 8(4):e60209 – concerns the analyses of the biosynthetic pathways of amino acids and vitamins. The published version of this manuscript can be found in the Appendix B.

In this study, we used DNA pyrosequencing and a reference-guided assembly to generate reads that predicted 16,960 and 12,162 open reading frames (ORFs) in two symbiont-bearing trypanosomatids, *Angomonas deanei* and *Strigomonas culicis*, respectively, in an effort to better understand such symbiotic association. Identification of each ORF was based primarily on TriTrypDB using tblastn, and each ORF was confirmed by employing getorf from EMBOSS and Newbler 2.6 when necessary. The monoxenic organisms revealed conserved housekeeping functions when compared to other trypanosomatids, especially *Leishmania major*. However, major differences were found in the ORFs corresponding to the cytoskeleton, the kinetoplast, and the paraflagellar structure. The monoxenic organisms also contain a large number of genes for cytosolic calpain-like and surface gp63 metalloproteases and a reduced number of compartmentalised cysteine proteases in comparison to other TriTryp organisms, reflecting adaptations to the presence of the symbiont. The assembled bacterial endosymbiont sequences exhibit a high A+T content with a total of 787 and 769 ORFs for the endosymbionts of *Angomonas deanei* and *Strigomonas culicis*, respectively, and indicate that these organisms have a common ancestor related to the Alcaligenaceae family. Importantly, both symbionts contain enzymes that complement essential host cell biosynthetic pathways, such as those for amino acid, lipid and purine/pyrimidine metabolism.

Detailed analyses of the synthesis of amino acids and vitamins in symbiont-bearing trypanosomatids are presented in the next section.

2.2 Biosynthetic pathways of amino acids and vitamins

In this section, two manuscripts are presented with some slight modifications to avoid redundancy. In both cases, I share the first authorship with J.M.P. Alves:

(i) J.M.P. Alves, C.C. Klein, F.M. da Silva, A.G. Costa-Martins, M.G. Serrano, G.A. Buck, A.T.R. Vasconcelos, M.-F. Sagot, M.M.G. Teixeira, M.C.M. Motta and E.P. Camargo. 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*, 13(1):190+, 2013;

(ii) C.C. Klein, J.M.P. Alves, M.G. Serrano, G.A. Buck, A.T.R. Vasconcelos, M.-F. Sagot, M.M.G. Teixeira, E.P. Camargo, M.C.M. Motta. Biosynthesis of vitamins and cofactors in bacterium-harbouring trypanosomatids depends on the symbiotic association as revealed by genomic analyses, *PLoS One*, 8 (11), 2013.

These studies were performed in collaboration with M.C.M. Motta from the Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro, Brazil; and with members of the Dept. of Parasitology, Institute of Biomedical Sciences, University of São Paulo, Brazil and of the Virginia Commonwealth University, Richmond, VA, USA.

Overview of both studies In addition to the heme biosynthetic pathway, the synthesis of essential amino acids and vitamins represents the known contribution of the bacterial symbionts of trypanosomatids to their respective hosts, based on nutritional data. These metabolic pathway analyses are the starting point of a wider and still ongoing investigation of this intricate relationship done in the context of the whole metabolic networks (Section 2.3).

In both studies (of amino acids and of vitamins), we investigate the entire genomes of five symbiont-harbouring trypanosomatids of the genera *Angomonas* and *Strigomonas* (SHTs) and their respective bacteria (TPEs), as well as two regular trypanosomatids without symbionts (RTs), for the presence of genes of the classical pathways for amino acid and vitamin biosynthesis. Most of the genes responsible for those routes were found in the genome of the symbionts, comprising the synthesis of lysine, branched-chain and aromatic amino acids, as well as four vitamins of the B complex: riboflavin, pantothenate, vitamin B_6 and folate. The fewer genes found in the host genomes were inspected for the possibility of horizontal gene transfer (HGT) from bacteria using phylogenetic analyses. This investigation is motivated by the fact that these genes could have been transferred from the symbiont to the trypanosomatid nuclei in the course of the bacterium genome reduction and co-evolution of these partners as happened in the case of the mitochondrion and chloroplast (a more extensive discussion on this topic can be found in Section 1.2.3). While these candidate HGTs were few in the case of vitamin synthesis, they were quite numerous as concerns amino acids. The vast majority of those HGTs were potentially transferred from diverse bacterial taxa not comprising betaproteobacteria. These findings suggest that HGT events played a fundamental role in the genomic evolution of the Trypanosomatidae analysed, which was also previously found in the heme biosynthetic pathway (Alves *et al.*, 2011). However, as previously seen in the case of *Buchnera* and the pea aphid, no massive transfer of the symbiont genes to the host seems to have happened, at least not concerning these pathways. Further phylogenetic studies of the whole host genomes should show the complete extent of this process.

2.2.1 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

BACKGROUND

Previous comparative studies on SHTs (wild and cured strains, obtained after antibiotic treatment) and RTs, often involving trace experiments using radioactive compounds, reported the requirement, substitution, and sparing of amino acids in culture media (Mundim *et al.*, 1974; Alfieri et Camargo, 1982; Cowperthwaite *et al.*, 1953; Kidder et Dutta, 1958; Guttman, 1966, 1967; Gutteridge *et al.*, 1969; Krassner et Flory, 1971; Kidder et Dewey, 1972; Cross *et al.*, 1975b; Anderson et Krassner, 1975; Cross *et al.*, 1975a; Mundim et Roitman, 1977; Roitman *et al.*, 1977; Yoshida et Camargo, 1978; Hutner *et al.*, 1979). 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 RTs. However, similar analyses showed that SHTs require only methionine or tyrosine in culture media, suggesting that they possess the necessary enzymatic equipment to synthesize most amino acids (Newton, 1957; Mundim *et al.*, 1974; de Menezes et Roitmanz, 1991; Chang et Trager, 1974). Unfortunately, besides the SHTs, 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 (Kidder *et al.*, 1966; Alfieri et Camargo, 1982; Fair et Krassner, 1971; Camargo et Freymuller, 1977; Figueiredo *et al.*, 1978b; Yoshida *et al.*, 1978; Camargo *et al.*, 1978; Galinari et Camargo, 1978, 1979; Gutteridge *et al.*, 1969; Camargo *et al.*, 1987). Thus, in contrast to the advanced state of knowledge of genes involved in amino acid biosynthesis in many microorganisms (Bono *et al.*, 1998), the potential for amino acid synthesis in trypanosomatids remains largely unknown. In SHTs, 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 SHTs and RTs of different genera (see Methods), through the characterization of each gene as identified by similarity searches and protein domain analyses. We applied extensive phylogenetic inferences to determine the most likely origin 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 (Alves *et al.*, 2011). 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 contribution of TPEs to the amino acid metabolism of their trypanosomatid hosts, which is related to high amounts of lateral transfer of genes from diverse bacterial groups to the 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 TPEs, SHTs, and RTs and then compared to the available nutritional and enzymatic data of 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 RT, SHT, and TPE genomes that have been sequenced here. Although the genomes of all available SHTs and TPEs have been examined, only a very limited sample of RT genomes (*H. muscarum* and *C. acanthocephali*) were 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 (Additional file C.1 in Appendix C) show 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 RTs 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 C.1 in Appendix C) and sequencing coverage (Additional file C.2 in Appendix C) 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 (Bhattacharjee, 1985; Nishida, 2001; Velasco *et al.*, 2002; Hudson *et al.*, 2005; Torruella *et al.*, 2009).

Early nutritional studies (Gutteridge *et al.*, 1969) showed that lysine is essential for the growth of RTs, but could be efficiently replaced by DAP. In agreement with this, 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* (Gutteridge *et al.*, 1969). 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, SHTs required neither lysine nor DAP to grow in defined medium (Newton, 1956, 1957; Kidder *et al.*, 1966; Mundim *et al.*, 1974; de Menezes et Roitmanz, 1991). 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 TPEs (Figure 2.1). In contrast, only the final gene of the DAP pathway was found in the genomes of the SHTs, and the final two found in one RT examined (*H. muscarum*), which explains why DAP could substitute for lysine in growth media of some RTs. 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 TPE, SHT, or RT genomes analyzed.

In summary, our findings using comparative genomics are in agreement with the data from previous nutritional and enzymatic studies, showing that only SHTs, and not RTs, are autotrophic for lysine and that this autonomy is provided by the DAP pathway present in their symbionts. The presence of DAP-decarboxylase in SHTs 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.

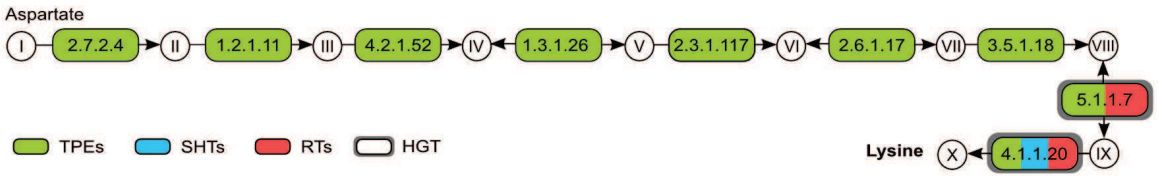


Figure 2.1: **DAP pathway for lysine biosynthesis.** Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – I: L-aspartate; II: 4-aspartyl-phosphate; III: aspartate 4-semialdehyde; IV: 2,3-dihydrodipicolinate; V: 2,3,4,5-tetrahydrodipicolinate; 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.

Methionine and cysteine Methionine is included in all defined media designed for the growth of trypanosomatids with or without symbionts (Newton, 1957; Mundim *et al.*, 1974; Kidder et Dutta, 1958), 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 (Kidder et Dutta, 1958; Guttman, 1967; Hutner et Provasoli, 1965).

Our analyses suggest that RTs and SHTs have the necessary genes to produce cystathionine, homocysteine, and methionine from homoserine (Figure 2.2), whereas the TPE 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 TPEs, SHTs, and RTs examined.

With respect to cysteine synthesis, it has been shown that the incubation of cell homogenates of *C. fasciculata* with 35S-methionine produced radioactive adenosyl-methionine (SAM), adenosyl-homocysteine (SAH), homocysteine, cystathionine, and cysteine (Guttman, 1967). Thus, this trypanosomatid is fully equipped to methylate methionine to produce ho-

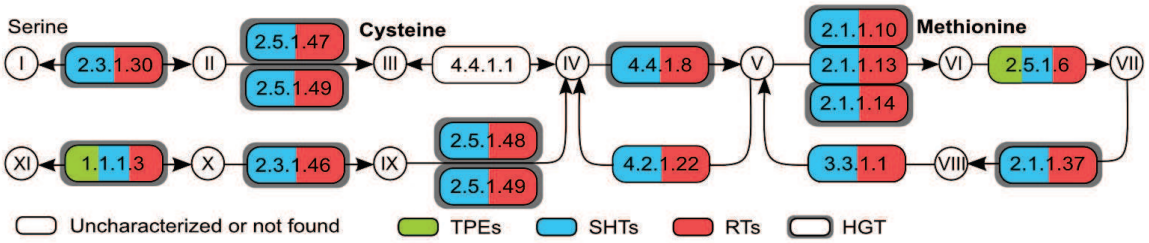


Figure 2.2: **Cysteine and methionine synthesis and interconversion pathway.** Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – I: L-serine; II: O-acetyl-serine; III: cysteine; IV: cystathionine; V: homocysteine; VI: methionine; 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.

mocysteine 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 RTs. 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 2.2, compounds III-X), all SHTs and RTs examined had the genes to produce cysteine from serine in a simple two-step reaction, with acetylserine as an intermediate (Fig 2, I-III).

In summary, if RTs and SHTs are capable of interconverting methionine and cysteine, as shown for *C. fasciculata* (Kidder et Dutta, 1958), 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.

Threonine In trypanosomatids, initial investigations about the nutritional requirements for threonine were controversial. Most results suggested that this amino acid is essential (Kidder et Dutta, 1958; Guttman, 1967; Kidder et Dewey, 1972; Hutner et Provasoli, 1965; Nathan et Cowperthwaite, 1954; Janakidevi *et al.*, 1966), but other studies considered the addition of threonine to the growth media of RTs unnecessary (Alfieri et Camargo, 1982). 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 2.3. 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 TPEs but present in the genomes of SHTs 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).

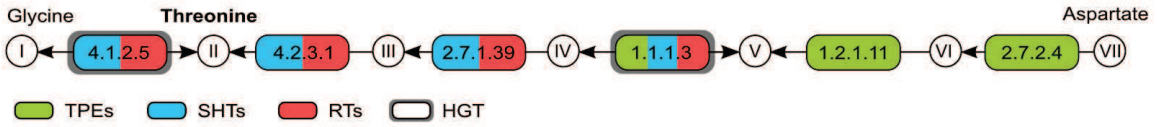


Figure 2.3: **Threonine synthesis pathway.** Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – I: glycine; II: threonine; III: phosphohomoserine; IV: homoserine; V: aspartate 4-semialdehyde; VI: 4-aspartyl-phosphate; VII: L-aspartate. 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.

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 TPEs, SHTs, and RTs. 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 SHTs), of *Trypanosoma* spp., and *Leishmania* spp., but totally absent from the TPE genomes.

Thus, the genetic constitution of RTs is consistent with earlier nutritional data showing the insect trypanosomatids, with or without symbionts, to be autotrophic for threonine. This observation suggests that TPEs 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 (Hutner *et al.*, 1980). 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 SHTs. The canonic pathway for the synthesis of isoleucine is depicted in Figure 2.4. Oxobutanoate (alpha-ketoglutaric acid) is the starting point of the pathway, and can be produced in two ways: from threonine (Figure 2.4, compounds II-III) or from pyruvate (Figure 2.4, 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 SHT homogenates than in any other cell fraction or in the cytosol, suggesting that this enzyme was located in

the symbiont (Alfieri et Camargo, 1982). However, genes for EC:4.3.1.19 are present in the genomes of TPEs, as well as those of SHTs and RTs (except *Leishmania* and *Trypanosoma*), contrasting with enzymatic determinations showing the absence of enzyme activity in RTs (Alfieri et Camargo, 1982). 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 TPEs. 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, valine, and leucine, is present in the genomes of SHTs and RTs, but not TPEs.

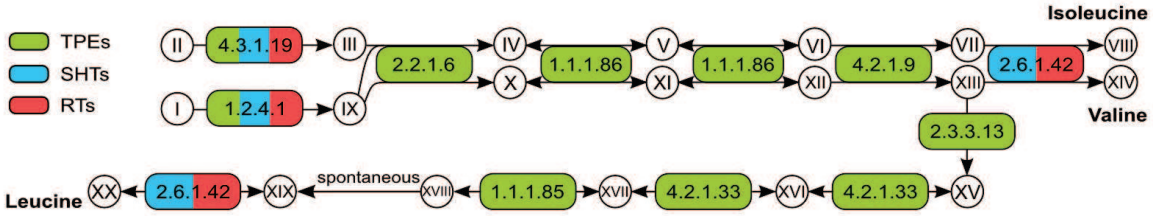


Figure 2.4: **Isoleucine, valine, and leucine synthesis pathway.** Metabolites – I: pyruvate; II: threonine; III: 2-oxobutanoate; IV: (S)-2-aceto-2-hydroxybutanoate; V: (R)-3-hydroxy-3-methyl-2-oxopentanoate; VI: (R)-2,3-dihydroxy-3-methylpentanoate; VII: (S)-3-methyl-2-oxopentanoate; VIII: isoleucine; 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.

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 SHTs, TPEs, and RTs. 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 TPEs. The reactions that follow from acetoacetate into valine involve the same TPE 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 TPEs – 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 RTs 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 RT synthetic culture media (Kidder et Dutta, 1958). Interestingly, this gene is present in all the SHT and RT genomes examined,

but absent from the TPE genomes (Figure 2.4). 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 RTs 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 (Shigenobu *et al.*, 2000; Macdonald *et al.*, 2012). 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 (Wilson *et al.*, 2010). 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 (Hansen et Moran, 2011). 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 (Hansen et Moran, 2011; Sandström et Moran, 1999). In summary, the presence in TPEs of most genes involved in isoleucine, valine and leucine synthesis explains why SHTs, but not RTs, 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 RTs, but not for SHTs (Newton, 1957; Mundim *et al.*, 1974; Kidder et Dutta, 1958; Guttman, 1966). 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 TPEs contain the genes for this route, while the genomes of SHTs and RTs do not (Figure 2.5).

The genes for the enzymes converting chorismate into prephenate and for transforming this compound into phenylalanine and tyrosine are present in all TPE genomes. The SHT and RT 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 SHTs and RTs, including the leishmaniae, but not in TPEs. Similarly, this enzyme is present only in the aphid. Furthermore, the gene encoding this enzyme is up-regulated in bacteriocytes, thus enhancing the production and interconversion of such amino acids (Hansen et Moran, 2011). On the other hand, TPEs have an additional route for the synthesis of phenylalanine from prephenate, involving the enzymes aromatic-amino-acid aminotransferase (EC:2.6.1.57) and prephenate dehydratase (EC:4.2.1.51), whose genes are absent in the SHT and RT 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 (Dunn *et al.*, 2008). Both subunits are present in the TPEs, whereas the genomes of SHTs and *H. muscarum* have only the beta subunit. None of the other trypanosomatid genomes examined presented either subunit of

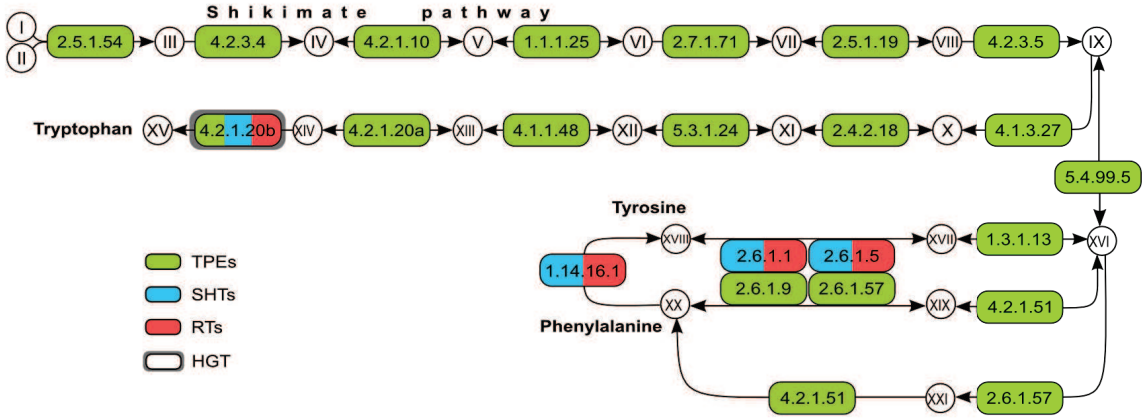


Figure 2.5: **Phenylalanine, tyrosine, and tryptophan synthesis pathway.** Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – I: D-erythrose 4-phosphate ; II: phosphoenolpyruvate; III: 7-phosphate-2-dehydro-3-deoxy-D-arabinoheptonate; 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.

tryptophan synthase.

In summary, the TPEs have all the genes for the different routes leading from chorismate to tryptophan, tyrosine, and phenylalanine, which are absent from the SHT and RT genomes. This obviously prevents RTs 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 TPEs indicates that the host might control the production of tyrosine.

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 RT growth (Mundim *et al.*, 1974; Kidder et Dutta, 1958; Guttman, 1966). Accordingly, the SHT and RT genomes do not seem to carry a single gene for histidine synthesis (Figure 2.6). All genes for the enzymes that participate in its biosynthesis, except the gene for histidinol-phosphate phos-

phatase (HPP, EC:3.1.3.15), which converts histidinol phosphate into histidinol, are present in the TPE genomes. Since SHTs 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 section 2.2.8).

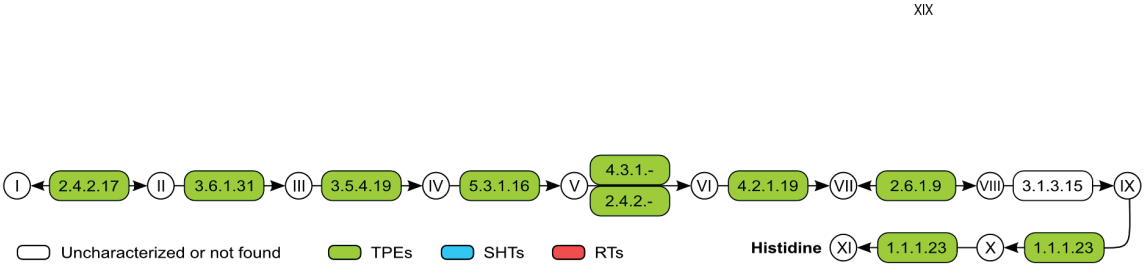


Figure 2.6: **Histidine synthesis pathway.** Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – I: 5-phosphoribosyl diphosphate (PRPP); II: phosphoribosyl-ATP; III: phosphoribosyl-AMP; IV: phosphoribosyl-formimino-AICAR phosphate; 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.

Arginine and ornithine Organisms autotrophic for ornithine use the glutamate pathway (Meister, 1965) for its synthesis via acetylated compounds as represented in Figure 2.7 (I-VI). All genes for this pathway are present in the genomes of TPEs. 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 SHTs and RTs, but not TPEs.

As represented in Figure 2.7, organisms lacking the glutamate pathway for the synthesis of ornithine can nevertheless produce it by different routes utilizing either citrulline or arginine (Figueiredo *et al.*, 1978b; Camargo *et al.*, 1978; Yoshida et Camargo, 1978). 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 SHTs and some RTs (*Leishmania* and *C. acanthocephali*), but not in the genomes of TPEs or *H. muscarum* – although a fragment was found in the latter.

Arginine can be synthesized from ornithine through a recognized universal enzymatic pathway (Meister, 1965), 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 TPEs and also in *Herpetomonas*, but was absent from the SHT and the other RT genomes examined. These findings confirm earlier immunocytochemical ultrastructural experiments showing the presence of OCT in the

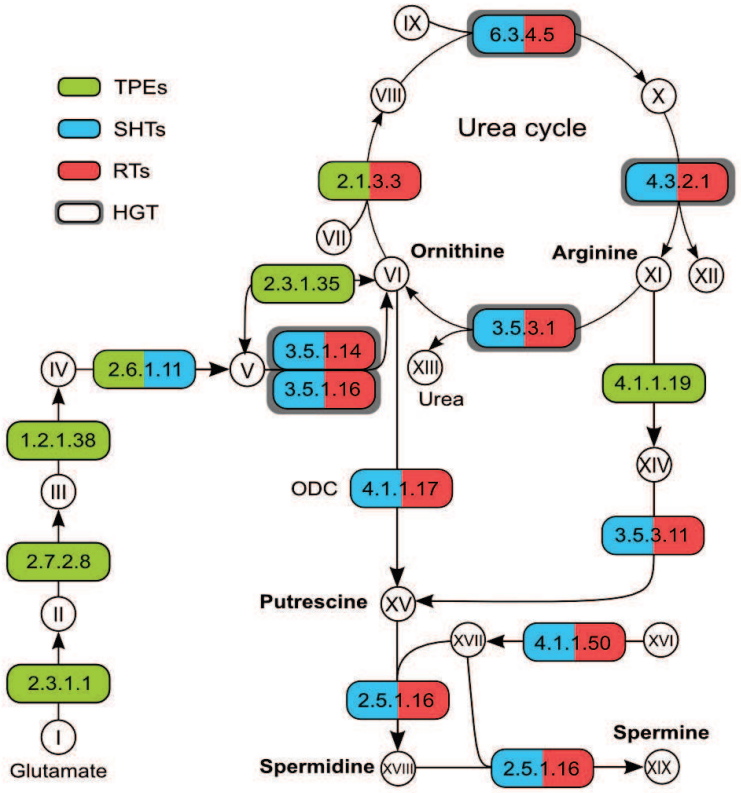


Figure 2.7: **Arginine, ornithine, and polyamine synthesis pathway.** Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – I: glutamate; II: N-acetylglutamate; III: N-acetylglutamyl-phosphate; IV: N-acetyl-glutamate semialdehyde; 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.

symbiont of *Angomonas deanei* (Camargo et Freymuller, 1977). The absence of the OCT gene renders most trypanosomatids unable to make citrulline from ornithine (Beutin et Eisen, 1983). However, the genes for the remaining enzymes leading from citrulline into arginine are all present in the genomes of all RTs and SHTs, but absent from the TPE genomes. These data are in full agreement 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 (Yoshida *et al.*, 1978; Camargo *et al.*, 1978, 1987).

Taking all these data together, we can conclude that RTs require exogenous sources of arginine or citrulline in their culture medium to produce ornithine. This is related to the fact that RTs lack the glutamate pathway for ornithine synthesis. Furthermore, ornithine

cannot substitute for arginine or citrulline because most RTs lack OCT. Conversely, SHTs 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 2.7, 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 genus *Angomonas* and in RTs, but not in TPEs or *Strigomonas*. Interestingly, it was proposed that the symbiont can enhance the ODC activity of *A. deanei* by producing protein factors that, in turn, increase the production of polyamines in the host trypanosomatid (Frossard *et al.*, 2006). Such high ODC activity may be directly connected to the lowest generation time described for trypanosomatids that is equivalent to 6 hours (Motta *et al.*, 2010). Putrescine could also be produced from agmatine since the genomes of RTs and SHTs 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 TPEs, thus completing the biosynthetic route for this polyamine, via agmatinase, in SHTs. Putrescine is then converted to spermidine and spermine by the 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 RTs and SHTs, but not in TPEs (Figure 2.7). The 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 SHTs and RTs examined (Additional file C.3 in Appendix C) , although there are questions regarding the step catalyzed by acireductone synthase (EC:3.1.3.77).

Phylogenetic analyses Our data on the phylogeny of the genes for essential amino acids biosynthesis have clearly shown that the genes present in the symbionts are of betaproteobacterial origin (for an illustrative example, see Figure 2.8), as shown before for the genes of heme synthesis (Alves *et al.*, 2011) and many others across the TPE genomes (Alves *et al.*, 2013b). The SHT and RT 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 results of the phylogenetic analyses). 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, which 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.

As concerns 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, in decreasing order, 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.

The 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 2.1, 2.2, 2.3, 2.5 and 2.7) and also the synthesis of a few non-essential amino acids such as glycine, serine, and proline. A 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 2.1, 2.2, and 2.3). The gene for EC:1.1.1.3 present in trypanosomatid genomes (both SHT and RT) 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 2.8). On the other hand, the TPE 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 TPE genomes.

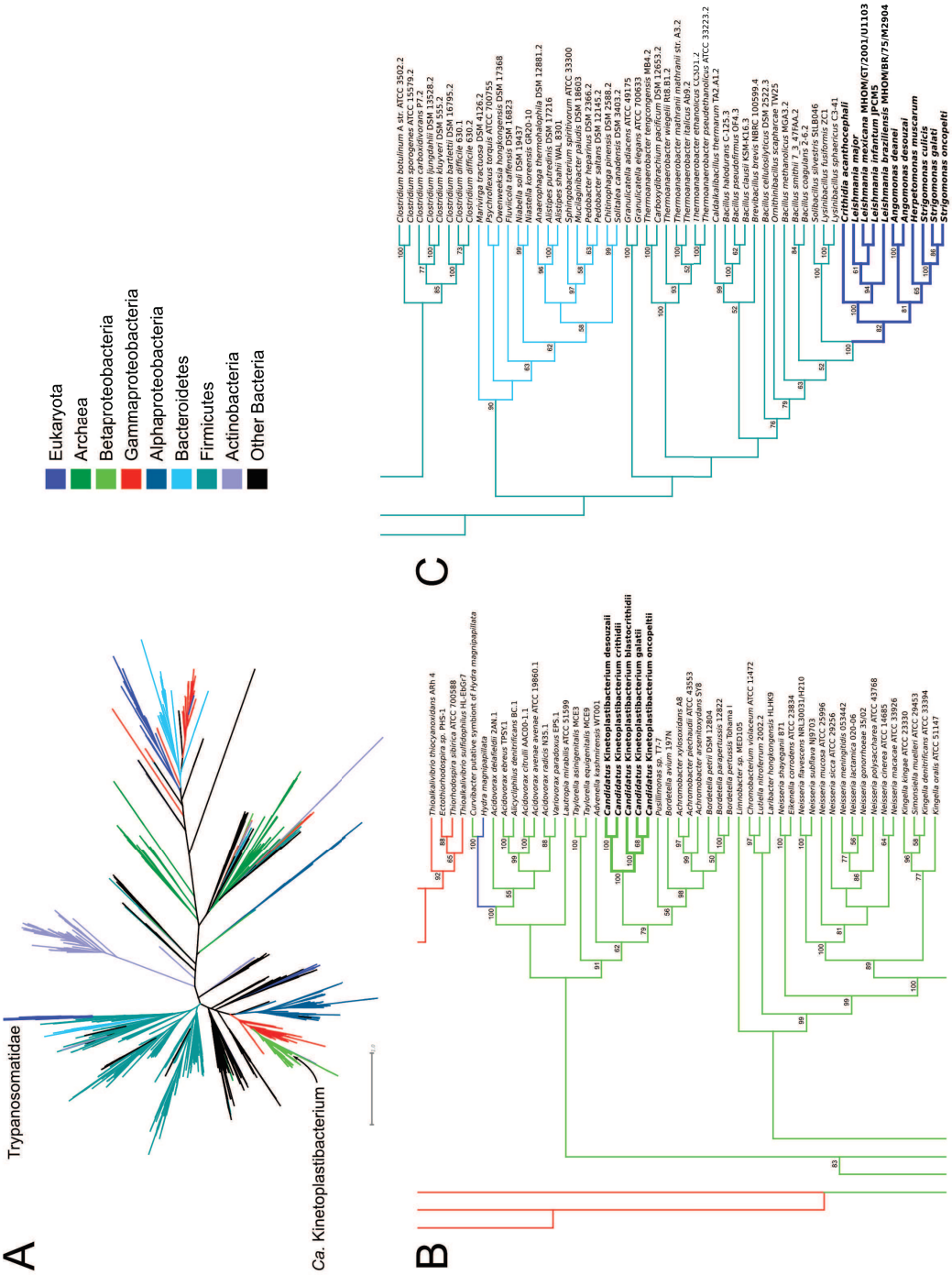


Figure 2.8: **Maximum likelihood phylogenetic tree of homoserine dehydrogenase (EC:1.1.1.3).** A – overall tree, colored according to the taxonomic affiliation of each taxon, as per the legend on the right; the distance bar only applies to panel A. B – details of the region of the tree where the *Ca.* Kinetoplastibacterium spp. are placed. C – details of the region of the tree where the Trypanosomatidae are placed. The values on the 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 the estimated distances between the sequences.

HGT and lysine biosynthesis The two genes of the lysine pathway (Figure 2.1) that were found in the 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 phylogenetically clusters strongly with the phylum Bacteroidetes, with BSV of 99 (Additional file C.4 in Appendix C) . The last gene, diaminopimelate decarboxylase (EC:4.1.1.20), was present in the SHTs and RTs. 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 C.5 in Appendix C) . 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.

Manual search using the *C. acanthocephali* gene for EC:4.1.1.20 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. A search against predicted *L. major* proteins yielded no results. What remains of this sequence suggests 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 2.2) 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 SHTs and RTs grouped inside a large cluster of diverse Bacteria (predominantly Bacteroidetes and Betaproteobacteria), with high BSV of 80 (Additional file C.6 in Appendix C) . 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 cannot be considered a case of HGT given the current results. Its gene is present in symbiont-harboring trypanosomatids 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 C.7 in Appendix C). 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 observed (reviewed in Andersson (2009)).

The gene for EC:2.5.1.47 (Additional file C.7 in Appendix C) is present in SHTs and RTs (including one sequence from *T. cruzi* CL Brener) and clusters as a sister group of Actinobacteria, although with low BSV.

The gene for EC:2.3.1.46, the first in the pathway converting homoserine to cystathionine, is present in all SHTs and *Herpetomonas*, but in no other RT 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 C.8 in Appendix C) .

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 SHTs and RTs, 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 C.9 in Appendix C) , with the few Eukaryota present forming a weakly supported clade. KEGG shows that many Eukaryota do possess the 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 bacte-

rial 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 Andersson (2009)), 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 in the trypanosomatids (Additional file C.10 in Appendix C). EC:2.5.1.48 is present in all SHTs and RTs 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 SHTs and *Herpetomonas*, but in none of the other RTs 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 SHTs and RTs, 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 (2.1.2 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 C.11 and C.12 in Appendix C), are present in all RTs and SHTs (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, the BSV is low and this gene cannot be considered a case of HGT given the 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 C.3 in Appendix C) , 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 SHTs and TPEs, seems to have originated in a bacterial group (Additional file C.13 in Appendix C) . These two organisms' enzymes group deep within the Gammaproteobacteria, with BSV of 97.

The enzyme acireductone synthase (EC:3.1.3.77) presents an intriguing case, being the only methionine salvage pathway enzyme absent from the SHT 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 RTs 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 2.3), EC:4.1.2.5, was identified in all SHTs and RTs (except *Herpetomonas*), but the evolutionary histories of SHT and RT genes are very different (Additional file C.14 in Appendix C). The gene found in the RTs *Leishmania* sp. and *C. acanthocephali* groups deep within the Firmicutes, most closely to *Clostridium*, with BSV of 63. The SHT 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 BSVs are low, there is a large group of bacteria from diverse phyla and a few other eukaryotic groups between the SHTs and the other eukaryotes in this part of the tree. It is therefore difficult to conclude whether the SHT gene is of bacterial or eukaryotic origin.

HGT and tryptophan biosynthesis The tryptophan synthase beta subunit (EC:4.2.1.20), present in the SHTs 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 C.15 in Appendix C) . 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 2.7). We have found that the genes for both enzymes, present in all SHT and RT 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 C.16 in Appendix C) . 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 SHTs and RTs group together in a monophyletic clade (Additional file C.17 in Appendix C) , which clusters within a large group of mostly Betaproteobacteria with BSV of 80, including the Alcaligenaceae, the family to which the TPEs belong. However, it seems highly unlikely that this sequence has been transferred from the TPE genomes to their host genomes because the nuclear sequences are clearly removed from the Alcaligenaceae, and many RTs (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 SHTs utilize the OCT provided by their endosymbionts, and their OCT genes group clearly 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 the TPE genomes and present in all the SHT and RT genomes – the only exception being the lack of the latter in *Leishmania* spp. The trypanosomatid genes form monophyletic

groups in their respective trees, within the Firmicutes in both cases (Additional files C.18 and C.19 in Appendix C). 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, the support is weak for deeper branches in the trees. Although the genomic sequences of the hosts 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 SHTs (Additional file C.1 in Appendix C) . The flanking genes are eukaryotic: terbinafine resistance locus protein and a multidrug resistance ABC transporter. As seen in the genome browser 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 SHTs and RTs 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 SHTs and RTs 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 except for the one from *Herpetomonas* cluster together with very high bootstrap support of 98, within Eukaryota (Additional file C.20 in Appendix C). 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 the SHT genomes is very similar to those from Betaproteobacteria of the Alcaligenaceae family, to which the TPEs belong. The RT and TPE genomes do not contain the gene for this enzyme. Accordingly, the phylogeny shows the SHT gene grouping close to several Alcaligenaceae, although the clade is not monophyletic and presents low BSV (Additional file C.21 in Appendix C). This grouping, together with the gene presence in the SHT genomes only, raises the possibility that EC:4.3.1.12 has been transferred from the ancestral TPE to the corresponding host, before the radiation of SHTs into the two genera and five species analyzed here.

Other observations on peculiarities of some of the amino acid pathways 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 2.4), was identified in all bacteria of the Alcaligenaceae family present in KEGG, except for the closest relatives of TPEs, *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 TPEs,

which are sister groups (Alves *et al.*, 2013b), had already lost the gene. Another possibility is that independent losses occurred in TPEs, *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 the 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 TPEs 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 (Mormann *et al.*, 2006; Petersen *et al.*, 2010). Our searches for the novel *C. glutamicum* HPP (cg0910, an inositol monophosphatase-like gene) have identified two possible candidate genes in the TPEs (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.

CONCLUSION

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. In fact, amino acid biosynthetic pathways in SHTs 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 (Figure 2.9), as previously seen in other systems (McCutcheon et von Dohlen, 2011).

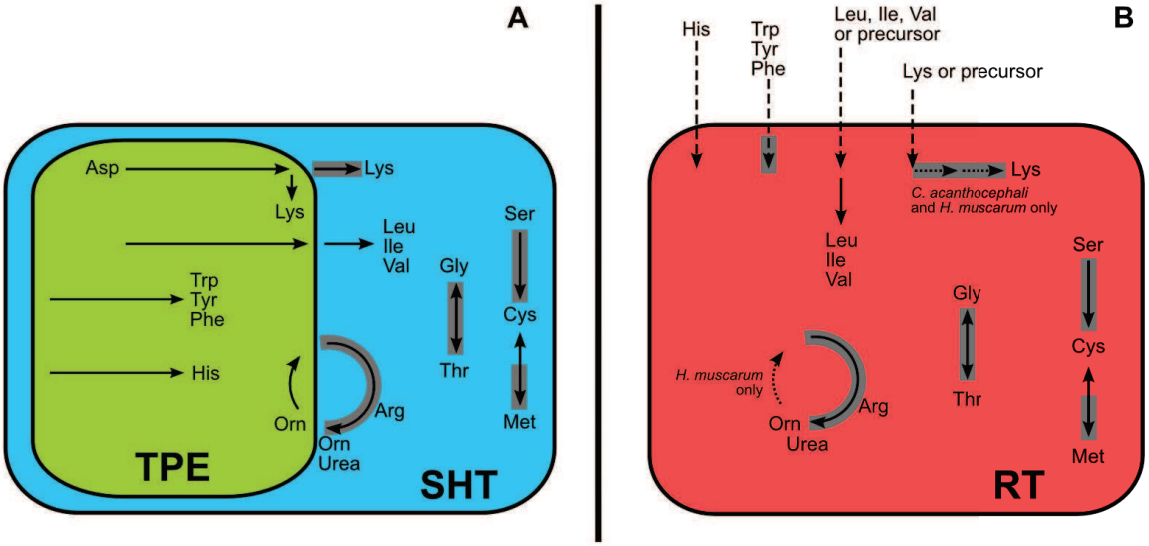


Figure 2.9: **Overview of the biosynthetic pathways of essential amino acids in trypanosomatids.** Dashed arrows: metabolite import; dotted arrows: reaction present in only some of the organisms analyzed; solid arrows: other reactions (a single arrow can summarize multiple steps); arrows surrounded by a gray box: enzymes possibly acquired through horizontal transfer from Bacteria to trypanosomatids (see main text). A. Contribution of SHTs and TPEs 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, based on the analysis of genomic data from *H. muscarum*, *C. acanthocephali*, and *L. major*.

Sometimes, as in the lysine and histidine synthesis, the symbionts contain all the genes for enzymes that compose the metabolic route. By contrast, in the cysteine and methionine pathway, the bacteria lack most genes involved in amino acid interconversion, which are present in the 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 TPE genome. This phenomenon has also been observed in the synthesis of heme (Alves *et al.*, 2011; Korený *et al.*, 2010), but the reasons for this peculiarity remain obscure. However, we have to consider the possibility that HGT events preceded the colonization of SHTs by TPEs, 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 some function, such as the control of amino acid production by the host trypanosomatid. This same strategy can be considered in the production of isoleucine, valine, and leucine, but in this case TPEs

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 SHTs 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 (Newton, 1956; Mundim *et al.*, 1974; Guttman, 1966; Mundim et Roitman, 1977). Symbiont-bearing trypanosomatids contain genes for all enzymes leading from glutamate to arginine. The corresponding genes are located partly in the genomes of their TPEs and partly in the protozoan nucleus. In this last case, the 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, TPEs 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 SHTs. The results in this study confirm previous findings (Alves *et al.*, 2013b, 2011) showing the betaproteobacterial origin of the genes of TPEs. 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, e.g. methionine/cysteine and arginine/ornithine synthesis. Transferred genes originated preferentially from three bacterial phyla, namely Firmicutes, Bacteroidetes, and Gammaproteobacteria (in decreasing order of occurrence), 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 the nuclear genomes of SHTs and not in any of the currently sequenced RT genomes. During the revision process of this work, a very recent report was published ((Husnik *et al.*, 2013)) of a similar situation of multiple lineages contributing to the metabolism in the symbiosis of mealybugs, involving the three interacting partners and genes acquired by the insect host through HGT from other bacterial sources (mainly Alphaproteobacteria, but also Gammaproteobacteria and Bacteroidetes). This suggests that this phenomenon could be widespread and of great importance in the evolution of genomes and metabolisms.

Having been detected in more than 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.*, in submitted), heme, and amino acids have already been shown to benefit from bacterial-trypanosomatid HGT; many other processes in the metabolism of Trypanosomatidae might also be subjected to this evolutionary process.

METHODS

Organisms and growth conditions The genomes of the symbiont-harboring trypanosomatid species sequenced here were: *Strigomonas oncopelti* TCC290E, *S. culicis* TCC012E, *S. galati* TCC219, *Angomonas deanei* TCC036E, and *A. desouzai* TCC079E. These SHTs harbor, respectively, the symbionts: *Candidatus* Kinetoplastibacterium oncopeltii, *Ca.* K. blastocrithidii, *Ca.* K. galatii, *Ca.* K. crithidii and *Ca.* K. desouzaii (Teixeira *et al.*, 2011), which were previously sequenced (Alves *et al.*, 2013b). In addition, we have also sequenced the genomes of two RT organisms, 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. SHTs were grown in Graces' medium (Gibco). RTs were grown in LIT media (Camargo, 1964).

DNA extraction and sequencing Total genomic DNA was extracted by the phenol-chloroform method (Ozaki et Cseko, 1984). We applied kDNA depletion methods to minimize the presence of this type of molecule, as previously described (Alves *et al.*, 2011), which result in less than about 5% of remaining kDNA in the sample. After kDNA depletion, about 5μg 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 under BioProject IDs PRJNA203418 and PRJNA203515-203520. The endosymbiont genomes were finished to a closed circle as previously described (Alves *et al.*, 2013b).

Gene discovery and annotation Endosymbiont genes were used as previously published (Alves *et al.*, 2013b). In an initial scan of the genome, trypanosomatid genes were discovered and mapped to the metabolic pathways using ASGARD (Alves et Buck, 2007), employing as reference the UniRef100 (Suzek *et al.*, 2007) and the Kyoto Encyclopedia of Genes and Genomes, KEGG (Ogata *et al.*, 1999) 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 (Stein *et al.*, 2002). Putative sequence functions were confirmed by domain searches against NCBI's Conserved Domain Database (Marchler-Bauer *et al.*, 2011). 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 C.22 in Appendix C. All endosymbiont genes analyzed here have been previously sequenced (Alves *et al.*, 2013b); gene identifiers are available from Additional file C.23 in Appendix C.

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 (Stein *et al.*, 2002) 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 taxonomic groups as widespread 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 (Edgar, 2004). Phylogenetic inferences were performed by the maximum likelihood method, using RAxML v. 7.2.8 (Stamatakis, 2006) and employing the WAG amino acid substitution model (Whelan et Goldman, 2001), 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 (Stover et Muller, 2010) and Dendroscope (Huson *et al.*, 2007), 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 C.2 in Appendix C) if the BSV was 80 or greater, and moderate if the BSV was between 50 and 80 – with one exception, EC:2.1.1.37, described in the results.

2.2.2 Biosynthesis of vitamins and cofactors in bacterium-harbouring trypanosomatids depends on the symbiotic association as revealed by genomic analyses

Introduction

As concerns the need for vitamins by RTs, very little is known mainly because their growth media are very complex, making it difficult to define their specific nutritional requirements. Despite this, various papers addressed indirect aspects of vitamin metabolism (Cowperthwaite *et al.*, 1953; Hutner *et al.*, 1956; Nathan *et al.*, 1960; Nathan et Cowperthwaite, 1955; Guttman, 1962; Hutner *et al.*, 1979, 1980; Fiorini, 1989). The development of a defined medium for RTs from insects had initially established that seven vitamins are essential to sustain protozoan growth in culture medium: riboflavin, pantothenic acid, pyridoxamine, folic acid, thiamine, nicotinic acid, and biotin (Roitman *et al.*, 1972). Studies on the nutritional requirements of insect trypanosomatids did not progress in a satisfactory way, but interestingly demonstrated that SHTs of the genus *Angomonas* are nutritionally much less exigent than RTs (Mundim *et al.*, 1974). Thus, while the autotrophy of SHTs for most of the B vitamins was evidenced, nothing was known about pathways for the synthesis of other vitamins. Furthermore, any direct evidence of the symbiont contribution to the vitamin synthetic capabilities of the host trypanosomatid was missing.

In recent studies, we reported on the sequencing of the entire genomes of five species of TPEs (Alves *et al.*, 2013b) and we also annotated the proteins of two SHT species and their respective symbionts (Motta *et al.*, 2013). Moreover, we sequenced to a draft-level the genomes of the five host species as well as of two RTs (Alves *et al.*, 2013a). In this paper, we analyze these genomes for the presence of genes involved in the synthesis of vitamins. The participation of both host and symbiont in the production of vitamins is presented and discussed in association with previous data on the nutritional requirements of RTs and SHTs. In order to get a broader view, we compared our findings with other trypanosomatids and bacteria from the Alcaligenaceae family based on KEGG (Ogata *et al.*, 1999).

Materials and methods

Analyzed organisms and their genome sequences The genomes of the following SHTs and of the respective symbionts were examined: *Strigomonas oncopelti* TCC290E (accession number AUXK00000000), *S. culicis* TCC012E (AUXH00000000), *S. galati* TCC219 (AUXN00000000), *Angomonas deanei* TCC036E (AUXM00000000), and *A. desouzai* TCC079E (AUXL00000000) (Alves *et al.*, 2013a). Their corresponding symbionts are referred to as: "*Candidatus* Kinetoplastibacterium oncopeltii", "*Ca.* K. blastocrithidii", "*Ca.* K. galatii", "*Ca.* K. crithidii", and "*Ca.* K. desouzaii" (Teixeira *et al.*, 2011). The endosymbiont genomes were finished to a closed circle as previously described (Alves *et al.*, 2013b).

The genomes of two RTs were also analyzed: *Herpetomonas muscarum* TCC001E (AUXJ00000000) and *Crithidia acanthocephali* TCC037E (AUXI00000000) (Alves *et al.*, 2013a).

Gene discovery and annotation Initially, the trypanosomatid genes were discovered and mapped to metabolic pathways using ASGARD (Alves et Buck, 2007), using as reference the UniRef100 (Suzek *et al.*, 2007) and KEGG (Ogata *et al.*, 1999) databases. The identified segments of DNA were then extracted from the genomes and manually curated for completion and proper location of start and stop codons by using the GBrowse genome browser (Donlin, 2009). Putative sequence functions were confirmed by domain searches against the NCBI's

CDD (conserved domain database) (Marchler-Bauer *et al.*, 2011). For 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 taxonomic groups as widespread as possible and keeping one from each species (or genus, if the tree was too large) for subsequent phylogenetic analysis. Only sequences that were complete and aligned along at least 75% of the length of the query were selected.

All trypanosomatid genes characterized in this study have been submitted to NCBI's GenBank; accession numbers are available in Table 2.10. All endosymbiont genes analyzed here have been previously sequenced (Alves *et al.*, 2013b); gene identifiers are available in Table 2.11.

For comparison, we used in our analyses the genome annotations of trypanosomatids (*Trypanosoma brucei*, *T. cruzi*, *Leishmania major*, *L. infantum*, *L. donovani*, *L. mexicana*, *L. braziliensis*) and bacteria from the Alcaligenaceae family (*Bordetella pertussis* Tohama II, *B. pertussis* CS, *B. pertussis* 18323, *B. parapertussis* 12822, *B. parapertussis* Bpp5, *B. bronchiseptica* RB50, *B. bronchiseptica* MO149, *B. bronchiseptica* 253, *B. petrii*, *B. avium*, *Achromobacter xylosoxidans*, *Taylorella equigenitalis* MCE9, *T. equigenitalis* ATCC 35865, *T. asinigenitalis*, *Pusillimonas* sp. T7-7, *Advenella kashmirensis*) available in KEGG (Ogata *et al.*, 1999). However, care should be taken since these data may lack manual curation. As concerns information on metabolic pathways, we used KEGG (Ogata *et al.*, 1999) and MetaCyc (Caspi *et al.*, 2012).

Phylogenetic analyses All analyses were performed at the protein sequence level. Sequences were aligned by using MUSCLE (Edgar, 2004) and phylogenetic inferences were performed by the maximum likelihood (ML) method using RAxML v. 7.2.8 (Stamatakis, 2006) and the WAG amino acid substitution model (Whelan et Goldman, 2001), 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. We also performed phylogenetic inferences by the neighbor joining (NJ) method using the seqboot and neighbor programs from PHYLIP v. 3.69 (Felsenstein, 1989) and RAxML v. 7.2.8 for the distance matrix calculation (in order to use the same amino acid substitution model) and for drawing the bootstrap support values (100 replicates) in the NJ tree. Trees were initially drawn and formatted using TreeGraph2 (Stover et Muller, 2010) and Dendroscope (Huson *et al.*, 2007), with subsequent cosmetic adjustments performed with the Inkscape vector image editor (http://inkscape.org). CodonW (Peden, 2006) was used to perform correspondence analyses of codon usage and to calculate codon adaptation index scores for the candidate HGT genes using an endosymbiont gene as a negative control.

	Accession numbers						
EC number	*Angomonas deanei*	*Angomonas desouzai*	*Strigomonas culicis*	*Strigomonas oncopelti*	*Strigomonas galati*	*Crithidia acanthocephali*	*Herpetomonas muscarum*
1.1.1.100	KF160081	KF160098	KF160036	KF160137	KF160252	KF160178	KF160225
1.1.1.169	KF160064	KF160119, KF160120, KF160121	KF160045	KF160154	KF160291		KF160217
1.5.1.3	KF160067	KF160109	KF160029	KF160142	KF160283	KF160198	KF160213
1.5.1.34	KF160060	KF160118	KF160033	KF160143	KF160243	KF160176	KF160215
2.1.1.201	KF160070	KF160094	KF160053	KF160147	KF160248	KF160180	KF160207
2.3.1.41	KF160084	KF160107	KF160031	KF160161	KF160286	KF160182	KF160235
2.4.2.11	KF160090	KF160110	KF160044	KF160140	KF160277	KF160181	KF160232
2.6.1.5	KF160061, KF160062	KF160101	KF160018, KF160019, KF160020	KF160149, KF160150, KF160151	KF160271, KF160272, KF160273, KF160274, KF160290	KF160193, KF160194, KF160195, KF160196	KF160239
2.7.1.23	KF160066	KF160125	KF160025	KF160163	KF160292	KF160205	KF160240
2.7.1.24	KF160086	KF160093	KF160028	KF160157	KF160246	KF160202	KF160222
2.7.1.26	KF160059	KF160096	KF160032	KF160155	KF160275	KF160179	KF160242
2.7.1.33	KF160085	KF160122	KF160039	KF160129	KF160276	KF160191	KF160219
2.7.1.35	KF160091	KF160113	KF160030	KF160165	KF160287	KF160192	KF160228
2.7.7.1	KF160080	KF160092	KF160037	KF160139	KF160267	KF160186	KF160229
2.7.7.2	KF160073	KF160097	KF160048	KF160146	KF160249	KF160168	KF160216
2.7.7.3	KF160063	KF160095	KF160027	KF160148	KF160279	KF160189	KF160220
2.8.1.7	KF160079	KF160102	KF160021	KF160144	KF160253	KF160175	KF160227
3.1.3.1	KF160074	KF160108	KF160052	KF160138	KF160278	KF160177	KF160234
3.2.2.1	KF160083	KF160099	KF160055	KF160141	KF160269	KF160170	KF160214
3.5.1.19						KF160171	KF160241
3.6.1.22	KF160076	KF160106	KF160043	KF160158	KF160288	KF160169	KF160230, KF160231
3.7.1.3	KF160068	KF160126	KF160051	KF160164	KF160293	KF160167	KF160237
4.1.1.36	KF160082	KF160117	KF160035	KF160159	KF160244	KF160187	KF160218
4.1.3.40			KF160050	KF160156	KF160289		
6.2.1.12	KF160077, KF160078	KF160112	KF160046, KF160047	KF160134, KF160135, KF160136	KF160254, KF160255, KF160256, KF160257, KF160258, KF160259, KF160260, KF160261, KF160262, KF160263, KF160264, KF160265, KF160266	KF160190	KF160223, KF160224
6.3.2.17	KF160071	KF160111	KF160026	KF160128	KF160245	KF160184	KF160238
6.3.2.5	KF160075	KF160100	KF160022	KF160153	KF160284	KF160166	KF160226
6.3.4.15	KF160069		KF160034	KF160160	KF160268	KF160172, KF160173, KF160174	KF160233
6.3.5.1	KF160065	KF160103	KF160049	KF160152	KF160285	KF160188	KF160206
Coq7	KF160089	KF160104	KF160054	KF160162	KF160270	KF160185	KF160212
UbiACoq2	KF160087	KF160123	KF160038	KF160127	KF160250	KF160204	KF160211
UbiB	KF160056, KF160057, KF160058	KF160114, KF160115, KF160116	KF160040, KF160041, KF160042	KF160130, KF160131, KF160132	KF160280, KF160281, KF160282	KF160199, KF160200, KF160201	KF160208, KF160209, KF160210
UbiG	KF160088	KF160105	KF160024	KF160133	KF160247	KF160197	KF160236
UbiH	KF160072	KF160124	KF160023	KF160145	KF160251	KF160183	KF160221

Figure 2.10: **Trypanosomatidae genes characterized in this study.**

Pathway	EC number	Ca. K. crithidii	Ca. K. desouzaii	Ca. K. blastocrithidii	Ca. K. oncopeltii	Ca. K. galatii
Riboflavin and FAD	3.5.4.25	CDEe_0522	CDSe_0515	BCUe_0506	CONe_0495	ST1e_0570
Riboflavin and FAD	4.1.99.12	CDEe_0522	CDSe_0515	BCUe_0506	CONe_0495	ST1e_0570
Riboflavin and FAD	3.5.4.26	CDEe_0029	CDSe_0026	BCUe_0022	CONe_0027	ST1e_0028
Riboflavin and FAD	1.1.1.193	CDEe_0029	CDSe_0026	BCUe_0022	CONe_0027	ST1e_0028
Riboflavin and FAD	2.5.1.78	CDEe_0523	CDSe_0516	BCUe_0507	CONe_0496	ST1e_0571
Riboflavin and FAD	2.5.1.9	CDEe_0028	CDSe_0024	BCUe_0021	CONe_0026	ST1e_0027
Riboflavin and FAD	2.7.1.26 / 2.7.7.2	CDEe_0479	CDSe_0471	BCUe_0467	CONe_0450	ST1e_0518
Pantothenic acid and CoA	2.1.2.11	CDEe_0142	CDSe_0138	BCUe_0139	CONe_0133	ST1e_0148
Pantothenic acid and CoA	6.3.2.1	CDEe_0061	CDSe_0057	BCUe_0050	CONe_0056	ST1e_0057
Pantothenic acid and CoA	2.7.1.33	CDEe_0137	CDSe_0133	BCUe_0131	CONe_0127	ST1e_0141
Pantothenic acid and CoA	6.3.2.5 / 4.1.1.36	CDEe_0474	CDSe_0468	BCUe_0463	CONe_0447	ST1e_0513
Pantothenic acid and CoA	2.7.7.3	CDEe_0327	CDSe_0321	BCUe_0324	CONe_0317	ST1e_0358
Pantothenic acid and CoA	2.7.1.24	CDEe_0284	CDSe_0277	BCUe_0286	CONe_0277	ST1e_0309
Vitamin B6	2.6.1.52	CDEe_0716	CDSe_0702	BCUe_0692	CONe_0673	ST1e_0774
Vitamin B6	2.2.1.7	CDEe_0650	CDSe_0637	BCUe_0626	CONe_0609	ST1e_0707
Vitamin B6	1.1.1.262	CDEe_0643	CDSe_0630	BCUe_0618	CONe_0602	ST1e_0699
Vitamin B6	2.6.99.2	CDEe_0412	CDSe_0411	BCUe_0408	CONe_0394	ST1e_0449
Vitamin B6	1.4.3.5	CDEe_0868	CDSe_0853	BCUe_0842	CONe_0814	ST1e_0951
Folic acid	3.5.4.16	CDEe_0649	CDSe_0636	BCUe_0625	CONe_0608	ST1e_0706
Folic acid	4.1.2.25	CDEe_0263	CDSe_0257	BCUe_0260	CONe_0256	ST1e_0281
Folic acid	2.7.6.3	CDEe_0882	CDSe_0865	BCUe_0856	CONe_0827	ST1e_0967
Folic acid	2.5.1.15	CDEe_0375	CDSe_0373	BCUe_0377	CONe_0369	ST1e_0419
Folic acid	6.3.2.12 / 6.3.2.17	CDEe_0571	CDSe_0563	BCUe_0549	CONe_0542	ST1e_0621
Folic acid	1.5.1.3	CDEe_0315	CDSe_0305	BCUe_0311	CONe_0304	ST1e_0341
Thiamine	2.7.1.49 / 2.7.4.7	CDEe_0282	CDSe_0275			
Thiamine	2.5.1.3	CDEe_0059	CDSe_0055			
Thiamine	2.8.1.7	CDEe_0496	CDSe_0490	BCUe_0483	CONe_0470	ST1e_0541
Thiamine	2.8.1.10	CDEe_0283	CDSe_0276			
Thiamine	2.7.4.16	CDEe_0525	CDSe_0518			
Thiamine	2.7.7.73	CDEe_0066	CDSe_0064			
Nicotinic acid and NAD	2.7.7.18	CDEe_0441	CDSe_0436	BCUe_0433	CONe_0419	ST1e_0482
Nicotinic acid and NAD	6.3.5.1	CDEe_0559	CDSe_0552	BCUe_0540	CONe_0530	ST1e_0611
Nicotinic acid and NAD	2.7.1.23	CDEe_0836	CDSe_0825	BCUe_0807	CONe_0791	ST1e_0912
Nicotinic acid and NAD	2.4.2.11	CDEe_0771	CDSe_0758	BCUe_0746	CONe_0724	ST1e_0843
Biotin	2.1.1.197	CDEe_0074	CDSe_0071	BCUe_0067	CONe_0070	ST1e_0072
Biotin	2.3.1.179	CDEe_0758	CDSe_0743	BCUe_0731	CONe_0710	ST1e_0829
Biotin	1.1.1.100	CDEe_0756	CDSe_0741	BCUe_0729	CONe_0708	ST1e_0826
Biotin	4.2.1.59	CDEe_0590	CDSe_0580	BCUe_0567	CONe_0561	ST1e_0643
Biotin	1.3.1.10	CDEe_0024	CDSe_0021	BCUe_0019	CONe_0022	ST1e_0023
Ubiquinone	UbiA / Coq2			BCUe_0137	CONe_0132	ST1e_0147
Ubiquinone	UbiD / UbiX			BCUe_0348; BCUe_0294	CONe_0341; CONe_0287	ST1e_0388; ST1e_0320
Ubiquinone	UbiB			BCUe_0285	CONe_0276	ST1e_0308
Ubiquinone	UbiG			BCUe_0695	CONe_0676	ST1e_0779
Ubiquinone	UbiH			BCUe_0821; BCUe_0258	CONe_0799; CONe_0255	ST1e_0931; ST1e_0280
Ubiquinone	UbiE			BCUe_0280	CONe_0271	ST1e_0304
Ubiquinone	UbiF / Coq7			BCUe_0313	CONe_0306	ST1e_0343

Figure 2.11: ***Ca.*** **Kinetoplastibacterium genes analyzed in this study.**

Results and Discussion

We analyzed the genomes of five species of SHTs and of their TPEs for the presence/absence of genes from the metabolic pathways for essential vitamin synthesis. The genomes of two RTs, *C. acanthocephali* and *H. muscarum*, were examined in detail, however these data do not fully represent the genomic diversity of insect trypanosomatids in general. Indeed, the enormous diversity present in the Trypanosomatidae family is sometimes not fully appreciated, leading to apparent conflicts in the interpretation of metabolic data, as happened with the early studies on the nutrition of *Crithidia* species. Data on the nutritional requirements of *C. fasciculata* strongly disagreed with those obtained for *C. oncopelti* (recently renamed as *Strigomonas oncopelti*) as concerns the necessity for amino acids and vitamins which is quite different for both organisms (Kidder et Dutta, 1958; Newton, 1956). Many years elapsed until it was realized that these organisms were quite distinct phylogenetically, and in fact belonged to different genera (Teixeira *et al.*, 2011). It became clear that *S. oncopelti*, as well as other trypanosomatids which were later isolated, carried a bacterial symbiont that probably endowed the host with enhanced biosynthetic capabilities. According to our present data, these extra nutritional capabilities largely result from the contribution of the endosymbiont to the metabolism of their trypanosomatid hosts as will be discussed here when analyzing vitamin biosynthesis in SHTs.

Autotrophy of SHTs for the synthesis of riboflavin, pantothenic acid, vitamin B_6 and folic acid

Riboflavin (Vitamin B_2) Riboflavin is essential for the growth of RTs, as well as for the aposymbiotic strains of SHTs (Cowperthwaite *et al.*, 1953; Kidder et Dutta, 1958; Mundim *et al.*, 1974), but not for the symbiont-carrying strains of SHTs, which are autotrophic for this vitamin (Newton, 1956; Mundim *et al.*, 1974; Roitman *et al.*, 1972). Riboflavin is synthesized from guanosine 5'-triphosphate (GTP) and ribulose 5'-phosphate (Figure 2.12), and is the precursor for the essential flavin cofactors of redox reactions: FMN (flavin mononucleotide) and FAD (flavin adenine dinucleotide) (Bacher *et al.*, 2001). The genomes of SHTs and RTs have none of the genes for the enzymes involved in riboflavin synthesis. On the other hand, TPEs have all the genes responsible for such synthesis, except for a poorly characterized step in the pathway, probably involving a phosphoric monoester hydrolase (Figure 2.12, IV-V). However, it is uncertain which enzyme is responsible for this dephosphorylation process although it was suggested that a phosphatase of low substrate specificity might be involved (Bacher *et al.*, 2001; Wu *et al.*, 2006). Bacteria from the Alcaligenaceae family have all the enzymes for the synthesis of riboflavin as is the case for TPEs, missing only the uncharacterized one (Figure 2.13). Since SHTs do not require riboflavin, it can be assumed that the dephosphorylation reaction is catalyzed by any of a cohort of phosphatases of broad substrate range.

Further along the riboflavin biosynthetic pathway, it can be seen that all trypanosomatids, with or without symbionts, have the genes for the conversion of riboflavin into FMN and of the latter into FAD. Those genes are also present in other trypanosomatids (Figure 2.13). It is worth considering that in SHTs the presence of such genes in the trypanosomatid host may be related to the control of the production of FMN and FAD. FMN acts as a coenzyme in oxidative enzymes, including NADH dehydrogenase while FAD forms the prosthetic group of certain oxidases, both serving as electron carriers. Recently, we proposed that the presence of the symbiont influences the energetic metabolism of *A. deanei* (unpublished data). This analysis reinforces this idea and reveals that, thanks to the genes of the symbiont, SHTs

are fully capable of riboflavin synthesis, corroborating the nutritional data that point to this vitamin as unnecessary for the growth of SHTs, although indispensable for the growth of RTs (Kidder et Dutta, 1958; Mundim *et al.*, 1974; Roitman *et al.*, 1972).

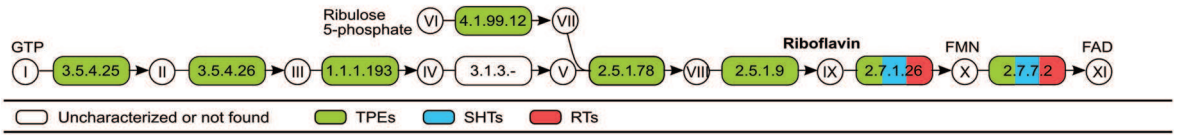


Figure 2.12: **Biosynthesis of riboflavin and FAD** Metabolites - I: Guanosine 5'-triphosphate; II: 2,5-Diamino-6-(5-phospho-D-ribosylamino)pyrimidin-4(3H)-one; III: 5-Amino-6-(5'-phosphoribosylamino)uracil; IV: 5-Amino-6-(5'-phospho-D-ribitylamino)uracil; V: 5-Amino-6-(1-D-ribitylamino)uracil; VI: D-Ribulose 5-phosphate; VII: 2-Hydroxy-3-oxobutyl phosphate; VIII: 6,7-Dimethyl-8-(D-ribityl)lumazine; IX: Riboflavin; X: Flavin mononucleotide; XI: Flavin adenine dinucleotide Enzymes - 3.5.4.25: GTP cyclohydrolase II; 3.5.4.26: diaminohydroxyphosphoribosylaminopyrimidine deaminase; 1.1.1.193: 5-amino-6-(5-phosphoribosylamino)uracil reductase; 3.1.3.-: Phosphoric monoester hydrolases; 4.1.99.12: 3,4-dihydroxy 2-butanone 4-phosphate synthase; 2.5.1.78: 6,7-dimethyl-8-ribityllumazine synthase; 2.5.1.9: riboflavin synthase; 2.7.1.26: riboflavin kinase; 2.7.7.2: FAD synthetase.

Pantothenic acid (Vitamin B_5) Early nutritional studies considered pantothenic acid as an absolute requirement for the growth of trypanosomatids (Cowperthwaite *et al.*, 1953; Kidder et Dutta, 1958). Later reports confirmed these observations, but showed also that pantothenate is not at all necessary for the cultivation of SHTs such as *S. oncopelti* and *A. deanei* (Newton, 1956; Mundim *et al.*, 1974; Roitman *et al.*, 1972). Bacteria synthesize coenzyme A (CoA) via pantothenic acid from aspartate and α-ketoisovalerate (Figure 2.14), while CoA is an acyl carrier required for a multitude of reactions for both biosynthetic and degradation pathways (Begley *et al.*, 2001b). The CoA biosynthetic route requires nine enzymes: four to synthesize pantothenic acid (Figure 2.14, I-VI) and five to produce CoA (Figure 2.14, VI-XI).

As concerns the first half, the enzyme aspartate 1-decarboxylase (EC:4.1.1.11), required for the conversion of aspartate into β-alanine (Figure 2.14, I-II), was not identified in TPEs, SHTs nor RTs. Moreover, the two latter groups possess the enzymes to catalyze the synthesis of β-alanine from malonyl-CoA. TPEs have two enzymes responsible for the synthesis of pantothenic acid (3-methyl-2-oxobutanoate hydroxymethyltransferase EC:2.1.2.11; and pantoate–beta-alanine ligase EC:6.3.2.1; Figure 2.14, III-IV and V-VI) which were not found in SHTs nor in RTs. Moreover, the genes necessary to convert pyruvate into α-ketoisovalerate (one precursor of this biosynthetic pathway) were only identified in TPEs, but neither in SHTs nor in RTs (Alves *et al.*, 2013a). These steps take part in the biosynthetic route of valine. The remaining step is the production of pantoate mediated by ketopantoate reductase (EC:1.1.1.169, Figure 2.14, VI-V), that participates exclusively in this pathway and was identified in all the SHTs analyzed and also in *H. muscarum*. In SHTs, its presence would be meaningful for the synthesis of pantothenic acid complemented by the TPEs, however the presence of this gene in *H. muscarum* is puzzling since RTs lack the remaining genes of the pantothenic acid biosynthetic pathway. As discussed in detail below in the Section *phylogenetic analyses*, this gene is likely a relic from a past lateral gene transfer event from a bacterium to a common ancestor of the Trypanosomatidae family.

Other trypanosomatids available in KEGG lack all the enzymes for the production of pantothenic acid. Most bacteria from the Alcaligenaceae family have all the machinery for

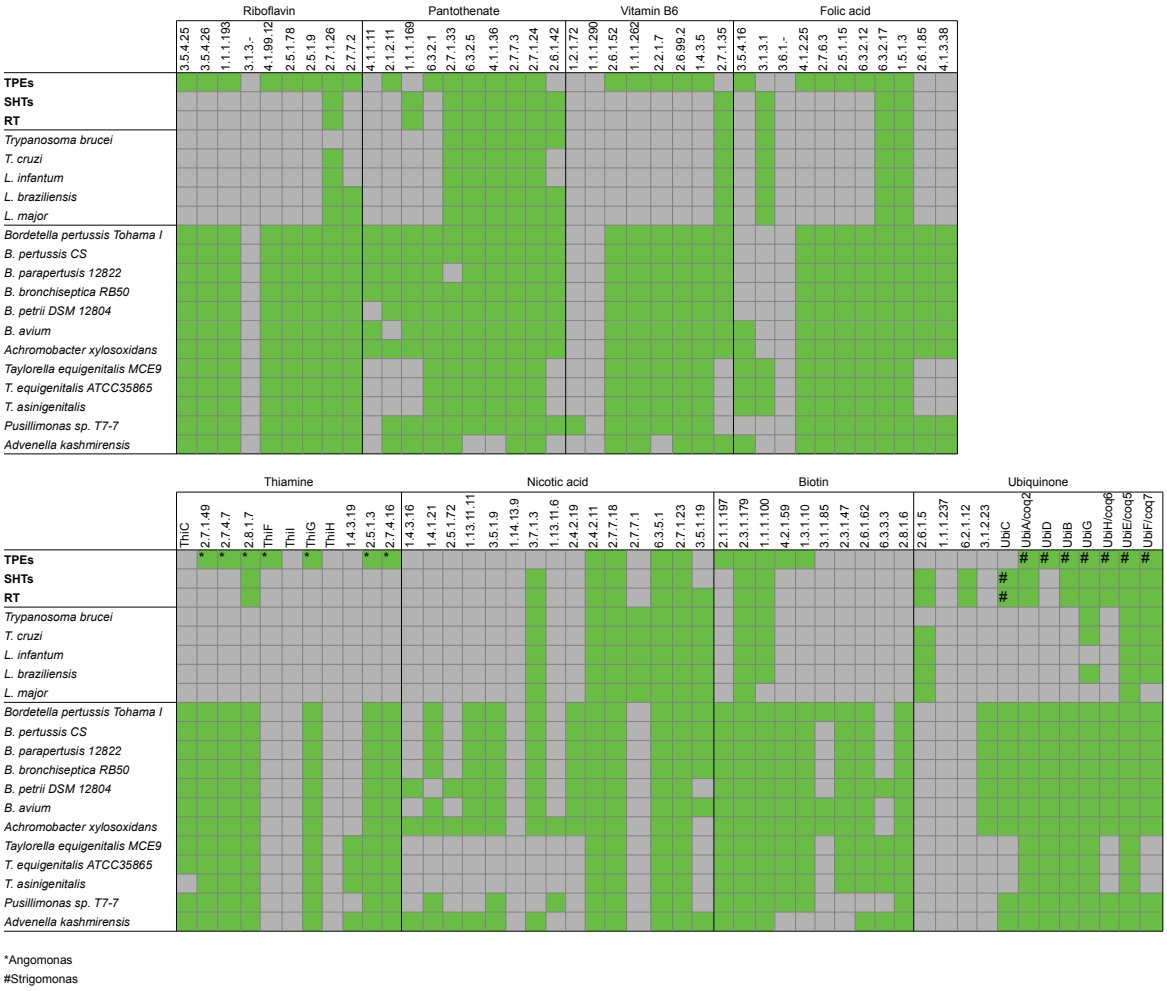


Figure 2.13: **Comparative analysis of trypanosomatids and bacteria from the Alcaligenaceae family and the species analyzed in the present study** Green squares indicate presence while gray ones indicate absence of the enzyme (column) in the respective organism (line). Information of trypanosomatids and bacteria other than the ones analyzed in the present study are from KEGG. *Only *Angomonas* spp. #Only *Strigomonas* spp.

this biosynthesis (*Bordetella* spp. and *A. xylosoxidans*) while *Pusillimonas* sp. and *A. kashmirensis* lack only enzyme EC:4.1.1.11. On the other hand, *Taylorella* spp, which have the most reduced genomes, lack the first 3 steps (Figure 2.13).

The second half of the pathway, which is the production of CoA from pantothenic acid, is accomplished by TPEs, SHTs, and RTs. In SHTs, we can assume that CoA formation is optimized by TPEs. This coenzyme is related to the synthesis and oxidation of fatty acids and oxidation of pyruvate in the Krebs cycle, which is directly associated with amino acid production. Recently, our genomic searches showed that the symbionts of the trypanosomatids contain the genes coding for enzymes that complete the host pathways for the synthesis of essential amino acids (Alves *et al.*, 2013a).

Based on those findings, the association of host/symbiont makes SHTs autotrophic for pantothenic acid production (Newton, 1956; Mundim *et al.*, 1974; Roitman *et al.*, 1972), whereas RTs depend on an exogenous source of this vitamin since they lack some genes for

the synthesis of pantothenic acid (Cowperthwaite *et al.*, 1953; Kidder et Dutta, 1958).

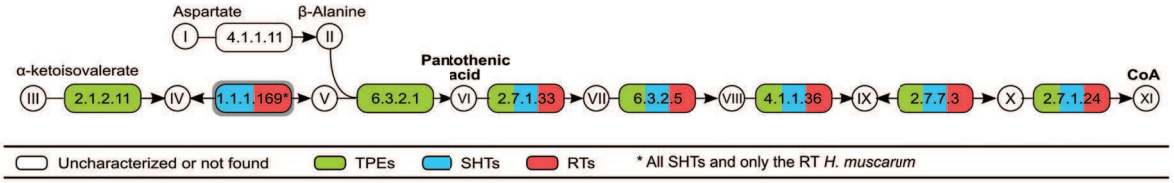


Figure 2.14: **Biosynthesis of pantothenic acid and coenzyme A** Enzymes surrounded by a gray box were possibly acquired through horizontal transfer from Bacteria to trypanosomatids (see main text). Metabolites - I: Aspartate; II: β-Alanine; III: α-ketoisovalerate; IV: 2-Dehydropantoate; V: Pantoate; VI: Pantothenic acid; VII: D-4'-Phosphopantothenate; VIII: (R)-4'-Phosphopantothenoyl-L-cysteine; IX: Pantetheine 4'-phosphate; X: Dephospho-coenzyme A; XI: Coenzyme A. Enzymes - 4.1.1.11: aspartate 1-decarboxylase; 2.1.2.11: 3-methyl-2-oxobutanoate hydroxymethyltransferase; 1.1.1.169: ketopantoate reductase; 6.3.2.1: pantoate–beta-alanine ligase; 2.7.1.33: pantothenate kinase; 6.3.2.5: phosphopantothenate-cysteine ligase; 4.1.1.36: phosphopantothenoylcysteine decarboxylase; 2.7.7.3: pantetheine-phosphate adenylyltransferase; 2.7.1.24: dephospho-CoA kinase.

Pyridoxal, pyridoxine and pyridoxamine (Vitamin B_6) Vitamin B_6 refers collectively to pyridoxal, pyridoxine, pyridoxamine and their corresponding phosphate esters. Its catalytically active forms are pyridoxal-5'-phosphate (PLP) and pyridoxamine 5'-phosphate (PMP) (Drewke et Leistner, 2001). This vitamin is essential for all organisms while PLP is an extremely versatile coenzyme necessary for over 100 enzymatic reactions, predominantly in the metabolism of amino acids (Drewke et Leistner, 2001; Eliot et Kirsch, 2004). Pyridoxal or pyridoxamine was described as an essential growth factor for RTs, as well as for the aposymbiotic strain of *A. deanei* (Kidder et Dutta, 1958; Mundim et Roitman, 1977). On the other hand, it was identified as not required by SHTs despite the fact that its presence doubled the growth rate of *S. oncopelti* (Newton, 1956; Mundim *et al.*, 1974).

As shown in Figure 2.15, the precursors for the de novo biosynthesis of PLP are D-erythrose-4-phosphate, glyceraldehyde-3-phosphate (GAP), and pyruvate (Drewke et Leistner, 2001). The genomes of RTs and SHTs have none of the enzymes for the synthesis of PLP, whereas TPEs have most of them, except for the first two steps mediated by the enzymes D-erythrose 4-phosphate dehydrogenase and erythronate-4-phosphate dehydrogenase (Epd EC:1.2.1.72 and PdxB EC:1.1.1.290, respectively), which convert D-erythrose-4-phosphate into 2-Oxo-3-hydroxy-4-phosphobutanoate (Figure 2.15, I-III). Since SHTs are autotrophic for PLP, these steps might be mediated by other distinct and unknown enzymes, or TPEs might use a precursor different from D-erythrose-4-phosphate. These same two steps are also missing in bacteria from the Alcaligenaceae family (Figure 2.13). The gene coding for Epd shares a high sequence similarity with gapA (gene coding for glyceraldehyde 3'-phosphate dehydrogenase, involved in glycolysis). Based on mutant essays, GapA was shown to be able to replace the Epd activity under certain conditions (Yang *et al.*, 1998). Since Epd was the only enzyme not identified in this pathway in *Ca.* B. cicadellinicola (endosymbiont of the sharpshooter), GapA was suggested as a candidate (Wu *et al.*, 2006). GapA is present in TPEs and also in all bacteria from the Alcaligenaceae family.

On the other hand, the RT and SHT genomes have the gene for pyridoxal kinase (EC:2.7.1.35), which converts pyridoxal, pyridoxine, and pyridoxamine into their respective phosphate es-

ters (salvage pathway), including PLP, the active principle of the B6 complex. However, they lack the oxidase responsible for the interconversion of the different forms of vitamin B_6 (see PdxH in Figure 2.15). Considering other trypanosomatids, there is no enzyme involved in this biosynthetic pathway, only the kinase above mentioned which is involved in the salvage pathway of vitamin B_6 (Figure 2.13).

Together, these findings underline the auxotrophy of RTs and the autotrophy of SHTs for the B6 complex (Kidder et Dutta, 1958; Newton, 1956; Mundim *et al.*, 1974; de Menezes et Roitmanz, 1991; Roitman *et al.*, 1972). Furthermore, PLP is an active coenzyme that acts especially on the metabolism of amino acids. This can be directly related to the low nutritional requirement of SHTs since essential amino acids are synthesized by the symbiotic bacterium (Alves *et al.*, 2013a).

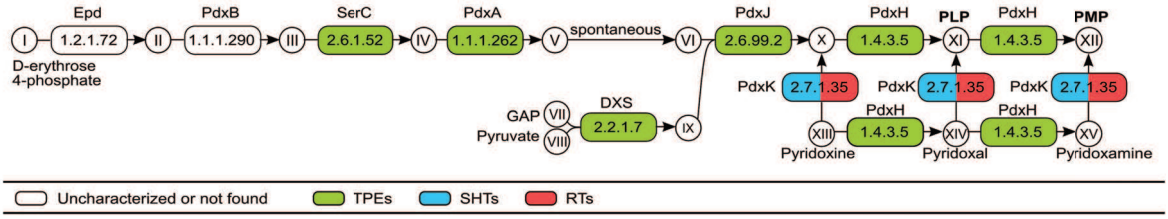


Figure 2.15: **Biosynthesis of vitamin B_6** Metabolites - I : D-Erythrose 4-phosphate; II: 4-Phospho-D-erythronate; III: 2-Oxo-3-hydroxy-4-phosphobutanoate; IV: 4-phospho-hydroxy-L-threonine; V: 2-amino-3-oxo-4-phosphonooxybutyrate; VI: 3-Amino-2-oxopropyl phosphate; VII: D-glyceraldehyde 3-phosphate; VIII: pyruvate; IX: 1-deoxy-D-xylulose 5-phosphate; X: Pyridoxine phosphate; XI: Pyridoxal 5'-phosphate (PLP); XII: Pyridoxamine phosphate; XIII: Pyridoxine; XIV: Pyridoxal; XV: Pyridoxamine. Enzymes - 1.2.1.72: D-erythrose 4-phosphate dehydrogenase; 1.1.1.290: erythronate-4-phosphate dehydrogenase; 2.6.1.52: phosphoserine aminotransferase; 1.1.1.262: 4-hydroxythreonine-4-phosphate dehydrogenase; 2.2.1.7: 1-deoxyxylulose-5-phosphate synthase;2.6.99.2: pyridoxine 5-phosphate synthase; 1.4.3.5: pyridoxamine 5'-phosphate oxidase; 2.7.1.35: pyridoxal kinase.

Folic acid (Vitamin B_9) Folic acid is considered an essential growth factor for RTs (Cowperthwaite *et al.*, 1953; Kidder et Dutta, 1958) despite the prevailing difficulties in defining essential requirements in complex culture media. However, after using a defined growth medium, it was confirmed that folic acid is indeed an absolute requirement for the growth of regular trypanosomatids (Roitman *et al.*, 1972). Conversely, it was shown that in SHTs, such as *S. oncopelti* (Newton, 1956) and *A. desouzai* (Fiorini, 1989), growth occurs in total absence of folic acid. The standard pathway for the synthesis of folic acid is shown in Figure 2.16. Folates are composed of pterin, para-aminobenzoate (pABA), and L-glutamate moieties. Pterin is synthesized from GTP (guanosine 5'-triphosphate), whereas pABA is obtained from chorismate (Begley *et al.*, 1998).

The genomes of all the TPEs examined carry the genes for the conversion of GTP, pABA, and L-glutamate into folate and tetrahydrofolate (THF), except for the step that removes the triphosphate motif of 7,8-dihydroneopterin triphosphate to produce dihydroneopterin (Figure 2.16, II-III). This step was for long unknown. It was recently shown in bacteria and in plants that this reaction is performed by an enzyme from the Nudix family called FolQ (or NudB, dATP pyrophosphohydrolase EC:3.6.1.-) (Klaus *et al.*, 2005; Gabelli *et al.*, 2007). The corresponding gene is not assigned in any bacteria of the Alcaligenaceae family, and it is not possible, based only on a sequence similarity search, to find a candidate for the step.

Ca. B. cicadellinicola, an endosymbiont of the sharpshooter, lacks only this gene for folate synthesis (Wu *et al.*, 2006). On the other hand, Nudix proteins are found in TPEs as well as in the members of the Alcaligenaceae family. Alcaligenaceae bacteria have the other genes for the conversion of GTP, pABA, and L-glutamate into folate and tetrahydrofolate, but most *Bordetella* spp. lack the first step of this pathway (GTP cyclohydrolase EC:3.5.4.16, Figure 2.13). In KEGG, we also found an alkaline phosphatase (EC:3.1.3.1) of broad spectrum as an option for the missing step (Figure 2.16, II-III), which is present in SHTs and RTs but not in TPEs.

Further down in the THF biosynthetic pathway, the genes coding for the last two enzymes of the folic acid and THF synthesis, folylpolyglutamate synthase (EC:6.3.1.17) and dihydrofolate reductase (EC:1.5.1.3), are present in the TPE, SHT, and RT genomes. They are also present in *Trypanosoma* and *Leishmania* spp. (Figure 2.13), and the latter genus is able to salvage folate and unconjugated pteridines from their hosts (Vickers et Beverley, 2011).

As mentioned above, pABA has been described as a nutritional requirement for *S. oncopelti* (Newton, 1956, 1957). Conversely, this metabolite is absent in the minimal medium for C. fasciculata (Kidder et Dutta, 1958), however it is interesting to observe that folate is required in this case. Since pABA is an intermediate for folic acid biosynthesis, it makes perfect sense that the uptake of folate from the diet dispenses the need for pABA. Its synthesis from chorismate requires pabAB (aminodeoxychorismate synthase EC:2.6.1.85) and pabC (aminodeoxychorismate lyase EC:4.1.3.38) (Begley *et al.*, 1998). These enzymes were not identified in TPEs, SHTs or RTs. Those steps are found in the Alcaligenaceae bacteria except for *Taylorella* spp., and they are absent in *Leishmania* and *Trypanosoma* spp. The inability of SHTs and TPEs to produce pABA agrees with the described need for this metabolite in the minimal medium of *S. oncopelti*; in other words, TPEs would be able to synthesize folate provided pABA is available. This corroborates the fact that folic acid was considered a nutritional requirement for *A. deanei* when pABA was not supplied (Mundim *et al.*, 1974).

As a result, TPEs potentially have the enzymatic machinery for folate synthesis but probably require an exogenous source of pABA, corroborating the fact that SHTs are autotrophic for folic acid (Newton, 1956, 1957; de Menezes et Roitmanz, 1991; Fiorini, 1989).

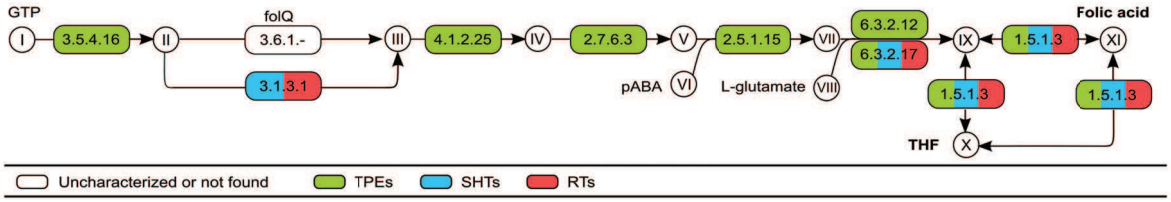


Figure 2.16: **Biosynthesis of folic acid.** Metabolites - I: Guanosine 5'-triphosphate; II: 7,8-Dihydroneopterin 3'-triphosphate; III: Dihydroneopterin; IV: 2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine; V: 2-Amino-7,8-dihydro-4-hydroxy-6-(diphosphooxymethyl)pteridine; VI: para-aminobenzoate; VII: Dihydropteroate; VIII: L-glutamate; IX: Dihydrofolate; X: Tetrahydrofolate; XI: Folic acid. Enzymes - 3.5.4.16: GTP cyclohydrolase I; 3.1.3.1: alkaline phosphatase; 3.6.1.-: Hydrolase acting on acid anhydrides in phosphorus-containing anhydrides; 4.2.1.25: dihydroneopterin aldolase; 2.7.6.3: 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase; 2.5.1.15: dihydropteroate synthase; 6.3.2.12: dihydrofolate synthase; 6.3.2.17: folylpolyglutamate synthase; 1.5.1.3: dihydrofolate reductase.

Auxotrophy of trypanosomatids for thiamine, nicotinic acid and biotin

Thiamine (Vitamin B_1) Thiamine is an essential growth factor for RTs, as well as for SHTs (Kidder et Dutta, 1958; Newton, 1956; Mundim *et al.*, 1974). Studies on the SHT requirement for thiamine were first performed with *S. oncopelti* (Newton, 1956) and only later were extended to *Angomonas* spp. (Mundim *et al.*, 1974). In both cases, thiamine was found to be an essential growth factor. This indicated that all or some of the genes for the biosynthesis of thiamine were missing from the genomes of both hosts and symbionts, which is in agreement with the genomic analysis performed in this work. Thiamine is particularly important for carbohydrate metabolism and its pathway involves the separate synthesis of thiazole and pyrimidine which are then coupled to form thiamine diphosphate (thiamine-PPi), which is the biologically active form of vitamin B_1 (Begley *et al.*, 1998).

Most genes related to the biosynthesis of thiamine are present only in the TPEs of *Angomonas* and totally absent from the genomes of RTs, SHTs, as well as from the TPEs from *Strigomonas* (Figure 2.17). However, even the symbionts of *Angomonas* lack the genes for key enzymes such as cysteine desulfurase (EC 4.1.99.17), thiamine biosynthesis protein ThiI (EC:2.8.1.4), and glycine oxidase (EC:1.4.3.19) that mediate the initial steps of any of the pathways leading to the synthesis of thiamine. Only the gene for cysteine desulfurase (EC:2.8.1.7) was found in all genomes including those of SHTs and RTs, but its presence may be related to its participation in the sulfur relay system. A few steps of the thiamine biosynthesis are missing in bacteria from the Alcaligenaceae family while the pathway is totally absent in *Leishmania* and *Trypanosoma* spp. (Figure 2.13). The genomic profile of RTs, SHTs, and TPEs is thus in perfect agreement with the absolute need of thiamine for the growth of trypanosomatids in general.

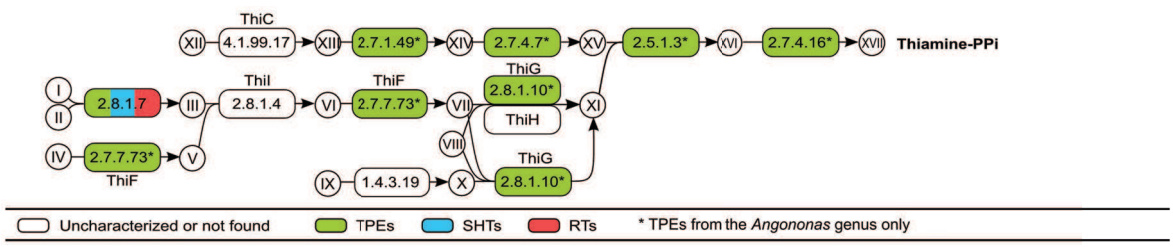


Figure 2.17: **Biosynthesis of thiamine** Metabolites - I: L-Cysteine; II: a [ThiI sulfur-carrier protein]-L-cysteine; III: a [ThiI sulfur-carrier protein]-S-sulfanylcysteine; IV: a ThiS sulfur carrier protein; V: a carboxy-adenylated-[ThiS sulfur-carrier protein]; VI: Thiamine biosynthesis intermediate 5; VII: a thiocarboxy-adenylated-[ThiS-protein]; VIII: L-Tyrosine; IX: Glycine; X: Iminoglycine; XI: 4-Methyl-5-(2-phosphoethyl)-thiazole; XII: 5'-Phosphoribosyl-5-aminoimidazole; XIII: 4-Amino-5-hydroxymethyl-2-methylpyrimidine; XIV: 4-Amino-2-methyl-5-phosphomethylpyrimidine; XV: 2-Methyl-4-amino-5-hydroxymethylpyrimidine diphosphate; XVI: Thiamine monophosphate; XVII: Thiamine diphosphate. Enzymes - 2.8.1.7: cysteine desulfurase; 2.7.7.73: sulfur carrier protein ThiS adenylyltransferase; 2.8.1.4: thiamine biosynthesis protein ThiI; 1.4.3.19: glycine oxidase; 2.8.1.10: thiamine biosynthesis ThiG; 4.1.99.19: thiamine biosynthesis ThiH; 4.1.99.17: thiamine biosynthesis protein ThiC; 2.7.1.49: hydroxymethylpyrimidine kinase; 2.7.4.7: phosphomethylpyrimidine kinase; 2.5.1.3: thiamine-phosphate pyrophosphorylase; 2.7.4.16: thiamine-monophosphate kinase.

Nicotinic acid (Vitamin B_3) Nicotinic acid is also essential for the growth of any kind of trypanosomatid, with or without endosymbionts (Cowperthwaite *et al.*, 1953; Kidder et Dutta, 1958; Newton, 1956; Mundim *et al.*, 1974; Mundim et Roitman, 1977; de Menezes et Roitmanz, 1991; Roitman *et al.*, 1972). The precursors for the de novo biosynthesis of nicotinamide adenine dinucleotide (NAD) are aspartate in prokaryotes and tryptophan in prokaryotes and eukaryotes (Begley *et al.*, 2001a; Kurnasov *et al.*, 2003). However, TPEs, SHTs, and RTs do not possess the enzymatic machinery for any of these processes. On the other hand, the genes responsible for the conversion of nicotinic acid into NAD+ and NADP+ are present in the genomes of all the TPEs, SHTs, and RTs examined (Figure 2.18). Interestingly, nicotinamidase (EC:3.5.1.19, Figure 2.18, XVI-XV), a key enzyme of this salvage pathway that catalyzes the conversion of nicotinamide to nicotinic acid, has been recently biochemically and functionally characterized in *L. infantum* (Gazanion *et al.*, 2011). Based on this sequence, we were able to identify candidates for this gene in the two RTs analyzed in the present study and in *Trypanosoma* and *Leishmania* spp., however not in SHTs or TPEs (Figure 2.13). This is in agreement with the fact that nicotinamide is frequently described in the minimal media of RTs (Kidder et Dutta, 1958), since it can be converted into nicotinic acid. There is also agreement that, once nicotinic acid is provided, all trypanosomatids are able to synthesize NAD, the essential coenzyme for the redox reactions of any living cell. As concerns the RT species, since they have the gene coding for nicotinamidase, they are also able to grow in culture medium containing only nicotinamide.

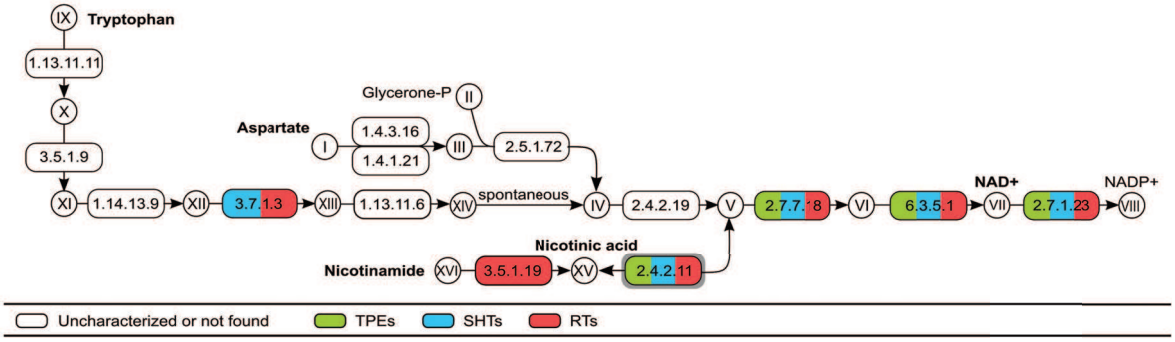


Figure 2.18: **Biosynthesis of nicotinic acid and NAD** Enzymes surrounded by a gray box were possibly acquired through horizontal transfer from Bacteria to trypanosomatids (see main text). Metabolites - I: Aspartate; II: Glycerone-phosphate; III: Iminoaspartate; IV: Quinolinate; V: Nicotinate D-ribonucleotide; VI: Deamino-NAD+; VII: Nicotinamide adenine dinucleotide; VIII: Nicotinamide adenine dinucleotide phosphate; IX: Tryptophan; X: L-Formylkynurenine; XI: L-Kynurenine; XII: 3-Hydroxy-L-kynurenine; XIII: 3-Hydroxyanthranilate; XIV: 2-Amino-3-carboxymuconate semialdehyde; XV: Nicotinic acid; XVI: Nicotinamide. Enzymes - 1.4.3.16: L-aspartate oxidase; 1.4.1.21: aspartate dehydrogenase; 2.5.1.72: quinolinate synthase; 2.4.2.19: nicotinate-nucleotide diphosphorylase; 2.7.7.18: ; 6.3.5.1: NAD+ synthase; 2.7.1.23: NAD+ kinase; 1.13.11.11: tryptophan 2,3-dioxygenase; 3.5.1.9: arylformamidase; 1.14.13.9: kynurenine 3-monooxygenase; 3.7.1.3 kynureninase; 1.13.11.6: 3-hydroxyanthranilate 3,4-dioxygenase; 2.4.2.11: nicotinate phosphoribosyltransferase (recently transferred to EC6.3.4.21); 3.5.1.19: nicotinamidase.

Biotin (Vitamin B_7) The need for biotin was demonstrated for RTs as well as for *A. deanei* (Cowperthwaite *et al.*, 1953; Kidder et Dutta, 1958). In the case of *S. oncopelti*, it

was described as a non-essential vitamin, although its growth rate doubled with the addition of biotin to the media (Newton, 1956).

Malonyl-CoA has been recently described as the precursor of the pimeloyl moiety of biotin in Escherichia coli by a modified fatty acid synthetic pathway (Lin *et al.*, 2010). The late steps of the biotin biosynthetic pathway (Figure 2.19, XI-XV) are responsible for forming the two rings in the structure of this coenzyme. The trypanosomatid genomes have a few genes of the upper part of the pathway, also identified in *Trypanosoma* and *Leishmania* spp. (Figure 2.19, Figure 2.13). On the other hand, the symbionts possess the genes for the first nine steps of the pathway starting from malonyl-CoA, but lack the remaining ones (Figure 2.19). Bacteria from the Alcaligenaceae family have most of the genes for the entire pathway (Figure 2.13).

This indicates that neither RTs nor SHTs are capable of biotin synthesis. The growth of *S. oncopelti* in the absence of exogenous biotin is thus puzzling, unless this protozoan synthesizes biotin via a distinct, unusual route. This would be the most probable alternative if the nutritional autotrophy of *S. oncopelti* is confirmed, which has not been the case so far.

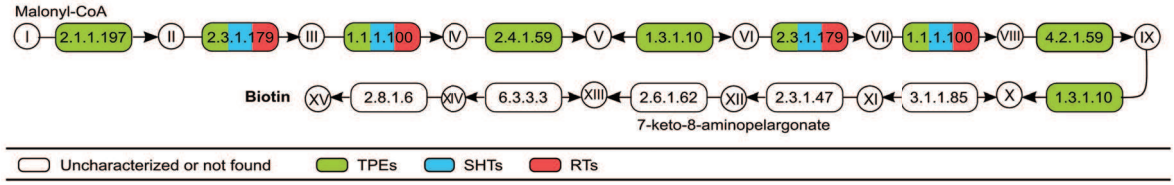


Figure 2.19: **Biosynthesis of biotin** Metabolites - I: malonyl-CoA; II: malonyl-CoA methyl ester; III: a 3-oxo-glutaryl-[acp] methyl ester; IV: a 3-hydroxyglutaryl-[acp] methyl ester; V: an enoylglutaryl-[acp] methyl ester; VI: a glutaryl-[acp] methyl ester; VII: a 3-oxo-pimelyl-[acp] methyl ester; VIII: a 3-hydroxypimelyl-[acp] methyl ester; IX: an enoylpimelyl-[acp] methyl ester; X: a pimelyl-[acp] methyl ester; XI: a pimelyl-[acp]; XII: 7-keto-8-aminopelargonate; XIII: 7,8-diaminopelargonate; XIV: dethiobiotin; XV: biotin. Enzymes - 2.1.1.197: malonyl-CoA methyltransferase; 2.3.1.180: β-ketoacyl-acyl carrier protein synthase III; 1.1.1.100: 3-oxo-acyl-[acyl-carrier-protein] reductase; 2.4.1.59: 3-hydroxy-acyl-[acyl-carrier-protein] dehydratase; 1.3.1.10: enoyl-[acyl-carrier-protein] reductase; 2.3.1.41: β-ketoacyl-ACP synthase I; 3.1.1.85: pimeloyl-[acp] methyl ester esterase; 2.3.1.47: 8-amino-7-oxononanoate synthase; 2.6.1.62: 7,8-diaminopelargonic acid synthase; 6.3.3.3: dethiobiotin synthetase; 2.8.1.6: biotin synthase.

Other cofactors Cofactors such as lipoic acid are produced by SHTs and RTs but not by TPEs. Conversely, the cobalamin (vitamin B_12) and menaquinone synthetic pathways are absent in all trypanosomatids and symbionts. Interestingly, the ubiquinone biosynthetic route is present in all RTs and SHTs as well as in the TPEs from the *Strigomonas* genus but absent in the TPEs from the *Angomonas* genus.

Ubiquinone functions as an electron carrier in membranes and is composed of a benzoquinone ring and an isoprene side chain which varies in the number of subunits in different organisms (Ranganathan et Mukkada, 1995). In *L. major*, the ubiquinone ring synthesis has been described as having either acetate (via chorismate as in prokaryotes) or aromatic amino acids (as in mammalian cells) as precursor (Ranganathan et Mukkada, 1995).

Most of the genes responsible for this biosynthetic pathway from tyrosine are present in SHTs and RTs, however the first steps of this route are still not well characterized in related species (Figure 2.20). As concerns the route from chorismate, the enzyme UbiC (EC:4.1.3.40),

which catalyzes the conversion of chorismate into 4-hydroxybenzoate, was identified only in the SHTs from the *Strigomonas* genus. In symbionts from this genus, we found this route with chorismate as precursor, however UbiC was not identified. Most bacteria from the Alcaligenaceae family have all the genes for ubiquinone production from chorismate while *Taylorella* spp. have some missing steps (Figure 2.13). The genes identified in all SHTs and RTs are also present in *Trypanosoma* and *Leishmania* spp. (Figure 2.13).

The presence of UbiC only in SHTs from the *Strigomonas* genus and of ubiquinone biosynthetic pathway only in their symbionts and not in other TPEs may indicate a higher production of ubiquinone in the *Strigomonas* host/symbiont system. As discussed in detail below in the Section *phylogenetic analyses*, the gene *ubi*C is closely related to those described in proteobacteria.

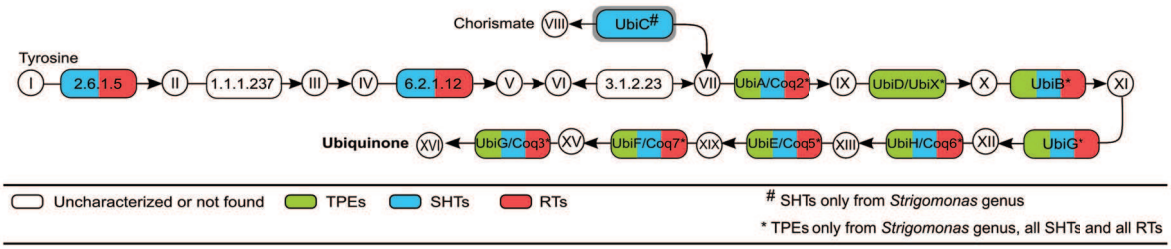


Figure 2.20: **Ubiquinone biosynthesis** Enzymes surrounded by a gray box were possibly acquired through horizontal transfer from Bacteria to trypanosomatids (see main text). Metabolites - I : L-Tyrosine; II: 4-Hydroxyphenylpyruvate; III: 4-Hydroxyphenyllactate; IV: 4-Coumarate; V: 4-Coumaroyl-CoA; VI: 4-Hydroxybenzoyl-CoA; VII: 4-Hydroxybenzoate; VIII: Chorismate; IX: 4-Hydroxy-3-polyprenylbenzoate; X: 2-Polyprenylphenol; XI: 2-Polyprenyl-6-hydroxyphenol; XII: 2-Polyprenyl-6-methoxyphenol; XIII: 2-Polyprenyl-6-methoxy-1,4-benzoquinone; XIV: 2-Polyprenyl-3-methyl-6-methoxy-1,4-benzoquinone; XV: 2-Polyprenyl-3-methyl-5-hydroxy-6-methoxy-1,4-benzoquinone; XVI: Ubiquinone. Enzymes - 2.6.1.5: tyrosine aminotransferase; 1.1.1.237: hydroxyphenylpyruvate reductase; 6.2.1.12: 4-coumarate–CoA ligase; 3.1.2.23: 4-hydroxybenzoyl-CoA thioesterase; UbiC: chorismate lyase; UbiA/Coq2: 4-hydroxybenzoate polyprenyltransferase; UbiD/UbiX: 3-octaprenyl-4-hydroxybenzoate carboxy-lyase; UbiB: ubiquinone biosynthesis protein; UbiG (EC:2.1.1.222): 2-polyprenyl-6-hydroxyphenyl methylase; UbiH/Coq6: 2-octaprenyl-6-methoxyphenol hydroxylase ; UbiE/Coq5: ubiquinone biosynthesis methyltransferase; UbiF/Coq7: 2-octaprenyl-3-methyl-6-methoxy-1,4-benzoquinol hydroxylase; UbiG/Coq3 (EC:2.1.1.64 / EC:2.1.1.114): 3-demethylubiquinol 3-O-methyltransferase / hexaprenyldihydroxybenzoate methyltransferase.

Phylogenetic analyses In trypanosomatids, most genes involved in the synthesis of vitamins are either of eukaryotic or of betaproteobacterial origin. In most cases, vitamin production benefits from the participation of the symbiotic bacterium whose genes are sister groups of the corresponding sequences described in *Bordetella* spp. and *Achromobacter* spp., both Betaproteobacteria that belong to the Alcaligenaceae family, as previously indicated for the heme biosynthesis genes (Alves *et al.*, 2011). As shown before in the whole genome analyses of these symbionts (Alves *et al.*, 2013b), the TPE genes involved in the synthesis of vitamins and the corresponding betaproteobacterial genes represent a monophyletic branch supported by bootstrap values close to 100 while they are distant from the equivalent genes in Alpha- and Gammaproteobacteria. The phylogenetic analyses of the trypanosomatid host genes were

carried out using the ML and NJ methods, which gave similar trees thus reinforcing the obtained results. Most genes were found to be of eukaryotic origin while three genes may have been transferred from bacterial groups to the trypanosomatid hosts (Table 2.21).

EC number	Enzyme name	Pathway	Nb Sequences	Nb Sites	# Distinct Alignment Patterns	Organisms	ML – Cluster / Sister group trypanosomatids
1.1.1.169	2-dehydropantoate 2-reductase	Pantothenate	607	1123	962	All SHTs and Herpetomonas, but not the other RTs.	They group within Firmicutes (BS=98).
2.4.2.11	nicotinate phosphoribosyltransferase	Nicotinate	630	1050	1010	All SHTs, RTs and TPEs	Trypanosomatid clade (BS=100) clusters within the Gammaproteobacteria (BS=93).
4.1.3.40	chorismate lyase	Ubiquinone	217	389	372	SHTs from Strigomonas genus	Strigomonas clade (BS=98) very similar to Pseudomonas, clusters within Gammaproteobacteria (BS=89), although this gene seems to diverge quite fast, making the identification of putative orthologs difficult.

EC number	NJ – Cluster / Sister group trypanosomatids	Cluster / Sister group TPEs	Figure	Average genomic coverage	Average contig coverage	Average gene coverage	Genome*
1.1.1.169	They group within Firmicutes and a few other groups of bacteria (BS=85).	-	9	23x	18x	18x	*A. deanei*
2.4.2.11	Trypanosomatid clade (BS=97) cluster within the Gammaproteobacteria (BS=91).	ML: Group with Alcaligenaceae (BS=90). NJ: Group with *Taylorella* and *Advenella* spp. (BS=99), and with the Alcaligenaceae (low BS).	10	24x	22x	27x	*A. desouzai*
4.1.3.40	Strigomonas clade (BS=97) clusters within Gammaproteobacteria (low BS).	-	11	28x	14x	14x	*S. galati*

Figure 2.21: **Summary of the phylogenetic and sequencing coverage analyses of the candidate HGT genes.** *Genome, contig, and gene average sequencing coverages were calculated for the organism indicated in the "Genome" column.

Possible horizontal gene transfer (HGT) from Firmicutes to trypanosomatids
The gene codifying for ketopantoate reductase (EC:1.1.1.169), involved in the synthesis of pantothenic acid (Figure 2.14), is present in the SHTs and in the genome of H. muscarum whereas it is absent in the TPE genomes. This gene is especially interesting due to the fact that all other steps for the synthesis of pantothenic acid are performed by enzymes coded by endosymbiont genes, including the enzymes necessary to synthesize the precursor α-ketoisovalerate. It is neither of proteobacterial nor of eukaryotic descent. With a high bootstrap support of 98 in the ML tree (85 in the NJ), its phylogeny indicates that it has been transferred to the SHTs and to *Herpetomonas* – or more probably to a common ancestor of these – from bacteria of the Firmicutes phylum (Figures 2.22 and 2.23).

It is interesting to note that the part of the phylogenetic tree containing the trypanosomatid gene, although mostly composed of Firmicutes, also includes genes of a few bacteria from other phyla interspersed amongst the Firmicutes genes. The phylum Firmicutes is divided in three major clades, with the clade containing the trypanosomatid genes separated from the other groups by a long branch (Figure 2.22). This could be due to different reasons: either the gene for EC:1.1.1.169 presents high evolutionary rates, leading to the long branch and low bootstrap values at deeper nodes of the tree and consequently to a difficulty in placing organisms in the tree; or there are multiple paralogs present in the tree due to ancient duplications. Our data do not permit to definitely distinguish between these two alternatives, although the much higher bootstrap support values at higher levels of the tree suggest the former.

Possible HGT from Gammaproteobacteria to trypanosomatids The gene for nicotinate phosphoribosyltransferase (EC:2.4.2.11), involved in the salvage pathway of nicotinic acid (Figure 2.18), is present in the SHT, RT, and TPE genomes. The trypanosomatids form a monophyletic group (bootstrap support of 100), and group within the Gammaproteobacteria with a high bootstrap support value of 93 in the ML tree (91 in the NJ, Figures 2.24 and 2.25). They are far from the other eukaryotes in the tree and overall form a monophyletic clade with moderate (66) and high (91) support values in the ML and NJ trees, respectively. The few other eukaryotes are placed within other bacterial groups; one such example concerns *Entamoeba* spp. placed within the *Bacteroidetes* (high support value of 95 and 81 in the ML and NJ trees, respectively). The TPEs group within the Alcaligenaceae family with high bootstrap support values of 90 and 99 (ML and NJ trees, respectively).

The gene for UbiC (EC:4.1.3.40), involved in the synthesis of ubiquinone (Figure 2.20), is present only in the SHTs of the *Strigomonas* genus, but is absent from the genome of any *Angomonas*, TPE, or RT genomes. The three *Strigomonas* form a monophyletic group, and are placed as the sister group of the genus *Pseudomonas* (Gammaproteobacteria) with a high bootstrap support value of 89 (Figures 2.26 and 2.27). The overall tree for UbiC contains almost only Beta- and Gammaproteobacteria, with a few Alphaproteobacteria of the *Bartonella* genus present within the Gammaproteobacteria.

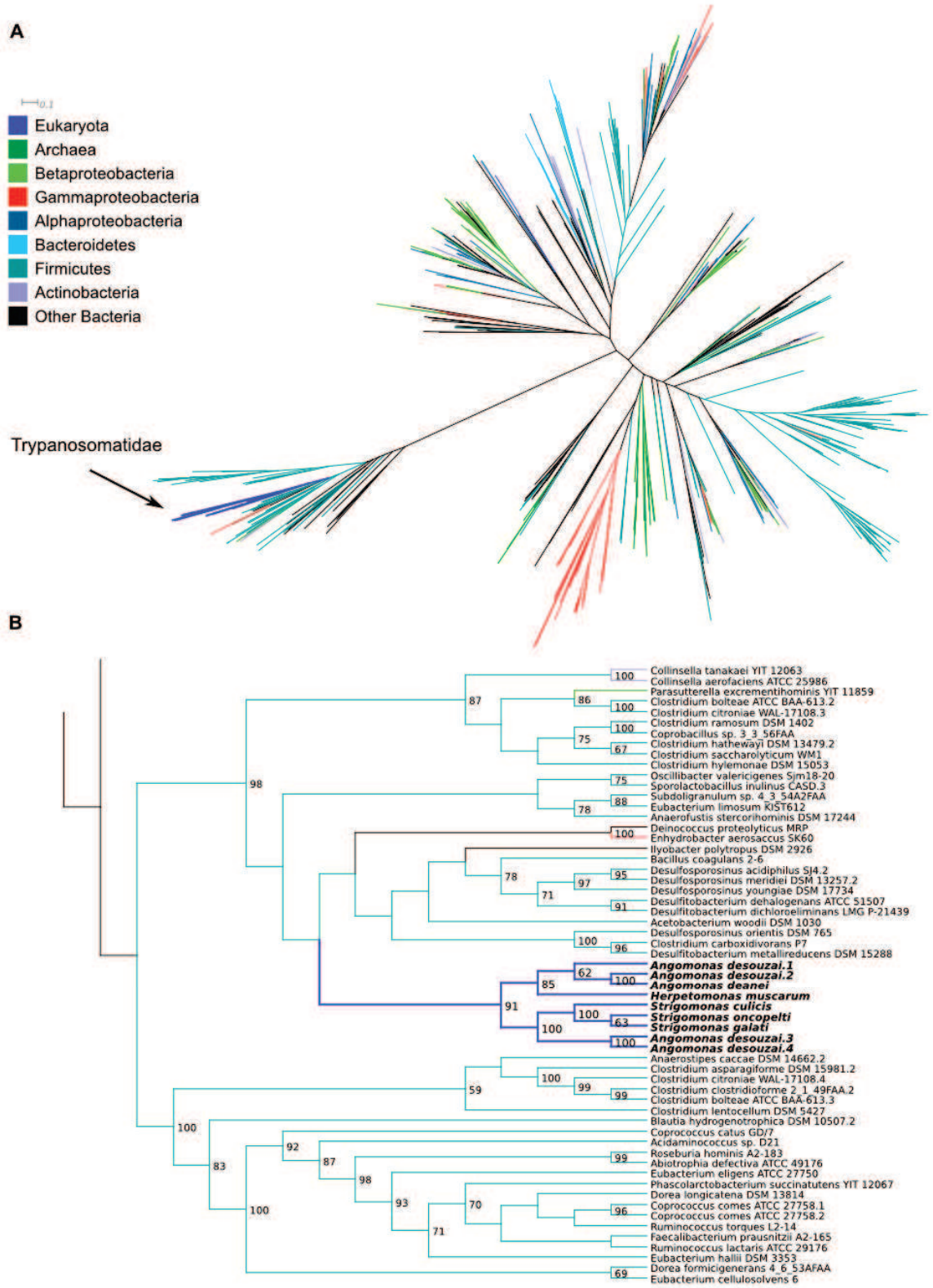


Figure 2.22: **Maximum likelihood phylogenetic tree of ketopantoate reductase (EC:1.1.1.169).** A - overall tree, colored according to taxonomic affiliation of each taxon, as per the legend on the left; distance bar only applies to panel A. B – details of the region of the tree where the Trypanosomatidae are placed. Values on nodes represent bootstrap support (only 50 or greater shown). Panel B is meant to only represent the branching patterns and do not portray estimated distances between sequences.

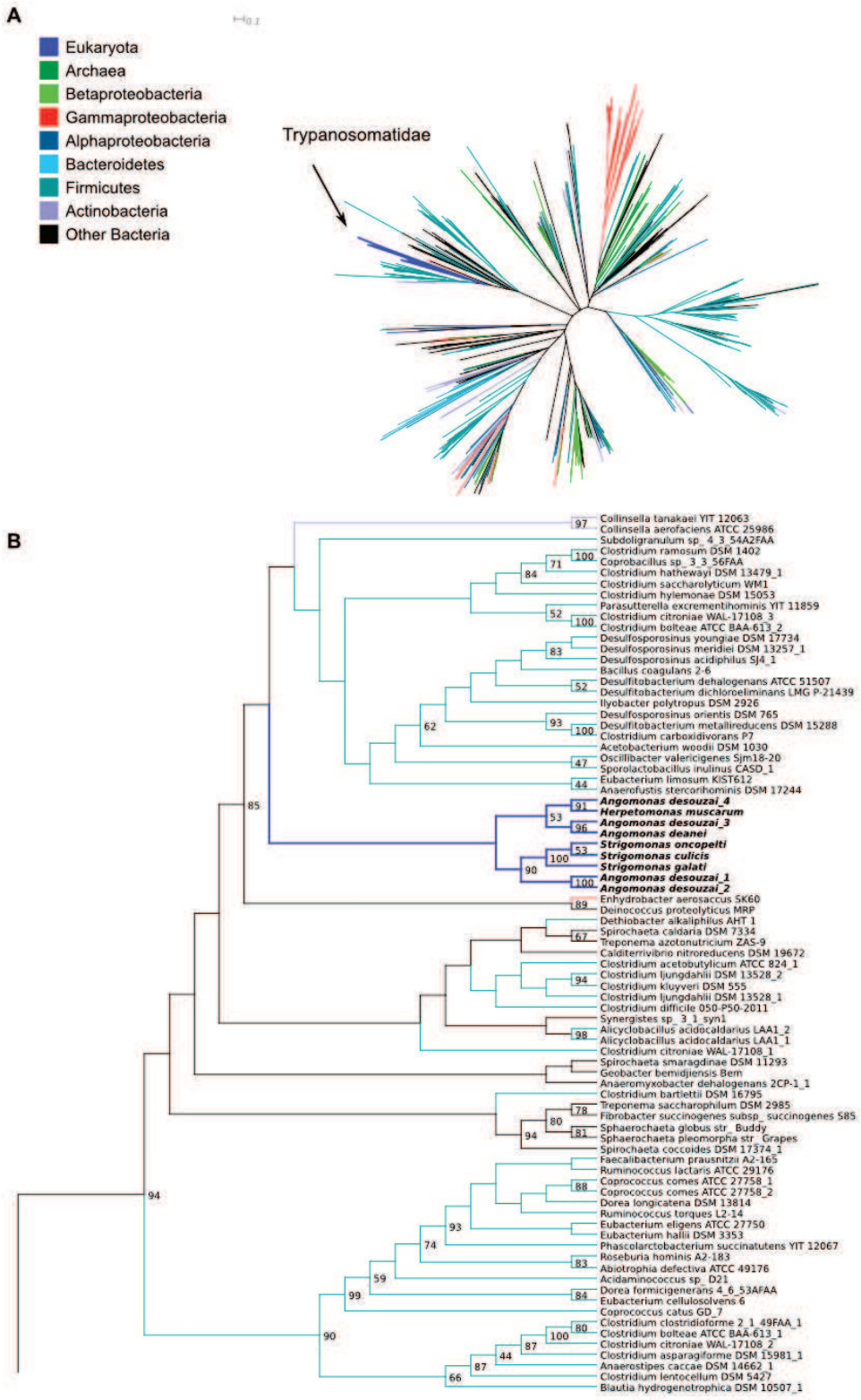


Figure 2.23: **Neighbor joining phylogenetic tree of ketopantoate reductase (EC:1.1.1.169)** A - overall tree, colored according to taxonomic affiliation of each taxon, as per the legend on the left; distance bar only applies to panel A. B – details of the region of the tree where the Trypanosomatidae are placed. Values on nodes represent bootstrap support (only 50 or greater shown). Panel B is meant to only represent the branching patterns and do not portray estimated distances between sequences.

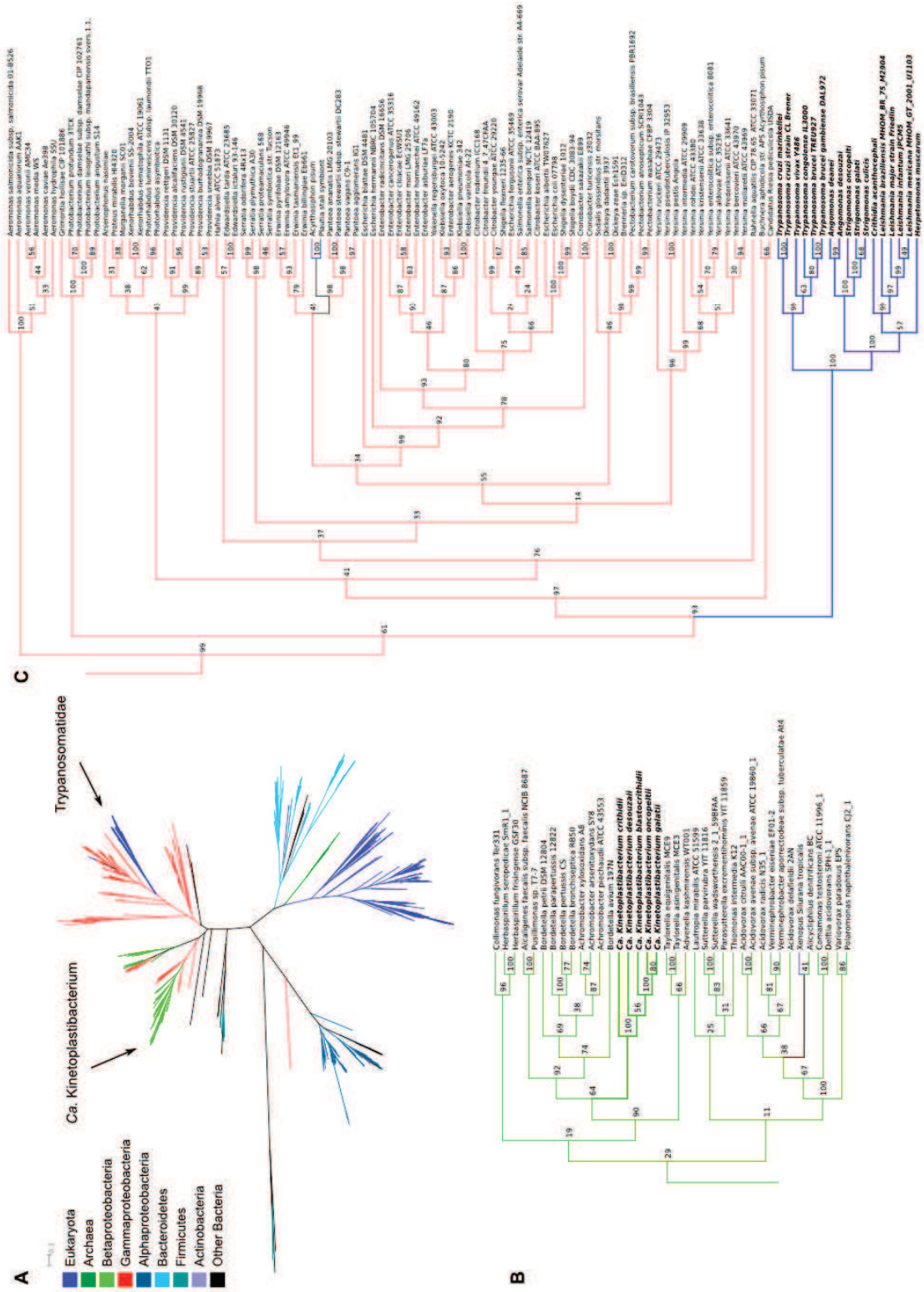


Figure 2.24: **Maximum likelihood phylogenetic tree of nicotinate phosphoribosyltransferase (EC:2.4.2.11).** A – overall tree, colored according to taxonomic affiliation of each taxon, as per the legend on the left; distance bar only applies to panel A. B – details of the region of the tree where the *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.

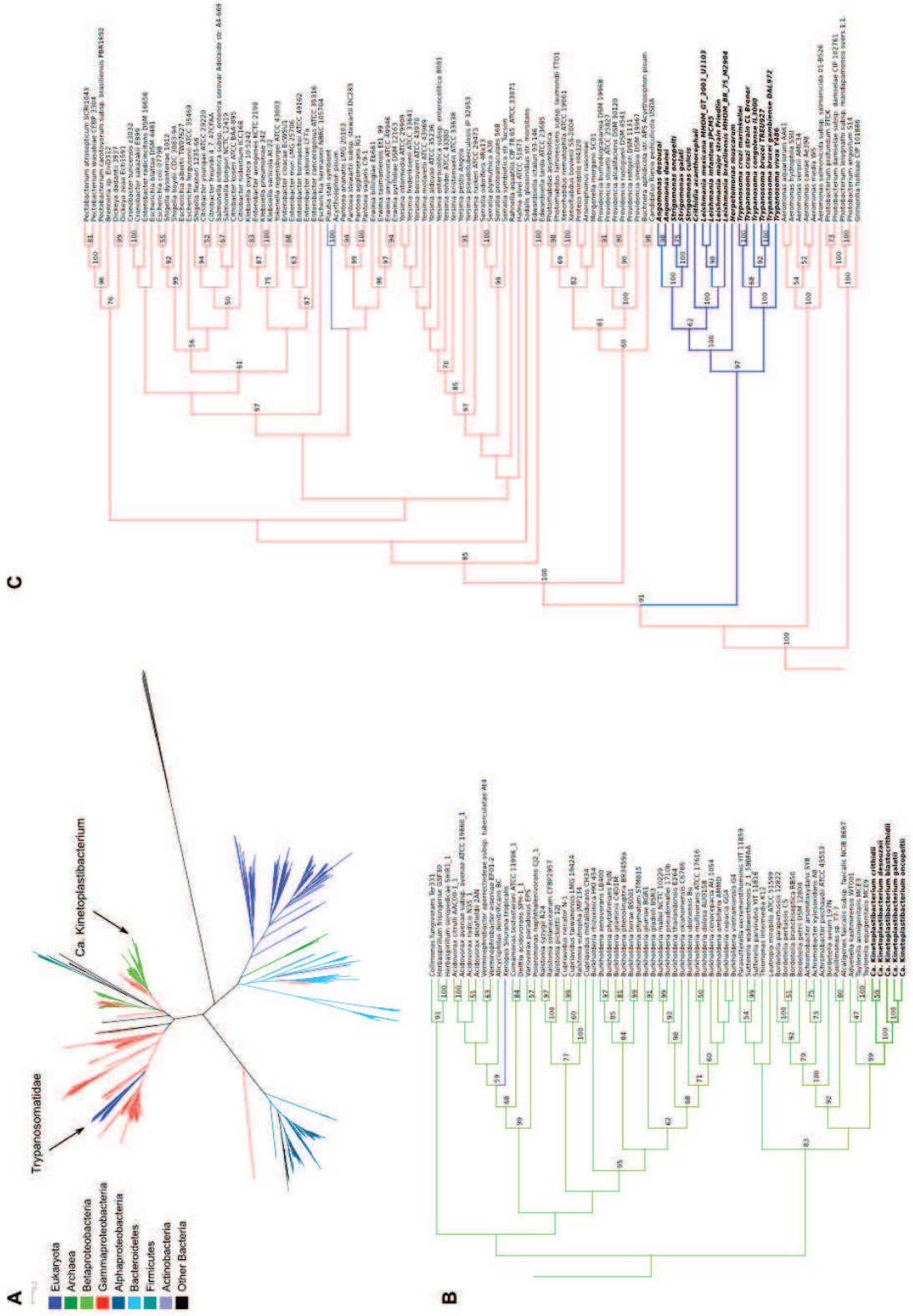


Figure 2.25: **Neighbor joining phylogenetic tree of nicotinate phosphoribosyltransferase (EC:2.4.2.11)** A – overall tree, colored according to taxonomic affiliation of each taxon, as per the legend on the left; distance bar only applies to panel A. B – details of the region of the tree where the Ca. 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.

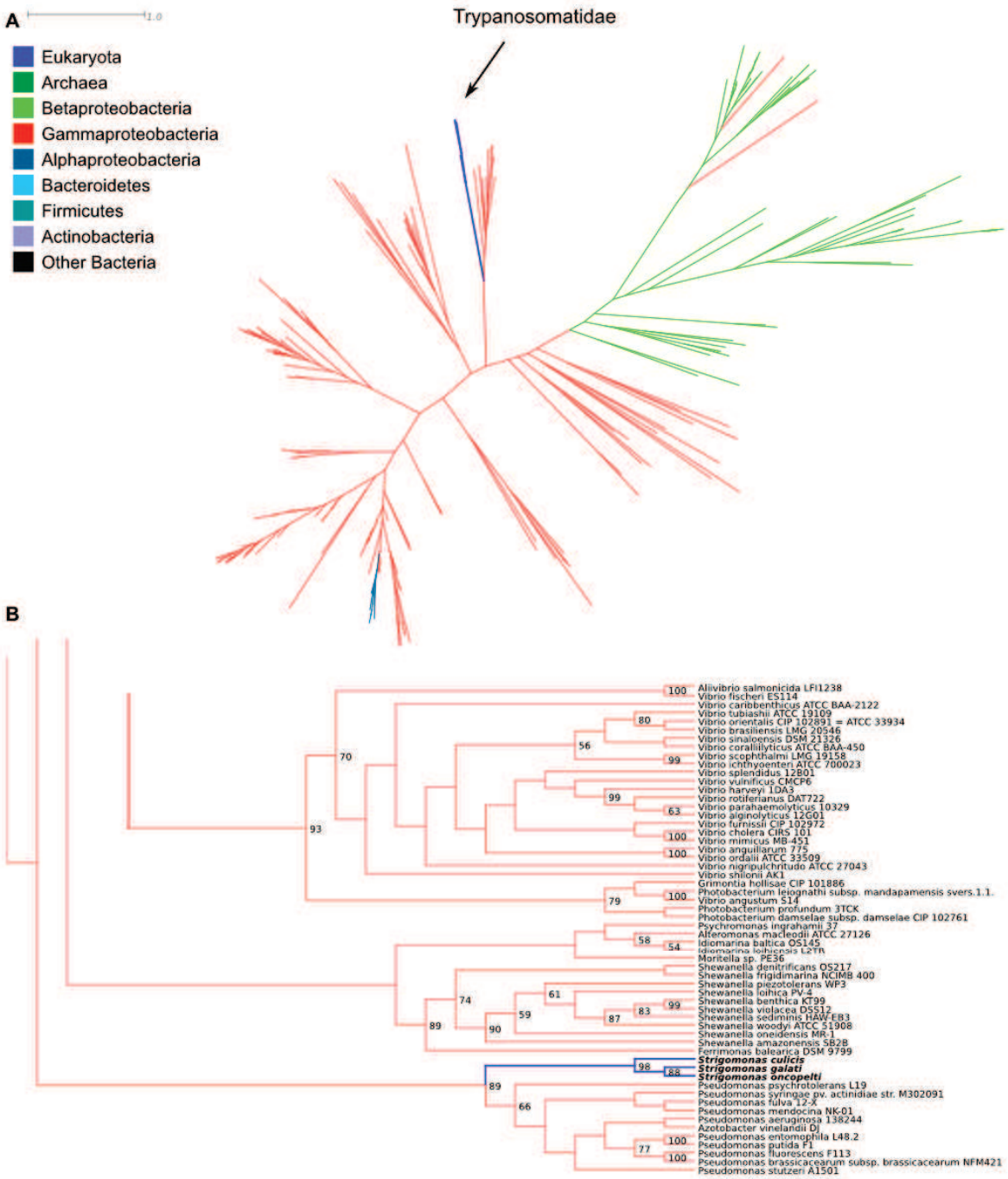


Figure 2.26: **Maximum likelihood phylogenetic tree of UbiC (EC:4.1.3.40).** A – overall tree, colored according to taxonomic affiliation of each taxon, as per the legend on the left; distance bar only applies to panel A. B – details of the region of the tree where the Trypanosomatidae are placed. Values on nodes represent bootstrap support (only 50 or greater shown). Panel B is meant to only represent the branching patterns and do not portray estimated distances between sequences.

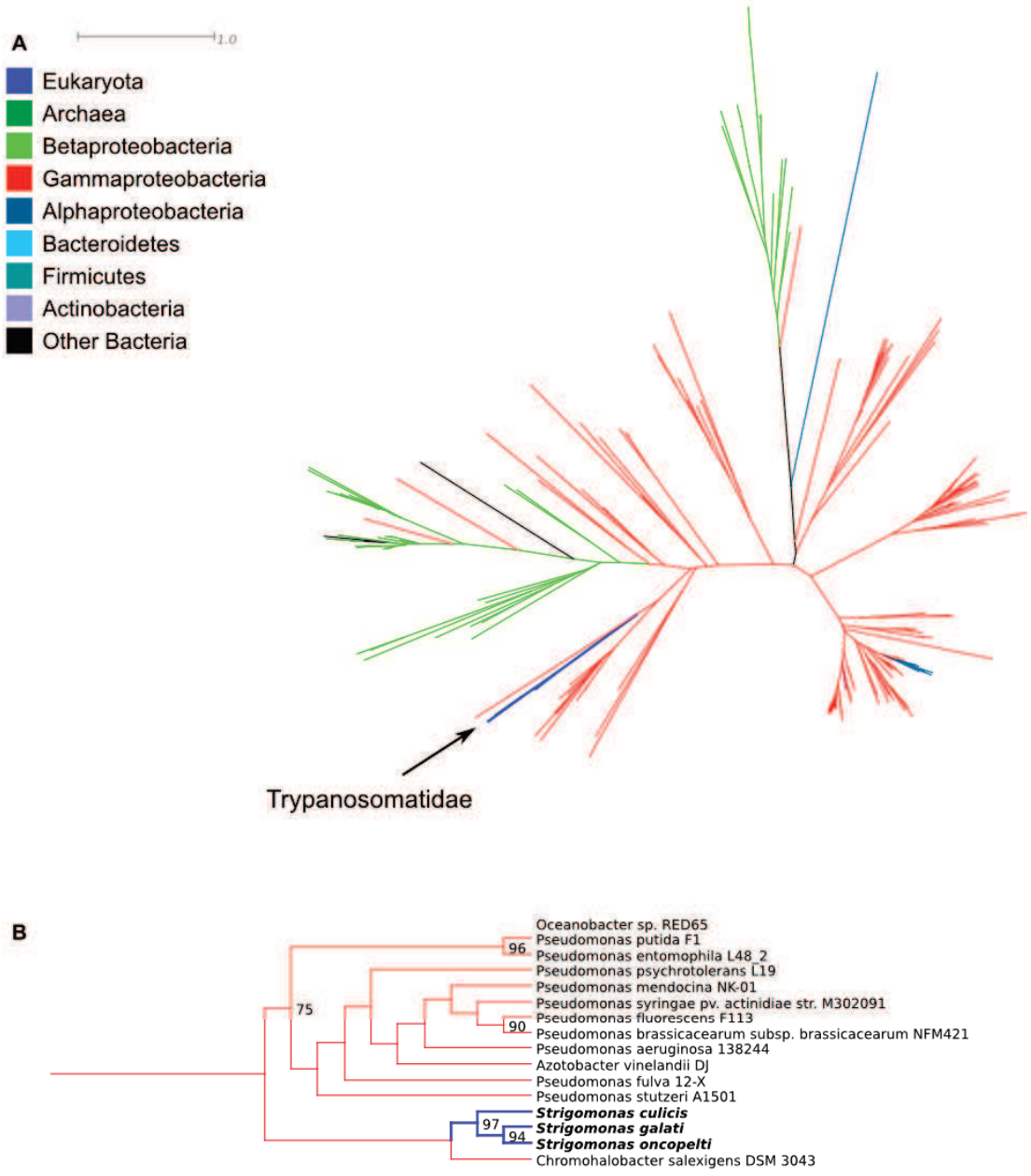


Figure 2.27: **Neighbor joining phylogenetic tree of UbiC (EC:4.1.3.40)** A – overall tree, colored according to taxonomic affiliation of each taxon, as per the legend on the left; distance bar only applies to panel A. B – details of the region of the tree where the Trypanosomatidae are placed. Values on nodes represent bootstrap support (only 50 or greater shown). Panel B is meant to only represent the branching patterns and do not portray estimated distances between sequences.

Genomic context and possible acquisition of HGTs These potentially transferred genes are mainly located in contigs presenting the typical trypanosomatid architecture of long stretches of genes in the same orientation (Figure 2.28). One such example is the upstream gene of ketopantoate reductase (EC:1.1.1.169) which is the one codifying for pyruvate kinase (EC:2.7.1.40), which is involved in the glycolytic pathway. This same genomic context was found in the previously sequenced strain of *A. deanei* (Motta *et al.*, 2013). In addition to that, the presence / absence of these three genes (codifying for EC:1.1.1.169, EC:2.4.2.11, EC:4.1.3.40) in the previously sequenced genomes of *A. deanei* and *S. culicis* are in agreement with the findings herein presented (Motta *et al.*, 2013). The GC percent (Figure 2.28) and sequencing coverage (Table 2.21) analyses also show that these genes present statistics typical of other genes from these organisms. The HGT genes analyzed show a codon usage consistent with that of about 125 other nuclear genes of the trypanosomatid based on the codon adaptation index and the correspondence analysis performed using a TPE gene as negative control (Figure 2.29).

The association of the betaproteobacterial symbionts and of the trypanosomatid hosts is very ancient, estimated to have occurred in the late Cretaceous (Du *et al.*, 1994a; Teixeira *et al.*, 2011) and to have perpetuated since by vertical transmission. No dating or any other kind of information is available about the acquisition of genes for vitamin synthesis of bacterial origin by trypanosomatids. It may be presumed, as is the case for similar instances of genes involved in the synthesis of heme or amino acids, that lateral gene transfer occurred in a common ancestor of several extant Trypanosomatidae clades, being subsequently lost in those where it was no longer necessary for the metabolism of the organism (Alves *et al.*, 2011, 2013a). Since the gene for the enzyme EC:1.1.1.169 is identified as being in a monophyletic group with the SHTs and *Herpetomonas* with a high bootstrap value of 91, this indicates that it was acquired by a common ancestor of these flagellates, and that other related genera and species might have been involved. However, the precise point in the tree of the family, or higher taxonomic category, where this gene was acquired remains obscure. Therefore, more studies are needed on the composition of the genes involved in the synthesis of vitamins, in more distantly related, non-parasitic Kinetoplastida, in order to try to elucidate this point of their genomic evolution.

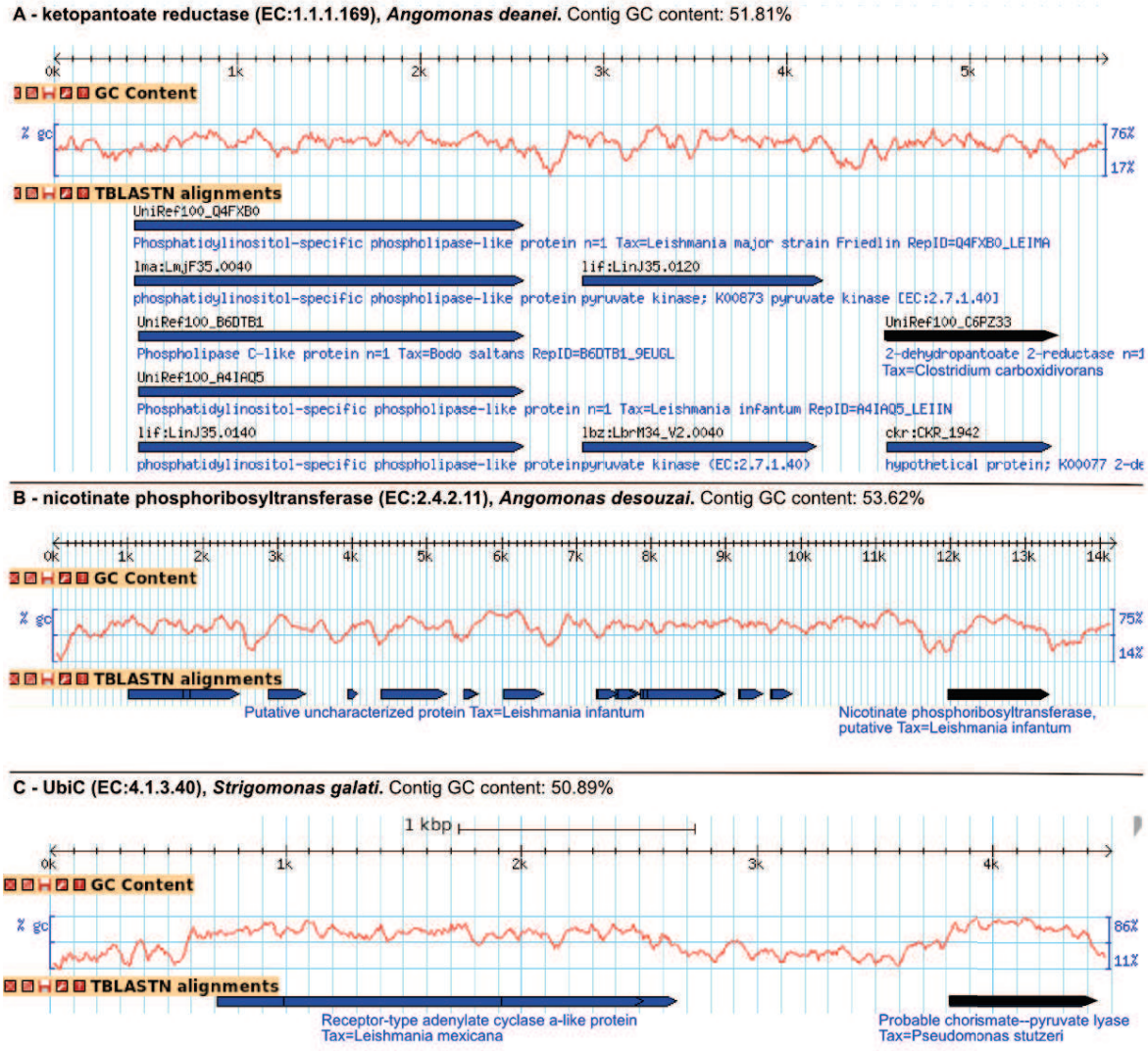


Figure 2.28: **Genomic context for candidate HGT genes in the Trypanosomatidae analyzed in this work.** Arrows show TBLASTN alignments of the genome against UniRef100 and KEGG proteins, as displayed by GBrowse and edited for clarity of presentation. The gene currently in focus is colored black. Coordinates are in kilobases.

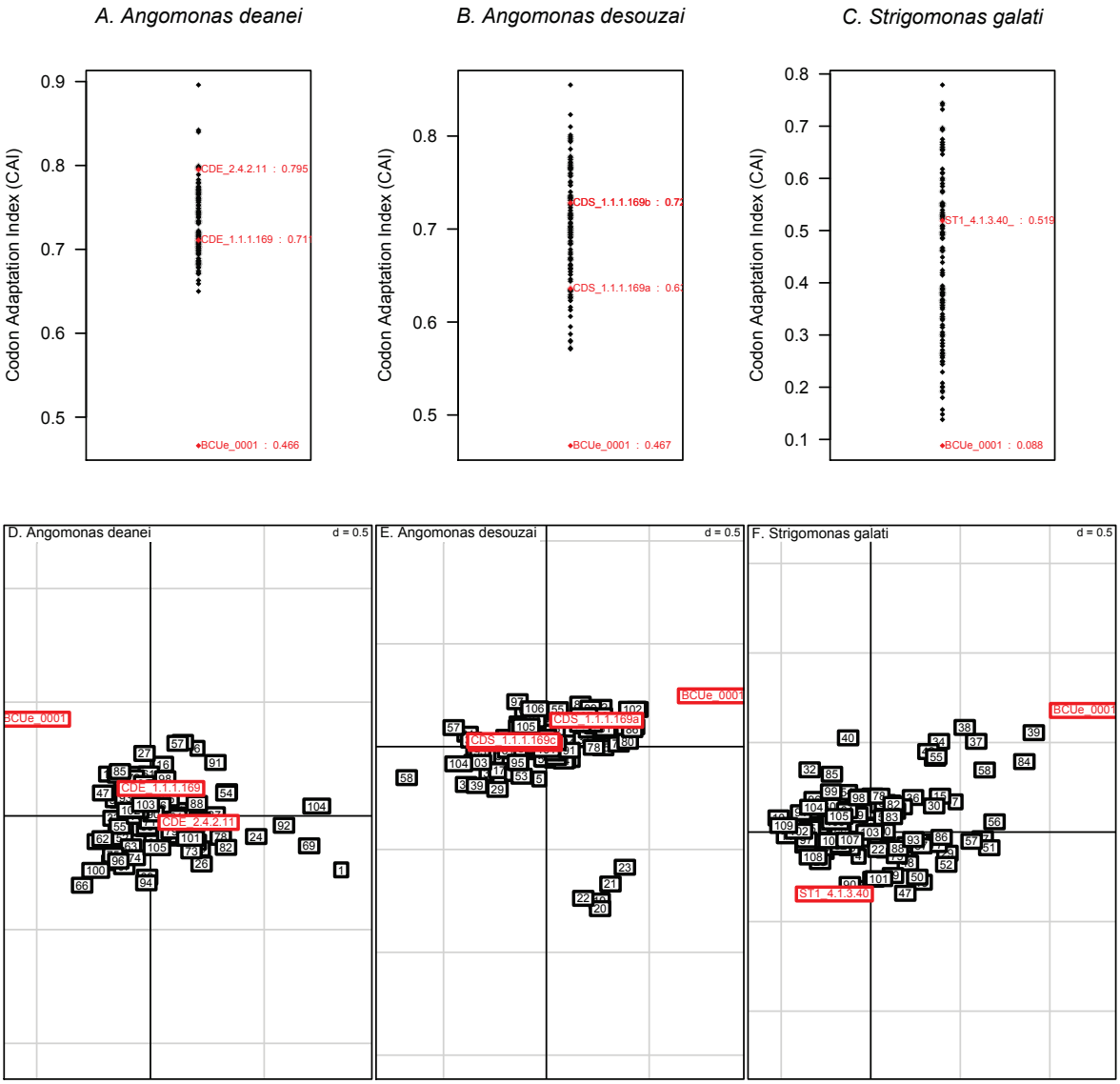


Figure 2.29: **Codon adaptation index and correspondence analysis of codon usage for candidate HGT genes.** Red: candidate HGT genes of the Trypanosomatidae analyzed in this work and the negative control which is the endosymbiont gene BCUe_0001. Codon adaptation index for *A. deanei* genes (A), for *A. desouzai* genes (B) and for *S. galati* (C). Correspondence analysis of codon usage for *A. deanei* genes (D), for *A. desouzai* genes (E) and for *S. galati* (F).

Conclusions

The results obtained in this work are in agreement with earlier nutritional studies (Cowperthwaite *et al.*, 1953; Kidder et Dutta, 1958; Mundim *et al.*, 1974; Mundim et Roitman, 1977), which indicated that trypanosomatids require seven vitamins in the culture media: folic and pantothenic acid, biotin, vitamin B_6, riboflavin, thiamine, and nicotinic acid (Figure 2.30). As shown in the present study, this is related to the fact that such protozoa lack the complete set of genes that codify for the enzymes involved in these essential biosynthetic pathways. However, this nutritional requirement does not apply to trypanosomatids carrying a cytoplasmic endosymbiont. SHTs have the necessary enzymes to produce most vitamins, with the exception of thiamine, biotin, and nicotinic acid, which represent absolute nutritional requirements for trypanosomatids in general. Most of the genes related to the synthesis of riboflavin, vitamin B_6, and folic acid were identified only in the symbiont genomes. This indicates the presence of complete biosynthetic routes in the TPEs with an the exchange of metabolites between host and bacterium in the extremities of the pathway, *i.e.* precursors and end products. On the other hand, the same is not observed in the synthesis of pantothenic acid, as suggested by our analyses. This pathway might have a more intricate participation of both partners in intermediate steps. SHTs and TPEs are able to perform the conversion of the vitamins riboflavin and pantothenic acid into the essential metabolites FAD and CoA, which indicates that possibly the symbiont enhances the production of these metabolites which may be controlled by the host in a way that is not yet fully elucidated.

According to the phylogenetic analyses, some genes coding for the enzymes involved in the biosynthetic and salvage pathways of vitamins and cofactors are in the host genome and are of eukaryotic origin, while most genes are localized in the genomes of the symbionts and are of betaproteobacterial ancestry. On the other hand, three genes were possibly transferred from bacteria to the trypanosomatid nuclei. Such is the case of the ketopantoate reductase gene (EC:1.1.1.169) involved in the de novo biosynthesis of pantothenic acid, which was probably transferred from a Firmicutes bacterium to an ancestor of the SHT host and of *H. muscarum*. The two other sequences may have been acquired from Gammaproteobacteria: nicotinate phosphoribosyltransferase (EC:2.4.2.11), which is involved in the salvage pathway of nicotic acid, and UbiC (EC:4.1.3.40), which is involved in the synthesis of ubiquinone.

Taken together, the nutritional data and our genomic analysis show that SHTs are autotrophic for riboflavin, pantothenic acid, vitamin B_6, and folic acid (Newton, 1956, 1957; Mundim *et al.*, 1974; de Menezes et Roitmanz, 1991). As a result, we can assume that the shared participation of the trypanosomatid host and of its symbiont in the synthesis of vitamins evidences an extensive metabolic exchange between both partners, at the extremities of the pathways or maybe even at intermediate steps, and that this exchange has an essential role in the maintenance of this mutualistic association.

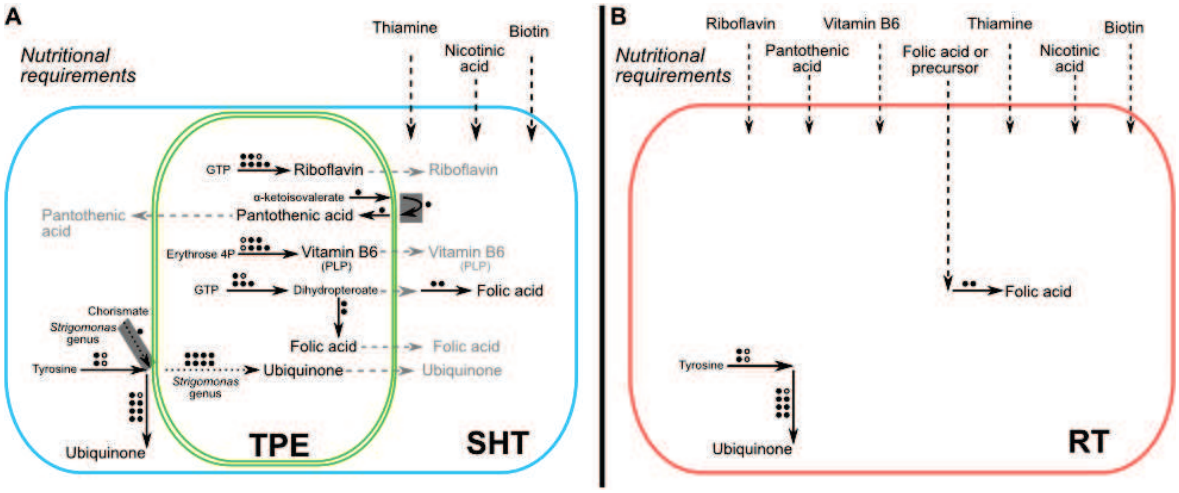


Figure 2.30: **Overview of the biosynthetic pathways of essential vitamins and co-factors in trypanosomatids.** Dashed arrows: metabolite import/exchange; dotted arrows: reaction present in only some of the organisms analyzed; solid arrows: other reactions (circles on the top of the arrows indicate number of steps and fulfilled circles indicate presence of enzyme); arrows surrounded by a gray box: enzymes possibly acquired through horizontal transfer from Bacteria to trypanosomatids (see main text). A - Contribution of SHTs and TPEs 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 based on the analysis of genomic data of *H. muscarum*, *C. acanthocephali* and *L. major*.

2.3 Metabolic networks of host and symbiont

2.3.1 Overview

This section introduces our ongoing analyses of the whole metabolic networks of *A. deanei* and its endosymbiont *Ca.* K. crithidii. Such analyses will be further extended to *S. culicis* and its endosymbiont. The aim is to have a global view of the exchanges between host and symbiont that goes beyond the scope offered by a study of pathways only. At first, the main focus is on the metabolic exchanges between the host mitochondrial, glycosomal and cytoplasmic metabolism and the symbiont. This choice is based on the previous findings of a physical proximity of the bacterium and the glycosomes as well as on the energetic metabolism since the symbiont possibly depends on the host supply of ATP (Motta *et al.*, 1997a; Faria-e Silva *et al.*, 2000; Motta *et al.*, 2010). The results presented hereafter are preliminary since they are mainly based on the early steps of a manual refinement of the reconstructed metabolic networks.

2.3.2 Reconstruction of the metabolic networks

The chosen method for the draft-level reconstruction was AUTOGRAPH (Notebaart *et al.*, 2006) (see detailed description in Section 1.3.2). The reasoning for that was the availability of a well-curated metabolic network of a phylogenetically close species, *Leishmania major* Friedlin (SBML model version iAC560; Chavali *et al.*, 2008), as well as the complexity of an eukaryotic network due to its compartments and size. For the symbiont, two reference networks were selected, one from the model species *Escherichia coli* k-12 MG1655 (SBML model version iJO1366; Orth *et al.*, 2011) and the other from the α-proteobacterium nitrogen fixing symbiont of leguminous plants, *Rhizobium etli* CFN42 (SBML model version iOR363; Resendis-Antonio *et al.*, 2007). Besides being also a mutualistic symbiont, the latter was chosen due to known similarities in its phospholipid metabolism with the endosymbiont of trypanosomatids, which are related to the interaction with eukaryotic cells (for further description see Section 1.2.3; Palmié-Peixoto *et al.*, 2006; de Azevedo-Martins *et al.*, 2007; Aktas *et al.*, 2010).

The reconstruction workflow based on the AUTOGRAPH method is recalled below in summarised form:

1. Choose one or more well-curated metabolic models for propagation.
2. Search for orthologs between the model organism and the organism of interest.
3. Manually check of ortholog pairs if the protein name or EC number does not match in the annotation of both organisms.
4. Propagate the model network based on the ortholog pairs.
5. Merge the networks propagated from different references.
6. Model the network as a compound graph and filter it removing cofactors and ubiquitous compounds to analyse the topological inputs of the network.
7. Validate the propagated reactions:
    - Reactions with no gene assigned: define them as mainly transport reactions.

- Enzymes composed of more than one subunit (*i.e.* needing more than one gene): Check whether all subunits were identified.

8. Annotate the genes of the organism of interest not used in the propagation.
9. Refine, in an iterative pathway-by-pathway manner, the reconstruction and functional analyses.

Steps 7-9 are still ongoing and the results presented here are preliminary and will be further refined.

The search for orthologs was performed using a LIPM version of InParanoid (Ostlund *et al.*, 2010) developed by L. Cottret (unpublished). The multi-fasta of proteins of *L. major* was obtained from Tritryp (in november/2012) whereas the remaining ones were obtained from GenBank NCBI. The propagation step was based on an algorithm developed by L. Cottret (unpublished). Network modelling and filtering were performed using MetExplore (Cottret *et al.*, 2010). The topological analysis was based on an implemented version of the Borenstein method (Borenstein *et al.*, 2008) using the Igraph package (Csardi et Nepusz, 2006), in order to identify which metabolites are potentially acquired from the environment (*i.e.*, are potential inputs). Cytoscape (Shannon *et al.*, 2003) was used for visualising the metabolic networks.

2.3.3 Metabolic network of *A. deanei*

Search for orthologs and propagation

From 7912 and 8412 protein sequences of *A. deanei* and of *L. major* respectively, 2690 groups of orthologs (which correspond to 4643 pairs of orthologs) were found. After filtering these pairs to only keep the proteins that are observed in the metabolic model of *L. major*, we ended up with a total of 690 pairs of orthologs after manual curation.

During the propagation step, the number of reactions evolved as follows. From the 1112 reactions present in the original SBML model of *L. major*, 1073 reactions were propagated, based on the orthologous pairs of proteins, to the new draft reconstruction of the metabolic network of *A. deanei*. Based on the boolean system of the Gene-Protein-Reaction (GPR) associations (for further information on this topic see Section 1.3.2 and Figure 1.8), these reactions can be divided into three categories depending on the confidence level accorded to their propagation:

1. The reaction was propagated with complete GPR, *i.e.* all the genes required to synthesize the enzyme that catalyses this reaction were found.
2. The reaction was propagated with incomplete GPR, *i.e.* at least one gene required to synthesize the protein complex that catalyses this reaction was missing.
3. The reaction was propagated due to an absence of GPR in the original SBML model, *i.e.* it was included during the refinement process of the reference network for which no gene was identified. Common examples are transport reactions and gaps in biochemical knowledge (e.g., nutritional data indicate that the organism is able to synthesise riboflavin, however the hydrolase that catalyses one step of this pathway has never been characterised in any species).

In the case of the propagation of our protozoan host, about 58% of the reactions were propagated due to a complete or incomplete GPR (481 and 143, respectively). The remaining 449 reactions were propagated because of no GPR.

General information

The overall characteristics of the propagated network can be found in Table 2.1. It has 8 compartments: acidocalcisome, cytosol, extracellular space, mitochondria, flagellum, glycosome, endoplasmic reticulum and nucleus. The overall distribution of the reactions in the different metabolic pathways are described in Table 2.2. The routes with more reactions assigned are those involved in the purine and pyrimidine metabolism, fatty acid biosynthesis and degradation, glycerophospholipid metabolism and steroid biosynthesis.

There is a high number of propagated reactions due to a lack of GPR and about 84% of them are associated to transport or unassigned pathways. Most of the remaining 16% are involved in fatty acid and steroid biosynthesis. Among the 39 reactions that were not propagated from *L. major*, most are involved in the metabolism of amino acids.

Number of compounds	1152
Number of reactions	1073
Number of reversible reactions	622
Number of irreversible reactions	451
Number of reactions with no gene assigned	449
Number of genes	425
Number of reactions with no pathway	65
Number of pathways	68

Table 2.1: **General information on the reconstructed network of *A. deanei***

Pathway	Number of reactions
Arginine and Proline Metabolism	17
Citrate Cycle (TCA)	15
Fatty Acid Biosynthesis	51
Fatty Acid Degradation	35
Fatty Acid Synthesis	21
Galactose metabolism	11
Glutamate Metabolism	18
Glycerophospholipid metabolism	47
Glycolysis/Gluconeogenesis	23
Inositol Phosphate metabolism	11
Methionine Metabolism	19
Oxidative phosphorylation	13
Pentose Phosphate Pathway	14
Purine Metabolism	71
Pyrimidine Metabolism	41
Sphingolipid Metabolism	17
Steroid Biosynthesis	37
Transport, Endoplasmic Reticular	39
Transport, Extracellular	62
Transport, Mitochondrial	133
Transport, Nuclear	33
Transport, Peroxisomal	70
Tryptophan Metabolism	10
Valine, leucine, and isoleucine degradation	13
Total number of reactions	821

Table 2.2: **Metabolic pathways with at least 10 reactions assigned in the network of *A. deanei*.**

2.3.4 Metabolic reconstruction of the endosymbiont

Search for orthologs and propagation

Ca.* K. crithidii and *E. coli From 734 and 4146 protein sequences of *Ca.* K. crithidii and *E. coli* respectively, 572 groups of orthologs (which correspond to 575 pairs of orthologs) were found. After filtering these pairs to only the proteins that appear in the metabolic model of *E. coli*, we ended up with a total of 245 pairs of orthologs after manual curation.

Ca.* K. crithidii and *R. etli From 734 and 5963 protein sequences of *Ca.* K. crithidii and *R. etli* respectively, 518 groups of orthologs (which correspond to 521 pairs of orthologs) were found. After filtering these pairs to only the proteins that appear in the metabolic model of *R. etli*, we ended up with a total of 121 pairs of orthologs after manual curation.

	E. coli	*R. etli*
Reactions in the original SBML file	2583	388
Reactions propagated	971	214
Reactions propagated with complete GPR	347	110
Reactions propagated with incomplete GPR	164	34
Reactions propagated because of no GPR	460	70

Table 2.3: **General information on the propagation process from the reference networks of *E. coli* and *R. etli*.**

Merged metabolic network *Ca.* K. crithidii

The overall characteristics of the merged network can be found in Table 2.4. It has 3 compartments: cytosol, periplasm and extracellular space. The overall distribution of the reactions in the different metabolic pathways are described in Table 2.5. The routes with more reactions assigned are those involved in the biosynthesis of cell envelope, cofactor and prosthetic group, and in the metabolism and transport of glycerophospholipid, purine and pyrimidine.

There is a high number of propagated reactions due to a lack of GPR and about 68% of them are associated to unassigned pathways.

Number of compounds	1153
Number of reactions	1070
Number of reversible reactions	530
Number of irreversible reactions	540
Number of reactions with no gene assigned	516
Number of genes	276
Number of reactions no pathway assigned	349
Number of pathways	69

Table 2.4: **General information on the reconstructed network of the symbiont of *A. deanei*.**

Pathway	Number of reactions
Alternate Carbon Metabolism	12
Arginine and Proline Metabolism	10
Cell Envelope Biosynthesis	84
Cofactor and Prosthetic Group Biosynthesis	83
Glycerophospholipid Metabolism	63
Histidine Metabolism	10
Inorganic Ion Transport and Metabolism	32
Lipopolysaccharide Biosynthesis / Recycling	20
Murein Biosynthesis	15
Murein Recycling	10
Nucleotide Salvage Pathway	60
Oxidative Phosphorylation	13
Purine and Pyrimidine Metabolism	38
Valine, Leucine, and Isoleucine Metabolism	18
Threonine and Lysine Metabolism	10
Transport, Inner Membrane	101
tRNA Charging	18
Tyrosine, Tryptophan, and Phenylalanine Metabolism	17
Total number of reactions	614

Table 2.5: **Metabolic pathways with at least 10 reactions assigned in the network of *Ca.* K. crithidii**.

2.3.5 Potential metabolic exchanges between the host and its symbiont

The aim of this section is to start discussing about the carbon sources in the protozoan host and its endosymbiont since we began the manual refinement of these networks by their central carbon metabolism. It is however important to note that new insights may change this picture as the whole network continues to be manually refined and analysed.

An overview of the central carbohydrate metabolism of the endosymbiont and its links with the synthesis of amino acids and vitamins can be found in Figure 2.31. As concerns glycolysis/gluconeogenesis, it is remarkable that it apparently works only in the gluconeogenetic direction, from pyruvate (PYR) to D-fructose-6-phosphate (F6P). This is due to irreversible steps that differentiate glycolysis and gluconeogenesis, such as the conversion of phosphoenolpyruvate (PEP) into pyruvate (PYR) and the conversion of fructose-6-phosphate (F6P) into D-fuctose-1,6-phosphate, for which only the genes coding for the enzymes catalysing reactions in the glucogenetic direction were found, *i.e.*, phosphoenolpyruvate synthetase and fructose-1,6-bisphosphatase, respectively. Moreover, the TCA cycle is incomplete. Comparing these results with those found for the endosymbionts of insects, *Buchnera*, *Ca.* Blochmannia and *Wigglesworthia*, the first two symbionts oxidize glucose to acetyl-CoA, while in the latter the pathway works in the opposite, glucogenetic direction (Zientz *et al.*, 2004). As concerns the TCA cycle, in *Buchnera* it is reduced to the step of 2-ketoglutarate (2KG) to succinyl-CoA (SCA) whereas most energy-yielding steps are conserved in the remaining two bacteria (Zientz *et al.*, 2004). The branching of the amino acids and vitamins are mainly from phosphoenolpyruvate (PEP), pyruvate (PYR), D-ribose-5-phosphate (R5P), D-erythrose-4-phosphate (E4P) and 2-ketoglutarate (2KG). Concerning the three branched-chain amino acids, leucine, isoleucine and valine, the amino acid transaminase responsible for the last step of their synthesis was found only in the host (for further information see Section 2.2.1). Two options are possible to explain this: either another enzyme in the symbiont performs this task or the immediate precursors of such amino acids are transported to the host and possibly further re-imported into the endosymbiont, for instance for protein synthesis.

As concerns the possible carbon sources supplied to the bacterial endosymbiont, we can start by analysing the environment of the host protozoan. *A. deanei* resides during all its life cycle in its insect host, possibly in the midgut. Similarly, the procyclic form of *T. brucei* lives in the midgut of the tsetse fly that is an environment where D-glucose and other sugars are usually scarce and the trypanosomatid thus relies mainly on L-proline that is the principal carbon and energy source available in the hemolymph of this fly (Coustou *et al.*, 2008). An overview of the metabolism of proline in the insect form of *T. brucei* is presented in Figure 2.32 (for reviews on this topic, refer to Bringaud *et al.* (2006, 2012)). The overall metabolic steps shown for this protozoan are also found in *A. deanei*. The most probable carbon sources for the endosymbiont are the metabolites found in the protozoan cytosol, and not the ones enclosed in the glycosomes, since less transport of compounds through the membrane of different cellular compartments is required. In that sense and considering the depleted glucose scenario (Figure 2.32A), the most prominent candidates would be malate, phosphoenolpyruvate (PEP) and pyruvate (PYR). Based on the overview in Figure 2.31, any of the three would be enough for the glucogenetic and pentose phosphate pathways and the downstream end products that branch out of those pathways. Disconnected from the previously mentioned metabolic routes and lacking a precursor in this preliminary view of the central carbon metabolism, there remains succinate, succinyl-CoA (SCA) and 2-ketoglutarate (2KG) for which enzymes catalysing reactions interconverting them were found (Figure 2.31). Directedly connected by a reversible reaction to the latter metabolite, there is glutamate

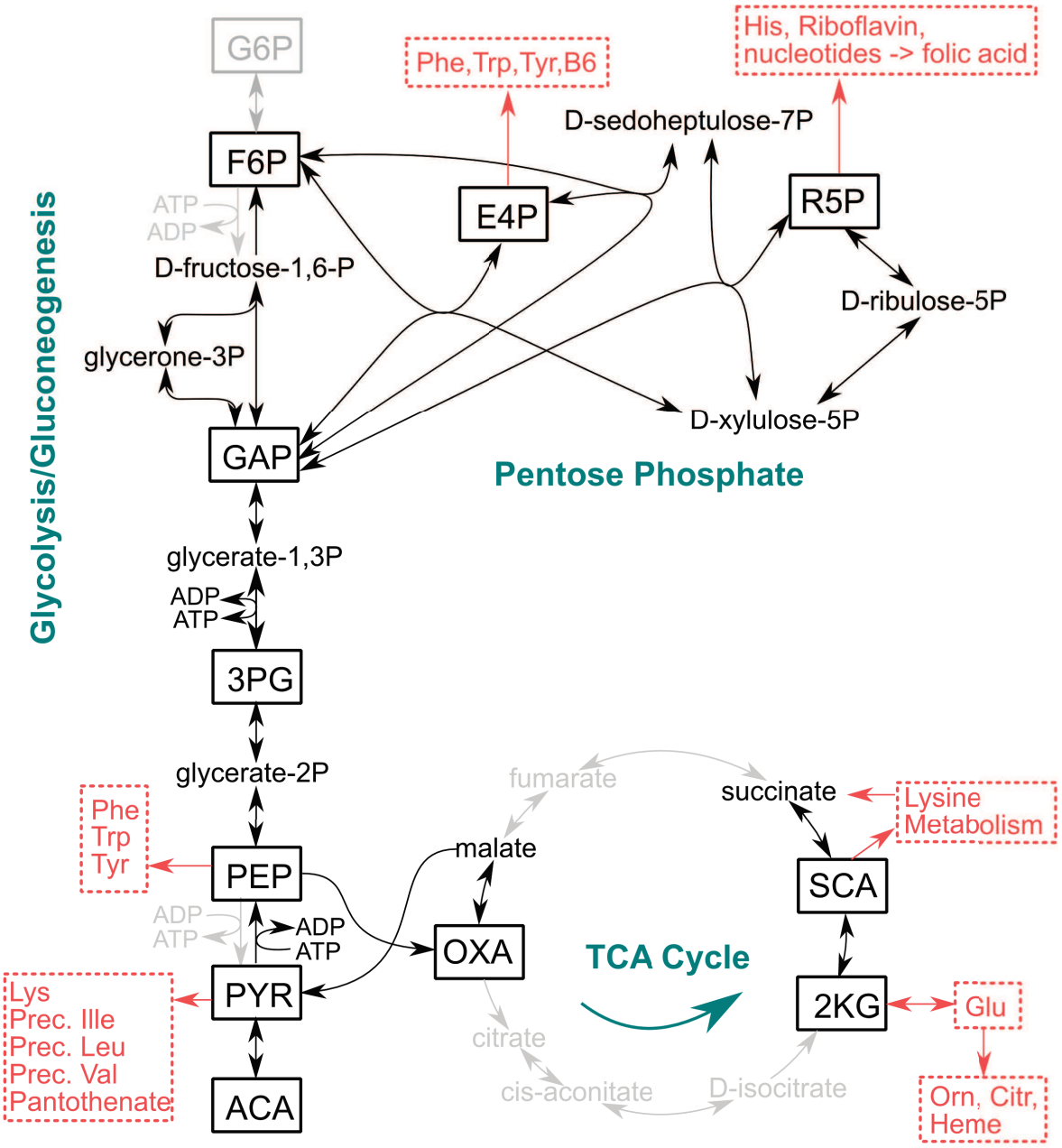


Figure 2.31: **The central carbohydrate metabolism and its links with the synthesis of amino acids and vitamins of the *Ca.* K. crithidii metabolic network.** Solid squares show 12 precursor metabolites, D-glucose-6-phosphate (G6P), D-fructose-6-phosphate (F6P), D-ribose-5-phosphate (R5P), D-erythrose-4-phosphate (E4P), D-glyceraldehyde-3-phosphate (GAP), glycerate-3P (3PG), phosphoenolpyruvate (PEP), pyruvate (PYR), acetyl-CoA (ACA), 2-ketoglutarate (2KG), succinyl-CoA (SCA) and oxaloacetate (OXA), and glycerate-1,3P (BPG), essential for a net gain of ATP in glycolysis in *E. coli*. Steps in grey were not found in the metabolic network of the endosymbiont of *A. deanei*. The red squares contain the downstream end products of the biosynthetic pathways of amino acids and vitamins that branch out from the central carbon metabolism. Abbreviations: phenylalanine (Phe), tryptophan (Trp), tyrosine (Tyr), lysine (Lys), isoleucine (Ille), leucine (Leu), valine (Val), vitamin B_6 (B6), histidine (His), glutamate (Glu), ornithine (Orn) and citrulline (Citr). **The structure of this figure is based on Figure 2 from** Noor *et al.* (2010).

which, for instance, might be a potential candidate precursor for the other ones since it is more probably found in the cytosol of the trypanosomatid as compared to the remaining ones that are possibly only found in the mitochondrion.

Thus, the endosymbiont might also have adapted to the depletion of glucose, maintaining only the gluconeogenetic pathway until the production of fructose-6-phosphate (F6P) which seems to be essential for the biosynthesis of key metabolites in the metabolic network of this bacterium.

2.3.6 Perspectives

As previously mentioned, the results presented above are preliminary and such analyses will continue in order to explore in more detail the metabolic exchanges of the host protozoan and its endosymbiont in the context of the entirely refined metabolic networks of these organisms. Additional experimental and transcriptomic data will be included as much as possible in this refinement step in order to have a well-curated metabolic model for predictions and simulations. The valuable interaction between computational and experimental analyses is possible through our collaboration with M.C.M. Motta and her team from Laboratório de Ultraestrutura Celular Hertha Meyer, Universidade Federal do Rio de Janeiro, Brazil. They have for long been studying symbiont-harbouring trypanosomatids, specially *A. deanei* and *S. culicis*, and this partnership plays a key role for the reconstruction of a well-curated metabolic model.

Figures 3 and 4 pages 357-8 from Bringaud *et al.* (2012).

Figure 2.32: **Schematic representation of the L-proline metabolism in procyclic *T. brucei* growing in glucose-depleted medium (A) and in glucose-rich medium (B).** Blue arrows represent enzymatic steps of the L-proline metabolism. While red arrows represent enzymatic steps of the glucose metabolism. **Extracted from** Bringaud *et al.* (2012). *(Caption continues on the next page.)*

Figure 2.32: *(Caption continued from the previous page.)* Excreted end products are in white characters on a blue background (major end products: L-alanine, L-glutamate and CO2) or in black characters on a light blue background (minor end products: acetate and succinate). At reversible steps, only the presumed or demonstrated direction of the reaction is represented. Dashed arrows indicate steps considered to occur at background level or not at all under glucose-depleted growth conditions. The glycosomal and mitochondrial compartments, the tricarboxylic acid cycle (TCA cycle) and gluconeogenesis are indicated. Abbreviations: C, cytochrome c; Cit, citrate; CoASH, coenzyme A; DHAP, dihydroxyacetone phosphate; G-6-P, glucose-6-phosphate; GLUT, glutamate; Gly-3-P, glycerol-3-phosphate; IsoCit, isocitrate; 2Ket, 2-ketoglutarate; Oxac, oxaloacetate; P5C, pyrroline-5-carboxylate; PEP, phosphoenolpyruvate; Pi, inorganic phosphate; PPi, inorganic pyrophosphate; g SAG, glutamate g -semialdehyde; SucCoA, succinyl-CoA; UQ, ubiquinone pool, 1,3BPGA, 1,3-bisphosphoglycerate; F-6-P, fructose-6-phosphate; FBP, fructose-1,6-bisphosphate; G-3-P, glyceraldehyde-3-phosphate; Gly-3-P, glycerol-3-phosphate; 3-PGA, 3-phosphoglycerate. Enzymes: 1, proline dehydrogenase (PRODH); 2, spontaneous reaction; 3, pyrroline-5 carboxylate dehydrogenase (P5CDH); 4, L-alanine aminotransferase (AAT); 5, glutamate dehydrogenase (GDH); 6, a -ketoglutarate dehydrogenase complex; 7, succinyl-CoA synthetase (SCoAS); 8, succinate dehydrogenase (SDH; complex II of the respiratory chain); 9, mitochondrial fumarase; 10, mitochondrial malate dehydrogenase; 11, citrate synthase; 12, aconitase; 13, NADP-dependent isocitrate dehydrogenase; 14, mitochondrial NADH-dependent fumarate reductase (FRDm); 15, mitochondrial malic enzyme (MEm); 16, cytosolic malic enzyme MEc); 17, glycosomal malate dehydrogenase; 18, phosphoenolpyruvate carboxykinase (PEPCK); 19, pyruvate phosphate dikinase (PPDK); 20, pyruvate kinase (PYK); 21, pyruvate dehydrogenase complex; 22, unknown enzyme; 23, acetate:succinate CoA-transferase (ASCT); 25, complex I of the respiratory chain; 26, rotenone-insensitive NADH dehydrogenase; 27, alternative oxidase (AOX); 28, complex III of the respiratory chain; 29, complex IV of the respiratory chain; 30, F0F1-ATP synthase (ATPe); 31, hexokinase: 32, glucose-6-phosphate isomerase; 33, phosphofructokinase; 34, aldolase; 35, triose-phosphate isomerase; 36, glyceraldehyde-3-phosphate dehydrogenase; 37, glycosomal phosphoglycerate kinase; 38, cytosolic phosphoglycerate kinase; 39, phosphoglycerate mutase; 40, enolase; 41, NADH-dependent glycerol-3-phosphate dehydrogenase; 42, FAD-dependent glycerol-3-phosphate dehydrogenase; 43, glycerol kinase; 44, cytosolic fumarase; 45, glycosomal NADH-dependent fumarate reductase (FRDg); 46, nonenzymatic reaction; 47, NADPH-dependent methylglyoxal reductase; 48, NAD+-dependent L-lactaldehyde dehydrogenase.

Chapter 3

Comparative analyses of metabolic networks

Contents

3.1 Exploration of the core metabolism of symbiotic bacteria 94
3.1.1 Background . . . 94
3.1.2 Methods . . . 95
3.1.3 Results . . . 100
3.1.4 Discussion . . . 110
3.1.5 Conclusions . . . 113
3.2 The extended core of metabolic networks . . . 115
3.2.1 Overview . . . 115
3.2.2 Dataset description . . . 115
3.2.3 Computation of the core/periphery reactions . . . 116
3.2.4 Results and discussion . . . 119
3.2.5 Conclusion and perspectives . . . 124

This chapter is dedicated to the comparative analyses of metabolic networks. Section 3.1 presents the exploration of common capabilities of symbiotic bacteria, of the contribution of each lifestyle group to the reduction of this core metabolism as well as the composition of this core in the different groups (Klein *et al.* 2012b, BMC Genomics 13:438). It is followed by Section 3.2 which introduces the ongoing investigation of an extended metabolic core in two datasets of bacteria from the *Escherichia* and *Pseudomonas* genera. The aim is to propose a methodology to compute an extended common set of metabolic capabilities more stable in size and content when compared to the traditional core, where only omnipresence determines this set which is quite unstable to the addition or removal of one organism.

3.1 Exploration of the core metabolism of symbiotic bacteria

3.1.1 Background

We now have at our disposal the full metabolic network based on genomic data for hundreds of species, mostly bacteria. The level of annotation is however widely heterogeneous across species, making it crucial for any comparative analysis to carefully choose a set of species for which we can guarantee a good enough annotation, and a same procedure for inferring the metabolic network from the annotated genomes.

One question commonly raised by the availability of many complete genome sequences is the number and content of a minimal set of protein-coding genes necessary to sustain a living cell (Mushegian et Koonin, 1996; Mushegian, 1999; Koonin, 2000), which has been investigated using experimental and computational approaches (Forsyth *et al.*, 2002; Koonin, 2003; Gerdes *et al.*, 2003; Klasson et Andersson, 2004; Gil *et al.*, 2004; Charlebois et Doolittle, 2004; Glass *et al.*, 2006; Gerdes *et al.*, 2006; Zhang et Zhang, 2008; Azuma et Ota, 2009; Juhas *et al.*, 2011; Gao et Zhang, 2011). One such method identifies essential genes based on those shared among genomes in a comparative analysis of diverse taxa (Mushegian et Koonin, 1996; Koonin, 2000; Klasson et Andersson, 2004; Gil *et al.*, 2004; Juhas *et al.*, 2011). Some studies included obligate host-dependent bacteria as a possibility for defining minimal gene sets in more specific and naturally occurring conditions (Klasson et Andersson, 2004; Gil *et al.*, 2004). The minimal gene sets proposed were not enriched in metabolic genes (Koonin, 2000; Klasson et Andersson, 2004; Gil *et al.*, 2004; Zhang et Zhang, 2008; Juhas *et al.*, 2011) and the corresponding pathways often presented missing steps (Koonin *et al.*, 1996; Gil *et al.*, 2004). These gaps may be due to non-orthologous gene displacement (NOGD) (*i.e.*, the presence of non-orthologous, paralogous or unrelated, genes for the same function in different organisms) (Koonin *et al.*, 1996) whose encoded enzymes have been defined as analogous (as opposed to homologous) and may be structurally unrelated (Galperin *et al.*, 1998). In comparative analyses of reaction sets instead of genes, NOGD has a reduced impact because different orthologous families encoding a single enzymatic capability are often represented by a same reaction. Another possible explanation for incomplete pathways is the use of different alternative routes, which recently have been defined as alternologs (*i.e.*, branches that proceed via different metabolites and converge to the same end product) (Hernández-Montes *et al.*, 2008). Their origin is closely related to different environmental metabolite sources and lifestyles among species (Hernández-Montes *et al.*, 2008). Since metabolism is a core function expected to be required for sustaining life (Danchin, 1989), and the core size may continue decreasing as more genome sequences appear (Charlebois et Doolittle, 2004; Danchin *et al.*, 2007), alternative approaches relaxing the requirement for ubiquity were proposed for analysing either prokaryotes (Charlebois et Doolittle, 2004; Gabaldón *et al.*, 2007; Azuma et Ota, 2009; Barve *et al.*, 2012) or species from the three domains of life (Danchin *et al.*, 2007; Peregrín-Alvarez *et al.*, 2009; Kim et Caetano-Anollés, 2010). One such example is the search for proteins commonly present (persistent) instead of strictly conserved everywhere (Danchin *et al.*, 2007). On the other hand, conserved portions of metabolism are found in lifestyle groups of bacteria (Koonin, 2000).

Small-scale comparative analyses of a selection of metabolic pathways were performed investigating each one individually (Zientz *et al.*, 2004; Tamas *et al.*, 2001) or grouped in one functional module (Nerima *et al.*, 2010). On the other hand, larger-scale comparative analysis were carried out in other papers but the question put in each case was different, related either to the proportion of metabolic genes in an organism, in absolute (van Nimwegen, 2003) or

classified according to lifestyle (Cases *et al.*, 2003; Merhej *et al.*, 2009), or related to the association between ecological strategies and growth rate (Freilich *et al.*, 2005). The notion of a core metabolism, meaning common elements, has been previously studied. However, this was done by comparing all known strains of a same species, namely, *Escherichia coli* (Vieira *et al.*, 2011). This approach of analysing metabolism as a single network allows a global view of functional processes, which was enabled by metabolic reconstruction methods based on genomic data (Reed *et al.*, 2006; Lacroix *et al.*, 2008; Durot *et al.*, 2009; Feist *et al.*, 2009; Cottret et Jourdan, 2010).

Here, we work at the level of whole metabolic networks for each organism and we analyse the core small molecule metabolism (*i.e.*, its conserved portion) of different lifestyle bacteria, aiming to characterise the contribution of each lifestyle group in the reduction of the common set of metabolic capabilities shared by the whole dataset. As concerns the impact of the obligate intracellular group, the question could be reformulated as the reactions which could not be dispensed and/or outsourced to the host in the course of genome compaction. Our major goals were to have a representative diversity in the symbiotic associations, a balanced amount of organisms in each lifestyle group, and as few biases as possible that might be related to the use of different annotation pipelines which is important when performing comparative analyses. We address this by comparing the presence of metabolic reactions as well as biochemical capabilities based on a partial Enzyme Commission (EC) number analysis at level 3 (*e.g.*, 2.5.1.-) (Webb, 1992). The purpose of the first is to be stringent although partially dealing with NOGD (see Methods for an example), while the purpose of the latter is to be more relaxed and to compare common functional capabilities in a broader sense. There are two possible advantages to this. One is to deal with enzymatic activities for which it was not possible to assign a full EC number during the functional annotation of a genome which resulted in partial EC numbers that do not denote a specific reaction. The second reason is to try to address the issue of alternologs, *e.g.*, alternative amino acid biosynthetic pathways that are often composed of enzymes which have the same partial EC numbers at level 3 (in the two alternative phenylalanine biosynthetic pathways, the partial EC numbers are ec:5.4.99, ec:4.2.1 and ec:2.6.1). We also analysed the metabolites that each bacterium potentially acquires from its environment in order to relate them to the set of common metabolic functions found for each lifestyle group.

3.1.2 Methods

Dataset

We selected 58 bacteria from the MicroScope platform (Vallenet *et al.*, 2009) and we carefully classified them according to their lifestyle based on the HAMAP information on interactions (Lima *et al.*, 2009) and the information provided in the literature. The broader lifestyle groups take into account the location of the bacterium in its host, constituting four groups: *obligate intracellular* INTRA, *cell associated* CA, *extracellular* EXTRA (16, 17 and 19 organisms, respectively) and the control group *free-living* FL. We further grouped them in subcategories on the basis of the association type and transmission mode. The lifestyle groups and the abbreviations are given in Figure 1. The full list of bacteria selected and their detailed classification is given in Additional file 2. The data on genes, metabolites and reactions were obtained from MicroCyc/MicroScope (Vallenet *et al.*, 2009). MicroCyc is a collection of microbial Pathway/Genome Databases which were generated using the PathoLogic module from the Pathway tools software (Karp *et al.*, 2010) which computes an initial set of pathways by comparing a genome annotation to the metabolic reference database MetaCyc (Caspi

et al., 2008). Using these databases as input, the metabolic networks of the 58 bacteria were obtained from MetExplore (Cottret *et al.*, 2010). It is important to notice that the completeness of metabolic network reconstructions is a current limitation as some reactions remain to be discovered and will be missing in the model while some false positive reactions may be wrongly included in the network. On the other hand, reactions shared by most bacteria are less likely to be missing in current datasets than organism-specific reactions, favouring the kind of analyses performed in the present work. The data on metabolic pathways were obtained from MetaCyc (Caspi *et al.*, 2008).

Core metabolism and core enzymatic function

Our analysis is restricted to the small molecule metabolism as defined in the MetaCyc/BioCyc databases (Caspi *et al.*, 2008; Karp *et al.*, 2005), *i.e.* small molecule reactions are those in which all participants are small molecules, hence reactions involving one or more macromolecules such as proteins or nucleic acids are not represented. The comparisons of compound and reaction sets are based on the BioCyc labels (Karp *et al.*, 2005), *e.g.*, the last reaction of glycolysis consists in the transformation of phosphoenolpyruvate, ADP and H^+ into pyruvate and water, and its label is PEPDEPHOS-RXN. The compounds found in the metabolic networks are those which are involved as substrates or products in the inferred reactions. All metabolites directly provided by the environment and not involved in any reaction as substrate are not included.

The presence of a metabolic core, *i.e.*, of a conserved set of elements in bacteria with different lifestyles, was analysed in terms of common compounds, common reactions and common partial EC number sets. The core metabolism was obtained by computing the intersection of the sets of reactions (resp. compounds and partial EC numbers) for each species. The panmetabolism was obtained by computing the union of these sets. The variable metabolism is the difference between pan- and core metabolism, *i.e.*, the set of elements that are missing from at least one bacterium. These definitions were introduced by Vieira *et al.* (Vieira *et al.*, 2011), however they worked with strains of a same species whereas here we compare different species.

The metabolic networks of the 58 bacteria were obtained from MetExplore (Cottret *et al.*, 2010). The macromolecules, as defined by BioCyc, were filtered out for all the analyses. The analyses were performed using R (R Development Core Team, 2009), as were the graphics. The IGRAPH package (Csardi et Nepusz, 2006) was adopted for analysing graphs.

For the analysis of the core enzymatic functions we used the EC number classification (Webb, 1992) which consists in a specific numerical identifier (*e.g.* 2.5.1.3) based on the chemical reactions a given enzyme catalyses. We worked with partial EC number sets at level 3 (*e.g.* 2.5.1.-), leaving the fourth digit open. The first digit represents which of the six main classes the enzyme belongs to (*e.g.* 1 for oxidoreductases; 2 for transferases). The following 3 digits provide a more detailed description of the enzymatic activity.

Exemplifying the issue of NOGD

We partially deal with NOGD because different orthologous families encoding one enzymatic capability often turn out to be a same reaction. One such example is an intermediate enzyme in the glycolytic pathway phosphoglycerate mutase (pgm), which has no sequence similarity between *Mycoplasma genitalium* and *Haemophilus influenzae* (Koonin *et al.*, 1996), and is represented by only one enzymatic reaction in our dataset. In fact, we have both analogous enzymes in our dataset: the cofactor-independent pgm (*e.g.*, *M. genitalium*, *Agrobacterium*,

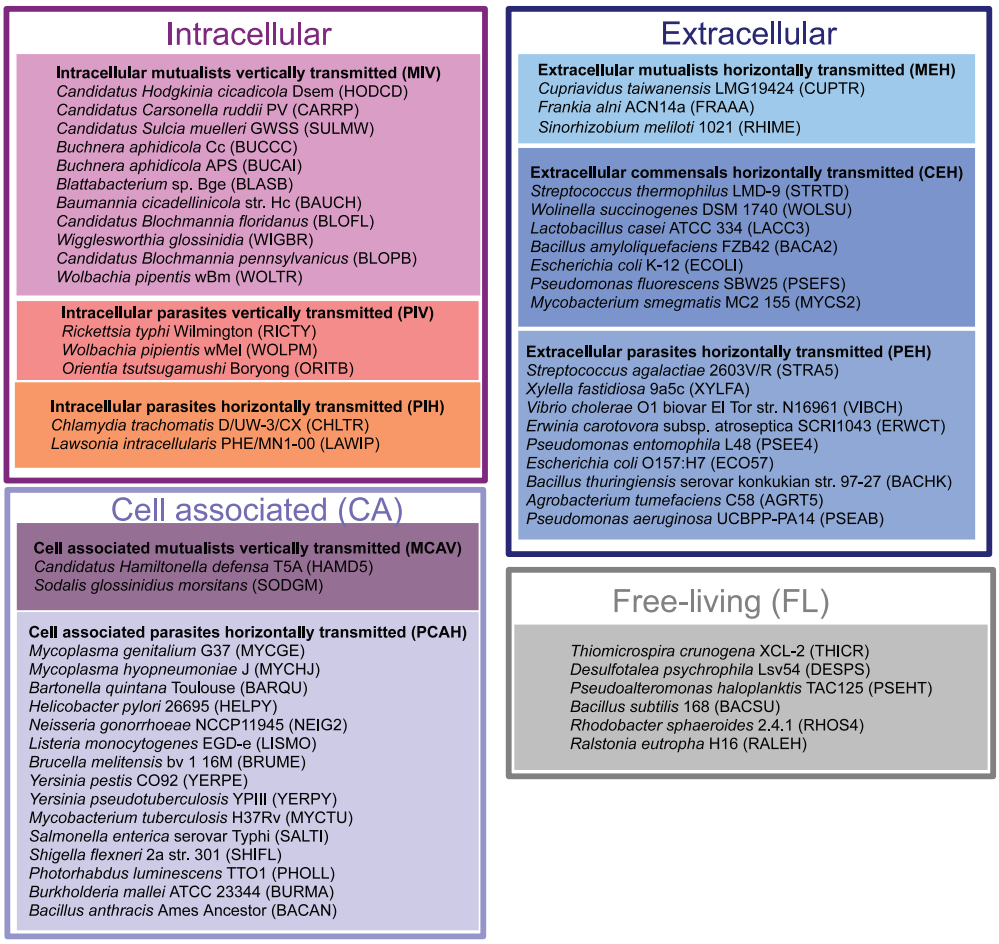


Figure 3.1: The lifestyle dataset consists of 58 bacteria. They were classified in 4 broader lifestyle groups based on the location of the bacterium in its host: *obligate intracellular* INTRA, *cell associated* CA, *extracellular* EXTRA (16, 17 and 19 organisms, respectively) and the control group FL. We further grouped them in subcategories on the basis of the association type (*mutualism* M, *commensalism* C, *parasitism* P) and the transmission mode (*vertical* V, *horizontal* H). INTRA includes only obligate intracellular bacteria, whereas CA includes bacteria which are facultatively intracellular, live on the surface of the host cell or are extracellular with a described intracellular step. These two groups have bacteria with obligate associations with their host (the totality of the individuals of the first group and 76% of the second one). EXTRA presents only bacteria that are facultatively associated with their host and are also free-living. The group FL has been used as a control group for all the analyses performed and includes organisms that have none of the three types of associations used for the grouping in the HAMAP information on interactions (Lima *et al.*, 2009). For that reason we included in FL one representative of each taxonomic class present in the dataset depending on the availability in the Microscope platform (Vallenet *et al.*, 2009). Two *Mycoplasma* species were classified in the CA group as they live on the surface of the host cell, although in another study they have been grouped together with the INTRA because of their reduced genome and the invariant environment within the hosts (Mendonça *et al.*, 2011). The codes for the organisms are the ones from HAMAP (Lima *et al.*, 2009).

Pseudomonas) and the cofactor-dependent pgm (*e.g.*, *Streptococcus*, *Bartonella*, *Buchnera*) based on the analysis of Foster *et al.* (Foster *et al.*, 2010). The reasoning for only one reaction representing both analogous enzymes is a many-to-many correspondence of enzyme-reaction, *i.e.*, one protein may catalyse more than one reaction, while the same reaction may be catalysed by more than one protein, and these multiple proteins catalysing a same reaction may or may not show sequence homology (Karp et Riley, 1993). Moreover, one should

be aware that some cases of NOGD do not result in only one reaction, such as thymidylate synthase which is folate-dependent (ec:2.1.1.45) in *M. genitalium* and in most bacteria, while it is flavin-dependent (ec:2.1.1.148) in *Actinobacteria*, *Rickettsia* and *Chlamydia* (Koonin, 2003). These enzymatic capabilities are classified with different EC numbers at level 4, however in our analysis of partial EC numbers at level 3, they are classified in the same way (2.1.1).

Connectivity in the reaction graph

We analysed the connectivity of the core metabolic network to check if the common reactions would be connected among themselves, *i.e.*, the produced metabolites would be consumed by other reactions in a chain of biochemical transformations. For that, the metabolic networks of the dataset were modelled as reaction graphs. In such a graph, nodes represent reactions, and arcs (*i.e.*, directed edges) between two reactions represent a compound which is produced by one reaction and consumed by the other. We set filters to exclude pairs of co-factors (*i.e.*, $ADP{+}P_i \rightarrow ATP$, $NAD^+{+}H^+ \rightarrow NADH$; for the full list see the MetExplore documentation) and current compounds (*e.g.*, water, proton, CO_2, phosphate, diphosphate, NH_3, H_2O_2 and O_2), which otherwise would connect unrelated reactions (Ma et Zeng, 2003).

Since we were working with the common reactions of a group of organisms, we computed the union graph of all the metabolic networks modelled as reaction graphs. We then calculated the graph induced by the common set of reactions, *i.e.*, the subgraph containing the nodes corresponding to these reactions as well as the arcs that link them. After that, we checked for the presence of connected components, *i.e.*, whether for every pair of nodes there is an undirected path.

In the case of the common partial EC number sets, we checked whether the reactions corresponding to each one of the partial EC numbers, *i.e.*, one reaction for each partial EC number, are connected in the metabolic networks. We analysed this in the union of all metabolic networks of the dataset (or of lifestyle groups) modelled as a reaction graph, as well as in the graph of each organism. This analysis was performed using MOTUS (Lacroix *et al.*, 2006).

Controlling for the impact of small networks

We controlled for the impact of bacteria with very reduced genomes on the size of the common set of reactions (resp. partial EC number sets). The six organisms which possess the smallest reaction sets (resp. partial EC number sets) were successively removed (*i.e.*, by forming subgroups from 57 to 53 bacteria) and the intersections of the remaining subgroups were recomputed. These six organisms are: "*Candidatus* Hodgkinia cicadicola" (HODCD), "*Candidatus* Carsonella ruddii" (CARRP), "*Candidatus* Sulcia mueller" GWSS (SULMW), *M. genitalium* (MYCGE), *Buchnera aphidicola* Cc (BUCCC) and *Mycoplasma hyopneumoniae* (MYCHJ). All possible orders for removing them were tested, and then the mean of the intersection sizes for each subset size of organisms was calculated. We also performed the same analysis by removing the eight bacteria with the smallest sets of reactions (resp. partial EC numbers).

Controlling for the structuring of MIV

In order to further explore the structure of MIV (Mutualistic Intracellular Vertically transmitted, see Figure 1 for abbreviations of group names) as they have the smallest genome sizes

of the dataset and they have specific symbiotic functions, we performed a multiple correspondence analysis (MCA) using the R (R Development Core Team, 2009) package ADE4 (Dray et Dufour, 2007). The input data was the contingency table of the presence and absence of reactions for each organism. We analysed the reactions with correlation ratio greater than 85% and 50% on the first and second axis (respectively) of the MCA.

Decay of the common reactions in the different lifestyle groups

Next, we checked whether there were reactions common to subsets of organisms within the same lifestyle group. To do so, for each lifestyle group l (l=*INTRA, CA, EXTRA*) having n_l organisms, we randomly drew x ($2 \leq x \leq n_l - 1$) organisms and computed the intersection (y) of their reaction sets. This was repeated 1000 times. In order to test if, when adding more species, the size of the intersection of reaction sets was expected to decrease to zero, we fitted exponential (E_l) and logistic (L_l) models to the data obtained for each lifestyle group l. Assuming normally distributed residuals, $\varepsilon \sim \mathcal{N}(0, \sigma)$, these models are given by:

$$E_l \quad : \quad \bar{y}_l = N_l * exp(-r_l * x_l) + \alpha_l + \varepsilon_l \tag{3.1}$$

$$L_l \quad : \quad \bar{y}_l = \frac{r_l}{(\frac{r_l}{N_l} - \frac{r_l}{\alpha_l}) * \exp(-r_l * x_l) + \frac{r_l}{\alpha_l}} + \varepsilon_l \tag{3.2}$$

where $\bar{y}$ represents the mean of the intersection of the reaction set over the 1000 simulations, x is the subset size (*i.e.*, the number of organisms drawn), α_l is the asymptote, r_l is the decay rate and ε_l is the residual of the l^{th} lifestyle group. N_l is theoretically defined as the mean of the reaction sets for an empty subset size ($\bar{y}_l$ for $x_l = 0$). A null intersection of the reaction sets corresponds to an asymptote $\alpha = 0$.

Preliminary analyses showed the strong impact of the two *Mycoplasma* species on the intersection size due to their reduced genomes (data not shown). Thus, both species were removed from the CA group for this simulation. We used the R package NLSTOOLS (Baty et Delignette-Muller, 2011) for model parameter estimation.

Differential random loss of enzymes

In order to rule out the possibility that the small intersection of partial EC number sets could be simply explained by a differential random loss of enzymes during genome reduction of the intracellular symbionts, we simulated the MIV (Mutualistic Intracellular Vertically transmitted, see Figure 1 for group names) partial EC number sets starting from bacteria of the EXTRA group. This was restricted to the *Gammaproteobacteria* of both groups. To do so, for each *Gammaproteobacteria* of the MIV group (7 organisms), we randomly picked a corresponding EXTRA *Gammaproteobacterium* and we randomly removed reactions from its set of reactions, until we reached the size of the corresponding MIV metabolic network. Then, we replaced each remaining reaction by its partial EC number at level 3, and removed redundant partial EC numbers from this set. We therefore obtained a group of simulated MIV networks for which we computed the union, intersection and average size of their partial EC number sets. This whole procedure was repeated 1000 times. Additionally, we aimed to test the differential random loss of biochemical capabilities, meaning the loss of partial EC numbers (at level 3). For that, we performed a similar procedure to the one explained above, however we stopped removing reactions when we reached the size of the MIV partial EC number set. We used a Monte-Carlo test from the R ADE4 package (Dray et Dufour, 2007) to compare simulated and observed values.

Metabolites potentially acquired from the environment

In order to identify which metabolites each bacterium potentially acquires from its environment (*i.e.*, potential inputs), we used the Borenstein method (Borenstein *et al.*, 2008). For this, the metabolic network of each bacterium was modelled as a directed compound graph, whose nodes are metabolites and arcs link a substrate to a product of a reaction. The cofactors and current compounds were filtered. We implemented a version of the Borenstein method using the IGRAPH package (Csardi et Nepusz, 2006). In order to cope with possible common inputs missed by the metabolic network reconstruction, we allowed distance one from the topological precursors if they were already assigned as input in another bacterium, and we grouped and compared them among organisms. In this analysis, the following compounds were removed since they are only produced by reactions which also involve macromolecules: dADP, dCDP, dUDP, dGDP. Hence, a systematic search for the inputs in the small molecule metabolism would indicate these compounds as potential inputs, whereas they in fact can be produced by the cell.

3.1.3 Results

Data overview

The total number of genes varies greatly among the 58 bacteria, ranging from 203 genes for "*Ca.* Hodgkinia cicadicola" (HODCD) to 7279 genes for *Ralstonia eutropha* (RALEH) (Figure 2 and Additional file 3). We noticed that the number of genes is greater in the EXTRA than in the INTRA, in agreement with the reduced genomes related to the intracellular lifestyle (Andersson et Kurland, 1998; Wernegreen, 2005).

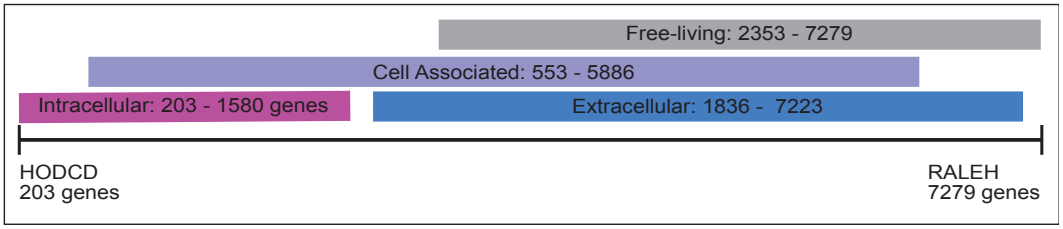


Figure 3.2: Range of the total number of genes in the different lifestyle groups. Abbreviations of group names are those from 3.1.

The three most frequent taxonomic classes among all the organisms analysed are *Gammaproteobacteria*, *Alphaproteobacteria* and *Bacilli*. These classes are well distributed in relation to the number of genes (Additional file 3), with no correlation observed between the two factors (Kruskal-Wallis test, $p = 0.65$).

The number of metabolic genes ranges from 49 for "*Ca.* Hodgkinia cicadicola" (HODCD) to 1970 for *Mycobacterium smegmatis* (MYCS2). As for the total number of genes, the number of metabolic genes is greater in the EXTRA as compared to the INTRA bacteria. However, the ratio of the number of metabolic genes over the total number of genes shows important differences depending on the organism (Figure 3). For the organisms in the groups other than MIV, the mean ratio of metabolic genes is 0.21 ± 0.08. By contrast, the mean ratio of metabolic genes for the MIV bacteria is 0.38 ± 0.18 and even reaches 0.48 for "*Candidatus* Baumannia cicadellinicola" (BAUCH), being significantly different from the bacteria of the other groups (Wilcoxon test, $p < 0.001$).

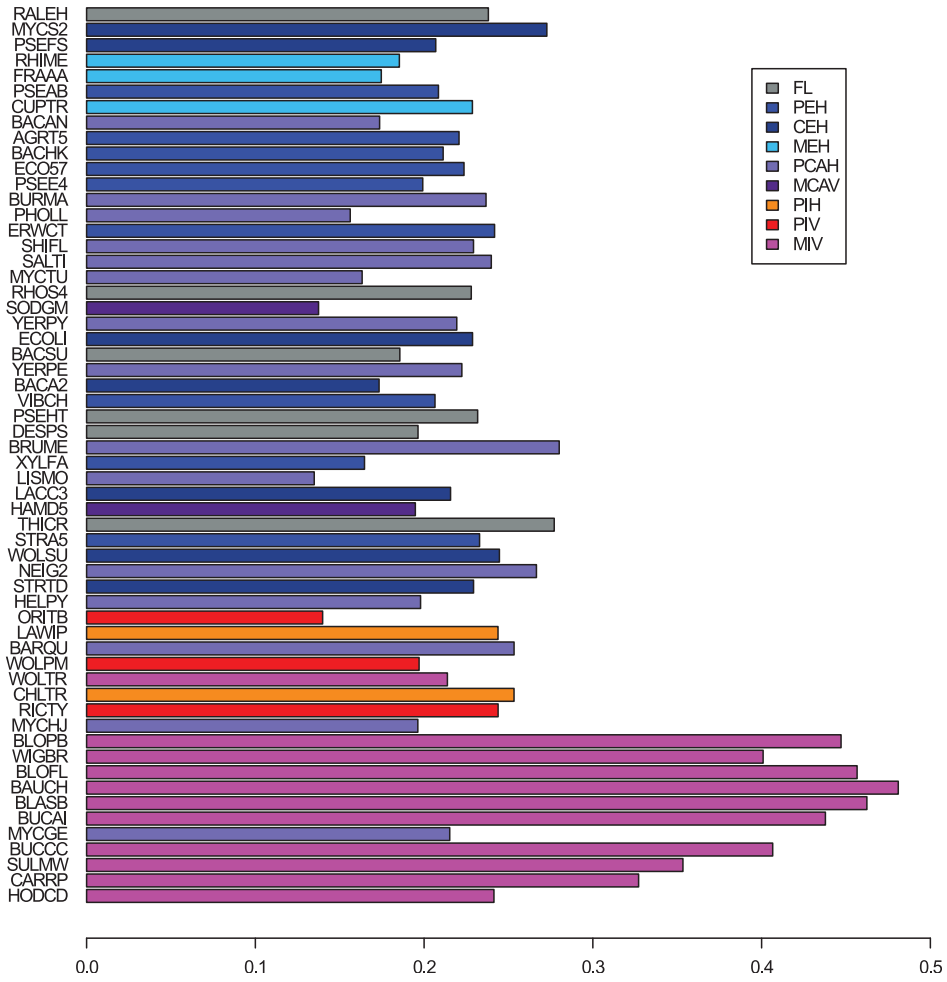


Figure 3.3: Ratio of the number of metabolic genes over the total number of genes. The colours correspond to the lifestyle groups (see Figure 3.1 for the abbreviations of group names) and bacteria are ordered by total number of genes.

Among the 58 bacteria analysed, the number of compounds and reactions follows almost the same trend as the number of metabolic genes, from 98 compounds and 42 reactions for "*Ca.* Hodgkinia cicadicola" (HODCD) to 1381 compounds and 1166 reactions for *M. smegmatis* (MYCS2) (Figure 4).

Core metabolism in the whole set of bacteria

Shared compounds and reactions

For the whole set of organisms, there are only 16 common compounds (Additional file 4) which correspond to amino acids, cofactors, ions and metabolites involved in the synthesis of nucleic acids. No small-molecule metabolic reaction is common to every organism of the dataset (Figure 5 and Additional file 5). The 16 compounds shared by the 58 organisms are therefore not involved in the same reactions in each organism. The full list of reactions analysed with the number of bacteria that possess them and the list of the organisms that lack them is presented in the Additional file 6.

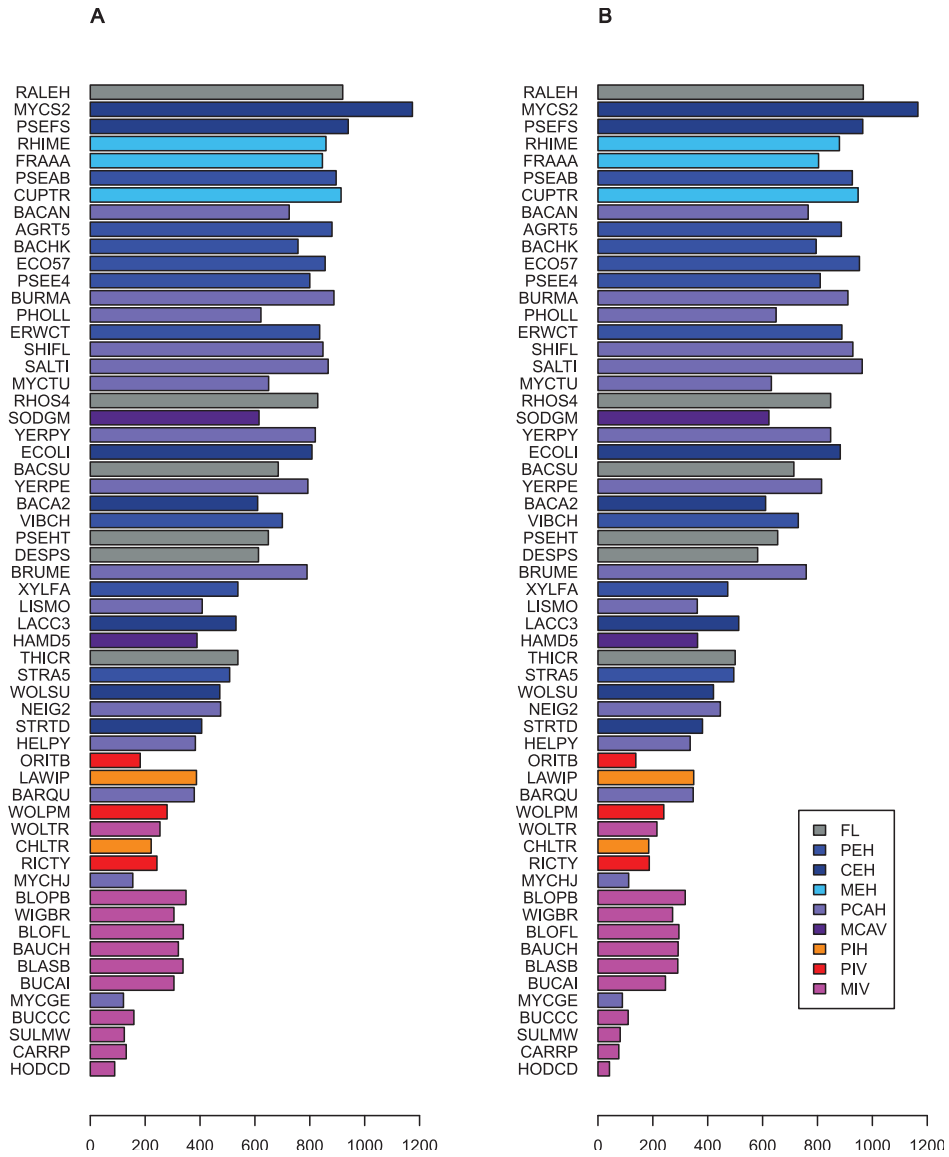


Figure 3.4: Total number of elements in the metabolic network of each bacterium in the dataset. **A** Total number of compounds. **B** Total number of reactions. The colours correspond to the lifestyle groups (see Figure 3.1 for the abbreviations of group names). Bacteria are ordered by total number of genes.

Core enzymatic function based on an EC number analysis

We found only four partial EC numbers common to all 58 bacteria: two transferases (2.3.1 and 2.5.1), one hydrolase (3.5.1) and one lyase (4.2.1) (Table 1, Figure 6a and Additional file 7). They correspond to 235 reactions in the union of all the reactions of our dataset. We searched for any four reactions, each corresponding to one of the four partial EC numbers, that are connected in the metabolic networks. In the union of all metabolic networks of the dataset, the graph induced by these 235 reactions has 418 arcs and forms 20 connected components apart from 84 isolated reactions. There are 30 occurrences of the four reactions (one for each of the four partial EC numbers) connected in the union of the metabolic networks. In the reaction graph of each organism, we found this connected pattern of four reactions in

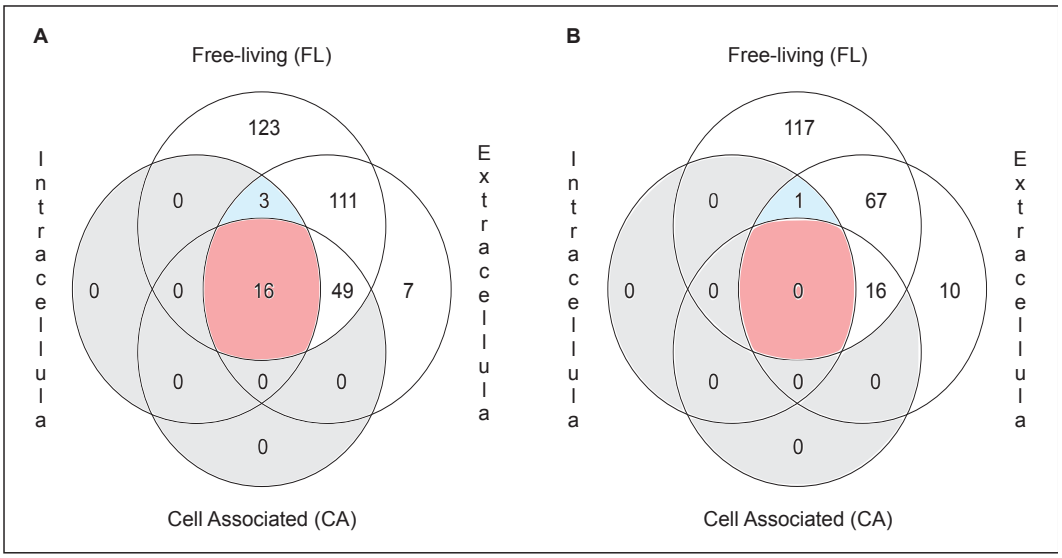


Figure 3.5: Size of the intersection of compound (A) and reaction (B) sets among groups of lifestyle. **Red**: number of compounds/reactions common to the whole dataset. **Blue**: number of compounds/reactions common the intracellular, extracellular and FL groups.

the graphs of 28 organisms. The common set of partial EC numbers in the whole dataset thus corresponds to a connected portion of the metabolic network of 28 bacteria out of the 58 analysed. This subset of bacteria represents most lifestyle groups described in this work, ranging from obligate intracellular to free-living as well as from mutualists to parasites.

Table 3.1: **Partial EC numbers common to the whole dataset**

EC number	Classification
2	**Transferases**
2.3	Acyltransferases
2.3.1	Transferring groups other than aminoacyl groups
2.5	Transferring alkyl or aryl groups, other than methyl groups
2.5.1	Transferring alkyl or aryl groups, other than methyl groups (only subclass identified to date)
3	**Hydrolases**
3.5	Acting on carbon-nitrogen bonds, other than peptide bonds
3.5.1	In linear amides
4	**Lyases**
4.2	Carbon-oxygen lyases
4.2.1	Hydro-lyases

The detailed description of the 4 partial EC numbers shared by the whole dataset includes 3 classes of enzymes: transferase, hydrolase and lyase.

Controlling for the impact of small networks

Clearly, we can expect that the inclusion of bacteria with very reduced genomes will have a large impact on the size of the intersection. However, it remains unclear if the small size of the intersection could be explained only by this. We found that the shared reaction sets, obtained

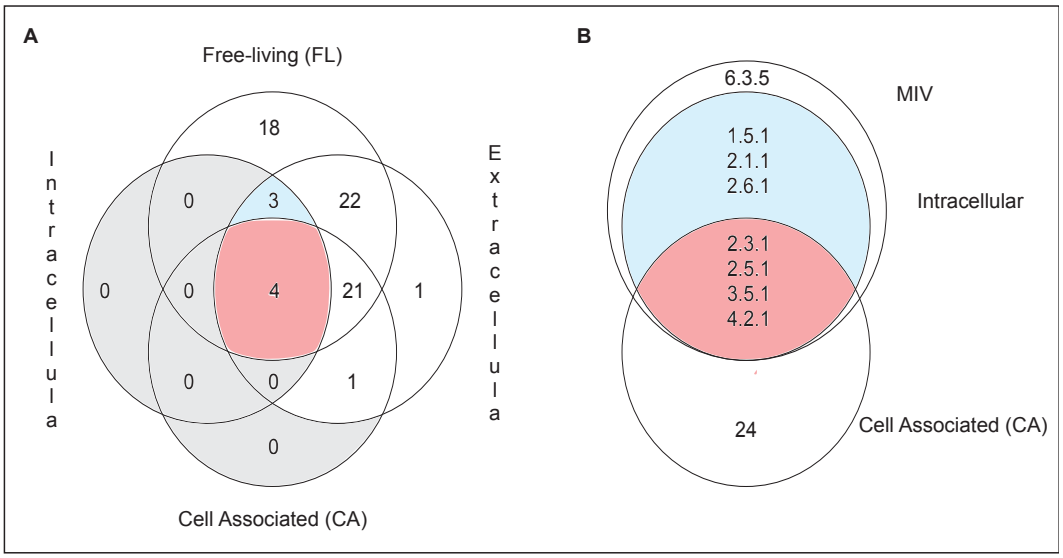


Figure 3.6: Size of the intersection of partial EC number sets. **A** Intersection of the partial EC number sets among groups of lifestyle. **Red**: number of compounds/reactions common to the whole dataset. **Blue**: number of compounds/reactions common to the intracellular, extracellular and FL groups. (**B**) Detailed intersection of the partial EC number sets of the MIV, intracellular and CA groups. **Red**: the 4 partial EC numbers common to the whole dataset. **Blue**: 3 partial EC numbers present in the intracellular, extracellular and FL groups and absent in the CA group.

when decreasing the number of organisms, range from zero to five and the mean varies from 0.2 to 2.5. There are 12 common reactions without the six bacteria with the smallest reaction sets, 7 take part in the biosynthesis of peptidoglycan, which is a cell wall precursor. These reactions do not complete this biosynthetic pathway (there are 4 missing steps, two present in 51 organisms and the other two in less than half of the dataset). Among the other reactions, there is an inorganic pyrophosphatase, a reaction involved in folate transformations and a couple of reactions which take part in purine nucleotides *de novo* biosynthesis. Removing the eight bacteria with the smallest reaction sets resulted in similar intersection sizes ranging from zero to 13 reactions and mean varying from zero to 5.4. Hence, the intersection sizes did not increase much without the bacteria with reduced metabolism.

On the other hand, these bacteria had a greater impact on the common partial EC number sets. The size of the intersections when decreasing the number of organisms ranges from 4 to 19 and the mean varies from 5 to 16. This size increases to 23 partial EC numbers without the same six bacteria. When removing the eight bacteria with the smallest sets, the sizes of the intersections increase, ranging from 4 to 27, reaching 30 partial EC numbers. Therefore, a core of biochemical capabilities composed of 30 partial EC numbers is present in 50 bacteria out of the 58 analysed. This set is made of all classes of enzymes: 3 oxidoreductases, 13 transferases, 4 hydrolases, 3 lyases, 5 isomerases and 2 ligases, which correspond to 958 reactions in our dataset (Table 2).

Table 3.2: **Partial EC number set common to 50 bacteria of the dataset**

Classes	partial EC	N^o reactions
Oxidoreductases	1.1.1	134
	1.2.1	50
	1.5.1	15
Transferases	2.1.1	46
	2.1.2	6
	2.2.1	8
	2.3.1	47
	2.4.1	69
	2.5.1	51
	2.6.1	51
	2.7.1	71
	2.7.2	8
	2.7.4	19
	2.7.6	4
	2.7.7	23
	2.7.8	15
Hydrolase	3.1.3	39
	3.5.1	32
	3.5.4	15
	3.6.1	26
Lyases	4.1.1	49
	4.1.2	22
	4.2.1	62
Isomerase	5.1.1	10
	5.1.3	16
	5.3.1	21
	5.4.2	9
	5.4.99	10
Ligase	6.3.2	19
	6.3.4	11
Total	30	958

Partial EC number set common to 50 bacteria out of the 58 present in our dataset and the number of reactions in the dataset corresponding to each partial EC number. The 8 bacteria which possess the smallest partial EC number sets were removed to calculate the common set.

Core metabolism according to lifestyle groups

Shared compounds and reactions

The same analyses performed for the whole dataset were also applied to the different lifestyle groups (Figure 5 and 6). The aim was to describe the influence of these groups on the size and composition of the common sets of compounds, reactions and partial EC numbers as well as on the union of each of these sets of elements. As a first overview of the two opposing groups in terms of the size of the metabolic networks, *i.e.*, the INTRA and the EXTRA groups, we notice that the size of the union of the compound sets (*i.e.*, the pan-metabolome) is quite large when compared to the mean number of compounds, indicating a relative diversity of the metabolome in these organisms (Table 3). By contrast, the sizes of the intersections of compound and reaction sets for the same groups are considerably different. Therefore, the universe of compounds and reactions of the intracellular bacteria is quite diverse, while common elements are far less abundant.

Table 3.3: **Compound and reaction sets among lifestyle groups**

	Compounds		Reactions	
	Mean / Union	**Intersection / Mean**	**Mean / Union**	**Intersection / Mean**
Intracellular	39%	8%	30%	0.5%
Cell Associated (CA)	39%	11%	34%	3%
Extracellular	41%	25%	36%	12%
Free-living (FL)	52%	43%	46%	28%

Ratio of the mean size over the union size, as well as the ratio of the intersection size over the mean size, of the compound and the reaction sets among the different lifestyle groups of bacteria.

Core metabolism in the extracellular bacteria

We found 186 compounds and 94 reactions shared by the 19 extracellular bacteria. The compounds include nucleosides, amino acids, carbohydrates, cofactors, while the reactions are involved in metabolic pathways, such as glycolysis, nucleotide and amino acid biosynthesis and degradation pathways, and peptidoglycan biosynthesis (Additional file 8). Most of them (88%) are classified as biosynthetic processes according to the metabolic processes defined in the BioCyc databases (Additional file 9). These reactions shared by the EXTRA are not connected in the reaction graph induced by these 94 reactions, which is composed of 10 connected components apart from 17 isolated reactions. The largest component has 26 reactions which are involved in pyrimidine ribonucleotides *de novo* biosynthesis, peptidoglycan and amino acid biosynthesis.

Core metabolism in the cell associated bacteria

The CA bacteria showed a considerable reduction in the common elements which are 67 compounds and 17 reactions. Even with this reduction, similar categories of compounds as for the EXTRA were found, whereas the reactions observed take part in fewer metabolic pathways: glycolysis and nucleotide biosynthesis and degradation pathways. Most of them (82%) are classified as biosynthetic processes (Additional file 9). This group is supposed to be intermediate between the INTRA and the EXTRA ones, thus presenting a broad diversity of genome sizes. In this group, the two bacteria with smallest genomes are *M. genitalium* (MYCGE) and *M. hyopneumoniae* J (MYCHJ) which are obligate parasites that have undergone extreme reductive genome evolution (Glass *et al.*, 2006; Yus *et al.*, 2009; McCutcheon, 2010). This pair of organisms is the one that most influences the small intersection of the CA group. Hence, the intersection of the elements of the CA bacteria without the two *Mycoplasma* species increases to 167 compounds and 88 reactions. These values are similar to the ones found for the EXTRA, the reactions take part in the same metabolic pathways observed for this group and the classification into biosynthetic and degradation processes present similar ratios (Additional file 9).

Core metabolism in the obligate intracellular bacteria

The INTRA share 19 compounds and one reaction (3.5.1.88-RXN, MetaCyc (Caspi *et al.*, 2008)). Indeed, the MIV is the group mainly responsible for this reduction. The common compounds still include the same ones mentioned for the EXTRA group. The only shared reaction is not assigned to participate in any metabolic pathway in MetaCyc.

As there is only one reaction common to INTRA, it is not possible to analyse whether there is a majority of biosynthesis reactions in their core as we found in the EXTRA and CA. Instead, we analyse the content of biosynthesis and degradation reactions in the variable metabolism (see Methods for definition). The total number of reactions in the variable metabolism is 704 (62% in biosynthetic and 24% in degradation processes) for the intracellular group while it is 2049 (38% in biosynthesis and 35% in degradation) in the EXTRA (Additional file 9). The variable metabolism of intracellular bacteria is therefore enriched in biosynthetic reactions (Fisher exact test, $p < 10^{-15}$) and depleted in degradation reactions (Fisher exact test, $p < 10^{-8}$).

Controlling for the structuring of MIV

As mentioned, the absence of a metabolic core common to all symbionts is mainly caused by the absence of such a core within the MIV group. We further analysed this group and we found two subgroups with opposite patterns of reaction presence/absence (Figure **??**), which can be directly related to the role of the symbiont in the mutualistic relationship (Akman *et al.*, 2002; Zientz *et al.*, 2004; Foster *et al.*, 2005; McCutcheon et Moran, 2007; Baumann *et al.*, 1995; Shigenobu *et al.*, 2000; Gil *et al.*, 2003; Degnan *et al.*, 2005; López-Sánchez *et al.*, 2009). Subgroup A presented mainly reactions involved in the biosynthesis of amino acids, whereas subgroup B showed reactions involved in heme synthesis. Indeed, *Wigglesworthia glossinidia* (WIGBR) and *Wolbachia pipientis* wBm (WOLTR) are known to supply heme to their hosts (Akman *et al.*, 2002; Zientz *et al.*, 2004; Foster *et al.*, 2005), while "*Candidatus* Sulcia muelleri" (SULMW) and *Buchnera aphidicola* APS (BUCAI) are known to provide amino acids (McCutcheon et Moran, 2007; Baumann *et al.*, 1995; Shigenobu *et al.*, 2000). In the case of *Candidatus* Blochmannia floridanus (BLOFL), *Candidatus* Blochmannia pennsylvanicus (BLOPB) and *Blattabacterium* sp. Bge (BLASB), the main symbiotic function is the metabolism of nitrogen, but the conservation of most of the pathways for the synthesis of essential amino acids indicates that they may also have an important role in the symbiosis (Gil *et al.*, 2003; Degnan *et al.*, 2005; Zientz *et al.*, 2004; López-Sánchez *et al.*, 2009).

Other than the impact of the small networks, this structuring of the MIV symbionts in two subgroups could also have an impact on the size of the intersection of the reaction sets. To test that, we calculated the intersection of the reaction sets between each subgroup (A or B) and the other bacteria which are not in the MIV group (47 organisms). This resulted in null for group A and 5 reactions for group B. Removing the two species of *Mycoplasma*, the intersection sizes are 2 reactions for the group A plus the non MIV bacteria and 27 for the group B. As the size of the intersection remains small in either subgroup of MIV symbionts with the non MIV organisms, the structuring of the MIV symbionts does not explain the reduced number of reactions shared by the bacteria analysed.

Decay of the common reactions in the different lifestyle groups

As there is a clear trend in the INTRA group indicating that common reactions decrease to none rapidly, it is important to address the question whether the other groups follow the same rule. This analysis is based on a simulation of the number of common reactions for different subset sizes of the organisms (Figure 7). The exponential model fitted best the data of the INTRA group, however it did not fit well the data of the CA (without the *Mycoplasma* species, see Methods) and EXTRA groups (results not shown). We therefore tested a logistic model, which has a smoother decay than the exponential model, and found that it fitted much better this data. As expected, the asymptote for the INTRA group (α_{INTRA}) was not

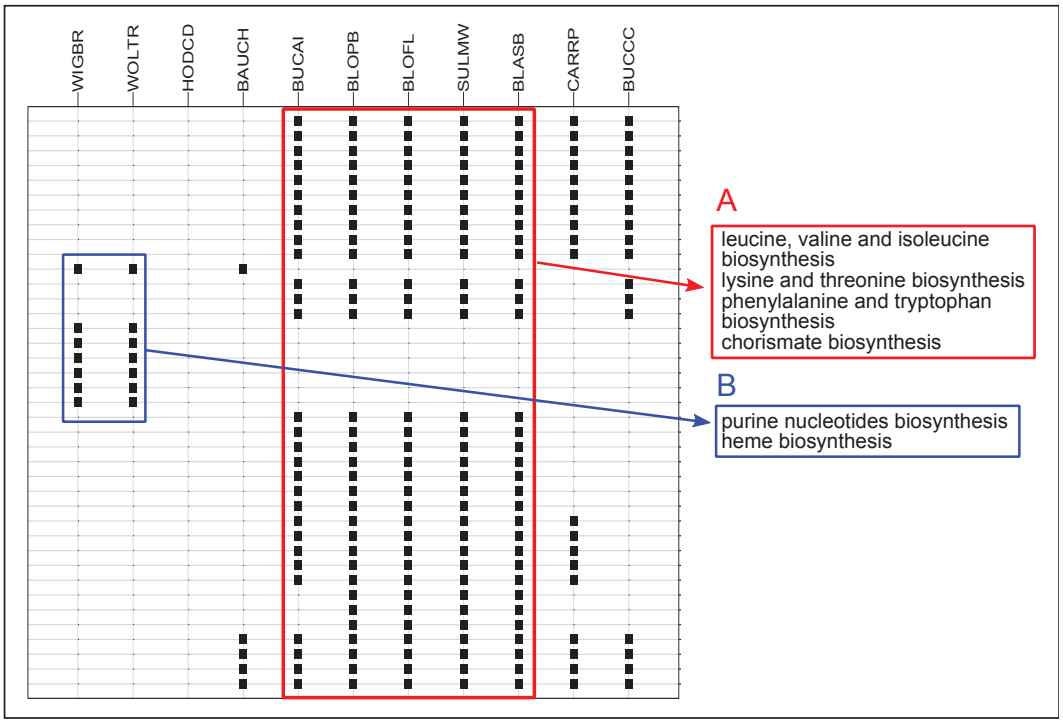


Figure 3.7: Two subgroups of MIV symbionts which present opposite patterns of reaction presence/absence. Table of the presence/absence of reactions with correlation ratio greater than 50% on the second axis of the MCA. The square indicates the presence of the reaction in the organism of the respective column. The red rectangle indicates subgroup A: *Buchnera aphidicola* APS (BUCAI), "*Candidatus* Blochmannia pennsylvanicus" (BLOPB), "*Candidatus* Blochmannia floridanus" (BLOFL), "*Candidatus* Sulcia muelleri" (SULMW) and *Blattabacterium* sp. Bge (BLASB) which presented mainly reactions involved in the biosynthesis of amino acids. The blue rectangle indicates subgroup B: *Wigglesworthia glossinidia* (WIGBR) and *Wolbachia pipientis* wBm (WOLTR) which showed reactions involved in heme synthesis.

significantly different from zero (Table 4). In contrast, the asymptote was estimated at 54 for the CA and 66 for the EXTRA. Thus, based on the analysed dataset, neither the CA nor the EXTRA group is expected to have an empty common set of reactions.

Table 3.4: **Parameters estimated for the fitting models**

Lifestyle group (l)	**Model**	α_l	r_l	N_l
INTRA	Exponential	1.70±5.17	0.56±0.19	269.48±122.30
CA	Logistic	53.84±4.37	0.06±0.01	1643.00±193.08
EXTRA	Logistic	65.85±3.06	0.06±0.00	1748.00±137.10

Estimates were obtained by fitting the exponential and logistic models to the corresponding data sets. α_l is the asymptote, r_l is the decay rate and N_l is theoretically defined as the mean of the reaction sets for an empty subset size of the l^{th} lifestyle group. Point estimates are given with ±2×standard error.

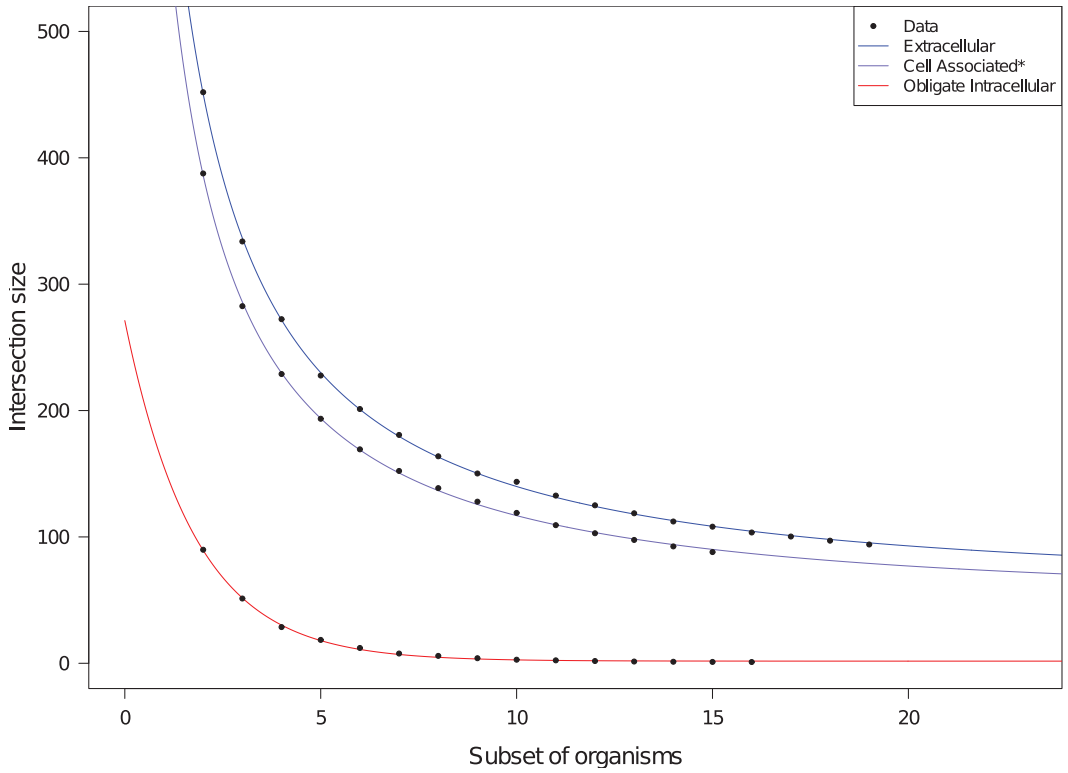


Figure 3.8: Decay of the common reactions in the different lifestyle groups. Data is the mean value of the 1000 simulations of the intersection of reaction sets for the different subsets of organisms for each lifestyle group. The red curve corresponds to the fitted exponential model to the data obtained for the obligate intracellular group, while the blue and violet curves correspond to the fitted logistic model to the data obtained for the extracellular and CA groups, respectively. The two *Mycoplasma* were removed from the CA groups for this analysis (see Methods)

Core enzymatic function based on an EC number analysis for lifestyle groups

The number of shared partial EC numbers depends on the lifestyle groups: 7 were found for the INTRA, 28 for the CA and 52 for the EXTRA (Figure 6a and Additional file 7). These values represent 6%, 20% and 34% of the respective union in each group. The common set for the INTRA and CA groups is exactly the same as for the whole set of bacteria. The set for the INTRA group adds 3 more partial EC numbers when compared to the common partial EC numbers of the whole dataset: one oxidoreductase (1.5.1) and 2 transferases (2.1.1 and 2.6.1) (Figure 6b). The MIV bacteria are the ones which mainly account for the small size of the intersections in the INTRA group, and they share 8 partial EC numbers which means adding the ligase 6.3.5 to the common set. Hence, the common set of EC numbers for the MIV group comprises all classes of EC numbers, except for isomerase. Both the CA and the EXTRA shared sets have all the 6 classes of enzymes; the number of different EC numbers of each class ranges from two subclasses of lyases to 17 subclasses of transferases.

Connectivity of the partial EC number set for obligate intracellular and extracellular bacteria

We searched for connected reactions corresponding to the common partial EC numbers for the INTRA and EXTRA groups (7 and 52, respectively). The 7 partial EC numbers common to the intracellular bacteria correspond to 154 reactions in the union of all reactions of this group, whereas the 52 of the EXTRA bacteria correspond to 1253 reactions. The same procedure of searching for connected reactions corresponding to the common partial EC numbers in the INTRA and EXTRA groups was performed in the union reaction graph of each group. The induced reaction graph from the 154 reactions in the intracellular group is composed of 174 arcs and 19 connected components apart from 47 isolated reactions. There is no connected set of 7 reactions labelled with the 7 partial EC numbers common to the intracellular symbionts. In the case of the EXTRA group, the induced reaction graph from the 1253 reactions has 5288 arcs and 38 connected components apart from 181 isolated reactions. There is no occurrence of 52 reactions whose EC numbers are the 52 common to the EXTRA bacteria that is connected in this graph. Hence, the common set of partial EC numbers in the intracellular and EXTRA groups does not correspond to a connected portion of the metabolic network of these bacteria.

Differential random loss of enzymes

The differential random loss of enzymes or of biochemical capabilities (meaning partial EC numbers at level 3) was compared with the small intersection size of the partial EC number sets. Comparing the simulated values with the ones of the real MIV *Gammaproteobacteria* (Table 5 and Additional file 10), we found that the MIV have lost a greater diversity of biochemical capabilities than expected by simulation (Monte Carlo test, $p < 0.001$).

Table 3.5: **Partial EC number sets for the MIV *Gammaproteobacteria* and for the simulated MIV**

	Mean	**Union**	**Intersection**
MIV *Gammaproteobacteria*	59	81	20
Differential random loss of reactions	69 ± 2.9	122 ± 6.2	21 ± 4.9
Estimated $p - value$	≤ 0.001	≤ 0.001	≤ 0.692
Differential random loss of biochemical capabilities	59 ± 0	114 ± 6.1	13 ± 4.4
Estimated $p - value$	1	≤ 0.001	≤ 0.006

Size of the mean, union and intersections of the sets of partial EC number for the MIV *Gammaproteobacteria* and for the simulated MIV.

Metabolites potentially acquired from the environment

The absence of a metabolic core in the INTRA might be linked to the differences in their environment. The number of metabolites that each bacterium potentially acquires from its environment (*i.e.*, potential inputs) ranged from 29 in "*Candidatus* Sulcia muelleri" (SULMW) to 341 in *M. smegmatis* (MYCS2) with a mean of 133 (Additional file 11). There are no potential inputs common to the 58 bacteria and the union of inputs is 1191 (Additional file 12). The intersection is null in the INTRA and the CA groups, while it is 2 in the EXTRA group. These two inputs are isolated from the rest of the network, and are linked together by one reaction which is catalysed by an enzyme that accelerates the folding of proteins (by catalysing the *cis-trans* isomerisation) (Caspi *et al.*, 2008). Overall, we found no common

inputs to the whole metabolic network of EXTRA bacteria. The mean values of the inputs in each group are 39, 133 and 190, respectively. Taking into account classes of compounds, the intracellular bacteria have in common ions, cofactors and nucleosides as potential inputs, while the EXTRA add vitamins and carbohydrates.

When we allowed distance one from the topological precursors (see Methods for details), the number of common inputs increased inside the lifestyle groups that have less organisms, such as PIV. In the broader lifestyle groups, the number of shared inputs remained equal. The number of bacteria that has glucose as input increased from 3 to 40. Furthermore, the size of the intersection augmented between the groups, such as MEH and PEH.

Overall, we find that neither EXTRA nor INTRA symbionts exhibit a common core of input metabolites. The absence of such a core is an intuitive explanation for the absence of a metabolic core of degradation pathways: different metabolic environments imply different metabolic pathways. However, this observation alone does not explain the total lack of a metabolic core for the INTRA symbionts. In this case, the specificity of the symbiosis with the host has to be considered.

3.1.4 Discussion

In this paper, we investigated to what extent there is any reaction common to a set of bacteria, including obligate intracellular symbionts, as well as the influence and the trend of each lifestyle group concerning shared reactions or biochemical capabilities. In order to do this, we considered 58 bacteria carefully selected to represent a wide range of lifestyles.

Existence of a metabolic core

Previous studies have found small sets of common metabolic genes even when including bacteria with reduced genomes (Mushegian et Koonin, 1996; Klasson et Andersson, 2004). Based on that and on the fact that we analysed reactions instead of genes (partially addressing the issue of NOGD), we therefore expected to find a small core of functional capabilities. Our analyses of the small molecule metabolism of 58 bacteria revealed however that they share no reaction, 16 compounds and 4 partial EC numbers.

Even though there was no reaction common to all bacteria, we actually found one reaction (3.5.1.88-RXN, MetaCyc (Caspi *et al.*, 2008)) present in all the dataset except in *M. hyopneumoniae* (MYCHJ). It is catalysed by the hydrolase peptide deformylase (Def), which releases the formyl group from the N-terminal methionine residue of most nascent polypeptides (Adams, 1968), an obligatory step during protein maturation in eubacteria (Rajagopalan *et al.*, 1997). The absence of Def in this bacterium apparently leaves it unable to formylate Met-tRNAi (Vasconcelos *et al.*, 2005), and it has been described as absent or nonessential in *Phytoplasma* sp. and *Mycoplasma arthritidis* (Vasconcelos *et al.*, 2005; Dybvig *et al.*, 2008). For long, peptide deformylase was believed to be exclusively present in bacteria, however Giglione *et al.* (Giglione *et al.*, 2000) identified eukaryotic deformylases which were localized in the organelles only. In our dataset, even the symbiont with most reduced genome ("*Ca.* Hodgkinia cicadicola" (HODCD)) is potentially capable to code for this enzyme. Nevertheless, recently an even smaller cellular genome (approx. 139 base pairs and 121 protein-coding genes) of "*Candidatus* Tremblaya princeps" has been described (McCutcheon et von Dohlen, 2011) which is missing homologs for Def. The presence of this enzyme in almost the whole dataset is justified by the fact that it is mostly related to information processing which is expected to be among the minimal functions required for sustaining life (Danchin, 1989; Mushegian et Koonin, 1996; Koonin, 2000; Klasson et Andersson, 2004; Gil *et al.*, 2004).

Such small sets found raised the question whether they could be explained only by the (6 or 8) bacteria with the smallest genomes. These bacteria had a weak impact on the number of shared reactions, while they had a strong effect on the common partial EC number set. Removing them, the shared set increased to 12 reactions mainly involved in the synthesis of a cell wall precursor, which is not considered as an essential pathway (Juhas *et al.*, 2011) and is known to be absent or reduced in host-dependent bacteria (Moya *et al.*, 2008; Pérez-Brocal *et al.*, 2006). Conversely, the common partial EC number set increased to 30 without those bacteria which is a quite broad set of biochemical capabilities. All six classes of enzymes are included in this set, and are similar to the ones described for a minimal metabolism (Gabaldón *et al.*, 2007). Only two partial EC numbers at level 3 (2.4.2 and 1.17.4) from this minimal metabolism are not included in our partial EC number set, however the latter partial EC number should not be in our analyses because it involves macromolecules and we work strictly with the small molecule metabolism. Furthermore, 8 of the 30 shared partial EC numbers are not included in this minimal metabolism, and four of them are transferases which are enriched in our common partial EC number set (43%).

The reduced set of common partial EC numbers raised the question whether it could be simply explained by a differential random loss of enzymes. This was not the case. We further identified the MIV *Gammaproteobacteria* as having lost a greater diversity of biochemical capabilities. This indicates that there is a set of partial EC numbers (capabilities) which are kept in subsets of organisms (not in every bacteria, *i.e.* it is not included in the shared set) and accounts for a reduced union.

Hence, we did not find a core of metabolic reactions shared by the symbiotic bacteria which agrees with the idea that searching for ubiquity as more genomes are included may ultimately reduce to nothing (Danchin *et al.*, 2007). Conversely, using a more relaxed approach we found a core of biochemical capabilities which is similar to a minimal metabolism previously described (Gabaldón *et al.*, 2007).

Impact of the lifestyle groups on the existence of a metabolic core

Among the different types of classification that we considered – (i) obligate intracellular, extracellular, cell associated, (ii) mutualistic, commensalist, parasitic, (iii) vertically or horizontally transmitted – the first is by far the one that explains best the differences in terms of metabolism. The CA group also accounted for the small common sets exclusively because of the *Mycoplasma* species. Even if this group presents other host-dependent bacteria, their genome sizes at least double when compared to the *Mycoplasma* species, and a core of reactions similar in size to the EXTRA is found. The other lifestyle groups (EXTRA and FL), which include just free-living bacteria, did not contribute to the size of the common set.

Furthermore, the impact of the INTRA and of the *Mycoplasma* species in the small sets can be directly related to their extremely reduced genomes (Wernegreen, 2002; Gil *et al.*, 2002; McCutcheon et Moran, 2012). They also have much fewer metabolic genes, even though this category is much less affected by the reduction in the INTRA group specially in the MIV. These bacteria (except for *W. pipientis* wBm (WOLTR)) are the most integrated (Nardon et Grenier, 1993) and are those for which the association with the host is essentially nutritional (Akman *et al.*, 2002; Zientz *et al.*, 2004; Foster *et al.*, 2005; McCutcheon et Moran, 2007; Baumann *et al.*, 1995; Shigenobu *et al.*, 2000; Gil *et al.*, 2003; Degnan *et al.*, 2005; López-Sánchez *et al.*, 2009). Indeed, the ratio of metabolic genes is significantly higher for MIV, indicating that the loss of genes primarily concerns the non metabolic ones (Moran, 2007; Moran *et al.*, 2008; Moya *et al.*, 2008). The loss of metabolic genes is affected by the requirements for host

survival, and to some extent by the presence of other symbionts in the same environment (Moya *et al.*, 2008).

Content and connectivity of the core metabolism of CA and EXTRA

In the analyses of each lifestyle group, we did not find a core of reactions for the INTRA, however we found it for the EXTRA and CA (the latter group without the two *Mycoplasma* species - the CA mentioned henceforward is without these bacteria). The shared reactions are involved in metabolic pathways that are also included in the minimal metabolism described by (Gil *et al.*, 2004; Gabaldón *et al.*, 2007), such as glycolysis and nucleotide biosynthesis. The cores found also include amino acid biosynthesis pathways which are not present in the minimal metabolism because they assumed a nutrient-rich medium with amino acids unlimitedly available for the minimal cell (Gil *et al.*, 2004; Gabaldón *et al.*, 2007).

The common sets of reactions of the CA and EXTRA groups are enriched in biosynthesis (approx. 88%) according to the metabolic processes defined in the BioCyc databases. In the core metabolism of *E. coli*, biosynthetic reactions are also overrepresented (57%) (Vieira *et al.*, 2011), thus our study enables to confirm and extend this result to multiple species. Overall, the core-metabolism of the CA and EXTRA bacteria is therefore much smaller than the one of the strains of *E. coli*, but at the same time, it is even more enriched in biosynthetic reactions. The reason for such an enrichment could be that, while the needs of the CA and the EXTRA symbionts are very similar in terms of building blocks for protein and DNA synthesis, the nutrients they uptake in their respective environment may be extremely variable. When variable environments are considered, degradation pathways, which are closer to the inputs of the network, are the first to be modified. This explanation is also corroborated by our observations on the lack of common inputs to all bacteria.

Considering now the proportion of biosynthesis and degradation reactions in the variable metabolism, we find that it is quite similar in *E. coli* (36% biosynthesis and 35% degradation) and the CA and EXTRA bacteria (approx. 39% biosynthesis and approx. 35% degradation), but the numbers are quite different for obligate intracellular bacteria (62% biosynthesis and 24% degradation). A possible explanation for this is that degradation pathways have largely disappeared in obligate intracellular bacteria, as the host provides an interface between the environment and the bacterium, while synthetic routes have not all disappeared but have been selected for, depending on the nature of the symbiosis (Moran, 2007; Moran *et al.*, 2008; Moya *et al.*, 2008; McCutcheon et Moran, 2012).

Here, we worked with whole metabolic networks enabling to check whether the metabolic core would represent chains of biochemical reactions regardless of specific metabolic pathways. The core of reactions found was not entirely connected, most likely because of the existence of alternative pathways as highlighted by Gil *et al.* (Gil *et al.*, 2004). This means that searching for ubiquity even inside lifestyle groups does not result in one functional metabolic network.

Persistent metabolic core of CA and EXTRA

We found a core of metabolic reactions for the CA and EXTRA, however we did not find one for the INTRA. This raised the question whether, as we add organisms, the decay of shared reactions and its limit was the same in these groups. First, we fitted the exponential model with asymptote to the data of all groups. This model described well the decay of shared reactions in the INTRA group. However, it was not appropriate to fit the EXTRA and CA data, since their behaviour of decay was not the same as that for the INTRA. Conversely, the logistic model was well adapted for these two groups. We also tested for

common parameters for the two groups, but model fitting was better with each group having its separate parameter values. The decay rates (r_{CA} and r_{EXTRA}) were similar, while the two other parameters were different. In principle we cannot give a direct biological interpretation to N_l (it corresponds to the mean of the reaction sets for an empty subset size of organisms), we found its estimates are close to the size of the union of reactions of the corresponding lifestyle group, *e.g.*, N_{EXTRA} was estimated at 1643, while the size of the union of EXTRA was 1725 reactions. As expected, the asymptote estimated for the INTRA was not significantly different from zero, which agrees with the absence of a core of metabolic reactions found for this group. Conversely, the asymptotes estimated for the CA and the EXTRA groups were significantly different from zero; thus, based on the analysed dataset, neither group is expected to have an empty common set of reactions when more genomes of these groups are added. One should be aware that adding one organism that has a very particular niche could certainly change this trend. This result is nevertheless interesting given the fact that there are organisms from distinct taxonomic classes inside these groups, that moreover present different types of association with their hosts. To have an idea of the subset of reactions that would be "asymptotically" kept in organisms with lifestyles similar to those two groups, we analysed the reactions shared by the EXTRA and CA groups in our dataset. These 62 reactions are involved in the synthesis of purine and pyrimidine, of peptidoglycan and glycolysis. These findings are similar in number of enzymatic steps and in the content of pathways to the minimal metabolism described by Gabaldón *et al.* (Gabaldón *et al.*, 2007).

3.1.5 Conclusions

In this paper, we explored to which extent each lifestyle group contributes to the reduction of a core metabolism as well as the composition of this core in the different groups, with a special focus on bacterial species only, in particular those that entertain a symbiotic relationship with a host. Moreover, we considered reactions instead of genes. Although we might then have expected to find a core, none common to all bacteria was observed. Symbionts with the most reduced genomes in our dataset had a weak impact on the number of shared reactions, but had a strong effect on the common partial EC number set which increased to 30 without those bacteria, covering a quite broad set of biochemical capabilities similar to those described for a minimal metabolism, with however an enrichment in transferases.

Obligate intracellular symbionts appeared as the main reason for such absence of a core of metabolic reactions due to their high specialisation. However, host-dependence alone is not an explanation for this absence. Indeed, although the cell associated group contained host-dependent bacteria, their core of reactions was observed to be similar in size to the one of extracellular bacteria once the two *Mycoplasma* species were eliminated from the group. Extremely reduced genomes such as those of the two *Mycoplasma* and of the intracellular group remain thus the main factor behind the absence of a core, even though the loss of genes primarily concerns the non metabolic ones.

A core of reactions was found for the cell-associated and the extracellular bacteria. This core roughly corresponds to the minimal metabolism previously described in the literature. It is not entirely connected and therefore does not result in one functional metabolic network. Although smaller than the core previously identified for strains of *E. coli*, we observed that it is even more enriched in biosynthetic reactions, which might be due to the extreme variability of the nutrients that cell-associated and extracellular bacteria uptake in their respective environment. On the other hand, the proportion of biosynthesis and degradation reactions in the variable metabolism appears quite similar to the one found in *E. coli*. The same is not the

case for obligate intracellular bacteria where degradation pathways have largely disappeared but synthetic routes appear instead to have been selected for depending on the nature of the symbiosis.

Finally, by using simulation, we tested whether the decay of shared reactions and its limit would be the same for cell-associated and extracellular bacteria as for the intracellular ones. Although one should be aware that adding one organism that has a very particular niche could certainly change the result observed, it appears that a subset of around 60 reactions would be "asymptotically" kept in cell-associated and extracellular bacteria. These are involved in the synthesis of purine and pyrimidine, of peptidoglycan and glycolysis, and are similar in number of enzymatic steps and content of pathways to the minimal metabolism described in the literature.

3.2 The extended core of metabolic networks

3.2.1 Overview

This section introduces an ongoing investigation of the common metabolic capabilities shared by a group of species (not requiring omnipresence) which we called *extended metabolic core* and that represent a larger set of capabilities when compared to the *traditional core* (where omnipresence is required). The traditional one tends to be highly dependent on the addition or removal of a single organism since the absence of a reaction in one organism is enough to remove such capability from the core. In order to introduce some flexibility into this set, possibly making it more stable in size and content as one organism is included or removed, one could, for instance, allow in the core reactions that are absent in one or two species. It is however not easy to determine a limit for such an extended core of metabolic capabilities. To address this problem, we propose a new approach where common and group-specific reactions are split automatically. The method was developed in collaboration with statisticians from the Laboratoire Statistique et Génome, INRA, Évry, France, namely Christophe Ambroise, Yolande Diaz and Catherine Matias.

We propose two different definitions of an extended core of reactions as opposed to the group of reactions that are organism-specific (which we call *periphery*): (i) one that is based on the presence/absence of a reaction in an organism, and (ii) a second that relies on a neighbour relationship between reactions where two reactions are considered neighbours if they share a metabolite. It is important to note that we are not using the term *periphery* in the topological sense, but rather in the sense of parts that are "marginal" in relation to the functional core. The second may be seen as a smoothing or refinement of the first. The first is robust with respect to the set of organisms under consideration, in the sense that reactions do not need to be present in *all* organisms but only in a *large enough proportion* in order to belong to the core. The first advantage of the approach that we propose is that the threshold to decide what is considered *large enough* is not set by the user (thus relying on a subjective choice), but is rather automatically selected by the method based on the information contained in the data and the two definitions given above.

Actually, we developed two methods, one that is based only on the first definition above, and the second that is based on both. The latter will tend to classify in a same group (core or periphery) a reaction for which a majority of its neighbours belong to this group. As mentioned, this results into clusters which are *smoothed* or *regularised* with respect to the first method. In particular, a typical example of this smoothing effect is obtained when one reaction which is present in a majority of the organisms is the neighbour of only one other reaction (or of very few), which appears only in one organism (or in very few). In this case, our second method will classify the first reaction in the periphery, even if it is present in a majority of the organisms. Indeed, it is reasonable to think that since it interacts only with a small number of other reactions which are not present in all organisms, it is in fact a *potential* rather than an *effective* reaction. In this sense, we argue that our first method will detect potential core reactions while the second one focuses on effective core reactions.

3.2.2 Dataset description

The dataset is composed of: (I) 13 organisms from the *Escherichia/Shigella* genera and (II) 13 from the *Pseudomonas* genus (Table 3.6). The aim is to have two groups with similar phylogenetic distance, however distinct in terms of ecological niches: group (I) narrow and (II) wide. The habitat of the first group is mainly in the human intestinal microflora, whereas the habi-

tat of the second includes different hosts as well as soil, fresh and waste water. We selected all available species from the above mentioned genera from the MicroCyc/Genoscope platform (Vallenet *et al.*, 2009) (accession in February 2013). MicroCyc is a collection of microbial Pathway/Genome Databases which were generated using the PathoLogic module from the Pathway tools software (Karp *et al.*, 2010) which computes an initial set of pathways by comparing a genome annotation to the metabolic reference database MetaCyc (Caspi *et al.*, 2012). Since there were 13 *Pseudomonas* available, the same number of *Escherichia/Shigella* were selected with the aim to cover as best as possible a wide diversity of the various *E. coli* phylogroups (Chaudhuri et Henderson, 2012) and types of host interactions (Hamap (Lima *et al.*, 2009) information on interactions). Using these databases as input, the metabolic networks of the bacteria were obtained from MetExplore (Cottret *et al.*, 2010). The data on metabolic pathways were obtained from MetaCyc (Caspi *et al.*, 2012).

Our analysis is restricted to the small molecule metabolism as defined in the MetaCyc/BioCyc databases (Caspi *et al.*, 2012; Karp *et al.*, 2005), *i.e.* small molecule reactions are those in which all participants are small molecules, hence reactions involving one or more macromolecules such as proteins or nucleic acids are not represented. The macromolecules were filtered out from all the analyses using MetExplore (Cottret *et al.*, 2010). The compounds found in the metabolic networks are those which are involved as substrates or products in the inferred reactions. All metabolites directly provided by the environment and not involved in any reaction as substrate are not included.

The metabolic networks of the dataset were modelled as reaction graphs. In such a graph, nodes represent reactions, and arcs (*i.e.*, directed edges) between two reactions represent a compound which is produced by one reaction and consumed by the other. Each set of organisms was modelled twice, applying or not the following filters: one that excludes pairs of cofactors (*i.e.*, $ADP + Pi \rightarrow ATP$, $NAD^+ + H^+ \rightarrow NADH$; for the full list see the MetExplore documentation) and a second that eliminates ubiquitous compounds (*e.g.*, water, proton, CO_2, phosphate, diphosphate, NH_3, H_2O_2 and O_2).

The sets of metabolic reactions and organisms are first listed and then ordered in an arbitrary way. For each of the R possible reactions, we have a binary vector of length N (the number of organisms) whose ith coordinate indicates whether the reaction is present or not in organism number i. These data are gathered in a matrix D of size $R \times N$.

Following the work of (Mithani *et al.*, 2009), a neighbour relationship is defined among reactions in the following way: two reactions that share at least one metabolite are considered neighbours. This neighbour relationship induces a graph structure, called *union graph*, on the set of reactions. This graph is described by its corresponding adjacency matrix A, namely a symmetric binary matrix of size $R \times R$, with null diagonal entries and off-diagonal entries (i, j) indicating whether reactions i and j are neighbours (*i.e.* share at least one metabolite). It is important to note that the adjacency matrix A containing the neighbour relationships is independent from the data matrix D. The reactions are thus described by both their frequency profile (data matrix D) and their neighbour profile (adjacency matrix A).

In our first clustering method, we rely only on the data matrix D while the second approach also takes into account the neighbour relationship (encoded in the adjacency matrix A) in order to obtain a finer classification of the core and periphery reactions.

3.2.3 Computation of the core/periphery reactions

We compared the two different methods for classifying the reactions into two different groups of core and periphery reactions that we proposed (see Section 3.2.1). We recall that the

Table 3.6: **Dataset used in this study.**

Organism	Interaction
Escherichia albertii TW07627	Human pathogen
E. coli 042	Human pathogen (EAEC)
E. coli CFT073	Human pathogen (UPEC)
E. coli E24377A	Human pathogen (ETEC)
E. coli IAI1	Human commensal
E. coli IAI39	Human pathogen (UPEC)
E. coli K12	Human commensal
E. coli O157:H7 Sakai	Human pathogen (EHEC)
E. fergusonii ATCC 35469T	Human commensal
Shigella boydii Sb227	Human pathogen
S. dysenteriae Sd197	Human pathogen
S. flexneri 2a 301	Human pathogen
S. sonnei Ss046	Human pathogen
Pseudomonas aeruginosa LESB58	Human opportunistic pathogen
P. aeruginosa PAO1	Human opportunistic pathogen
P. aeruginosa UCBPP-PA14	Human opportunistic pathogen
P. entomophila L48	Insect pathogen
P. fluorescens Pf-5	Plant commensal
P. fluorescens Pf0-1	Plant commensal
P. fluorescens SBW25	Plant saprophyte
P. putida F1	Plant pathogen
P. putida GB-1	Plant pathogen
P. putida KT2440	Plant pathogen
P. syringae pv. phaseolicola 1448A	Plant saprophyte
P. syringae pv. syringae B728a	- (soil, water)
P. syringae pv. tomato DC3000	- (soil, water)

EAEC - Enteroaggregative *E. coli*; UPEC - Uropathogenic *E. coli*; ETEC - Enterotoxigenic *E. coli*; EHEC - Enterohemorrhagic *E. coli*.

first detects potential core reactions, while the second detects what we called effective core reactions.

The first method relies only on the data matrix D containing the information on the presence/absence of a reaction within an organism. It clusters the reactions into two groups, using for this a multivariate Bernoulli mixture model with two components (see (Allman *et al.*, 2009; Carreira-Perpiñán et Renals, 2000) for the issue of parameter identifiability in these models). We call this method BinEm as it is based on the Expectation-Maximisation (EM) algorithm (Dempster *et al.*, 1977) applied to binary data. More precisely, the data vectors $D_i = (D_{ij})_{1\leq j\leq N}$ describing the presence/absence of the reactions ($1 \leq i \leq R$) in each organism are assumed to be i.i.d. vectors with mixture distribution:

$$D_i \sim \lambda_c \prod_{j=1}^{N} \alpha_{c,j}^{D_{ij}} (1-\alpha_{c,j})^{1-D_{ij}} + (1-\lambda_c) \prod_{j=1}^{N} \alpha_{p,j}^{D_{ij}} (1-\alpha_{p,j})^{1-D_{ij}},$$

where $\lambda_c \in (0,1)$ is the unknown proportion of reactions belonging to the core group ($1-\lambda_c$ being the unknown proportion of reactions belonging to the periphery group), $\boldsymbol{\alpha}_c = (\alpha_{c,j})_{1\leq j\leq N} \in (0,1)^N$ and $\boldsymbol{\alpha}_p = (\alpha_{p,j})_{1\leq j\leq N} \in (0,1)^N$ are the unknown vectors of probabilities that a reaction belonging, respectively to the core or periphery group is present in species j. The parameters of this model, as well as corresponding clusters are estimated with

an EM algorithm. Note that in our context, the number of groups is fixed (and equal to 2) as we want to impose a description of the dataset into core and periphery groups.

The second approach combines the data matrix D containing the observations of presence/absence of reactions within organisms with the adjacency matrix A of neighbour relationships between reactions. It is based on a hidden Markov random field model, with dependency structure among the reactions given by the adjacency matrix A. In this model, each reaction belongs to some unobserved (hidden) group and these groups are distributed among reactions according to a Markov random field. This means that two reactions which are neighbours are more likely to belong to the same group. Conditional on this hidden structure, the binary vectors of presence/absence of each reaction within the set of organisms are independent and follow a multivariate Bernoulli distribution, with proportion vector depending on the group to which the reaction belongs. We call this method NEM, as it relies on the neighbour expectation-maximization NEM algorithm (Ambroise *et al.*, 1997; Ambroise et Govaert, 1998; Dang et Govaert, 1998) developed for clustering in hidden Markov random fields.

More precisely, the model is as follows. We introduce latent variables $\{Z_i\}_{1\leq i\leq R}$ with state space $\{c, p\}$ that indicate the group (core or periphery) to which each reaction belongs. These random variables follow a Markov random field distribution, given by the following Gibbs distribution:

$$\mathbb{P}(\{Z_{i\cdot}\}_{1\leq i\leq R}) = \frac{1}{W}(\beta)\exp\Big(\beta\sum_{i\sim j}1_{Z_i=Z_j}\Big),$$

where 1_A is the indicator function of event A, the previous sum concerns every pair (i, j) of neighbour reactions (a relation denoted by $i \sim j$), the parameter $\beta > 0$ represents the inverse of the temperature and:

$$W(\beta) = \sum_{\{z_i\}_{1\leq i\leq R}}\exp\Big(\beta\sum_{i\sim j}1_{z_i=z_j}\Big),$$

is a normalising constant. Note that $W(\beta)$ may not be computed, due to the large number of possible configurations. The degree of dependence between the reactions is controlled by the parameter β: the higher its value, the smoother the clustering will be (with neighbour reactions tending to belong to the same group). Now, the data vectors $(D_i)_{1\leq i\leq R}$ are no more independent. However, conditional on the latent groups $\{Z_i\}_{1\leq i\leq R}$, they are independent and follow the multivariate Bernoulli distribution:

$$\mathbb{P}(\{D_i\}_{1\leq i\leq R}|\{Z_i\}_{1\leq i\leq R}) = \prod_{i=1}^{R}\prod_{j=1}^{N}\alpha_{Z_i,j}^{D_{ij}}(1-\alpha_{Z_i,j})^{1-D_{ij}}.$$

Many different techniques may be used to approximate the maximum likelihood estimator in hidden Markov random fields. The NEM algorithm is based on a mean-field approximation for the distribution of the latent random variables $\{Z_i\}_{1\leq i\leq R}$ conditional on the observations. The algorithm is fully described in (Dang et Govaert, 1998). Note that the parameter β may be either fixed to a default value or optimised within the algorithm. We tested the two different approaches, setting the default value to $\beta = 1$ or optimising it, and they gave similar results.

3.2.4 Results and discussion

The results presented and discussed here are based on the terminology given in Table 3.7. As mentioned in the *Dataset description*, our goal is to compare two groups with similar phylogenetic distance and distinct in terms of ecological niche. For a similar phylogenetic distance, we worked at the level of the genus of the two selected groups, being however aware of the limitations of such a selection since there is no perfect standard definition at this taxonomic level. The group presenting a wider ecological niche is the one of the *Pseudomonas* genus which is found in association with different hosts as well as in the soil, in fresh and in waste water. We selected all available *Pseudomonas* species present in MicroCyc (Vallenet *et al.*, 2009), that is a total of 13 organisms (for a detailed information, refer to Section 3.2.2). The group with a narrow ecological niche that was chosen is one composed of *Escherichia/Shigella* species that mainly colonise the human intestinal microflora for which there was a higher number of strains of *E. coli* available. In order to have exactly the same number of organisms as the dataset of the *Pseudomonas* genus, we kept all other species of *Escherichia* (*i.e. E. albertii* and *E. fergusonii*), and we restricted the number of *E. coli* strains to have a wide diversity based on the phylogroups and types of host interaction.

In these preliminary analyses, we focused on comparing these two datasets aiming to find important differences in terms of ecological niche. In this sense, we compared the sizes of their core and periphery with the idea that a larger core might indicate less diverse metabolic capabilities and a narrow ecological niche. We also investigated the content of those sets of reactions by analysing the metabolic processes in which these reactions are involved. In all these analyses, we focused on the differences between the proposed extended core and the traditional one. Moreover, we compared our findings, especially the group of *Escherichia/Shigella* organisms, to the work of Vieira *et al.* (2011) where the authors analysed the core metabolism of 29 strains *E. coli* and *Shigella*. They used for this the traditional definition of core as opposed to the *variable metabolism* that is the group containing strain-specific or group-specific reactions which would roughly correspond to our periphery.

Table 3.7: **Terminology for the groups presented hereafter.**

Group	Meaning
traditionalCore	Rections present in all organisms
extendedCore	BinEm and Nem* place these reactions in the core
neighbourCore	Reactions added in the core based on neighbour relationship, *i.e.* BinEm classifies them in the periphery while Nem places them in the core
neighbourPeriphery	opposite of neighbourCore, *i.e.* BinEm places them in the core while Nem classify them in the periphery
periphery	both BinEm and Nem place them in the periphery
Nem	in all analyses below we used the Nem parameter β set as half

*see Section 3.2.3 for description.

The *Escherichia/Shigella* dataset

The results of *Escherichia/Shigella* are presented in a comparative way to the investigation of Vieira *et al.* (2011). We first did a comparison of the organisms included in each study since this is key to determine which reactions are in the traditional core (*i.e.* are found in all organisms). The metabolic networks in our analyses were obtained from MicroCyc where Vieira *et al.* (2011) made available their genomic and metabolic data. Thus, our dataset of

Escherichia/Shigella organisms includes a subset of the 29 strains of *E. coli/Shigella* from Vieira *et al.* (2011), in addition to the two other species of *Escherichia*, namely *E. albertii* and *E. fergusonii*. Since we are studying a smaller set of organisms, our traditional core tends to be larger than the one presented by Vieira *et al.* (2011). On the other hand, the inclusion of the two extra species of *Escherichia*, which might have more distinct metabolic capabilities, may eventually reduce the set of common reactions.

Vieira *et al.* (2011) found 885 reactions belonging to the traditional core metabolism (57% of the total number of 1545 reactions) and 660 reactions belonging to the variable metabolism (43% of the total number of reactions). Our results for the *Escherichia* dataset are shown in Table 3.8 for which we found a slightly larger total number of reactions included in the traditional core (62%). This variation in the size of the traditional core is somehow expected since this set is quite unstable depending on the addition or removal of one organism. We included about 15% of the total number of reactions to the core based on our first proposed approach (BinEm) to which were added 6% more of the total number of reactions with the second one (Nem), based on the results for the filtered metabolic networks. Our second method is currently not dealing well with the non-filtered network reducing considerably the periphery, which can be more clearly evidenced in the dataset of *Pseudomonas* that we will present in the next Section 3.10. Our results are therefore detailed only for the filtered networks.

Table 3.8: **Distribution of the reactions and the number of organisms where they appear for the *Escherichia* dataset.**

	***Escherichia* Filter**		***Escherichia* No Filter**	
	NbOrganisms	NbReactions	NbOrganisms	NbReactions
traditionalCore	13..13	1061	13..13	1086
extendedCore	9..12	250	9..12	261
sumTcoreEcore	9..13	1311	9..13	1347
neighbourCore	1..10	101	1..11	336
sumTotalcore	1..13	1412	1..13	1683
neighbourPeriphery		0		0
periphery	1..6	288	1..6	59
sumTotalPeriphery	1..6	288	1..6	59
TotalNbReactions	1..13	1700	1..13	1742

The core of *E. coli* strains is mainly composed of biosynthesis (57%) (Vieira *et al.*, 2011) rather than degradation based on the MetaCyc metabolic processes. The variable metabolism, on the other hand, is composed of similar proportions of biosynthesis and degradation processes (36% biosynthesis and 35% degradation) (Vieira *et al.*, 2011). The results on metabolic processes for the *Escherichia* dataset are presented in Table 3.9. The traditional core is well enriched in biosynthesis (46% of the total number of reactions in this group) while the periphery has more reactions involved in degradation processes (41%). Our two approaches included to the core more reactions involved in degradation than in biosynthesis as well as some transport reactions. Even including the reactions with our first and second approaches, the core remains with enhanced biosynthesis capabilities, with around 40% of biosynthesis and 20% of degradation processes.

Table 3.9: **Metabolic processes in which the reactions from each group are classified.**

	***Escherichia* Filter**					
	traditionalCore	extendedCore	*sumTcoreEcore*	neighbourCore	*sumTotalCore*	periphery
A-I-I*;Biosynthesis	2		2		2	3
Biosynthesis	340	19	359	8	367	45
Biosynthesis;Energy-Metabolism	143	35	178	12	190	17
Total Biosynthesis	**485 (46%)**	**54 (22%)**	**540 (41%)**	**20 (20%)**	**560 (40%)**	**65 (22%)**
A-I-I*;Degradation	2	2	4	1	5	
Degradation	118	92	210	41	251	111
Degradation;Detoxification	2	1	3		3	1
Degradation;Energy-Metabolism	20	8	28	2	30	6
Total Degradation	**142 (13%)**	**103 (41%)**	**245 (19%)**	**44 (44%)**	**289 (20%)**	**118 (41%)**
A-I-I*;Biosynthesis;Degradation	9		9		9	1
Biosynthesis;Degradation	47	9	56	2	58	12
Biosynthesis;Degradation;Detoxification	1		1		1	
Biosynthesis;Degradation;Energy-Metabolism	24	1	25	1	26	3
Total Biosynthesis;Degradation	**81 (8%)**	**10 (4%)**	**91 (7%)**	**3 (3%)**	**94 (7%)**	**16 (5%)**
Cofactor	33	3	36	4	40	8
Detoxification	3		3		3	1
Energy-Metabolism	17	1	18	1	19	4
Transport	151	49	200	17	217	18
No pathway assigned	149	30	179	12	191	58
Total number of reactions	**1061**	**250**	**1311**	**101**	**1412**	**288**

A-I-I stands for Activation-Inactivation-Interconversion.

The *Pseudomonas* dataset and the comparison of the two datasets

The results for the *Pseudomonas* group are presented in Table 3.10. The wider ecological niche, and consequently wider also metabolic diversity appears clearly in the results when compared to the *Escherichia* dataset. This can be evidenced by the smaller number of reactions included in the traditional core (28% of the total number of reactions), the extended core (48%) and the neighbour core (60%) of the *Pseudomonas* dataset, when compared to the corresponding groups of the *Escherichia* dataset. The smaller number of reactions in the core metabolism, and consequently higher number of reactions included in the periphery, indicate more organism or group-specific reactions and less common metabolic capabilities. The reactions are more evenly distributed between the core and the periphery (40%), almost half of the reactions in each group. Since these *Pseudomonas* species live in distinct and possibly more diverse environments, their metabolic capabilities are as well more variable from one organism to another. The metabolic processes in which the reactions are involved are shown in Table 3.11. The traditional and the extended cores are enriched in biosynthesis while the periphery presents similar amounts of both (30% biosynthesis and 34% degradation). The reactions added to the traditional core by our method are slightly more frequently involved in biosynthesis than in degradation processes.

Table 3.10: **Distribution of reactions and the number of organisms which they appear for *Pseudomonas* dataset.**

	***Pseudomonas* Filter**		***Pseudomonas* No Filter**	
	NbOrganisms	NbReactions	NbOrganisms	NbReactions
traditionalCore	13..13	569	13..13	585
extendedCore	8..12	388	9..12	391
sumTcoreEcore	8..13	957	9..13	976
neighbourCore	1..9	251	1..9	1016
sumTotalcore	1..13	1208	1..13	1992
neighbourPeriphery		0	9..13	17*
periphery	1..7	801	1..7	54
sumTotalPeriphery	1..7	801	1..13	71
TotalNbReactions	1..13	2009	1..13	2063

*10 reactions are present in 13 organisms.

Table 3.11: **Metabolic processes in which the reactions from each group are classified.**

	Pseudomonas Filter					
	extendedCore	traditionalCore	*sumTcoreEcore*	neighbourCore		periphery
A-I-I*	1		1		1	1
A-I-I*;Biosynthesis	1	1	2	2	4	
Biosynthesis	263	105	268	59	327	181
Biosynthesis;Energy-Metabolism	29	28	57	25	79	58
Total Biosynthesis	**293 (51%)**	**134 (34%)**	**427 (%)**	**86 (34%)**	**513 (42%)**	**239 (30%)**
A-I-I*;Degradation	2	2	4		4	2
Degradation	83	94	177	67	144	259
Degradation;Detoxification	1	2	3	1	4	
Degradation;Energy-Metabolism	13	8	21	3	24	12
Total Degradation	**98 (17%)**	**106 (27%)**	**204 (%)**	**71 (28%)**	**275 (23%)**	**273 (34%)**
A-I-I*;Biosynthesis;Degradation	5	3	8	1	9	3
Biosynthesis;Degradation	38	21	59	11	70	36
Biosynthesis;Degradation;Detoxification	1		1		1	
Biosynthesis;Degradation;Energy-Metabolism	21	2	23	5	28	1
Total Biosynthesis;Degradation	**65 (11%)**	**26 (7%)**	**91 (%)**	**17 (7%)**	**108 (9%)**	**40 (5%)**
Cofactor	22	13	35	4	39	13
Detoxification	3		3	1	4	5
Energy-Metabolism	11	9	20	1	21	10
Transport	2	10	12	7	19	7
No pathway assigned	73	90	163	64	227	213
Total number of reactions	**569**	**388**	**957**	**251**	**1208**	**801**

A-I-I stands for Activation-Inactivation-Interconversion.

3.2.5 Conclusion and perspectives

The proposed approaches already give some insights on the structure and content of the common metabolic capabilities, highlighting differences between the two datasets chosen. The wider ecological niche of the *Pseudomonas* species can be evidenced in our preliminary results. It is important to note that we are still working on the method in order to tune the algorithm that takes into account the neighbour relationship to allow the use of non filtered networks. The reactions included in the proposed extended core will be further analysed as concerns completeness of the metabolic routes and alternative pathways. We will investigate the connectivity of the extended and neighbour core to check if the common reactions are connected among themselves, *i.e.* the produced metabolites are consumed by other reactions in a chain of biochemical transformations, perhaps forming an entirely connected core network.

Chapter 4

Exploring metabolomics data

Contents

4.1 Overview 125
4.2 Metabolic stories 125
4.3 Yeast response to cadmium exposure 126
4.4 Perspectives 128

4.1 Overview

The increasing availability of metabolomics data enables to better understand the metabolic processes involved in the immediate response of an organism to environmental changes and stress. The data usually comes in the form of a list of metabolites whose concentrations significantly changed under some conditions, and are thus not easy to interpret without being able to precisely visualise how such metabolites are interconnected. We present a method that enables to organise the data from any metabolomics experiment into what we called *metabolic stories*. We detail an application of the method to the response of yeast to cadmium exposure. This closing chapter of the thesis is entirely dedicated to this investigation, for which my main contribution comprises the analyses and interpretation of the application of the method to the response of yeast to cadmium exposure. This study resulted in the following manuscript: Milreu *et al.*, Telling metabolic stories to explore metabolomics data – A case study on the Yeast response to cadmium exposure, *Bioinformatics (Oxford, England)*, 30(1):61–70, 2014. A brief description of this investigation will be presented hereafter and a complete version of the paper can be found in the Appendix D.

4.2 Metabolic stories

Each story corresponds to a possible scenario explaining the flow of matter between the metabolites of interest. These scenarios may then be ranked in different ways depending on which interpretation one wishes to emphasise for the causal link between two affected metabolites: enzyme activation, enzyme inhibition, or domino effect on the concentration changes of substrates and products. Equally probable stories under any selected ranking scheme can be further grouped into a single anthology that summarises, in a unique subnetwork, all equivalently plausible alternative stories.

4.3 Yeast response to cadmium exposure

In order to illustrate how to use our method, we concentrated on the study of the exposition of *Saccharomyces cerevisiae* to the toxic cadmium (Cd2+) reported in (Madalinski *et al.*, 2008). A widely studied metabolic pathway in yeast is the one responsible for glutathione biosynthesis (Figure 4.1), since it is related to the detoxification process of the cell when exposed to high concentrations of cadmium (Fauchon *et al.*, 2002; Lafaye *et al.*, 2005; Madalinski *et al.*, 2008). Previous studies demonstrated that the presence of such a metal in the environment has a huge impact in terms of gene expression and metabolism, showing that there is a strong response both at the metabolomic and proteomic levels. Basically, glutathione needs to be produced because it is a thiol metabolite linked to the detoxification of cadmium through a process called chelation (Li *et al.*, 1997).

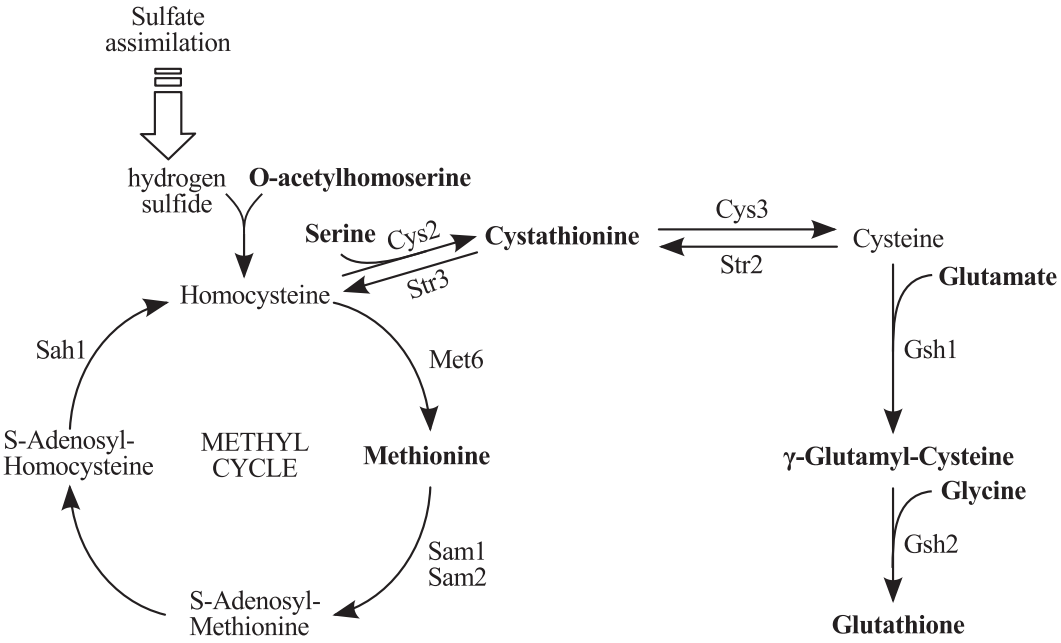


Figure 4.1: Glutathione biosynthetic pathway. Compounds in bold are discriminating in (Madalinski *et al.*, 2008) and are involved in the synthesis of glutathione. **Adapted from Figure 1 in** Lafaye *et al.* (2005).

Using only the metabolomics experiment data reported in (Madalinski *et al.*, 2008) to choose the discriminating compounds (i.e., the set of metabolites identified in the metabolomic experiment that have significantly changed their concentration) and to rank the stories, we are able to obtain stories that correspond very well to the current biological understanding of the system under study, as well as to propose new alternatives that could serve as a basis for further experimental validations. Such results can be illustrated by the anthology corresponding to the 20 stories with the maximal score computed (Figure 4.2), where the reactions corresponding to the glutathione biosynthesis are highlighted in grey. This is a strong point of our method since it allows exploring alternative but close scenarios through the analysis of these (and possibly other) stories altogether, which might provide new insights on the underlying processes that took place under the given conditions.

We discuss several interpretations for the changes we see and we suggest hypotheses which could in principle be experimentally tested. Such interpretations include the link between

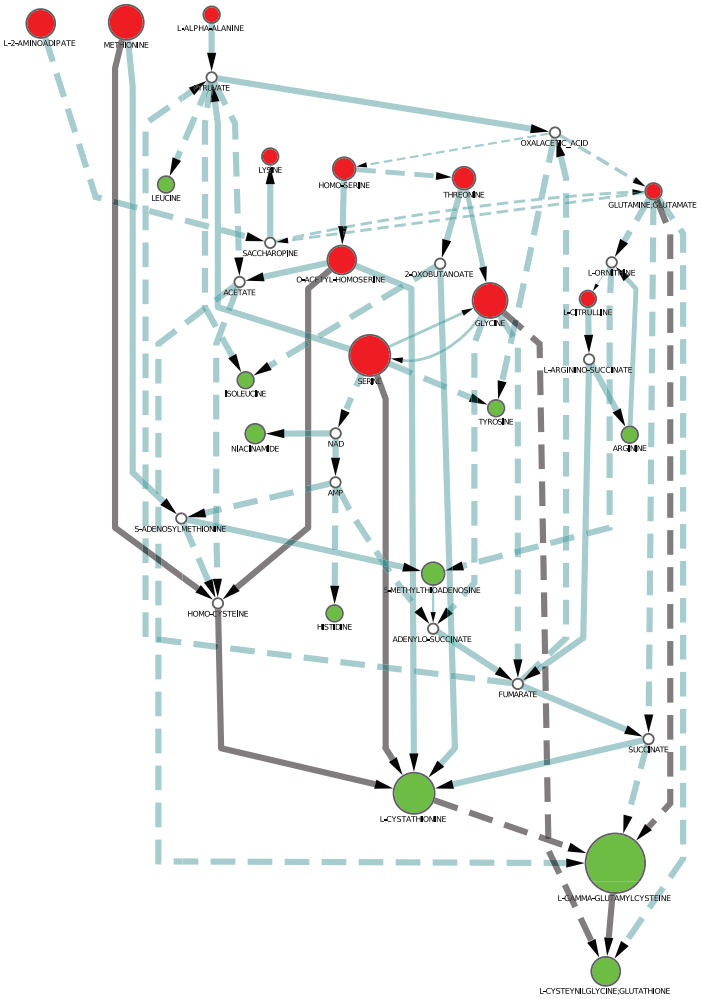


Figure 4.2: Anthology corresponding to the 20 stories with the maximal score computed for the experiment on yeast s288c exposed to cadmium. Red nodes correspond to metabolites whose concentration decreased and green nodes to those whose concentration increased in the metabolomics experiment. White nodes have their concentration unchanged or it could not be measured. The diameter of the nodes is proportional to the concentration change. Solid arcs represent single reactions connecting the two compounds while dashed ones correspond to a chain of at least two reactions. The arcs thickness represents the frequency of the arc in the stories making up the anthology while grey arcs correspond to reactions known to be part of the response to cadmium.

arginine to sulfur metabolism. By using the metabolic stories based approach, the increased levels of arginine may be related to decreased concentrations of citrulline, which has not been formally identified in our experimental conditions, and which is itself linked to glutamate. Besides, citrulline was identified as a discriminating compound in (Madalinski *et al.*, 2008), but was only indicated as putative, requiring more analysis for final identification. Our results seem to confirm that citrulline was correctly identified. This emphasizes the relevance of using this kind of approach to generate biological hypotheses that have to be further investigated by biologists. Of note, such a link between arginine and sulfur metabolism has been noticed in other organisms (Sekowska *et al.*, 2001) and links between nitric oxide and

polyamines have been established with cadmium toxicity in wheat roots (Groppa *et al.*, 2008). Furthermore, this global view of the discriminating compounds links the sulfur metabolism to non-sulfur amino acids and other metabolites through intermediates of the central metabolism. The amino acids that are precursors to the glutathione synthesis have their levels reduced as expected, whereas most of the others increased. This agrees with the fact that global protein synthesis rapidly drops after cadmium exposure (Lafaye *et al.*, 2005), reducing the consumption of amino acids not directly connected to glutathione synthesis.

4.4 Perspectives

We presented a generic method which enables to analyse metabolomics data. This method requires very simple input and can be applied to a wide variety of situations. Together with other omics data, analysis of metabolomics experiments is essential to further learn about the large unexplored portion of the metabolic map, which includes organism-specific and condition-specific activities (Breitling *et al.*, 2008). Moreover, the interaction of symbiotic partners is currently under-explored as concerns metabolomics analyses, which may give important insights on the establishment and maintenance of those associations. One question, that could possibly be explored and analysed using the method presented, is whether the host regulates the gene expression of the symbiont through larger or smaller amounts of compounds provided to the symbiont. Thus, the development of methods that allow for the extraction of such knowledge from whole metabolic networks plays an important role in the future investigations concerning metabolism as well as the metabolic complementarity in symbiotic associations.

Conclusion and Perspectives

In this thesis, we presented three main types of analyses of metabolism, most of which involved symbiosis: metabolic dialogue between a trypanosomatid and its symbiont, comparative analyses of metabolic networks and exploration of metabolomics data. All of them were essentially based on genomics data where metabolic capabilities were predicted from the annotated genes of the target organism, and were further refined with other types of data depending on the aim and scope of each investigation. In addition to genomics, the last study presented in this thesis focused on metabolomics data which were mapped into the genome-scale metabolic network of the target organism and metabolic stories were then extracted through our method which thus provides an approach to treat metabolomics data and give insights into the metabolism of the organism of interest in some condition-specific situations such as a stress response.

The metabolic dialogue between a trypanosomatid and its symbiont was originally the main topic of this thesis giving it its title and was first investigated by means of the classically defined metabolic pathways. The selected routes corresponded to the biosynthesis of essential amino acids and vitamins for which there were nutritional data indicating that the symbiotic bacterium had an important contribution. Five pairs of trypanosomatids and endosymbionts were investigated. Most of the genes coding for the enzymes involved in such processes were found in the endosymbionts suggesting some intricate metabolic exchanges between the bacterium and its host protozoan. Based on genomic data, we were able to indicate the potential metabolic contributions of the endosymbiont rendering the host protozoan less nutritional exigent when compared to the trypanosomatids that do not harbour a bacterium in their cytoplasm. Assuming the massive gene transfer from symbiont-derived organelles to the host nucleus, we investigated such an event in our symbiotic partners using phylogenetic analyses. The fewer genes involved in the synthesis of amino acids and vitamins found in the host genomes possibly show an important influence of bacterial genes horizontally transferred to the trypanosomatid, however those genes are closer to other bacterial lineages than the one of the betaproteobacterial endosymbiont. Thus, the pattern of massive gene transfer found in symbiont-derived organelles does not seem to take place in the analysed symbiosis of trypanosomatids, at least as concerns those metabolic routes. We are now reconstructing and refining the genome-scale metabolic models of a pair of host trypanosomatid and bacterial endosymbiont with an aim to investigate the minimal sets of metabolites exchanged by both partners. We plan to have transcriptomic and experimental data that will be important for the manual refinement of the model which may allow to perform model-driven simulations and predictions. Still concerning metabolic exchanges, the integration of metabolomics data in our model might give initial insights about the levels of metabolites exchanged in the different stages of the life cycle of the trypanosomatid, and if such levels are related to the regulation of gene expression in the bacterium. We will therefore continue studying such symbiotic associations in the light of the evolutionary perspective of cellular evolution and the transition from symbionts to organelles.

The comparative analyses performed, and still ongoing, focused on the common metabolic capabilities of different lifestyle groups of bacteria. The first investigation presented in this thesis aimed at exploring the metabolism of symbiotic bacteria at different levels of integration with the host, including the cellular location and the genome reduction associated to such level of integration. We confirmed the strong impact of the inclusion of a single organism in the size and content of the common metabolic capabilities, and how such a set tends to be very reduced or empty as more genomes are sequenced and included in the analyses. In that sense, our second comparative study focused on a method to automatically establish the common and the group-specific activities. Such approach is based on the pattern of presence/absence of reactions in each organism and may also include information on the neighbour relationships in a metabolic network. In this second work, we selected the dataset to represent diverse ecological niches that are host association, either pathogenic or commensal, or free-living. These results are still preliminary and the goal is to have an approach that deals with non filtered metabolic networks and that propose a pertinent common set of metabolic capabilities not requiring: (i) omnipresence of a reaction and (ii) manual setting of a threshold for the number of organisms in which the reaction is present to classify it either in the common or in the group-specific set of metabolic activities. We will therefore continue developing this method and conducting the associated analyses in collaboration with the statisticians who are our partners.

The application of our method on metabolic stories enumeration to the yeast response to cadmium exposure was a validation of this approach on a well-studied biological response to stress. The purpose was to show that the method captured well the underlying knowledge as it extracts stories allowing for further interpretations of the metabolomics data mapped into the genome-scale metabolic model of yeast. This method requires a simple input and can be applied to a wide variety of situations such as the comparison of aposymbiotic and wild strains of trypanosomatids to investigate the metabolic complementarity of these symbiotic associations.

The analyses herein presented were enabled by the recent high-throughput technologies for which it is essential to develop methods and approaches on how to extract useful information, perform simulations and predictions and draw conclusions in order to treat such enormous amount of information. Comparative analyses and propagation of model species knowledge have been currently largely used and explored, stressing the importance of improving such approaches in order to reduce as much as possible their limitations. Being able to deal with genome-scale metabolic models and having the possibility to integrate other omics data such as metabolomics is quite promising to advance on new findings concerning the under-explored parts of metabolism such as condition and organism-specific metabolic capabilities. In that sense, both experimental and computational advances, possibly in an iterative manner, will play key roles in the future research concerning metabolism, regulation and the evolution of symbiotic associations. We therefore believe to have contributed to the research field on symbiosis with our findings on the metabolic complementarity and horizontal gene transfers in such a unique model where a single symbiotic bacterium divides synchronously with the host protozoan nucleus implying a strict control over the endosymbiont division. Such control of cellular division is found in organelles and is important for the persistence of the interaction between the symbiotic partners. In addition to that, we proposed, applied to biological data and analysed methods to explore common metabolic capabilities and to treat metabolomics data. Such approaches may give valuable insights on metabolic properties and capabilities as we presented throughout this thesis.

Bibliography

Adams, D. W. et Errington, J. (2009). Bacterial cell division: assembly, maintenance and disassembly of the Z ring. *Nature reviews. Microbiology*, 7(9):642–53.

Adams, J. M. (1968). On the release of the formyl group from nascent protein. *Journal of Molecular Biology*, 33(3):571–574.

Akman, L., Yamashita, A., Watanabe, H., Oshima, K., Shiba, T., Hattori, M. et Aksoy, S. (2002). Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nature genetics*, 32(3):402–407.

Aktas, M., Wessel, M., Hacker, S., Klüsener, S., Gleichenhagen, J. et Narberhaus, F. (2010). Phosphatidylcholine biosynthesis and its significance in bacteria interacting with eukaryotic cells. *European journal of cell biology*, 89(12):888–94.

Alfieri, S. C. et Camargo, E. P. (1982). Trypanosomatidae: isoleucine requirement and threonine deaminase in species with and without endosymbionts. *Experimental parasitology*, 53(3):371–380.

Allman, E. S., Matias, C. et Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132.

Alves, J. a. M., Voegtly, L., Matveyev, A. V., Lara, A. M., da Silva, F. M. M., Serrano, M. G., Buck, G. A., Teixeira, M. M. et Camargo, E. P. (2011). Identification and phylogenetic analysis of heme synthesis genes in trypanosomatids and their bacterial endosymbionts. *PloS one*, 6(8).

Alves, J. M. et Buck, G. A. (2007). Automated system for gene annotation and metabolic pathway reconstruction using general sequence databases. *Chemistry & biodiversity*, 4(11): 2593–2602.

Alves, J. M., Klein, C. C., da Silva, F. M., Costa-Martins, A. G., Serrano, M. G., Buck, G. A., Vasconcelos, A. T. R., Sagot, M.-F., Teixeira, M. M. G., Motta, M. C. M. et Camargo, E. P. (2013a). 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*, 13(1):190+.

Alves, J. M., Serrano, M. G., Maia da Silva, F., Voegtly, L. J., Matveyev, A. V., Teixeira, M. M., Camargo, E. P. et Buck, G. A. (2013b). Genome evolution and phylogenomic analysis of *Candidatus* kinetoplastibacterium, the betaproteobacterial endosymbionts of *Strigomonas* and *Angomonas*. *Genome biology and evolution*, 5(2):338–350.

Ambroise, C., Dang, M. V. et Govaert, G. (1997). Clustering of spatial data by the EM algorithm. *In* Soares, A. Gómez-Hernandez, J. et Froidevaux, R., éditeurs : *geoENV I - Geostatistics for Environmental Applications*, volume 9 de *Quantitative Geology and Geostatistics*, pages 493–504. Kluwer Academic Publisher.

Ambroise, C. et Govaert, G. (1998). Convergence proof of an EM-type algorithm for spatial clustering. *Pattern Recognition Letters*, 19:919–927.

Anderson, S. J. et Krassner, S. M. (1975). Axenic culture of *Trypanosoma cruzi* in a chemically defined medium. *The Journal of parasitology*, 61(1):144–145.

Andersson, J. O. (2009). Horizontal gene transfer between microbial eukaryotes. *Methods in molecular biology (Clifton, N.J.)*, 532:473–487.

Andersson, S. G. et Kurland, C. G. (1998). Reductive evolution of resident genomes. *Trends in microbiology*, 6(7):263–268.

Azuma, Y. et Ota, M. (2009). An evaluation of minimal cellular functions to sustain a bacterial cell. *BMC systems biology*, 3(1):111+.

Bacher, A., Eberhardt, S., Eisenreich, W., Fischer, M., Herz, S., Illarionov, B., Kis, K. et Richter, G. (2001). Biosynthesis of riboflavin. *Vitamins and hormones*, 61:1–49.

Barreto-de-Souza, V., Medeiros, T. X., Motta, M. C. M., Bou-Habib, D. C. et Saraiva, E. M. (2008). HIV-1 infection and HIV-1 Tat protein permit the survival and replication of a non-pathogenic trypanosomatid in macrophages through TGF-beta1 production. *Microbes and infection / Institut Pasteur*, 10(6):642–9.

Barve, A., Rodrigues, J. F. M. F. et Wagner, A. (2012). Superessential reactions in metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, 109(18):E1121–E1130.

Baty, F. et Delignette-Muller, M.-L. (2011). *nlstools: tools for nonlinear regression diagnostics*. R package version 0.0-11.

Baumann, P., Baumann, L., Lai, C. Y., Rouhbakhsh, D., Moran, N. A. et Clark, M. A. (1995). Genetics, physiology, and evolutionary relationships of the genus *Buchnera*: intracellular symbionts of aphids. *Annual review of microbiology*, 49(1):55–94.

Baumann, P., Moran, N. A. et Baumann, L. (1997). The evolution and genetics of aphid endosymbionts. *BioScience*, 47(1):12–20.

Begley, T., Kinsland, C., Taylor, S., Tandon, M., Nicewonger, R., Wu, M., Chiu, H.-J., Kelleher, N., Campobasso, N. et Zhang, Y. (1998). Cofactor biosynthesis: A mechanistic perspective. *In* Leeper, F. et Vederas, J., éditeurs : *Biosynthesis*, volume 195 de *Topics in Current Chemistry*, pages 93–142. Springer Berlin Heidelberg.

Begley, T. P., Kinsland, C., Mehl, R. A., Osterman, A. et Dorrestein, P. (2001a). The biosynthesis of nicotinamide adenine dinucleotides in bacteria. *Vitamins and hormones*, 61:103–119.

Begley, T. P., Kinsland, C. et Strauss, E. (2001b). The biosynthesis of coenzyme a in bacteria. *Vitamins and hormones*, 61:157–171.

Bennett, G. M. et Moran, N. A. (2013). Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a phloem-feeding insect. *Genome biology and evolution*, 5(9): 1675–1688.

Bernander, R. et Ettema, T. J. (2010). FtsZ-less cell division in archaea and bacteria. *Current opinion in microbiology*, 13(6):747–52.

Beutin, L. et Eisen, H. (1983). Regulation of enzymes involved in ornithine/arginine metabolism in the parasitic trypanosomatid *Herpetomonas samuelpessoai*. *Molecular & general genetics : MGG*, 190(2):278–283.

Bhattacharjee, J. K. (1985). alpha-Aminoadipate pathway for the biosynthesis of lysine in lower eukaryotes. *Critical reviews in microbiology*, 12(2):131–151.

Bono, H., Ogata, H., Goto, S. et Kanehisa, M. (1998). Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. *Genome research*, 8(3):203–210.

Borenstein, E., Kupiec, M., Feldman, M. W. et Ruppin, E. (2008). Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proceedings of the National Academy of Sciences of the United States of America / PNAS*, 105(38):14482–14487.

Borghesan, T. C., Ferreira, R. C., Takata, C. S. a., Campaner, M., Borda, C. C., Paiva, F., Milder, R. V., Teixeira, M. M. G. et Camargo, E. P. (2013). Molecular phylogenetic redefinition of *Herpetomonas* (Kinetoplastea, Trypanosomatidae), a genus of insect parasites associated with flies. *Protist*, 164(1):129–52.

Breitling, R., Vitkup, D. et Barrett, M. P. (2008). New surveyor tools for charting microbial metabolic maps. *Nature reviews. Microbiology*, 6(2):156–61.

Bringaud, F., Barrett, M. P. et Zilberstein, D. (2012). Multiple roles of proline transport and metabolism in trypanosomatids. *Frontiers in bioscience (Landmark edition)*, 17:349–374.

Bringaud, F., Rivière, L. et Coustou, V. (2006). Energy metabolism of trypanosomatids: adaptation to available carbon sources. *Molecular and biochemical parasitology*, 149(1):1–9.

Brooker, B. (1971). *Fine Structure of Bodo Saltans and Bodo Caudatus Zoomastigophora: Protozoa and Their Affinities with the Trypanosomatidae*. Bulletin of the British Museum Nat. History. Zoology. Trustees of the Brit. Mus.

Camargo, E. P. (1964). Growth and differentiation in *Trypanosoma cruzi*. i. origin of metacyclic *Trypanosomes* in liquid media. *Revista do Instituto de Medicina Tropical de São Paulo*, 6:93–100.

Camargo, E. P., Coelho, J. A., Moraes, G. et Figueiredo, E. N. (1978). *Trypanosoma* spp., *Leishmania* spp. and *Leptomonas* spp.: enzymes of ornithine-arginine metabolism. *Experimental parasitology*, 46(2):141–144.

Camargo, E. P. et Freymuller, E. (1977). Endosymbiont as supplier of ornithine carbamoyltransferase in a trypanosomatid. *Nature*, 270(5632):52–53.

Camargo, E. P., Silva, S., Roitman, I., de Souza, W., Jankevicius, J. V. et Dollet, M. (1987). Enzymes of Ornithine-Arginine metabolism in trypanosomatids of the genus *Phytomonas*. *Journal of Eukaryotic Microbiology*, 34(4):439–441.

Campbell, D., Thomas, S. et Sturm, N. R. (2003). Transcription in kinetoplastid protozoa: why be normal? *Microbes and Infection*, 5(13):1231–1240.

Carreira-Perpiñán, M. A. et Renals, S. (2000). Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Comp.*, 12(1):141–152.

Cases, I., de Lorenzo, V. et Ouzounis, C. A. (2003). Transcription regulation and environmental adaptation in bacteria. *Trends in microbiology*, 11(6):248–253.

Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Pujar, A., Shearer, A. G., Travers, M., Weerasinghe, D., Zhang, P. et Karp, P. D. (2012). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research*, 40(Database issue):D742–D753.

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, A. G., Tissier, C., Walk, T. C., Zhang, P. et Karp, P. D. (2008). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucl. Acids Res.*, 36(suppl_1):D623–631.

Cavalcanti, D. P., Thiry, M., de Souza, W. et Motta, M. C. M. (2008). The kinetoplast ultrastructural organization of endosymbiont-bearing trypanosomatids as revealed by deep-etching, cytochemical and immunocytochemical analysis. *Histochemistry and cell biology*, 130(6):1177–85.

Chang, K. P. (1975). Reduced growth of *Blastocrithidia culicis* and *Crithidia oncopelti* freed of intracellular symbiotes by chloramphenicol. *The Journal of protozoology*, 22(2):271–276.

Chang, K. P., Chang, C. S. et Sassa, S. (1975). Heme biosynthesis in bacterium-protozoon symbioses: enzymic defects in host hemoflagellates and complemental role of their intracellular symbiotes. *Proceedings of the National Academy of Sciences*, 72(8):2979–2983.

Chang, K. P. et Trager, W. (1974). Nutritional significance of symbiotic bacteria in two species of hemoflagellates. *Science (New York, N.Y.)*, 183(124):531–532.

Charlebois, R. L. et Doolittle, W. F. (2004). Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome research*, 14(12):2469–2477.

Chaudhuri, R. R. et Henderson, I. R. (2012). The evolution of the escherichia coli phylogeny. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 12(2):214–226.

Chavali, A. K., Whittemore, J. D., Eddy, J. a., Williams, K. T. et Papin, J. a. (2008). Systems analysis of metabolism in the pathogenic trypanosomatid Leishmania major. *Molecular systems biology*, 4(177):177.

Chen, N., del Val, I. J., Kyriakopoulos, S., Polizzi, K. M. et Kontoravdi, C. (2012). Metabolic network reconstruction: advances in in silico interpretation of analytical information. *Current opinion in biotechnology*, 23(1):77–82.

Cornish-Bowden, A. (2004). *The Pursuit of Perfection: Aspects of Biochemical Evolution*. Oxford University Press.

Corrêa-da-silva, M. S., Fampa, P. et Motta, M. C. M. (2006). Colonization of *Aedes aegypti* midgut by the endosymbiont-bearing trypanosomatid *Blastocrithidia culicis*. *Parasitology Research*, 99:384–391.

Cottret, L. et Jourdan, F. (2010). Graph methods for the investigation of metabolic networks in parasitology. *Parasitology*, 137(9):1393–1407.

Cottret, L., Wildridge, D., Vinson, F., Barrett, M. P., Charles, H., Sagot, M.-F. et Jourdan, F. (2010). Metexplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic acids research*, 38:W132-7.

Coulomb, S., Bauer, M., Bernard, D. et Marsolier-Kergoat, M.-C. C. (2005). Gene essentiality and the topology of protein interaction networks. *Proceedings. Biological sciences / The Royal Society*, 272(1573):1721–1725.

Coustou, V., Biran, M., Breton, M., Guegan, F., Rivière, L., Plazolles, N., Nolan, D., Barrett, M. P., Franconi, J.-M. et Bringaud, F. (2008). Glucose-induced remodeling of intermediary and energy metabolism in procyclic Trypanosoma brucei. *The Journal of biological chemistry*, 283(24):16342–54.

Covert, M. W. et Palsson, B. O. (2003). Constraints-based models: regulation of gene expression reduces the steady-state solution space. *Journal of theoretical biology*, 221(3): 309–325.

Cowperthwaite, J., Weber, M. M., Packer, L. et Hutner, S. H. (1953). Nutrition of *Herpetomonas* (*Strigomonas*) *culicidarum*. *Annals of the New York Academy of Sciences*, 56(5):972–981.

Cross, G. A., Klein, R. A. et Linstead, D. J. (1975a). Utilization of amino acids by *Trypanosoma brucei* in culture: L-threonine as a precursor for acetate. *Parasitology*, 71(2): 311–326.

Cross, G. A. M., Klein, R. A. et Baker, J. R. (1975b). *Trypanosoma cruzi*: growth, amino acid utilization and drug action in a defined medium. *Annals of Tropical Medicine and Parasitology*, 69(4):513–514.

Csardi, G. et Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Dale, J., Popescu, L. et Karp, P. (2010). Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics*, 11(1):15+.

Danchin, A. (1989). Homeotopic transformation and the origin of translation. *Progress in biophysics and molecular biology*, 54(1):81–86.

Danchin, A., Fang, G. et Noria, S. (2007). The extant core bacterial proteome is an archive of the origin of life. *Proteomics*, 7(6):875–889.

Dang, M. V. et Govaert, G. (1998). Spatial fuzzy clustering using EM and Markov random fields. *In International Journal of System Research and Information Science*, pages 183–202.

D'Avila-Levy, C. M., Santos, L. O., Marinho, F. A., Matteoli, F. P., Lopes, A. H. C. S., Motta, M. C. M. et Branquinha, M. H. (2008). *Crithidia deanei*: Influence of parasite gp63 homologue on the interaction of endosymbiont-harboring and aposymbiotic strains with *Aedes aegypti* midgut. *Experimental Parasitology*, 118:345–353.

D'Avila-Levy, C. M., Silva, B. a., Hayashi, E. a., Vermelho, A. B., Alviano, C. S., Saraiva, E. M. B., Branquinha, M. H. et Santos, A. L. S. (2005). Influence of the endosymbiont of *Blastocrithidia culicis* and *Crithidia deanei* on the glycoconjugate expression and on *Aedes aegypti* interaction. *FEMS microbiology letters*, 252(2):279–86.

de Azevedo-Martins, A. C., Frossard, M. L. L., de Souza, W., Einicker-Lamas, M. et Motta, M. C. M. C. (2007). Phosphatidylcholine synthesis in *Crithidia deanei*: the influence of the endosymbiont. *FEMS microbiology letters*, 275(2):229–236.

de Freitas-Junior, P. R. G., Catta-Preta, C. M. C., Andrade, I., Cavalcanti, D. P., de Souza, W., Einicker-Lamas, M. et Motta, M. C. (2012). Effects of miltefosine on the proliferation, ultrastructure, and phospholipid composition of *Angomonas deanei*, a trypanosomatid protozoan that harbors a symbiotic bacterium. *FEMS Microbiol Lett*, 333(2):129–137.

de Menezes, M. C. et Roitmanz, I. (1991). Nutritional requirements of *Blastocrithidia culicis*, a trypanosomatid with an endosymbiont. *Journal of Eukaryotic Microbiology*, 38(2): 122–123.

de Oliveira Dal'Molin, C. G. et Nielsen, L. K. (2013). Plant genome-scale metabolic reconstruction and modelling. *Current opinion in biotechnology*, 24(2):271–7.

de Souza, W. (2002). Special organelles of some pathogenic protozoa. *Parasitology research*, 88(12):1013–25.

de Souza, W. et da Cunha-e Silva, N. L. (2003). Cell fractionation of parasitic protozoa: a review. *Memórias do Instituto Oswaldo Cruz*, 98(2):151–70.

Degnan, P. H., Lazarus, A. B. et Wernegreen, J. J. (2005). Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects. *Genome research*, 15(8):1023–1033.

Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38.

Deschamps, P., Lara, E., Marande, W., López-García, P., Ekelund, F. et Moreira, D. (2011). Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids. *Molecular biology and evolution*, 28(1):53–58.

Docampo, R., de Souza, W., Miranda, K., Rohloff, P. et Moreno, S. N. J. (2005). Acidocalcisomes - conserved from bacteria to man. *Nature reviews. Microbiology*, 3(3):251–61.

Docampo, R., Jimenez, V., King-Keller, S., Li, Z.-h. et Moreno, S. N. J. (2011). The Role of Acidocalcisomes in the Stress Response of *Trypanosoma cruzi*. *Advances in parasitology*, 75:307–324.

Docampo, R. et Moreno, S. N. J. (2011). Acidocalcisomes. *Cell calcium*, 50(2):113–9.

Docampo, R., Ulrich, P. et Moreno, S. N. J. (2010). Evolution of acidocalcisomes and their role in polyphosphate storage and osmoregulation in eukaryotic microbes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1541): 775–84.

Donlin, M. J. (2009). Using the generic genome browser (GBrowse). *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 9.

Douglas, A. E. (2010). *The Symbiotic Habit*. Princeton University Press.

Douglas, A. E. et Raven, J. a. (2003). Genomes at the interface between bacteria and organelles. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 358(1429):5–17; discussion 517–8.

Dray, S. et Dufour, A. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4):1–20.

Drewke, C. et Leistner, E. (2001). Biosynthesis of vitamin b6 and structurally related derivatives. *Vitamins and hormones*, 61:121–155.

Du, Y., Maslov, D. A. et Chang, K. P. (1994a). Monophyletic origin of beta-division proteobacterial endosymbionts and their coevolution with insect trypanosomatid protozoa *Blastocrithidia culicis* and *Crithidia* spp. *Proceedings of the National Academy of Sciences of the United States of America*, 91(18):8437–8441.

Du, Y., McLaughlin, G. et Chang, K. P. (1994b). 16S ribosomal DNA sequence identities of beta-proteobacterial endosymbionts in three *Crithidia* species. *Journal of bacteriology*, 176(10):3081–3084.

Dunn, M. F., Niks, D., Ngo, H., Barends, T. R. et Schlichting, I. (2008). Tryptophan synthase: the workings of a channeling nanomachine. *Trends in biochemical sciences*, 33(6): 254–264.

Durot, M., Bourguignon, P.-Y. Y. et Schachter, V. (2009). Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS microbiology reviews*, 33(1):164–190.

Dwyer, D. M. et Chang, K. P. (1976). Surface membrane carbohydrate alterations of a flagellated protozoan mediated by bacterial endosymbiotes. *Proceedings of the National Academy of Sciences of the United States of America*, 73(3):852–856.

Dybvig, K., Zuhua, C., Lao, P., Jordan, D. S., French, C. T., Tu, A.-H. H. et Loraine, A. E. (2008). Genome of mycoplasma arthritidis. *Infection and immunity*, 76(9):4000–4008.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5(1):113+.

El-Sayed, N. M., Myler, P. J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renauld, H., Worthey, E. a., Hertz-Fowler, C., Ghedin, E., Peacock, C., Bartholomeu, D. C., Haas, B. J., Tran, A.-N., Wortman, J. R., Alsmark, U. C. M., Angiuoli, S., Anupama, A., Badger, J., Bringaud, F., Cadag, E., Carlton, J. M., Cerqueira, G. C., Creasy, T., Delcher, A. L., Djikeng, A., Embley, T. M., Hauser, C., Ivens, A. C., Kummerfeld, S. K., Pereira-Leal, J. B., Nilsson,

D., Peterson, J., Salzberg, S. L., Shallom, J., Silva, J. C., Sundaram, J., Westenberger, S., White, O., Melville, S. E., Donelson, J. E., Andersson, B., Stuart, K. D. et Hall, N. (2005). Comparative genomics of trypanosomatid parasitic protozoa. *Science (New York, N.Y.)*, 309(5733):404–9.

Eliot, A. C. et Kirsch, J. F. (2004). Pyridoxal phosphate enzymes: mechanistic, structural, and evolutionary considerations. *Annual review of biochemistry*, 73:383–415.

Fair, D. S. et Krassner, S. M. (1971). Alanine aminotransferase and aspartate aminotransferase in *Leishmania tarentolae*. *Journal of Eukaryotic Microbiology*, 18(3):441–444.

Fampa, P., Corrêa-da Silva, M. S., Lima, D. C., Oliveira, S. M., Motta, M. C. M. et Saraiva, E. M. (2003). Interaction of insect trypanosomatids with mosquitoes, sand fly and the respective insect cell lines. *International Journal for Parasitology*, 33(10):1019–1026.

Faria-e Silva, P. M., Attias, M. et de Souza, W. (2000). Biochemical and ultrastructural changes in *Herpetomonas roitmani* related to the energy metabolism. *Biology of the cell / under the auspices of the European Cell Biology Organization*, 92(1):39–47.

Faria e Silva, P. M., Solé-Cava, A. M., Soares, M. J., Motta, M. C., Fiorini, J. E. et de Souza, W. (1991). *Herpetomonas roitmani* (fiorini et al., 1989) n. comb.: a trypanosomatid with a bacterium-like endosymbiont in the cytoplasm. *The Journal of protozoology*, 38(5): 489–494.

Fauchon, M., Lagniel, G., Aude, J.-C., Lombardia, L., Soularue, P., Petat, C., Marguerie, G., Sentenac, A., Werner, M. et Labarre, J. (2002). Sulfur sparing in the yeast proteome in response to sulfur demand. *Molecular Cell*, 9:713–723.

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L. et Palsson, B. Ø. (2009). Reconstruction of biochemical networks in microorganisms. *Nature reviews. Microbiology*, 7(2):129–143.

Feldhaar, H., Straka, J., Krischke, M., Berthold, K., Stoll, S., Mueller, M. J. et Gross, R. (2007). Nutritional upgrading for omnivorous carpenter ants by the endosymbiont *Blochmannia*. *BMC biology*, 5:48.

Felsenstein, J. (1989). PHYLIP - phylogeny inference package (version 3.2). *Cladistics*, 5:164–166.

Figueiredo, E. N., Yoshida, N., Roitman, C. et Camargo, E. P. (1978a). Enzymes of ornithine-Arginine Metabolism of Trypanosomatids of the genus *Crithidia*. *The Journal of Protozoology*, 25(4):546–549.

Figueiredo, E. N., Yoshida, N., Roitman, C. et Camargo, E. P. (1978b). Enzymes of the Ornithine-Arginine metabolism of trypanosomatids of the genus *Crithidia*. *Journal of Eukaryotic Microbiology*, 25(4):546–549.

Fiorini, J. E. (1989). Três novas espécies de tripanosomatídeos de insetos isolados em alfenas, minas gerais, brasil. *Memórias do instituto Oswaldo Cruz*, 84(1):69–74.

Fong, S. S., Burgard, A. P., Herring, C. D., Knight, E. M., Blattner, F. R., Maranas, C. D. et Palsson, B. O. (2005). In silico design and adaptive evolution of escherichia coli for production of lactic acid. *Biotechnology and bioengineering*, 91(5):643–648.

Forsyth, R. A., Haselbeck, R. J., Ohlsen, K. L., Yamamoto, R. T., Xu, H., Trawick, J. D., Wall, D., Wang, L., Brown-Driver, V., Froelich, J. M., C, K. G., King, P., McCarthy, M., Malone, C., Misiner, B., Robbins, D., Tan, Z., Zhu Zy, Z.-y. Y., Carr, G., Mosca, D. A., Zamudio, C., Foulkes, J. G. et Zyskind, J. W. (2002). A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Molecular microbiology*, 43(6):1387–1400.

Foster, J., Ganatra, M., Kamal, I., Ware, J., Makarova, K., Ivanova, N., Bhattacharyya, A., Kapatral, V., Kumar, S., Posfai, J., Vincze, T., Ingram, J., Moran, L., Lapidus, A., Omelchenko, M., Kyrpides, N., Ghedin, E., Wang, S., Goltsman, E., Joukov, V., Ostrovskaya, O., Tsukerman, K., Mazur, M., Comb, D., Koonin, E. et Slatko, B. (2005). The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS biology*, 3(4).

Foster, J. M., Davis, P. J., Raverdy, S., Sibley, M. H., Raleigh, E. A., Kumar, S. et Carlow, C. K. (2010). Evolution of bacterial phosphoglycerate mutases: non-homologous isofunctional enzymes undergoing gene losses, gains and lateral transfers. *PloS one*, 5(10).

Francke, C., Siezen, R. J. et Teusink, B. (2005). Reconstructing the metabolic network of a bacterium from its genome. *Trends in microbiology*, 13(11):550–8.

Freilich, S., Spriggs, R. V., George, R. A., Al-Lazikani, B., Swindells, M. et Thornton, J. M. (2005). The complement of enzymatic sets in different species. *Journal of molecular biology*, 349(4):745–763.

Freymuller, E. et Camargo, E. P. (1981). Ultrastructural differences between species of trypanosomatids with and without endosymbionts. *The Journal of protozoology*, 28(2):175–182.

Frossard, M. L. L., Seabra, S. H. H., DaMatta, R. A. A., de Souza, W., de Mello, F. G. G. et Machado Motta, M. C. C. (2006). An endosymbiont positively modulates ornithine decarboxylase in host trypanosomatids. *Biochemical and biophysical research communications*, 343(2):443–449.

Gabaldón, T. et Huynen, M. A. (2007). From endosymbiont to host-controlled organelle: the hijacking of mitochondrial protein synthesis and metabolism. *PLoS computational biology*, 3(11).

Gabaldón, T., Peretó, J., Montero, F., Gil, R., Latorre, A. et Moya, A. (2007). Structural analyses of a hypothetical minimal metabolism. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 362(1486):1751–1762.

Gabelli, S. B., Bianchet, M. A., Xu, W., Dunn, C. A., Niu, Z.-D. D., Amzel, L. M. et Bessman, M. J. (2007). Structure and function of the *E. coli* dihydroneopterin triphosphate pyrophosphatase: a nudix enzyme involved in folate biosynthesis. *Structure (London, England : 1993)*, 15(8):1014–1022.

Gadelha, C., Wickstead, B., Souza, W. D., Gull, K. et Cunha-e silva, N. (2005). Cryptic Paraflagellar Rod in Endosymbiont-Containing Kinetoplastid Protozoa. *Eukaryotic cell*, 4(3):516–525.

Galinari, S. et Camargo, E. P. (1978). Trypanosomatid protozoa: survey of acetylornithinase and ornithine acetyltransferase. *Experimental parasitology*, 46(2):277–282.

Galinari, S. et Camargo, E. P. (1979). Urea cycle enzymes in wild and aposymbiotic strains of *Blastocrithidia culicis*. *Journal of Parasitology*, 5(1):88+.

Galperin, M. Y., Walker, D. R. et Koonin, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome research*, 8(8):779–790.

Galvez Rojas, R. L. L., Frossard, M. L. L., Machado Motta, M. C. C. et Silber, A. M. M. (2008). L-Proline uptake in *Crithidia deanei* is influenced by its endosymbiont bacterium. *FEMS microbiology letters*, 283(1):15–22.

Gao, F. et Zhang, R. R. R. (2011). Enzymes are enriched in bacterial essential genes. *PloS one*, 6(6):e21683+.

Gazanion, E., Garcia, D., Silvestre, R., Gérard, C., Guichou, J. F., Labesse, G., Seveno, M., Cordeiro-Da-Silva, A., Ouaissi, A., Sereno, D. et Vergnes, B. (2011). The *Leishmania* nicotinamidase is essential for NAD+ production and parasite proliferation. *Molecular microbiology*, 82(1):21–38.

Gerdes, S., Edwards, R., Kubal, M., Fonstein, M., Stevens, R. et Osterman, A. (2006). Essential genes on metabolic maps. *Current opinion in biotechnology*, 17(5):448–456.

Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balázsi, G., Ravasz, E., Daugherty, M. D., Somera, A. L., Kyrpides, N. C., Anderson, I., Gelfand, M. S., Bhattacharya, A., Kapatral, V., D'Souza, M., Baev, M. V., Grechkin, Y., Mseeh, F., Fonstein, M. Y., Overbeek, R., Barabási, A.-L. L., Oltvai, Z. N. et Osterman, A. L. (2003). Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *Journal of bacteriology*, 185(19):5673–5684.

Giglione, C., Serero, A., Pierre, M., Boisson, B. et Meinnel, T. (2000). Identification of eukaryotic peptide deformylases reveals universality of N-terminal protein processing mechanisms. *The EMBO journal*, 19(21):5916–5929.

Gil, R., Sabater-Muñoz, B., Latorre, A., Silva, F. J. et Moya, A. (2002). Extreme genome reduction in *Buchnera* spp.: toward the minimal genome needed for symbiotic life. *Proceedings of the National Academy of Sciences of the United States of America*, 99(7):4454–4458.

Gil, R., Silva, F. J., Peretó, J. et Moya, A. (2004). Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev*, 68(3).

Gil, R., Silva, F. J., Zientz, E., Delmotte, F., González-Candelas, F., Latorre, A., Rausell, C., Kamerbeek, J., Gadau, J., Hölldobler, B., van Ham, R. C., Gross, R. et Moya, A. (2003). The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9388–9393.

Gill, J. W. et Vogel, H. J. (1963). A bacterial endosymbiote in *Crithidia* (*Strigomonas*) *oncopelti*: Biochemical and morphological aspects. *Journal of Eukaryotic Microbiology*, 10(2):148–152.

Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., Maruf, M., Hutchison, C. A., Smith, H. O. et Venter, J. C. (2006). Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences of the United States of America*, 103(2):425–430.

Gómez-Valero, L., Silva, F. J., Christophe Simon, J. et Latorre, A. (2007). Genome reduction of the aphid endosymbiont *Buchnera aphidicola* in a recent evolutionary time scale. *Gene*, 389(1):87–95.

Green, M. et Karp, P. (2004). A bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5(1):76+.

Groppa, M. D., Rosales, E. P., Iannone, M. F. et Benavides, M. P. (2008). Nitric oxide, polyamines and cd-induced phytotoxicity in wheat roots. *Phytochemistry*, 69(14):2609–2615.

Gualdrón-López, M., Brennand, A., Hannaert, V., Quiñones, W., Cáceres, A. J., Bringaud, F., Concepción, J. L. et Michels, P. a. M. (2012). When, how and why glycolysis became compartmentalised in the Kinetoplastea. A new look at an ancient organelle. *International journal for parasitology*, 42(1):1–20.

Gutteridge, W. E., McCormack, J. J. et Jaffe, J. J. (1969). Presence and properties of dihydrofolate reductases within the genus *crithidia*. *Biochimica et Biophysica Acta (BBA) - Enzymology*, 178(3):453–458.

Guttman, H. N. (1962). *Crithidia* assays for unconjugated pteridines. *In* Pfleiderer, W. et Taylor, E. C., éditeurs : *Third Intern. Symp. Pteridines*, pages 255–266.

Guttman, H. N. (1966). First defined media for *Leptomonas* spp. from insects. *Journal of Eukaryotic Microbiology*, 13(3):390–392.

Guttman, H. N. (1967). Patterns of methionine and lysine biosynthesis in the trypanosomatidae during growth. *The Journal of protozoology*, 14(2):267–271.

Guttman, H. N. et Eisenman, R. N. (1965). 'cure' of *Crithidia* (*Strigomonas*) *Oncopelti* of its bacterial endosymbiote. *Nature*, 206:113–114.

Haggart, C. R., Bartell, J. a., Saucerman, J. J. et Papin, J. a. (2011). *Whole-genome metabolic network reconstruction and constraint-based modeling.*, volume 500. Elsevier Inc., 1 édition.

Hannaert, V., Bringaud, F., Opperdoes, F. R. et Michels, P. A. M. (2003). Evolution of energy metabolism and its compartmentation in Kinetoplastida. *Kinetoplastid Biology and Disease*, 30:1–30.

Hansen, A. K. et Moran, N. A. (2011). Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7):2849–2854.

Heinz, E., Kolarov, I., Kästner, C., Toenshoff, E. R., Wagner, M. et Horn, M. (2007). An *Acanthamoeba* sp. containing two phylogenetically different bacterial endosymbionts. *Environmental microbiology*, 9(6):1604–1609.

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B. et Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, 28(9):977–982.

Hernández-Montes, G., Díaz-Mejía, J. J., Pérez-Rueda, E. et Segovia, L. (2008). The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. *Genome biology*, 9(6):R95+.

Hoare, C. A. (1972). *Herpetosoma from man and other mammals*, pages 288–314. Blackwell Scientific Publications, Oxford.

Hollar, L., Lukes, J. et Maslov, D. a. (1998). Monophyly of endosymbiont containing trypanosomatids: phylogeny versus taxonomy. *The Journal of eukaryotic microbiology*, 45(3):293–7.

Horn, M. et Wagner, M. (2004). Bacterial endosymbionts of free-living amoebae. *The Journal of eukaryotic microbiology*, 51(5):509–514.

Hotopp, J. C. D., Clark, M. E., Oliveira, D. C. S. G., Foster, J. M., Fischer, P., Torres, M. C. M. n., Giebel, J. D., Kumar, N., Ishmael, N., Wang, S., Ingram, J., Nene, R. V., Shepard, J., Tomkins, J., Richards, S., Spiro, D. J., Ghedin, E., Slatko, B. E., Tettelin, H. et Werren, J. H. (2007). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science (New York, N.Y.)*, 317(5845):1753–6.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J.-H. H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Le Novère, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J. et SBML Forum (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics (Oxford, England)*, 19(4):524–531.

Hudson, A. O., Bless, C., Macedo, P., Chatterjee, S. P., Singh, B. K., Gilvarg, C. et Leustek, T. (2005). Biosynthesis of lysine in plants: evidence for a variant of the known bacterial pathways. *Biochimica et biophysica acta*, 1721(1-3):27–36.

Husnik, F., Nikoh, N., Koga, R., Ross, L., Duncan, R. P., Fujie, M., Tanaka, M., Satoh, N., Bachtrog, D., Wilson, A. C., von Dohlen, C. D., Fukatsu, T. et McCutcheon, J. P. (2013). Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis. *Cell*, 153(7):1567–1578.

Huson, D., Richter, D., Rausch, C., Dezulian, T., Franz, M. et Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, 8(1):460+.

Hutner, S. H., Bacchi, C. J. et Baker, H. (1979). Nutrition of the kinetoplastida. *In* Lumsden, W. H. R. et Evans, D. A., éditeurs : *Biology of the Kinetoplastida*, volume 2, pages 645–691. Academic Press, London & New York.

Hutner, S. H., Bacchi, C. J., Shapiro, A. et Baker, H. (1980). Protozoa as tools for nutrition research. *Nutrition Reviews*, 38(11):361–364.

Hutner, S. H., Levin, H. L. et Nathan, H. A. (1956). Independent requirements for crithidia factor and folic acid in a trypanosomid flagellate. *Nature*, 178(4536):741–742.

Hutner, S. H. et Provasoli, L. (1965). Comparative physiology: Nutrition. *Annual Review of Physiology*, 27(1):19–48.

International Aphid Genomics Consortium (2010). Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS biology*, 8(2):e1000313+.

Itoh, T., Martin, W. et Nei, M. (2002). Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12944–12948.

Janakidevi, K., Dewey, V. C. et Kidder, G. W. (1966). Serotonin in protozoa. *Archives of biochemistry and biophysics*, 113(3):758–759.

Jansen, A. M., Carreira, J. C. et Deane, M. P. (1988). Infection of a Mammal by Monogenetic Insect Trypanosomatids. *Memórias do Instituto Oswaldo Cruz*, 83(3):271–272.

Jensen, R. E. et Englund, P. T. (2012). Network news: the replication of kinetoplast DNA. *Annual review of microbiology*, 66:473–91.

Juhas, M., Eberl, L. et Glass, J. I. (2011). Essence of life: essential genes of minimal genomes. *Trends in cell biology*, 21(10):562–568.

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

Karlberg, O., Canbäck, B., Kurland, C. G. et Andersson, S. G. (2000). The dual origin of the yeast mitochondrial proteome. *Yeast (Chichester, England)*, 17(3):170–187.

Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., Tsoka, S., Darzentas, N., Kunin, V. et Lopez-Bigas, N. (2005). Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucl. Acids Res.*, 33(19):6083–6089.

Karp, P. D. et Paley, S. (1996). Integrated access to metabolic and genomic data. *Journal of computational biology : a journal of computational molecular cell biology*, 3(1):191–212.

Karp, P. D., Paley, S. et Romero, P. (2002). The pathway tools software. *Bioinformatics (Oxford, England)*, 18 Suppl 1.

Karp, P. D., Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I. M. et Caspi, R. (2010). Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, 11(1):40–79.

Karp, P. D. et Riley, M. (1993). Representations of metabolic knowledge. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 1:207–215.

Keeling, P. J. (2010). The endosymbiotic origin, diversification and fate of plastids. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1541):729–48.

Keeling, P. J. (2011). Endosymbiosis: bacteria sharing the load. *Current biology : CB*, 21(16):R623–4.

Keeling, P. J. et Archibald, J. M. (2008). Organelle evolution: what's in a name? *Current biology : CB*, 18(8):R345–R347.

Keeling, P. J. et Corradi, N. (2011). Shrink it or lose it: balancing loss of function with shrinking genomes in the microsporidia. *Virulence*, 2(1):67–70.

Keseler, I. M., Bonavides-Martinez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R. P., Johnson, D. A., Krummenacker, M., Nolan, L. M., Paley, S., Paulsen, I. T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A. G. et Karp, P. D. (2009). Ecocyc: A comprehensive view of *Escherichia coli* biology. *Nucl. Acids Res.*, 37(suppl_1):D464–470.

Kidder, G. W., Davis, J. S. et Cousens, K. (1966). Citrulline utilization in crithidia. *Biochemical and biophysical research communications*, 24(3):365–369.

Kidder, G. W. et Dewey, V. C. (1972). Methionine or folate and phosphoenolpyruvate in the biosynthesis of threonine in *Crithidia fasciculata*. *Journal of Eukaryotic Microbiology*, 19(1):93–98.

Kidder, G. W. et Dutta, B. N. (1958). The growth and nutrition of *Crithidia fasciculata*. *Journal of general microbiology*, 18(3):621–638.

Kim, K. M. M. et Caetano-Anollés, G. (2010). Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data. *Molecular biology and evolution*, 27(7):1710–1733.

Kim, T. Y., Sohn, S. B., Kim, Y. B., Kim, W. J. et Lee, S. Y. (2012). Recent advances in reconstruction and applications of genome-scale metabolic models. *Current opinion in biotechnology*, 23(4):617–23.

Kirkness, E. F., Haas, B. J., Sun, W., Braig, H. R., Perotti, M. A., Clark, J. M., Lee, H., Robertson, H. M., Kennedy, R. C., Gerlach, D., Kriventseva, E. V., Elsik, G., Graur, D., Hill, C. A., Veenstra, J. A., Walenz, B., Tubío, J. M. C., Ribeiro, J. M. C., Rozas, J., Johnston, J. S., Reese, J. T., Popadic, A., Tojo, M., Raoult, D., Reed, D. L., Tomoyasu, Y., Kraus, E., Mittapalli, O., Margam, V. M., Li, H.-m., Meyer, J. M., Johnson, R. M., Romero-severson, J., Vanzee, J. P., Alvarez-ponce, D., Vieira, F. G., Aguadé, M., Guirao-rico, S., Anzola, J. M., Yoon, S., Strycharz, J. P., Unger, M. F., Christley, S., Lobo, N. F., Seufferheld, M. J., Wang, N., Dasch, A., Struchiner, C. J., Madey, G., Hannick, L. I., Bidwell, S., Joardar, V., Caler, E., Shao, R., Barker, S. C., Cameron, S., Bruggner, R. V., Regier, A., Johnson, J., Viswanathan, L., Utterback, R., Sutton, G. G., Lawson,

D., Waterhouse, R. M., Venter, J. C., Strausberg, R. L., Berenbaum, M. R., Collins, F. H., Zdobnov, E. M. et Pittendrigh, B. R. (2010). Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences*, 107(27):12168–12173.

Klamt, S., Haus, U.-U. et Theis, F. (2009). Hypergraphs and cellular networks. *PLoS computational biology*, 5(5):e1000385.

Klasson, L. et Andersson, S. G. (2004). Evolution of minimal-gene-sets in host-dependent bacteria. *Trends in microbiology*, 12(1):37–43.

Klaus, S. M., Wegkamp, A., Sybesma, W., Hugenholtz, J., Gregory, J. F. et Hanson, A. D. (2005). A nudix enzyme removes pyrophosphate from dihydroneopterin triphosphate in the folate synthesis pathway of bacteria and plants. *The Journal of biological chemistry*, 280(7):5274–5280.

Klein, C., Marino, A., Sagot, M.-F., Vieira Milreu, P. et Brilli, M. (2012a). Structural and dynamical analysis of biological networks. *Briefings in Functional Genomics*, 11(6):420–433.

Klein, C. C., Alves, J. M. P., Serrano, M. G., Buck, G. A., Vasconcelos, A. T. R., Sagot, M.-F., Teixeira, M. M. G., Camargo, E. P. et Motta, M. C. M. (2013). Biosynthesis of vitamins and cofactors in bacterium-harbouring trypanosomatids depends on the symbiotic association as revealed by genomic analyses. *PLoS One*, 8(11).

Klein, C. C., Cottret, L., Kielbassa, J., Charles, H., Gautier, C., Vasconcelos, A. T. T., Lacroix, V. et Sagot, M.-F. F. (2012b). Exploration of the core metabolism of symbiotic bacteria. *BMC genomics*, 13(1).

Kondo, N., Nikoh, N., Ijichi, N., Shimada, M. et Fukatsu, T. (2002). Genome fragment of *Wolbachia* endosymbiont. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22):14280–14285.

Koonin, E. V. (2000). How many genes can make a cell: the minimal-gene-set concept. *Annual review of genomics and human genetics*, 1:99–116.

Koonin, E. V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature reviews. Microbiology*, 1(2):127–136.

Koonin, E. V., Mushegian, A. R. et Bork, P. (1996). Non-orthologous gene displacement. *Trends in genetics : TIG*, 12(9):334–336.

Korený, L., Lukes, J. et Oborník, M. (2010). Evolution of the haem synthetic pathway in kinetoplastid flagellates: an essential pathway that is not essential after all? *International journal for parasitology*, 40(2):149–156.

Kornberg, A., Rao, N. N. et Ault-riché, D. (1999). Inorganic polyphosphate: a molecule of many functions. *Annual review of biochemistry*, 68:89–125.

Krassner, S. M. et Flory, B. (1971). Essential amino acids in the culture of *Leishmania tarentolae*. *The Journal of parasitology*, 57(4):917–920.

Kurland, C. G. et Andersson, S. G. (2000). Origin and evolution of the mitochondrial proteome. *Microbiology and molecular biology reviews : MMBR*, 64(4):786–820.

Kurnasov, O., Goral, V., Colabroy, K., Gerdes, S., Anantha, S., Osterman, A. et Begley, T. P. (2003). NAD biosynthesis: identification of the tryptophan to quinolinate pathway in bacteria. *Chemistry & biology*, 10(12):1195–1204.

Lacroix, V., Cottret, L., Thébault, P. et Sagot, M.-F. (2008). An introduction to metabolic networks and their structural analysis. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 5(4):594–617.

Lacroix, V., Fernandes, C. G. et Sagot, M.-F. (2006). Motif search in graphs: application to metabolic networks. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 3(4):360–368.

Lafaye, A., Junot, C., Pereira, Y., Lagniel, G., Tabet, J.-C., Ezan, E. et Labarre, J. (2005). Combined proteome and metabolite-profiling analyses reveal surprising insights into yeast sulfur metabolism. *J. of Biol. Chemistry*, 280:24723–24730.

Lanyon-Hogg, T., Warriner, S. L. et Baker, A. (2010). Getting a camel through the eye of a needle: the import of folded proteins by peroxisomes. *Biology of the cell / under the auspices of the European Cell Biology Organization*, 102(4):245–63.

Latendresse, M. et Karp, P. (2011). Web-based metabolic network visualization with a zooming user interface. *BMC Bioinformatics*, 12(1):176+.

Latendresse, M., Krummenacker, M., Trupp, M. et Karp, P. D. (2012). Construction and completion of flux balance models from pathway databases. *Bioinformatics (Oxford, England)*, 28(3):388–396.

Lee, T. J., Paulsen, I. et Karp, P. (2008). Annotation-based inference of transporter function. *Bioinformatics*, 24(13):i259–i267.

Lemercier, G., Espiau, B., Ruiz, F. a., Vieira, M., Luo, S., Baltz, T., Docampo, R. et Bakalara, N. (2004). A pyrophosphatase regulating polyphosphate metabolism in acidocalcisomes is essential for *Trypanosoma brucei* virulence in mice. *The Journal of biological chemistry*, 279(5):3420–5.

Lewis, S. et Ball, S. (1981). Micro-organisms in *Trypanosoma cobitis*. *International Journal for Parasitology*, 11(2):121 – 125.

Li, Z.-S., Lu, Y.-P., Zhen, R.-G., Szczypka, M., Thiele, D. J. et Rea, P. A. (1997). A new pathway for vacuolar cadmium sequestration in saccharomyces cerevisiae: Ycf1-catalyzed transport of glutathionato cadmium. *PNAS*, 94.

Lim, L. et McFadden, G. I. (2010). The evolution, metabolism and functions of the apicoplast. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1541):749–63.

Lima, T., Auchincloss, A. H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C., Baratin, D., Phan, I., Bougueleret, L. et Bairoch, A. (2009). Hamap: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in uniprotkb/swiss-prot. *Nucleic acids research*, 37(Database issue):gkn661+.

Lin, S., Hanson, R. E. et Cronan, J. E. (2010). Biotin synthesis begins by hijacking the fatty acid synthetic pathway. *Nature chemical biology*, 6(9):682–688.

López-Sánchez, M. J., Neef, A., Peretó, J., Patiño Navarrete, R., Pignatelli, M., Latorre, A. et Moya, A. (2009). Evolutionary convergence and nitrogen metabolism in *Blattabacterium* strain bge, primary endosymbiont of the cockroach *Blattella germanica*. *PLoS genetics*, 5(11):e1000721+.

Luo, S., Ruiz, F. a. et Moreno, S. N. J. (2005). The acidocalcisome Ca2+-ATPase (TgA1) of *Toxoplasma gondii* is required for polyphosphate storage, intracellular calcium homeostasis and virulence. *Molecular microbiology*, 55(4):1034–45.

Lwoff, M. (1940). *Recherches sur le pouvoir de synthèse des flagellés trypanosomides*. Thèse de doctorat, Institut Pasteur.

Ma, H. et Zeng, A.-P. (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270–277.

Macdonald, S. J., Lin, G. G., Russell, C. W., Thomas, G. H. et Douglas, A. E. (2012). The central role of the host cell in symbiotic nitrogen metabolism. *Proceedings. Biological sciences / The Royal Society*, 279(1740):2965–2973.

Madalinski, G., Godat, E., Alves, S., Lesage, D., Genin, E., Levi, P., Labarre, J., Tabet, J.-C., Ezan, E. et Junot, C. (2008). Direct introduction of biological samples into a ltq-orbitrap hybrid mass spectrometer as a tool for fast metabolome analysis. *Analytical Chemistry*, 80(9):3291–3303. PMID: 18351782.

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Lu, F., Marchler, G. H., Mullokandov, M., Omelchenko, M. V., Robertson, C. L., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., Zhang, N., Zheng, C. et Bryant, S. H. (2011). CDD: a conserved domain database for the functional annotation of proteins. *Nucleic acids research*, 39(Database issue):D225–D229.

Margolin, W. (2005). FtsZ and the division of prokaryotic cells and organelles. *Nature reviews. Molecular cell biology*, 6(11):862–71.

Martínez-calvillo, S., Nguyen, D., Stuart, K. et Myler, P. J. (2004). Transcription Initiation and Termination on *Leishmania major* Chromosome 3. *Eukaryotic cell*, 3(2):506–517.

Martínez-Calvillo, S., Vizuet-de Rueda, J. C., Florencio-Martínez, L. E., Manning-Cela, R. G. et Figueroa-Angulo, E. E. (2010). Gene expression in trypanosomatid parasites. *Journal of biomedicine & biotechnology*, 2010:525241.

McCutcheon, J. P. (2010). The bacterial essence of tiny symbiont genomes. *Current opinion in microbiology*, 13(1):73–78.

McCutcheon, J. P., McDonald, B. R. et Moran, N. A. (2009). Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36):15394–15399.

McCutcheon, J. P. et Moran, N. A. (2007). Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proceedings of the National Academy of Sciences of the United States of America*, 104(49):19392–19397.

McCutcheon, J. P. et Moran, N. a. (2010). Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. *Genome biology and evolution*, 2:708–18.

McCutcheon, J. P. et Moran, N. A. (2012). Extreme genome reduction in symbiotic bacteria. *Nature reviews. Microbiology*, 10(1):13–26.

McCutcheon, J. P. et von Dohlen, C. D. (2011). An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Current biology : CB*, 21(16):1366–1372.

McDonald, A. G., Boyce, S. et Tipton, K. F. (2009). ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic acids research*, 37(Database issue):D593–7.

McFadden, G. I. (2011). The apicoplast. *Protoplasma*, 248(4):641–50.

McGhee, R. B. (1957). Infection of Chick Embryos by *Crithidia* from a Phytophagous Hemipteron Homotransplantation of Human Cell Lines. *Science*, 125(January):157–158.

Meister, A. (1965). *Biochemistry of the amino acids*. Academic Press.

Mendonça, A. G., Alves, R. J. et Pereira-Leal, J. B. (2011). Loss of genetic redundancy in reductive genome evolution. *PLoS computational biology*, 7(2):e1001082+.

Merhej, V., Royer-Carenzi, M., Pontarotti, P. et Raoult, D. (2009). Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biology direct*, 4:13+.

Michels, P. a., Hannaert, V. et Bringaud, F. (2000). Metabolic aspects of glycosomes in trypanosomatidae - new data and views. *Parasitology today (Personal ed.)*, 16(11):482–9.

Michels, P. a. M., Bringaud, F., Herman, M. et Hannaert, V. (2006). Metabolic functions of glycosomes in trypanosomatids. *Biochimica et biophysica acta*, 1763(12):1463–77.

Milreu, P. V., Klein, C. C., Cottret, L., Acuña, V., Birmelé, E., Borassi, M., Junot, C., Marchetti-Spaccamela, A., Marino, A., Stougie, L., Jourdan, F., Crescenzi, P., Lacroix, V. et Sagot, M.-F. (2014). Telling metabolic stories to explore metabolomics data: a case study on the yeast response to cadmium exposure. *Bioinformatics (Oxford, England)*, 30(1):61–70.

Mithani, A., Preston, G. M. et Hein, J. (2009). A stochastic model for the evolution of metabolic networks with neighbor dependence. *Bioinformatics*, 25(12):1528–1535.

Moran, N. A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 93(April): 2873–2878.

Moran, N. A. (2006). Symbiosis. *Current biology : CB*, 16(20).

Moran, N. A. (2007). Symbiosis as an adaptive process and source of phenotypic complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 104 Suppl 1(Suppl 1):8627–8633.

Moran, N. A., McCutcheon, J. P. et Nakabachi, A. (2008). Genomics and evolution of heritable bacterial symbionts. *Annual review of genetics*, 42(1):165–190.

Moran, N. A., Munson, M. A., Baumann, P. et Ishikawa, H. (1993). A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. *Proceedings. Biological sciences / The Royal Society*, 253(1337):167–171.

Moreno, S. N. J. et Docampo, R. (2009). The Role of Acidocalcisomes in Parasitic Protists. *The Journal of eukaryotic microbiology*, 56(3):208–213.

Morio, F., Reynes, J., Dollet, M., Pratlong, F., Dedet, J.-P. et Ravel, C. (2008). Isolation of a protozoan parasite genetically related to the insect trypanosomatid Herpetomonas samuelpessoai from a human immunodeficiency virus-positive patient. *Journal of clinical microbiology*, 46(11):3845–7.

Mormann, S., Lomker, A., Ruckert, C., Gaigalat, L., Tauch, A., Puhler, A. et Kalinowski, J. (2006). 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*, 7(1):205+.

Motta, M. C., Catta-Preta, C. M., Schenkman, S., Martins, A. C., Miranda, K., de Souza, W. et Elias, M. C. (2010). The bacterium endosymbiont of *Crithidia deanei* undergoes coordinated division with the host cell nucleus. *PLoS ONE*, 5(8):e12415+.

Motta, M. C., Martins, A. C., de Souza, S. S., Catta-Preta, C. M., Silva, R., Klein, C. C., de Almeida, L. G., de Lima Cunha, O., Ciapina, L. P., Brocchi, M., Colabardini, A. C., de Araujo Lima, B., Machado, C. R., de Almeida Soares, C. M., Probst, C. M., de Menezes, C. B., Thompson, C. E., Bartholomeu, D. C., Gradia, D. F., Pavoni, D. P., Grisard, E. C., Fantinatti-Garboggini, F., Marchini, F. K., Rodrigues-Luiz, G. F., Wagner, G., Goldman, G. H., Fietto, J. L., Elias, M. C., Goldman, M. H., Sagot, M.-F., Pereira, M., Stoco, P. H., de Mendonça-Neto, R. P., Teixeira, S. M., Maciel, T. E., de Oliveira Mendes, T. A., Ürményi, T. P., de Souza, W., Schenkman, S. et de Vasconcelos, A. T. (2013). Predicting the proteins of angomonas deanei, strigomonas culicis and their respective endosymbionts reveals new aspects of the trypanosomatidae family. *PLoS ONE*, 8(4):e60209+.

Motta, M. C., Soares, M. J., Attias, M., Morgado, J., Lemos, A. P., Saad-Nehme, J., Meyer-Fernandes, J. R. et De Souza, W. (1997a). Ultrastructural and biochemical analysis of the relationship of *Crithidia deanei* with its endosymbiont. *European journal of cell biology*, 72(4):370–377.

Motta, M. C. M. (2010). Endosymbiosis in Trypanosomatids as a Model to Study Cell Evolution. *The Open Parasitology Journal*, 4(1):139–147.

Motta, M. C. M., Leal, L. H. M., Souza, W. d., de Almeida, D. F. et Ferreira, L. C. S. (1997b). Detection of penicillin-binding proteins in the endosymbiont of the trypanosomatid *Crithidia deanei*. *Journal of Eukaryotic Microbiology*, 44(5):492–496.

Motta, M. C. M., Picchi, G. F. a., Palmié-Peixoto, I. V., Rocha, M. R., de Carvalho, T. M. U., Morgado-Diaz, J., de Souza, W., Goldenberg, S. et Fragoso, S. P. (2004). The microtubule analog protein, FtsZ, in the endosymbiont of trypanosomatid protozoa. *The Journal of eukaryotic microbiology*, 51(4):394–401.

Moya, a., Gil, R. et Latorre, a. (2009). The evolutionary history of symbiotic associations among bacteria and their animal hosts: a model. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 15 Suppl 1:11–3.

Moya, A., Peretó, J., Gil, R. et Latorre, A. (2008). Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nature reviews. Genetics*, 9(3):218–229.

Mundim, M. H. et Roitman, I. (1977). Extra nutritional requirements of artificially aposymbiotic *Crithidia deanei*. *Journal of Eukaryotic Microbiology*, 24(2):329–331.

Mundim, M. H., Roitman, I., Hermans, M. A. et Kitajima, E. W. (1974). Simple nutrition of *Crithidia deanei*, a reduviid trypanosomatid with an endosymbiont. *The Journal of protozoology*, 21(4):518–521.

Mushegian, A. (1999). The minimal genome concept. *Current opinion in genetics & development*, 9(6):709–714.

Mushegian, A. R. et Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19):10268–10273.

Nardon, P. et Grenier, A. M. (1993). Symbiose et évolution. *Annales de la Société entomologique de France*, 29(2):113–140.

Nascimento, M. T. T., Garcia, M. C. C., da Silva, K. P. P., Pinto-da Silva, L. H. H., Atella, G. C., Motta, M. C. C. et Saraiva, E. M. (2010). Interaction of the monoxenic trypanosomatid *Blastocrithidia culicis* with the *Aedes aegypti* salivary gland. *Acta tropica*, 113(3):269–278.

Nathan, H. A., Baker, H. et Frank, O. (1960). Influence of pteridines on the production of vitamin b12 by trypanosomid flagellates. *Nature*, 188:35–37.

Nathan, H. A. et Cowperthwaite, J. (1954). Use of the trypanosomid flagellate, *Crithidia fasciculata*, for evaluating antimalarials. *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine (New York, N.Y.)*, 85(1):117–119.

Nathan, H. A. et Cowperthwaite, J. (1955). ,Äùcrithidia factor,Äù,Äî a new member of the folic acid group of vitamins. *Journal of Eukaryotic Microbiology*, 2(2):37–42.

Nerima, B., Nilsson, D. et Mäser, P. (2010). Comparative genomics of metabolic networks of free-living and parasitic eukaryotes. *BMC genomics*, 11:217+.

Newton, B. A. (1956). A synthetic growth medium for the trypanosomid flagellate *Strigomonas* (*Herpetomonas*) *oncopelti*. *Nature*, 177(4502):279–280.

Newton, B. A. et Horne, R. W. (1957). Intracellular structures in *Strigomonas oncopelti*. i. cytoplasmic structures containing ribonucleoprotein. *Experimental cell research*, 13(3):563–574.

Newton, B. S. (1957). Nutritional requirements and biosynthetic capabilities of the parasitic flagellate *Strigomonas oncopelti*. *Journal of general microbiology*, 17(3):708–717.

Nikoh, N., McCutcheon, J. P., Kudo, T., Miyagishima, S.-y. Y., Moran, N. A. et Nakabachi, A. (2010). Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. *PLoS genetics*, 6(2):e1000827+.

Nikoh, N., Tanaka, K., Shibata, F., Kondo, N., Hizume, M., Shimada, M. et Fukatsu, T. (2008). *Wolbachia* genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes. *Genome research*, 18(2):272–80.

Nishida, H. (2001). Distribution of genes for lysine biosynthesis through the aminoadipate pathway among prokaryotic genomes. *Bioinformatics (Oxford, England)*, 17(2):189–191.

Noguchi, H. et Tilden, E. B. (1926). Comparative studies of herpetomonads and leishmanias : I. cultivation of herpetomonads from insects and plants. *The Journal of experimental medicine*, 44(3):307–325.

Nogueira de Melo, A. C., D'Avila-Levy, C. M., Dias, F. a., Armada, J. L. a., Silva, H. D., Lopes, A. H. C. S., Santos, A. L. S., Branquinha, M. H. et Vermelho, A. B. (2006). Peptidases and gp63-like proteins in *Herpetomonas megaseliae*: possible involvement in the adhesion to the invertebrate host. *International journal for parasitology*, 36(4):415–22.

Noor, E., Eden, E., Milo, R. et Alon, U. (2010). Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. *Molecular cell*, 39(5):809–20.

Notebaart, R. a., van Enckevort, F. H. J., Francke, C., Siezen, R. J. et Teusink, B. (2006). Accelerating the reconstruction of genome-scale metabolic networks. *BMC bioinformatics*, 7:296.

Novy, F. G., MacNeal, W. J. et Torrey, H. N. (1907). The trypanosomes of mosquitoes and other insects. *Journal of Infectious Diseases*, 4(2):223–276.

Nowack, E. C. M. et Melkonian, M. (2010). Endosymbiotic associations within protists. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1541):699–712.

Oda, L. M., Alviano, C. S., Filho, F. C., Angluster, J., Roitman, I. et Souza, W. (1984). Surface anionic groups in Symbiote-Bearing and Symbiote-Free strains of *Crithidia deanei*. *Journal of Eukaryotic Microbiology*, 31(1):131–134.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. et Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29–34.

Opperdoes, F. R. et Borst, P. (1977). Localization of nine glycolytic enzymes in a microbody-like organelle in *Trypanosoma brucei*: the glycosome. *FEBS letters*, 80(2).

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. a., Nam, H., Feist, A. M. et Palsson, B. O. (2011). A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Molecular systems biology*, 7(535):535.

Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D. N., Roopra, S., Frings, O. et Sonnhammer, E. L. L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic acids research*, 38(Database issue):D196–203.

Ozaki, L. S. et Cseko, Y. M. T. (1984). *Genomic DNA Cloning and Related Techniques*. Fundação Oswaldo cruz.

Pacheco, R. S., Marzochi, M. C., Pires, M. Q., Brito, C. M., Madeira, M. D. F. et Barbosa-Santos, E. G. (1998). Parasite genotypically related to a monoxenous trypanosomatid of dog's flea causing opportunistic infection in an HIV positive patient. *Memórias do Instituto Oswaldo Cruz*, 93(4):531–7.

Paley, S. M. et Karp, P. D. (2006). The pathway tools cellular overview diagram and omics viewer. *Nucleic Acids Research*, 34(13):3771–3778.

Palmié-Peixoto, I. V. V., Rocha, M. R. R., Urbina, J. A., de Souza, W., Einicker-Lamas, M. et Motta, M. C. M. C. (2006). Effects of sterol biosynthesis inhibitors on endosymbiont-bearing trypanosomatids. *FEMS microbiology letters*, 255(1):33–42.

Palsson, B. (2000). The challenges of in silico biology. *Nature biotechnology*, 18(11):1147–1150.

Palsson, B. (2006). *Systems Biology*. Cambridge University Press.

Parsons, M., Furuya, T., Pal, S. et Kessler, P. (2001). Biogenesis and function of peroxisomes and glycosomes. *Molecular and biochemical parasitology*, 115(1):19–28.

Peden, J. F. (2006). CodonW software distributed by the author.

Peregrín-Alvarez, J. M., Sanford, C. et Parkinson, J. (2009). The conservation and evolutionary modularity of metabolism. *Genome biology*, 10(6):R63+.

Pereira, F. M., Bernardo, P. S., Dias Junior, P. F. F., Silva, B. a., Romanos, M. T. V., D'Avila-Levy, C. M., Branquinha, M. H. et Santos, a. L. S. (2009). Differential influence of gp63-like molecules in three distinct Leptomonas species on the adhesion to insect cells. *Parasitology research*, 104(2):347–53.

Pérez-Brocal, V., Gil, R., Ramos, S., Lamelas, A., Postigo, M., Michelena, J. M. M., Silva, F. J., Moya, A. et Latorre, A. (2006). A small microbial genome: the end of a long symbiotic relationship? *Science (New York, N.Y.)*, 314(5797):312–313.

Petersen, L. N., Marineo, S., Mandalà, S., Davids, F., Sewell, B. T. et Ingle, R. A. (2010). The missing link in plant histidine biosynthesis: Arabidopsis myoinositol monophosphatase-like2 encodes a functional histidinol-phosphate phosphatase. *Plant physiology*, 152(3):1186–1196.

Pharkya, P., Burgard, A. P. et Maranas, C. D. (2003). Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. *Biotechnology and bioengineering*, 84(7):887–99.

Pharkya, P., Burgard, A. P. et Maranas, C. D. (2004). OptStrain: a computational framework for redesign of microbial production systems. *Genome research*, 14(11):2367–76.

Pinney, J. W., Papp, B., Hyland, C., Wambua, L., Westhead, D. R. et McConkey, G. a. (2007). Metabolic reconstruction and analysis for parasite genomes. *Trends in parasitology*, 23(11):548–54.

Pitkänen, E., Rousu, J. et Ukkonen, E. (2010). Computational methods for metabolic reconstruction. *Current opinion in biotechnology*, 21(1):70–7.

Podlipaev, S. a. (2000). Insect trypanosomatids: the need to know more. *Memórias do Instituto Oswaldo Cruz*, 95(4):517–22.

Podlipaev, S. a., Sturm, N. R., Fiala, I., Fernandes, O., Westenberger, S. J., Dollet, M., Campbell, D. a. et Lukes, J. (2004). Diversity of insect trypanosomatids assessed from the spliced leader RNA and 5S rRNA genes and intergenic regions. *The Journal of eukaryotic microbiology*, 51(3):283–90.

Poliakov, A., Russell, C. W., Ponnala, L., Hoops, H. J., Sun, Q., Douglas, A. E. et van Wijk, K. J. (2011). Large-scale label-free quantitative proteomics of the pea aphid-*Buchnera* symbiosis. *Molecular & cellular proteomics : MCP*, 10(6):M110.007039.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rajagopalan, P. T., Datta, A. et Pei, D. (1997). Purification, characterization, and inhibition of peptide deformylase from *Escherichia coli*. *Biochemistry*, 36(45):13910–13918.

Raman, K. et Chandra, N. (2009). Flux balance analysis of biological systems: applications and challenges. *Briefings in bioinformatics*, 10(4):435–49.

Ranganathan, G. et Mukkada, A. J. (1995). Ubiquinone biosynthesis in *Leishmania major* promastigotes. *International journal for parasitology*, 25(3):279–284.

Rao, N. N., Gómez-García, M. R. et Kornberg, A. (2009). Inorganic polyphosphate: essential for growth and survival. *Annual review of biochemistry*, 78:605–47.

Reed, J. L., Famili, I., Thiele, I. et Palsson, B. O. (2006). Towards multidimensional genome annotation. *Nature reviews. Genetics*, 7(2):130–141.

Ren, Q., Chen, K. et Paulsen, I. T. (2007). TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic acids research*, 35(Database issue):D274–9.

Resendis-Antonio, O., Reed, J. L., Encarnación, S., Collado-Vides, J. et Palsson, B. O. (2007). Metabolic reconstruction and modeling of nitrogen fixation in Rhizobium etli. *PLoS computational biology*, 3(10):1887–95.

Rocha, I., Förster, J. et Nielsen, J. (2008). Design and application of genome-scale reconstructed metabolic models. *Methods in molecular biology (Clifton, N.J.)*, 416:409–431.

Roitman, C., Roitman, I. et de AZEVEDO, H. P. E. I. X. O. T. O. (1972). Growth of an insect trypanosomatid at 37 c in a defined medium. *Journal of Eukaryotic Microbiology*, 19(2):346–349.

Roitman, I. et Camargo, E. P. (1985). Endosymbionts of trypanosomatidae. *Parasitology today (Personal ed.)*, 1(5):143–144.

Roitman, I., Mundim, M. H., Azevedo, H. P. et Kitajima, E. W. (1977). Growth of crithidia at high temperature: *Crithidia hutneri* sp. n. and *Crithidia luciliae* thermophila s. sp. n. *Journal of Eukaryotic Microbiology*, 24(4):553–556.

Romero, P. R. et Karp, P. D. (2004). Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. *Bioinformatics*, 20(5):709–717.

Rucktäschel, R., Girzalsky, W. et Erdmann, R. (2011). Protein import machineries of peroxisomes. *Biochimica et biophysica acta*, 1808(3):892–900.

Saier, M. H., Yen, M. R., Noto, K., Tamang, D. G. et Elkan, C. (2009). The Transporter Classification Database: recent advances. *Nucleic acids research*, 37(Database issue):D274–8.

Salzman, T. A., Batlle, A. M., Angluster, J. et de Souza, W. (1985). Heme synthesis in *Crithidia deanei*: influence of the endosymbiote. *The International Journal of Biochemistry*, 17(12):1343–7.

Sandström, J. et Moran, N. (1999). How nutritionally imbalanced is phloem sap for aphids? *Entomologia Experimentalis et Applicata*, 91(1):203–210.

Santos, F., Boele, J. et Teusink, B. (2011). A practical guide to genome-scale metabolic models and their analysis. *Methods in enzymology*, 500:509–32.

Schaub, G. A. (1988). Parasite-host interrelationships of *Blastocrithidia triatomae* end Triatomines. *Memórias do Instituto Oswaldo Cruz*, 83.

Schaub, G. A. et Sehnitker, A. (1988). Influence of *Blastocrithidia triatomae* (Trypanosomatidae) on the reduviid bug *Triatoma infestans*: alterations in the Malpigbian tubules. *Parasitology research*, 75(2):88–97.

Schellenberger, J., Que, R., Fleming, R. M., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R. et Palsson, B. Ø. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox v2.0. *Nature protocols*, 6(9):1290–1307.

Schomburg, I., Chang, A., Placzek, S., Söhngen, C., Rother, M., Lang, M., Munaretto, C., Ulas, S., Stelzer, M., Grote, A., Scheer, M. et Schomburg, D. (2013). BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic acids research*, 41(Database issue):D764–72.

Sekowska, A., Robin, S., Daudin, J. J., Hénaut, A. et Danchin, A. (2001). Extracting biological information from DNA arrays: an unexpected link between arginine and methionine metabolism in bacillus subtilis. *Genome biology*, 2(6).

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. et Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. et Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. aps. *Nature*, 407(6800):81–86.

Silva, F. J., Latorre, A. et Moya, A. (2003). Why are the genomes of endosymbiotic bacteria so stable? *Trends in genetics : TIG*, 19(4):176–.

Simpson, A. G., Stevens, J. R. et Lukes, J. (2006). The evolution and diversity of kinetoplastid flagellates. *Trends in parasitology*, 22(4):168–174.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A. et Lewis, S. (2002). The generic genome browser: a building block for a model organism system database. *Genome research*, 12(10):1599–1610.

Stoll, S., Feldhaar, H. et Gross, R. (2009). Promoter characterization in the AT-rich genome of the obligate endosymbiont *Candidatus* Blochmannia floridanus. *Journal of bacteriology*, 191(11):3747–51.

Stover, B. et Muller, K. (2010). TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics*, 11(1):7+.

Suthers, P. F., Zomorrodi, A. et Maranas, C. D. (2009). Genome-scale gene/reaction essentiality and synthetic lethality analysis. *Molecular systems biology*, 5(301):301.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. et Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics (Oxford, England)*, 23(10):1282–1288.

Tamas, I., Klasson, L., Canbäck, B., Näslund, A. K., Eriksson, A.-S. S., Wernegreen, J. J., Sandström, J. P., Moran, N. A. et Andersson, S. G. (2002). 50 million years of genomic stasis in endosymbiotic bacteria. *Science (New York, N.Y.)*, 296(5577):2376–2379.

Tamas, I., Klasson, L. M., Sandström, J. P. et Andersson, S. G. (2001). Mutualists and parasites: how to paint yourself into a (metabolic) corner. *FEBS letters*, 498(2-3):135–139.

Teixeira, M. M., Borghesan, T. C., Ferreira, R. C., Santos, M. A., Takata, C. S., Campaner, M., Nunes, V. L., Milder, R. V., de Souza, W. et Camargo, E. P. (2011). 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*, 162(3):503–524.

Theodoulou, F. L., Bernhardt, K., Linka, N. et Baker, A. (2013). Peroxisome membrane proteins: multiple trafficking routes and multiple functions? *The Biochemical journal*, 451(3):345–52.

Thiele, I. et Palsson, B. O. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121.

Timmis, J. N., Ayliffe, M. A., Huang, C. Y. et Martin, W. (2004). Endosymbiotic gene Transfer: Organelle Genomes Forge Eukaryotic Chromosomes. *Nature reviews. Genetics*, 5(February):123–135.

Toft, C. et Andersson, S. G. E. (2010). Evolutionary microbial genomics: insights into bacterial host adaptation. *Nature reviews. Genetics*, 11(7):465–75.

Torruella, G., Suga, H., Riutort, M., Peretó, J. et Ruiz-Trillo, I. n. (2009). The evolutionary history of lysine biosynthesis pathways within eukaryotes. *Journal of molecular evolution*, 69(3):240–248.

Trinh, C. T., Unrean, P. et Srienc, F. (2008). Minimal Escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses. *Applied and environmental microbiology*, 74(12):3634–43.

Vaishnava, S. et Striepen, B. (2006). The cell biology of secondary endosymbiosis–how parasites build, divide and segregate the apicoplast. *Molecular microbiology*, 61(6):1380–7.

Vallenet, D., Engelen, S., Mornico, D., Cruveiller, S., Fleury, L., Lajus, A., Rouy, Z., Roche, D., Salvignol, G., Scarpelli, C. et Médigue, C. (2009). Microscope: a platform for microbial genome annotation and comparative genomics. *Database : the journal of biological databases and curation*, 2009(0):bap021.

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. *Trends in genetics : TIG*, 19(9):479–484.

Vasconcelos, A. T. R., Ferreira, H. B., Bizarro, C. V., Bonatto, S. L., Carvalho, M. O., Pinto, P. M., Almeida, D. F., Almeida, L. G., Almeida, R., Alves-Filho, L., Assunção, E. N., Azevedo, V. A., Bogo, M. R., Brigido, M. M., Brocchi, M., Burity, H. A., Camargo, A. A., Camargo, S. S., Carepo, M. S., Carraro, D. M., de Mattos Cascardo, J. C., Castro, L. A., Cavalcanti, G., Chemale, G., Collevatti, R. G., Cunha, C. W., Dallagiovanna, B., Dambrós, B. P., Dellagostin, O. A., Falcão, C. et et al. (2005). Swine and poultry pathogens: the complete genome sequences of two strains of mycoplasma hyopneumoniae and a strain of mycoplasma synoviae. *Journal of bacteriology*, 187(16):5568–5577.

Velasco, A. M., Leguina, J. I. et Lazcano, A. (2002). Molecular evolution of the lysine biosynthetic pathways. *Journal of molecular evolution*, 55(4):445–459.

Vercesi, A. E., Moreno, S. N. et Docampo, R. (1994). Ca2+/H+ exchange in acidic vacuoles of *Trypanosoma brucei*. *The Biochemical journal*, 304 (Pt 1):227–33.

Vickerman, K. (1994). The evolutionary expansion of the trypanosomatid flagellates. *International journal for parasitology*, 24(8):1317–1331.

Vickers, T. J. et Beverley, S. M. (2011). Folate metabolic pathways in *Leishmania*. *Essays in biochemistry*, 51:63–80.

Vieira, G., Sabarly, V., Bourguignon, P.-Y. Y., Durot, M., Le Fèvre, F., Mornico, D., Vallenet, D., Bouvet, O., Denamur, E., Schachter, V. et Médigue, C. (2011). Core and panmetabolism in *Escherichia coli*. *Journal of bacteriology*, 193(6):1461–1472.

von Dohlen, C. D., Kohler, S., Alsop, S. T. et McManus, W. R. (2001). Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. *Nature*, 412(6845):433–6.

Wallace, F. G. (1966). The trypanosomatid parasites of insects and arachnids. *Experimental parasitology*, 18(1):124–193.

Webb, E. C. (1992). *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. Published for the International Union of Biochemistry and Molecular Biology by Academic Press (San Diego), 1 édition.

Wenyon, C. M. (1926). *Protozoology - A manual for Medical Men, Veterinarians and Zoologists*. London.

Wernegreen, J. J. (2002). Genome evolution in bacterial endosymbionts of insects. *Nature reviews. Genetics*, 3(11):850–861.

Wernegreen, J. J. (2004). Endosymbiosis: Lessons in conflict resolution. *PLoS Biol*, 2(3):e68+.

Wernegreen, J. J. (2005). For better or worse: genomic consequences of intracellular mutualism and parasitism. *Current opinion in genetics & development*, 15(6):572–583.

Wernegreen, J. J. (2012). Strategies of genomic integration within insect-bacterial mutualisms. *The Biological bulletin*, 223(1):112–122.

Whelan, S. et Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, 18(5):691–699.

Willert, E. et Phillips, M. A. (2012). Regulation and function of polyamines in African trypanosomes. *Trends in parasitology*, 28(2):66–72.

Williams, L. E. et Wernegreen, J. J. (2010). Unprecedented loss of ammonia assimilation capability in a urease-encoding bacterial mutualist. *BMC genomics*, 11(1):687.

Wilson, A. C. C., Ashton, P. D., Calevro, F., Charles, H., Colella, S., Febvay, G., Jander, G., Kushlan, P. F., Macdonald, S. J., Schwartz, J. F., Thomas, G. H. et Douglas, A. E. (2010). Genomic insight into the amino acid relations of the pea aphid, *Acyrthosiphon pisum*, with its symbiotic bacterium *Buchnera aphidicola*. *Insect Molecular Biology*, 19:249–258.

Wu, D., Daugherty, S. C., Van Aken, S. E., Pai, G. H., Watkins, K. L., Khouri, H., Tallon, L. J., Zaborsky, J. M., Dunbar, H. E., Tran, P. L., Moran, N. A. et Eisen, J. A. (2006). Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS biology*, 4(6).

Wunderlich, Z. et Mirny, L. a. (2006). Using the topology of metabolic networks to predict viability of mutant strains. *Biophysical journal*, 91(6):2304–11.

Yang, Y., Zhao, G., Man, T. K. et Winkler, M. E. (1998). Involvement of the gapA- and epd (gapB)-encoded dehydrogenases in pyridoxal 5'-phosphate coenzyme biosynthesis in *Escherichia coli* k-12. *Journal of bacteriology*, 180(16):4294–4299.

Yoshida, N. et Camargo, E. P. (1978). Ureotelism and ammonotelism in trypanosomatids. *Journal of bacteriology*, 136(3):1184–1186.

Yoshida, N., Jankevicius, J. V., Roitman, I. et Camargo, E. P. (1978). Enzymes of the Ornithine-Arginine metabolism of trypanosomatids of the genus *Herpetomonas*. *Journal of Eukaryotic Microbiology*, 25(4):550–555.

Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W.-H. H., Wodke, J. A., Güell, M., Martínez, S., Bourgeois, R., Kühner, S., Raineri, E., Letunic, I., Kalinina, O. V., Rode, M., Herrmann, R., Gutiérrez-Gallego, R., Russell, R. B., Gavin, A.-C. C., Bork, P. et Serrano, L. (2009). Impact of genome reduction on bacterial metabolism and its regulation. *Science (New York, N.Y.)*, 326(5957): 1263–1268.

Zhang, C.-T. T. et Zhang, R. (2008). Gene essentiality analysis based on DEG, a database of essential genes. *Methods in molecular biology (Clifton, N.J.)*, 416:391–400.

Zientz, E., Beyaert, I., Gross, R. et Feldhaar, H. (2006). Relevance of the endosymbiosis of *Blochmannia floridanus* and carpenter ants at different stages of the life cycle of the host. *Applied and environmental microbiology*, 72(9):6027–33.

Zientz, E., Dandekar, T. et Gross, R. (2004). Metabolic interdependence of obligate intracellular bacteria and their insect hosts. *Microbiology and Molecular Biology Reviews*, 68(4):745–770.

Appendix A

Published article: Structural and dynamical analysis of biological networks

BRIEFINGS IN FUNCTIONAL GENOMICS. VOL II. NO 6. 420–433 doi:10.1093/bfgp/els030

Structural and dynamical analysis of biological networks

Cecilia Klein, Andrea Marino, Marie-France Sagot, Paulo Vieira Milreu and Matteo Brilli

Advance Access publication date 20 August 2012

Abstract

Biological networks are currently being studied with approaches derived from the mathematical and physical sciences. Their structural analysis enables to highlight nodes with special properties that have sometimes been correlated with the biological importance of a gene or a protein. However, biological networks are dynamic both on the evolutionary time-scale, and on the much shorter time-scale of physiological processes. There is therefore no unique network for a given cellular process, but potentially many realizations, each with different properties as a consequence of regulatory mechanisms. Such realizations provide snapshots of a same network in different conditions, enabling the study of condition-dependent structural properties. True dynamical analysis can be obtained through detailed mathematical modeling techniques that are not easily scalable to full network models.

Keywords: *networks; structural analysis; centrality; mathematical modeling; flux balance analysis*

INTRODUCTION

High-throughput technologies have recently led to a new perspective in biology, where the cell is interpreted as a large and complex system composed of highly integrated subsystems. Interpretation of these systems as networks of interactions has spurred the application of analytical tools developed since long by mathematicians and physicists to analyze biological networks.

Different biological networks can be defined; detailed descriptions in addition to the approaches to their reconstruction are treated exhaustively in several publications (Supplementary Material File 1). In this review, we focus on gene regulatory, metabolic and protein–protein interaction networks (PPINs), which are at the basis of all cellular processes, sparsely citing other kinds of networks when interesting for the discussion. A few technical definitions are provided in the Supplementary Material File 2 for the terms underlined in the text.

A PPIN (Figure 1A) has nodes corresponding to proteins and edges indicating their physical interaction. When a protein has more than one partner, the network is not able to tell if the different interactions take place together (as in a protein complex), or if they correspond to interactions taking place at different times.

An MN may be interpreted and built in various ways (Figure 1B): nodes can be metabolites or reactions (respectively giving rise to the compound and

Corresponding author. Matteo Brilli, INRIA, Université de Lyon, F-69000, Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne cedex, France. E-mail: matteo.brilli@univ-lyon1.fr; matteo.brilli.bip@gmail.com

Cecilia Klein is a PhD student at the University of Lyon 1 and at the INRIA, co-advised by Ana Tereza Vasconcelos and Marie-France Sagot. She is interested in computational biology and is currently studying the evolution of metabolic dialogs among symbiotic partners.

Andrea Marino is a PhD student at the University of Florence (Italy), advised by Pierluigi Crescenzi; he is visiting the INRIA Bamboo Team (Lyon, France). He is interested in algorithms and complexity, bioinformatics and complex networks analysis in general.

Marie-France Sagot is Director of Research at the French National Institute of Research in Computer Science and Control (INRIA) and head of the INRIA-CNRS-UCBL BAOBAB-BAMBOO teams.

Paulo Vieira Milreu is developing his PhD work at the UCBL (INRIA Bamboo Team), under the co-supervision of Vincent Lacroix, Christian Gautier and Marie-France Sagot. His interests are currently focused on understanding metabolic mechanisms through mathematical modeling strategies and algorithm design.

Matteo Brilli (post-doc INRIA, equipe Bamboo) is a biologist converted to bioinformatics: he is interested in comparative genomics, mainly to study the evolution of metabolism and gene regulation, mathematical modeling of integrated systems and biological network analysis.

the reaction graphs), and arcs (i.e. directed edges) can be reactions or shared metabolites. In both cases, the reconstruction may lead to a loss of fundamental information (Figure 1B). These limitations ask for a full treatment of complex reactions in an MN (discussed in detail e.g. in [1,2]): bipartite graphs and hypergraphs help to overcome these problems at the price of a higher algorithmic complexity. Hypergraphs are indeed generalizations of graphs and thus problems may become harder to solve (see [3] for some examples of hypergraphs applied to biological questions and the associated computational problems).

In a gene regulatory network (GRN; Figure 1C), nodes representing transcriptional regulators are connected to the nodes corresponding to their targets by signed arcs. The sign or weight of such arcs indicates the effect of the control. Because of combinatorial regulation whose output depends on the architecture of promoters which is not encoded in a basic GRN, an hypergraph representation could also represent a better choice for these networks [4–6].

With a biological network in hand, we can inspect many properties of the nodes or the edges/arcs searching for interesting features. Network metrics were mainly developed for nonbiological purposes, but in some cases they provided meaningful biological information (see sections below and Supplementary Material File 1). A more thorough description of the use of network metrics in biology is given in the following sections. Different measures focus on distinct properties of nodes or edges/arcs; hence, the choice of a meaningful metric depends on the type of network and on the question(s) asked. This task requires some knowledge on the biological processes modeled by the network because they strongly affect the interpretation or even the usefulness of a measure.

MNs can also be studied using quantitative constraint-based models that are able to identify the optimal distribution of fluxes in the network in a defined growth condition, at the expense of neglecting the dynamics to reach steady state [7]. The accessible structure of the network can therefore be proficiently used to obtain quantitative and testable information on the physiological state of a bacterium.

Although informative, the analysis of a static structure has its drawbacks. The first one is that we completely neglect any additional property the nodes (genes and proteins) may have, asking for an integration of those features into meaningful network metrics inspired by biology. The second drawback concerns the highly dynamic nature of biological networks: regulatory mechanisms active in different physiological states change the connectivity of the network, so that structural properties may be condition dependent. Another problem arises because a structural analysis is not always able to take into account regulatory mechanisms: the activity of enzymes is often regulated by one or more effector metabolites but since the latter are not consumed, the MN neglects such regulations (Figure 1B). This can have profound consequences because these regulations have important roles in stabilizing the metabolic states and in generating complex and biologically important dynamic behaviors [8–10].

These effectors are moreover able to cross the boundaries between different biological levels, such as metabolism and gene regulation. Building integrated models taking these cross-talks into account therefore represents a major challenge in systems biology. Previous modeling efforts have demonstrated that none of the different biological layers is truly isolated [11–13] and that enzymes also have regulatory functions, exerted through their control over the concentration of particular metabolites.

These considerations lead to a view of the cell as a network of networks, whose understanding requires considering regulatory interactions not only within, but also between biological networks.

STRUCTURAL ANALYSIS

In this section, we explore some topological metrics often used to analyze biological networks. In particular, we focus on centrality measures to predict essential genes, average distance (AD) and diameter to inspect the compactness of the network, assortativity and dyadicity to study the modularity of a network and any correlations between the properties of the nodes.

Before discussing these measures, let us stress that biases in the network reconstructions or manipulation can strongly affect the results of the analysis, confounding (if any exist) the observed correlations of biological and topological properties [14]. Consequently, we need to carefully interpret the topological measures obtained given that we only have a partial reconstruction in hand, and that

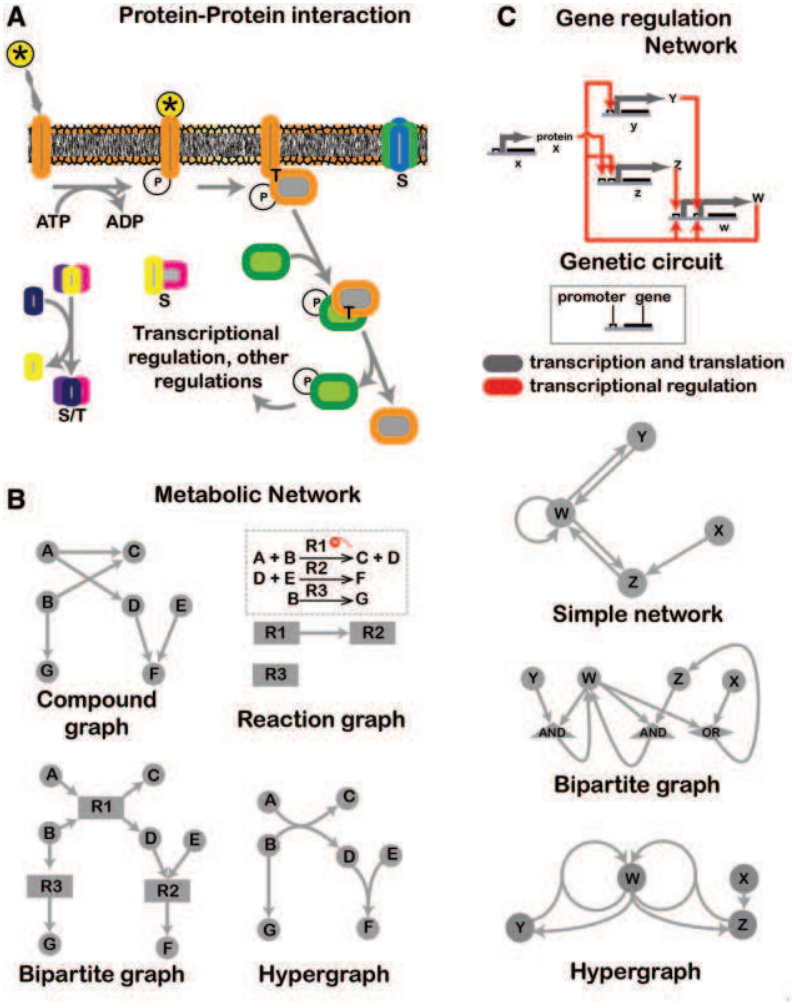


Figure 1: (**A**) An example of different kinds of interactions that build up a PPIN. A signal (asterisk) activates a receptor, which auto-phosphorylates and then passes the phosphate group to another protein (in Bacteria usually a Response Regulator), which is then able to regulate the activity of other proteins, or activate and repress gene expression. Interactions during this process are transient (T), therefore they are more difficult to detect using high-throughput technologies. Consequently, the PPIN is enriched in stable (S) interactions. (**B**) Graph models to represent an MN. Given three biochemical reactions (RI, R2, R3), metabolic graphs are built with metabolites as round nodes and reactions as square nodes. The enzyme catalyzing reaction RI has a metabolic regulatory feedback from compound C. The same system can be represented using different kinds of networks. Compound graph, where nodes are metabolites and there is an arc between a substrate and a product of a reaction; reaction graph, where nodes correspond to reactions and are connected when a product of one reaction is a substrate of the next one; bipartite graph: nodes are either compounds or reactions in which there is an arc between the substrate/reaction and reaction/product; hypergraph: nodes are compounds and a hyperarc links the substrate(s) to the product(s) of a reaction. The feedback from C to the enzyme catalyzing reaction RI is lost in all of these representations. Also, the compound and reaction graphs account for loss of information, e.g. reaction RI has two substrates (A and B) and two products (C and D), however, by looking at the corresponding compound graph one could imagine that the production of C only requires A, and by looking at the corresponding reaction graph we notice that the arc between RI and R2 exists only because of the compound D regardless of the presence of E. (**C**) A genetic circuit is a visual representation of a biological system and we provide three of its possible mathematical translations. The bipartite graph has nodes for proteins (circles) and different logical gates for combinatorial regulation: AND (triangle) requires the presence of both regulators to have transcription, while OR (diamond) can be activated by one of the regulators alone. The information on the promoter logics is lost in the Simple representation, while it is encoded in the hypergraph. The difference between these representations is evident if we suppose to remove regulator Z. By analyzing the Simple network, one may infer that the autoregulation of W continues to take place, which is not true, as correctly predicted by the bipartite graph and the hypergraph.

some of the measures described below are strongly affected by the sampling [15,16].

Centrality analysis

Given a network, it is natural to wonder how important each node is to its functionality. A number of graph measures have been developed for evaluating node centrality [17–21] and several tools allow to compute diverse network metrics, like CentiBiN [17], VisANT [22], Visone [23], Pajek [23], CentiScaPe [21] and CentiLib [24].

Centrality measures can be local (or neighborhood based) or global (distance or feedback based).

Local measures

With neighborhood-based measures, such as degree, the importance of the nodes is inferred from their local connectivity: the more connections a node has, the more central it is. Highly connected nodes (hubs) were found to possess special properties in the yeast PPIN: they are more often essential than non-hub proteins [25,26]), they tend to play a central role in the modular organization of a PPIN [27,28] and they seem to be evolutionarily more conserved [29]. Nevertheless, since then, several works have raised doubts on some of these associations [30,31].

There is no consensus in the literature on how to define a hub, and different criteria have been used: a given fraction of the highest degree nodes [32]; nodes with a given fraction of the total connectivity [33]; and a degree greater than an arbitrary threshold [28,34,35]. Recently, Vallabhajosyula *et al*. [36] proposed three objective functions allowing to define hubs in a PPIN in a rigorous way; unfortunately these are based on previous results on the properties of hubs in PPINs, limiting their applicability to other types of networks.

In order to have an indication about the homogeneity of the nodes of a network, it could be interesting to study the degree distribution that for most biological networks is well fitted by a power-law ($P(k) \sim k^{-\gamma}$) with $\gamma \sim 2$, where k is the degree. In these networks, a few hubs play a fundamental role for the integrity and navigability of the network [27], whereas a vast majority of the nodes has only a few connections. This degree distribution has been associated with robustness against random node removal. Robustness to the loss of a node in an MN indicates the presence of alternative pathways bypassing the missing reaction; in GRNs it may correspond to the presence of alternative ways of transmitting and controlling information. On the contrary, these networks are highly sensitive to attacks directed on hubs, because their removal deeply affects network functionality [37]. Even though much research has been done on the power-law distribution and its universality in biological networks, criticisms have been raised [38]. Power-law degree distributions indeed can be obtained through random sampling of networks with different topologies, indicating that it might not be possible to infer the true degree distribution from biological networks, for which complete reconstructions are usually not available [39].

The local connectivity of nodes can be studied in further detail by using either assortativity or dyadicity. The first measure represents the correlation between the degree of adjacent nodes [40]. Maslov and Sneppen [41] found that hubs in the yeast PPIN are mostly connected to non-hubs, and are therefore well separated from each other. Dyadicity [42] measures the degree to which the nodes of a network are connected to other nodes that share some characteristic (functional classification, essentiality, involvement in a disease and so on) and is therefore able to characterize the modular structure of a network by considering the distribution of the functions over the nodes and their connectivity [43]. A network is called heterophilic (heterophobic) when different categories are connected more (less) often than expected under a random model. It was recently used to study the potential coupling between structure and functionality in transcriptional and noncoding (nc) RNA–protein interaction networks [44]. The results showed that most transcriptional regulators and ncRNAs tend to connect to genes/proteins of other functional classes, suggesting that regulators do not really belong to a functional class but tend instead to coordinate several of them [44]. On the converse, in PPINs and MNs, the connections more often involve proteins of a same functional category.

Global measures

Closeness [45] and shortest path-based betweenness [46] reflect global properties of a network and use a distance measure between nodes, often the shortest path. The closeness of a node depends on its AD from the others and is of particular interest for information networks (such as signaling network and GRNs) as it measures how fast information flows from a node of interest to all the reachable nodes

[47]. It has been recently integrated with biological information in a parameter-free gene prioritization approach that computes the interconnectedness (ICN) between genes in a network [48]. ICN measures closeness of each candidate gene to genes possessing an interesting property by considering alternative paths in addition to the direct link and shortest one.

Shortest path-based betweenness depends on the number of shortest paths crossing a node. In PPINs, betweenness can be interpreted as the relevance of a protein to be intermediary in the interaction between other proteins, assuming that this interaction passes through shortest paths [21]. Bottlenecks are nodes with high betweenness centrality and were found to be key connectors with surprising functional and dynamical properties, often essential [49]. Bottleneck and hub genes were identified in coexpression networks inferred from experimental data, and found to be often essential for virulence in *Salmonella typhimurium* with the role of mediators of transitions between different cellular states or of sentinels that reflect the dynamics of these transitions [50]. Cell cycle checkpoints were found to be bottlenecks in a gene coexpression network of cell cycle regulated genes in the fission yeast [51].

Network metrics in general [52–54], and betweenness centrality in particular are also used for the rational prediction of drug targets [55]. Essential genes are preferred targets for drug design and central genes are more likely to be essential. Another constraint was imposed in this particular case: the gene must be essential for the pathogen but not for the host to reduce any side effects of the drug.

One problem with shortest path-based measures is that communication between biological entities is assumed to pass along those paths, which is often not plausible: from the point of view of MNs, the shortest path might be defined on the basis of the energy/cofactor requirements instead of the number of steps, whereas in GRNs and PPINs all active connections will take place, not only the shortest ones. In the case of GRNs, the targets with different shortest paths to a common regulator may exhibit hierarchical gene expression patterns as is the case for flagellar genes [56].

To overcome the limitation of shortest paths, a node can be considered central when it is crossed by many random walks: this is the case of the random walk-based betweenness centrality [57]. Some feedback-based measures implicitly rely on random walks, like eigenvector [58] and spectral centrality [59]. Eigenvector centrality has been applied to several MNs [60] and was shown to outperform other metrics for the identification of essential proteins in the PPIN of yeast [61], together with subgraph centrality [62].

Distance analysis

The diameter of a network is an overall indication of its compactness. Despite the fact that real networks sometimes exhibit the small-world property and that shorter diameters may be beneficial to some networks (e.g. for rapid information flow), it was shown that several biological networks have larger diameters than their randomizations. One possible reason for this is their modular nature [63] leading to the suggestion that modularity may be a universal characteristic of real networks, due to the advantages it brings to multi-functionality, robustness and evolvability. On one hand, high modularity reduces pleiotropic effects improving the evolvability of the system. On the other, numerical experiments also demonstrated that modularization provides robustness against random perturbations in network structure, i.e. evolutionary change [64].

The distribution of distances and the AD may be more informative than the diameter about the global properties of a network [63]. The small AD commonly observed in biological networks pertains to the so-called small-world effect [65]. The AD ranged between 3 and 5 in 43 MNs of 200–800 nodes [66], showing that all nodes are quite close to each other. Although several groups confirmed the small-world property of the MN of different organisms [67–71], Arita [72] heavily criticized the way the pathways are computed in those works since they do not conserve their structural moieties. When this problem is accounted for correctly, the analysis revealed that the average path length of the *Escherichia coli* metabolism is much longer than previously thought [72,73].

Quantitative structural analysis

Flux Balance Analysis (FBA; Figure 2) is a quantitative modeling technique that relies on a validated reconstruction of an MN, the steady-state assumption and additional constraints [74–76].

The target of the method is obtaining the flux distribution within the MN under specified growth conditions (Figure 2).

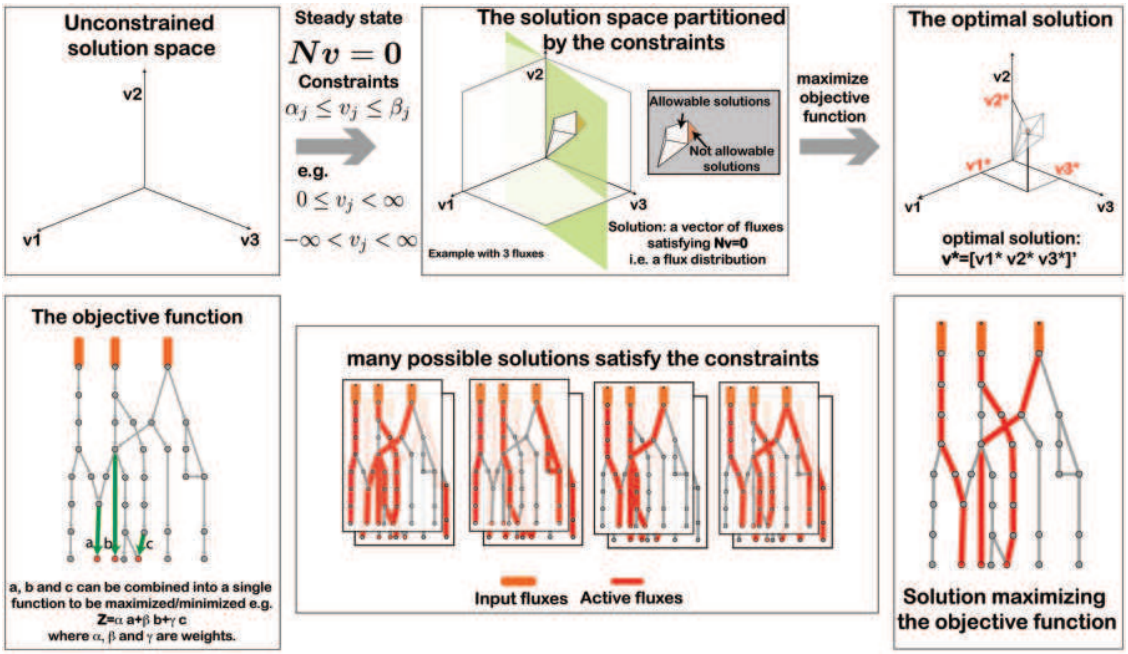


Figure 2: FBA is a constraint-based model based on the stoichiometric modeling of an MN, a (quasi) steady-state condition and an objective function. The constraints are the reaction set of the network encoded in the stoichiometric matrix N and additional thermodynamic and environmental constraints. The steady-state condition for MNs corresponds to a regime where the intracellular fluxes and metabolite concentrations are constant in time ($Nv = 0$), where v is a vector representing a flux distribution for the reactions. There are many flux distributions satisfying the steady-state condition and the other constraints. In FBA experiments, the interest is the identification of the flux distribution that maximizes/minimizes a given objective function.

The stoichiometry of the reactions encode the mass conservation rules, and a modeling of the environment through transport reactions impose constraints on the possible flux distributions satisfying the steady-state condition; additional constraints may also be added such as reaction reversibility and maximum velocity of enzymes. Since the solution space for such models is very large even under the constraints used, FBA seeks an optimal flux distribution with respect to a carefully chosen objective function using optimization techniques. The assumption behind FBA is that metabolism maximizes some objective, but there may exist many suboptimal flux distributions that help the organism during adaptation to specific environmental conditions. This led to elementary mode analysis [77], which seeks for the solutions satisfying the above constraints regardless of the objective function. Elementary modes can be loosely defined as the smallest subnetworks allowing an MN to function in steady state [78,79]. According to Stelling *et al.* [79], they can be used to understand cellular objectives for an overall MN.

The objective function plays a fundamental role in FBA as it provides a way to choose one optimal solution: assuming that the objective of *E. coli* in rich medium is to grow at maximum speed, we may formulate an objective function that combines fluxes exiting the MN to produce biomass. Optimization through integer linear programming [7,80] then allows to identify one optimal solution which is a physiological steady state of the MN of an organism in that condition. When the target is maximization of the production of some compound, the compound is usually included in the objective function to enforce solutions where its production is active. Other formulations for the objective function may be designed to mimic disparate growth conditions, not necessarily focusing on fast growth [81–91].

Biologically speaking, solutions obtained through FBA describe a partition of the input fluxes into the different branches of the network to produce the compounds required by growth (through the objective function).

One of the most appealing properties of constraint-based models is that they provide a way to explore the consequences of genetic manipulations on the whole MN: one or more reactions can be eliminated (simulating knock-out mutants) [92–95]

or otherwise manipulated, and simulations can be run to see if and how the objective function can be improved with respect to the wild-type model [96]. By coupling two levels of optimization, it is possible to predict the best engineering strategy to have mutants that maximize some by-product of interest, such as ethanol [96] or lactate [97], while growing. A recent survey on FBA and its applications can be found in [98].

Dynamic analysis

Dynamic analysis of structural properties

In general, we look at biological networks as static entities, but it should be stressed that they are instead very dynamic at widely different time-scales. They are dynamic in evolutionary time like any other biological structure, and even more on short time-scales, since regulatory connections and feedbacks change the connectivity of the network depending on the physiological state (Figure 3). Consequently, we should interpret most of the currently available biological network reconstructions as potential networks, where all the possible connections are indicated. By the term potential, we highlight the fact that edges/arcs and nodes in this network will be hardly present all together *in vivo*. If we consider for instance a PPIN, not all interaction partners of a protein will be expressed in a given condition, reducing the number of actual partners. Conversely, we may speak of network realizations when focusing on the active subgraph of a potential network, defined on the basis of experimental data [28,99–101]. The dynamic nature of biological networks is also at the basis of differential network analysis [102], which aims at capturing the subgraphs specific of a given network realization.

These considerations are important since they affect the analysis of biological networks. As there are many condition-specific realizations of a biological network, they plausibly have different structural properties. It was indeed shown that random subgraphs of a network do not necessarily maintain the same-degree distribution as the entire network [103], suggesting that other structural properties may also change (Figure 4).

Therefore, it is not clear if we can look for 'universal' properties of biological networks by analyzing potential networks, or whether we should instead define as 'universal' those properties that characterize most realizations.

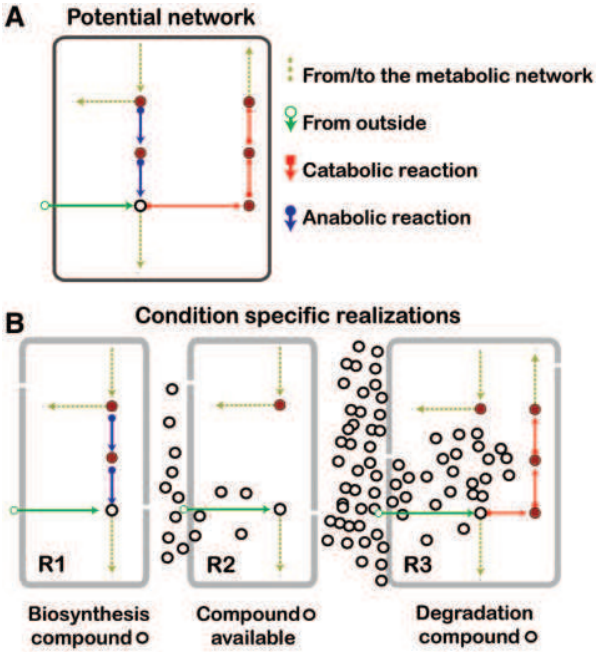


Figure 3: Illustrative example on the potential and realization concept concerning the anabolic and catabolic pathways of a same compound (4). (**A**) The potential network. (**B**) The realizations are shown for different physiological states: RI, biosynthetic state for compound o. R2 compound o is available and its biosynthetic route is off. R3 catabolic state: a degradation pathway is activated to reduce the intracellular concentration of the compound.

Han *et al.* [28] estimated the temporal connectivity of hubs in the yeast PPIN by using gene expression data: the correlation in gene expression between two connected nodes in the potential network allowed to define two types of hubs: party hubs, interacting with their partners simultaneously; and date hubs, which bind their different partners at different times or locations. It is then plausible to do the same for other measures: genes may be central in the potential network and frequently or not in the realizations (party and date centers); party and date bottlenecks may be defined in the same way, and so on. This additional level of complexity may allow a deeper understanding of how physiological transitions are driven by topological changes.

Gene expression was integrated in a centrality measure called Pec [104], which was used to identify essential genes in yeast. This measure exploits the strength of the connectivity between two adjacent nodes based on an Edge Clustering Coefficient [105], weighted by the co-expression between genes in experimental data.

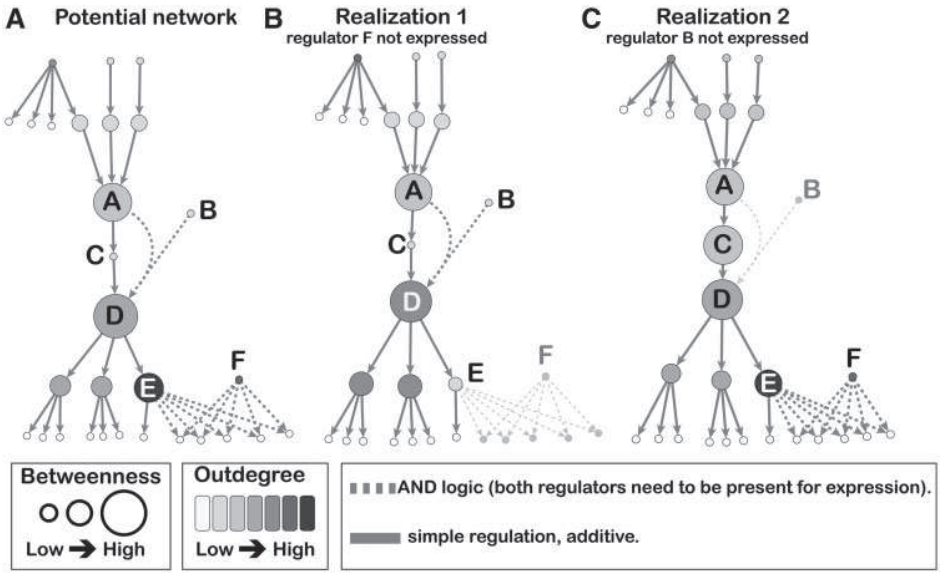


Figure 4: Centrality measures change in GRN realizations. Nodes have a size proportional to the betweenness centrality measure and the color of a node changes according to the outdegree. The pairs of regulators A and B as well as E and F are both required for the activation of the target gene(s). (**A**) The potential network, where regulators A and D are central following betweenness centrality, and E with respect to outdegree centrality. Now let us suppose to use experimental data to obtain two realizations of this potential network. In (**B**) regulator F is not expressed, and regulator E has consequently a low outdegree. In (**C**) regulator B is inactive, imposing a remarkable change in the betweenness centrality value of regulator C.

This reasoning also affects the evolutionary interpretation of network properties, for instance when concluding that evolution promoted the fixation of a given structural feature of the potential network. Luscombe *et al.* [99] analyzed the structural properties of the yeast GRN in different conditions. Starting from a validated GRN, they used gene expression data to extract the subnetworks supposed to be active during environmental stress or the cell cycle, highlighting important differences: the cell cycle subnetwork has long shortest paths and combinatorial regulation is common, whereas short paths and mainly single-input regulations characterize the stress condition. The length of a path may be relevant in the context of a GRN because it can be interpreted as a measure of the delay to have a response once the top regulator is activated (Figure 1B). The short paths for the stress conditions suggest evolution of a fast response to stressors, whereas cell cycle evolved under the necessity for fine regulations giving the correct temporal ordering of events, which explains the combinatorial regulation (information integration) and the longer paths (check points). Performing the analysis on the potential network, these differences would not have been noticed. The previous work has however been heavily criticized [99], but both studies conclude that realization networks can be largely different in their structural properties (see also [28,101]).

The use of realization networks is currently limited by the need for high-quality and high-throughput experimental data, today available only for a few organisms. Nevertheless, large-scale experimental data will be more easily obtained in the future, giving the occasion to develop the algorithms required for a similar approach.

Kinetic modeling of full-scale networks

In the previous section, we discussed how to explore the structural properties of a biological network using experimental data to define the active subgraphs in a potential network. However, the analysis is not really dynamic, but gives instead only a snapshot of the steady states of a network in different conditions. To move forward with the dynamic analysis of networks, we discuss the mathematical modeling of biochemical reaction networks from the perspective of building large, network-scale models able to predict the dynamics between different states. Many different modeling strategies were devised and

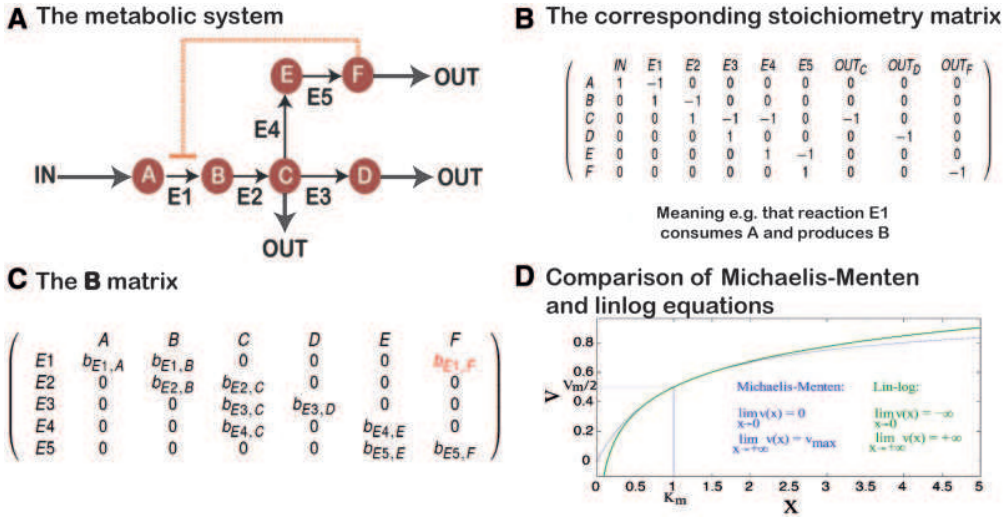


Figure 5: (**A**) A metabolic system. (**B**). The corresponding stoichiometry matrix N. The evolution in time of the six metabolite concentrations is given by: $dx/dt = Nv(e,f(x,p))$, where x is the vector of metabolite concentrations and $v(e,f(x,p))$ is a vector of rates, functions depending on enzyme levels e and on metabolites in x, including the effectors. The latter dependencies are not encoded in the stoichiometry matrix. $f(x,p)$ can take many different forms, e.g. mass action, Michaelis–Menten or linlog. (**C**) The parameter matrix of the linlog approximation of the entire system; all the rate functions have the same standard format, a linear combination of logarithmic metabolite concentrations i.e. $v = \mathrm{diag}(e)\,(A + B \log X)$, with A and B a vector and a matrix of parameters, respectively. (**D**) Comparison of the irreversible Michaelis–Menten (V_{max} [S]/(K_m+[S])) and corresponding linlog: linlog is not saturable for large substrate concentrations, and gives minus infinite fluxes when one of the metabolites in a given reaction goes to zero.

described elsewhere [4,8,106–119]; here we briefly discuss the modeling of biochemical networks (MN and GRN) and its application to cellular scale systems. Some of the discussions also apply to signaling systems, which combine different types of regulation (protein–protein interaction, phosphorylation and transcriptional regulation).

Kinetic metabolic models are traditionally based on systems of ordinary differential equations where the rates modeling the activity of an enzyme are mechanistic, nonlinear and more or less precisely describe the catalytic mechanism of an enzyme. The activity of promoters in gene regulation is usually modeled using sigmoid functions as suggested by experimental data [120,121], and combination thereof in the presence of combinatorial regulation [4]. The parameters of these models are usually derived from *in vitro* (rarely *in vivo*) experiments but the large differences between *in vivo* and *in vitro* conditions have called into question this approach [122–125], and *in vivo* experiments should be preferred [126]. The main drawback of building such detailed models is therefore that it is very time-consuming for the amount of good quality and informative experimental data required to perform parameter identification. Mechanistic models have been consequently applied mainly to well-studied systems, and only recently models for less studied ones have started being implemented [127–131].

All these limitations make it impossible at the moment to build mechanistic models at a full network scale. The only exception for MNs is a work by Jamshidi and Palsson [132], who use mass action kinetics to build a model of the MN of red blood cells with 100 chemical reactions (catalytic or regulatory), and 95 variables. To overcome the limits imposed by mechanistic models, approximative nonmechanistic rate equations have been developed for both metabolic (e.g. [113–115]) and gene regulation systems [4]. The main advantage of approximated formalisms is that they require less parameters, reducing as well the experimental effort for parameter identification. One of these approximations is called linlog, and was recently used to model a network-scale MN of yeast [133]. The parameters were obtained from a model repository (see Figure 5 for more details on this

approximation). The resulting model contains 956 metabolic reactions and 820 metabolites; the key steps were identified using metabolic control analysis. This modeling framework may be considered a stepping-stone towards the long-term goal of a fully parameterized model of genome-scale metabolism even if its performance needs to be improved.

GRNs also cannot be modeled at a full scale, since much of the information required is not available, and approximated formalisms were proposed [4]. We stress that obtaining a GRN is much more difficult than obtaining an MN; the methods give moreover very partial reconstructions that strongly affect the structural analysis [16].

Modeling network scale integrated systems

An important and ambitious challenge in systems biology is building integrated models where the interactions between different biological layers are explicitly taken into account. We here consider the case of integrated models where metabolism is modeled together with the gene regulation system, but it should be noticed that increasing experimental evidence suggests further integration of signaling pathways and GRNs with regulation mediated by ncRNAs [134–138]. On one hand, integration of metabolism and gene regulation might allow to study a much wider range of situations using a same model, and on the other, it allows to study more in detail the importance of the cross-talk between the two systems. A first effort to measure the effect of regulation in FBA predictions through the addition of Boolean logic time-dependent constraints modeling transcriptional regulatory events is regulatory FBA (rFBA; [139]). rFBA changes the shape of the solution space considerably with respect to FBA, finding physiologically relevant solutions [139]. These initial methods were improved by several recent works such as steady-state regulatory FBA (SR-FBA), which is an integrated regulatory-metabolic model for predicting gene expression and metabolic fluxes [140], integrated FBA (iFBA) that combines rFBA and inferred ordinary differential equations [141], OptFlux which is a software for strain prediction through metabolic/regulatory integrated data [142], and hybrid modeling [143]. For a more detailed review on different coupled regulatory/metabolic models, we refer to [144].

CONCLUSIONS

Structural analysis allows the identification of important nodes within a network and for this reason, has become very popular in many disciplines. However, in the biological domain, the importance of a node can be defined in many different ways so that identifying the most appropriate network measures is an important preliminary step that can radically change the output of an analysis. It is then essential to understand the meaning of a given measure with respect to the specific network at hand.

Besides discussing some of the most informative metrics for biological networks analysis, we stress the importance of a biologically meaningful interpretation of any measure, which is not always intuitive and can change for different networks.

The dynamical nature of biological networks indicates that it may be better to perform structural analysis on what we have defined as the realizations of a network. The risk when studying a potential network is confounding the signals encoded in the network by putting everything together. Are we sure that a metabolic hub is a hub in every realization of the network? What if it is lowly connected with different nodes in every realization? This approach is today limited by the availability of experimental data, but databases are growing fast and a similar analysis would be feasible for several prokaryotes, as well as for a few eukaryotes.

Concerning the more biologically oriented interpretation of the metrics, it requires to move the collaboration between computational and experimental biologists to a higher level. It would also contribute to the integration of biological information in network analysis, which is a topical challenge in the field. Let us take the example of hubs in a GRN. From the biological point of view, it is clearly different if the hub controls a single cellular function or affects widely different processes. Since a GRN transmits information, a similar approach would require being able to define the scope of a regulator by also taking into account indirect targets (similarly to [6]). This example illustrates the need for biologically oriented network metrics that are able to take into account the heterogeneous information associated with biological entities. As pointed out by Keller [145], Watts and Strogatz (65) have proficiently used simple mathematical models to study social networks, but some of their most interesting results emerged only after they took into account the

property that sociologists consider as fundamental to social dynamics: social identity. The challenge is to do the same with biological networks, which requires an effort to develop meaningful metrics able to account for and integrate biological properties.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bfg.oxfordjournals.org/.

Key Points

- Structural analysis of biological networks allows to identify genes and proteins playing important roles in cellular physiology.
- Biological networks are dynamic; the structural properties of genes and proteins are consequently also dynamic, i.e. the importance of a protein might change depending on the growth condition.
- The dynamics of biological systems can be studied using detailed mathematical modeling, but they are not easily scalable at the network level and approximations have been provided that might simplify the task.

FUNDING

This work was funded by the French project ANR MIRI BLAN08-1335497, and the ERC Advanced Grant Sisyphe held by M.-F.S. It was supported by the INRIA International Partnership AMICI.

References

1. Lacroix V, Cottret L, Thébault P, *et al*. An introduction to metabolic networks and their structural analysis. *IEEE/ACM Transact Comput Biol Bioinform* 2008;**5**:594–617.
2. Cottret L, Jourdan F. Graph methods for the investigation of metabolic networks in parasitology. *Parasitology* 2010;**137**:1393–1407.
3. Klamt S, Haus U-U, Theis F. Hypergraphs and cellular networks. *PLoS Comput Biol* 2009;**5**:e1000385.
4. de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol* 2002;**9**:67–103.
5. Klamt S, Saez-Rodriguez J, Lindquist JA, *et al*. A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics* 2006;**7**:56.
6. Wang R, Albert R. Elementary signaling modes predict the essentiality of signal transduction network components. *BMC Syst Biol* 2011;**5**:44.
7. Edwards JS, Ibarra RU, Palsson BØ. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol* 2001;**19**:125–30.
8. Steuer R, Gross T, Selbig J, *et al*. Structural kinetic modeling of metabolic networks. *Proc Natl Acad Sci USA* 2006;**103**:11868–73.
9. Grimbs S, Selbig J, Bulik S, *et al*. The stability and robustness of metabolic states: identifying stabilizing sites in metabolic networks. *Mol Syst Biol* 2007;**3**:146.
10. Steuer R. Computational approaches to the topology, stability and dynamics of metabolic networks. *Phytochemistry* 2007;**68**:2139–51.
11. Baldazzi V, Ropers D, Markowicz Y, *et al*. The carbon assimilation network in Escherichia coli is densely connected and largely sign-determined by directions of metabolic fluxes. *PLoS Comput Biol* 2010;**6**:e1000812.
12. Baldazzi V, Ropers D, Geiselmann J, *et al*. Importance of metabolic coupling for the dynamics of gene expression following a diauxic shift in Escherichia coli. *J Theor Biol* 2011;**295**:100–15.
13. Kotte O, Zaugg JB, Heinemann M. Bacterial adaptation through distributed sensing of metabolic fluxes. *Mol Syst Biol* 2010;**6**:355.
14. Coulomb S, Bauer M, Bernard D, *et al*. Gene essentiality and the topology of protein interaction networks. *Proc Biol Sci Roy Soc* 2005;**272**:1721–5.
15. Costenbader E, Valente TW. The stability of centrality measures when networks are sampled. *Soc Network* 2003;**25**:283–307.
16. de Silva E, Thorne T, Ingram P, *et al*. The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol* 2006;**4**:39.
17. Junker BH, Koschützki D, Schreiber F. Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics* 2006;**7**:219.
18. Koschützki D, Schreiber F. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regulat Syst Biol* 2008;**2**:193–201.
19. Pavlopoulos GA, Secrier M, Moschopoulos CN, *et al*. Using graph theory to analyze biological networks. *BioData Mining* 2011;**4**:10.
20. Mason O, Verwoerd M. Graph theory and networks in Biology. *IET Syst Biol* 2007;**1**:89.
21. Scardoni G, Laudanna C. Centralities based analysis of complex networks. In: Zhang Y (ed). *New Frontiers in Graph Theory*. InTech Open, 2012.
22. Hu Z, Hung J-H, Wang Y, *et al*. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucl Acids Res* 2009;**37**:W115–21.
23. Baur M, Benkert M, Brandes U, *et al*. Visone - software for visual social network analysis. *Proceeings of the 9th Internatinal Symposium on Graph Drawing (GD '01)*, 2001.
24. Grassler J, Koschutzki D, Schreiber F. CentiLib: comprehensive analysis and exploration of network centralities. *Bioinformatics* 2012;**28**:1178–9.
25. Jeong H, Mason SP, Barabási A-L, *et al*. Lethality and centrality in protein networks. *Nature* 2001;**411**:41–2.
26. He X, Zhang J. Why do hubs tend to be essential in protein networks? *PLoS Genet* 2006;**2**:e88.
27. Albert R, Jeong H, Barabasi A. Error and attack tolerance of complex networks. *Nature* 2000;**406**:378–82.
28. Han JDJ, Bertin N, Hao T, *et al*. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* 2004;**430**:88–93.
29. Wuchty S, Almaas E. Peeling the yeast protein network. *Proteomics* 2005;**5**:444–9.

30. Wuchty S. Interaction and domain networks of yeast. *Proteomics* 2002;**2**:1715–23.
31. Zotenko E, Mestre J, O'Leary DP, *et al*. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol* 2008;**4**: e1000140.
32. Batada NN, Reguly T, Breitkreutz A, *et al*. Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS Biol* 2006;**4**:e317.
33. Reguly T, Breitkreutz A, Boucher L, *et al*. Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol* 2006;**5.**
34. Ekman D, Light S, Björklund AK, *et al*. What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? *Genome Biol* 2006;**7**:R45.
35. Aragues R, Sali A, Bonet J, *et al*. Characterization of protein hubs by inferring interacting motifs from protein interactions. *PLoS Comput Biol* 2007;**3**:1761–71.
36. Vallabhajosyula RR, Chakravarti D, Lutfeali S, *et al*. Identifying hubs in protein interaction networks. *PloS One* 2009;**4**:e5344.
37. Barabási A-lászló, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13.
38. Lima-Mendez G, van Helden J. The powerful law of the power law and other myths in network biology. *Mol BioSys* 2009;**5**:1482–93.
39. Han J-DJ, Dupuy D, Bertin N, *et al*. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol* 2005;**23**:839–44.
40. Newman MJ. Assortative mixing in networks. *Phys Rev Lett* 2002;**89**:1–4.
41. Maslov S, Sneppen K. Specificity and stability in topology of protein networks. *Science* 2002;**296**:910–3.
42. Park J, Barabási A-L. Distribution of node characteristics in complex networks. *Proc Natl Acad Sci USA* 2007;**104**: 17916–20.
43. Jiang X, Liu B, Jiang J, *et al*. Modularity in the genetic disease-phenotype network. *FEBS Lett* 2008;**582**: 2549–54.
44. Nacher J, Araki N. On the relation between structure and biological function in transcriptional networks and ncRNA-mediated interactions. *Intl Conf Biosci Biochem Bioinform* 2011;**5**:348–352.
45. Latora V, Marchiori M. A measure of centrality based on network efficiency. *New J Phys* 2007;**9**:188.
46. Freeman L. A set of measures of centrality based on betweenness. *Sociometry* 1977;**40**:35–40.
47. Li X-li, Ng S-K. Biological data mining in protein interaction networks. Hershey: IGI Global, 2009.
48. Hsu C-L, Huang Y-H, Hsu C-T, *et al*. Prioritizing disease candidate genes by a gene interconnectedness-based approach. *BMC Genomics* 2011;**12**(Suppl. 3):S25.
49. Yu H, Kim PM, Sprecher E, *et al*. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* 2007;**3**:e59.
50. McDermott JE, Taylor RC, Yoon H, *et al*. Bottlenecks and hubs in inferred networks are important for virulence in Salmonella typhimurium. *J Comput Biol* 2009;**16**:169–80.
51. Caretta-Cartozo C, De Los Rios P, Piazza F, *et al*. Bottleneck genes and community structure in the cell cycle network of S. pombe. *PLoS Comput Biol* 2007;**3**:e103.
52. Vallabhajosyula RR, Raval A. Computational modeling in systems biology. *Syst Biol Drug Discov Dev* 2010;**662.**
53. Chavali AK, Blazier AS, Tlaxca JL, *et al*. Metabolic network analysis predicts efficacy of FDA-approved drugs targeting the causative agent of a neglected tropical disease. *BMC Syst Biol* 2012;**6**:27.
54. Chen L-C, Yeh H-Y, Yeh C-Y, *et al*. Identifying co-targets to fight drug resistance based on a random walk model. *BMC Syst Biol* 2012;**6**:5.
55. Rahman SA, Schomburg D. Observing local and global properties of metabolic pathways: "load points" and "choke points" in the metabolic networks. *Bioinformatics* 2006;**22**:1767–74.
56. Smith TG, Hoover TR. Deciphering bacterial flagellar gene regulatory networks in the genomic era. *Adv Appl Microbiol* 2009;**67**:257–95.
57. Newman MJ. A measure of betweenness centrality based on random walks. *Soc Network* 2005;**27**:39–54.
58. Bonacich P. Some unique properties of eigenvector centrality. *Soc Net* 2007;**29**:555–64.
59. Perra N, Fortunato S. Spectral centrality measures in complex networks. *Phys Rev E* 2008;**78**:1–10.
60. Ding D-wu, He X-qing, Science C. Application of eigenvector centrality in metabolic networks. In *2010 2nd International Conference on Computer Engineering and Technology*, 2010. 89–91.
61. Estrada E. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics* 2006;**6**: 35–40.
62. Estrada E, Rodríguez-Velázquez JA. Subgraph centrality in complex networks. *Phys Rev E* 2005;**71**:056103.
63. Zhang Z, Zhang J. A big world inside small-world networks. *PloS One* 2009;**4**:e5686.
64. Kashtan N, Alon U. Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci USA* 2005;**102**: 13773–8.
65. Watts DJ, Strogatz SH. Collective dynamics of "small-world" networks. *Nature* 1998;**393**:440–2.
66. Jeong H, Tombor B, Albert R, *et al*. The large-scale organization of metabolic networks. *Nature* 2000;**407**:651–4.
67. Fell DA, Wagner A. The small world of metabolism. *Nat Biotechnol* 2000;**18**:1121–2.
68. Wagner A, Fell DA. The small world inside large metabolic networks. *Proc Biol Sci Roy Soc* 2001;**268**:1803–10.
69. Ravasz E, Somera AL, Mongru DA, *et al*. Hierarchical organization of modularity in metabolic networks. *Science* 2002;**297**:1551–5.
70. Ma H, Zeng AP. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* 2003;**19**:270.
71. Yook S-H, Oltvai ZN, Barabási A-L. Functional and topological characterization of protein interaction networks. *Proteomics* 2004;**4**:928–42.
72. Arita M. The metabolic world of Escherichia coli is not small. *Proc Natl Acad Sci USA* 2004;**101**:1543–7.
73. Pitakänen E, Rantanen A, Rousu J, *et al*. Finding feasible pathways in metabolic networks. *Proceedings of the 10th*

Panhellenic Conference on Informatics (PCI'2005), Lecture Notes in Computer Science, 2005.

74. Price N. Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol* 2003;**21**: 162–9.
75. Varma A, Palsson BO. Metabolic flux balancing: basic concepts, scientific and practical use. *Nat Biotechnol* 1994;**12**: 994–8.
76. Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. *Curr Opin Biotechnol* 2003;**14**:491–6.
77. Schuster S, Dandekar T, Fell DA. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol* 1999;**17**:53–60.
78. Schuster S, Fell DA, Dandekar T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol* 2000;**18**:326–32.
79. Stelling J, Klamt S, Bettenbrock K, *et al*. Metabolic network structure determines key aspects of functionality and regulation. *Nature* 2002;**420**:190–3.
80. Chvatal V. Linear Programming. New York: W. H. Freeman, 1983.
81. Perumal D, Samal A, Sakharkar KR, *et al*. Targeting multiple targets in Pseudomonas aeruginosa PAO1 using flux balance analysis of a reconstructed genome-scale metabolic network. *J Drug Target* 2011;**19**:1–13.
82. Boyle NR, Morgan JA. Flux balance analysis of primary metabolism in Chlamydomonas reinhardtii. *BMC Syst Biol* 2009;**3**:4.
83. Puchalka J, Oberhardt MA, Godinho M, *et al*. Genome-scale reconstruction and analysis of the Pseudomonas putida KT2440 metabolic network facilitates applications in biotechnology. *PLoS Comput Biol* 2008;**4**:e1000210.
84. Resendis-Antonio O, Reed JL, Encarnación S, *et al*. Metabolic reconstruction and modeling of nitrogen fixation in Rhizobium etli. *PLoS Comput Biol* 2007;**3**:1887–95.
85. Radhakrishnan D, Rajvanshi M, Venkatesh KV. Phenotypic characterization of Corynebacterium glutamicum using elementary modes towards synthesis of amino acids. *Syst Synth Biol* 2010;**4**:281–91.
86. Oh Y-K, Palsson BØ, Park SM, *et al*. Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J Biol Chem* 2007;**282**:28791–9.
87. Mahadevan R, Palsson BØ, Lovley DR. In situ to in silico and back: elucidating the physiology and ecology of Geobacter spp. using genome-scale modelling. *Nat Rev Microbiol* 2010;**9.**
88. Wittmann C, Heinzle E. Modeling and experimental design for metabolic flux analysis of lysine-producing Corynebacteria by mass spectrometry. *Metabolic Eng* 2001; **3**:173–91.
89. Poolman MG, Venkatesh KV, Pidcock MK, *et al*. A method for the determination of flux in elementary modes, and its application to Lactobacillus rhamnosus. *Biotechnol Bioeng* 2004;**88**:601–12.
90. Risso C, Sun J, Zhuang K, *et al*. Genome-scale comparison and constraint-based metabolic reconstruction of the facultative anaerobic Fe(III)-reducer Rhodoferax ferrireducens. *BMC Genomics* 2009;**10**:447.
91. Sun J, Sayyar B, Butler JE, *et al*. Genome-scale constraint-based modeling of Geobacter metallireducens. *BMC Syst Biol* 2009;**3**:15.
92. Suthers PF, Zomorrodi A, Maranas CD. Genome-scale gene/reaction essentiality and synthetic lethality analysis. *Mol Syst Biol* 2009;**5**:301.
93. Wunderlich Z, Mirny L. Using topology of the metabolic network to predict viability of mutant strains. *Genome Biol* 2005;**6**:P15.
94. Pharkya P, Burgard AP, Maranas CD. OptStrain: a computational framework for redesign of microbial production systems. *Genome Res* 2004;**14**:2367–76.
95. Pharkya P, Burgard AP, Maranas CD. Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. *Biotechnol Bioeng* 2003;**84**:887–99.
96. Trinh CT, Unrean P, Srienc F. Minimal Escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses. *Appl Env Microbiol* 2008;**74**:3634–43.
97. Fong SS, Burgard AP, Herring CD, *et al*. In silico design and adaptive evolution of Escherichia coli for production of lactic acid. *Biotechnol Bioeng* 2005;**91**:643–8.
98. Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. *Brief Bioinform* 2009;**10**: 435–49.
99. Luscombe NM, Babu MM, Yu H, *et al*. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 2004;**431**:308–12.
100. Konagurthu AS, Lesk AM. Single and multiple input modules in regulatory networks. *Proteins* 2008;**73**:320–4.
101. Gopalacharyulu PV, Velagapudi V, Lindfors E. Dynamic network topology changes in functional modules predict responses to oxidative stress in yeast. *Mol Bio Syst* 2009.
102. Ideker T, Krogan NJ. Differential network biology. *Mol Syst Biol* 2012;**8**:1–9.
103. Stumpf MPH, Wiuf C, May RM. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci USA* 2005;**102**:4221–4.
104. Li M, Zhang H, Wang J, *et al*. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Syst Biol* 2012;**6**:15.
105. Wang J, Li M, Wang H, *et al*. Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Transact Comput Biol Bioinform* 2011;**6.**
106. Zheng Y, Sriram G. Mathematical modeling: bridging the gap between concept and realization in synthetic biology. *J Biomed Biotechnol* 2010;**2010**:541609.
107. Bellouquid A, Delitala M. Mathematical modeling of complex biological systems - a kinetik theory approach 2006;**188.**
108. Allman ES, Rhodes JA. Mathematical models in biology, an introduction. Cambridge (UK): Cambridge University press, 2004.
109. Goutsias J, Kim S. A nonlinear discrete dynamical model for transcriptional regulation: construction and properties. *Biophys J* 2004;**86**:1922–45.
110. Gao J, Li L, Wu X, *et al*. BioNetSim: a Petri net-based modeling tool for simulations of biochemical processes. *Protein Cell* 2012;**3**:225–9.
111. Pozo C, Marin-Sanguino A, Alves R, *et al*. Steady-state global optimization of metabolic non-linear dynamic

models through recasting into power-law canonical models. *BMC Syst Biol* 2011;**5**:137.

112. Maus C, Rybacki S, Uhrmacher AM. Rule-based multi-level modeling of cell biological systems. *BMC Syst Biol* 2011;**5**:166.
113. Visser D, Schmid JW, Mauch K, *et al*. Optimal re-design of primary metabolism in Escherichia coli using linlog kinetics. *Metabolic Eng* 2004;**6**:378–90.
114. Alves R, Sorribas A. In silico pathway reconstruction: Iron-sulfur cluster biogenesis in Saccharomyces cerevisiae. *BMC Syst Biol* 2007;**1**:10.
115. Liebermeister W, Uhlendorf J, Klipp E. Modular rate laws for enzymatic reactions: thermodynamics, elasticities and implementation. *Bioinformatics* 2010;**26**:1528–34.
116. Ay A, Arnosti DN. Mathematical modeling of gene expression: a guide for the perplexed biologist. *Crit Rev Biochem Mol Biol* 2011;**46**:137–51.
117. Savageau M. Biochemical systems analysis. 3. Dynamic solutions using a power-law approximation. *J Theor Biol* 1970;**26**:215–26.
118. Savageau M. Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. *J Theor Biol* 1969;**25**:365–9.
119. Savageau M. Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation. *J Theor Biol* 1969;**25**:370–9.
120. Yagil G, Yagil E. On the relation between effector concentration and the rate of induced enzyme synthesis. *Biophys J* 1971;**11**:11–27.
121. Yagil G. Quantitative aspects of protein induction. *Curr Topic Cell Reg* 1975;**9**:183–236.
122. Teusink B, Passarge J, Reijenga CA, *et al*. Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur J Biochem/FEBS* 2000;**267**:5313–29.
123. Wright B, Kelly P. Kinetic models of metabolism in intact cells, tissues, and organisms. *Curr Topic Cell Reg* 1981;**19**:103–58.
124. Rizzi M, Baltes M, Theobald U, *et al*. In vivo analysis of metabolic dynamics in Saccharomyces cerevisiae: II. Mathematical model. *Biotechnol Bioeng* 1997;**55**:592–608.
125. Vaseghi S, Baumeister A, Rizzi M, *et al*. In vivo dynamics of the pentose phosphate pathway. *Metabol Eng* 1999;**140.**
126. Visser D, Heijnen JJ. Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics. *Metabol Eng* 2003;**5**:164–76.
127. Anbumathi P, Bhartiya S, Venkatesh KV. Mathematical modeling of fission yeast Schizosaccharomyces pombe cell cycle: exploring the role of multiple phosphatases. *Syst Synth Biol* 2011;**5**:115–29.
128. Caldara M, Dupont G, Leroy F, *et al*. Arginine biosynthesis in Escherichia coli: experimental perturbation and mathematical modeling. *J Biol Chem* 2008;**283**:6347–58.
129. Rabouille S, Staal M, Stal LJ, *et al*. Modeling the dynamic regulation of nitrogen fixation in the cyanobacterium Trichodesmium sp. *Appl Environ Microbiol* 2006;**72**:3217–27.
130. Dräger A, Kronfeld M, Ziller MJ, *et al*. Modeling metabolic networks in C. glutamicum: a comparison of rate laws in combination with various parameter optimization strategies. *BMC Syst Biol* 2009;**3**:5.
131. Singh VK, Ghosh I. Kinetic modeling of tricarboxylic acid cycle and glyoxylate bypass in Mycobacterium tuberculosis, and its application to assessment of drug targets. *Theor Biol Med Model* 2006;**3**:27.
132. Jamshidi N, Palsson BØ. Mass action stoichiometric simulation models: incorporating kinetics and regulation into stoichiometric models. *Biophys J* 2010;**98**:175–85.
133. Smallbone K, Simeonidis E, Swainston N, *et al*. Towards a genome-scale kinetic model of cellular metabolism. *Syst Biol* 2010;**4.**
134. Beisel CL, Storz G. Base pairing small RNAs and their roles in global regulatory networks. *FEMS Microbiol Rev* 2010;**34**:866–82.
135. DiChiara JM, Contreras-Martinez LM, Livny J, *et al*. Multiple small RNAs identified in Mycobacterium bovis BCG are also expressed in Mycobacterium tuberculosis and Mycobacterium smegmatis. *Nucleic Acids Res* 2010;**38**:4067–78.
136. Mraheil MA, Billion A, Kuenne C, *et al*. Comparative genome-wide analysis of small RNAs of major Gram-positive pathogens: from identification to application. *Microbial Biotechnol* 2010;**3**:658–76.
137. Storz G, Vogel J, Wassarman KM. Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell* 2011;**43**:880–91.
138. Bradley ES, Bodi K, Ismail AM, *et al*. A genome-wide approach to discovery of small RNAs involved in regulation of virulence in Vibrio cholerae. *PLoS Pathogen* 2011;**7**:e1002126.
139. Covert MW, Palsson BØ. Transcriptional regulation in constraints-based metabolic models of Escherichia coli. *J Biol Chem* 2002;**277**:28058–64.
140. Shlomi T, Eisenberg Y, Sharan R, *et al*. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol Syst Biol* 2007;**3**:101.
141. Covert MW, Xiao N, Chen TJ, *et al*. Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. *Bioinformatics* 2008;**24**:2044–50.
142. Vilaça P, Rocha I, Rocha M. A computational tool for the simulation and optimization of microbial strains accounting integrated metabolic/regulatory information. *Bio Syst* 2011;**103**:435–41.
143. Kim J, Varner J, Ramkrishna D. A hybrid model of anaerobic E. coli GJT001: combination of elementary flux modes and cybernetic variables. *Biotechnol Prog* 2008;**24**:993–1006.
144. Tenazinha N, Vinga S. A survey on methods for modeling and analyzing integrated biological networks. *IEEE/ACM Transact Comput Biol Bioinform* 2011;**8**:943–58.
145. Keller EF. Revisiting "scale-free" networks. *BioEssays: news and reviews in molecular, cellular and developmental biology* 2005;**27**:1060–8.

Appendix B

Published article: Predicting the Proteins of *Angomonas deanei*, *Strigomonas culicis* and Their Respective Endosymbionts Reveals New Aspects of the Trypanosomatidae Family

OPEN ACCESS Freely available online

PLOS ONE

Predicting the Proteins of *Angomonas deanei*, *Strigomonas culicis* and Their Respective Endosymbionts Reveals New Aspects of the Trypanosomatidae Family

Maria Cristina Machado Motta[1], Allan Cezar de Azevedo Martins[1], Silvana Sant'Anna de Souza[1,2], Carolina Moura Costa Catta-Preta[1], Rosane Silva[2], Cecilia Coimbra Klein[3,4,5], Luiz Gonzaga Paula de Almeida[3], Oberdan de Lima Cunha[3], Luciane Prioli Ciapina[3], Marcelo Brocchi[6], Ana Cristina Colabardini[7], Bruna de Araujo Lima[6], Carlos Renato Machado[9], Célia Maria de Almeida Soares[10], Christian Macagnan Probst[11,12], Claudia Beatriz Afonso de Menezes[13], Claudia Elizabeth Thompson[3], Daniella Castanheira Bartholomeu[14], Daniela Fiori Gradia[11], Daniela Parada Pavoni[12], Edmundo C. Grisard[15], Fabiana Fantinatti-Garboggini[13], Fabricio Klerynton Marchini[12], Gabriela Flávia Rodrigues-Luiz[14], Glauber Wagner[15], Gustavo Henrique Goldman[7], Juliana Lopes Rangel Fietto[16], Maria Carolina Elias[17], Maria Helena S. Goldman[18], Marie-France Sagot[4,5], Maristela Pereira[10], Patrícia H. Stoco[15], Rondon Pessoa de Mendonça-Neto[9], Santuza Maria Ribeiro Teixeira[9], Talles Eduardo Ferreira Maciel[16], Tiago Antônio de Oliveira Mendes[14], Turán P. Ürményi[2], Wanderley de Souza[1], Sergio Schenkman[19]*, Ana Tereza Ribeiro de Vasconcelos[3]*

1 Laboratório de Ultraestrutura Celular Hertha Meyer, Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brazil, **2** Laboratório de Metabolismo Macromolecular Firmino Torres de Castro, Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brazil, **3** Laboratório Nacional de Computação Científica, Laboratório de Bioinformática, Petrópolis, Rio de Janeiro, Brazil, **4** BAMBOO Team, INRIA Grenoble-Rhône-Alpes, Villeurbanne, France, **5** Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558, Villeurbanne, France, **6** Departamento de Genética, Evolução e Bioagentes, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, São Paulo, Brazil, **7** Departamento de Ciências Farmacêuticas, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brazil, **8** Laboratório Nacional de Ciência e Tecnologia do Bioetanol, Campinas, São Paulo, Brazil, **9** Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **10** Laboratório de Biologia Molecular, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiânia, Goiás, Brazil, **11** Laboratório de Biologia Molecular de Tripanossomatídeos, Instituto Carlos Chagas/Fundação Oswaldo Cruz, Curitiba, Paraná, Brazil, **12** Laboratório de Genômica Funcional, Instituto Carlos Chagas/Fundação Oswaldo Cruz, Curitiba, Paraná, Brazil, **13** Centro Pluridisciplinar de Pesquisas Químicas, Biológicas e Agrícolas, Universidade Estadual de Campinas, Campinas, São Paulo, Brazil, **14** Departamento de Parasitologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **15** Laboratórios de Protozoologia e de Bioinformática, Departamento de Microbiologia, Imunologia e Parasitologia, Centro de Ciências Biológicas, Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, Brazil, **16** Departamento de Bioquímica e Biologia Molecular, Centro de Ciências Biológicas e da Saúde, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil, **17** Laboratório Especial de Ciclo Celular, Instituto Butantan, São Paulo, São Paulo, Brazil, **18** Departamento de Biologia, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brazil, **19** Departamento de Microbiologia, Imunologia e Parasitologia, Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo, São Paulo, Brazil

Abstract

Endosymbiont-bearing trypanosomatids have been considered excellent models for the study of cell evolution because the host protozoan co-evolves with an intracellular bacterium in a mutualistic relationship. Such protozoa inhabit a single invertebrate host during their entire life cycle and exhibit special characteristics that group them in a particular phylogenetic cluster of the Trypanosomatidae family, thus classified as monoxenics. In an effort to better understand such symbiotic association, we used DNA pyrosequencing and a reference-guided assembly to generate reads that predicted 16,960 and 12,162 open reading frames (ORFs) in two symbiont-bearing trypanosomatids, *Angomonas deanei* (previously named as *Crithidia deanei*) and *Strigomonas culicis* (first known as *Blastocrithidia culicis*), respectively. Identification of each ORF was based primarily on TriTrypDB using tblastn, and each ORF was confirmed by employing getorf from EMBOSS and Newbler 2.6 when necessary. The monoxenic organisms revealed conserved housekeeping functions when compared to other trypanosomatids, especially compared with *Leishmania major*. However, major differences were found in ORFs corresponding to the cytoskeleton, the kinetoplast, and the paraflagellar structure. The monoxenic organisms also contain a large number of genes for cytosolic calpain-like and surface gp63 metalloproteases and a reduced number of compartmentalized cysteine proteases in comparison to other TriTryp organisms, reflecting adaptations to the presence of the symbiont. The assembled bacterial endosymbiont sequences exhibit a high A+T content with a total of 787 and 769 ORFs for the *Angomonas deanei* and *Strigomonas culicis* endosymbionts, respectively, and indicate that these organisms hold a common ancestor related to the Alcaligenaceae family. Importantly, both symbionts contain enzymes that complement essential host cell biosynthetic pathways, such as those for amino acid, lipid and purine/pyrimidine metabolism. These findings increase our understanding of the intricate symbiotic relationship between the bacterium and the trypanosomatid host and provide clues to better understand eukaryotic cell evolution.

Citation: Motta MCM, Martins ACdA, de Souza SS, Catta-Preta CMC, Silva R, et al. (2013) Predicting the Proteins of *Angomonas deanei*, *Strigomonas culicis* and Their Respective Endosymbionts Reveals New Aspects of the Trypanosomatidae Family. PLoS ONE 8(4): e60209. doi:10.1371/journal.pone.0060209

Editor: John Parkinson, Hospital for Sick Children, Canada

Received October 16, 2012; **Accepted** February 22, 2013; **Published** April 3, 2013

Funding: This work was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). The work of CCK as part of her PhD is funded by the ERC AdG SISYPHE coordinated by MFS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The co-author Maria Carolina Elias is a PLOS ONE Editorial Board member. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: atrv@lncc.br (ATRdV); sschenkman@unifesp.br (SS)

Introduction

Protists of the Trypanosomatidae family have been intensively studied because some of them are agents of human illnesses such as Chagas' disease, African sleeping sickness, and leishmaniasis, which have a high incidence in Latin America, Sub-Saharan Africa, and parts of Asia and Europe, together affecting approximately 33 million people. Some species are also important in veterinary medicine, seriously affecting animals of economic interest such as horses and cattle. In addition, some members of the *Phytomonas* genus infect and kill plants of considerable economical interest such as coconut, oil palm, and cassava. These organisms circulate between invertebrate and vertebrate or plant hosts. In contrast, monoxenic species, which predominate in this family, inhabit a single invertebrate host during their entire life cycle [1].

Among the trypanosomatids, six species found in insects bear a single obligate intracellular bacterium in their cytoplasm [2], with *Angomonas deanei* and *Strigomonas culicis* (previously named as *Crithidia deanei* and *Blastocrithidia culicis*, respectively) representing the species better characterized by ultrastructural and biochemical approaches [3]. In this obligatory association, the endosymbiont is unable to survive and replicate once isolated from the host, whereas aposymbiotic protozoa are unable to colonize insects [4,5]. The symbiont is surrounded by two membrane units and presents a reduced peptidoglycan layer, which is essential for cell division and morphological maintenance [6]. The lack of a typical gram-negative cell wall could facilitate the intense metabolic exchange between the host cell and the symbiotic bacterium.

Biochemical studies revealed that the endosymbiont contains enzymes that complete essential metabolic pathways of the host protozoan for amino acid production and heme biosynthesis, such as the enzymes of the urea cycle that are absent in the protozoan [7,8,9,10,11]. Furthermore, the bacterium enhances the formation of polyamines, which results in high rates of cell proliferation in endosymbiont-bearing trypanosomatids compared to other species of the family [12]. Conversely, the host cell supplies phosphatidylcholine, which composes the endosymbiont envelope [5], and ATP produced through the activity of protozoan glycosomes [13].

The synchrony in cellular division is another striking feature of this symbiotic relationship. The bacterium divides in coordination with the host cell structures, especially the nucleus, with each daughter cell carrying only one symbiont [14]. The presence of the prokaryote causes ultrastructural alterations in the host trypanosomatid, which exhibits a reduced paraflagellar structure and a typical kinetoplast DNA network [15,16,17]. The endosymbiont-harboring strains exhibit a differential surface charge and carbohydrate composition than the aposymbiotic cells obtained after antibiotic treatment [18,19]. Furthermore, the presence of the symbiotic bacterium influences the protozoan interaction with the insect host, which seems to be mediated by gp63 proteases, sialomolecules, and mannose-rich glycoconjugates [20,21].

Molecular data support the grouping of all endosymbiont-containing trypanosomatids together in a single phylogenetic branch. Moreover, studies based on rRNA sequencing suggest that symbionts from different protozoan species share high identities and are most likely derived from an ancestor of a β-proteobacterium of the genus *Bordetella*, which belongs to the Alcaligenaceae family [2,22,23]. Taken together, these results suggest that a single evolutionary event gave rise to all endosymbiont-bearing trypanosomatids, recapitulating the process that led to the formation of the mitochondrion in eukaryotic cells [24].

In this work, we analyzed the predicted protein sequences of *A. deanei* and *S. culicis* and their respective symbionts. This is the first time that genome databases have been generated from endosymbiont-containing trypanosomatids, which represent an excellent biological model to study eukaryotic cell evolution and the bacterial origin of organelles. The analysis presented here also clarifies aspects of the evolutionary history of the Trypanosomatidae family and helps us to understand how these protozoa maintain a close symbiotic relationship.

Materials and Methods

Materials and methods are described in the Text S1.

Nucleotide Sequence Accession Numbers

The sequences of *Angomonas deanei*, *Strigomonas culicis*, *Candidatus* Kinetoplastibacterium crithidii and *Candidatus* Kinetoplastibacterium blastocrithidii were assigned as PRJNA169008, PRJNA170971, CP003978 and CP003733, respectively, in the DDBJ/EMBL/GenBank.

Results and Discussion

General Characteristics

A 454-based pyrosequencing generated a total of 3,624,411 reads with an average length of 365 bp for *A. deanei* and a total of 2,666,239 reads with an average length of 379 bp for *S. culicis* (Table 1). A total of 16,957 and 12,157 ORFs were obtained for *A. deanei* and *S. culicis* genomes using this strategy, while their respective endosymbionts held a total of 787 and 769 ORFs, respectively. The total number of ORFs includes non-coding protein tRNA and rRNA genes. Tables 1 and 2 present the number of known proteins, hypothetical and partial ORFs for the two trypanosomatids and their endosymbionts, respectively.

The tRNA genes representing all 20 amino acids were identified in both trypanosomatids and their respective symbionts. At least one copy of the rRNA genes (18S, 5.8S and 28S) was identified in the genomes of *A. deanei* and *S. culicis*. We found that bacterial

Table 1. Protein Reference Sequence-Guided Assembly data of *A. deanei* and *S. culicis* genomes.

Parameter	*A. deanei*	*S. culicis*
Reads	3,624,411	2,666,239
Average reads length (bp)	365	379
Steps	3	5
Genes in contigs (protein reference sequence)	12,469	9,902
Genes in exclusive contigs	4,435	2,202
Number of known protein ORFs	7,912	6,192
Number of hypothetical ORFs	8,791	5,700
Number of partial ORFs	206	217
Total number of genes (including tRNAs and rRNAs)	16,957	12,157

doi:10.1371/journal.pone.0060209.t001

endosymbiont genomes also contain at least three copies of the rRNA operon.

General Protein Cluster Analysis

A total of 16,648 clusters were identified. Of those, 2,616 (16.4%) contained proteins from all species analyzed. To provide a more comprehensive coverage of the phylogenetic distribution, we have separated the species into three groups: endosymbiont-bearing trypanosomatids (A, s = 2 species), *Leishmania* sp. (B, s = 5) and *Trypanosoma* sp. (C, s = 4), and we considered a protein cluster to be present in the group even if zero, two or one species were missing, respectively. The protein cluster distribution is shown in Figure 1.

In this way, 2,979 protein clusters (17.9%) were identified in all groups, with 130 (0.8%) identified only in groups A and B (AB group), 31 (0.2%) only in groups A and C (AC group), and 501 (3.2%) only in groups B and C (BC group). The AB group represents the proteins that are absent in the *Trypanosoma* sp. branch. These proteins are mainly related to general metabolic function (p = 46 proteins), hypothetical conserved (p = 37) or

Table 2. General characteristics of the *A. deanei* and *S. culicis* symbionts.

Parameter	*A. deanei* symbiont	*S. culicis* symbiont
Length (BP)	821,813	820,037
G+C (%)	30.96%	32.55%
Number of known protein CDSs	640	637
Number of hypothetical CDSs	94	78
Coding region (% of genome size)	88	87
Average CDSs length (bp)	987 bp	1,004 bp
rRNA	9	9
rRNA 16 s	3	3
rRNA 23 s	3	3
rRNA 5 s	3	3
tRNA	44	45
Total number of genes	787	769

doi:10.1371/journal.pone.0060209.t002

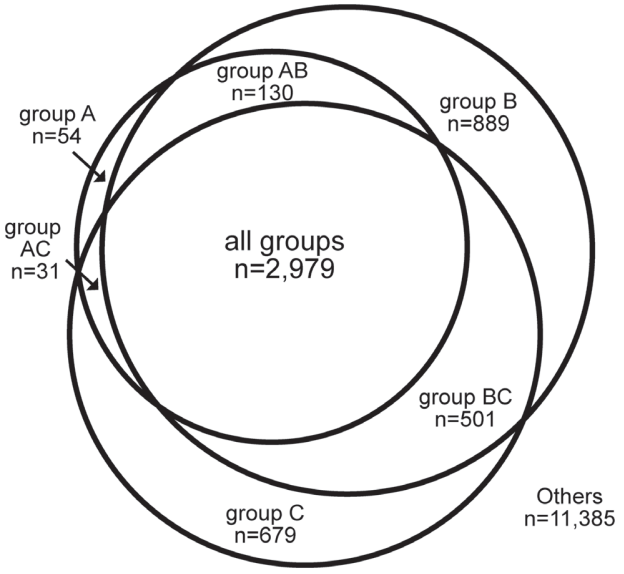


Figure 1. Venn diagram illustrating the distribution of MCL protein clusters. The diagram shows the cluster distribution comparing endosymbiont-bearing trypanosomatids (group A), *Leishmania* sp. (group B) and *Trypanosoma* sp. (group C). Protein clusters with less clear phylogenetic distributions are identified as others.
doi:10.1371/journal.pone.0060209.g001

transmembrane/surface proteins (p = 33). The AC group is four-fold smaller than the AB group, in accordance with the closer relationship between endosymbiont-bearing trypanosomatids and *Leishmania* sp [25]. The proteins in the AC group are mainly related to general metabolic function (p = 11), transmembrane/surface proteins (p = 8) and hypothetical conserved proteins (p = 7), and the relative distribution between these categories is very similar to the distribution in the AB group. The BC group is almost four-fold larger than the AB group, and mainly consists of conserved hypothetical proteins. One hypothesis to explain these different levels of conservation could be that organisms from the genera *Trypanosoma* and *Leishmania* inhabit insect and mammalian hosts, while the symbiont-bearing protozoa are mainly insect parasites. Thus, different surface proteins would be involved in host/protozoa interactions and distinct metabolic proteins are required for survival in these diverse environments.

Only a small fraction of protein clusters (n = 54, 0.3%) was identified in group A. This finding is in striking contrast to protein clusters identified only in group B (n = 889, 5.3%) or only in group C (n = 679, 4.5%), which represent specializations of the *Leishmania* or *Trypanosoma* branches. This small set is mainly composed of hypothetical proteins without similar proteins in the GenBank database. Only three of the group A clusters are similar to bacterial proteins, with two of these similar to *Bordetella* (clusters 04518 and 05756). The third one is similar to the bacterial-type glycerol dehydrogenase of *Crithidia sp.* (cluster 07344).

Of all the clusters that are present in all species except for one (n = 1,274, 7.6%), 694 (54.5%) are missing in *S. culicis*, followed by *T. congolense* (n = 211, 16.6%), *A. deanei* (n = 201, 15.8%) and *T. vivax* (n = 104, 8.0%). The fact that endosymbiont-bearing species are better represented in these sets could be due to unidentified proteins in the assembly and/or cluster analysis. This is reinforced by the fact that among clusters containing proteins from just one species (n = 9,477; 56.9%), most (73.9%) are from species with genomes that are not completely assembled (*T. vivax*, n = 1,881, 19.8%; *T. congolense*, n = 1,845, 19.5%; *A. deanei*, n = 1,745, 18.4%;

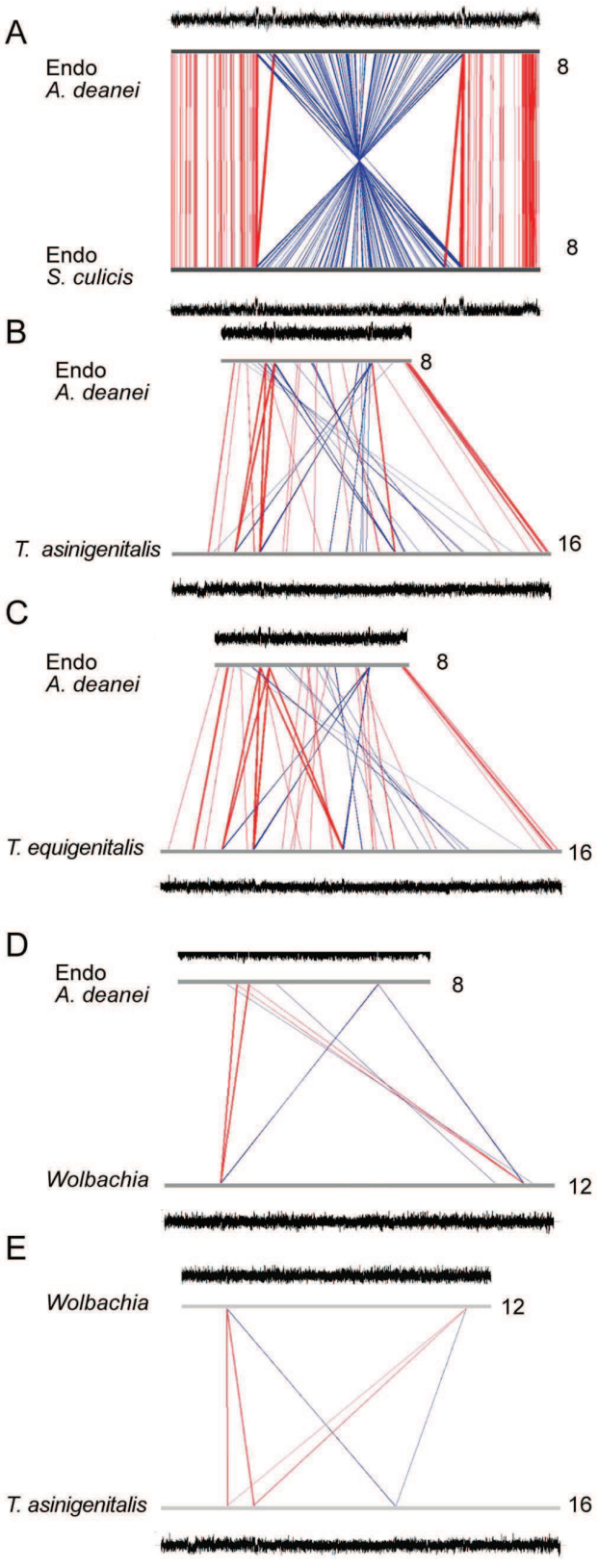


Figure 2. Genome alignments. The figure shows the alignment of the *A. deanei* endosymbiont (Endo-*A. deanei*) and the *S. culicis* endosymbiont (Endo-*S. culicis*) (A); between Endo-*A. deanei* and *T. asinigenitalis* (B), *T. equigenitalis* (C), or *Wolbachia* (D); and between *Wolbachia* and *T. asinigenitalis* (E). Alignments were performed with the ACT program based on tblastx analyses. Red (direct similarity) and blue lines (indirect similarity) connect similar regions with at least 700 bp and a score cutoff of 700. The numbers on the right indicate the size of the entire sequence for each organism.
doi:10.1371/journal.pone.0060209.g002

S. culicis, n = 1,530, 16.1%). *T. brucei* and *T. cruzi* also account for significant numbers of clusters with only a single species (n = 1,094, 11.5% and n = 1,071, 11.3%, respectively), and these clusters mainly consist of multigenic surface proteins.

Our data support the idea that endosymbiont-bearing trypanosomatids share a larger proportion of their genes with the *Leishmania* sp. in accordance with previous phylogenetic studies [2,25]. Only one fifth of all trypanosomatid protein clusters are shared among most of the species analyzed here. This proportion increases to one fourth if we only analyze the *Leishmania* and *Trypanosoma* genera; however, the number of clusters specific for endosymbiont-bearing kinetoplastids is a relatively small proportion (0.6%) of all clusters, indicating that the specialization of genes in the species following this evolutionary process was relatively small.

Genomic Characteristics of the *A. deanei* and *S. culicis* Endosymbionts

The endosymbiont genomes. Table 2 summarizes the genome analyses of both symbionts. The genome of the *A. deanei* endosymbiont contains 821,813 bp, with almost 31% G+C content and 787 CDSs. Of these, 640 (81.3%) were characterized as known CDSs, 94 (11.9%) as hypothetical, and 53 (6.7%) as rRNA or tRNA. The average CDS length is 987 bp, and coding regions account for 88% of the genome, indicating that the genome is highly compact. There are three copies of each rRNA and 44 tRNAs, suggesting a functional translation metabolism. The endosymbiont of *S. culicis* has a genome composed of 820,037 bps and 769 CDSs, 637 (83.5%) coding for known proteins, 78 (9.5%) annotated as hypothetical proteins, and 54 (6.0%) as rRNA or tRNA. The G+C content (32.6%) is similar to but slightly higher than that of the *A. deanei* endosymbiont (30.96%). *A. deanei* and *S. culicis* endosymbiont genomes are composed of 88 and 87% of CDSs with few regions formed by non-coding sequences.

A direct comparison between the two endosymbionts indicated that they share 507 genes that meet the criteria for inclusion in a cluster as described in the Materials and Methods. This represents approximately 70% of the annotated genes in both genomes, indicating a certain degree of genetic similarity. Figure 2A shows the full alignment of the *A. deanei* and *S. culicis* symbionts. This alignment indicates the occurrence of an inversion involving approximately one half of the genomes. However, this inversion would be validated by experimental work. The observed differences agree with phylogenetic analyses suggesting the classification of these symbionts as different species, *Candidatus* Kinetoplastibacterium crithidii and *Candidatus* Kinetoplastibacterium blastocrithidii [2,23].

The origins of symbionts in trypanosomatids. Previous phylogenetic studies based on sequencing of the small-subunit ribosomal DNA suggested that symbionts of trypanosomatids descended from a common ancestor, a β-proteobacteria of the *Bordetella* genus [2,22,23]. Comparisons of the endosymbiont genomes with the KEGG database revealed eight organisms that share high numbers of similar CDSs: *Bordetella petrii*, *A. xylosoxidans*, *Bordetella avium*, *Bordetella parapertussis*, *Pusillimonas*, *Bordetella bronchiseptica* and *Taylorella equigenitalis*. All these species are phylogenet-

ically related to β-proteobacteria belonging to the Alcaligenaceae family. The genus *Taylorella* consists of two species, *T. equigenitalis* and *T. asinigenitalis*, which are microaerophilic, slow-growing gram-negative bacteria belonging to the family Alcaligenaceae [26,27]. *T. equigenitalis* is an intracellular facultative pathogen in horses that causes contagious equine metritis (CEM), a sexually transmitted infection [28].

Based on these facts, clustering analysis was performed to compare these genomes and establish the genetic similarity among them. The clustering analysis compared the genomes of *A. deanei* and *S. culicis* endosymbionts, *T. equigenitalis* MCE9, *T. asinigenitalis* MCE3, *B. petrii* DSM 12804, *A. xylosoxidans* A8 and *Wolbachia pipiens* (WMel). For the *A. deanei* endosymbiont, the highest numbers of shared clusters are observed for *A. xylosoxidans* (490 clusters) and *B. petrii* (483 clusters), followed by *T. asinigenitalis* (376 clusters) and *T. equigenitalis* (375 clusters). However, considering the genome length, *T. equigenitalis* and *T. asinigenitalis* had the greater proportion of genes in clusters (24.1 and 24.67% of the annotated genes, respectively). The values for *A. xylosoxidans* and *B. petrii* are 7.59 and 9.61%, respectively. Note that the *A. xylosoxidans* plasmids pA81 and pA82 are not included in these comparisons. The *S. culicis* endosymbiont shares a high number of clusters (74%) with other genomes; considering 714 annotated genes (rRNA and tRNA genes were not taken into account), 544 (76.19%) were similar to genes of the other microorganisms. The highest number of clusters is shared between *A. xylosoxidans* (501 clusters) and *B. petrii* (495 clusters), followed by *T. asinigenitalis* (390) and *T. equigenitalis* (388 clusters). Using *W. pipiens* (wMel), an endosymbiont of *Drosophila melanogaster*, as an out-group, we found 70 clusters for *A. deanei* and 73 clusters for *S. culicis*. *Wolbachia* also shares a lower number of clusters with *T. asinigenitalis* (79) and *T. equigenitalis* (81).

T. equigenitalis MCE9 and *T. asinigenitalis* MCE3 contain 1,695,860 and 1,638,559 bps, respectively. Therefore, the *A. deanei* and *S. culicis* symbiont genomes are reduced when compared to *Taylorella,* which also have reduced genomes when compared to *Bordetella* or *Achromobacter* [26,27]. Alignments indicate the existence of similar sequences between the *Taylorella* and the kinetoplastid symbionts (Figure 2B and C), corroborating the results obtained in the clustering analyses. Much less similarity is observed between *A. deanei* and *W. pipientis* wMel, as well as between *W. pipientis* and *T. asinigenitalis* using the same alignment parameters (Figure 2D and E). Both *Taylorella* genomes are AT-rich (37.4 and 38.3% for *T. equigenitalis* and *T. asinigenitalis*, respectively), a characteristic also shared with both symbionts. Therefore, it is possible that the process of adaptation to intracellular life involved substantial base-composition modification, as most symbiotic bacteria are AT-rich [29,30].

The degree of similarity and even identity of the endosymbionts with *Taylorella* genomes and even with genomes of other species such as *Bordetella* and *Achromobacter* reinforce the origin of both endosymbionts from an ancestor of the Alcaligenaceae group. Both endosymbionts are similar to *T. equigenitalis*, *T. asinigenitalis*, *B. petrii,* and *A. xylosoxidans* and to other species of this family to different degrees. In absolute numbers, *B. petrii* and *A. xylosoxidans* have the highest numbers of clusters in common with the symbionts. However, considering the genome length, *Taylorella* species have the highest proportions of clusters in common with the *A. deanei* and *S. culicis* endosymbionts. A phylogenomic analysis using 235 orthologs was performed in order to establish the evolutionary history among *A. xylosoxidans* A8, *B. petrii* DSM 12804, *T. asinigenitalis* MCE3, *T. equigenitalis* MCE9, *Ca.* K. blastocrithidii and *Ca.* K. crithidii. The results indicated that symbionts present in both trypanosomatid species are closely related to the Alcaligenaceae family (Figure S1). *Pseudomonas aeruginosa* PA7 was the Gammaproteobacteria used as outgroup. These data corroborate the results from Alves et al. 2011 [11].

Although the genome lengths of both trypanosomatid bacteria are slightly larger than those of *Buchnera* sp. [31], they are several fold larger than those of symbiotic bacteria, which have extremely reduced genomes [32]. Analysis of the *B. pertussis* and *B. parapertussis* genomes revealed a process of gene loss during host adaptation [33,34]. This process was proposed to be associated with mobile DNA elements such as Insertion Sequences (IS) and the presence of pseudo genes [33,34]. However, the mechanism(s) involved in the length reduction observed for the genomes of the two symbionts studied here needs further investigation. Our data enable future studies examining the relationship between endosymbiosis in trypanosomatids and the origin of organelles in eukaryotic cells.

Host Trypanosomatid Characteristics

The microtubule cytoskeleton and flagellum of the host trypanosomatids. The cytoskeleton is composed of structures such as the microtubular subpelicular corset, the axoneme, the basal body, and the paraflagellar rod [35]. Thus, the cytoskeleton controls several characteristics of trypanosomatids such as their shape, the positions of structures, the flagellar beating and the host colonization. The presence of the symbiont has been related to unique characteristics of the host trypanosomatid.

Six members of the tubulin superfamily (α, β, δ, γ, ε and ζ) are present in *A. deanei* and *S. culicis*. Accordingly, δ and ε-tubulins are present in organisms that possess basal bodies and flagella [36]. γ-tubulin is localized in the basal body of *A. deanei* [14] as in other trypanosomatids [35]. Additionally, in common with other trypanosomatids, five centrins were identified in *A. deanei* and *S. culicis*. Furthermore, symbiont-containing trypanosomatids contain ε-tubulin, as in algae genomes, which can be related to the replication and inheritance of the centriole and basal bodies [37,38]. Interestingly, the absence of microtubules that form the subpelicular corset in areas where the mitochondrion touches the plasma membrane is unique to symbiont-containing trypanosomatids [15]. However, we cannot explain this atypical microtubule distribution based on database searches. Moreover, no classical eukaryotic microtubule associated proteins (MAPs) or intermediate filament homologues were identified in symbiont-bearing or other trypanosomatids, except for TOG/MOR1 and Asp.

Actin and other protein homologues that play roles in the binding and nucleation of actin filaments are present in *A. deanei* and *S. culicis*. However, the ARP 2/3 complex, which is involved in the nucleation of actin, is absent in symbiont-bearing species. As actin seems to be necessary for endocytosis in trypanosomatids [39], the absence of some proteins involved in actin nucleation may be related to the low rates of endocytosis of these protozoa (unpublished data). Indeed, both symbiont-bearing trypanosomatids have low nutritional requirements, as the symbiotic bacterium completes essential metabolic routes of the host cell [3].

Trypanosomatids are the only organisms from the orders Euglenida and Kinetoplastida that have a paraflagellar rod. This structure is continuously associated with axoneme and it contains two major proteins designated PFR1 and PFR2 [35]. Importantly, only PFR1 was identified in *A. deanei* and *S. culicis*. Perhaps we missed PFR2 since these PFR proteins are highly repetitive and their assemblies are difficult. Nevertheless, these species have a reduced paraflagellar rod located at the proximal area of the flagellum [15,16], although the same pattern of flagellar beating described for other trypanosomatids is observed for *A. deanei* [40]. The paraflagellar rod components (PFC) 4, PFC 10, PFC 16, and

PFC 18 were detected in the *A. deanei* database, whereas in *S. culicis* PFC 11 was also identified. Other minor components of the paraflagellar rod could not be detected. Accordingly, RNA interference (RNAi) knockdown of PFCs such as PFC3 does not impair the flagellar movement of *T. brucei* [41], differently from PFC4 and PFC6 depletion [42].

Several other minor flagellar proteins detected in these and other trypanosomatids are absent in *A. deanei* and *S. culicis*, especially the flagellar membrane proteins and those involved in intraflagellar transport (kinesins). Symbiont-containing species had adenylate kinase B (ADKB) but not ADKA, in contrast to other trypanosomatids, which express both. These proteins are involved in the maintenance of ATP supply to the distal portion of the flagellum [43,44].

Taken together, the differences in the composition and function of the cytoskeleton in symbiont-containing trypanosomatids seem to represent adaptations to incorporate the endosymbiont. Further exploration of these differences could enable a better understanding of how endosymbiosis was established.

The kinetoplast. The kinetoplast is an enlarged portion of the single mitochondrion that contains the mitochondrial DNA, which exhibits an unusual arrangement of catenated circles that form a network. The kinetoplast shape and the kDNA topology vary according to species and developmental stage. Endosymbiont-containing trypanosomatids show differences in the morphology and topology of the kDNA network when compared to other species of the same family. Both species present a loose kDNA arrangement, but in *A. deanei*, the kinetoplast has a trapezoid-like shape with a characteristic transversal electron-dense band, whereas in *S. culicis* the disk shape structure is wider at the center in relation to the extremities [2,17].

Differences in kDNA arrangement are related to low molecular weight basic proteins such as kinetoplast-associated protein (KAP), taking part in the organization and segregation of the kDNA network [45,46]. Our data indicate that KAP4 and KAP3 homologues are present in *A. deanei*, while KAP4, KAP2 homologues, and ScKAP-like protein are found in *S. culicis* (Table S1). In addition, a conserved nine amino acid domain in the N-terminal region, most likely a mitochondrial import signal [47,48], is found in AdKAP4 and ScKAP4 (amino acid positions 10 to 16) (Figure S2). Furthermore, ScKAP2 has a conserved domain called the High Mobility Group (HMG), indicating that this protein may be involved in protein-protein interactions. These KAPs might be related to the typical kDNA condensation of symbiont-bearing trypanosomatids.

Housekeeping genes. Histones, which are responsible for structuring the chromatin, are highly conserved proteins that appeared in the eukaryotic branch of evolution. Although well conserved, Trypanosomatidae histones display differences in the N and C-terminal sequences, sites of post-translational modifications, when compared to other eukaryotes. Phylogenetic analysis revealed that histones and their variants in both *A. deanei* and *S. culicis* are clustered in a separate branch, between the *Trypanosoma* and *Leishmania* species (Figure 3A). Similar phylogenetic distribution is seen for the dihydrofolate reductase-thymidylate synthase when we performed the analysis using nucleotide sequences (Figure 3B). Nevertheless, the symbiont-bearing species show conservation in the sites of post-translation when compared to other trypanosomes as shown in supplementary Figure S3. In *A. deanei* and *S. culicis* the proteins related to the chromatin assembly are also maintained, including histones and histone-modifying enzymes as shown in Tables S2–S7 and Figure S4 of the supporting information. For a more detailed analysis about housekeeping genes of *A. deanei* and *S. culicis* see Text S1.

DNA replication, repair, transcription, translation and signal transduction in *A. deanei* and *S. culicis* functions can be respectively attributed at least to 914 ORFs and 643 ORFs (Table 3). Most of the genes are exclusive to the protozoan and are absent in the endosymbiont (Table 4), thus indicating that these processes are exclusive to the host organism as shown in the supplementary Tables S8–S13, typically containing a conserved spliced-leader RNA as found in other trypanosomes (see Figure S5 for more information). A total of 133 and 130 proteins with similar functions are detectable in the endosymbionts of both species, with up to 95% amino acid identity to proteins of *Bordetella* sp. and *A. xylosoxidans*.

Similar DNA repair proteins are present in both eukaryote and prokaryote predicted sequences. These findings demonstrate that the endosymbionts conserved essential housekeeping proteins despite their genome reduction. Some differences were found in mismatch repair (MMR) between symbiont-bearing trypanosomatid genomes. As microsatellite instability is considered the molecular fingerprint of the MMR system, we compared the abundance of tandem repeats in the genomes of *A. deanei* and *S. culicis* and their respective endosymbionts. We noticed that the genomes of *S. culicis* and its endosymbiont are more repetitive than the genomes of *A. deanei* and its endosymbiont (Figure 4A). However, the higher repetitive content of the genomes of *S. culicis* and its endosymbiont is not only due to the higher number of microsatellite loci (Figure 4B) but also to the expansion of the size of the microsatellite sequences. These data suggest that microsatellites of *S. culicis* and its endosymbiont evolved faster than those of *A. deanei* and its endosymbiont. Interestingly, we identified some missing components of the MMR machinery in *S. culicis* that are present in *A. deanei*, such as exonuclease I (Exo I), a 5′-3′ exonuclease that is implicated in the excision step of the DNA mismatch repair pathway (Table S9). Several studies have correlated the silencing of the ExoI protein and/or mutations of the ExoI gene and microsatellite instability with development of lymphomas and colorectal cancer [49,50,51]. Therefore, we speculate that deficiencies in the MMR machinery in *S. culicis* may be related to the high proportion of microsatellites in its genome. The association between microsatellite instability and MMR deficiency has already been described for *T. cruzi* strains [52,53]. The same variability pattern is observed for each symbiont, despite the fact that the MMR machinery seems to be complete in both symbiotic bacteria (Table S10). It is tempting to speculate that this finding may indicate that the parasite and its endosymbiont are exposed to the same environment and therefore may be subjected to similar selective pressures imposed by an external oxidative condition.

A. deanei and *S. culicis* have 607 and 421 putative kinase-encoding genes, respectively (Table 5). Thirty one of the *A. deanei* kinases were classified in the AGC family, 31 as atypical, 49 as CAMK, 15 as CK1, 108 as CMGC, 64 as STE, 1 as TKL, 81 as others, and 227 that could not be classified in any of these families. No typical tyrosine kinases (TK) are present in *A. deanei* or *S. culicis*, as in other trypanosomes, although tyrosine residues are subjected to phosphorylation [54,55]. Several phosphatases have also been described in trypanosomes, pointing toward their regulatory role in the development of these organisms. The *T. brucei* PTP (*Tb*PTP1) is associated with the cytoskeleton and has been reported to be intrinsically involved in this parasite's cycle [56]. Similar sequences are found in the *A. deanei* genome, including PTP1, which is not found in the *S. culicis* database. Additionally, a large number of other PTPs appear in both genomes, including ectophosphatases (Table S14).

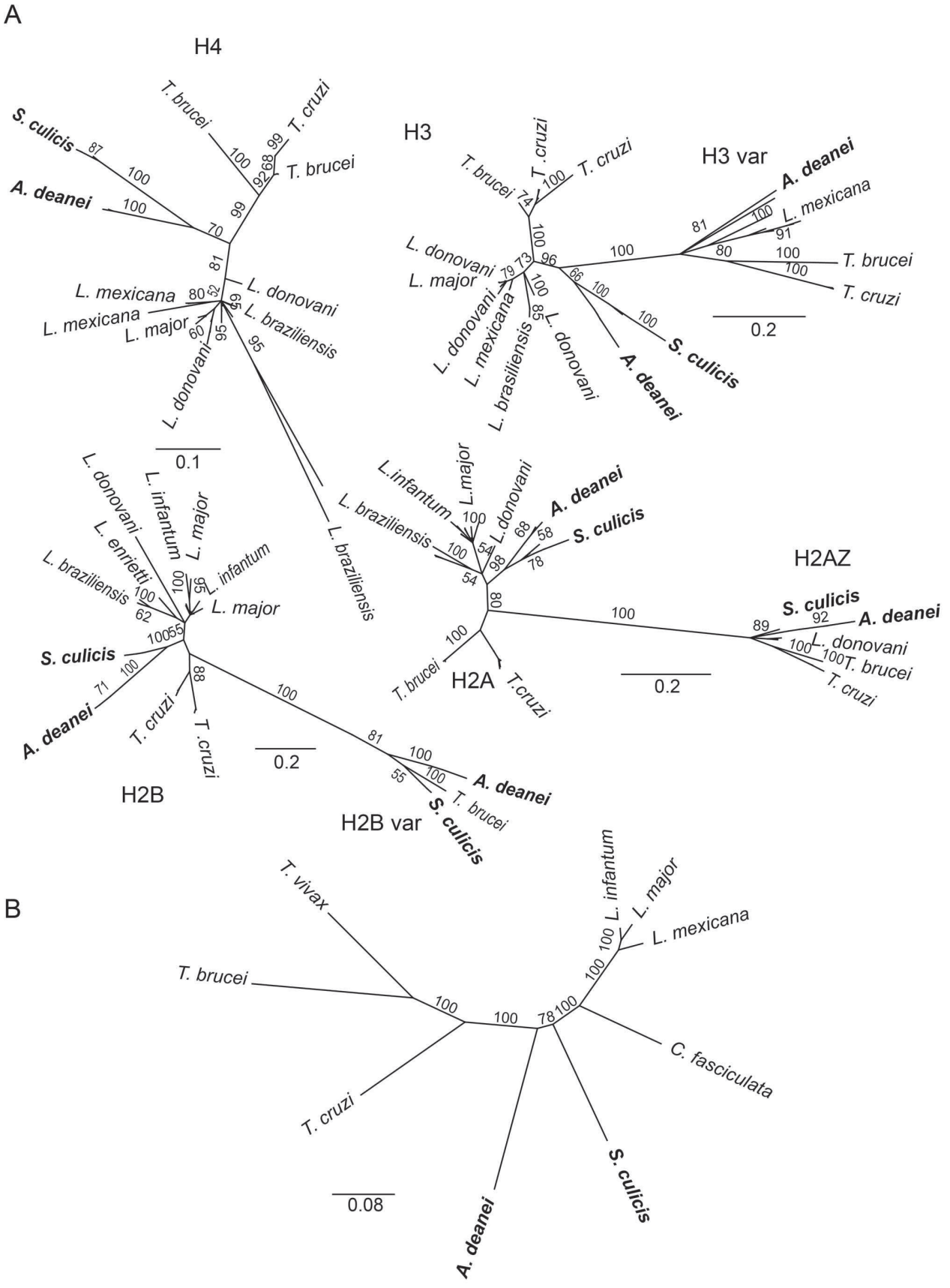

A
H4
T. brucei
T. cruzi
S. culicis
A. deanei
T. brucei
L. mexicana
L. donovani
L. mexicana
L. major
L. braziliensis
L. donovani
0.1
H3
T. brucei
T. cruzi
T. cruzi
L. donovani
L. major
L. donovani
L. mexicana
L. brasiliensis
L. donovani
A. deanei
S. culicis
H3 var
A. deanei
L. mexicana
T. brucei
T. cruzi
0.2
L. donovani
L. infantum
L. major
L. enrietti
L. braziliensis
L. infantum
L. major
S. culicis
A. deanei
T. cruzi
T. cruzi
0.2
H2B
L. braziliensis
L.infantum
L.major
L. donovani
A. deanei
S. culicis
T. brucei
T.cruzi
H2A
H2AZ
S. culicis
A. deanei
L. donovani
T. brucei
T. cruzi
0.2
A. deanei
T. brucei
S. culicis
H2B var
B
T. vivax
T. brucei
T. cruzi
L. infantum
L. major
L. mexicana
C. fasciculata
S. culicis
A. deanei
0.08

Figure 3. Phylogenetic of histones of *A. deanei, S. culicis,* and other trypanosomatids. Histone protein (panel A) and nucleotide (panel B) sequences were generated by MUSCLE tool using 10 iterations in the Geneious package [120]. Trees were constructed using the Geneious Tree Builder, by employing Jukes-Cantor genetic distance model with a neighbor-joining method and no out-groups. The consensus trees were generated from 100 bootstrap replicates of all detected histone genes, as shown below. Scale bars are indicated for each consensus tree. The trees in panel A are based in a collection of sequences of all trypanosomatids. The nucleotide sequences used for dihydrofolate reductase-thymidylate synthase are: *T. cruzi,* XM_810234; *T. brucei*, XM_841078; *T. vivax,* HE573023; *L. mexicana*, FR799559; *L. major*, XM_001680805; *L. infantum*, XM_001680805; and *C. fasciculata*, M22852.
doi:10.1371/journal.pone.0060209.g003

Two major signal transduction pathways are described in trypanosomatids: one is the cyclic AMP-dependent route and the other is the mitogen-activated protein kinase pathway [57]. The major components of these pathways, including phosphatidylinositol signaling, mTOR and MAPK signaling pathways are identified in *A. deanei* and *S. culicis*. These pathways may regulate cellular activities such as gene expression, mitosis, differentiation, and cell survival/apoptosis (Table 6).

Most genes encoding heat shock proteins are present in symbiont-bearing species, as was previously described in other trypanosomatids (Table S15). Genes for redox molecules and antioxidant enzymes, which are part of the oxidative stress response, are also present in the *A. deanei* and *S. culicis* genomes. Both contain slightly more copies of ascorbate peroxidase, methionine sulfoxide reductase, glucose-6-phosphate dehydrogenase, and trypanothione reductase genes than *L. major*. In particular, several genes related to the oxidative stress response are present in higher copy numbers in symbiont-bearing trypanosomatids than in *L. major* (Figure 5).

A. deanei sequences codify enzymes involved in RNAi, a mechanism described in various organisms that promotes the specific degradation of mRNA. RNAi is initiated by the recognition of double-stranded RNA through the action of endoribonucleases known as Dicer and Slicer, members of the Argonaut (Ago) protein family (RNase H-type) [58]. The cleavage of double-stranded RNA results in a complex that specifically cleaves mRNA molecules that are homologous to the double-stranded sequence. *A. deanei* contains the gene coding Dicer-like protein II (AGDE14022) and Ago1 (AGDE11548), homologous to enzymes in *T. brucei* and *Leishmania braziliensis* (Ngo *et al.*, 1998; Lye *et al.*, 2010). In addition, *A. deanei* contains the RNA interference factor (RIF) 4 (AGDE09645) with an exonuclease domain of the DnaQ superfamily, as described in *T. brucei*. A fragmented RIF5 sequence was also found in the sequence AGDE15656. These proteins were shown to interact with Ago1 as was recently demonstrated in *T. brucei* [59], suggesting that RNAi might be active in *A. deanei*. None of these sequences were found in the *S. culicis* database.

Table 3. Numbers of ORFs identified in *A. deanei* and *S. culicis* and their symbionts, according to the mechanisms of DNA replication and repair, signal transduction, transcription and translation.

	Number of ORFs			
Mechanism	***A. deanei***	***S. culicis***	***A. deanei* symbiont**	***S. culicis* symbiont**
Replication and Repair	178	148	56	54
Base excision repair	34	34	9	9
DNA replication	54	32	11	11
Homologous recombination	11	11	16	15
Mismatch repair	28	29	12	12
Non-homologous end-joining	8	7	–	–
Nucleotide excision repair	43	35	8	7
Signal Transduction	136	46	1	1
Phosphatidylinositol signaling system	23	17	–	–
mTOR signaling pathway	113	29	–	–
Two component system	–	–	1	1
Transcription	96	61	3	3
Basal transcription factors	15	4	–	–
RNA polymerase	28	16	3	3
Spliceosome	53	41	–	–
Translation	504	388	73	72
Aminoacyl-tRNA biosynthesis	63	56	25	25
mRNA surveillance pathway	43	45	–	–
Ribosome proteins	231	152	48	47
Ribosome biogenesis in eukaryotes	84	66	–	–
RNA transport	83	69	–	–
TOTAL	914	643	133	130

doi:10.1371/journal.pone.0060209.t003

Table 4. Summary of the origin of ORFs found in *A. deanei* and *S. culicis*.

	A. deanei		Symbiont
Functional Classification	**Prokaryotes***	**Eukaryotes****	**P/E*****
Replication and Repair			
Base excision repair	5	11	4/0
Nucleotide excision repair	2	16	9/0
Non-homologous end-joining	1	5	N
Mismatch repair	2	13	8/0
Homologous recombination	2	9	10/0
DNA replication	3	22	10/0
Signal Transduction			
Two-component system	N	N	1
Phosphatidylinositol signaling system	0	16	N
mTOR signaling pathway	0	8	N
MAPK signaling pahway - yeast	0	1	N
Transcription			
Spliceosome	0	20	N
RNA polymerase	0	16	3/0
Basal transcription factors	0	5	N
Translation			
RNA transport	0	31	N
Ribosome biogenesis in eukaryotes	0	27	N
Ribosome	0	75	48/0
mRNA surveillance pathway	0	17	N
Aminoacyl-tRNA biosynthesis	0	22	23
	S. culicis		Symbiont
Functional Classification	**Prokaryotes**	**Eukaryotes**	**P/E**
Replication and Repair			
Base excision repair	2	6	5/0
Nucleotide excision repair	2	10	7/0
Non-homologous end-joining	1	1	N
Mismatch repair	1	5	8/0
Homologous recombination	1	4	11/0
DNA replication	2	15	9/0
Signal Transduction			
Two-component system	N	N	1
Phosphatidylinositol signaling system	0	11	N
mTOR signaling pathway	0	8	N
MAPK signaling pathway - yeast	0	0	N
Transcription			
Spliceosome	0	13	
RNA polymerase	0	11	3/0
Basal transcription factors	0	2	
Translation			
RNA transport	0	19	N
Ribosome biogenesis in eukaryotes	0	20	N
Ribosome	0	53	46/0
mRNA surveillance pathway	0	16	N
Aminoacyl-tRNA biosynthesis	0	18	23

*Number of genes with identity to Prokaryotes.
**Number of genes with identity to Eukaryotes.
***Ratio of the number of genes with identity to Prokaryotes/Eukaryotes.
doi:10.1371/journal.pone.0060209.t004

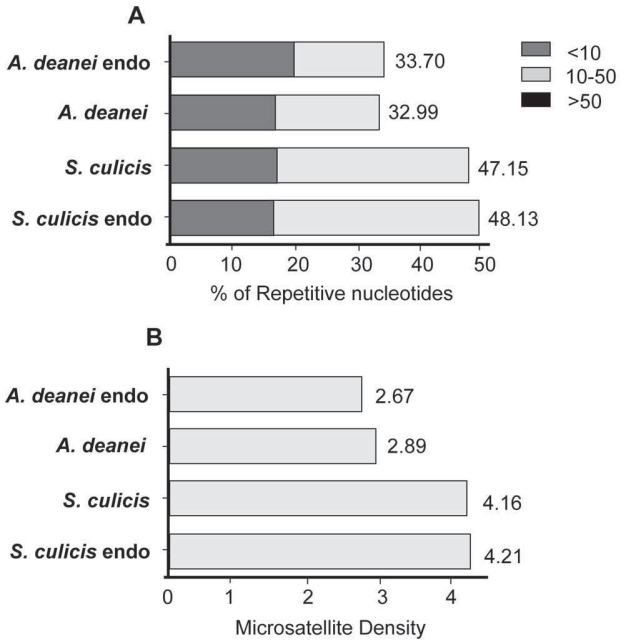


Figure 4. Microsatellite content in the genomes of *A. deanei*, *S. culicis*, and their endosymbionts. Panel (A) shows the percentage of repetitive nucleotides for each repeat length. The total numbers of nucleotides are derived from microsatellite sequences divided by the total number of assembled nucleotides. Panel (B) shows the microsatellite density. The values indicate the number of microsatellite *loci* divided by the genome length×100.
doi:10.1371/journal.pone.0060209.g004

The Coordinated Division of the Bacterium during the Host Protozoan Cell Cycle

Cell cycle control in host trypanosomes. In eukaryotes, DNA replication is coordinated with cell division by a cyclin-CDK complex that triggers DNA duplication during the S phase of the cell cycle. Multiple copies of the CRK gene (cdc2-related protein kinase) are found in *A. deanei* and four genes coding for two different CRKs are present in *S. culicis*. Both proteins exhibit structural features of the kinase subunits that make up the CDK complex, as they contain the cyclin-binding PSTAIRE motif, an ATP-binding domain and a catalytic domain. These motifs and domains are not the same in different CRKs (Figure S6), strongly suggesting that these CRKs might control different stages of the cell cycle. *A. deanei* contains four genes coding for cyclins. Three of these genes are homologues to mitotic cyclin from *S. cerevisiae* and *T. brucei*. However, none of them contain the typical destruction domain present in *T. brucei* mitotic cyclin [60]. The fourth codes for a *S. cerevisiae* Clb5 homolog, an S-phase cyclin. These data indicate that more than one CRK and more than one cyclin would be involved in the cell cycle control of symbiont-containing trypanosomatids, suggesting that tight regulation must occur to guarantee the precise maintenance of only one symbiont per cell [14].

Table 5. Kinase families identified in trypanosomatids.

Kinase family	*A. deanei*	*S. culicis*
AGC	31	23
Atypical	31	21
CAMK	49	39
CK1	15	8
CMGC	108	77
STE	64	31
TKL	1	0
Other	81	58
No hits found	227	164
TOTAL	607	421

doi:10.1371/journal.pone.0060209.t005

Table 6. Representative ORFs involved in the signal transduction pathways in *A. deanei* and *S. culicis*.

Product	*A. deanei*	*S. culicis*
Calmodulin	AGDE02036	STCU01612
Diacylglycerol kinase	AGDE02361	STCU00226
CDP-diacylglycerol-inositol-3-phosphatidyltransferase	AGDE04835	STCU01286
Myo-inositol-1(or 4) monophosphatase	AGDE08470	STCU02993
Phospholipase C	AGDE12052	STCU02439
Phosphatidylinositol 4-phosphate 5-kinase alpha	AGDE09669	STCU03909
Inositol-1,4,5-trisphosphate (IP3) 5-phosphatase	AGDE06690	nd
phosphatidate cytidylyltransferase	AGDE09922	nd
Mitogen-activated protein kinase 5	AGDE00259	STCU00603
Protein kinase A	AGDE06073	STCU01525
TP53 regulating kinase	AGDE08400	nd
Serine/threonine-protein kinase CTR1	AGDE00613	nd
Casein kinase	AGDE11868	STCU01611
Phosphoinositide-specific phospholipase C	nd	STCU09903

nd: not determined.
doi:10.1371/journal.pone.0060209.t006

Cell cycle control in the endosymbionts. Bacterial cell division is a highly regulated event that mainly depends on two structures, the peptidoglycan layer and the Z ring. The first step in the segregation of the bacterium is the formation of a polymerized Z ring at the middle of the cell. This structure acts as a platform for the recruitment of other essential proteins named Filament Temperature Sensitive (Fts), which are mainly involved in the formation and stabilization of the Z ring [61,62] and in establishing the peptidoglycan septum formation site in most bacteria [63] (Figure 6A).

Two *fts* sequences were identified in *A. deanei* and *S. culicis* symbionts based on *Bordetella* genes (Table 7). One of them is FtsZ, which requires integral membrane proteins such as Zip A and FtsA for anchoring. However, these sequences are absent in the symbionts. FtsZ should also interact with FtsE, which is absent in both symbionts. This protein is homologous to the ATP-binding cassette of ABC transporters and co-localizes with the division septum [64]. The lack of these proteins could be related to the absence of a classical Z ring in these symbionts. The other sequence is FtsK that docks FtsQ, FtsB and FtsL, which are related to the formation of the peptidoglycan layer in *E. coli* and *B. subtilis* [65,66,67], but these proteins are absent in symbionts, as in most bacteria that exhibit reduced peptidoglycan production [64]. RodA, a homologous integral membrane protein involved in bacterial cell growth, is detected in the endosymbionts. RodA could replace FtsW, which is absent in both symbionts. FtsW is

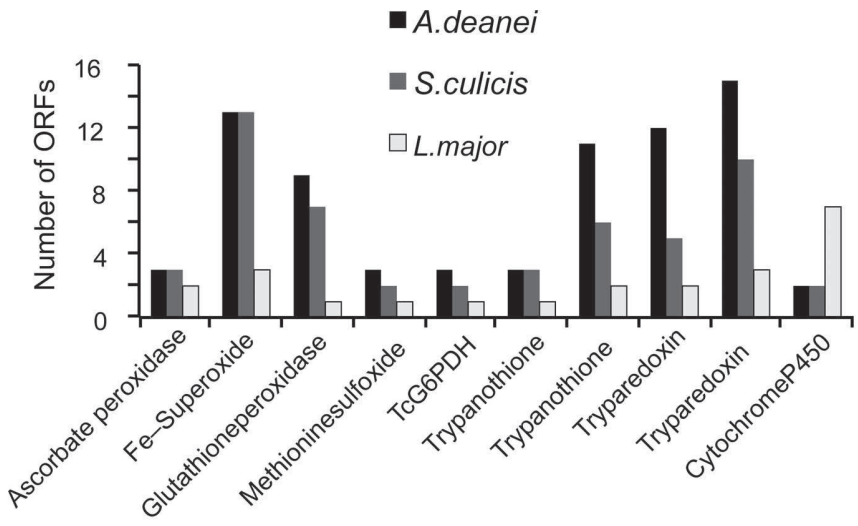


Figure 5. Oxidative stress-related genes in the genomes of *A. deanei*, *S. culicis* and *L. major*. The figure shows the number of ORFs for the indicated enzymes for each species.
doi:10.1371/journal.pone.0060209.g005

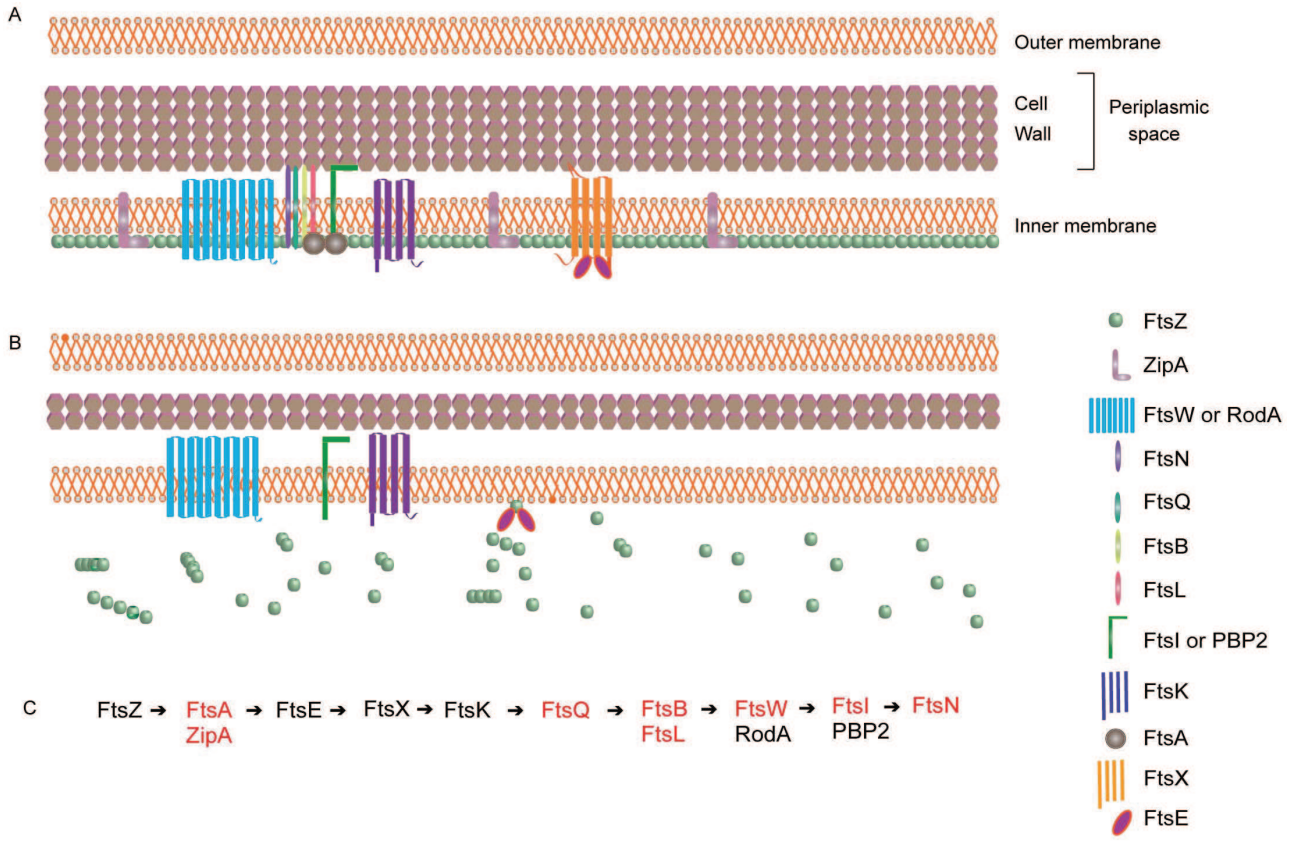


Figure 6. Schematic representation of the cell division machinery found in the endosymbionts. Panel (A) indicates the basic model derived from a gram-negative bacterium with the localization of each component (shown on the right). Panel (B) represents the components found in the endosymbiont of *A. deanei*, and Panel (C) shows the steps in the assembly of the Z-ring. The missing components of the *A. deanei* endosymbiont are drawn in red.
doi:10.1371/journal.pone.0060209.g006

Table 7. Members of the Fts family and PBPs that are present in endosymbionts of *A. deanei* and *S. culicis*.

Function	Protein	*A. deanei*	*S. culicis*
Stabilization and attachment of FtsZ polymers to the inner membrane	FtsA	nd	nd
	FtsE	nd	nd
	ZipA	nd	nd
	FtsK	CKCE00084	CKBE00632
Interaction with peptidogycan synthases PBPs	FtsQ	nd	nd
	FtsB	Nd	nd
	FtsL	nd	nd
	FtsN	nd	nd
Lipid II flippase	FtsW(RodA)	CKCE 00486	CKBE00079
Forms a dynamic cytoplasmic ring structure at midcell	FtsZ	CKCE00034	CKBE00683
Penicillin binding proteins (PBPs)	PBP1A	CKCE00524	CKBE00119
	PBP2	CKCE00487	CKBE00080
	FtsI/PBP3	CKCE00487	CKBE00080
	PBP4	nd	nd
	PBP5/dacC	CKCE00510	CKBE00105
	PBP6	nd	nd
	PBP6B	nd	nd
	PBP7	nd	nd

nd: not determined.

doi:10.1371/journal.pone.0060209.t007

essential for the localization of FtsI (PBP3) in the Z ring [68], which is absent in the symbiotic bacteria.

Endosymbionts have only one bifunctional synthase (PBP1A), while *E. coli* has PBP1A, PBP1B, and PBP1C. Cells require at least one of these synthases for viability. The peptidoglycan layer is functional in trypanosomatid symbionts, as shown by treatment with β-lactam antibiotics affecting the division of the bacterium, generating filamentous structures and culminating in cell lysis. PBP1 and PBP2 have also been detected at the symbiont envelope [6]. PBP1B interacts with the two essential division proteins, FtsN and PBP3/FtsI, which are absent in the symbiont. PBP1B can also interact with PBP2 that is identified in both symbiont databases (see Table 7).

A sequence encoding a minor PBP described in *E. coli* was also identified in the symbionts. This protein is known as a putative PBP precursor (PBP5/dacC). This PBP is involved in the regulation of the peptidoglycan structure, along with 3 other minor PBPs described in *E. coli*, but these are absent from the symbiont (Table 7). On the other hand, all the enzymes involved in the synthesis of activated nucleotide precursors for the assembly of the peptidoglycan layer are present in the symbiont genome, except for Braun's lipoprotein (Lpp), which forms the lipid-anchored disaccharide-pentapeptide monomer subunit [69]. In *E. coli* strains, mutations in Lpp genes result in a significant reduction of the permeability barrier, although small effects on the maintenance of the cell growth and metabolism were observed in these cells [70,71].

Taken together, we consider that gene loss in the *dcw* cluster [72] (represented in Figure 6) explains the lack of the FtsZ ring in the endosymbiont during its division process [73]. Moreover, the symbiont envelope contains a reduced peptidoglycan layer and lacks a septum during its division process, which can be related to the facilitation of metabolic exchanges, as well as to the control of division by the host protozoan [6]. These losses could be understood since the host trypanosomatid is controlling the number of symbiotic bacteria per cell. This phenomenon has been described for obligatory intracellular bacteria that co-evolve in eukaryotic cells, as well as for the organelles of prokaryotic origin, the chloroplast and the mitochondrion [74,75].

Metabolic Co-evolution of the Bacterium and the Host Trypanosomatid

Symbiosis in trypanosomatids is characterized as a mutual association where both partners benefit. These symbiont-bearing protozoa have low nutritional requirements, as intense metabolic exchanges occur. Our data corroborate previous biochemical and ultrastructural analyses showing that the bacterium has enzymes and metabolic precursors that complete important biosynthetic pathways of the host [76].

Oxidative phosphorylation. FoF1-ATP synthase and the entire mitochondrial electron transport chain are present in *A. deanei* and *S. culicis*, although some subunits are missing (Table 8). These species have a rotenone-insensitive NADH:ubiquinone oxidoredutase in complex I, as do other trypanosomatids [77]. Ten complex II (succinate:ubiquinone reductase) subunits of the twelve identified in *T. cruzi* [78] are also present in both trypanosomatids. Many subunits from complex III, composed of cytochrome c reductase, are found in *A. deanei* and *S. culicis*. In addition, these protozoa contain genes for cytochrome c, as previously suggested by biochemical studies in other symbiont-containing trypanosomatids [3,79].

Both symbionts contain sequences with hits for all subunits of complex I, NADH:ubiquinone oxidoredutase, similar to *E. coli* (Table 8). Complexes II and III, including cytochrome *c*, and complex IV (cytochrome *c* oxidase, succinate:ubiquinone reductase and cytochrome c reductase, respectively) are not found in

Table 8. Respiratory chain complexes identified in the predicted proteome of *A. deanei*, *S. culicis* and their respective endosymbionts.

	A. deanei	*A. deanei* **endosymbiont**	*S. culicis*	*S. culicis* **endosymbiont**
Complex I	33	0	33	0
Complex II	10	0	10	0
Complex III	5	0	4	0
Complex IV	10	2*	2	2*
Complex V	10	8	3	8

*The complex IV of the endosymbionts might be a cytochrome *d* ubiquinol oxidase identified in both organisms, instead a classical cytochrome *c* oxidase.
doi:10.1371/journal.pone.0060209.t008

either symbiont. However, we detected the presence of cytochrome *d* as found in *Allochromatium vinosum*, and also a cytochrome *d* oxidase with a sequence close to that of *B. parapertussis*. All portions of the FoF1-ATP synthase were identified in symbionts, although not every subunit of each portion was found.

Lipid metabolism. The sphingophospholipid (SPL) content in *A. deanei* and its symbiont has been previously described, with phosphatidylcholine (PC) representing the major SPL in the host, whereas cardiolipin predominates in the symbiotic bacterium [5,80]. The synthetic pathway of phosphatidylglycerol from glycerol phosphate is present in both host trypanosomatids (Table S16). The biosynthetic pathways of PC and PE from CDP-choline and CDP-ethanolamine (Kennedy pathways), that synthesize PC and PE respectively, are incomplete in *A. deanei* and *S. culicis*. Nevertheless, the methylation pathway (Greenberg pathway), which converts PE in PC, seems to be absent in both trypanosomatids, even though one enzyme sequence was identified in *A. deanei*.

The symbiont of *A. deanei* exhibits two routes for phosphatidylethanolamine (PE) synthesis, starting from CDP-diacylglycerol and producing phosphatidylserine as an intermediate (Table S17). Interestingly, this last step of the pathway is not found in the *S. culicis* endosymbiont. Importantly, both symbionts lack genes that encode proteins of PC biosynthetic pathways, reinforcing the idea that this phospholipid is mainly obtained from the host protozoa [5]. Remarkably, phoshpatidylglycerophosphatase A, which produces the intermediate phosphatidylglycerol necessary for cardiolipin biosynthesis, was not found in either protozoa but is present in both symbionts. As cardiolipin is present in the inner membranes of host mitochondria, the symbionts may complete cardiolipin biosynthesis.

Pathways for sphingolipid production, including the synthesis of ceramide from sphingosine-1P, are present in *A. deanei*, while *S. culicis* lacks enzymes of this pathway (Table S16). Both host trypanosomatids have glycerol kinase and 3-glycerophosphate acyltransferase, enzymes for the synthesis of 1,2-diacyl-sn-glycerol and triacylglycerol from D-glycerate. In endosymbionts, glycerolipid metabolism seems to be reduced to two enzymes: 3-glycerophosphate acyltransferase and 1-acylglycerol-3-phosphate O-acyltransferase (Table S17), suggesting metabolic complementation between partners.

Furthermore, both hosts contain enzymes of the biosynthesis pathway for ergosterol production from zymosterol, as well as the pathway of sterol biosynthesis that produces lanosterol from farnesyl-PP. These pathways are only complete in *A. deanei*. The symbionts do not have enzymes for sterol biosynthesis, in accordance with our previous biochemical analysis [80].

Metabolism of amino acids, vitamins, cofactors and hemin. Symbiosis in trypanosomatids is characterized by intensive metabolic exchanges, reducing the nutritional requirements of these trypanosomatids when compared to species without the symbiotic bacterium, or to aposymbiotic strains. Several biochemical studies have been carried out analyzing the biosynthetic pathways involved in this intricate relationship as recently reviewed [76], and our genomic data corroborate these findings. A schematic description of the potential metabolic interactions concerning the metabolism of amino acids, vitamins, cofactors, and hemin is provided in Figure 7.

Both symbiotic bacteria have genes potentially encoding for all necessary enzymes for lysine, phenylalanine, tryptophan and tyrosine synthesis, in agreement with previous experimental data [40]. Tyrosine is required in the growth medium of *A. deanei* [81], but it is not essential for *S. oncolpelti* or *S. culicis* [41,82,83]. Here, in the symbiotic bacteria, we found enzymes involved in tyrosine synthesis, as well as indications that phenylalanine and tyrosine can be interconverted. In fact, protozoan growth is very slow in absence of phenylalanine and tryptophan [81], which may

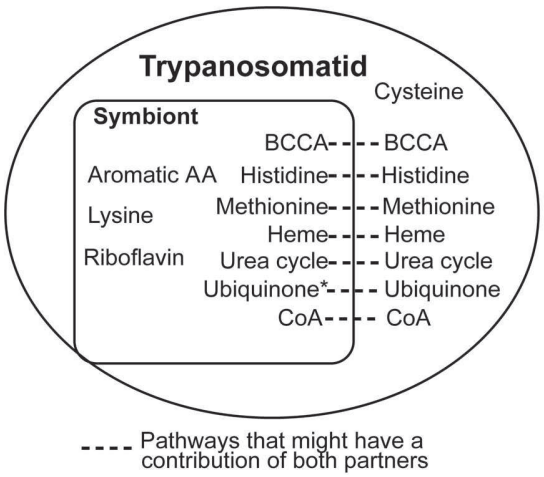


Figure 7. Main metabolic exchanges between host and endosymbionts. Schematic representation of the amino acids, vitamins, and cofactors exchanged between *A. deanei* and *S. culicis* and their respective symbionts. Dotted lines indicate pathways that have or might have contributions from both partners, whereas metabolites inside one of the circles, representing the symbiont or host, indicate that one partner holds candidate genes coding for enzymes of the whole biosynthetic pathway. *Candidate genes were only found for the symbiont of *S. culicis* and not for the symbiont of *A. deanei*. BCAA (branched-chain amino acids) are leucine, isoleucine and valine.
doi:10.1371/journal.pone.0060209.g007

indicate that larger amounts of these amino acids are required for rapid cell proliferation.

Our data indicate that branched-chain amino acid (BCAA) synthesis mainly occurs in the symbionts except for the last step, with the branched-chain amino acid aminotransferase found in the host protozoan.

Among the pathways that (might) involve contributions from both partners, two have previously been characterized in detail, the urea cycle and heme synthesis. The urea cycle is complete in both symbiont-harboring trypanosomatids. Symbiotic bacteria contribute with ornithine carbamoyltransferase, which converts ornithine to citrulline, and with ornithine acetyltransferase, which transforms acetylornithine in ornithine. Conversely, aposymbiotic strains and symbiont-free *Crithidia* species need exogenous arginine or citrulline for cell proliferation [8] [68]. Our genomic data corroborate these studies.

Contrary to symbiont-free trypanosomatids, *A. deanei* and *S. culicis* do not require any source of heme for growth because the bacterium contains the required enzymes to produce heme precursors that complete the heme synthesis pathway in the host cell [7,9,10,11,84]. Our results support the idea that heme biosynthesis is mainly accomplished by the endosymbiont, with the last three steps of this pathway performed by the host trypanosomatid, and in most cases also by the bacterium as described in [11]. Furthermore, this metabolic route may represent the result of extensive gene loss and multiple lateral gene transfer events in trypanosomatids [11].

According to our genomic analyses, the symbiotic bacteria also perform the synthesis of histidine, folate, riboflavin, and coenzyme A, but one step is missing in the middle of each pathway, making them candidates for metabolic interchange with the host. In the case of folate and coenzyme A biosynthesis, one candidate gene was found in the host trypanosomatid. Moreover, none of these four metabolites are required in the growth medium of *A. deanei* and *S. culicis* [85], suggesting that these pathways are fully functional.

Candidate genes for the ubiquinone biosynthetic pathway were found in *S. culicis* but none for *A. deanei* endosymbionts. For the route with chorismate as precursor, only the first out of nine steps is missing in the *S. culicis* endosymbiont; moreover a candidate gene for that step is found in *S. culicis* genome. Only a few steps of these pathways are absent in *A. deanei* and *S. culicis* host organisms. In *L. major*, the ubiquinone ring synthesis has been described as having either acetate (via chorismate as in prokaryotes) or aromatic amino acids (as in mammalian cells) as precursors [45].

Methionine is considered essential for the growth of *A. deanei*, *S. culicis* and *S. oncopelti* [41,81,82]. We were not able to identify one enzyme among the four involved in the synthesis of methionine from either pyruvate or serine via cysteine in the genomes of *A. deanei* and *S. culicis*. No candidate to complement this pathway was found in the symbiotic bacteria.

Purine and pyrimidine metabolism for nucleotide production. Trypanosomatids are not able to synthesize the purine ring *de novo* [86,87,88]. We observed that endosymbiont-bearing trypanosomatids contain sequences encoding ectonucleotidases from the E-NTPDase family and the adenosine deaminase family (Table S18), which are required for the hydrolysis and deamination of extracellular nucleotides [89,90]. Interestingly, sequences encoding 5′-nucleotidases are not found in either symbiont-bearing trypanosomatid. The absence of this enzyme can be related to the presence of the endosymbiont, which can supply adenosine to the host cell, as we found all genes involved in the *de novo* pathway in the symbionts, indicating that they are able to complement the purine requirements of the host (Figure 8). However, we cannot discard the possibility that adenosine is transported to the intracellular medium by carriers of monophosphate nucleoside or by the presence of other enzymes that have the same function as 5′-nucleotidase. On the other hand, the lack of 5′-nucleotidase in *A. deanei* and *S. culicis* can be related to the fact that such protozoa are only insect parasites. According to this idea, several studies have shown the importance of ectonucleotidases in the establishment of infection by some trypanosomatid species [91]. The high activity of ectonucleotidases with concomitant production of adenosine, a known immune system inhibitor, lead to high susceptibility to *Leishmania* infection because adenosine can induce anti-inflammatory effects on the host [92,93].

Nucleoside transporters can take up nucleosides and nucleobases generated by ectonucleotidase activity. Genes encoding nucleoside transporters are present in both trypanosomatid genomes (Table S19), enabling cells to obtain exogenous purines from the medium. Furthermore, *A. deanei* and *S. culicis* contain intracellular enzymes that can convert purines to nucleotides, such as adenine phosphoribosyltransferase, hypoxanthine-guanine phosphoribosyltransferase, adenylate kinase, AMP deaminase, inosine monophosphate dehydrogenase and GMP synthetase. These data indicate that these organisms can interconvert intracellular purines into nucleotides. In contrast, both endosymbionts lack all the genes encoding enzymes related to purine salvage. Nevertheless, the symbiotic bacteria have genes encoding all the enzymes expected to participate in the *de novo* synthesis of purine nucleotides as previously proposed [94,95]. One interesting possibility is that the symbiotic bacterium is able to supply the host trypanosomatid with purines. According to this idea, the endosymbiont participates in the *de novo* purine nucleotide pathway of *A. deanei*, as the aposymbiotic strain is unable to utilize glycine for the synthesis of purine nucleotides, only for pyrimidine nucleotide production [87].

Protozoa are generally, but not universally considered to be capable of synthesizing pyrimidines from glutamine and aspartic acid, which are used as precursors. Our results indicate that both symbiont-bearing trypanosomatids carry out *de novo* pyrimidine synthesis (Table S19). Interestingly, *in silico* analyses also revealed the presence of all the genes for *de novo* pyrimidine synthesis in both symbiont genomes, but not for the pyrimidine salvage pathway. A previous report indicated that *A. deanei* was able to synthesize purine and pyrimidine nucleotides from glycine (“de novo” pathway) and purine nucleotides from adenine and guanine (“salvage” pathway). Adenine would be incorporated into both adenine and guanine nucleotides, whereas guanine was only incorporated into guanine nucleotides, suggesting a metabolic block at the level of GMP reductase [87].

Deoxyribonucleotides are derived from the corresponding ribonucleotides by reactions in which the 2′-carbon atom of the D-ribose portion of the ribonucleotide is directly reduced to form the 2′-deoxy derivative. This reaction requires a pair of hydrogen atoms that are donated by NADPH via the intermediate-carrying protein thioredoxin. The disulfide thioredoxin is reduced by NADPH in a reaction catalyzed by thioredoxin reductase, providing the reducing equivalents for the ribonucleotide reductase, as observed for the endosymbionts that could provide 2′-deoxy derivatives. In folate metabolism, the formation of thymine nucleotides requires methylation of dUMP to produce dTMP, a reaction catalyzed by thymidilate kinase, which is present in *A. deanei*, *S. culicis*, and their respective endosymbionts. Figure 8 summarizes the purine and pyrimidine metabolisms in *A. deanei* and *S. culicis* considering the metabolic complementarity between the protozoan and the endosymbiont.

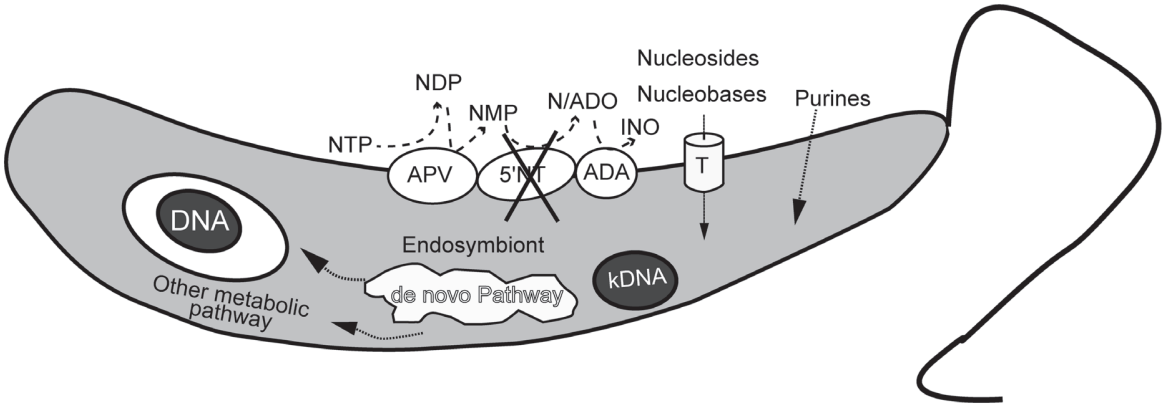


Figure 8. Purine production, acquisition, and utilization in *A. deanei* and *S. culicis*. The figure illustrates the production, acquisition and utilization of purines in the host trypanosomes considering the presence of endosymbiont enzymes. This model suggests that the trypanosomatid acquires purines from the symbiont, which synthesizes them *de novo*. Some ecto-localized proteins, such as apyrase (APY) and adenosine deaminase (ADA), could be responsible for the generation of extracellular nucleosides, nucleobases, and purines. Nucleobases and purines could be acquired by the parasite through membrane transporters (T) or diffusion and could be incorporated into DNA, RNA, and kDNA molecules after "purine salvage pathway" processing. Abbreviations: NTP (nucleoside tri-phosphate), NDP (nucleoside di-phosphate), NMP (nucleoside mono- phosphate), N (nucleobase), ADO (adenosine), INO (inosine).
doi:10.1371/journal.pone.0060209.g008

In this way, both symbiont-containing protozoa express a unique complement of nutritionally indispensable salvage and interconversion enzymes that enable the acquisition of purines from the medium. The intracellular purines can be acquired through the medium by the action of ectonucleotidases and nucleoside transporters.

Factors Involved in Protozoa-host Interactions

Monoxenic trypanosomatids only parasitize invertebrates, especially insects belonging to the orders Diptera and Hemiptera [1]. These organisms have been found in Malphigian tubules, in the hemolymph and hemocoel, and in the midgut, which is considered the preferential site for protozoal multiplication and colonization [1,96,97]. *S. culicis*, for example, is able to colonize the insect midgut, to invade the hemocoel and to reach the salivary glands [97,98]. The presence of the symbiotic bacterium has been shown to influence the interactions between trypanosomatid cells and insect cell lines, explanted guts and host insects [4,20]. This seems to occur because the endosymbiont influences the glycoprotein and polysaccharide composition of the host, the exposure of carbohydrates on the protozoan plasma membrane, and the surface charge [18,19,20,21].

Several glycosyltransferases from the two major families (GT-A and GT-B [99]) and members of the family 25 (glycosyltransferases involved in lipo-oligosaccharide protein biosynthesis) are present in both *A. deanei* and *S. culicis* genomes (Table S20). Other glycosyltransferases with no characteristic domains that are thus not classified as belonging to the GT-A or GT-B families are also found in the *A. deanei* and *S. culicis* genomes. Importantly, 1,2-fucosyltransferase transferase is present in *A. deanei* but not in the *S. culicis* dataset, and fucose residues were found in high amounts on glycoinositolphospholipid (GIPL) molecules of *A. deanei*, different from the observations for other trypanosomatids (data not published). Although the role of fucose is unknown, fucose and arabinose transfer to lipophosphoglycan (LPG) of *Leishmania* is noticed when the culture medium is supplemented with this carbohydrate [100], suggesting that fucose might have a specific role in *A. deanei*-insect interactions.

Another glycosyltransferase found in both *A. deanei* and *S. culicis* genomes and involved in the N-glycosylation of asparagine residues is the dolichyl-diphosphooligosaccharide-protein glycosyltransferase (DDOST), an oligosaccharyltransferase (OST) that is not classified in any of the above-mentioned families. The *A. deanei* and *S. culicis* DDOSTs contain the STT3 domain, a subunit required to establish the activity of the oligosaccharyl transferase (OTase) complex of proteins, and they are orthologous to the human DDOST. These OTase complexes are responsible for transferring lipid-linked oligosaccharides to the asparagine side chain of the acceptor polypeptides in the endoplasmic reticulum [101], suggesting a conserved N-glycosylation among the trypanosomatids.

Five different GalfT sequences are also present in the endosymbiont-bearing trypanosomatids, and all of them contain the proposed catalytic site, indicating genetic redundancy. Redundancy of GalfTs is commonly observed in many different trypanosomatid species, as different transferases are used for each linkage type [102]. As β-galactofuranose (β-Gal*f*) has been shown to participate in trypanosome-host interactions [103], their presence in *A. deanei* and *S. culicis* might also indicate a role in the interaction with the insect host. However, no enzymes involved in synthesis of β-Galf-containing glycoconjugates are detected in our *A. deanei* dataset, despite reports of enzymes involved in β-Gal*f* synthesis in *Crithidia* spp. [104,105,106].

Surface proteins and protease gene families. One remarkable characteristic of trypanosomatid genomes is the large expansion of gene families encoding surface proteins [107]. Experimental data indicated that these genes encode surface proteins involved in interactions with the hosts. We selected eight gene families encoding surface proteins present in *T. cruzi*, *T. brucei* and *Leishmania* spp. to search for homologous sequences in the genomes of the two symbiont-bearing trypanosomatids. Because the draft assemblies of these genomes are still fragmented, we also used a read-based analysis to search for sequences with homology to these multigene families. It is well known that misassemblies frequently occur for tandemly repeated genes, as most repetitive copies collapse into only one or two copies. A total of 3,624,411 reads (corresponding to 1,595 Mb of sequences) from the *A. deanei* genome and 2,666,239 reads (corresponding to 924 Mb) from the *S. culicis* genome were used in this comparison. In *A. deanei* and *S. culicis*, we identified gene families encoding amastins, gp63, and

cysteine peptidases (Table S21). As expected, we could not identify sequences homologous to mucin-like glycoproteins typical of *T. cruzi* [108], variant surface glycoprotein (VSG) characteristic of African trypanosomes, or trans-sialidases present in the genomes of all *Trypanosoma* species.

Calpain-like cysteine peptidases constitute the largest gene family identified in the *A. deanei* (85 members) and *S. culicis* (62 members) genomes, and they are also abundant in trypanosomatids [46]. The presence of the N-terminal fatty acid acylation motif was found in some members of calpain-like cysteine peptidases, indicating that some of these peptidases are associated with membranes, as has also been shown for other members of the family [109,110]. The relatively large amount of calpain-like peptidases may be related to the presence of the endosymbiont, which would require a more complex regulation of the cell cycle and intracellular organelle distribution [14], as cytosolic calpains were found to regulate cytoskeletal remodeling, signal transduction, and cell differentiation [46].

A second large gene family in the *A. deanei* and *S. culicis* genomes encoding surface proteins with proteolytic activity is gp63. In our genomic analyses, we identified 37 and 9 genes containing sequences homologous to the gp63 of *Leishmania* and *Trypanosoma* spp. in the genomes of *A. deanei* and *S. culicis*, respectively. Proteins belonging to this group of zinc metalloproteases, also known as major surface protease (MSP) or leishmanolysin, have been characterized in various species of *Leishmania* and *Trypanosoma* [111]. Extensive studies on the role of this family in *Leishmania* indicate that they are involved in several aspects of host-parasite interaction including resistance to complement-mediated lysis, cell attachment, entry, and survival in macrophages [112]. Gene deletion studies in *T. brucei* indicated that the TbMSP of bloodstream trypanosomes acts in concert with phospholipase C to remove the variant surface protein from the membrane, required for parasite differentiation into the procyclic insect form [113]. Gp63-like molecules have been observed on the cell surface of symbiont-harboring trypanosomatids [114]. Importantly, the symbiont containing *A. deanei* displays a higher amount (2-fold) of leishmanolysin-like molecules at the surface compared to the aposymbiotic strain, which are unable to colonize insects [4]. As anti-gp63 antibodies decrease protozoan-insect interactions [21], our results reinforce the idea that the presence of such interactions caused the expansion of this gene family in endosymbiont-bearing organisms.

In contrast, only two copies of lysosomal cathepsin-like cysteine peptidases were identified in the *A. deanei* (AGDE05983 and AGDE10254) and *S. culicis* genomes (STCU01417 and STCU06430). The two *A. deanei* sequences encode identical cathepsin-B-like proteins, whereas the two *S. culicis* genes encode proteases of the cathepsin-L-like group. This class of cysteine peptidase is represented by cruzain or cruzipain, major lysosomal proteinases of *T. cruzi* expressed by parasites found in insect and vertebrate hosts, and encoded by a large gene family [115,116]. In *T. cruzi,* these enzymes have important roles in various aspects of the host/parasite relationship and in intracellular digestion as a nutrient source [115]. Conversely, the low copy number of this class of lysosomal peptidase in symbiont-containing trypanosomatids seems to be related to their low nutritional requirements.

Amastins constitute a third large gene family in the *A. deanei* and *S. culicis* genomes that encodes surface proteins. Initially described in *T. cruzi* [117], amastin genes have also been identified in various *Leishmania* species [118], in *A. deanei* and in another related insect parasite, *Leptomonas seymouri* [119]. In *Leishmania*, amastins constitute the largest gene family with gene expression that is regulated during the parasite life cycle. As amastin has no sequence similarity to any other known protein, its function remains unknown. In this work, we identified 31 genes with sequences belonging to all four sub-families of amastins in the genome of *A. deanei* and 14 copies of amastin genes in *S. culicis*. Similar to *Leishmania*, members of all four amastin subfamilies were identified in symbiont-containing species (see Figure S7).

Conclusion

The putative proteome of symbiont-bearing trypanosomatids revealed that these microorganisms exhibit unique features when compared to other protozoa of the same family and that they are most closely related to *Leishmania* species. Most relevant are the differences in the genes related to cytoskeleton, paraflagellar and kinetoplast structures, along with a unique pattern of peptidase gene organization that may be related to the presence of the symbiont and of the monoxenic life style. The symbiotic bacteria of *A. deanei* and *S. culicis* are phylogenetically related with a common ancestor, most likely a β-proteobacteria of the Alcaligenaceae family. The genomic content of these symbionts is highly reduced, indicating gene loss and/or transfer to the host cell nucleus. In addition, we confirmed that both bacteria contain genes that encode enzymes that complement several metabolic routes of the host trypanosomatids, supporting the fitness of the symbiotic relationship.

Supporting Information

Figure S1 Evolutionary history of endosymbionts obtained through a phylogenomic approach. The figure indicates analysis using the Neighbor joining (NJ) **(A)** and Maximum parsimony (MP) **(B)** methods. For NJ and MP, the percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1,500 replicates) is shown next to the branches. The scale bar represents amino acids substitutions per site.
(TIF)

Figure S2 Amino acid alignment of Kinetoplast Associated Proteins. Panel (A) shows the KAP4 ClustalW alignment of *A. deanei* (AdKAP-4), *S. culicis* (ScKAP-4) and *C. fasciculata* (CfKAP-4). Panel (B) shows the ClustalW alignment of KAP2 of *S. culicis* and *C. fasciculata* (CfKAP2-2, GenBank Q9TY84 and CfKAP2-1 GeneBank Q9TY83). Black color highlight is 100% similar gray is 80 to 99% similar light gray is 60 to 79% similar white is less than 59% similar.
(TIF)

Figure S3 Comparison of the histone sequences of *A. deanei* and *S. culicis* with other trypanosomes. Residues indicated in red correspond to lysines that are acetylated and green, methylated in *T. cruzi* and *T. brucei* [121]. Residues indicated in blue are predicted site for phosphorylation upon DNA damage as shown in *T. brucei* [122].
(TIF)

Figure S4 Phylogenetic tree of sirtuins from Trypanosomatids. The numbers represent bootstrap values. The proteins from each species are grouped in nuclear and mitochondrial Sir2 based on the sequences of *S. cerevisiae* (nuclear), and the similarity with *S. coelicolor* and *S. enterica.*
(TIF)

Figure S5 Phylogenetic tree of spliced leader (SL) sequences of *A. deanei* and *S. culicis*. A neighbor-joining tree (1000 bootstraps) obtained by MEGA 5.0 using the SL gene from the *A. deanei* and *S. culicis* genome sequences and sequences

retrieved from GenBank (*S. culicis* DQ860203.1, *L. pyrrhocoris* JF950600.1, *H. samuelpessoai* X62331.1, *H. mariadeanei* AY547468.1, *A. deanei* EU099545.1, *T. rangeli* AF083351 and *T. cruzi* AY367127).
(TIF)

Figure S6 Comparison between the amino acid sequences of *S. culicis* CRK sequences. The figure shows a ClustalW alignment with the ATP binding domains boxed in yellow, PSTAIRE motifs boxed in blue, and the catalytic domain boxed in pink. Red residues indicate the observed variations in the amino acids involved in the activity.
(TIF)

Figure S7 Tree showing the distribution of amastin subfamilies in *A. deanei*. The amastins are grouped as delta-amastin (red), gamma-amastins (yellow), alpha-amastins (dark blue) and beta-amastins (light blue).
(TIF)

Table S1 ORFs identified as Kinetoplast-associated protein (KAPs) in *A. deanei* and *S. culicis*.
(DOC)

Table S2 Histone acetyltransferases of the MYST family present in *A. deanei* and *S. culicis* compared to other trypanosomes.
(DOC)

Table S3 Distribution of Sirtuins in the protozoan and endosymbiont species.
(DOC)

Table S4 Histone deacetylase identified in *A. deanei* and *S. culicis*.
(DOC)

Table S5 Histone methyltransferase in *A. deanei* and *S. culicis*.
(DOC)

Table S6 Histone chaperones identified in *A. deanei* and *S. culicis*.
(DOC)

Table S7 Bromodomain proteins found in *A. deanei* and *S. culicis*.
(DOC)

Table S8 Components of replication mechanism of the kDNA identified in *A. deanei* and *S. culicis* and similar endosymbionts ORFs.
(DOC)

Table S9 Identified ORFs related to DNA replication and DNA repair in *A. deanei* and *S. culicis*.
(DOC)

Table S10 DNA replication and repair ORFs found in the *A. deanei* and *S. culicis* endosymbionts.
(DOC)

Table S11 Identified ORFs involved in DNA transcription and RNA splicing in the genome of *A. deanei* and *S. culicis*.
(DOC)

Table S12 Transcription related proteins in the endosymbionts of *A. deanei* and *S. culicis*.
(DOC)

Table S13 Main ORFs detected participating in ribosomal biogenesis and translation *in A. deanei* and *S. culicis*.
(DOC)

Table S14 Identified phosphatases in *A. deanei* and *S. culicis*.
(DOC)

Table S15 Number of heat shock and stress response proteins in *A. deanei* and *S.culicis*.
(DOC)

Table S16 Glycerophospholipids (GPL) enzymes of *A. deanei* and *S. culicis*[1].
(DOC)

Table S17 Glycerophospholipids (GPL) enzymes of *A. deanei and S. culicis* endosymbionts.
(DOC)

Table S18 Ectonucleotidases families and identification of ORFs found in *A. deane*i and *S. culicis*.
(DOC)

Table S19 ORFs encoding enzymes involved in purine and pyrimidine metabolism of *A. deanei*, *S. culicis* and their symbionts.
(DOC)

Table S20 Glysosyltransferases found in *A. deanei* and *S. culicis*.
(DOC)

Table S21 Surface proteins of *A. deanei* e *S. culicis*.
(DOC)

Text S1
(DOC)

Acknowledgments

We would like to dedicate this paper to professors Erney Camargo and Marta Teixeira who have made important contributions related to the study of basic aspects of the biology of trypanosomatids, especially those harboring an endosymbiont, and identified several new species of this relevant and interesting group of eukaryotic microorganism.

Author Contributions

Conceived and designed the experiments: MCMM WS SS ATRV. Analyzed the data: MCMM ACAM SSAS CMCCP RS CCK LGPA OLC LPC MB ACC BAL CRM CMAS CMP CBAM CET DCB DFG DPP ECG FFG FKM GFRL GW GHG JLRF MCE MHSG MFS MP PHS RPMN SMRT TEFM TAOM TPÜ WS SS ATRV. Contributed reagents/materials/analysis tools: ATRV LGPA OLC WS. Wrote the paper: MCMM SS ATRV.

References

1. Wallace FG (1966) The trypanosomatid parasites of insects and arachnids. Experimental Parasitology 18: 124–193.
2. Teixeira MM, Borghesan TC, Ferreira RC, Santos MA, Takata CS, et al. (2011) 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 162: 503–524.
3. Edwards C, Chance B (1982) Evidence for the presence of two terminal oxidases in the trypanosomatid *Crithidia oncopelti*. Journal of General Microbiology 128: 1409–1414.
4. Fampa P, Correa-da-Silva MS, Lima DC, Oliveira SM, Motta MC, et al. (2003) Interaction of insect trypanosomatids with mosquitoes, sand fly and the respective insect cell lines. International Journal for Parasitology 33: 1019–1026.

5. de Azevedo-Martins AC, Frossard ML, de Souza W, Einicker-Lamas M, Motta MC (2007) Phosphatidylcholine synthesis in *Crithidia deanei*: the influence of the endosymbiont. FEMS Microbiology Letters 275: 229–236.
6. Motta MCM, Leal LHM, Souza WD, De Almeida DF, Ferreira LCS (1997) Detection of Penicillin-binding Proteins in the Endosymbiont of the Trypanosomatid *Crithidia deanei*. The Journal of Eukaryotic Microbiology 44: 492–496.
7. Chang KP, Chang CS, Sassa S (1975) Heme biosynthesis in bacterium-protozoon symbioses: enzymic defects in host hemoflagellates and complemental role of their intracellular symbiotes. Proceedings of the National Academy of Sciences of the United States of America 72: 2979–2983.
8. Camargo EP, Freymuller E (1977) Endosymbiont as supplier of ornithine carbamoyltransferase in a trypanosomatid. Nature 270: 52–53.
9. Galinari S, Camargo EP (1978) Trypanosomatid protozoa: survey of acetylornithinase and ornithine acetyltransferase. Experimental Parasitology 46: 277–282.
10. Salzman TA, Batlle AM, Angluster J, de Souza W (1985) Heme synthesis in *Crithidia deanei*: influence of the endosymbiote. The International Journal of Biochemistry 17: 1343–1347.
11. Alves JM, Voegtly L, Matveyev AV, Lara AM, da Silva FM, et al. (2011) Identification and phylogenetic analysis of heme synthesis genes in trypanosomatids and their bacterial endosymbionts. PLoS One 6: e23518.
12. Frossard ML, Seabra SH, DaMatta RA, de Souza W, de Mello FG, et al. (2006) An endosymbiont positively modulates ornithine decarboxylase in host trypanosomatids. Biochemical and Biophysical Research Communications 343: 443–449.
13. Motta MC, Soares MJ, Attias M, Morgado J, Lemos AP, et al. (1997) Ultrastructural and biochemical analysis of the relationship of *Crithidia deanei* with its endosymbiont. European Journal of Cell Biology 72: 370–377.
14. Motta MC, Catta-Preta CM, Schenkman S, Azevedo Martins AC, Miranda K, et al. (2010) The bacterium endosymbiont of *Crithidia deanei* undergoes coordinated division with the host cell nucleus. PLoS One 5: e12415.
15. Freymuller E, Camargo EP (1981) Ultrastructural differences between species of trypanosomatids with and without endosymbionts. The Journal of Protozoology 28: 175–182.
16. Gadelha C, Wickstead B, de Souza W, Gull K, Cunha-e-Silva N (2005) Cryptic paraflagellar rod in endosymbiont-containing kinetoplastid protozoa. Eukaryotic Cell 4: 516–525.
17. Cavalcanti DP, Thiry M, de Souza W, Motta MC (2008) The kinetoplast ultrastructural organization of endosymbiont-bearing trypanosomatids as revealed by deep-etching, cytochemical and immunocytochemical analysis. Histochemistry and Cell Biology 130: 1177–1185.
18. Dwyer DM, Chang KP (1976) Surface membrane carbohydrate alterations of a flagellated protozoan mediated by bacterial endosymbiotes. Proceedings of the National Academy of Sciences of the United States of America 73: 852–856.
19. Oda LM, Alviano CS, Filho FCS, Angluster J, Roitman I, et al. (1984) Surface Anionic Groups in Symbiote-Bearing and Symbiote-Free Strains of *Crithidia deanei*. The Journal of Eukaryotic Microbiology 31: 131–134.
20. d'Avila-Levy CM, Silva BA, Hayashi EA, Vermelho AB, Alviano CS, et al. (2005) Influence of the endosymbiont of *Blastocrithidia culicis* and *Crithidia deanei* on the glycoconjugate expression and on *Aedes aegypti* interaction. FEMS Microbiology Letters 252: 279–286.
21. d'Avila-Levy CM, Santos LO, Marinho FA, Matteoli FP, Lopes AH, et al. (2008) *Crithidia deanei*: influence of parasite gp63 homologue on the interaction of endosymbiont-harboring and aposymbiotic strains with *Aedes aegypti* midgut. Experimental Parasitology 118: 345–353.
22. Du Y, Maslov DA, Chang KP (1994) Monophyletic origin of beta-division proteobacterial endosymbionts and their coevolution with insect trypanosomatid protozoa *Blastocrithidia culicis* and *Crithidia* spp. Proceedings of the National Academy of Sciences of the United States of America 91: 8437–8441.
23. Du Y, McLaughlin G, Chang KP (1994) 16S ribosomal DNA sequence identities of beta-proteobacterial endosymbionts in three *Crithidia* species. Journal of Bacteriology 176: 3081–3084.
24. Martin W, Hoffmeister M, Rotte C, Henze K (2001) An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. Biological chemistry 382: 1521–1539.
25. Hollar L, Lukes J, Maslov DA (1998) Monophyly of endosymbiont containing trypanosomatids: phylogeny versus taxonomy. The Journal of Eukaryotic Microbiology 45: 293–297.
26. Hebert L, Moumen B, Duquesne F, Breuil MF, Laugier C, et al. (2011) Genome sequence of *Taylorella equigenitalis* MCE9, the causative agent of contagious equine metritis. Journal of Bacteriology 193: 1785.
27. Hebert L, Moumen B, Pons N, Duquesne F, Breuil MF, et al. (2012) Genomic characterization of the *Taylorella* genus. PLoS One 7: e29953.
28. Sugimoto C, Isayama Y, Sakazaki R, Kuramochi S (1983) Transfer of *Haemophilus equigenitalis* Taylor et al. 1978 to the genus *Taylorella* gen. nov. as *Taylorella equigenitalis* comb. nov. Current Microbiology 9: 155–162.
29. Moran NA, McCutcheon JP, Nakabachi A (2008) Genomics and evolution of heritable bacterial symbionts. Annual Review of Genetics 42: 165–190.
30. Toft C, Andersson SG (2010) Evolutionary microbial genomics: insights into bacterial host adaptation. Nature Reviews Genetics 11: 465–475.
31. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature 407: 81–86.
32. McCutcheon JP, Moran NA (2012) Extreme genome reduction in symbiotic bacteria. Nature Reviews Microbiology 10: 13–26.
33. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, et al. (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nature Genetics 35: 32–40.
34. Cummings CA, Brinig MM, Lepp PW, van de Pas S, Relman DA (2004) Bordetella species are distinguished by patterns of substantial gene loss and host adaptation. Journal of Bacteriology 186: 1484–1492.
35. Gull K (1999) The cytoskeleton of trypanosomatid parasites. Annual Review of Microbiology 53: 629–655.
36. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, et al. (2005) The Genome of the African Trypanosome *Trypanosoma brucei*. Science 309: 416–422.
37. Beech PL, Heimann K, Melkonian M (1991) Development of the Flagellar Apparatus during the Cell-Cycle in Unicellular Algae. Protoplasma 164: 23–37.
38. Lange BM, Gull K (1996) Structure and function of the centriole in animal cells: progress and questions. Trends in Cell Biology 6: 348–352.
39. Garcia-Salcedo JA, Perez-Morga D, Gijon P, Dilbeck V, Pays E, et al. (2004) A differential role for actin during the life cycle of *Trypanosoma brucei*. The EMBO Journal 23: 780–789.
40. Gadelha C, Wickstead B, Gull K (2007) Flagellar and ciliary beating in trypanosome motility. Cell Motility and the Cytoskeleton 64: 629–643.
41. Portman N, Gull K (2010) The paraflagellar rod of kinetoplastid parasites: from structure to components and function. International Journal for Parasitology 40: 135–148.
42. Lacomble S, Vaughan S, Gadelha C, Morphew MK, Shaw MK, et al. (2009) Three-dimensional cellular architecture of the flagellar pocket and associated cytoskeleton in trypanosomes revealed by electron microscope tomography. Journal of Cell Science 122: 1081–1090.
43. Oberholzer M, Marti G, Baresic M, Kunz S, Hemphill A, et al. (2007) The *s* cAMP phosphodiesterases TbrPDEB1 and TbrPDEB2: flagellar enzymes that are essential for parasite virulence. The FASEB Journal 21: 720–731.
44. Ginger ML, Portman N, McKean PG (2008) Swimming with protists: perception, motility and flagellum assembly. Nature Reviews Microbiology 6: 838–850.
45. Xu C, Ray DS (1993) Isolation of proteins associated with kinetoplast DNA networks in vivo. Proceedings of the National Academy of Sciences of the United States of America 90: 1786–1789.
46. Ersfeld K, Barraclough H, Gull K (2005) Evolutionary relationships and protein domain architecture in an expanded calpain superfamily in kinetoplastid parasites. Journal of Molecular Evolution 61: 742–757.
47. Avliyakulov NK, Lukes J, Ray DS (2004) Mitochondrial histone-like DNA-binding proteins are essential for normal cell growth and mitochondrial function in *Crithidia fasciculata*. Eukaryotic Cell 3: 518–526.
48. Cavalcanti DP, Shimada MK, Probst CM, Souto-Padron TC, de Souza W, et al. (2009) Expression and subcellular localization of kinetoplast-associated proteins in the different developmental stages of *Trypanosoma cruzi*. BMC Microbiology 9: 120.
49. Wei K, Clark AB, Wong E, Kane MF, Mazur DJ, et al. (2003) Inactivation of Exonuclease 1 in mice results in DNA mismatch repair defects, increased cancer susceptibility, and male and female sterility. Genes & Development 17: 603–614.
50. Wu Y, Berends MJ, Post JG, Mensink RG, Verlind E, et al. (2001) Germline mutations of EXO1 gene in patients with hereditary nonpolyposis colorectal cancer (HNPCC) and atypical HNPCC forms. Gastroenterology 120: 1580–1587.
51. Kim YR, Yoo NJ, Lee SH (2010) Somatic mutation of EXO1 gene in gastric and colorectal cancers with microsatellite instability. Acta oncologica 49: 859–860.
52. Augusto-Pinto L, Teixeira SM, Pena SD, Machado CR (2003) Single-nucleotide polymorphisms of the *Trypanosoma cruzi* MSH2 gene support the existence of three phylogenetic lineages presenting differences in mismatch-repair efficiency. Genetics 164: 117–126.
53. Machado CR, Augusto-Pinto L, McCulloch R, Teixeira SM (2006) DNA metabolism and genetic diversity in Trypanosomes. Mutation Research 612: 40–57.
54. Andreeva AV, Kutuzov MA (2008) Protozoan protein tyrosine phosphatases. International Journal for Parasitology 38: 1279–1295.
55. Brenchley R, Tariq H, McElhinney H, Szoor B, Huxley-Jones J, et al. (2007) The TriTryp phosphatome: analysis of the protein phosphatase catalytic domains. BMC Genomics 8: 434.
56. Szoor B, Wilson J, McElhinney H, Tabernero L, Matthews KR (2006) Protein tyrosine phosphatase TbPTP1: a molecular switch controlling life cycle differentiation in trypanosomes. The Journal of Cell Biology 175: 293–303.
57. Huang H (2011) Signal transduction in *Trypanosoma cruzi*. Advances in Parasitology 75: 325–344.
58. Atayde VD, Tschudi C, Ullu E (2011) The emerging world of small silencing RNAs in protozoan parasites. Trends in Parasitology 27: 321–327.

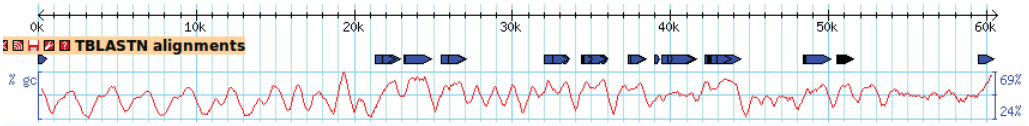
59. Barnes RL, Shi H, Kolev NG, Tschudi C, Ullu E (2012) Comparative genomics reveals two novel RNAi factors in *Trypanosoma brucei* and provides insight into the core machinery. PLoS Pathogens 8: e1002678.
60. Van Hellemond JJ, Neuville P, Schwarz RT, Matthews KR, Mottram JC (2000) Isolation of *Trypanosoma brucei* CYC2 and CYC3 cyclin genes by rescue of a yeast G(1) cyclin mutant. Functional characterization of CYC2. The Journal of Biological Chemistry 275: 8315–8323.
61. Carballido-Lopez R, Errington J (2003) A dynamic bacterial cytoskeleton. Trends in Cell Biology 13: 577–583.
62. Pichoff S, Lutkenhaus J (2002) Unique and overlapping roles for ZipA and FtsA in septal ring assembly in *Escherichia coli*. The EMBO Journal 21: 685–693.
63. Harry E, Monahan L, Thompson L (2006) Bacterial cell division: the mechanism and its precison. International Review of Cytology 253: 27–94.
64. Margolin W (2005) FtsZ and the division of prokaryotic cells and organelles. Nature Reviews Molecular Cell Biology 6: 862–871.
65. Buddelmeijer N, Beckwith J (2004) A complex of the *Escherichia coli* cell division proteins FtsL, FtsB and FtsQ forms independently of its localization to the septal region. Molecular Microbiology 52: 1315–1327.
66. Chen JC, Beckwith J (2001) FtsQ, FtsL and FtsI require FtsK, but not FtsN, for co-localization with FtsZ during *Escherichia coli* cell division. Molecular Microbiology 42: 395–413.
67. Chen JC, Weiss DS, Ghigo JM, Beckwith J (1999) Septal localization of FtsQ, an essential cell division protein in *Escherichia coli*. Journal of Bacteriology 181: 521–530.
68. Mercer KL, Weiss DS (2002) The *Escherichia coli* cell division protein FtsW is required to recruit its cognate transpeptidase, FtsI (PBP3), to the division site. Journal of Bacteriology 184: 904–912.
69. Bouhss A, Trunkfield AE, Bugg TD, Mengin-Lecreulx D (2008) The biosynthesis of peptidoglycan lipid-linked intermediates. FEMS microbiology reviews 32: 208–233.
70. Ni Y, Chen R (2009) Extracellular recombinant protein production from *Escherichia coli*. Biotechnology Letters 31: 1661–1670.
71. Ni Y, Reye J, Chen RR (2007) lpp deletion as a permeabilization method. Biotechnology and Bioengineering 97: 1347–1356.
72. Mingorance J, Tamames J, Vicente M (2004) Genomic channeling in bacterial cell division. Journal of molecular recognition 17: 481–487.
73. Motta MC, Picchi GF, Palmie-Peixoto IV, Rocha MR, de Carvalho TM, et al. (2004) The microtubule analog protein, FtsZ, in the endosymbiont of trypanosomatid protozoa. The Journal of Eukaryotic Microbiology 51: 394–401.
74. Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature Reviews Genetics 5: 123–135.
75. Pyke KA (2010) Plastid division. AoB plants 2010: plq016.
76. Motta MC (2010) Endosymbiosis in trypanosomatids as a model to study cell evolution. The Open Parasitology Journal 4: 139–147.
77. Opperdoes FR, Michels PA (2008) Complex I of Trypanosomatidae: does it exist? Trends in Parasitology 24: 310–317.
78. Morales J, Mogi T, Mineki S, Takashima E, Mineki R, et al. (2009) Novel mitochondrial complex II isolated from *Trypanosoma cruzi* is composed of 12 peptides including a heterodimeric Ip subunit. The Journal of Biological Chemistry 284: 7255–7263.
79. Edwards C (1984) Terminal oxidases of *Crithidia oncopelti*. FEMS Microbiology Letters 21: 319–322.
80. Palmie-Peixoto IV, Rocha MR, Urbina JA, de Souza W, Einicker-Lamas M, et al. (2006) Effects of sterol biosynthesis inhibitors on endosymbiont-bearing trypanosomatids. FEMS Microbiology Letters 255: 33–42.
81. Mundim MH, Roitman I, Hermans MA, Kitajima EW (1974) Simple nutrition of *Crithidia deanei*, a reduviid trypanosomatid with an endosymbiont. The Journal of Protozoology 21: 518–521.
82. Newton BA (1956) A synthetic growth medium for the trypanosomid flagellate *Strigomonas* (Herpetomonas) *oncopelti*. Nature 177: 279–280.
83. Newton BS (1957) Nutritional requirements and biosynthetic capabilities of the parasitic flagellate *Strigomonas oncopelti*. Journal of General Microbiology 17: 708–717.
84. Camargo EP, Coelho JA, Moraes G, Figueiredo EN (1978) *Trypanosoma* spp., *Leishmania* spp. and *Leptomonas* spp.: enzymes of ornithine-arginine metabolism. Experimental Parasitology 46: 141–144.
85. Gill JW, Vogel HJ (1963) A Bacterial Endosymbiote in *Crithidia* (Strigomonas) *oncopelti*: Biochemical and Morphological Aspects. The Journal of Eukaryotic Microbiology 10: 148–152.
86. Marr JJ, Berens RL, Nelson DJ (1978) Purine metabolism in *Leishmania donovani* and *Leishmania braziliensis*. Biochimica et Biophysica Acta 544: 360–371.
87. Ceron CR, Caldas RD, Felix CR, Mundim MH, Roitman I (1979) Purine metabolism in trypanosomatids. The Journal of Protozoology 26: 479–483.
88. Berens RL, Krugg EC, Marr JJ (1995) Purine and Pyrimidine Metabolism. In: Marr JJ, Muller M, editors. Biochemistry and Molecular Biology of Parasites. London: Academic Press. 89–117.
89. Zimmermann H (2000) Extracellular metabolism of ATP and other nucleotides. Naunyn-Schmiedeberg's Archives of Pharmacology 362: 299–309.
90. Plesner L (1995) Ecto-ATPases: identities and functions. International Review of Cytology 158: 141–214.
91. Sansom FM, Robson SC, Hartland EL (2008) Possible effects of microbial ecto-nucleoside triphosphate diphosphohydrolases on host-pathogen interactions. Microbiology and Molecular Biology Reviews 72: 765–781.
92. Maioli TU, Takane E, Arantes RM, Fietto JL, Afonso LC (2004) Immune response induced by New World Leishmania species in C57BL/6 mice. Parasitology Research 94: 207–212.
93. Marques da Silva C, Miranda Rodrigues L, Passos da Silva Gomes A, Mantuano Barradas M, Sarmento Vieira F, et al. (2008) Modulation of P2X7 receptor expression in macrophages from mineral oil-injected mice. Immunobiology 213: 481–492.
94. Rebora K, Desmoucelles C, Borne F, Pinson B, Daignan-Fornier B (2001) Yeast AMP pathway genes respond to adenine through regulated synthesis of a metabolic intermediate. Molecular and Cellular Biology 21: 7901–7912.
95. Zalkin H, Nygaard P (1996) Biosynthesis of purine nucleotides. In: Frederick Carl N, editor. *Escherichia coli* and *Salmonella* : cellular and molecular biology. 2 ed. Washington, D.C.: ASM Press. 561–579.
96. Podlipaev SA (2000) Insect trypanosomatids: the need to know more. Memorias do Instituto Oswaldo Cruz 95: 517–522.
97. Correa-da-Silva MS, Fampa P, Lessa LP, Silva Edos R, dos Santos Mallet JR, et al. (2006) Colonization of *Aedes aegypti* midgut by the endosymbiont-bearing trypanosomatid *Blastocrithidia culicis*. Parasitology Research 99: 384–391.
98. Nascimento MT, Garcia MC, da Silva KP, Pinto-da-Silva LH, Atella GC, et al. (2010) Interaction of the monoxenic trypanosomatid *Blastocrithidia culicis* with the *Aedes aegypti* salivary gland. Acta Tropica 113: 269–278.
99. Lairson LL, Henrissat B, Davies GJ, Withers SG (2008) Glycosyltransferases: structures, functions, and mechanisms. Annual Review of Biochemistry 77: 521–555.
100. Mengeling BJ, Turco SJ (1998) Microbial glycoconjugates. Current Opinion in Structural Biology 8: 572–577.
101. Schwarz F, Aebi M (2011) Mechanisms and principles of N-linked protein glycosylation. Current Opinion in Structural Biology 21: 576–582.
102. Oppenheimer M, Valenciano AL, Sobrado P (2011) Biosynthesis of galactofuranose in kinetoplastids: novel therapeutic targets for treating leishmaniasis and chagas' disease. Enzyme research 2011: 415976.
103. de Lederkremer RM, Colli W (1995) Galactofuranose-containing glycoconjugates in trypanosomatids. Glycobiology 5: 547–552.
104. Moraes CT, Bosch M, Parodi AJ (1988) Structural characterization of several galactofuranose-containing, high-mannose-type oligosaccharides present in glycoproteins of the trypanosomatid *Leptomonas samueli*. Biochemistry 27: 1543–1549.
105. Mendelzon DH, Previato JO, Parodi AJ (1986) Characterization of protein-linked oligosaccharides in trypanosomatid flagellates. Molecular and Biochemical Parasitology 18: 355–367.
106. Mendelzon DH, Parodi AJ (1986) N-linked high mannose-type oligosaccharides in the protozoa *Crithidia fasciculata* and *Crithidia harmosa* contain galactofuranose residues. The Journal of Biological Chemistry 261: 2129–2133.
107. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, et al. (2005) Comparative genomics of trypanosomatid parasitic protozoa. Science 309: 404–409.
108. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, et al. (2005) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science 309: 409–415.
109. Tull D, Vince JE, Callaghan JM, Naderer T, Spurck T, et al. (2004) SMP-1, a member of a new family of small myristoylated proteins in kinetoplastid parasites, is targeted to the flagellum membrane in *Leishmania*. Molecular Biology of the Cell 15: 4775–4786.
110. Galetovic A, Souza RT, Santos MR, Cordero EM, Bastos IM, et al. (2011) The repetitive cytoskeletal protein H49 of *Trypanosoma cruzi* is a calpain-like protein located at the flagellum attachment zone. PLoS One 6: e27634.
111. Yao C, Li Y, Donelson JE, Wilson ME (2010) Proteomic examination of *Leishmania chagasi* plasma membrane proteins: Contrast between avirulent and virulent (metacyclic) parasite forms. Proteomics Clinical applications 4: 4–16.
112. Yao C, Donelson JE, Wilson ME (2003) The major surface protease (MSP or GP63) of *Leishmania* sp. Biosynthesis, regulation of expression, and function. Molecular and Biochemical Parasitology 132: 1–16.
113. Grandgenett PM, Otsu K, Wilson HR, Wilson ME, Donelson JE (2007) A function for a specific zinc metalloprotease of African trypanosomes. PLoS Pathogens 3: 1432–1445.
114. Nogueira de Melo AC, d'Avila-Levy CM, Dias FA, Armada JL, Silva HD, et al. (2006) Peptidases and gp63-like proteins in *Herpetomonas megaseliae*: possible involvement in the adhesion to the invertebrate host. International Journal for Parasitology 36: 415–422.
115. Cazzulo JJ (2002) Proteinases of *Trypanosoma cruzi*: patential targets for the chemotherapy of Chagas desease. Current Topics in Medicinal Chemistry 2: 1261–1271.
116. Caffrey CR, Lima AP, Steverding D (2011) Cysteine peptidases of kinetoplastid parasites. Advances in experimental medicine and biology 712: 84–99.
117. Teixeira SM, Russell DG, Kirchhoff LV, Donelson JE (1994) A differentially expressed gene family encoding "amastin," a surface protein of *Trypanosoma cruzi* amastigotes. The Journal of Biological Chemistry 269: 20509–20516.
118. Wu Y, El Fakhry Y, Sereno D, Tamar S, Papadopoulou B (2000) A new developmentally regulated gene family in *Leishmania* amastigotes encoding a homolog of amastin surface proteins. Molecular and Biochemical Parasitology 110: 345–357.

119. Jackson AP (2010) The evolution of amastin surface glycoproteins in trypanosomatid parasites. Molecular Biology and Evolution 27: 33–45.
120. Drummond AJ, Ashton B, Buxton S, Cheung M, A C, et al. (2011) Geneious v5.5.
121. Schenkman S, Pascoalino Bdos S, Nardelli SC (2011) Nuclear Structure of *Trypanosoma cruzi*. Advances in Parasitology 75: 251–283.
122. Glover L, Horn D (2012) Trypanosomal histone gammaH2A and the DNA damage response. Molecular and Biochemical Parasitology 183: 78–83.

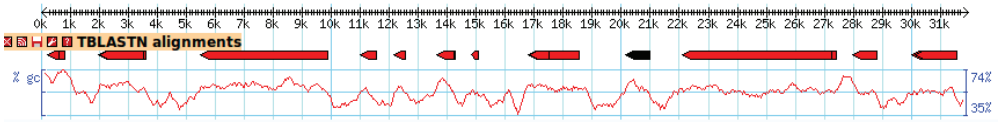
Appendix C

Additional material: 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

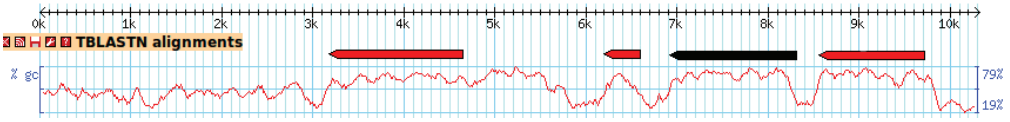
1.1.1.3 – Homoserine dehydrogenase, *Angomonas desouzai*. Contig's GC content: 46.15%.



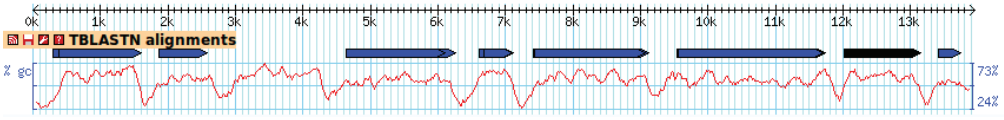
5.1.1.7 – Diaminopimelate epimerase, *Herpetomonas muscarum*. Contig's GC content: 54.84%.



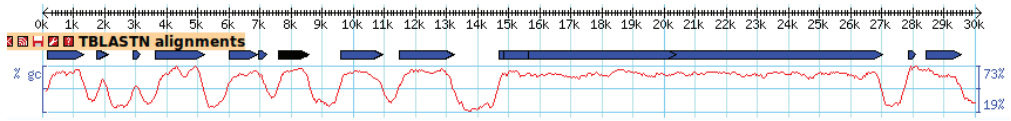
4.1.1.20 – Diaminopimelate decarboxylase, *Strigomonas galati*. Contig's GC content: 53.05%.



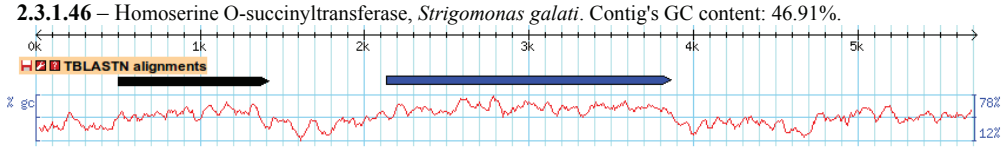
2.3.1.30 – Serine O-acetyltransferase, *Angomonas deanei*. Contig's GC content: 53.20%.



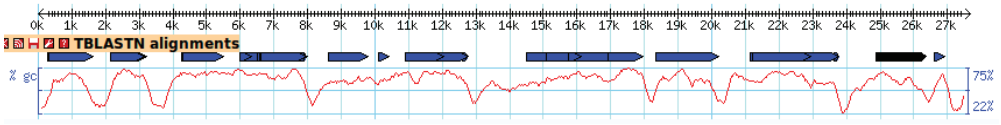
2.5.1.47 – Cysteine synthase, *Strigomonas galati*. Contig's GC content: 55.93%.



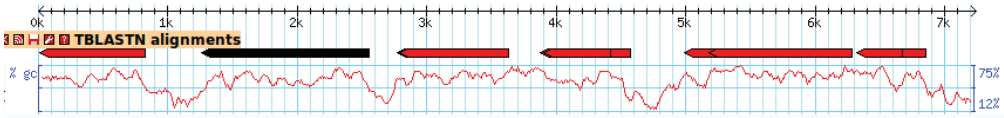
2.3.1.46 – Homoserine O-succinyltransferase, *Strigomonas galati*. Contig's GC content: 46.91%.



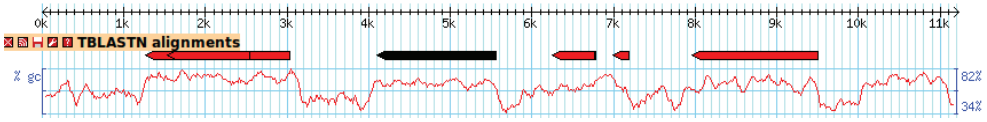
2.1.1.37 – DNA (cytosine-5-)-methyltransferase, *Strigomonas galati*. Contig's GC content: 59.10%.



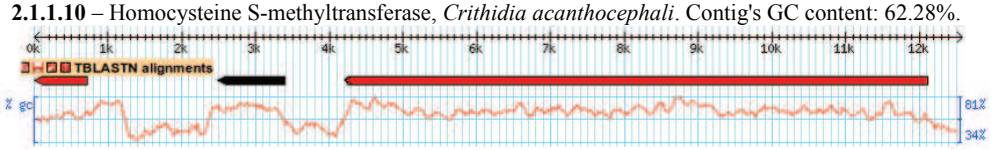
2.5.1.49 – O-acetylhomoserine aminocarboxypropyltransferase, *Angomonas deanei*. Contig's GC content: 54.02%.



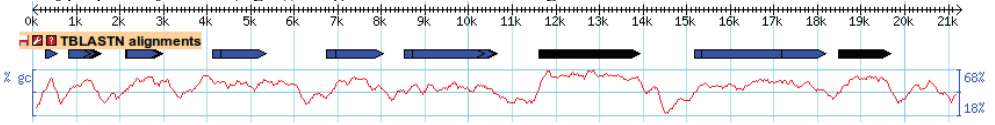
4.4.1.8 – Cystathionine beta-lyase, *Herpetomonas muscarum*. Contig's GC content: 61.10%.



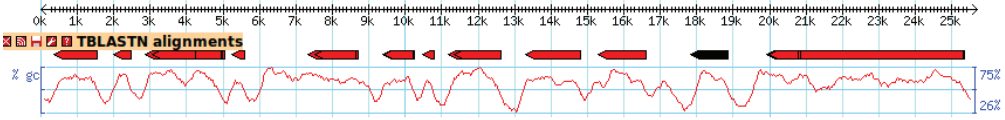
2.1.1.10 – Homocysteine S-methyltransferase, *Crithidia acanthocephali*. Contig's GC content: 62.28%.



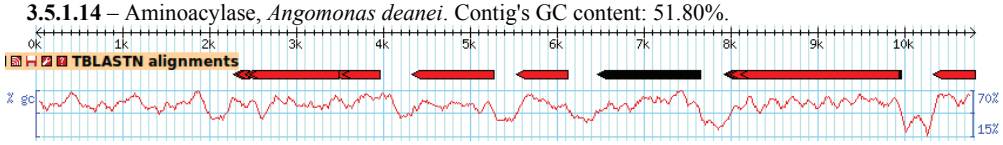
2.1.1.14 – 5-methyltetrahydropteroyltriglutamate-homocysteine S-methyltransferase (left) and **4.2.1.20** – tryptophan synthase (right), *Strigomonas culicis*. Contig's GC content: 48.69%.



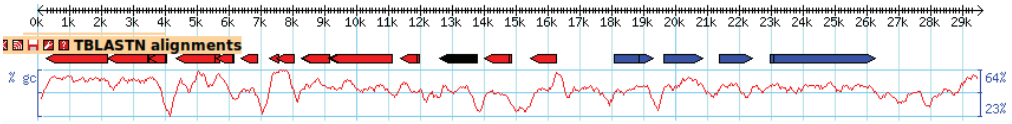
4.1.2.5 – Threonine aldolase, *Strigomonas culicis*. Contig's GC content: 57.13%.



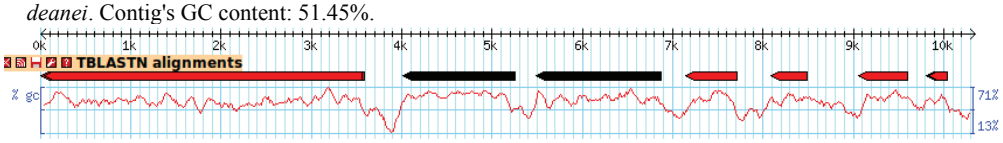
3.5.1.14 – Aminoacylase, *Angomonas deanei*. Contig's GC content: 51.80%.



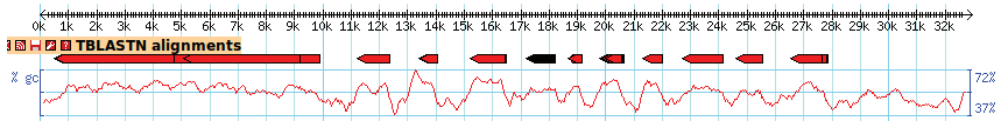
3.5.1.16 – Acetylornithine deacetylase, *Angomonas desouzai*. Contig's GC content: 47.01%.



6.3.4.5 – Argininosuccinate synthase (left) and **4.3.2.1** – Argininosuccinate lyase (right), *Angomonas deanei*. Contig's GC content: 51.45%.



3.5.3.1 – Arginase, *Herpetomonas muscarum*. Contig's GC content: 54.07%.



4.3.1.12 – Ornithine cyclodeaminase, *Angomonas deanei*. Contig's GC content: 48.93%.

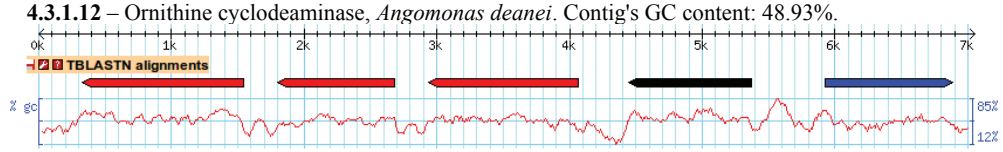


Figure C.1: **Genomic context and GC content for candidate HGT genes in the Trypanosomatidae analyzed in this work.** Three-page figure. Arrows show TBLASTN alignments of the genome against UniRef100 and KEGG proteins. Alignment orientation is displayed in blue or red, except for the alignment for the gene currently in focus, which is colored black. Coordinates are in kilobases.

Figure C.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.

Summary of the phylogenetic and sequencing coverage analyses

EC number	Enzyme name	Pathway	Alignment length	# distinct aligment patterns	Presence	Cluster / sister group of trypanosomatids	Cluster / sister group of TPEs	Figure / Additional file	Genome's average coverage*	Contig's average coverage*	Gene's average coverage*	Genome*
5.1.1.7	diaminopimelate epimerase	Lysine	1661	1409	Only in *Herpetomonas* and TPEs	Group within Bacteroidetes (BSV=99).	Group with the Alcaligenaceae family (BS=98).	Additional file 2	18x	17x	21x	*H. muscarum*
4.1.1.20	diaminopimelate decarboxylase	Lysine	2171	1869	SHTs (*A. desouzai* not in the tree, incomplete sequence similar to *A. deanei*) and *Herpetomonas* (but not the other RTs) and TPEs.	Group with a few other Eukaryota (Dictyostelium, Polysphondylium, and Capsaspora, BSV=65), all as the sister group of a large group of Actinobacteria (BSV=79).	Group within the Alcaligenaceae family (BS=95).	Additional file 3	28x	19x	21x	*S. galati*
2.3.1.30	serine O-acetyltransferase	Cysteine	1500	1248	SHTs and RTs	Group inside a group of diverse bacterial taxa (BS=80), among them there are Bacteroidetes, Betaproteobacteria, Gammaproteobacteria and Firmicutes.	-	Additional file 4	23x	19x	31x	*A. deanei*
2.5.1.47	cysteine synthase	Cysteine	2809	2577	SHTs and RTs	SHTs, RTs (one copy of *T. cruzi* CL Brener) and one *Entamoeba* sp. clade has Actinobacteria as sister group (low BS), and far from the other eukaryotic groups.	-	Additional file 5	28x	23x	28x	*S. galati*
2.3.1.46	homoserine O-succinyltransferase	Cysteine – Methionine	975	808	SHTs and *Herpetomonas*, but not the other RTs	Trypanosomatid clade (BS=91). Group within the Bacteroidetes (BS=53).	-	Additional file 6	28x	19x	16x	*S. galati*
2.1.1.37	DNA (cytosine-5-)-methyltransferase	Cysteine – Methionine	2123	2095	SHTs and RTs	The few Eukaryota in the tree form a clade (low BS) among the several different phyla of Bacteria (low BS).	-	Additional file 7	28x	23x	22x	*S. galati*
2.5.1.48	cystathionine gamma-synthase	Cysteine – Methionine	2848	2522	SHTs and RTs	Group with *Trypanosoma* sp. and a few other Eukaryota, mostly Apicomplexa and Stramenopiles (BS=100).	-	Additional file 8	ND	ND	ND	
2.5.1.49	O-acetylhomoserine aminocarboxypropyltransferase	Cysteine – Methionine	2848	2522	SHTs and *Herpetomonas*	Cluster with diverse groups of Bacteria (low BS).	-	Additional file 8	24x	27x	53x	*A. desouzai*
4.4.1.8	cystathionine beta-lyase	Cysteine – Methionine	2848	2522	SHTs and RTs	Two copies: one clusters with eukaryotes (BS=95) and the other seems to be of bacterial descent, grouping mostly with Alphaproteobacteria of the Rhizobiales order (BS=99).	-	Additional file 8	18x	20x	21x	*H. muscarum*
2.1.1.10	homocysteine S-methyltransferase	Methionine	2771	2445	all SHTs and RTs	Clade SHTs and RTs (BS=96) grouped next to mostly Gammaproteobacteria (low BS).	-	Additional file 9	18x	14x	11x	*C. acanthocephali*
2.1.1.14	5-methyltetrahydropteroyltriglutamate---homocysteine S-methyltransferase	Methionine	1811	1662	SHTs and RTs (except *Herpetomonas*)	The *Angomonas* species group with *C. acanthocephali* (BS=98) while the *Strigomonas* group with *Leishmania* (BS=100). The trypanosomatids group deep within the Gammaproteobacteria (BS=74).	-	Additional file 10	23x	15x	15x	*S. culicis*
2.7.1.100	S-methyl-5-thioribose kinase	Methionine salvage	1358	1080	*C. acanthocephali* and *Herpetomonas*	Group deep within the Gammaproteobacteria (BS=97).	-	Additional file 11	18x	15x	24x	*H. muscarum*
1.1.1.3	homoserine dehydrogenase	Cysteine – Methionine Threonine	2189	1921	SHTs, RTs and TPEs	Cluster within the Firmicutes, with *Solibacillus silverstris* and *Lysinibacillus fusiformis* and *L. sphaericus* as sister group (BS=100).	Placed in the Alcaligenaceae family (BS=91).	Figure 8	24x	26x	34x	*A. desouzai*
4.1.2.5	L-threonine aldolase	Glycine <=> Threonine	2083	1700	SHTs and RTs (except *Herpetomonas*)	SHTs and RTs group in very distant clades: *Leishmania* and *C. acanthocephali* group within Firmicutes, specially *Clostridium* (BS=63), while the SHT group basally with Eukaryota, interrupted by an assorted group of Bacteria (low BS). SHTs and RTs are in opposite sides of the tree, separated by a realtively long branch.	-	Additional file 12	23x	16x	12x	*S. culicis*
4.2.1.20	tryptophan synthase	Tryptophan	1392	1339	SHTs, *Herpetomonas* and TPEs	SHTs and *Herpetomonas* group together (BS=50) and group robustly with the Bacteroidetes phylum (BS=97).	Group with the Alcaligenaceae family (BS=90).	Additional file 13	23x	15x	21x	*S. culicis*
3.5.1.14	aminoacylase	Ornithine	2397	1920	SHTs and RTs	*Angomonas* has two different copies (in different clades), while *Strigomonas* have one (grouped with *Angomonas*, BS=96). RTs also have multiple different copies. All gene copies group as one clade and have as neasrest siter group a Gammaproteobacterium (BS=98). Group with Bacteria of different phyla (low BS).	-	Additional file 14	23x	36x	36x	*A. deanei*
3.5.1.16	acetylornithine deacetylase	Ornithine	920	831	SHTs and RTs	The gene copies of SHTs and RTs group together (BS=84). Group within mainly Betaproteobacteria (BS= 80), and no other Eukaryota seem to have these orthologs.	-	Additional file 15	24x	17x	13x	*A. desouzai*
6.3.4.5	argininosuccinate synthase	Arginine	2322	1311	SHTs and RTs	Clade trypanosomatid (BS=100) group within mainly Firmicutes (BS=69).	-	Additional file 16	23x	28x	35x	*A. deanei*
4.3.2.1	argininosuccinate lyase	Arginine	1960	1520	SHTs and RTs	Clade trypanosomatid (BS=100) group within Firmicutes (BS=82).	-	Additional file 17	23x	28x	31x	*A. deanei*
3.5.3.1	arginase	Ornithine	2403	1548	SHTs and RTs	All but the *Herpetomonas* ortholog are of eukaryotic origin (BS=61). *Herpetomonas* gene groups in a distant bacterial clade (BS=79) containing several different assorted phyla.	-	Additional file 18	18x	21x	15x	*H. muscarum*
4.3.1.12	ornithine cyclodeaminase	Ornithine <=> Proline	1108	990	Only in SHTs	Group close to several Alcaligenaceae (low BS).	-	Additional file 19	23x	15x	19x	*A. deanei*

* Genome, contig, and gene average sequencing coverages were calculated for the organism indicated in the “Genome” column. ND: not determined.

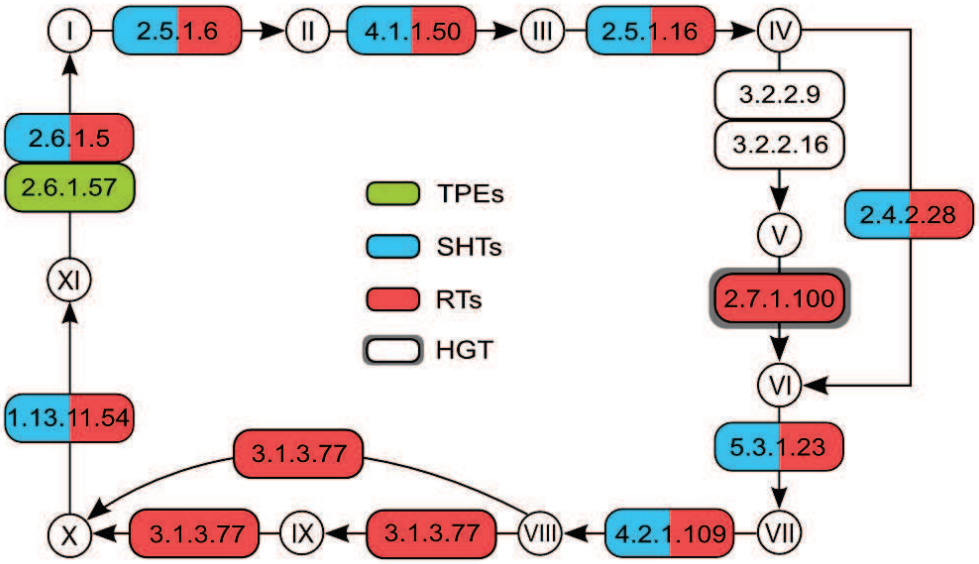


Figure C.3: **Methionine salvage pathway.** Enzymes surrounded by a thick gray box were shown to be horizontally transferred from Bacteria (see main text). Metabolites – I: methionine; II: S-adenosylmethionine; III: S-adenosylmethioninamine; IV: S-methyl-5-thioadenosine; V: S-methyl-5-thioribose; VI: 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-methylthiopentenyl-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.

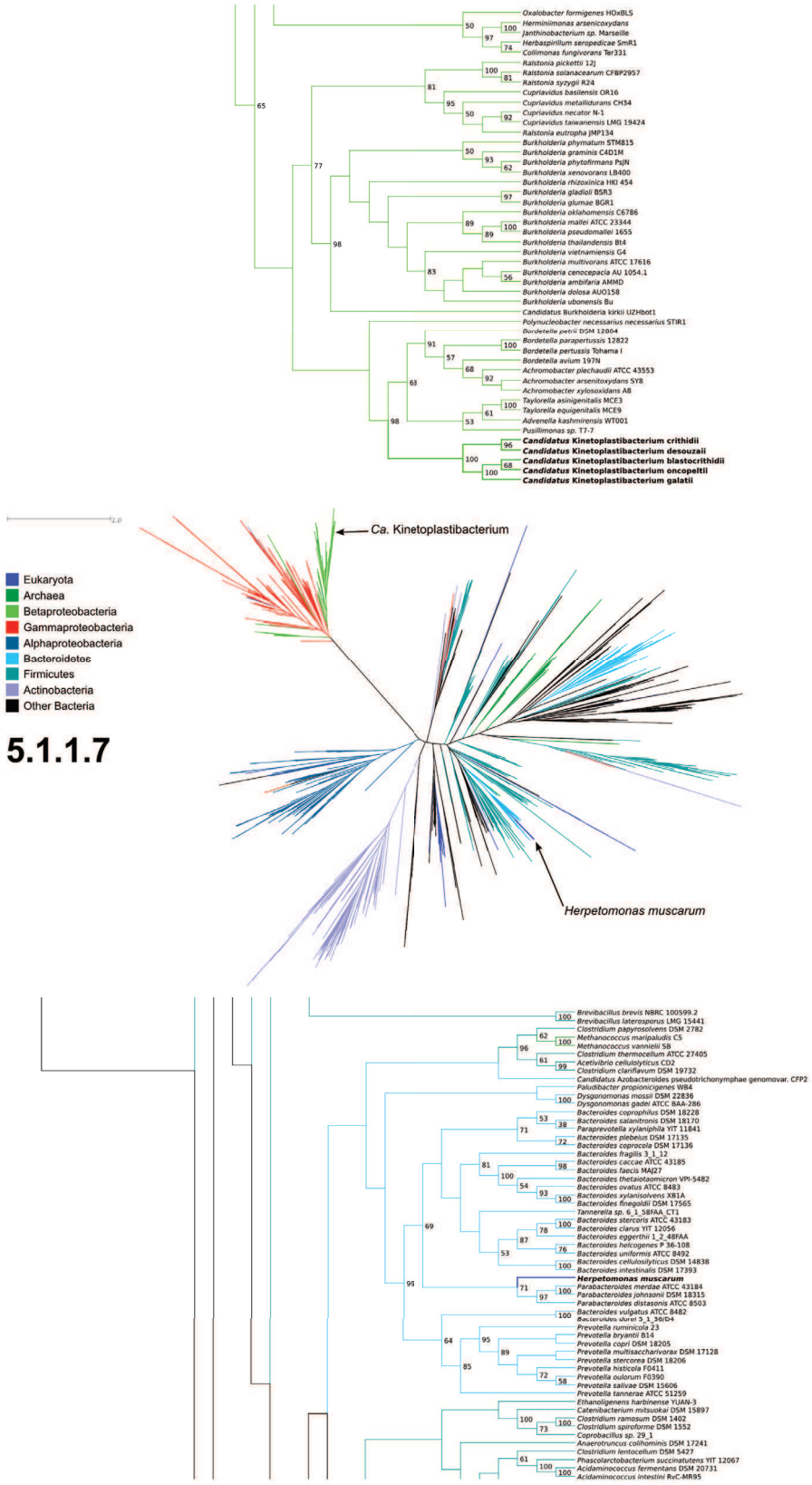


Figure C.4: **Maximum likelihood phylogeny of diaminopimelate epimerase (EC:5.1.1.7).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

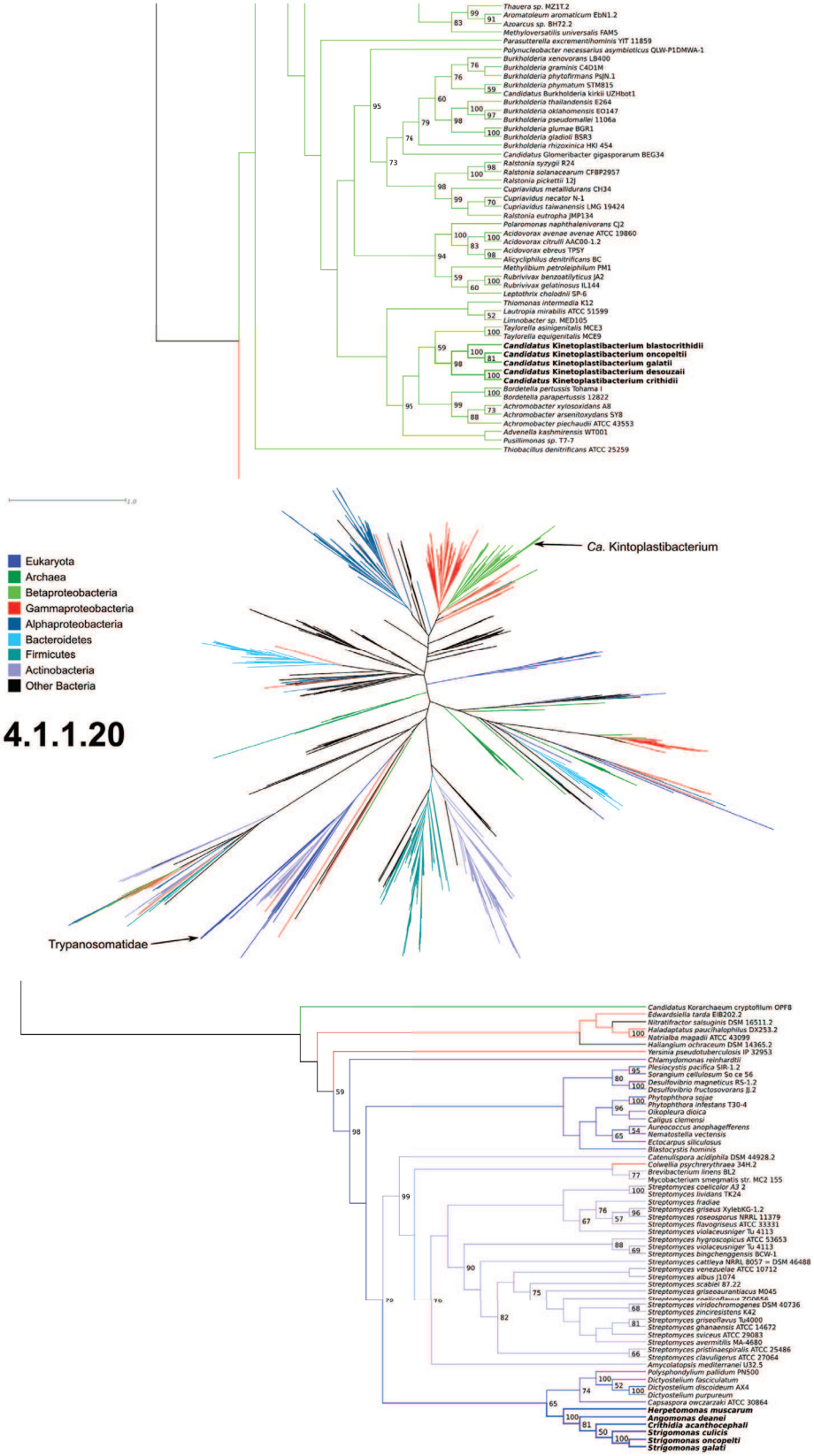


Figure C.5: **Maximum likelihood phylogeny of diaminopimelate decarboxylase (EC:4.1.1.20).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

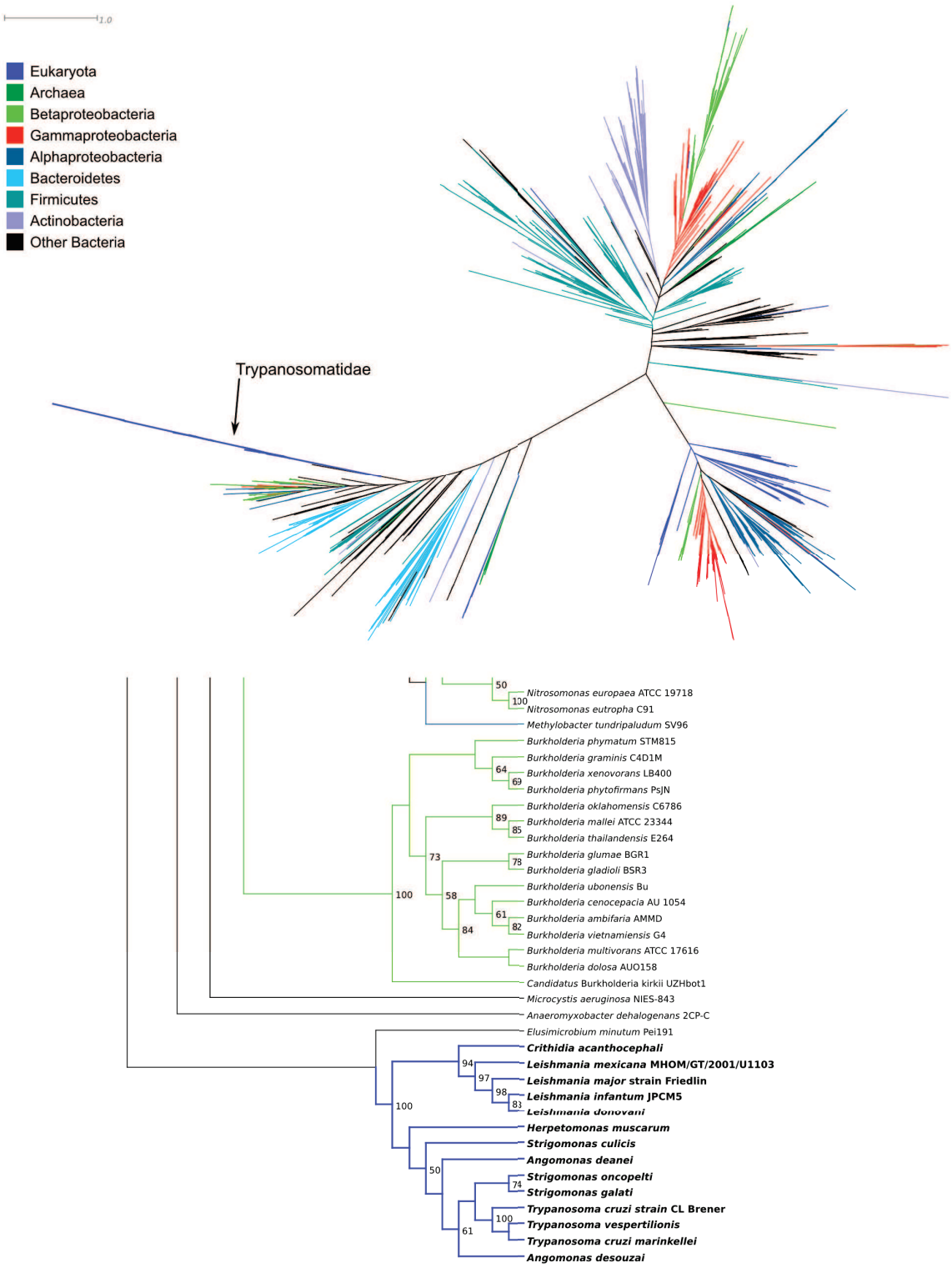


Figure C.6: **Maximum likelihood phylogeny of serine O-acetyltransferase (EC:2.3.1.30).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

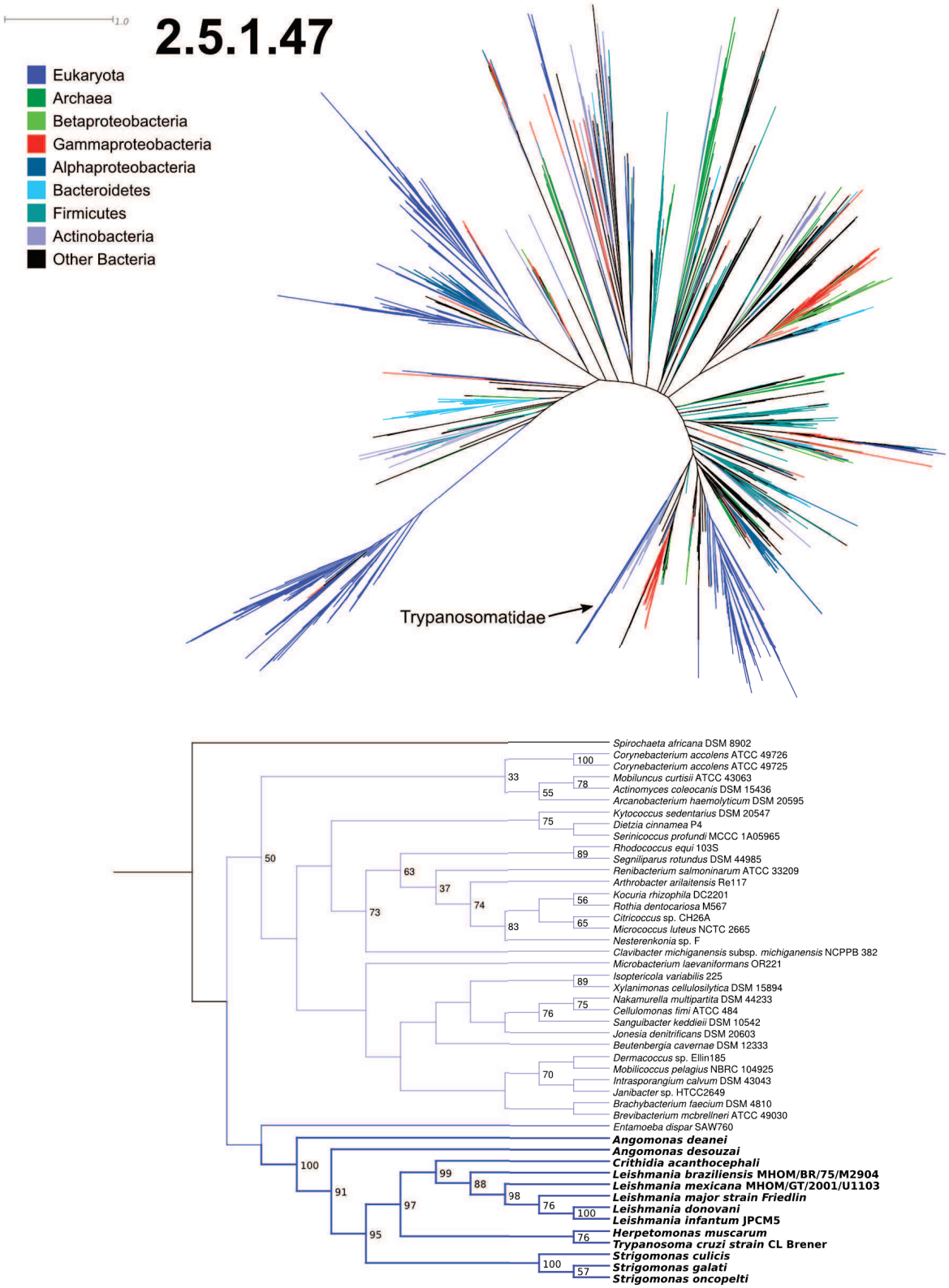


Figure C.7: **Maximum likelihood phylogeny of cysteine synthase (EC:2.5.1.47).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

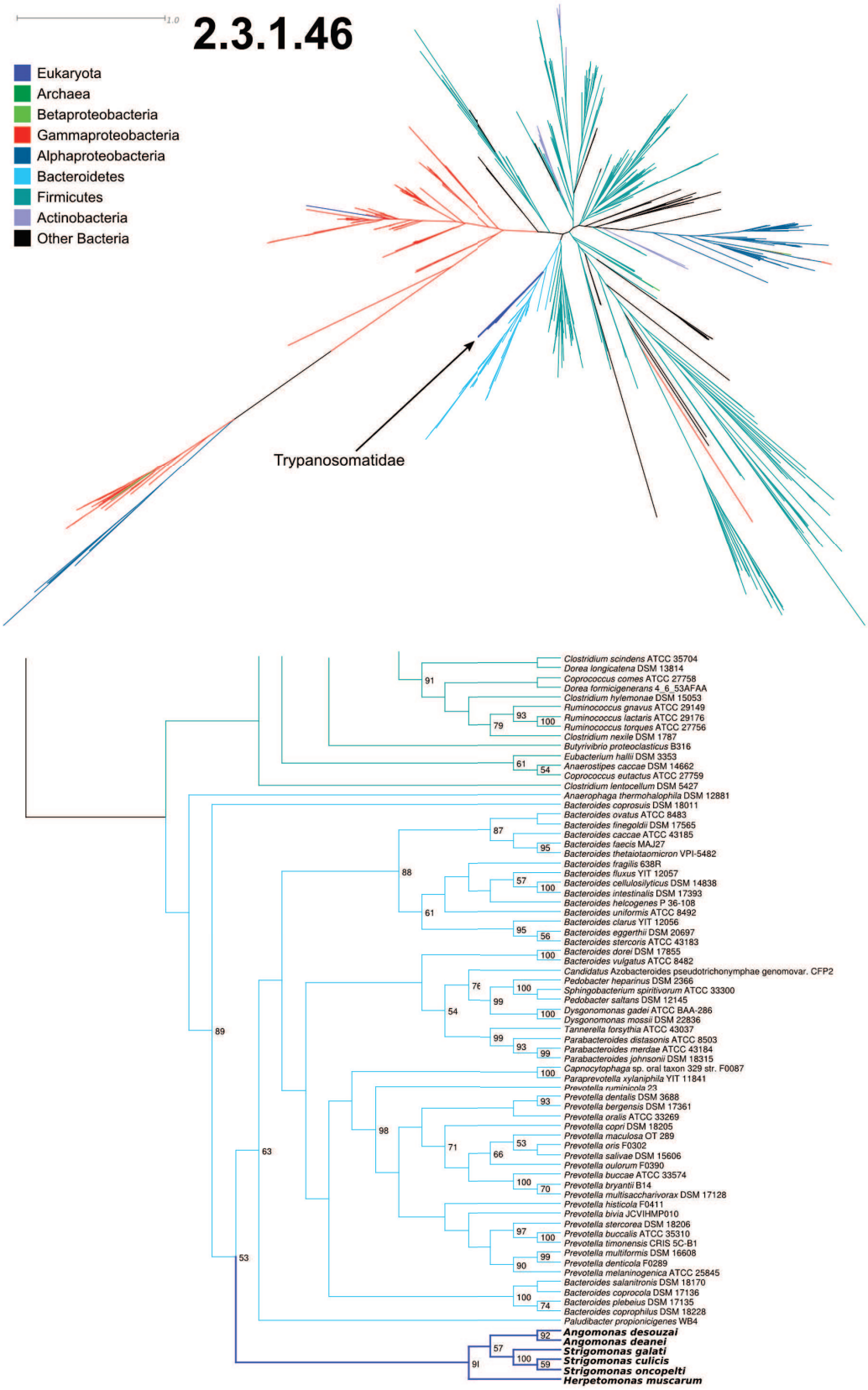


Figure C.8: **Maximum likelihood phylogeny of homoserine O-succinyltransferase (EC:2.3.1.46).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

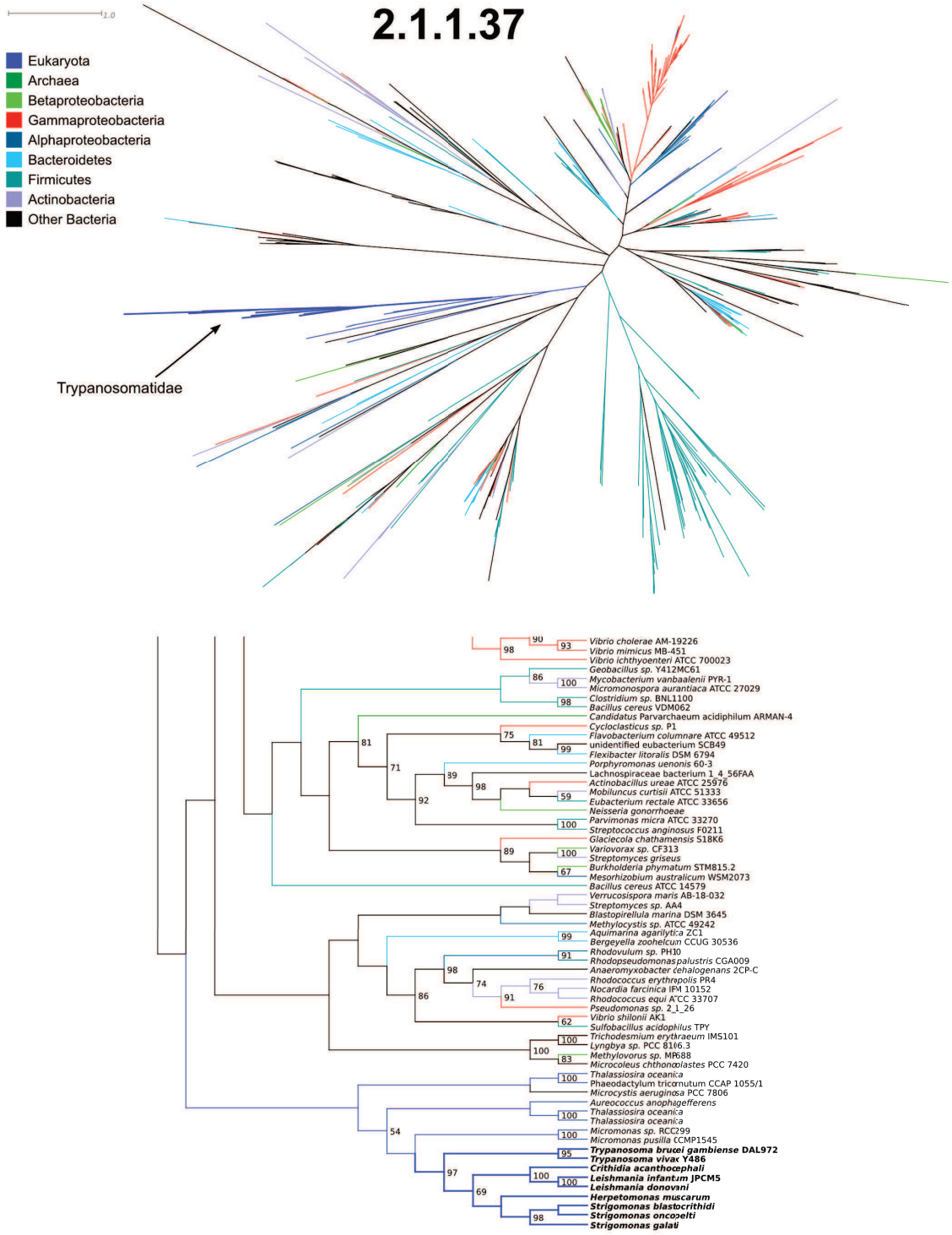


Figure C.9: **Maximum likelihood phylogeny of DNA (cytosine-5-)-methyltransferase (EC:2.1.1.37).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

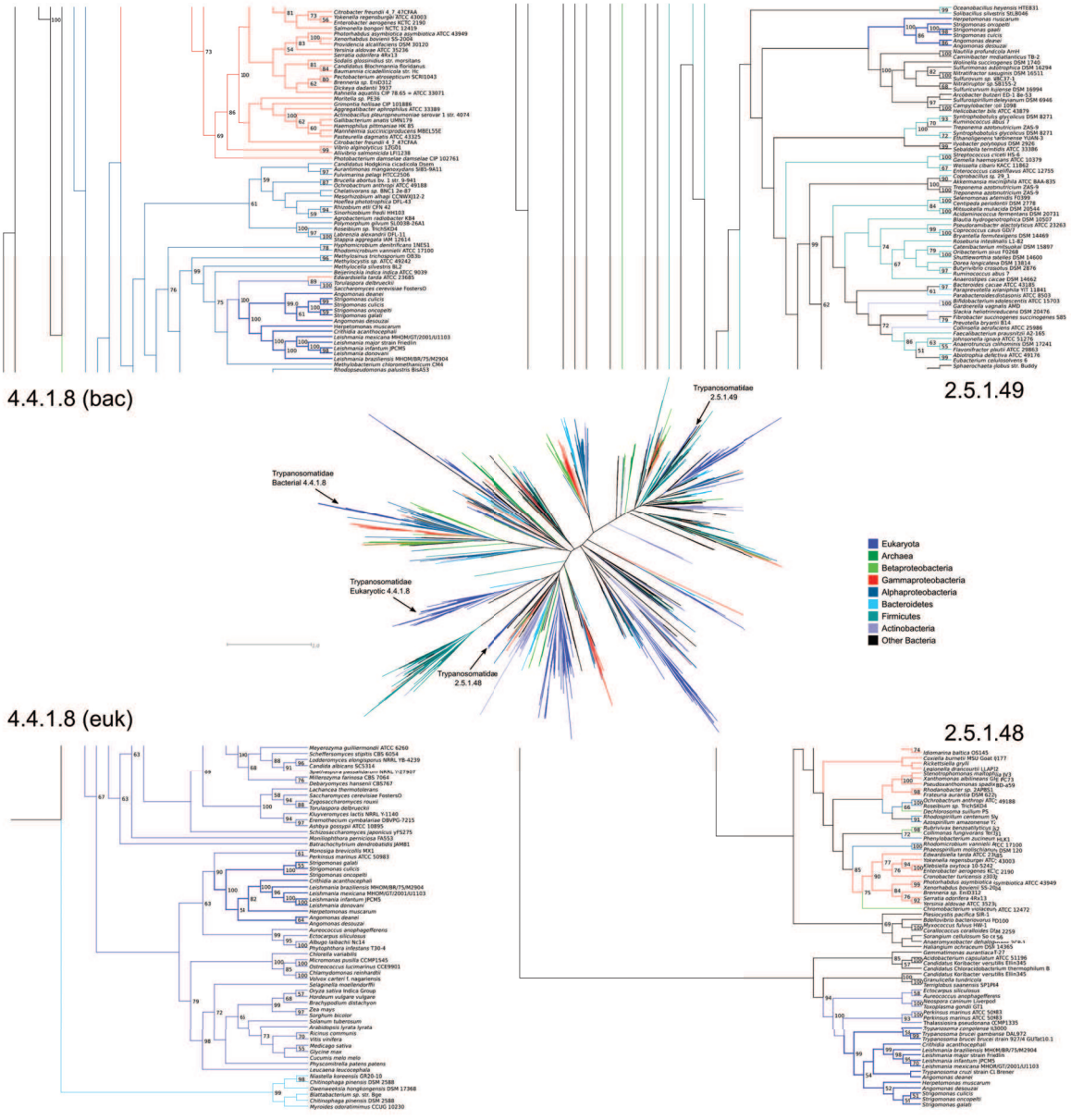


Figure C.10: **Maximum likelihood phylogeny of cystathionine gamma-synthase, O-acetylhomoserine aminocarboxypropyltransferase, and cystathionine beta-lyase (EC:2.5.1.48, EC:2.5.1.49, and EC:4.4.1.8).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

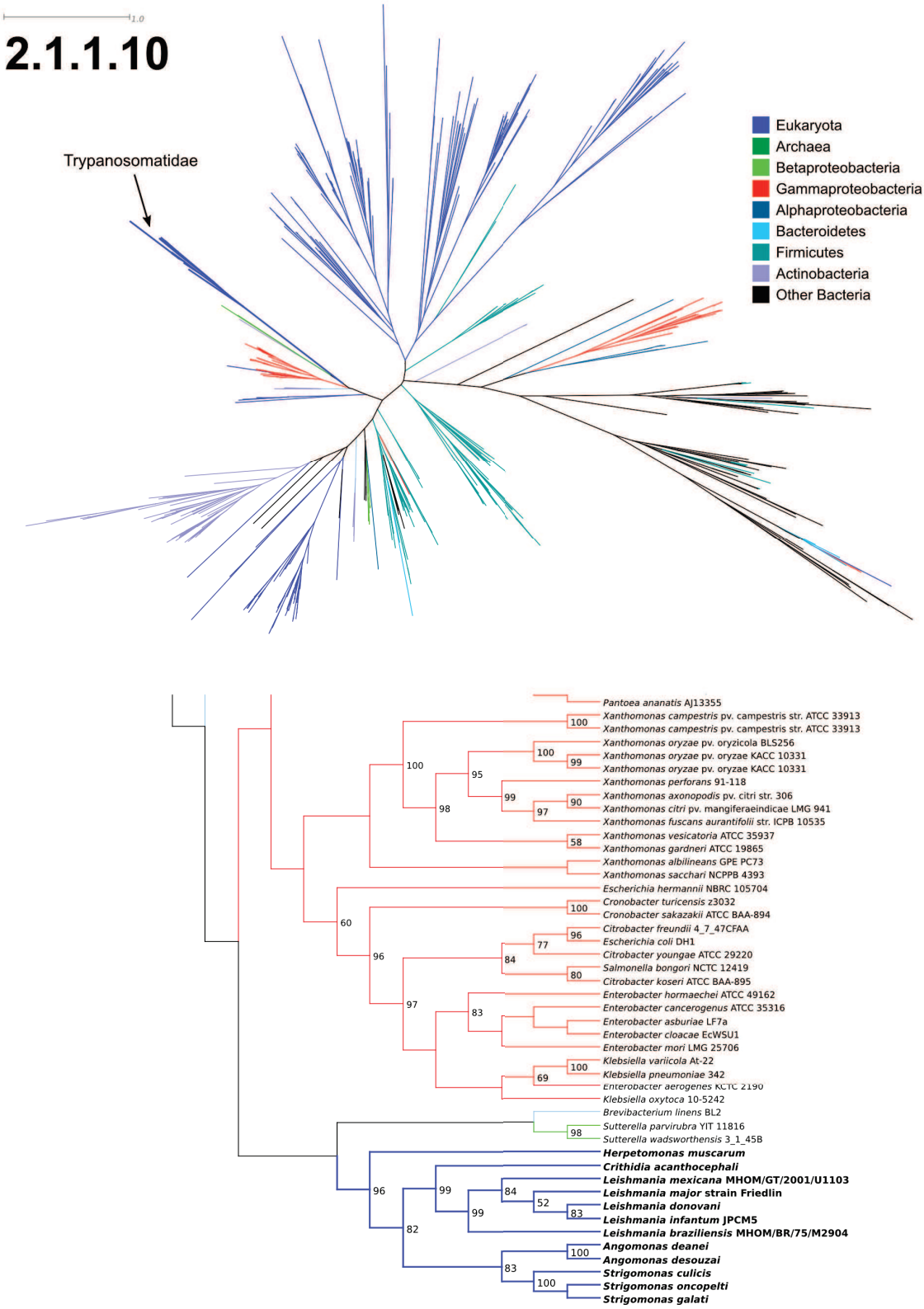


Figure C.11: **Maximum likelihood phylogeny of homocysteine S-methyltransferase (EC:2.1.1.10).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

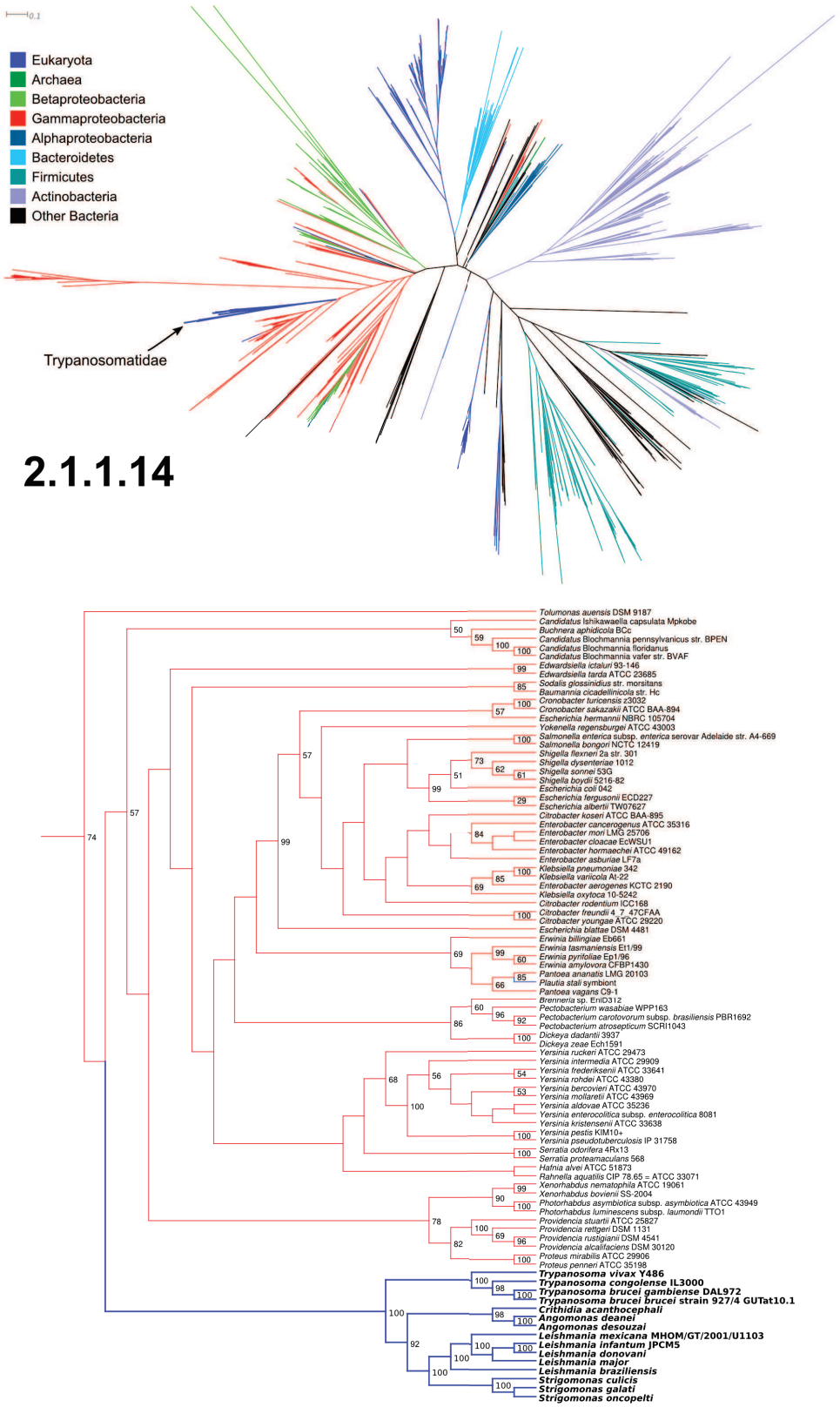


Figure C.12: **Maximum likelihood phylogeny of 5-methyltetrahydropteroyltriglutamate-homocysteine S-methyltransferase (EC:2.1.1.14).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

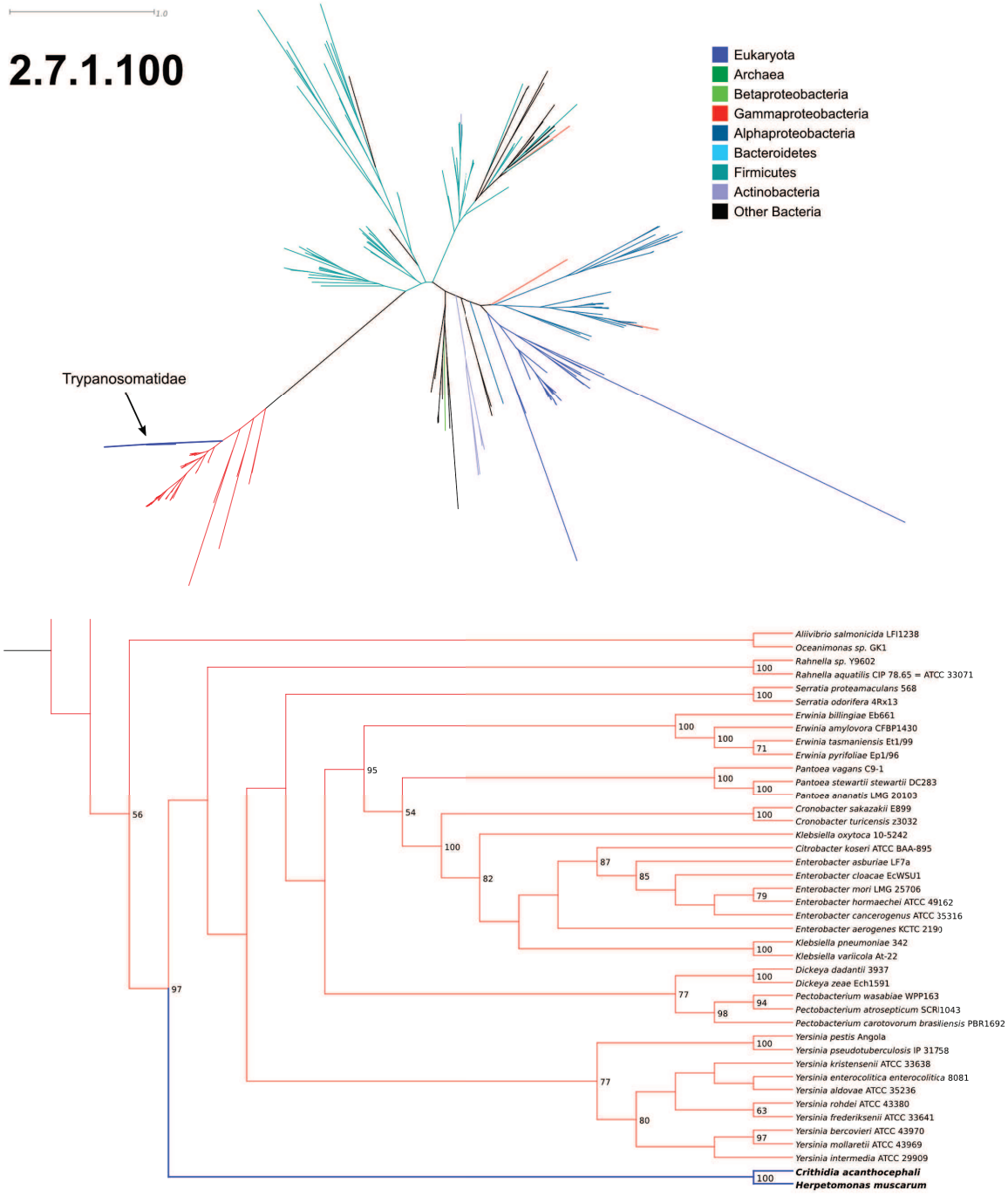


Figure C.13: **Maximum likelihood phylogeny of S-methyl-5-thioribose kinase (EC:2.7.1.100).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

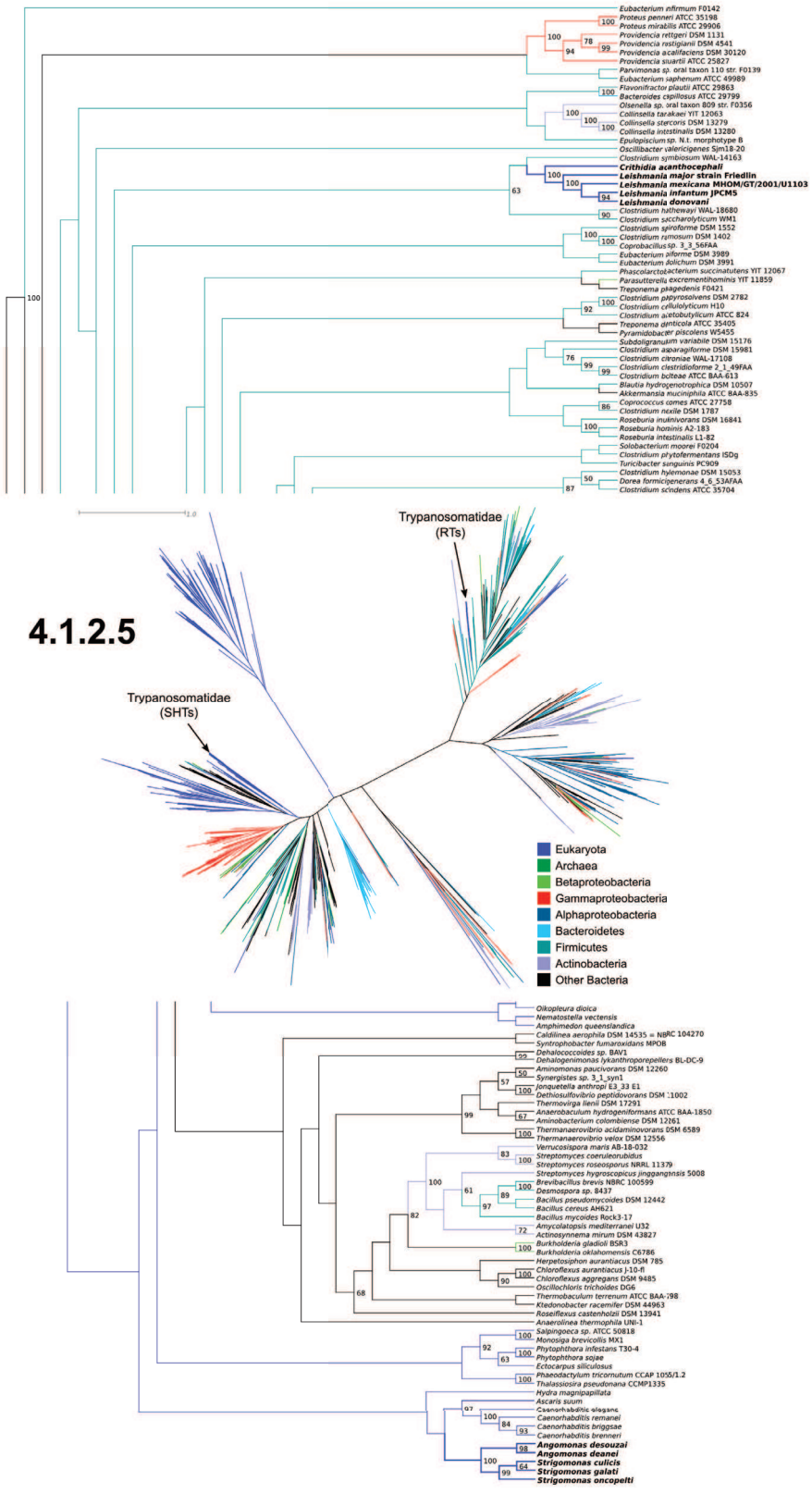


Figure C.14: **Maximum likelihood phylogeny of L-threonine aldolase (EC:4.1.2.5).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

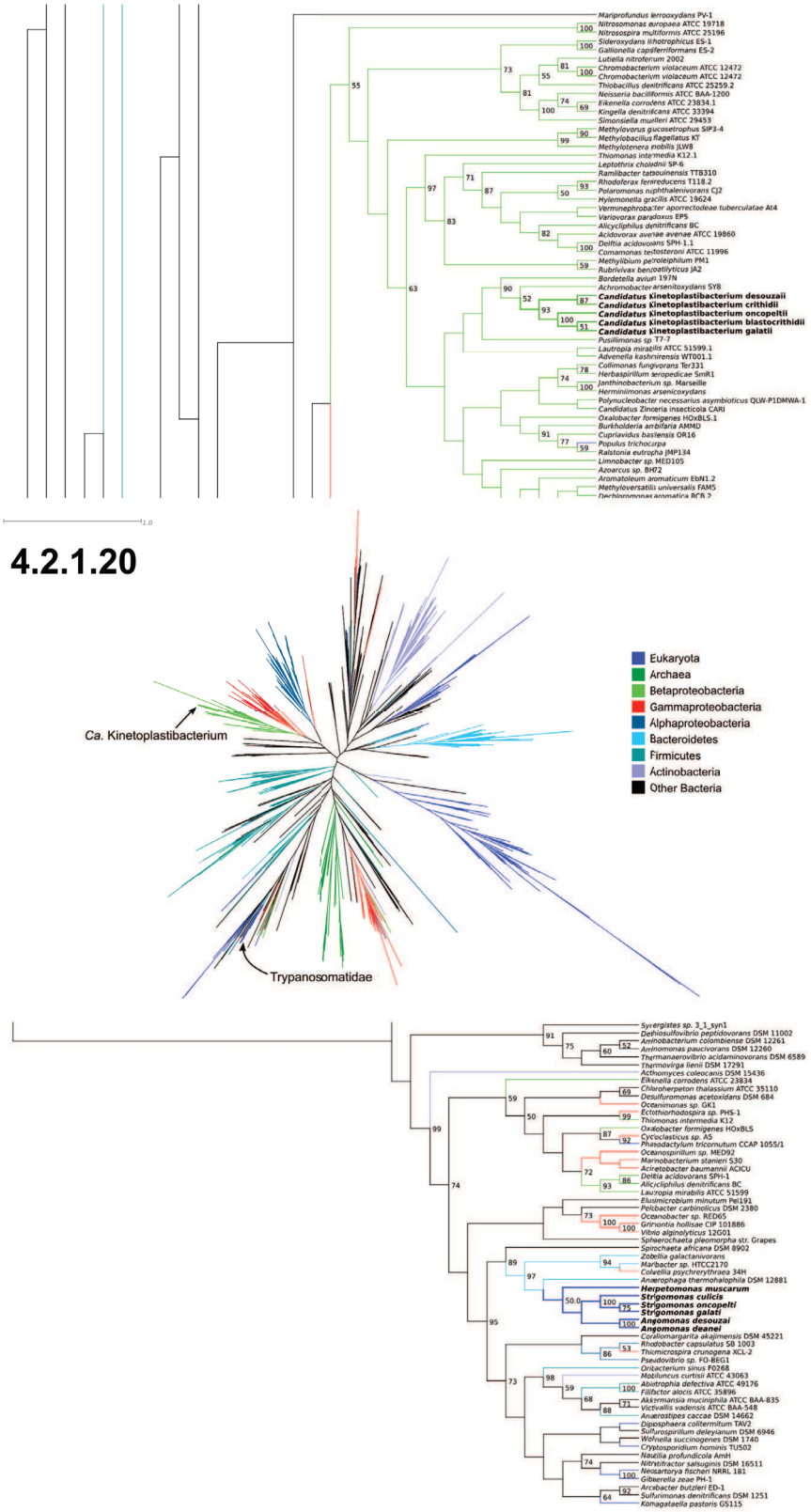


Figure C.15: **Maximum likelihood phylogeny of (EC:4.2.1.20).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

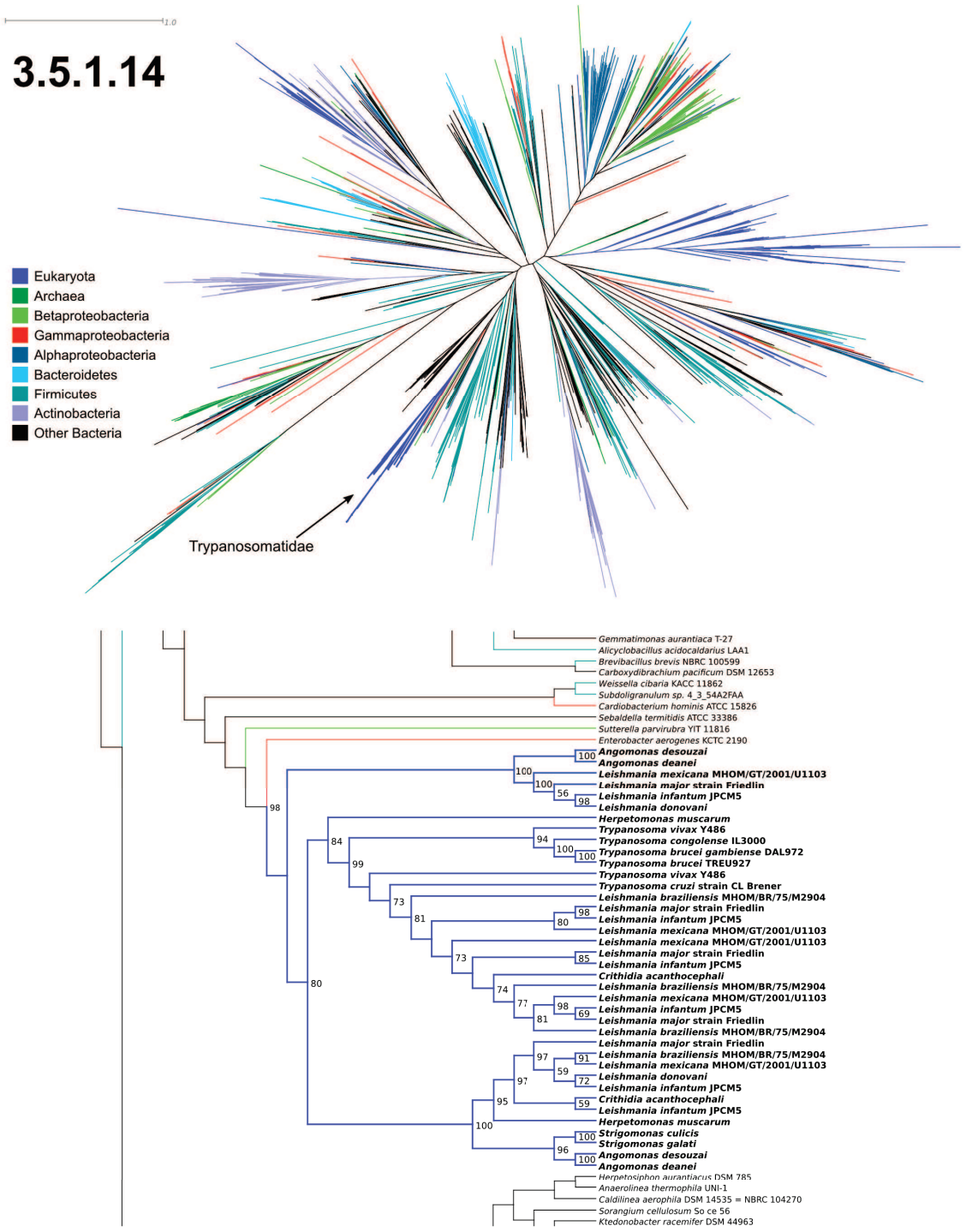


Figure C.16: **Maximum likelihood phylogeny of aminoacylase (EC:3.5.1.14).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

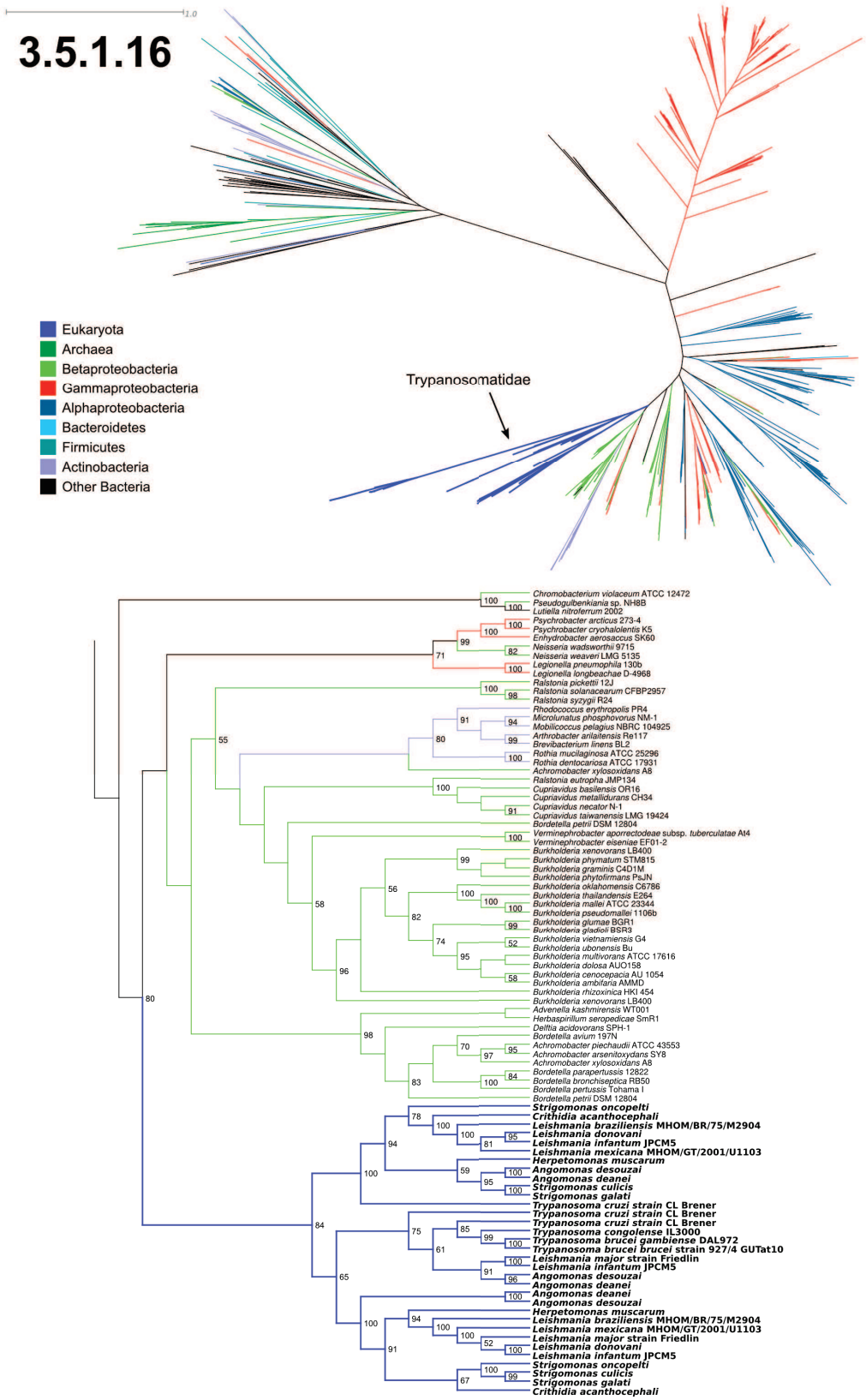


Figure C.17: **Maximum likelihood phylogeny of acetylornithine deacetylase (EC:3.5.1.16).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

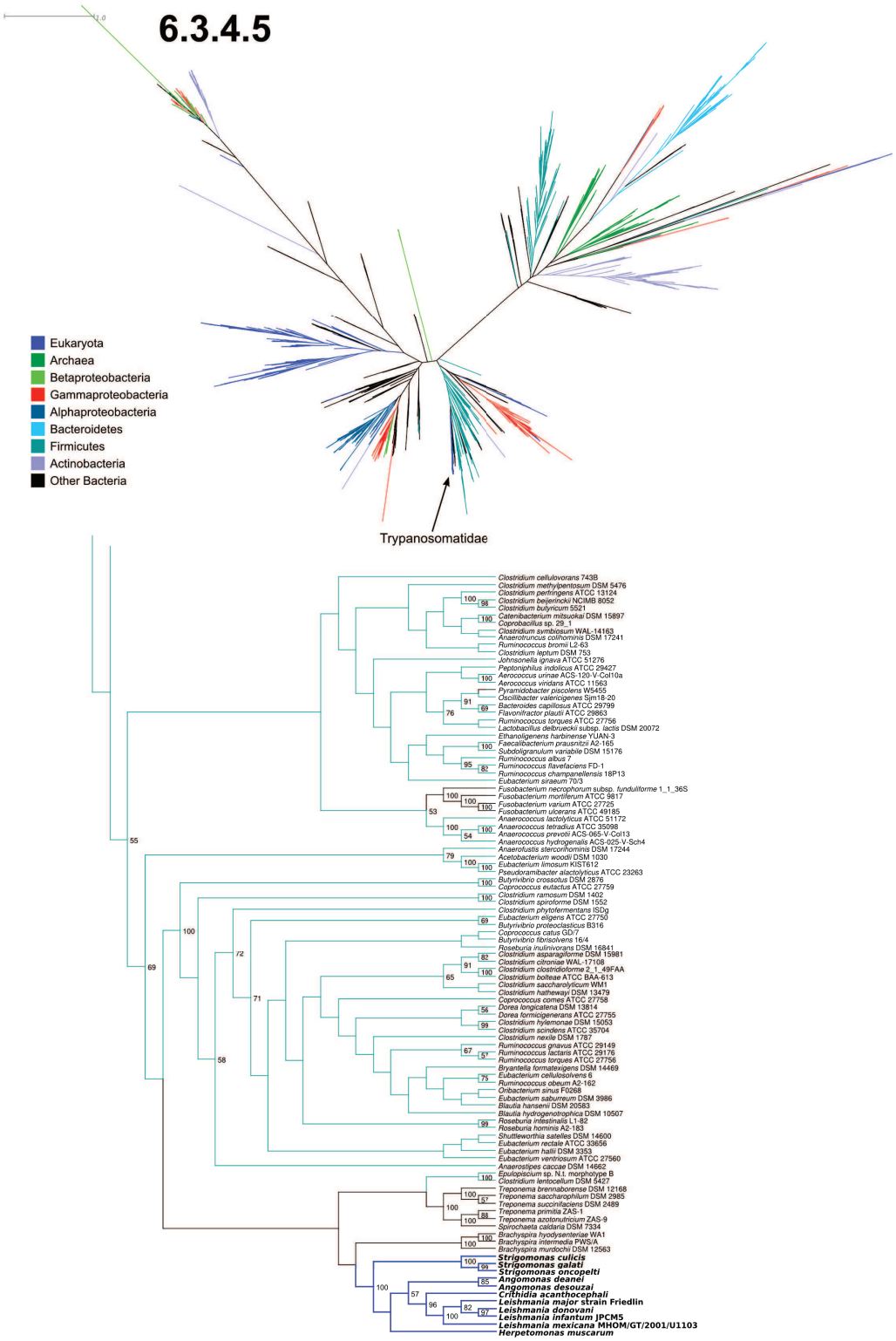


Figure C.18: **Maximum likelihood phylogeny of argininosuccinate synthase (EC:6.3.4.5).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

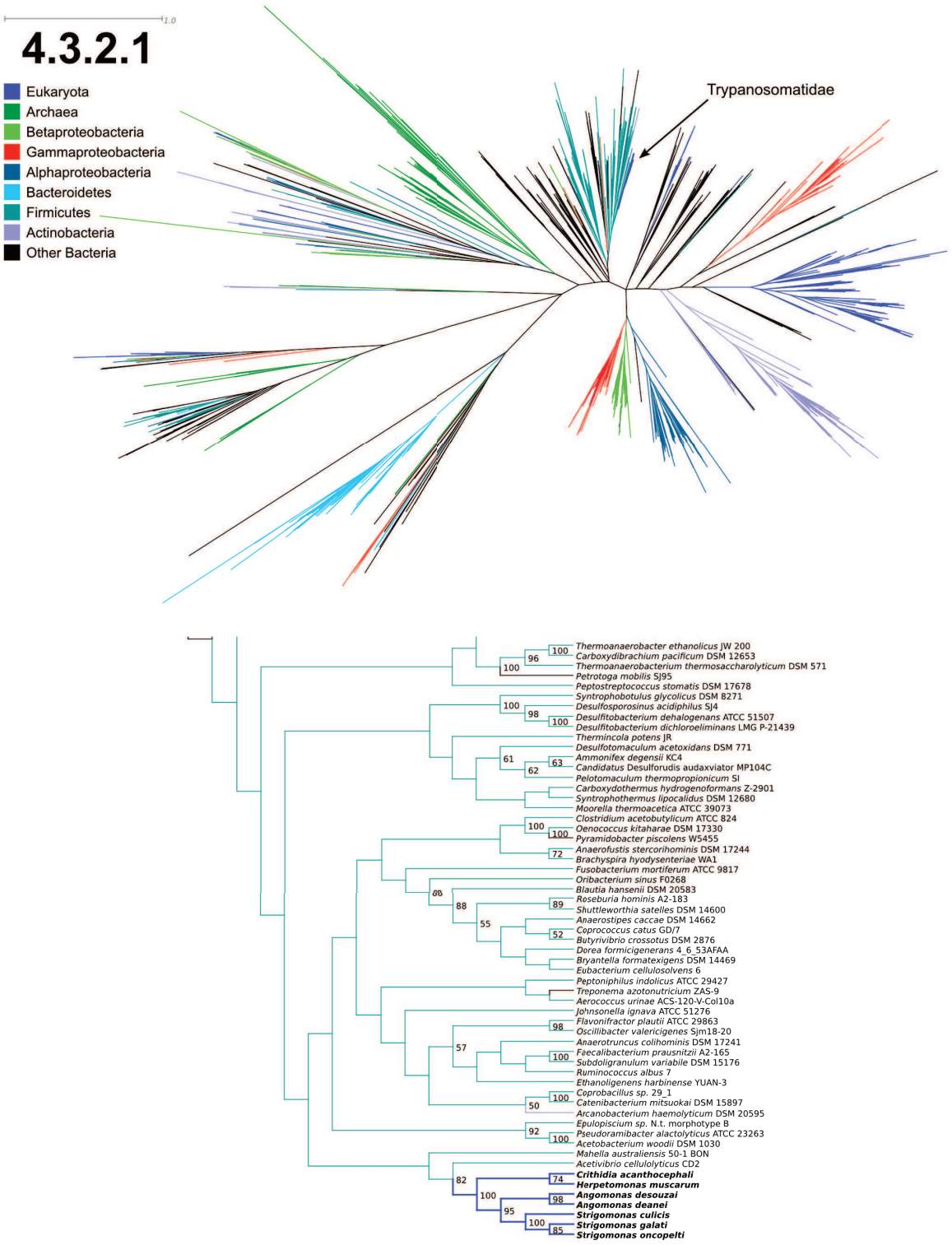


Figure C.19: **Maximum likelihood phylogeny of argininosuccinate lyase (EC:4.3.2.1).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

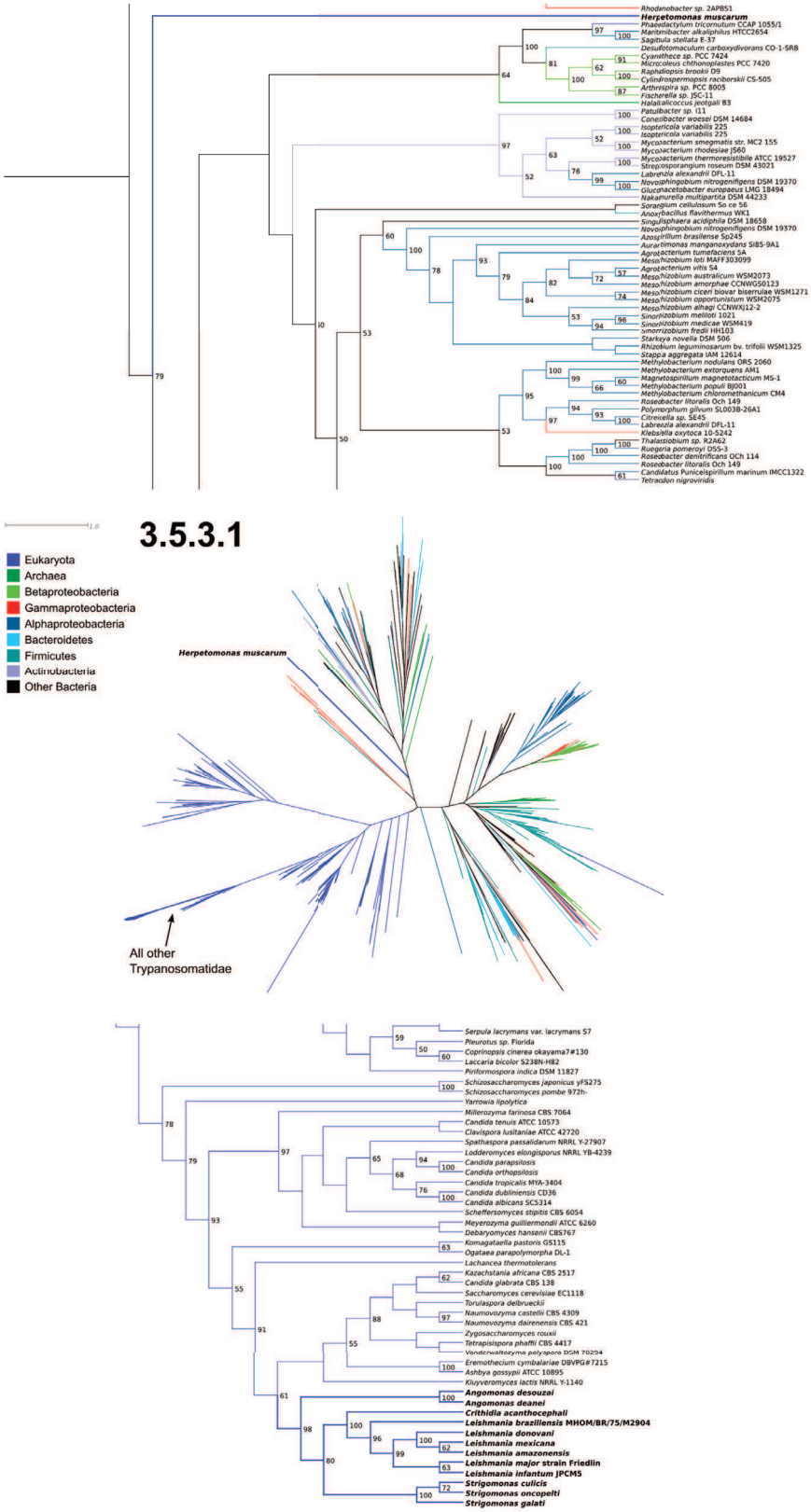


Figure C.20: **Maximum likelihood phylogeny of arginase (EC:3.5.3.1).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

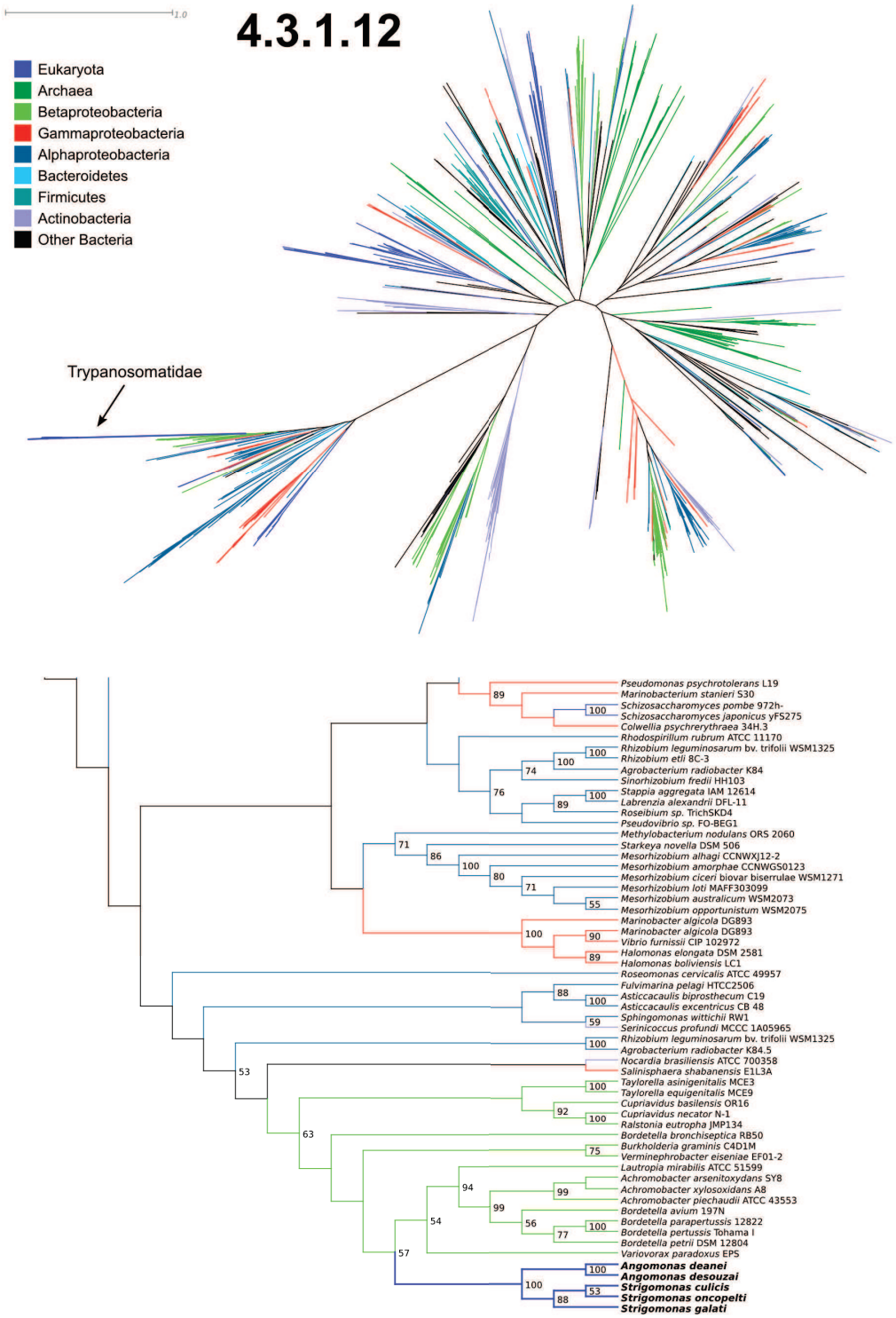


Figure C.21: **Maximum likelihood phylogeny of ornithine cyclodeaminase (EC:4.3.1.12).** Overall tree colored according to taxonomic affiliation of sequences. Values on nodes represent bootstrap support (only 50 or greater shown) and distance bar only applies to the overall tree and not to the detailed regions.

EC number	Accession Numbers						
	Strigomonas culicis	*Strigomonas galati*	*Strigomonas oncopelti*	*Angomonas deanei*	*Angomonas desouzai*	*Herpetomonas muscarum*	*Crithidia acanthocephali*
1.1.1.3	KC476503	KC545207	KC545098		KC584075	KC503395	KC545151
1.13.11.54	KC140182	KC545206	KC545090	KC503336	KC584069	KC503392	KC545145
1.14.16.1	KC476510	KC545215	KC545100	KC503346	KC584077	KC503406	KC545155
1.2.1.41	KC140162	KC545185	KC545068	KC503312	KC584044	KC503371	KC545121
1.2.4.1	KC476514, KC476515	KC545219, KC545220	KC545107, KC545106	KC503348, KC503349	KC584081, KC584082	KC503407, KC503408	KC545159, KC545160
1.4.1.2	KC005719	KC545170	KC545056	KC503303	KC584034	KC503356	KC545114
1.4.1.4	KC005720	KC545171	KC545057	KC503296	KC584035	KC503357, KC503358	KC545110
1.5.1.12	KC140160	KC545183	KC545058	KC503297	KC584036	KC503369	KC545125
1.5.1.2	KC140159	KC545182	KC545070	KC503311	KC584049	KC503368	KC545124
1.5.99.8	KC140163	KC545186	KC545071	KC503315	KC584050	KC503372	KC545127
2.1.1.10	KC140174	KC545198	KC545082	KC503324	KC584061	KC503384	KC545139
2.1.1.13	KC140175	KC545199	KC545084	KC503325	KC584052	KC503385	KC545140
2.1.1.14	KC140173	KC545197	KC545083	KC503326	KC584062		KC545132
2.1.1.37	KC140164	KC545188	KC545073			KC503375	KC545129
2.1.2.1	KC476505, KC476506	KC545209, KC545210	KC545094, KC545095	KC503343	KC584071	KC503396, KC503397	KC545148, KC545149
2.1.3.3						KC503373	
2.3.1.30	KC140166	KC545189	KC545074	KC503319	KC584053	KC503376	KC545130
2.3.1.46	KC140171	KC545194	KC545079	KC503323	KC584057	KC503381	
2.4.2.28	KC140179	KC545203	KC545087	KC503332	KC584066	KC503389	KC545142
2.5.1.16	KC140178	KC545202		KC503331	KC584065	KC503388	KC545134
2.5.1.47	KC140168	KC545191	KC545077	KC503320	KC584055	KC503378	KC545136
2.5.1.48	KC140170	KC545193	KC545078	KC503321	KC584056	KC503380	KC545131
2.5.1.49	KC140169	KC545192	KC545076	KC503322	KC584054	KC503379	
2.5.1.6	KC140176	KC545200	KC545085	KC503335	KC584063	KC503386	KC545133
2.6.1.1	KC476511, KC476512	KC545216, KC545217	KC545053, KC545054	KC503299, KC503300	KC584033, KC584032	KC503403, KC503404	KC545156, KC545157
2.6.1.11	KC140151	KC545174	KC545060		KC584038		
2.6.1.42	KC476516, KC476517	KC545221, KC545222	KC545104, KC545105	KC503347	KC584080	KC503409	KC545161, KC545162
2.6.1.2	KC005716	KC545167	KC545047			KC503353	
2.6.1.5	KC476513	KC545218	KC545101, KC54510, KC545103	KC503327, KC503328, KC503329, KC503330	KC584078	KC503405	KC545158
2.7.1.100						KC503394	KC545146
2.7.1.39	KC476509	KC545213	KC545097	KC503342	KC584074	KC503398	KC545153
2.7.2.11	KC140161	KC545184	KC545065, KC545066, KC545067	KC503305	KC584045, KC584046, KC584047	KC503370	KC545126
3.1.3.77						KC503393	KC545147
3.3.1.1	KC140165	KC545187	KC545072	KC503316	KC584051	KC503374	KC545128
3.5.1.1	KC005718	KC545169		KC503298	KC584031	KC503355	KC545113
3.5.1.14	KC140152	KC545175	KC545064	KC503306, KC503307	KC584039	KC503361, KC503362	KC545116, KC545117
3.5.1.16	KC140153, KC140154	KC545176, KC545177	KC545061, KC545062	KC503308, KC503309, KC503310	KC584040, KC584041, KC584042	KC503363, KC503364	KC545118, KC545119
3.5.3.1	KC140157	KC545180	KC545063	KC503313	KC584043	KC503367	KC545123
4.1.1.20	KC476502	KC545214	KC545099	KC503345	KC584076	KC503401	KC545154
4.1.1.50	KC140177	KC545201	KC545086	KC503337	KC584064	KC503387	KC545141
4.1.2.5	KC476507	KC545211	KC545091	KC503339	KC584072		KC545150
4.2.1.109	KC140181	KC545205	KC545089	KC503334	KC584068	KC503391	KC545144
4.2.1.20	KC476504	KC545208	KC545093	KC503340	KC584070	KC503399	
4.2.1.22	KC140167	KC545190	KC545075	KC503338	KC584060	KC503377	KC545135
4.2.3.1	KC476508	KC545212	KC545096	KC503341	KC584073	KC503400	KC545152
4.3.1.1	KC005715	KC545166	KC545048			KC503352	
4.3.1.12	KC140158	KC545181	KC545069	KC503314	KC584048		
4.3.1.19	KC476518	KC545223	KC545092	KC503344	KC584079	KC503410	KC545163
4.3.2.1	KC140155	KC545178	KC545050	KC503302	KC584029	KC503365	KC545122
4.3.2.2	KC005713	KC545164	KC545049	KC503293		KC503350	KC545108
4.4.1.8	KC140172	KC545195, KC545196	KC545080, KC545081	KC503317, KC503318	KC584058, KC584059	KC503382, KC503383	KC545137, KC545138
5.1.1.7						KC503402	
5.3.1.23	KC140180	KC545204	KC545088	KC503333	KC584067	KC503390	KC545143
6.3.1.1	KC005717	KC545168	KC545055	KC503295	KC584030	KC503354	KC545112
6.3.1.2	KC005721	KC545172		KC503304	KC584037	KC503359	KC545111
6.3.4.4	KC005714	KC545165	KC545052	KC503292	KC584026	KC503351	KC545109
6.3.4.5	KC140156	KC545179	KC545051	KC503301	KC584027	KC503366	KC545120
6.3.5.5	KC005722	KC545173	KC545059	KC503294	KC584028	KC503360	KC545115

Figure C.22: **Genbank accession numbers for Trypanosomatidae genes characterized in this study.**

EC number*	*Ca.* K. blastocrithidii	*Ca.* K. galatii	*Ca.* K. oncopeltii	*Ca.* K. crithidii	*Ca.* K. desouzaii	Gene names / notes
1.1.1.23	BCUe_0232	ST1e_0252	CONe_0227	CDEe_0237	CDSe_0230	hisD
1.1.1.25	BCUe_0042	ST1e_0050	CONe_0048	CDEe_0053	CDSe_0049	aroE
1.1.1.3	BCUe_0615	ST1e_0695	CONe_0599	CDEe_0640	CDSe_0625	thrA metL
1.1.1.85	BCUe_0453	ST1e_0503	CONe_0438	CDEe_0465	CDSe_0457	leuB
1.1.1.86	BCUe_0788	ST1e_0892	CONe_0770	CDEe_0817	CDSe_0803	ilvC
1.2.1.11	BCUe_0454	ST1e_0504	CONe_0439	CDEe_0466	CDSe_0458	asd
1.2.1.38	BCUe_0027	ST1e_0034	CONe_0032	CDEe_0037	CDSe_0033	argC
1.2.4.1	BCUe_0535	ST1e_0604	CONe_0525	CDEe_0553	CDSe_0547	aceE
1.3.1.13	BCUe_0689	ST1e_0730	CONe_0670	CDEe_0713	CDSe_0699	tyrA
1.3.1.26	BCUe_0811	ST1e_0920	CONe_0794	CDEe_0840	CDSe_0829	dapB
2.1.2.1	BCUe_0023	ST1e_0029	CONe_0028	CDEe_0031	CDSe_0028	glyA
2.1.3.3	BCUe_0601	ST1e_0676	CONe_0586	CDEe_0623	CDSe_0612	argF
2.2.1.6	BCUe_0789	ST1e_0893	CONe_0771	CDEe_0818	CDSe_0804	ilvH small (regulatory) subunit
2.2.1.6	BCUe_0790	ST1e_0895	CONe_0772	CDEe_0819	CDSe_0805	ilvB large (catalytic) subunit
2.3.1.1 2.3.1.35	BCUe_0289	ST1e_0312	CONe_0281	CDEe_0289	CDSe_0282	argJ
2.3.1.1	BCUe_0426	ST1e_0469	NF**	CDEe_0432	CDSe_0430	argB/A
2.3.1.117	BCUe_0472	ST1e_0523	CONe_0456	CDEe_0485	CDSe_0478	dapD
2.3.3.13	BCUe_0218	ST1e_0237	CONe_0211	CDEe_0220	CDSe_0215	leuA
2.4.2.-	BCUe_0229	ST1e_0249	CONe_0224	CDEe_0233	CDSe_0227	hisH
2.4.2.17	BCUe_0233	ST1e_0253	CONe_0228	CDEe_0238	CDSe_0231	hisG
2.4.2.18	BCUe_0250	ST1e_0273	CONe_0250	CDEe_0256	CDSe_0251	trpD
2.5.1.19	BCUe_0688	ST1e_0769	CONe_0669	CDEe_0712	CDSe_0698	aroA
2.5.1.54	BCUe_0609	ST1e_0688	CONe_0593	CDEe_0631	CDSe_0620	aroG
2.5.1.6	BCUe_0116	ST1e_0127	CONe_0113	CDEe_0125	CDSe_0118	metK
2.5.1.57	BCUe_0479	ST1e_0535	CONe_0466	CDEe_0492	CDSe_0486	tyrB
2.6.1.11	BCUe_0215	ST1e_0233	CONe_0208	CDEe_0217	CDSe_0211	argD
2.6.1.17	BCUe_0473	ST1e_0526	CONe_0458	CDEe_0486	CDSe_0480	dapC
2.6.1.57	BCUe_0479	ST1e_0535	CONe_0466	CDEe_0492	CDSe_0486	tyrB
2.6.1.9	BCUe_0231	ST1e_0251	CONe_0226	CDEe_0235	CDSe_0229	hisC
2.6.1.9	BCUe_0690	ST1e_0771	CONe_0671	CDEe_0714	CDSe_0700	hisC
2.7.1.71	BCUe_0145	ST1e_0153	CONe_0137	CDEe_0147	CDSe_0145	aroK
2.7.2.4	BCUe_0398	ST1e_0441	CONe_0384	CDEe_0401	CDSe_0397	lysC
2.7.2.8	BCUe_0090	ST1e_0099	CONe_0088	CDEe_0096	CDSe_0095	argB
3.5.1.18	BCUe_0471	ST1e_0522	CONe_0455	CDEe_0484	CDSe_0476	dapE
3.5.4.19	BCUe_0226	ST1e_0245	CONe_0221	CDEe_0229	CDSe_0224	hisI
3.6.1.31	BCUe_0225	ST1e_0244	CONe_0220	CDEe_0228	CDSe_0223	hisE
4.1.1.20	BCUe_0149	ST1e_0158	CONe_0143	CDEe_0152	CDSe_0149	lysA
4.1.1.48	BCUe_0249	ST1e_0272	CONe_0249	CDEe_0255	CDSe_0250	trpC
4.1.3.-	BCUe_0227	ST1e_0246	CONe_0222	CDEe_0230	CDSe_0225	hisF
4.1.3.27	BCUe_0251	ST1e_0274	CONe_0251	CDEe_0257	CDSe_0252	trpG component II
4.1.3.27	BCUe_0253	ST1e_0275	CONe_0252	CDEe_0258	CDSe_0253	trpE component I
4.2.1.10	BCUe_0040	ST1e_0048	CONe_0045	CDEe_0051	CDSe_0046	aroQ
4.2.1.19	BCUe_0230	ST1e_0250	CONe_0225	CDEe_0234	CDSe_0228	hisB
4.2.1.20	BCUe_0742	ST1e_0840	CONe_0720	CDEe_0768	CDSe_0755	alpha subunit
4.2.1.20	BCUe_0743	ST1e_0841	CONe_0722	CDEe_0770	CDSe_0756	beta subunit
4.2.1.33 4.2.1.35	BCUe_0451	ST1e_0501	CONe_0436	CDEe_0462	CDSe_0455	leuC large subunit
4.2.1.33 4.2.1.35	BCUe_0452	ST1e_0502	CONe_0437	CDEe_0463	CDSe_0456	leuD small subunit
4.3.3.7	BCUe_0380	ST1e_0422	CONe_0371	CDEe_0382	CDSe_0379	dapA
4.2.1.9	BCUe_0298	ST1e_0326	CONe_0291	CDEe_0297	CDSe_0291	ilvD
4.2.3.4	BCUe_0144	ST1e_0152	CONe_0136	CDEe_0146	CDSe_0143	aroB
4.2.3.5	BCUe_0450	ST1e_0499	CONe_0434	CDEe_0461	CDSe_0454	aroC
4.3.1.19	BCUe_0074	ST1e_0080	CONe_0076	CDEe_0080	CDSe_0079	ilvA two C-terminal threonine dehydratase domains
4.3.1.19	BCUe_0303	ST1e_0333	CONe_0296	CDEe_0304	CDSe_0297	ilvA
4.3.2.2	BCUe_0606	ST1e_0681	CONe_0590	CDEe_0628	CDSe_0618	purB
5.1.1.7	BCUe_0118	ST1e_0130	CONe_0116	CDEe_0126	CDSe_0120	dapF
5.3.1.16	BCUe_0228	ST1e_0248	CONe_0223	CDEe_0231	CDSe_0226	hisA
5.3.1.24	BCUe_0456	ST1e_0507	CONe_0441	CDEe_0468	CDSe_0461	trpF
5.4.99.5 4.2.1.51	BCUe_0691	ST1e_0772	CONe_0672	CDEe_0715	CDSe_0701	pheA
6.3.4.4	BCUe_0635	ST1e_0718	CONe_0615	CDEe_0658	CDSe_0644	purA
6.3.5.5	BCUe_0369	ST1e_0409	CONe_0363	CDEe_0369	CDSe_0366	carA small subunit
6.3.5.5	BCUe_0370	ST1e_0411	CONe_0364	CDEe_0370	CDSe_0367	carB large subunit

Figure C.23: **Locus tags for the *Ca.* Kinetoplastibacterium genes analyzed in this study.** * two EC numbers in one row indicate putative bifunctional enzymes. NF**: enzyme present in the genome as a putative pseudogene.

Appendix D

Published article: Telling metabolic stories to explore metabolomics data: A case study on the Yeast response to cadmium exposure

BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 1 2014, pages 61–70
doi:10.1093/bioinformatics/btt597

Structural bioinformatics

Advance Access publication October 27, 2013

Telling metabolic stories to explore metabolomics data: a case study on the yeast response to cadmium exposure

Paulo Vieira Milreu[1,2,*], Cecilia Coimbra Klein[1,2,3], Ludovic Cottret[4], Vicente Acuña[1,2,5], Etienne Birmelé[1,2,6], Michele Borassi[7], Christophe Junot[8], Alberto Marchetti-Spaccamela[9], Andrea Marino[10], Leen Stougie[11], Fabien Jourdan[12], Pierluigi Crescenzi[10], Vincent Lacroix[1,2,*] and Marie-France Sagot[1,2,*]

[1]INRIA Grenoble Rhône-Alpes & Université de Lyon, F-69000 Lyon, [2]Université Lyon 1; CNRS, UMR5558 LBBE, France, [3]Laboratório Nacional de Computação Científica (LNCC), Petrópolis, Brazil, [4]LISBP, UMR CNRS 5504 - INRA 792, Toulouse, France, [5]Mathomics, Center for Mathematical Modeling (UMI-2807 CNRS) and Center for Genome Regulation (Fondap 15090007), University of Chile, Santiago, Chile [6]Lab. Statistique et Génome, CNRS UMR8071 INRA1152, Université d'Évry, France, [7]Scuola Normale Superiore, 56126 Pisa, Italy, [8]Laboratoire d'Etude du Métabolisme des Médicaments, DSV/iBiTecS/SPI, CEA/Saclay, 91191 Gif-sur-Yvette, France, [9]La Sapienza University of Rome, Rome, [10]Dipartimento di Sistemi e Informatica, Università di Firenze, I-50134 Firenze, Italy, [11]VU University and CWI, Amsterdam, The Netherlands and [12]INRA UMR1331 - Toxalim, Toulouse, France

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: The increasing availability of metabolomics data enables to better understand the metabolic processes involved in the immediate response of an organism to environmental changes and stress. The data usually come in the form of a list of metabolites whose concentrations significantly changed under some conditions, and are thus not easy to interpret without being able to precisely visualize how such metabolites are interconnected.

Results: We present a method that enables to organize the data from any metabolomics experiment into metabolic stories. Each story corresponds to a possible scenario explaining the flow of matter between the metabolites of interest. These scenarios may then be ranked in different ways depending on which interpretation one wishes to emphasize for the causal link between two affected metabolites: enzyme activation, enzyme inhibition or domino effect on the concentration changes of substrates and products. Equally probable stories under any selected ranking scheme can be further grouped into a single anthology that summarizes, in a unique subnetwork, all equivalently plausible alternative stories. An anthology is simply a union of such stories. We detail an application of the method to the response of yeast to cadmium exposure. We use this system as a proof of concept for our method, and we show that we are able to find a story that reproduces very well the current knowledge about the yeast response to cadmium. We further show that this response is mostly based on enzyme activation. We also provide a framework for exploring the alternative pathways or side effects this local response is expected to have in the rest of the network. We discuss several interpretations for the changes we see, and we suggest hypotheses that could in principle be experimentally tested. Noticeably, our method requires simple input data and could be used in a wide variety of applications.

Availability and implementation: The code for the method presented in this article is available at http://gobbolino.gforge.inria.fr.

*To whom correspondence should be addressed.

Contact: pvmilreu@gmail.com; vincent.lacroix@univ-lyon1.fr; marie-france.sagot@inria.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 4, 2013; revised on September 19, 2013; accepted on October 14, 2013

1 INTRODUCTION

One of the main goals of metabolic studies is to understand the metabolic processes involved in the adaptation to an environmental change. Recently, metabolomic techniques gained the spotlight by providing a way to monitor metabolism by measuring the concentration of metabolites in different conditions or at different time points. A typical result from such an experiment is a list of metabolites whose concentrations significantly changed when the cell or organism was exposed to some stress. How to interpret this list became then a new research topic, consisting in identifying the metabolic processes that link the metabolites of interest, possibly explaining the observed variations in their concentrations. This topic goes in the literature by the name of 'metabolite set enrichment analysis', and is an extension to metabolism of work that was initiated in the context of transcriptomics and then proteomics under the name of 'gene set enrichment analysis' [see (Subramanian *et al.*, 2005) for what is possibly the first work on this and (Khatri *et al.*, 2012) for a recent survey]. The simplest idea one may think of is to highlight the set of metabolites identified in the experiment that have significantly changed their concentration, let us call them discriminating compounds, and then to visually analyze their interconnections. This is what is done notably in Xia *et al.* (2012). However, like a number of other approaches on metabolite set enrichment analysis, the projection of enriched metabolites is done on pathways instead of the whole network, thereby missing

(alternative) pathways not annotated in current databases, or more generally paths traversing several pathways.

For genome-scale networks, the metabolism of a whole organism is considered, which may be large (Thiele and Palsson, 2010), whereas a metabolic perturbation caused by some stress condition may impact only a small portion of this complex network. Even if it is sometimes possible to visually identify the pathways that explain some of the variations in the monitored metabolites, getting an overall explanation for all the observed variations usually cannot be performed by visual inspection.

Recently, automatic methods have been proposed to deal with this kind of data (Antonov *et al*., 2009; Dittrich *et al*., 2008; Faust *et al*., 2010; Leader *et al*., 2011). A natural idea is to try to link all discriminating compounds through chains of reactions. One possible model for this is by means of a Steiner tree, which is a minimum cost tree that connects all nodes belonging to a predefined subset called *terminals*, which in the case of metabolism would be the discriminating compounds (Dittrich *et al*., 2008; Scott *et al*., 2005). However, any pair of metabolites may be connected through several alternative paths within a network, and each of these paths may validly explain the observed changes of concentration. In this context, the extraction of subgraphs appears to be more relevant than the extraction of subtrees. The number of alternative paths between two metabolites may, however, be large and restricting the search to all the shortest or lightest (the weight is given by the sum of the out-degrees of the vertices in the path) paths between pairs of metabolites seems to be a realistic compromise.

This is the approach followed by (Croes *et al*., 2006; Faust *et al*., 2010; van Helden *et al*., 2002) where the authors concentrate on a pair of discriminating compounds and search for subgraphs corresponding to source-to-sink paths between them. In Antonov *et al*. (2009), this approach is pushed one step further as the authors consider all pairs of metabolites and unify all the shortest paths, this time with a maximum length k. In practice, this may lead to large networks (if k is too big) or to disconnected ones (if k is too small).

The aforementioned methods are based only on the topology of the network, but one could consider different approaches based on flux distributions over the set of reactions, such as elementary modes (Schuster and Hilgetag, 1994; Schuster *et al*., 1999) that are minimal subnetworks working at steady state. One difficulty in this case is that flux-based models need stoichiometric values as well as a definition of the boundaries of the system under analysis, which are not always simple to identify, particularly in the case of a metabolomics experiment in which the list of discriminating compounds does not directly define such boundaries. Moreover, flux approaches are focused on reactions and are not designed to take into account endogenous metabolite concentrations. The very same metabolites may play different roles in different metabolic processes, being source in one, intermediate in a second and target in a third one. The inability and the unwillingness to tell, *a priori*, the role of the discriminating compounds in each scenario to be proposed is a key factor of our approach: we are interested not only in connecting the discriminating compounds but also in establishing their individual role for each scenario.

Our approach is a subgraph extraction technique in which we want to find maximal directed acyclic subgraphs (DAGs) whose set of sources and targets are discriminating compounds. We call such subgraphs metabolic stories, or for short, simply stories. In practice, for a given set of discriminating compounds, the number of stories may be large. Because we do not have a clear criterion for choosing which of these stories is the most relevant, we first aim at enumerating them all. In a second step, we discuss ways to rank them based on how the concentration of the discriminating compounds is observed to vary in the experiment. This procedure allows a good filter of the solutions, selecting stories that best fit the experimental data.

2 MODELS

2.1 Modeling metabolic stories

In this section, we introduce the notion of story and give a rationale for its definition. Briefly, stories are subgraphs that summarize the flow of matter from a set of source metabolites to a set of target metabolites. The candidates to be the endpoints (sources or targets) of a story should belong to the set of discriminating compounds. To guarantee that stories will have at least one source and one target, we introduce the acyclicity constraint. These two combined constraints lead us to search for DAGs with sources and targets contained in the given set of discriminating compounds. Then, because there can be several paths connecting two discriminating compounds and we want the story to contain all these alternative paths, we impose a constraint of maximality, that is, we search for maximal DAGs, in the sense that alternative pathways between all the nodes should be included, if their addition does not create cycles. In other words, a DAG is maximal if by adding any arc makes it not a DAG anymore, meaning that it contains at least one cycle.

Our goal is to have an algorithm that enumerates all stories, i.e. to provide all possible scenarios that explain the observed transformations. Because our focus is on the relation between discriminating compounds, we use a representation of metabolic networks focused on metabolites, the so-called compound network (Lacroix *et al*., 2008), that is a directed graph in which vertices are compounds and there is an arc from a compound to another compound if there is a reaction that consumes the first to produce the second.

More formally, we introduce a constrained version of the problem of enumerating all maximal DAGs of a graph G (Schwikowski and Speckenmeyer, 2002). Let $G = (\mathbb{B} \cup \mathbb{W}, E)$ be a directed graph such that $\mathbb{B} \cap \mathbb{W} = \emptyset$. We write $V = \mathbb{B} \cup \mathbb{W}$. Nodes in $\mathbb{B}$ are said to be black nodes and correspond to the discriminating compounds, whereas those in $\mathbb{W}$ are said to be white nodes. Let $d^+(u)$ and $d^-(u)$ denote, respectively, the in-degree and the out-degree of a node u. Node u is called a *source* if $d^+(u) = 0$ and $d^-(u) > 0$ and a *target* if $d^-(u) = 0$ and $d^+(u) > 0$.

A metabolic story of G is a maximal acyclic subgraph $G' = (\mathbb{B} \cup \mathbb{W}', E')$ of G with $\mathbb{W}' \subseteq \mathbb{W}$ and $E' \subseteq E$ and such that, for each node $w \in \mathbb{W}'$, w is neither a source nor a target red in G'. Maximality means that it is not possible to add other arcs or nodes without creating cycles, or white sources or targets. We denote by $\Sigma(G)$ the set of stories of G.

2.2 Enumerating metabolic stories

A first step of our algorithm to enumerate $\Sigma(G)$ is to apply compression operations on the input graph obtaining a more

compact representation, which is equivalent in terms of story sets. The operations are (i) white source and target removal that consists in removing iteratively white nodes that are either sources or targets, as such nodes cannot appear in any story; (ii) self-loop removal that consists in removing all arcs of the form (u, u): because stories are acyclic, such arcs do not appear in any story; (iii) forward and backward bottleneck removal, that consists in removing a white node v whose out-degree (respectively, in-degree) is equal to 1, and directly connecting any predecessor (respectively, successor) of v to the unique successor (respectively predecessor) of v (without creating multiarcs). Our preprocessing algorithm consists in applying operations (i), (ii) and (iii) successively until no more white sources and targets, self-loops and bottlenecks are present in the graph. We call the resulting graph a compressed network.

In (Acuna *et al.*, 2012), we proposed a first method to enumerate stories based on a polynomial-time algorithm to compute one story. This is briefly recalled in Supplementary Material S1. More recently we developed a much faster enumerator for stories based on a linear-time enumeration algorithm for non-maximal stories (Borassi *et al.*, 2013) that allows us to explore the whole set of solutions even for genome-scale metabolic networks. This is the enumerator algorithm we use here.

2.3 Scoring function

From a formal point of view, there is no qualitative difference between any two stories. In this sense, whether a given discriminating compound is a source, an intermediate node or a target in a story is indifferent for the enumeration process, as all possible scenarios satisfying the three properties given by the definition, namely, maximality of paths, acyclicity and source/target constraint, have to be computed.

However, in practice, the number of stories can be large and being able to rank them greatly facilitates their analysis. To do this, we propose the following score function:

$$s(S) = \sum_{x \rightsquigarrow y \in S} \omega(x) \times \omega(y) \times \omega(x \rightsquigarrow y),$$

where the score $s(S)$ of a story S is the sum, for each black transformation $x \rightsquigarrow y$, of the product of the *node weights* $\omega(x)$ and $\omega(y)$ of the nodes x and y, times the *path weight* $\omega(x \rightsquigarrow y)$. A black transformation is defined as an arc or a simple white path between two black nodes. A simple white path is a simple path (i.e. containing no cycles) composed of only white nodes between two black ones. The values assigned to the node and path weights will depend on the data available and are thus perfectly suited for the integration of various omics data. For our analysis, we used the topology of the stories and additional data from the metabolomics experiments as described in more detail in the Section 3 (see Table 2).

2.4 Yeast metabolic network

For the analysis of the metabolics experiment (Madalinski *et al.*, 2008), we used the metabolic reconstruction of *Saccharomyces cerevisiae s288c* available in MetExplore (Cottret *et al.*, 2010) (the metabolic model was built based on the YeastCyc database). The procedure followed is briefly described in Supplementary Material S2.

3 RESULTS

3.1 Metabolic stories to analyze metabolomics data

To illustrate how to use our method, we concentrate on the study of the exposition of *S.cerevisiae* to the toxic cadmium (Cd^{2+}) reported in Madalinski *et al.* (2008). A widely studied metabolic pathway in *S.cerevisiae* is the one responsible for glutathione biosynthesis, as it is related to the detoxification process of the cell when exposed to high concentrations of cadmium (Fauchon *et al.*, 2002; Lafaye *et al.*, 2005; Madalinski *et al.*, 2008). Previous studies demonstrated that the presence of such a metal in the environment has a huge impact in terms of gene expression and metabolism, showing that there is a strong response both at the metabolomic and proteomic levels. Basically, glutathione needs to be produced because it is a thiol metabolite linked to the detoxification of cadmium through a process called chelation (Li *et al.*, 1997). Plants are the natural biotope of *S.cerevisiae* and it is known that they are able to tolerate cadmium and other metals up to 1% of their dry weight, which is believed to provide defense against herbivores and pathogenic microorganisms (Fauchon *et al.*, 2002). This exposition to cadmium in natural conditions provides a reason for yeast to keep a detoxification pathway. However, the biosynthesis of glutathione requires high quantities of sulfur. To save sulfur, there is a replacement of abundant sulfur-rich proteins related to other metabolic processes by sulfur-depleted isozymes (i.e. other enzymes that have the same function). Such is the case for the enzymes pyruvate decarboxylase (Pdc1p), enolase (Eno2p) and aldehyde dehydrogenase (Ald6p) that are replaced by isozymes containing less sulfur amino acids (i.e. methionine and cysteine) that are mobilized in the glutathione pathway and are less available for protein synthesis (Fauchon *et al.*, 2002). This response affects a large portion of the metabolic network and represents the mechanism used by the cell to survive under this specific stress condition. Sulfur limitation conditions slow down the growth rate but do not induce this same sulfur-sparing response (Fauchon *et al.*, 2002). A schema of the known glutathione biosynthesis metabolic pathway is presented in Figure 1.

The metabolic network used for this analysis (see the Section 2 for a description) contains 600 metabolites and 949 arcs. Madalinski *et al.* (2008) identified a list of 24 metabolites

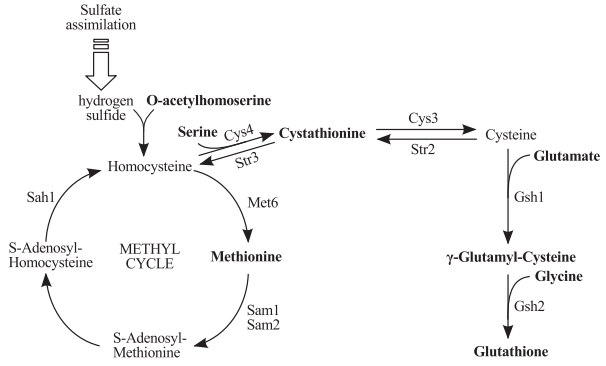


Fig. 1. Glutathione biosynthetic pathway. Compounds in bold are discriminating in Madalinski *et al.* (2008) and are involved in the synthesis of glutathione. Source: adapted from Figure 1 in Lafaye *et al.* (2005)

whose concentration significantly changed after cadmium exposure, shown in the table given in Supplementary Material S3.

It is important to notice here that identification of the metabolites that have changed their concentration is based on a minimum of two orthogonal criteria relative to an authentic compound analyzed under identical experimental conditions: retention time and mass spectrum or retention time and ^{1}H nuclear magnetic resonance (NMR) spectra, accurate mass and tandem mass spectra or accurate mass and related isotopic clusters or ^{1}H and/or ^{13}C NMR with 2D NMR spectrum (Sumner *et al.*, 2007). However, many metabolites are not commercially available and many of them may require tedious and expensive chemical synthesis, which often hampers their definitive metabolite identification. Thus, such compounds remain putatively annotated or characterized.

We decided to perform two analyses to explore the effect of cadmium exposure on *S.cerevisiae* cells. We first enumerated metabolic stories using a set of black nodes restricted to the measured metabolites that are known to participate to the biosynthesis of glutathione. The idea is to check whether our method is able to recover one or more stories that correspond to the known metabolic pathway. In a second step, we enumerated metabolic stories using the entire list of 24 discriminating compounds identified in the metabolomics experiments. In this case, the goal is to analyze both the response of glutathione biosynthesis, but also the potential response of other pathways and the side effects of these responses in the rest of the network.

3.2 First analysis: local response to cadmium exposure, biosynthetic pathway of glutathione

We first consider the aforementioned metabolic pathway directly involved in cadmium detoxification, namely, the glutathione biosynthetic pathway, to enumerate stories and check whether we are able to recover one that fits our current knowledge of the biological process. We thus selected as black nodes for this first analysis only the metabolites that were measured in the experiment (Madalinski *et al.*, 2008) and that are also known to participate in the glutathione biosynthetic pathway (Fauchon *et al.*, 2002). These eight compounds are presented in the table given in Supplementary Material S3 with the third column marked as 'yes': glutathione, O-acetylhomoserine, methionine, glutamate, glutamylcysteine, serine, glycine and cystathionine.

3.2.1. Compressed network A first practical result that follows directly from the properties of our definition of stories is the compressed representation of the subnetwork in which all interactions between the discriminating compounds are captured. The compression is obtained in two steps. In the first step, we extract all biologically relevant routes between the black nodes. In our case, we computed lightest paths between black nodes using the out-degree of a node as its weight, which has been defended as being more biologically sound than a simple shortest-path approach (Blum and Kohlbacher, 2008). The second step is to apply the four compression rules that were previously described briefly (see Section 2) and that are fully detailed in Acuna *et al.* (2012).

The compressed network obtained for the reduced set of black nodes contains 10 nodes and 25 arcs, i.e. represents >98% of compression in terms of nodes and >97% in terms of arcs with respect to the original input size of the *S.cerevisiae* metabolic

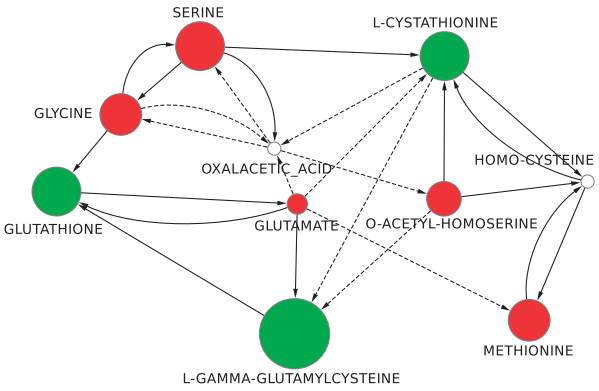


Fig. 2. The compressed network computed considering as black nodes the eight compounds of the table in Supplementary Material S3 marked as present in the glutathione biosynthetic pathway. Green nodes are the ones whose concentration significantly increased in the presence of cadmium, whereas the red are the ones whose concentration significantly decreased. The diameter of the nodes is proportional to the concentration change. Solid arcs represent single reactions connecting the two compounds, whereas dashed ones correspond to a chain of at least two reactions

network. The resulting compressed network is shown in Figure 2. This compression ratio is spectacular, as it is now much easier to visually inspect the network in which we can highlight the metabolites of interest. This type of visualization is, therefore, already a result in itself, which can readily be used to start proposing causal explanations for the changes of metabolite concentrations. To facilitate this, we further enrich this representation with the information on the direction of the change of concentration (whether the metabolite concentration increased or decreased) and the intensity of this change. Of the 8 metabolites considered, 3 had a significant increase of their concentration (reduced glutathione, cystathionine and glutamylcysteine), whereas the other 5 had a significant decrease of their concentration (methionine, O-acetylhomoserine, glutamate, serine and glycine). From now on, we will denote the first set as green nodes and the second set as red nodes. The other nodes, whose concentration did not change significantly, will remain identified as white nodes. We notice that this distinction between red and green nodes is only possible for applications where two conditions are compared. This is the case we consider in this article. When more than two conditions are compared, our methodology still applies, keeping the terminology of black and white nodes. We can produce the compressed network and enumerate the stories. The ranking scheme described later would, however, need to be adapted. Finally, during the preprocessing of the network, some paths are compressed into a single arc. To distinguish between reactions linking two compounds and these compressed paths, we used solid lines for the former and dashed lines for the latter. Importantly, the compression of the network is lossless as it is easily reversible, for instance if we need to have access to the full path of white nodes that indirectly link two black nodes. Interestingly, in practice, although most white nodes can be compressed, some remain. Their compression would prevent us from being able to enumerate the full set of stories. These compounds, although not detected as discriminating, seem to also play an important role in the studied process as

they are at the crossroads between at least two possible routes between discriminating compounds.

3.2.2. Enumerating and scoring the stories The compressed network is already a result *per se*, but its visual inspection remains difficult; the many cycles it contains allow for a reading of the flow of matter in many possible directions, thereby suggesting several possible causal scenarios. Therefore, we go one step further in the analysis and enumerate the metabolic stories. In this analysis, there are a total of 222 stories.

With the aim of classifying the set of computed stories, we have to define how to assign values to the node and arc weights needed by our score function scheme (see Section 2).

There are basically four kinds of interactions that may be observed in a metabolic story (see Table 1). In the following proposal for causal interpretation of each type of arc, we will make the simplifying assumption that each arc is independent from the other ones. In this context, an arc linking a red node to a green node will correspond to the consumption of the red node to the benefit of the green node. If we focus solely on this arc, this can only be explained by an activation of the enzyme catalyzing the reaction linking the two nodes. On the other hand, an arc linking a green node to a red node can be interpreted as the inhibition of the enzyme catalyzing the reaction linking the two nodes. Finally, an arc linking two red nodes can be explained by a domino effect. The simple fact that the substrate concentration decreases causes the product concentration to decrease. This domino effect does not require any enzyme change. It just corresponds to a change in concentration that propagates. The case of green to green arcs can be explained by a similar effect. We additionally need to assume that the enzyme is not present in a limiting amount.

We remind that in this section, our approach is local and focuses on single transformations. We always favor the most parsimonious explanation (the one with fewest enzyme changes), but, in practice, other plausible explanations could be proposed for each arc. Importantly, the notion of enzyme activation or inhibition as used in this article should be understood in a general sense as it captures allosteric regulation of the enzyme or transcriptional regulation of the gene(s) encoding the enzyme. In the application considered here, the time separating the measurements (before and after exposure to cadmium) is large enough to allow to interpret enzyme activations as a change in their concentration through a transcriptional response. Our methodology also applies when the time separating the measurements is shorter. In the following, we propose three ranking schemes for stories. In each of them we favor one type of arc, which means that we look for the stories with a large number of arcs of this type. Even if the individual explanation of each arc is not necessarily correct, the overall optimization of the total number of each arc type makes intuitive sense, and we show that in practice it enables to explore efficiently the space of all stories.

3.2.3 Three scoring schemes Let us start by defining the arc weights that are restricted to being -1, 0 or $+1$. The first scoring scheme privileges stories where green nodes are preferentially targets in the story (i.e. are produced) and, on the other hand, red nodes are preferentially sources in the story (i.e. are consumed). Let us call this score function enzyme-activation-first, as it should privilege arcs from red to green nodes and penalize the inverse as shown in Table 2a. Another possibility is to classify first stories in which the concentration change responses are privileged as shown in Table 2b. Let us call this score function concentration-change-first, as it should privilege arcs from red to red nodes or green to green nodes. Finally, we may define a score function in which we privilege arcs going from green nodes to red nodes; in such a case these arcs represent enzyme inhibition, as shown in Table 2c. Let us call this score function enzyme-inhibition-first.

Once an arc weighting scheme has been chosen, we define the node weights. For our experiments, we define the value $\omega(x)$ for a given node x as its *normalized intensity ratio*, which is its intensity ratio divided by the maximum intensity ratio observed in the experiment (if v is a green node) or the minimum intensity ratio observed in the experiment divided by the intensity ratio of the node (if v is a red node). An example is given in the figure in Supplementary Material S4.

3.2.4 Application to cadmium stress response in yeast Using the three presented score functions, we were able to rank the 222

Table 1. Biological interpretation for arcs in a story

Arc	To red	To green
From red	Concentration change	Enzyme activation
From green	Enzyme inhibition	Concentration change

Table 2. Weights for different score functions of a story

(a) Enzyme activation first			(b) Concentration Change first			(c) Enzyme inhibition first		
Outgoing arcs			Outgoing arcs			Outgoing arcs		
Arc	To red	To green	Arc	To red	To green	Arc	To red	To green
From red	0	1	From red	1	−1	From red	0	−1
From green	−1	0	From green	−1	1	From green	1	0

Note: Table exhibiting the arc weights for interactions between green and red nodes used for computing the score of a story in the context of a metabolomics experiment: (a) weights used to privilege enzyme activation, (b) weights used to privilege concentration change and (c) weights used to privilege enzyme inhibition.

stories previously computed and identify the top scoring stories for each one of the three functions. Figure 3a shows one of the six optimal stories according to the enzyme-activation-first scheme, Figure 3b shows the single optimal story according to the concentration-change-first scheme and Figure 3c shows one of the two optimal stories found according to the enzyme inhibition first scheme. The goal of this first analysis is to try to identify stories that could correspond to the current knowledge on the response of yeast to cadmium exposure, i.e. a story that corresponds to the glutathione biosynthetic pathway previously presented in Figure 1. Among the top scored stories, the one given by the enzyme-activation-first score function (see Figure 3a) agrees well with the current knowledge of yeast response to cadmium. The discussion in Madalinski *et al.* (2008) presents as a result a flux corresponding to the detoxification of cadmium by glutathione, explaining that the levels of metabolites involved in the glutathione biosynthesis pathway (homocysteine, cystathionine, glutamyl-cysteine and glutathione itself) were increased following cadmium exposure, which is the same flow of matter preserved in the story shown in Figure 3a. The story selected by the concentration-change-first score function, shown in Figure 3b, preferentially preserves arcs between two nodes of a same color. The idea is that an increase (or a decrease) of concentration of a given metabolite could be a side effect of the increase (or decrease) of another one. The goal is to minimize the number of arcs that suggest some enzyme activity change, i.e. arcs that involve red and green nodes. Interestingly, the story that scores best with this ranking scheme does not fit with the current knowledge of the response to cadmium exposure. This means that, in principle, there exists a scenario that uses fewer red to green or green to red arcs than the true response (and therefore fewer enzyme changes), but this scenario is not the one taken in practice. There can be a number of reasons why this optimal scenario is not taken. Although any enzyme can, in principle, be activated or inhibited, in practice, some have more degrees of freedom. In addition, some reactions annotated as reversible in general, happen to have one clearly favored direction in specific conditions. Finally, the story presented in Figure 3c preferentially preserves green to red arcs that could represent an enzyme inhibition. Again, this scenario does not fit with the current knowledge on yeast response to cadmium, which indicates that the response is probably not based mostly on enzyme inhibition.

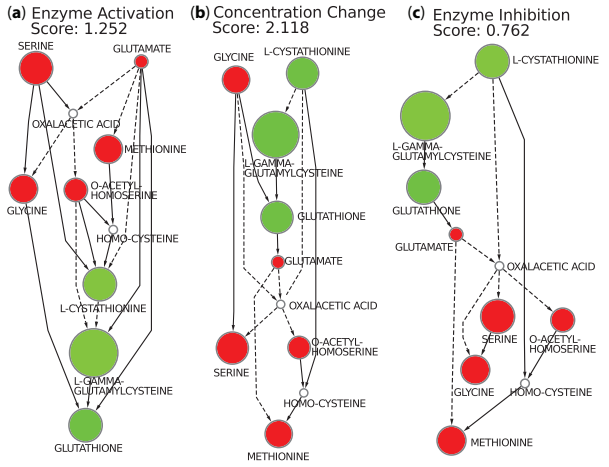


Fig. 3. The best stories generated for our analysis taking into account only the metabolites known to be present in the glutathione biosynthesis and whose concentration significantly changed after cadmium exposure. Green nodes are the ones whose concentration significantly increased in the presence of cadmium, whereas the red nodes are the ones whose concentration significantly decreased. The diameter of the nodes is proportional to the concentration change. Solid arcs represent single reactions connecting the two compounds, whereas dashed ones correspond to a chain of at least two reactions

3.2.5 Anthologies In Figure 3a, a story with score 1.252 for the enzyme-activation-first score function is presented. However, there are other five stories that achieved the same score. These *tied* optimal stories may be combined into a single graph representation to ease the analysis of their differences as presented in Figure 4. A unique graph representing the union of several different stories is called an anthology. Notice that differently from stories, which are maximal DAGs, an anthology contains at least one cycle. The sources and targets (sinks) of an anthology (if any remains) are, however, black nodes only, as with stories. In this case, the equivalent stories are due to the fact that serine, glycine and oxalacetic acid are all interconnected by reversible paths.

3.3 Second analysis: global response to cadmium exposure

For the second analysis, we decided to explore the global response to cadmium exposure and we considered all 24 discriminating compounds. One of them, pyroline-hydroxy-carboxylate, was eliminated when computing the lightest paths

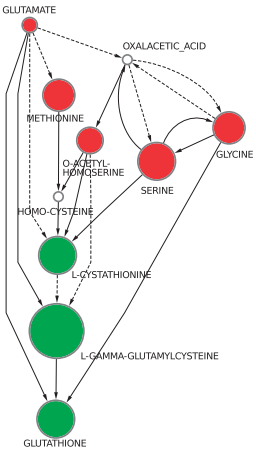


Fig. 4. The anthology combining the six maximal stories obtained with the enzyme-activation-first score function. Notice that the anthology preserves the flow of matter observed in the pathway known to be involved in cadmium detoxification by the yeast. Once more, green nodes are the ones whose concentration significantly increased in the presence of cadmium, whereas the red are the ones whose concentration significantly decreased. The diameter of the nodes is proportional to the magnitude of the concentration change as measured by the intensity ratio of the compound. Solid arcs represent single reactions connecting the two compounds, whereas dashed ones correspond to a chain of at least two reactions

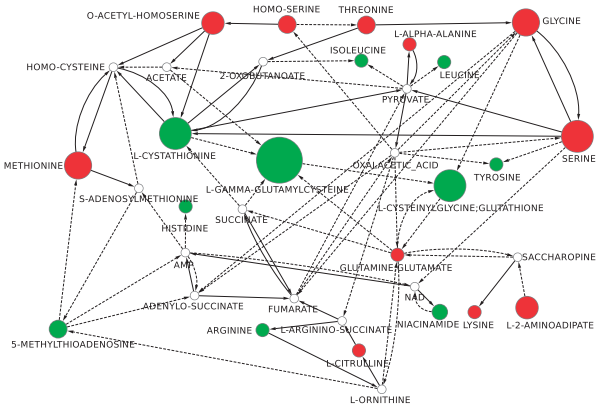


Fig. 5. The compressed network computed for the whole list of discriminating compounds of the table in Supplementary Material 3 and the metabolic network of the yeast strain s288c. Green nodes are the ones whose concentration significantly increased in the presence of cadmium, whereas the red are the ones whose concentration significantly decreased. The diameter of the nodes is proportional to the concentration change. Solid arcs represent single reactions connecting the two compounds, whereas dashed ones correspond to a chain of at least two reactions

between all pairs of black nodes, as it was part of a small disconnected component of the original input graph, most probably due to missing information in the metabolic network reconstruction as the metabolite was present in the metabolome of the strain. The computed compressed network contains 34 nodes and 76 arcs, i.e. a compression of 94% in terms of nodes and 92% in terms of arcs. The resulting compressed network is shown in Figure 5. Again, this compressed network is already a result *per se* as it enables to visualize jointly all the possible ways of explaining the flow of matter through the network. However, in this case again, and probably even more than before, the readability is complicated, and we, therefore, go one step further and compute all stories.

This time, the number of stories is much larger: there are 3 934 160 in total. In fact, this exact number could only be obtained with the recent improvement we proposed in Borassi *et al.* (2013). Before that, the computation would not end in reasonable time and we only had an approximate number. In our initial analysis, the score function that selected a story that best fitted the targeted known metabolic pathway of the glutathione biosynthesis was the enzyme-activation-first scoring scheme. For this reason, we used it also to analyze the larger dataset produced in this second analysis, obtaining 20 maximal stories presented as an anthology in Figure 6. Considering all the metabolites that were measured in the experiment as black nodes in our method allows us to have a more global view of the organism's response to cadmium exposure. This enables to explore whether the other identified paths, apart from the ones involved in the glutathione pathway, are part of this response or simply side effects of the sulfur redirection, as further discussed in the next section.

3.4 Analytic tools

All the compressed networks, stories and anthologies presented in this section were computed using our algorithm called Touché

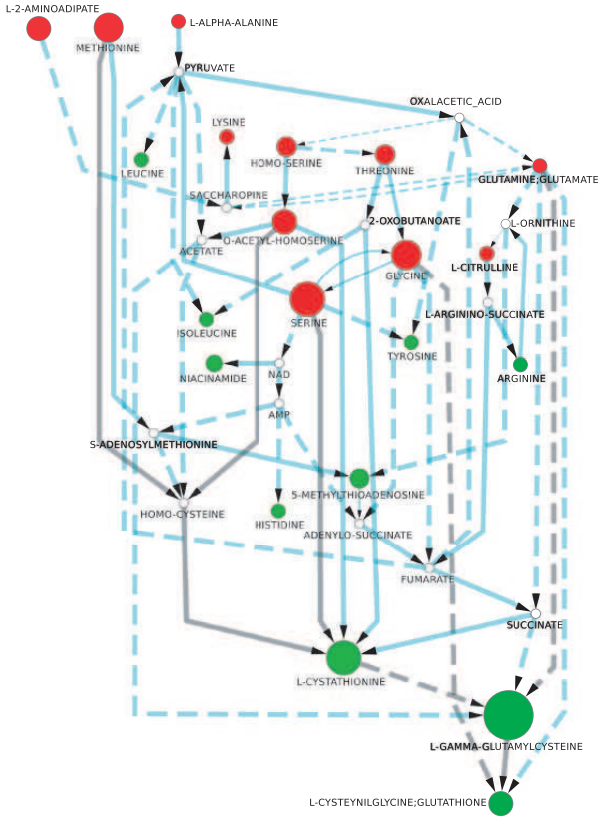


Fig. 6. Anthology corresponding to the 20 stories with the maximal score computed for the experiment on yeast s288c exposed to cadmium. Red nodes correspond to metabolites whose concentration decreased and green nodes to those whose concentration increased in the metabolomics experiment. White nodes have their concentration unchanged or it could not be measured. The diameter of the nodes is proportional to the concentration change. Solid arcs represent single reactions connecting the two compounds, whereas dashed ones correspond to a chain of at least two reactions. The arc's thickness represents the frequency of the arc in the stories making up the anthology, whereas gray arcs correspond to reactions known to be part of the response to cadmium

(Borassi *et al.*, 2013). For visualization and analytical purposes, we used Cytoscape (Shannon *et al.*, 2003), which is a software for network visualization, enriched with a plug-in we developed to enable loading, visualizing and inspecting the three aforementioned objects (compressed networks, stories, anthologies) inside Cytoscape. The plug-in applies the given visual properties corresponding to a metabolomics experiment (e.g. colour and diameter of the nodes, the thickness of an arc corresponding to the frequency of the arc in the stories composing the anthology) and allows a zoom-in in the dashed arcs, exhibiting the paths connecting the two nodes. Both Touché and the Cytoscape plug-in are available on demand.

4 DISCUSSION

Focusing specifically on the biological application presented in the previous section, we may see that exploring the topological properties of the stories through the preprocessing of the input

network creates a compressed network that captures all the relationships between the discriminating compounds in a much smaller graph than the whole network. This already allows a visual inspection of the observed variations, which is rather difficult, if not impossible, in the entire network. On the other hand, one may easily highlight in a whole metabolic pathway map those metabolites whose concentration were detected as having changed using the YeastCyc database. However, because the pathways are presented as disconnected, it is not possible to follow a path that traverses several pathways (see Figure in Supplementary Material S5). To demonstrate the utility of our approach, we used data from Madalinski *et al*. (2008) in which the authors monitor changes in metabolite concentration as a response of the yeast *S.cerevisiae* to cadmium, a toxic chemical. The aim of this study is to analyze the global response of an organism to a stress. Using only the metabolomics experiment data to choose the discriminating compounds and to rank the stories, we are able to obtain stories that correspond well to the current biological understanding of the system under study, as well as to propose new alternatives that could serve as a basis for further experimental validations. Because regulatory information and quantitative information are not needed by the method, this allows it to be used for metabolic network reconstructions even when they are not well refined and where these additional informations may be unavailable or incomplete.

The method herein presented allows visual inspection of a set of discriminating compounds (either local or broader) from metabolomics data in the compressed network, stories and/or anthology with no *a priori* selected pathways. The metabolic stories may be ranked in different ways depending on which interpretation one wishes to emphasize for the causal link between two affected metabolites: enzyme activation, enzyme inhibition or domino effect on the concentration changes of substrates and products. Equally probable stories under any selected ranking scheme can be further grouped into a single anthology that summarizes in a single subnetwork all equivalently plausible alternative stories.

4.1 First analysis: local response to cadmium exposure

The first analysis performed aimed at locally inspecting the yeast response to cadmium exposure limited to the biosynthetic pathway of glutathione, given in Figure 1. Of the 222 stories found, the ones favoring enzyme activation were clearly closer to our current understanding of this response, where an increased sulfur flux passes through homocysteine, cystathionine, cysteine and glutamyl-cysteine to yield high levels of glutathione (Madalinski *et al*., 2008). This same flow of matter is captured in the anthology combining the six best stories under this scoring scheme, shown in Figure 4.

Interestingly, we show that there exists one scenario that, in principle, uses fewer enzymes to explain the observed changes in concentration. This is the scenario that favors concentration changes, shown in Figure 5b. However, this scenario does not match the current knowledge of the main pathway of yeast response to cadmium. In fact, it even uses some reactions in the opposite direction. Because these reactions are annotated as reversible, they can be taken in both directions, at least in theory, and this explains that we found these alternative stories. Those are scenarios that are *a priori* possible. They are not necessarily 'chosen' in practice, possibly because the reactions are only reversible under some conditions that are not met in this experiment. Unfortunately, the precise conditions under which a reaction is reversible are in general not well known. The addition of such knowledge would for sure enable to reduce substantially the number of stories we output, as a large part of the combinatorial explosion we observe comes from these 'cycles'. Conversely, understanding why some possible scenarios are not taken in practice could help to better annotate the reversibility of reactions.

From the list of discriminating compounds identified in Madalinski *et al*. (2008), the ones that are involved in the glutathione biosynthetic pathway (as shown in bold in Fig. 1) are as follows: *O*-acetylhomoserine, methionine, serine, cystathionine, glutamate, γ-glutamyl-cysteine, glycine and glutathione. All of them are present in the compressed networks, stories and anthologies herein presented (Figs 2–6). As concerns cysteine and homocysteine, they were either not measured or not discriminating in Madalinski *et al*. (2008), thus in our analysis they appear as white nodes and may be compressed inside an arc (dashed arcs in Figs 2–6). Cysteine is included in the dashed arc linking cystathionine and *L*-gamma-glutamyl-cysteine that is its expected place based on the biosynthetic pathway of glutathione. Homocysteine is represented as a white node in all figures. Because it is at a crossroads between three black nodes (cystathione, methionine and O-acetylhomoserine), it could not be compressed.

Interestingly, the compounds involved in the methyl cycle, which is a sulfur salvage pathway (see Fig. 1), were not recovered in the highest score stories found in our first analysis. The reason is that the lightest path found between methionine and cystathione in that analysis passed through the reaction catalyzed by the enzyme homocysteine *S*-methyltransferase (Mht1), which is assigned as reversible in the YeastCyc database (Caspi *et al*., 2010) and in our data. This enzyme was described as recycling S-adenosylmethionine (AdoMet) to methionine (Thomas *et al*., 2000).

4.2 Second analysis: global response to cadmium exposure

This more local view of the behavior of the metabolic network of yeast in this stress condition may be contrasted with the second analysis, where the whole list of discriminating compounds was considered. The anthology combining the 20 best stories under the scoring scheme favoring enzyme activation is presented in Figure 6, where the reactions corresponding to the glutathione biosynthesis are highlighted in gray. This is a strong point of our method, as it allows exploring alternative but close scenarios through the analysis of these (and possibly other) stories altogether, which might provide new insights on the underlying processes that took place under the given conditions.

Among the 35 nodes presented in this anthology, eight have sulfur in their chemical structure: AdoMet, γ-glutamylcysteine, 5-methylthioadenosine (MTA), *O*-acetyl-*L*-homoserine, cysteinylglycine, glutathione, cystathionine and L-methionine. Among these sulfured metabolites, the only one that is not involved in the glutathione biosynthesis is MTA, which is instead involved in the MTA cycle, a sulfur salvage pathway (Thomas and Surdin-Kerjan, 1997). This recycles AdoMet to methionine

through a chain of reactions, whereas Mht1 (mentioned earlier in the text) can also perform it in one step, which is important for controlling the intracellular ratio between these two metabolites (Thomas *et al*., 2000) Although there is a redirection of sulfur flux to glutathione biosynthesis after cadmium exposure, the levels of MTA increased as well as those of arginine, which is a precursor for MTA. The metabolites in the methyl cycle are recovered, with the white nodes AdoMet and homocysteine present in the anthology and the metabolite *S*-adenosyl-homocysteine compressed into the arc between them. We have previously tried to link arginine to sulfur metabolism by emphasizing that it is a precursor of spermidine, a polyamine metabolite that is itself involved in the biosynthesis of MTA, a metabolite associated with the methyl cycle and whose levels are increased after cadmium exposure (Madalinski *et al*., 2008). However, experimental data lacked to support this assumption. By using the metabolic stories based approach, the increased levels of arginine are linked to decreased concentrations of citrulline, which has not been formally identified in our experimental conditions, and which is itself linked to glutamate. Besides, citrulline was identified as a discriminating compound in Madalinski *et al*. (2008), but was only indicated as putative, requiring more analysis for final identification. Our results seem to confirm that citrulline was correctly identified. This emphasizes the relevance of using this kind of approach to generate biological hypotheses that have to be further investigated by biologists. Of note, such a link between arginine and sulfur metabolism has been noticed in other organisms (Sekowska *et al*., 2001) and links between nitric oxide and polyamines have been established with cadmium toxicity in wheat roots (Groppa *et al*., 2008). Furthermore, this global view of the discriminating compounds links the sulfur metabolism to non-sulfur amino acids and other metabolites through intermediates of the central metabolism. The amino acids that are precursors to the glutathione synthesis have their levels reduced as expected, whereas most of the others increased. This agrees with the fact that global protein synthesis rapidly drops after cadmium exposure (Lafaye *et al*., 2005), reducing the consumption of amino acids not directly connected to glutathione synthesis.

4.3 Perspectives

We presented a generic method that enables to analyze metabolomics data. This method requires simple input and can be applied to a wide variety of situations. Clearly, the results of the method can be improved with the addition of other types of data. For instance, the use of carbon tracing experiments could help in focusing directly on the stories that are involved in the response to the stress condition, instead of considering the set of all possible stories. Besides, we assumed that the set of discriminating compounds did not need to be questioned. However, these are predicted based on the analysis of peaks in a spectrum. We remark that extracting such information is in itself a bioinformatics challenge. Therefore, a possible extension of the method could be to take into account noisy data, i.e. to deal with a level of confidence for the roles of discriminating compounds and non-discriminating compounds. From the modeling point of view, we enforce that each story corresponds to a flow of matter by the acyclicity constraint. We could relax this constraint by allowing internal cycles, and therefore computing, for each combination of sources and targets, a single story. This will lead to a completely different model and is beyond the scope of this article.

ACKNOWLEDGMENTS

The authors would like to thank S. Klamt and J.J. Heijnen for fruitful discussions.

Funding: European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. (247073)10; the French project (ANR MIRI BLAN08-1335497); and the ANR funded LabEx ECOFECT. It was partially supported by the Plateforme Bioinformatique de Toulouse, ANR-BBSRC Systryp, the CIRIC-INRIA Chile line Natural Resources, the NWO-CLS MEMESA project and the 'DISCO' PRIN National Research Project.

Conflict of Interest: none declared.

REFERENCES

Acuna,V. *et al.* (2012) Telling stories: enumerating maximal directed acyclic graphs with a constrained set of sources and targets. *Theor. Comput. Sci.*, **457**, 1–9.

Antonov,A.V. *et al.* (2009) Ticl – a web tool for network-based interpretation of compound lists inferred by high-throughput metabolomics. *FEBS J.*, **276**, 2084–2094.

Blum,T. and Kohlbacher,O. (2008) Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *J. Comput. Biol.*, **15**, 565–576.

Borassi,M. *et al.* (2013) Telling stories fast: via linear-time delay pitch enumeration. In: *12th International Symposium, SEA 2013*. Rome, Italy, June 5-7, 2013, Proceedings.

Caspi,R. *et al.* (2010) The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res.*, **38**, D473–D479.

Cottret,L. *et al.* (2010) Metexplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Res.*, **38**, W132–W137.

Croes,D. *et al.* (2006) Inferring meaningful pathways in weighted metabolic networks. *J. Mol. Biol.*, **356**, 222–236.

Dittrich,M.T. *et al.* (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, **24**, i223–i231.

Fauchon,M. *et al.* (2002) Sulfur sparing in the yeast proteome in response to sulfur demand. *Mol. Cell*, **9**, 713–723.

Faust,K. *et al.* (2010) Pathway discovery in metabolic networks by subgraph extraction. *Bioinformatics*, **26**, 1211–1218, 2010.

Groppa,M.D. *et al.* (2008) Benavides. Nitric oxide, polyamines and cd-induced phytotoxicity in wheat roots. *Phytochemistry*, **69**, 2609–2615.

Khatri,P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.

Lacroix,V. *et al.* (2008) An introduction to metabolic networks and their structural analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **5**, 594–617.

Lafaye,A. *et al.* (2005) Combined proteome and metabolite-profiling analyses reveal surprising insights into yeast sulfur metabolism. *J. Biol. Chem.*, **280**, 24723–24730.

Leader,D.P. *et al.* (2011) Barrett. Pathos: a web facility that uses metabolic maps to display experimental changes in metabolites identified by mass spectrometry. *Rapid Commun. Mass Spectrom.*, **25**, 3422–3426.

Li,Z.-S. *et al.* (1997) A new pathway for vacuolar cadmium sequestration in *Saccharomyces cerevisiae*: Ycf1-catalyzed transport of glutathionato cadmium. *Proc. Natl Acad. Sci. USA*, **94**, 42–47.

Madalinski,G. *et al.* (2008) Direct introduction of biological samples into a ltq-orbitrap hybrid mass spectrometer as a tool for fast metabolome analysis. *Anal. Chem.*, **80**, 3291–3303.

Schuster,S. *et al.* (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53–60.

Schuster,S. and Hilgetag,C. (1994) On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.*, **2**, 165–182.

Schwikowski,B. and Speckenmeyer,E. (2002) On enumerating all minimal solutions of feedback problems. *Discrete Appl. Math.*, **117**, 253–265.

Scott,M.S. *et al.* (2005) Identifying regulatory subnetworks for a set of genes. *Mol. Cell. Proteomics*, **4**, 683–692.

Sekowska,A. *et al.* (2001) Extracting biological information from DNA arrays: an unexpected link between arginine and methionine metabolism in bacillus subtilis. *Genome Biol.*, **2**.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Sumner,L.W. *et al.* (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics*, **3**, 211–221.

Thiele,I. and Palsson,B.O. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.*, **5**, 93–121.

Thomas,D. *et al.* (2000) Reverse methionine biosynthesis from s-adenosylmethionine in eukaryotic cells. *J. Biol. Chem.*, **275**, 40718–40724.

Thomas,D. and Surdin-Kerjan,Y. (1997) Metabolism of sulfur amino acids in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.*, **61**, 503–532.

van Helden,J. *et al.* (2002) Bioinformatics and Genome Analysis. In: Mewes,H.-W., Seidel,H. and Weiss,B. (eds) *Graph-Based Analysis of Metabolic Networks*. Vol. 38, Springer, Berlin Heidelberg, pp. 245–274.

Xia,J. *et al.* (2012) Metaboanalyst 2.0-a comprehensive server for metabolomic data analysis. *Nucleic Acids Res.*, **40**, W127–W133.

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–3

Telling metabolic stories to explore metabolomics data – A case study on the Yeast response to cadmium exposure (Supplementary material)

Paulo Vieira Milreu[1,*], Cecilia Coimbra Klein[1,2], Ludovic Cottret[3], Vicente Acuña[1,4], Etienne Birmelé[1,5], Michele Borassi[6], Christophe Junot[7], Alberto Marchetti-Spaccamela[8], Andrea Marino[9], Leen Stougie[10], Fabien Jourdan[11], Pierluigi Crescenzi[9], Vincent Lacroix[1*], Marie-France Sagot[1,*]

[1] INRIA Grenoble Rhône-Alpes & Université de Lyon, F-69000 Lyon; Université Lyon 1; CNRS, UMR5558, LBBE, France
[2] Laboratório Nacional de Computação Científica (LNCC), Petrópolis, Brazil;
[3] LISBP, UMR CNRS 5504 - INRA 792, Toulouse, France;
[4] Mathomics (UMI 2807 CNRS) and Center for Genome Regulation (Fondap 15090007), University of Chile, Santiago;
[5] Lab. Statistique et Génome, CNRS UMR8071 INRA1152, Université d'Évry, France;
[6] Scuola Normale Superiore, 56126 Pisa, Italy;
[7] Laboratoire d'Etude du Métabolisme des Médicaments, DSV/iBiTecS/SPI, CEA/Saclay, 91191 Gif-sur-Yvette, France;
[8] La Sapienza University of Rome, Italy;
[9] Università di Firenze, Dipartimento di Sistemi e Informatica, I-50134 Firenze, Italy;
[10] VU University and CWI, Amsterdam, The Netherlands;
[11] INRA UMR1331 - Toxalim, Toulouse, France
∗ Corresponding authors: pvmilreu@gmail.com, vincent.lacroix@univ-lyon1.fr and marie-france.sagot@inria.fr.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

SUPPLEMENTARY MATERIAL

1) Enumeration Algorithm

The algorithm to compute one story has as input a compressed network G and a total order π of the nodes and is illustrated in Figure 1. From π, we may easily compute what we call a **pitch**, which is defined as a story except for the maximality condition. Moving from a total order π to a pitch is done by keeping only arcs that are consistent with π and, after that, removing recursively any remaining white source or target. Completing a pitch into a story is done by adding paths between black nodes while avoiding cycles. The algorithm searches for extensions of the pitch following the order π of the nodes, moving to the next node when no new path may be added from the previous one. The resulting graph is a story. The enumeration was performed by examining all possible orderings of the nodes. More details on the mathematical modelling, the preprocessing step, the algorithms and their computational complexity are in [1].

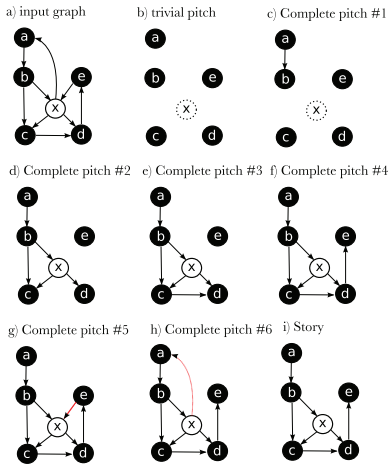


Fig. 1. a) The input graph with set of black nodes $B = \{a, b, c, d, e\}$ and white nodes $W = \{x\}$. b) The starting pitch, which is simply a graph $V = B$ and no arcs. c) The path $a \rightsquigarrow b$ is added to the pitch. d) Three paths starting from b are added to the pitch: $b \rightsquigarrow c$, $b \rightsquigarrow x \rightsquigarrow c$ and $b \rightsquigarrow x \rightsquigarrow d$. Notice that as x did not belong to the pitch at this point, the algorithm goes further and stops only when nodes in the pitch are found. e) The path $c \rightsquigarrow d$ is added to the pitch. f) The path $d \rightsquigarrow e$ is added to the pitch. g) The path $e \rightsquigarrow x$ is evaluated but cannot be added since e comes after x in the current partial order inferred from the pitch, and therefore such an addition creates at least one cycle, for instance $e \rightarrow x \rightarrow d \rightarrow e$. h) The path $x \rightsquigarrow a$ is evaluated but cannot be added since x comes after a in the current partial order inferred from the pitch, and therefore such an addition creates at least one cycle, for instance $x \rightarrow a \rightarrow b \rightarrow x$. i) There are no more nodes to traverse, the final object is a maximal pitch, *i.e.*, a story.

2) Yeast metabolic network

We retrieved the reconstruction of the metabolic network of *Saccharomyces cerevisiae s288c* from MetExplore. This platform allows applying different filters to the network. Herein it is restricted to the small molecule metabolism, *i.e.* reactions involving one or more macromolecules such as proteins or nucleic acids are not represented. In addition, reactions involving pairs of cofactors were split into two reactions, such as the following transformation (or reverse): *compound A* + ATP $\rightarrow$ *compound B* + ADP + Pi will be represented as reaction 1: *compound A* $\leftrightarrow$ *compound B*, and reaction 2: ATP $\rightarrow$ ADP + Pi. Ubiquitous compounds (*i.e.*, water, proton, carbon dioxide, phosphate, diphosphate, ammonia, hydrogen peroxide and oxygen) and cell compartments were as well removed from the network.

3) List of discriminating compounds

Table 1. List of discriminating compounds for the *S. cerevisiae* cell exposed to cadmium

Metabolite ID	intensity ratio	Present in the pathway
arginine	1.9	no
reduced glutathione	33.9	**yes**
O-acetylhomoserine (*)	0.5	**yes**
2-aminoadipate (*)	0.5	no
niacinamide (*)	4.8	no
pyridine-3-aldoxime (*)	4.8	no
pyrroline-hydroxy-carboxylate	0.7	no
methionine	0.3	**yes**
citrulline (*)	0.7	no
threonine	0.6	no
homoserine	0.6	no
glutamine	0.7	no
glutamate	0.8	**yes**
glutamylcysteine	192.2	**yes**
5-methylthioadenosine	11.0	no
serine	0.2	**yes**
glycine (*)	0.3	**yes**
cystathionine	50.5	**yes**
lysine	0.7	no
cysteinylglycine (*)	35.9	no
leucine/isoleucine	1.2	no
tyrosine	2.9	no
histidine	1.2	no
alanine	0.8	no

List of 24 metabolites from the yeast metabolic network whose concentration significantly varied under cadmium exposed. The intensity ratio column presents the ratio between the stress condition and the control. The 3rd column indicates whether the compound is present in the glutathione biosynthetic pathway (Fig. 1 of the main manuscript) or not. Metabolites identified with an (*) after their names were putative metabolites requiring more analysis for final identification.

4) Example of small metabolic story

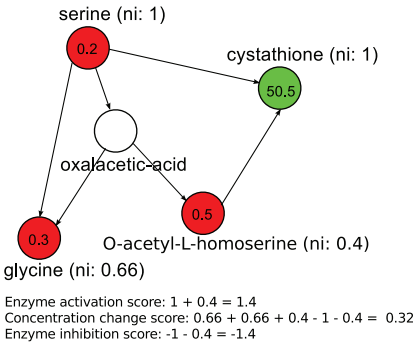


Fig. 2. Considering the story with 5 nodes presented in the figure, we may compute its score for the three different scoring schemes given in Table 2 of the main manuscript. The minimum concentration observed in the story for the red nodes is 0.2 and the maximum concentration observed for a green node is 50.5. Therefore, $ni(serine) = 0.2/0.2 = 1$, $ni(cystathione) = 50.5/50.5 = 1$, $ni(glycine) = 0.2/0.3 = 0.66$ and $ni(O-acetyl-L-homoserine) = 0.2/0.5 = 0.4$. Summing up the contribution of each arc as the product of the normalized intensity ratios of its extremities times the corresponding entrance in the score matrix, we obtain an enzyme activation score of 1.4, a concentration change score of 0.32 and an enzyme inhibition score of -1.4. Notice that these scores cannot be compared between them, their role is to enable us to compare different stories.

5) Mapping of metabolites on the YeastCyc overview diagram

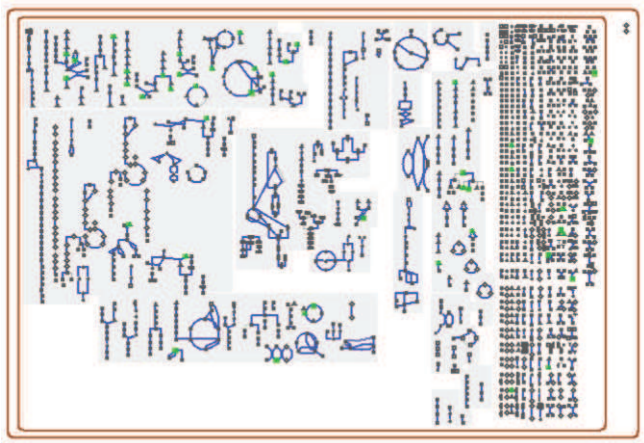


Fig. 3. Mapping of metabolites on the YeastCyc overview diagram. Notice that there are more highlighted metabolites than appear in the list. This is due to the fact that this view is pathway-oriented and metabolites are therefore duplicated. Moreover, there is no link between the pathways so the network connectivity is lost.

6) Glossary

Table 2. Main definitions used in the paper

Word	Definition
Arc	An arc of a graph $G(V, A)$ is an ordered pair $(u, v) \in A$, with $u, v \in V$. Such an arc is outgoing from u and incoming into v
Black nodes	Nodes corresponding to the discriminating compounds;
Directed graph (digraph)	A digraph is a pair (V, A), where V is a set of nodes and A, the arc set, is a binary relation on V
Discriminating compounds	Compounds measured and whose concentration change is statistically significant
Green nodes	Nodes corresponding to the discriminating compounds whose concentration significantly increased
Metabolic story	Maximal directed acyclic subgraph that contains only black nodes as sources and targets
Red nodes	Nodes corresponding to the discriminating compounds whose concentration significantly decreased
Source	A node that has no incoming arc
Target	A node that has no outgoing arc
White nodes	Nodes corresponding to non-discriminating compounds, *i.e.,* compounds that were not measured or whose concentration did not significantly change

REFERENCES

[1]V. Acuna, E. Birmelé, L. Cottret, P. Crescenzi, F. Jourdan, V. Lacroix, A. Marchetti-Spaccamela, A. Marino, P. V. Milreu, M.-F. Sagot, and L. Stougie. Telling stories: Enumerating maximal directed acyclic graphs with a constrained set of sources and targets. *Theor. Comput. Sci.*, 457:1–9, 2012.

TITRE en français

Une étude bioinformatique du dialogue métabolique entre trypanosome non pathogène et son endosymbiote à des buts évolutifs et fonctionnels

RÉSUMÉ en français

Lors de cette thèse, nous avons présenté trois principaux types d'analyses du métabolisme, dont la plupart impliquaient la symbiose : dialogue métabolique entre un trypanosomatide et son symbiote, analyses comparatives de réseaux métaboliques et exploration de données métabolomiques. Tous ont été essentiellement basés sur des données de génomique où les capacités métaboliques ont été prédites à partir des gènes annotés de l'organisme cible, et ont été affinées avec d'autres types de données en fonction de l'objectif et de la portée de chaque analyse. Le dialogue métabolique entre un trypanosomatide et son symbiote a été explorée avec des objectifs fonctionnels et évolutifs qui comprennaient une analyse des voies de synthèse des acides aminés essentiels et des vitamines telles que ces voies sont classiquement définies, une exploration de réseaux complets métaboliques et une recherche de potentiels transferts horizontaux de gènes des bactéries vers les trypanosomatides. Les analyses comparatives effectuées ont mis l'accent sur les capacités métaboliques communes de bactéries appartenant à différents groupes de vie, et nous avons proposé une méthode pour établir automatiquement les activités métaboliques communes ou spécifiques à chaque groupe. Nous avons appliqué notre méthode d'énumération d'histoires métaboliques à la réponse de la levure à une exposition au cadmium comme une validation de cette approche sur une réaction au stress bien étudiée. Nous avons montré que la méthode a bien capté la connaissance que nous avons de cette réponse en plus de permettre de nouvelles interprétations des données métabolomiques mappées sur le réseau métabolique complet de la levure.

MOTS-CLEFS en français

symbiose, voies métaboliques, réseaux métaboliques, mutualisme, trypanosomatids.

Title in english

Bioinformatic study of the metabolic dialog between a non-pathogenic trypanosomatid and its endosymbiont with evolutionary and functional goals

Abstract in english

In this thesis, we presented three main types of analyses of metabolism, most of which involved symbiosis: metabolic dialogue between a trypanosomatid and its symbiont, comparative analyses of metabolic networks and exploration of metabolomics data. All of them were essentially based on genomics data where metabolic capabilities were predicted from the annotated genes of the target organism, and were further refined with other types of data depending on the aim and scope of each investigation. The metabolic dialogue between a trypanosomatid and its symbiont was explored with functional and evolutionary goals which included analysing the classically defined pathways for the synthesis of essential amino acids and vitamins, exploring the genome-scale metabolic networks and searching for potential horizontal gene transfers from bacteria to the trypanosomatids. The comparative analyses performed focused on the common metabolic capabilities of different lifestyle groups of bacteria and we proposed a method to automatically establish the common and the group-specific activities. The application of our method on metabolic stories enumeration to the yeast response to cadmium exposure was a validation of this approach on a well-studied biological response to stress. We showed that the method captured well the underlying knowledge as it extracted stories allowing for further interpretations of the metabolomics data mapped into the genome-scale metabolic model of yeast.

Keywords in english

symbiosis, metabolic pathways, metabolic networks, mutualistic associations, trypanosomatids.