A Computational Investigation of Stage-specific and Species-specific Factors of the *Leishmania* Parasite to understand the Survival Strategies

Thesis Submitted to AcSIR for the Award of
the Degree of
DOCTOR OF PHILOSOPHY
In
Biological Sciences

By

Abhishek Subramanian

10BB12A26071

Under the guidance of
Dr. Ram Rup Sarkar

CSIR-National Chemical Laboratory, Pune, India

DECLARATION

I hereby declare that the thesis entitled "**A Computational Investigation of Stage-specific and Species-specific Factors of the *Leishmania* Parasite to understand the Survival Strategies**" submitted for the degree of Doctor of Philosophy in Biological Sciences to the Academy of Scientific & Innovative Research (AcSIR), has been carried out by me at the Chemical Engineering and Process Development Division of CSIR-National Chemical Laboratory, Pune under the supervision of Dr. Ram Rup Sarkar. Such material as has been obtained by other sources has been duly acknowledged in the thesis. The work is original and has not been submitted in part or full by me for any other degree or diploma to any other Institution or University.

Abhishek Subramanian

Ph. D candidate

Date: 01/02/2018

Place: CSIR-NCL, Pune

सीएसआयआर–राष्ट्रीय रासायनिक प्रयोगशाला
(वैज्ञानिक तथा औद्योगिक अनुसंधान परिषद)
डॉ. होमी भाभा मार्ग, पुणे – 411 008. भारत
CSIR-NATIONAL CHEMICAL LABORATORY
(Council of Scientific & Industrial Research)
Dr. Homi Bhabha Road, Pune - 411008. India

Certificate

This is to certify that the work incorporated in this Ph.D. thesis entitled "**A Computational Investigation of Stage-specific and Species-specific Factors of the *Leishmania* Parasite to understand the Survival Strategies"** submitted by **Mr. Abhishek Subramanian** to the Academy of Scientific and Innovative Research (AcSIR) is in fulfilment of the requirements for the award of the Degree of Philosophy and embodies original research work under my supervision and guidance. I further certify that this work has not been submitted to any other University or Institution in part or full for the award of any degree or diploma. Research material obtained from other sources has been duly acknowledged in the thesis. Any text, illustration, table etc., used in the thesis from other sources, have been duly cited and acknowledged.

Abhishek

(Student)

Mr. Abhishek Subramanian,
Senior Research Fellow,
Chemical Engineering & Process Development,
CSIR – National Chemical Laboratory,
Dr. Homi Bhabha Road, Pashan,
Pune – 411008

Ram Rup Sarkar

(Supervisor)

Dr. Ram Rup Sarkar,
Principal Scientist,
Chemical Engineering & Process Development,
CSIR-National Chemical Laboratory,
Dr. Homi Bhabha Road, Pashan,
Pune 411008

Dedicated to my family, friends, teachers and all the scientists who have inspired me to think rationally

Acknowledgements

I wish to express my heartfelt acknowledgments and gratitude to all those people who were instrumental in bringing out the best in me during my five years of Ph.D journey.

First and foremost, I owe my sincere thanks to Dr. Ram Rup Sarkar, my supervisor who has guided and supported me in the last 5 years to be confident, to be passionate about research and to explore new ideas fearlessly. It is because of him that I have learnt to be organized, at least to some extent. Most importantly, he has also accepted my mistakes graciously and provided timely solutions to improve myself. I thank him for all the scientific and non-scientific discussions that exposed me to new realizations and avenues. I hope he will continue to guide me in the future as well.

I am also grateful to my Doctoral Advisory Committee members Dr. Dhanasekharan Shanmugam, Dr. Chetan Gadgil and Dr. Anu Raghunathan for evaluating my progress and providing many useful comments & suggestions that improved my research outlook. I wish they will continue their support. From CSIR-NCL, I would like to thank the previous and the present directors Dr. S. Pal and Dr. A. K. Nangia and heads of CEPD division, Dr. V. V. Ranade and Dr. S. S. Tambe for providing the research infrastructure and other official formalities. I would like to thank the Student Academic Office, CSIR-NCL for their administrative support. I also officially thank AcSIR for providing a platform to facilitate the Ph.D process, DBT - BINC for providing the Junior and Senior Research Fellowships and Mr. Ajinkya of the Accounts Section for efficient, hassle-free processing of documents of the fellowship release.

I would like to thank my present labmates and friends, Sutanu, Arpit, Saikat, Noopur, Piyali, Rupa for patiently listening to my frustrations and all the fun we had in and out of NCL. I am also grateful to other lab alumni and trainees –Jitesh, Apurv, Devrat, Varsha, Shomeek, Prachi, Swarnabha, Vidhi, Pranali and Rohini for giving me the experience of a teacher and also, in turn teaching me a lot of new things. I specifically thank Mr. Jitesh Jhawar for the great scientific discussions, for making me realize my drawbacks and the conceptual understanding of bridging mathematics and biology. I express my gratitude to Mr. Apurv Mishra, a six month dissertation student who helped me with the genome-scale reconstruction studies. I learnt to be patient from him and also to keep small goals.

I would also express a deep gratitude to my friends Snehal, Ragini, Sneha, Madhura, and Lakhan who provided rightful advice and many opportunities to detach myself from research

work & spend time for myself. It was because of them I was able to avoid mental stresses and also be social. I specially thank my friend Snehal who has been listening to all my troubles in every step of my Ph.D.

I would like to thank my teacher, mentor Mr. Atish Gavit for introducing me to bioinformatics as a field of applied science in biology. It was because of him I appeared for the entrance exam at University of Pune and was able to graduate in Bioinformatics. Dr. Manali Joshi gave me the first-hand experience of research in computational biology. I thank her for moulding my excitement and curiosity into organized research. I am also lucky to get guidance from seniors like Dr. Priyabrata Panigrahi, Dr. Pandurang Kolekar and Dr. Shraddha Puntambekar who exposed me to different fields and have been a constant source of inspiration till today. I also would like to thank all my school & college teachers for imparting their knowledge and experiences.

Last, but never the least, I would like to thank my family who have sacrificed a lot, emotionally supported & accommodated me throughout. I also apologize to them for any stress I caused during my Ph.D journey. Even though he is younger to me, my brother Aditya took the emotional responsibility of the family and let me focus on my research. Most importantly, it was because of my parents, I have learnt the lessons of honesty, hard work and perseverance. Because of irregular fellowships, my father Mr. N. K. Subramanian had to sacrifice his retired life and still continues to work and earn for the family. I am really sorry for all the sacrifices you were forced to undertake & to make me achieve this feat. I promise to become more responsible and be at least half of what you are. Whatever achievements you achieve in life, no happiness can replace the happiness of being with your mother. My mother Mrs. Rajeswari Subramanian has supported me in every step of life. She has totally dedicated her life for us. It was because of her encouragement that I could follow my own path. Thank you very, very much.

Scientifically speaking, my Ph.D journey was a constraint-based multi-variable optimization problem. Without all of you in my life, I could not have found a solution.

THANK YOU !!!

- *Abhishek Subramanian*

Contents

List of Figures

Figures (Appendix A)

List of Tables

Tables (Appendix B)

Abstract

Leishmania are protozoan parasites that cause the neglected tropical disease leishmaniasis in humans, which spread through the bites of infected female sandfly vectors. The *Leishmania* parasite exhibits a digenetic lifecycle, where the motile promastigote forms of the parasite survive and proliferate in the midgut of the sandfly vector, whereas the non-motile amastigotes persists in the macrophage phagolysozome of the human host; the environments being principally antagonistic. This problem is aggravated by the presence of differential clinical manifestations - cutaneous and visceral leishmaniasis caused by a specific set of species. Experimental observations suggest that species-specific differences can largely be attributed to global translation regulation of the proteomes and the stage-specific differences to the adaptations of *Leishmania* metabolism to a stage-specific uptake of metabolites under the two environments. We hypothesize that, with respect to global translation regulation, codon usage bias is the most important mechanism that varies across *Leishmania* species and the unique organization of the *Leishmania* metabolism is the most influential factor that promotes adaptations under varying metabolic environments. Observing the pattern of variations in these factors across stages and species can help unravel the survival strategies of parasite within the hosts.

In *Leishmania* species, directional mutational pressure and translation selection acting on the synonymous position of codons to maintain a G or C nucleotide are the main evolutionary players that preserve frequent codons in genes. There is no effect of amino acid composition bias on choice of codons within a gene. Bias in usage of specific codons within a gene constrains the formation of mRNA secondary structures at 5'end of the mRNA and hence, can be a putative mechanism for global regulation of translation. The occurrence of codon context pairs follows codon usage bias and is related to efficient translation elongation. A high codon adaptation is observed in energy metabolism, translation and stress related genes of *L. infantum* and *L. donovani* as opposed to other species, suggesting species-specificity in codon usage.

Lately, a number of metabolomics studies have revived the interest to understand metabolic strategies utilized by the *Leishmania* parasite for optimal survival within its hosts. For the first time, we reconstructed the entire genome-scale metabolism of a *Leishmania infantum* strain speculated to cause infantile visceral leishmaniasis. Subjecting the reconstruction to flux-based analysis, it was identified that *L. infantum* possesses a unique metabolic organization, the functioning of which can be explained with respect to the

stoichiometric and reversibility constraints largely imposed by the underlying metabolic network structure alone. A dynamic non-essential amino acid motif exists within the network that promotes a restricted re-distribution of resources to yield required essential metabolites. Further, subcellular compartments regulate this metabolic re-routing by reinforcing the physiological coupling of specific reactions. These constraints create an increased flux through glycosomal succinate fermentation in promastigotes, glutamate & aspartate formation in both stages, increased formation of overflow metabolites in promastigotes, unique glycosomal fatty acid β-oxidation to maintain redox balance in amastigotes and re-routing of glucose only towards mannogen formation in amastigotes. The role of glucose and tyrosine was to supplement the quantity of output metabolites formed through non-essential amino acids, whose uptake is a bare essential for optimal production of biomass. This unique metabolic organization is robust against accidental errors and provides a wide array of alternatives for the parasite to achieve optimal survival.

Ultimately, the relative contribution of eight inter-correlated predictors representing translational and functional constraints on the evolutionary rates of singleton metabolic genes was measured and their effects compared across three *Leishmania* metabolomes. Our analysis revealed that codon adaptation, multi-functionality and flux-coupling potential of an enzyme are independent contributors of enzyme evolutionary rates and can together explain a large variation in enzyme evolutionary rates across species. For the first time, we document that a species-specific occurrence of duplicated genes in novel subcellular locations can create new flux routes through certain singleton flux-coupled enzymes, thereby constraining their evolution. A cross-species comparison of the contributory factors exposed both common and species-specific genes whose evolutionary divergence was constrained by multiple independent factors. Out of these, previously known pharmacological targets and virulence factors in *Leishmania* were identified, suggesting their evolutionary reasons for being important survival factors to the parasite. These highlight the importance of comprehensive multivariate studies in understanding the various causes and consequences of evolutionary divergence and conservation in *Leishmania* metabolism; thereby, identifying targetable mechanisms and genes, which can be further perused for designing novel strategies against parasite persistence.

The work performed in this thesis provides many experimentally testable hypotheses that can uncover the underlying mechanisms of adaptations implemented by the parasite within the host. Understanding these survival strategies can provide a fresh impetus to the identification of drug targets and their prioritization to combat this neglected, tropical disease.

Publications

Related to thesis

1. ***Abhishek Subramanian***, Ram Rup Sarkar (2015), "Comparison of codon usage bias across *Leishmania* and Trypanosomatids to understand mRNA secondary structure, relative protein abundance and pathway functions", *Genomics*, 106:232–241

2. ***Abhishek Subramanian***, Ram Rup Sarkar (2015), "Data in support of large scale comparative codon usage analysis in *Leishmania* and Trypanosomatids", *Data in Brief*, 4:269–272

3. ***Abhishek Subramanian***, Jitesh Jhawar, Ram Rup Sarkar (2015), "Dissecting *Leishmania infantum* Energy Metabolism - A Systems Perspective", *PLoS ONE*, 10(9): e0137976

4. ***Abhishek Subramanian*,** Ram Rup Sarkar: Network structure and enzymatic evolution in *Leishmania* metabolism: a computational study. *Proceedings of the International Symposium on Mathematical and Computational Biology - BIOMAT 2015*., *World Scientific*, ISBN: 978-981-3141-90-2, 1 – 20, 11/2015, DOI:10.1142/9789813141919_0001

5. ***Abhishek Subramanian***, Ram Rup Sarkar (2017), "Revealing the mystery of metabolic adaptations using a genome scale model of *Leishmania infantum*", *Scientific Reports*, 7(1):10262

6. ***Abhishek Subramanian***, Ram Rup Sarkar, "Identification of confounding genotype-phenotype features that constrain metabolic enzyme evolution in *Leishmania* species" (Manuscript submitted)

Other publications

7. ***Abhishek Subramanian***, Ram Rup Sarkar (2015), "Dynamics of GLI regulation and a strategy to control cancerous situation: hedgehog signaling pathway revisited", *J. Biol. Syst.* 23: 1550033

8. ***Abhishek Subramanian***, Vidhi Singh, Ram Rup Sarkar (2015), "Understanding Visceral Leishmaniasis Disease Transmission and its Control - A Study Based on Mathematical Modeling", *Mathematics*, 3:913-944

9. Rupa Bhowmick, ***Abhishek Subramanian***, Ram Rup Sarkar (2015), "Exploring the differences in metabolic behavior of astrocyte and glioblastoma: a flux balance analysis approach", *Syst. Synth. Biol.* 9:159–177

10. Santanu Biswas, ***Abhishek Subramanian***, Ibrahim M. ELMojtaba, Joydev Chattopadhyay, Ram Rup Sarkar (2017), "Optimal combinations of control strategies and cost-effective analysis for visceral leishmaniasis disease transmission", *PLoS ONE*, 12(2): e0172465

11. Sutanu Nandi, ***Abhishek Subramanian***, Ram Rup Sarkar (2017), "An integrative machine learning strategy for improved prediction of essential genes in *Escherichia coli* metabolism using flux-coupled features", *Mol. Biosyst.*, 13, 1584 – 1596

12. Swarnendu Banerjee, ***Abhishek Subramanian***, Joydev Chattopadhyay, Ram Rup Sarkar (2017), "Exploring the role of GS–GOGAT cycle in microcystin synthesis and regulation–a model based analysis", *Mol. BioSyst.*, 2017, 13, 2603-2614

Chapter 1 – Introduction

Members of the *Leishmania* genus are protozoan parasites that cause the neglected tropical disease leishmaniasis in humans. They spread from an infected human host to a susceptible human through the bites of infected female sandfly vectors, largely belonging to the *Phlebotomus* and *Lutzomyia* genera. The leishmaniasis infection can be clinically categorized into i) cutaneous leishmaniasis (CL): less severe form, well characterized by skin sores, which on long-term exposure may become a skin ulcer that heals very slowly and, ii) visceral leishmaniasis or kala-azar (VL): more severe form of the disease where the infection spreads to visceral organs like spleen, liver, kidney, etc. (http://www.who.int/mediacentre/factsheets/fs375/en/). Leishmaniasis in humans is collectively caused by more than 20 *Leishmania* species that are distributed across five continents (except Australia and Antarctica), while being heavily concentrated in the tropical countries. Around 1 billion people living in endemic areas are at a risk of the leishmaniasis infection globally, with 1 million reported cutaneous leishmaniasis cases in the last 5 years and 300,000 annual, estimated visceral leishmaniasis cases worldwide (http://www.who.int/leishmaniasis/en/). This problem is further intensified by the underestimation of actual number of cases (Singh et al. 2010), inappropriate application of vector controls (Biswas et al. 2017), unavailability of a vaccine (Zand and Narasu 2013; Kumar and Engwerda 2014; Schroeder and Aebischer 2014) and drug resistance (Croft et al. 2006; Monzote 2009). The wide prevalence of the disease and the diversity of clinical manifestations can be attributed to the adaptive mechanisms developed within the parasite for survival and the inherent species-specific nature of the infection, respectively.

Biologically, the *Leishmania* parasite exhibits a digenetic lifecycle namely, the promastigotes and amastigotes (Fig. 1.1). The motile promastigote forms of the parasite survive and proliferate in the midgut of the sandfly vector, whereas the non-motile amastigotes persists in the macrophage phagolysozome of the human host. Paleoparasitological evidence suggests that leishmaniasis might have existed in humans from 3500-2800 BC or long before this period (Tuon et al. 2008). This long-standing evolutionary association might have equipped the parasite with unique mechanisms that support their existence within the hosts. Furthermore, experimental observations of *Leishmania* parasite differentiation from the promastigote to amastigote being triggered by sudden changes in pH, temperature, and carbon sources supports this assertion (Zilberstein and Shapira 1994; Saunders et al. 2014).

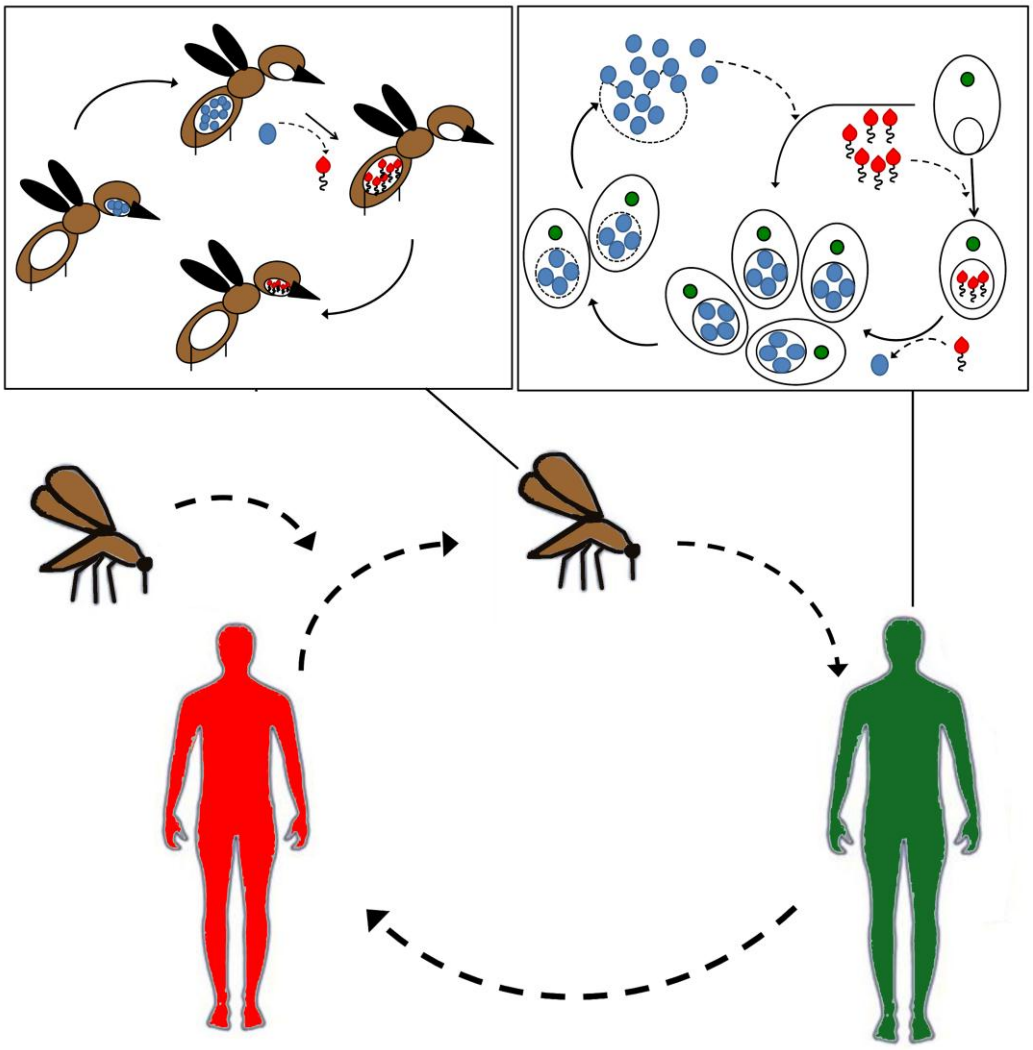

Figure 1.1. The lifecycle of the parasite – The parasite spreads from infected human (red) to susceptible human (green) by bites of female sandflies. Within the newly infected human, the promastigote stages (red with black tails) that are released in blood during the bite are engulfed by the macrophage (right panel). The promastigotes differentiate into the amastigote (blue ellipsoid) within macrophage phagolysosome. The macrophages in the vicinity are infected by the amastigotes. The phagolysosome disrupts along with the macrophage membrane (dotted outline), infects new uninfected macrophages and re-populates. When a susceptible sandfly bites an infected human, the amastigotes along with blood are engulfed by the sandfly. From the proboscis, the amastigotes travel to the midgut, where they grow and differentiate into promastigotes (left panel). The promastigotes being motile travel upstream back to the proboscis of the sandfly ready to bite a susceptible human and continue infection.

On the other hand, distinct species heterogeneity is demonstrated in clinical manifestations within the host. Cutaneous leishmaniasis is largely caused by *Leishmania major*, *Leishmania braziliensis* (muco-cutaneous), *Leishmania mexicana* (diffuse cutaneous), and *Leishmania amazonensis*; although, under cases of under-represented host immunity, visceral leishmaniasis can also occur. On the contrary, visceral leishmaniasis is caused

mainly by 2 species - *Leishmania infantum*, and *Leishmania donovani*. This heterogeneity in clinical presentation of different *Leishmania* species in the host has been attributed to 2 major frontiers (Fig, 1.2), (McCall et al. 2013):

A) The parasite frontier: This frontier argues that the adaptive species-specific phenotype of the parasite within the host is responsible for species-specific clinical manifestation. The parasite context of clinical presentation can be broadly understood through – i) the identification of species-specific genes and their role in parasite survival within the host; ii) differences in expression of the conserved genes between parasite species and iii) differential behavior of biochemical pathways.

B) The host frontier: Host genetic variability and generation of differential immune responses uniquely against a specific parasite species is responsible for species-specific clinical manifestation. The host context of clinical presentation is largely understood with respect to the interaction of parasite molecules with the host signaling pathways to suppress host immunity to disease.

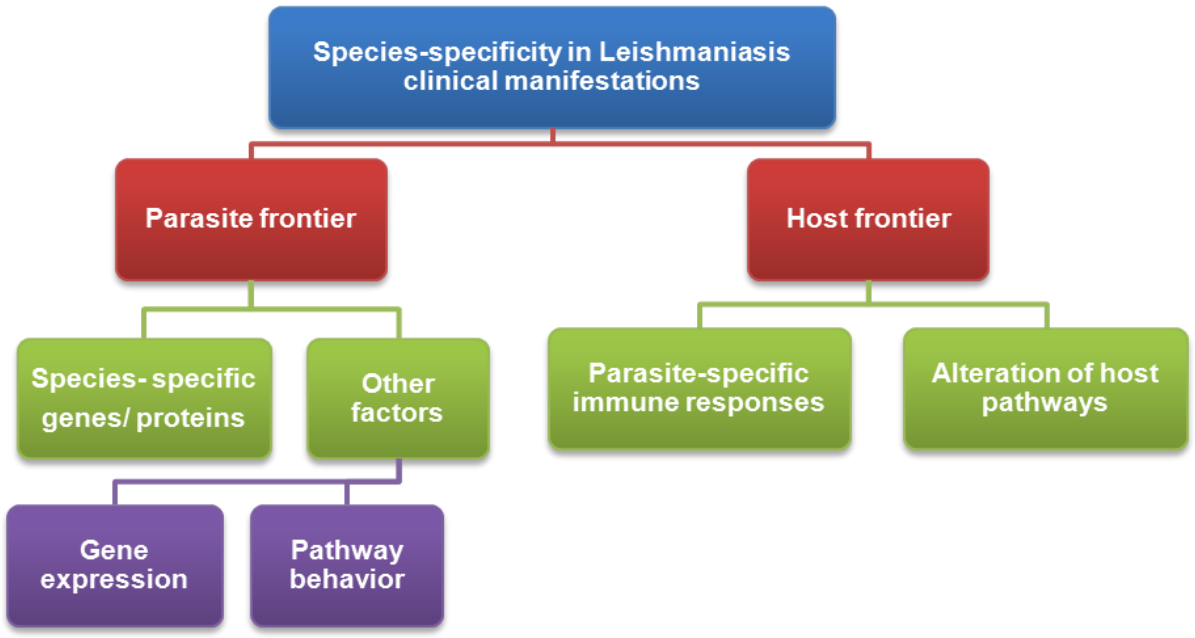

Figure 1.2. Perspectives of species-specific heterogeneity in clinical manifestations of Leishmaniasis - The parasite frontier discusses about the parasite-specific factors that promote survival within the host, whereas host frontier discusses about the host being able to detect different *Leishmania* species uniquely.

The main focus of the thesis is to identify the stage-specific and species-specific determinants that can employ the parasite with novel survival strategies of adaptation within the host. As the disease involves complexity at many layers, a systems-level perspective that

combines the viewpoints of the genotype and phenotype is very much essential. Hereafter, the genotype refers to the genome-based features like GC content, preferred usage of codons within the genes of the *Leishmania* genome, gene length, etc. and the phenotype specifically refers to the distribution of metabolites (flux) across metabolic pathways in a given *Leishmania* genome. The underlying mechanisms governing the genotype and phenotype characteristics can play an important role in uncovering the stage-specific and species-specific determinants of Leishmaniasis. Also, a significant overlap observed in the experimental data obtained for both stage and species-specificity highlights the requirement of a large-scale study that can unify various levels of information to study the successful survival strategies employed by the parasite within the host. With this pretext, in this chapter, we provide the review of the previous studies that have attempted to understand the determinants of parasite survival within the host at different levels, while providing a mechanistic basis, which forms the background for the research work underlying this thesis. These are discussed in a chronological order with respect to their nature and time of discovery. This chapter will also highlight the drawbacks of different technologies, novel discoveries of mechanisms and the non-obvious nature of the infection that makes it complex. This chapter will also emphasize the need of novel *in-silico* systems-based investigations for identifying the counter-intuitive mechanisms of adaptation which otherwise remain inconspicuous.

Numerous studies have been independently performed to identify the differences observed between different stages of the parasite lifecycle and across *Leishmania* species. Most of these observations were the results of different 'omics' approaches employing a parasite-specific protocol, which became only recently possible.

1.1. Identification of species-specific genes

Around 7 finished *Leishmania* genomes have been published and are publicly available till date, out of which 6 *Leishmania* species infect humans (El-Sayed et al. 2005; Ivens et al. 2005; Peacock et al. 2007; Downing et al. 2011; Raymond et al. 2011; Rogers et al. 2011; Real et al. 2013). Comparative genomics approaches applied to the completed *Leishmania* and Trypanosomatid genomes provided a vast variety of useful information regarding the architecture and content of the *Leishmania* genome (El-Sayed et al. 2005). The *Leishmania* genome is a highly intact genome and does not display major variations in organization or nucleotide content between different *Leishmania* species and maintains a conserved synteny

(El-Sayed et al. 2005; Smith et al. 2007). The genome structure of *Leishmania* like other Trypanosomatids consists of large collinear polycistronic cluster of genes distributed within a single strand of DNA commonly termed as directional gene clusters (DGCs) exhibiting a very few spliceosomal introns (Ghedin et al. 2004; El-Sayed et al. 2005; Peacock et al. 2007; Smith et al. 2007).

Architecturally, in comparison to other sequenced Trypanosomatid genomes, the *Leishmania* genomes lack extended subtelomeric regions that have been implicated in closely related *Trypanosoma* genomes to contain species-specific genes (Peacock et al. 2007). Also, to supplement the choice of retaining a conserved genome architecture, *Leishmania* genomes also lack transposable elements; an exception being *L. braziliensis* genome, which encompasses two types of transposons (TATEs) and retroposons (SLACS) that are absent in other *Leishmania* genomes (Peacock et al. 2007). Only fragments of ingi/L1Tc related transposons known as DIREs (degenerate ingi/L1Tc-like retrotransposable elements) are found to be interspersed in the non-coding regions of the *L. major* and *L. infantum* genomes (Peacock et al. 2007; Smith et al. 2007). Accordingly, only a few species-specific genes could be identified between the different *Leishmania* species; the reason for this difference being pseudogene formation in one genome as compared to presence of a functional gene in another genome. The numbers related to the genomic content between different sequenced *Leishmania* genomes are given in Table 1.1.

Table 1.1. Statistics of the differences in genomic content of 6 *Leishmania* species [Data obtained from TriTrypDB version 4.0 (Aslett et al. 2010)].

Species	Total gene count	Protein-coding genes	Non-coding genes	Pseudo-genes	Mega base pairs	Chromo-somes
Leishmania braziliensis MHOM/BR/75/M2904	8505	8357	148	156	32.09	35
Leishmania donovani BPK282A1	8195	8083	112	51	32.44	36
Leishmania infantum JPCM5	8381	8239	142	59	32.13	36
Leishmania major strain Friedlin	9378	8400	978	91	32.86	36
Leishmania mexicana MHOM/GT/2001/U1103	9063	8250	813	99	32.11	34
Leishmania tarentolae Parrot-TarII	8530	8452	78	0	31.63	36

As discussed above, the polycistronic gene groups remain conserved across species with around 8000 - 9000 genes distributed across 36 chromosomes. The variations in the numbers are only due to duplication of genes within a genome. Comparison of sequenced *Leishmania* whole genomes led to prediction of 19 *L. infantum* specific, 14 *L. major* specific,

67 *L. braziliensis* specific and 2 *L. mexicana* specific genes or paralogous groups suggesting a remarkably low number of genes unique to one species when compared with the coding capacity of the genomes (Rogers et al. 2011). Out of the above genes only a few genes have experimentally been proven to be functional species-specific virulence factors. Examples include the A2 gene family that plays a key role for survival of visceral *Leishmania* species in visceral organs (Zhang and Matlashewski 2001) and a hypothetical cytosolic protein (LinJ.28.0340) whose knockout led to decreased parasite survival in visceral organs (Zhang and Matlashewski 2010).

1.1.2. Gene copy numbers

Gene copy numbers have been postulated to control gene dosage and expression in many eukaryotic genomes (Birchler et al. 2005; Veitia et al. 2008). The possible diversity observed between the *Leishmania* genomes was attributed to events like gene duplication that cause dispersion of gene copies within the genome and pseudogene formation that leads to silencing of gene function (Mannaert et al. 2012). A high percentage of genes are represented in the *Leishmania* genome as tandem arrays of genes. A comparative genomics approach between the sequenced *Leishmania* genomes distinctly identified significant variation in gene copy numbers between 4 completely sequenced *Leishmania* genomes – *L. major*, *L. braziliensis*, *L. infantum*, and *L. mexicana* (Rogers et al. 2011). The gene copies in this study was found out through comparisons between the sequenced genomes and verified by the read depth coverage measured while sequencing the genomes. The total number of arrays considerably varied between the species – 200 in *L. major* Friedlin, 207 in *L. infantum* JPCM5, 214 in *L. braziliensis* M2904 and 132 in *L. mexicana* U1103. In *Leishmania donovani*, gene copy number was also shown to strongly correlate with transcription levels suggesting that there might be important functional consequences of copy number leading to variability in mRNA expression (Downing et al. 2011).

1.2. Regulation of gene expression

Paucity in the number of species-specific genes within *Leishmania* genomes and a frequent observation of gene duplication and gene loss, led to a broad speculation that the diversity in stage-specific and species-specific differences within and between *Leishmania* species is controlled at the gene expression level. Post-transcriptional processes like pre-mRNA processing and mRNA degradation were believed to cause significant variations in

transcriptome levels (Haile and Papadopoulou 2007). In Trypanosomatids, the pre-mRNA processing step consists of trans-splicing which involves the binding of a splice-leader RNA (SL-RNA), a 39-41 long nucleotide transcript having a hyper-modified capped structure to the pre-mRNA (polycistronic transcription) and a 3' polyadenylation, both occurring simultaneously (Liang et al. 2003). When these capped mRNAs are released into the cytoplasm, they are either translated or degraded. mRNA degradation in Trypanosomatids is an elaborate mechanism where the degradation occurs either immediately at the 5' and 3' ends (regulated degradation) or after de-capping and poly-deadenylation. Hence, the rate at which the 5' cap or poly A sites are added to the mRNA and degradation rates of unstable mRNAs being relatively high to that of stable mRNAs might cause differential gene expression. As an application, the SL-RNA addition step has been used as a primer for second strand synthesis in an mRNA sequencing protocol (Rastrojo et al. 2013).

These *Leishmania* specific trans-splicing and poly-adenylation signals can be predicted through *in-silico* techniques (Smith et al. 2008). As mentioned before, the stable genomes of *Leishmania* display a lack of retrotransposable elements (TEs) within their genomes and only fragments of ingi/L1Tc related transposons known as DIREs (degenerate ingi/L1Tc-like retrotransposable elements) were found to be interspersed in the non-coding regions of the *L. major* and *L. infantum* genomes (Smith et al. 2007). In *L. infantum* promastigotes, axenic amastigotes and intra-macrophage amastigotes, the SIDER1 family of TEs complemented by the 3' U-rich elements (UREs) was shown to stimulate translation initiation of amastigote specific mRNA in response to heat shock indicating its possible role in stage specific gene regulation (Haile and Papadopoulou 2007). While on the other hand, the SIDER2 family of TEs promotes destabilization of mRNAs and hence, their constitutive degradation (Bringaud et al. 2007). A large *in-silico* pipeline was recently proposed to identify cis-regulatory elements among conserved intercoding regions (LeishCICS: *Leishmania* conserved inter-coding sequences) in 3 *Leishmania* genomes – *Leishmania braziliensis, Leishmania major* and *Leishmania infantum* (Vasconcelos et al. 2012). The study could predict putative regulons, mainly comprising of the SIDER family that can possibly affect *Leishmania* gene expression.

Through microarray studies, up-regulation of amastin proteins, down-regulation of glycolytic enzymes, translation-associated genes and cell-cycle genes were identified to be specific to amastigotes commonly across species (Duncan et al. 2004; Holzer et al. 2006; Cohen-Freue et al. 2007). Gene expression comparison performed between promastigote-amastigote transcriptome profiles of *L. major*, *L. infantum*, and *L. braziliensis* led to

identification of 570-620 differentially regulated transcripts between the compared species (Depledge et al. 2009).

Although subtle differences were found in transcriptome profiles across stages, a study suggested that the *Leishmania* genome (tested in *L. major*) was constitutively expressed at the transcriptome level between the developmental stages of the parasite (Leifso et al. 2007). Even with an advanced RNA-seq protocol, only 14% of the total transcripts in *L. mexicana* demonstrated a greater than 2-fold difference between promastigote and amastigotes (Fiebig et al. 2015). Furthermore, a combined transcriptomic-proteomic study identified that variations in expression of only around 10-12% mRNA correlated with variations in the effective protein expression (Lahav et al. 2011), with changes in mRNA levels observable only during early stages of differentiation. These observations point towards the obvious weak degree of regulation exerted on the gene-expression phenotype.

1.3. Translation

The weak degree of regulation at the transcriptome pointed towards regulation occurring in protein abundances. A few mass spectrometry-based proteomics studies attempted to quantify the relative abundance of proteins among the two developmental stages. Two independent studies on the *L. donovani* proteome were able to detect only about 21 - 50 % of the total proteome, out of which 613 proteins (7.6% of the total proteome) unique to the amastigote and 1090 (13.5% of total proteome) unique to promastigotes (Rosenzweig et al. 2008; Nirujogi et al. 2014). Few glycolytic enzymes, TCA cycle enzymes, and flagellar proteins tend to be uniquely present in promastigotes, whereas enzymes of fatty acid β-oxidation, cytoskeletal proteins, translation-related enzymes and many amino acid transporters were uniquely detected in amastigotes (Rosenzweig et al. 2008; Paape et al. 2010). In *L. infantum*, only around 70 differentially expressed proteins could be detected through quantitative mass spectrometry (Brotherton et al. 2010). Among these, most of the differentially expressing proteins belonged to energy metabolism, amino acid metabolism and gluconeogenesis which specifically increase in amastigotes. As mentioned previously, mRNA levels vary only during early stages of differentiation (probably due to post-transcriptional processing and degradation); whereas translational and post-translational regulation in relatively varying protein abundance predominantly play a role in later stages of differentiation (Lahav et al. 2011). The detection of a low number of differentially regulated or stage-specific proteins can be argued upon either as a limitation of existing proteomics technologies or as a relatively

unchanged proteomic profile across stages. This stresses upon the dire need for more sensitive techniques of detecting proteins and their coverage. This also suggests that relatively small changes in protein abundances might possibly bring about a major systemic change while differentiation.

Very little is known about the translational regulatory mechanisms in *Leishmania* parasites. In *Leishmania infantum*, it was demonstrated that eIF2α phosphorylation provides translational control during its differentiation (Cloutier et al. 2012). Differentiation of promastigote to amastigote triggers eIF2α phosphorylation thereby decreasing global translation initiation. Also, it was observed that, in kinetoplastids - *Leishmania tarentolae* and *Phytomonas serpens*, triplets just before the start codon site named the pre-ATG triplets complemented by the post-ATG coding region of variable length together mediate gene expression of EGFP by inhibiting translation initiation in *Leishmania*-transfected constructs (Lukes et al. 2006). It can be hypothesized that probable mRNA secondary structures like hairpin loops develop around the ATG region as seen in yeast (Tuller et al. 2010) that causes delayed translation initiation thereby regulating translation efficiency; although, the universal presence of these secondary structures across genus *Leishmania* is still left to be demonstrated.

1.3.1. Codon usage as a mechanism of translation regulation

In other eukaryotes, it has been observed that efficiency of translation elongation can be governed by the codon usage bias (CUB) of a particular gene. Codon usage bias (CUB) essentially depicts the unequal usage of synonymous codons in genes where the preferred codon in the sequence primarily correlates with tRNA isoacceptor abundances. CUB largely explains selection for the translational efficiency of a gene; as an unequal choice of codons in a particular gene can affect the rate of translation elongation of that gene (Shah and Gilchrist 2011; Novoa and de Pouplana 2012). Although translation initiation has been proven to be the rate-limiting step that controls differential protein synthesis, codon bias in genes are demonstrated in many eukaryotes to correlate with protein expression levels probably indicating its role in translation efficiency under no or low influence of translation initiation regulation (Ghaemmaghami et al. 2003).

Very few codon usage bias studies have been performed on Genus *Leishmania*; most of these studies focusing on understanding the effect of selection or mutation bias on the *Leishmania* genes. The earliest study performed for 84 gene sequences from *Trypanosoma brucei*, 47 from *T. cruzi*, 12 from *Crithidia fasciculata*, 13 from *Leishmania major*, and 21

from *L. donovani* complex indicated that *Leishmania* genes being GC-biased, experience a high mutational pressure to maintain G or C at the 3rd (Wobble) position of the synonymous codon when compared to Trypanosomatid genomes (Alonso et al. 1992; Alvarez et al. 1994). Also, translation selection was demonstrated to be the most important force acting on codon usage within the Trypanosomatid genomes supported by a significant positive correlation of *T. cruzi* protein expression with the codon adaptation indices (CAI) of corresponding gene (Horn 2008). A comparative multivariate analysis of codon and amino acid usage performed in *L. major*, *L. infantum* and some genes of *L. donovani* revealed that putative lowly expressed genes might demonstrate a mutational bias towards A/T at the third position as compared to putative highly expressed genes that show mutational bias towards G/C justifying selection pressure for GC at third position of synonymous codon (Chauhan et al. 2011; Singh and Vidyarthi 2011). Another study that considers comparison of *L. major*, *L. braziliensis* and *L. infantum* have also demonstrated the effect of translation selection acting on *Leishmania* coding sequences by correlation of relative codon frequency with tRNA gene copy number and visualizing patterns of codon usage bias in putative high and low expressed genes (Rashmi and Swati 2013).

Although the above studies largely described the evolutionary forces of mutation and selection acting on the coding sequences of *Leishmania*, comparative codon usage analysis was performed only on a small sample of genes from very few genomes. The comparisons of CUB measures for enumerating the role of mutation pressure and translation selection along with comparison of putative codon usage (global expression) profiles in genes between the 6 sequenced *Leishmania* genomes and other also with other Trypanosomatid species still remains to be performed. As mentioned before, mRNA secondary structures at the 5' end of translation start site exist in certain genes of *L. tarentolae* that significantly reduce gene expression (Lukes et al. 2006). The choice of codons at the translational start site affected by secondary structure of mRNA thus, can further help us to understand the role of codon usage bias, if any, on translational regulation in *Leishmania*. Further, none of the aforementioned studies also investigate the role of amino acid compositional bias on codon usage.

1.4. The *Leishmania* interactome

Proteins do not function in isolation, but in context with other proteins by interacting with them. Protein-protein interactions (PPIs) govern a wide scale of biological process like cell-cell interactions & communication, metabolic and developmental regulation (Rao et al.

2014). High throughput protein-protein interaction experimental studies that can be used to trace interactome maps are still lacking in *Leishmania*; although certain experimental protein-protein interaction studies like co-immunoprecipitation have been performed for a specific set of proteins (parasite-parasite or parasite-host protein pairs) (Peters et al. 1997; Abu-Dayyeh et al. 2008; McCall and Matlashewski 2010).

Due to experimental limitations, computational techniques for the prediction of possible interactions were used. Florez et. al. presented the first PPI network study in *Leishmania major* (Flórez et al. 2010). The PPI network consisted of 1336 nodes and 33861 interactions. Network analyses could predict important drug targets that were essentially represented as hubs in the PPI network. The most important output of the work was that functional clusters were identified within the network and coupled with GO enrichment for biological process to annotate the pathway-level function. Another comprehensive study predicted PPI networks for *L. braziliensis*, *L. major*, and *L. infantum*, each network with a large set of 39,420, 43,351 and 45,235 predicted interactions respectively (Rezende et al. 2012). Extending the previous interactome study, this study also attempted to annotate around 50% of the total hypothetical proteins for their biological process along with prediction of important drug targets by exploring the phenomenon of network modularity. Both of these studies use a combination of tools and databases to identify interactions on the basis of existing homologues with known interacting partners in other closely related organisms. The predictions provided by both these studies are yet to be confirmed through wet-lab experiments. Nevertheless, these predictions indicate that the species-specific differences do not occur simply at the level of gene or protein expression, but at the system-level which can also comprise of the interacting phenotype.

1.5. The *Leishmania* metabolome

As introduced previously, *Leishmania* parasite differentiation from the promastigote to amastigote is largely triggered by sudden changes in pH, temperature, and carbon sources between these hostile environments (Zilberstein and Shapira 1994; Saunders et al. 2014). The sandfly midgut represents a predominantly alkaline environment abundant with glucose and other carbon sources whereas the macrophage phagolysozome is a highly acidic environment with scarcity of glucose, amino acids and a relatively higher availability of fatty acids (McConville and Naderer 2011). Also, the sandfly gut is characterized by a low temperature as compared to the human macrophage.

It can be presumed that the parasite across developmental stages might dynamically fluctuate its metabolism by varying the expression of certain metabolism genes to cope up with these extreme environments, thereby providing a relatively insensitive but flexible adaptation. Despite the amount of available information related to the gene/protein expression profiles of promastigotes and amastigotes, a very few studies exist that accurately delineate the molecular events that might be decisive for the parasite to adapt to different environments. Using a time-series mass spectrometric approach, Rosenzweig et. al. captured the expression profiles of around 21% proteins of the *L. donovani* proteome at different stages of parasite differentiation, eventually discovering that *L. donovani* promastigotes retool their metabolism by changing expression of few proteins during their transition into amastigotes (Rosenzweig et al. 2008). Proteins belonging to fatty acid oxidation, TCA cycle, and mitochondrial oxidative phosphorylation were significantly upregulated in the initial phase of amastigote differentiation demonstrating the shift of the promastigotes from utilizing glucose to amastigotes that utilize fatty acids. Similarly, gluconeogenic enzymes - phosphoenolpyruvate carboxykinase and fructose-1, 6-bisphosphatase were also upregulated indicating the utilization of glycerol and amino acids to form sugars. Also, a decrease in the expression of cytosolic glycolytic enzymes in the amastigotes suggested a hurdle for utilization of glucose from the environment and increase in dependency of the amastigotes on other carbon sources for their survival and proliferation. Two important studies through stepwise ^{13}C labeling of different carbon sources in promastigotes and amastigotes of *L. mexicana* have also explored the metabolic routes of the parasite metabolome (Saunders et al. 2011; Saunders et al. 2014). The results of these studies exhibited the essential role of succinate fermentation, glutamate biosynthesis and TCA anaplerosis in maintaining parasitic energy metabolism in promastigotes (Saunders et al. 2011). This was further corroborated in axenic amastigotes indicating the presence of conserved catabolic routes existing across stages (Saunders et al. 2014). The reduction in glucose and amino acid uptake rate, presence of an active TCA cycle and role of glutamate biosynthesis in energy metabolism are important characteristics that induce a stringent metabolic response in the intracellular amastigotes (Saunders et al. 2014). The observations suggest that parasite metabolism responds similarly in both stages with flux changes between pathways that can be attributed only to the changes in uptake and intracellular availability of specific metabolites. Apart from stage-specific comparisons, a comparative metabolomics approach carried out across *L. major*, *L. donovani* and *L. mexicana* identified that metabolism of essential amino acids like arginine, phenylalnine and tyrosine also varies across species (Westrop et al. 2015).

In silico computational approaches like metabolic reconstruction accompanied by flux balance analysis aim in a similar way to discover strategic metabolic routes adopted by the parasites in order to ensure flexible adaptation within the host. Apart from the above *in vivo* studies, genome-scale reconstructions of parasites belonging to two *Leishmania* species have been proposed to understand the stage-specific routes of metabolite utilization. The first study proposed a *L. major* specific reconstruction that accounts for 560 genes, 1112 reactions, 1101 metabolites and 8 subcellular compartments (Chavali et al. 2008). The model was analyzed using a constraint-based flux balance analysis approach. Important results of their study include annotation of a number of hypothetical proteins for their putative function in metabolism, discovery of an *in silico* minimal medium essential for parasite growth and identification of a comprehensive set of drug targets obtained from single and double reaction knockout analyses. As a follow-up study, the model predicted drug targets were also tested for the efficacy of FDA approved drugs in inhibiting them (Chavali et al. 2012). Recently, a genome-scale reconstruction of *Leishmania donovani* that accounts for 1159 reactions, 1135 metabolites and 604 genes also became available (Sharma et al. 2017). Similar to the earlier study, the metabolic network reconstruction was employed as a constraint-based model to identify synthetic lethal gene combinations that can be probable drug targets.

But, the aforementioned studies have restricted only towards primary reconstruction of the model and a routine analysis to identify probable drug targets giving less emphasis on the choice of common metabolic routes in the two stages as observed in metabolomics studies, the core enzymes that dictate them and the organization of metabolism. Furthermore, no genome-scale reconstruction for the well-understood *L. infantum* JPCM5 genome was proposed till date.

1.6. Genotype - phenotype relationships

As discussed above, *Leishmania* resides in a constrained environment within which it has adapted to survive. From metabolomics studies, it became clear that across different environments (sandfly and human) there are common metabolic pathways that work together for optimal survival of the parasite. Being an intracellular parasite, it needs to overcome resistance mechanisms posed by the host metabolic microenvironment for survival. Accordingly, the well-adapted parasite metabolism can be clearly subjected to relatively strong selection pressure. This probably imparts the parasite with mechanisms that can combat against any random perturbations. The reason for this incapability to eliminate the

parasite is due to the incomplete understanding of parasite metabolism and its ability to adapt to different stress conditions. Also, the role of the genotype in establishing an adaptive phenotypic landscape is still not appropriately studied in *Leishmania*. These diverse problems can be addressed only by analyzing the genotype and phenotype factors or their combinations that determine the evolution of metabolic enzymes under environmental pressure.

With the advent of evolutionary systems biology approaches (Koonin and Wolf 2006; Papp et al. 2011), the effect of the metabolic phenotype on enzyme evolution of organisms can further help to understand the coordinated functioning of metabolism so as to adapt to the environmental pressure. Previously, a large number of topology-based studies have been used to infer evolution of enzymes in metabolic networks of bacteria and yeast (Liu et al. 2007; Copley 2012). Also, from the function or behavioral point of view, few studies have also attempted to understand the role of flux carried by an enzyme and its role in enzyme evolution (Vitkup et al. 2006; Lu et al. 2007; Olson-Manning et al. 2012; Colombo et al. 2014). The aforementioned studies have been largely carried out in free-living species that experience different types of environmental stresses as compared to parasites which live in a constrained and less variable, nutrient-deprived stress environment. But, specifically for *Leishmania*, the genotype and phenotype has largely been analyzed independently without considering an explicit evolutionary linkage between them.

The above review suggests that there are many gaps in the understanding of how the parasite survives within its two hosts and the species-specific factors that expands the diversity of survival mechanisms. This further emphasizes the primary need of a unified understanding of the above discrete events considering the information available for different *Leishmania* genomes and integration of different types of available genotypic and phenotypic information for identification of the survival strategies employed by the parasite.

1.7. Organization of the thesis

In the present thesis, we have used bioinformatics and computational systems biology approaches to study the multi-level organization of mechanisms that govern the genotype-phenotype characteristics of the *Leishmania* parasite. The variations in the *Leishmania* genotype were studied initially by performing a comparative genomic approach. The variations in the metabolic phenotype were studied with respect to both stages and species. Lastly, the genotype and phenotype are integrated to predict evolutionary rates of metabolic genes in different *Leishmania* species.

With respect to this, the thesis is organized into six different chapters.

Chapter 1 or the present chapter provides a review about the previous studies that have attempted to understand the stage-specific and species-specific determinants of parasite survival within the host at different levels with an emphasis on the underlying mechanisms, which provides a background for the research work underlying this thesis.

Chapter 2 provides the methodological overview of the computational and statistical methods used to analyze the heterogeneous genotype and phenotype data obtained from different *Leishmania* species. This chapter encompasses all the methods used to analyze data discussed in the following chapters.

Chapter 3 attempts to study the causes and the consequences of codon usage bias in regulation of gene translation in *Leishmania*. This chapter provides insights into the roles of mutational pressure, translational selection and amino acid composition bias as causes of codon usage bias and the formation of mRNA secondary structures at the 5' end of the mRNA as a mechanism of translation regulation by comparison of codon usage profiles across *Leishmania* species.

Chapter 4 discusses the two new reconstructions of energy and genome-scale metabolism of *Leishmania infantum* JPCM5. Constraint-based analysis of these reconstructions provides insights into the organization of parasite metabolism in context of adaptation to the two host environments, the constraints on the choice of common metabolic routes of utilization and their role in parasite surivival within host metabolic microenvironment.

Chapter 5 investigates the relationships between the genotype & metabolic phenotype and their effects on the evolutionary rates of metabolic enzymes. This chapter attempts to define a logical relationship between the genotype and phenotype characteristics of the parasite, with the identification of gene-specific factors that affect enzyme evolution in *Leishmania* metabolism across different species.

Chapter 6 summarizes the most important observations provided in the previous chapters that identify the novel stage and species-specific factors and mechanisms. This chapter also provides new directions that can build upon the research avenues provided in this thesis.

Chapter 2 - Methodology

2.1. Comparative Codon usage analysis

With respect to the central dogma of molecular biology (Fig. 2.1A), the gene is transcribed into mRNA by transcription (Crick 1970). This mRNA contains the encoded information of the function of the gene. By the process of translation, the encoded information within the mRNA is translated into a sequence of amino acids, which further form the protein that govern that function. Codons are triplet nucleotides present within the coding sequence/mRNA of a gene that code for an amino acid. It is often observed that there is degeneracy in the choice of the codons coding for a particular amino acid. There are amino acids that are either encoded by 2, 3, 4 or 6 codons. This association of the amino acid and the codon is proposed by a codon usage table, which in case of *Leishmania* genomes follows the standard genetic code (Fig. 2.1B). Furthermore, it is also commonly observed that certain codons are chosen with a higher frequency to code for an amino acid as compared to other alternative codons (Plotkin and Kudla 2010). This phenomenon is broadly termed as codon usage bias (CUB). Evolutionary events like mutation pressure and translation selection are the most explored causes of CUB in many organisms. Mutation pressure suggests that codons with a high occurrence of specific nucleotides are chosen within a genome (Sueoka 1988). Translation selection suggests codons that benefit translational efficiency are chosen within a genome (Hershberg and Petrov 2008; Plotkin and Kudla 2010). The roles for mutation pressure and translation remained unknown for *Leishmania* and other Trypanosomatids. To investigate the causes and consequences of codon usage bias, a number of sequence-based measures were calculated and compared across 13 different species of Trypanosomatids to observe similarities and differences in codon patterns.

2.1.1. Dataset curation

For this study, the coding sequence (CDS) datasets from the available assembled genomes of 13 species of Trypanosomatids namely, *Crithidia fasciculata* Cf-cl (Cfas), *Leishmania braziliensis* MHOM BR75 M2904 (Lbraz), *Leishmania donovani* BPK282A1 (Ldon), *Leishmania infantum* JPCM5 (Linf), *Leishmania mexicana* MHOM GT 2001 U1103 (Lmex), *Leishmania major* Friedlin (Lmaj), *Leishmania tarentolae* Parrot Tar II (Ltar), *Trypanosoma brucei gambiense* DAL 972 (Tbg), *Trypanosoma brucei* TREU 927 (Tbt), *Trypanosoma congolense* IL3000 (Tcl), *Trypanosoma cruzi* CL Brener Esmeraldo-like (Tclbe), *Trypanosoma evansi* STIB805 (Tev), and *Trypanosoma vivax* Y486 (Tvx) in the TriTrypDB v.8.1 (updated as of 07-May-2014) were used for this study (Aslett et al. 2010).

*The bulk of this chapter has appeared in the Methods Section of the articles mentioned in the "Publications" list, co-authored by A. Subramanian and R. R. Sarkar

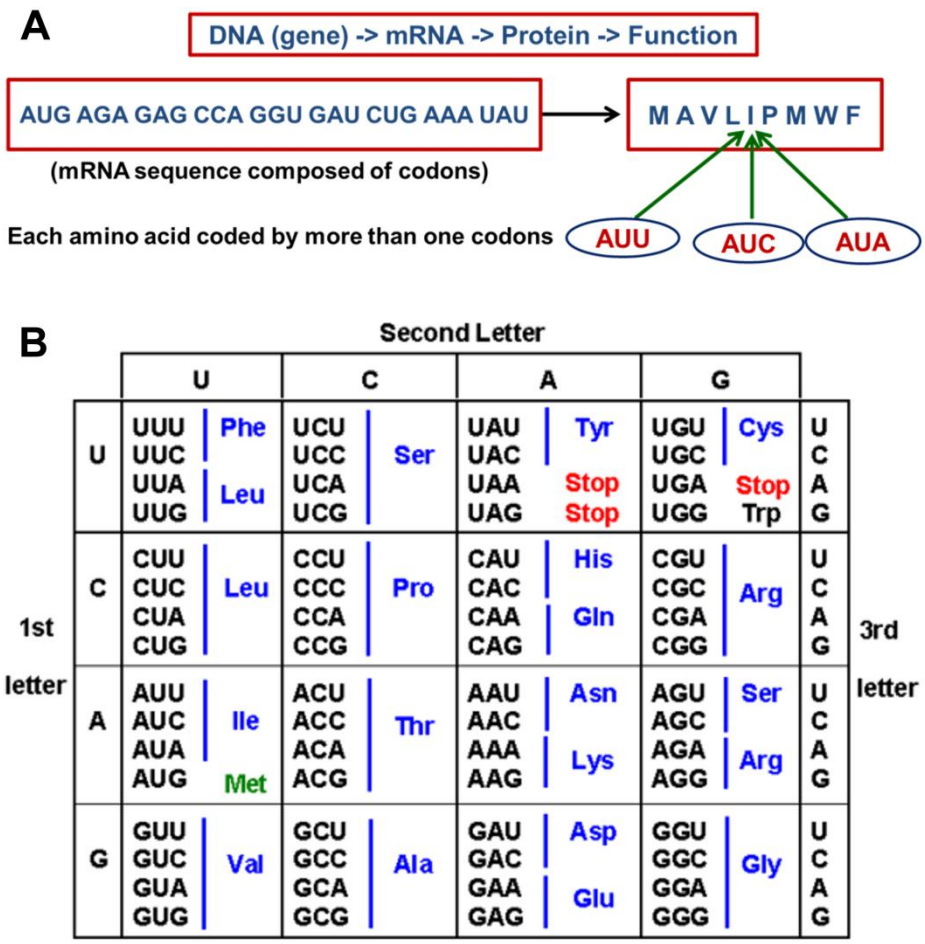

Figure 2.1. Degeneracy in the genetic code – A) Central dogma of molecular biology; B) Standard genetic code (*Courtesy*: https://openwetware.org/wiki/CH391L/S12/TranslationRBSandCodons).

The datasets were processed to remove sequences - that did not have a termination codon, that had more than one internal stop codons, and that were part of the mitochondrial genome. Also, in order to minimize the sampling error, sequences having more than 200 codons were only considered for the study. Proteome data extracted from amastigote stages of *L. mexicana* (Paape et al. 2010) was used for calculation of relative protein abundance. In these studies, 810 single copy genes out of the total 1666 genes in *L. mexicana*, for which proteins were detected in the aforementioned proteomic studies, were shortlisted for comparison with CAI.

2.1.2. Sequence-based measures of codon usage

To analyze the effects of different evolutionary events causing codon usage bias (CUB), diverse sequence-based measures were computed for the genes within the 13 Trypanosomatid genomes.

a) *Mutation pressure*: The mutation pressure towards maintaining AT or GC content in a genome can be calculated between a pair of orthologous genes (Sueoka 1988; Alonso et al. 1992); one gene of the pair belonging to the first genome and the second gene belonging to the second genome given by,

$$\text{Mutational pressure} = \frac{p}{q}, \qquad (1)$$

where p is the total number of substitutions of G or C to A or T between the pair of orthologous genes and q is the total number of substitutions of A or T to G or C between the same pair of genes. If p/q is greater than 1, mutation pressure towards AT is expected for the second gene as compared to first gene. If p/q is less than 1, mutation pressure towards GC is expected for the second gene as compared to first gene.

b) *GC content*: The frequency of G and C in different codon positions (GC1s, GC2s, and GC3s) was simply computed as the sum of counts of G and C nucleotides divided by the sum total of frequencies of A, T, G and C at the respective positions in the codons. GC content at different positions of the codon is given by,

$$GCn = \frac{\sum G_n + \sum C_n}{\sum A_n + \sum T_n + \sum G_n + \sum C_n}, \qquad (2)$$

where GCns is the frequency of the GC content at the n^{th} synonymous position of the codon (n=1, 2, 3) in a particular gene.

c) *Relative synonymous codon usage (RSCU)*: The relative synonymous codon usage (RSCU) is the observed frequency of a particular codon coding for an amino acid within a gene scaled by the number of synonymous codons for that amino acid (Sharp and Li 1987). RSCU is given by,

$$RSCU_{ij} = \frac{x_{ij}}{\frac{1}{k_i}\sum_{j=1}^{k_i} x_{ij}}, \qquad (3)$$

where $RSCU_{ij}$ is the relative synonymous codon usage for the j^{th} codon coding for the i^{th} amino acid, x_{ij} is the total count of codon 'j' coding for amino acid 'i', and k_i is the number of alternate codons coding for that amino acid.

d) *Codon Adaptation Index (CAI)*: Codon adaptation index is a widely used measure of codon usage that assesses the degree of translation selection acting upon a gene (Sharp and Li 1987). CAI is calculated for every gene relative to a known reference set of highly expressing genes (in this case, set of ribosomal genes) as a geometric mean of RSCU values. CAI is given by,

$$CAI = \exp \frac{1}{L} \sum_{i=1}^{18} \sum_{j=1}^{k_i} x_{ij} \ln(w_{ij}), \quad \text{where } w_{ij} = \frac{RSCU_{ij}}{RSCU_{i\max}} \qquad (4)$$

$RSCU_{imax}$ is the relative synonymous codon usage of the most frequently used codon for the ith amino acid and L is the length of the gene. The weight w_{ij} is calculated from a reference set of highly expressing genes. The values of CAI are scaled between 0 and 1. Further, CAI > 0.5 also, indicates a high degree of translation selection.

e) *Effective Number of Codons (ENC)*: The ENC index is a non-directional measure of codon usage bias which calculates the sum of contributions of codons for 2-, 3-, 4- and 6- fold degenerate amino acids to codon usage bias within a gene (Wright 1990). The ENC index has a strong relationship with the nucleotide composition at the 3rd synonymous position of the codon. The ENC index is calculated using,

$$ENC = 2 + \frac{9}{\bar{\hat{F}}_2} + \frac{1}{\bar{\hat{F}}_3} + \frac{5}{\bar{\hat{F}}_4} + \frac{3}{\bar{\hat{F}}_6}, \qquad (5)$$

where $\bar{\hat{F}}_i$ is the average codon homozygosity estimate for amino acids having degeneracy of '*i*' codons. The values of the ENC index are scaled between 20 and 61. A gene with ENC value of 20 indicates a high codon usage bias in the gene whereas an ENC value of 61 suggests an equal contribution of each codon within the gene to code for an amino acid.

f) *ENC for amino acid composition bias (N_C (AA))*: Not all amino acids within a degeneracy group tend to have equal amounts of frequent codons. Thus, an amino acid also, experiences a certain degree of bias exerted by biased codon usage. In other words, certain amino acid codons can be preferred over the others. To quantify this bias, the ENC index for individual amino acids called as N_c *(AA)* can be calculated (Fuglsang 2003).

$$N_c(AA) = \frac{1}{F_{AA}}, \text{ where } F_{AA} = \frac{(n\sum_{i=1}^{k} p_i^2) - 1}{n-1} \qquad (6)$$

F_{AA} is termed as the average homozygosity estimate for an amino acid where 'n' is the total number of codons that can possibly code for the particular amino acid in a gene and p_i is the frequency of i^{th} codon for that amino acid.

2.1.3. Codon usage and mRNA secondary structure formation

The mRNA folding energy profile was calculated for an overlapping sliding window of 20, 30 and 40 nucleotides run throughout each mRNA (coding) sequence in the *Leishmania* and *Trypanosoma* genomes. The folding energy was calculated using the minimum free energy algorithm, which yields a single optimal structure and its energy in kcal/mol (Zuker and Stiegler 1981). Further, average folding energy (in kcal/mol) was calculated for each gene in the 13 species by averaging the folding energy for all the overlapping windows in each sequence (of length 40).

2.1.4. Software and computational tools

RSCUs were calculated for each genome using an in-house PERL code. Hierarchical clustering of genomes using RSCUs was performed using Cluster 3.0 (Eisen et al. 1998). For finding GC to AT substitutions and vice versa, codon-based alignment followed by frequency calculations of the substitutions was performed. Single copy orthologous groups for identification of GC to AT substitutions and predicted GO process for the orthologous genes were extracted from TriTrypDB database v.8.1 (Aslett et al. 2010). The PRANK tool was used to perform codon-based alignment and its output was further processed to compute the average frequency of the substitutions (Löytynoja and Goldman 2008). CAI and ENC for each gene in every genome were computed using the EMBOSS package (Rice et al. 2000). $N_c(AA)$ was calculated using an in-house PERL code. For enumerating the mRNA folding energies, the RNAFold Software v.2.1.8 was used (Lorenz et al. 2011). Shuffling of each sequence (maintaining genome-specific CUB) was performed using the 'syncodon' model of the 'seqinr' package in R (Charif and Lobry 2007). Single copy orthologous genes across the *Leishmania* species for comparison of CAI values were identified using the Proteinortho orthology detection tool v.5.10 (Lechner et al. 2011).

2.1.5. Statistical analysis

Throughout the study to compare codon usage profiles at different levels, a number of statistical techniques were computed using R (Zar 1996). The Fisher's exact test was used to compare the codon RSCU values between the *Leishmania* and *Trypanosoma* groups.

Coefficient of determination (R^2) was used for finding goodness of fit between the codon GC content and average genome GC content, and also between CAI and relative protein abundance. For finding the association between ENC and CAI values Pearson correlation was used. To find out bias towards particular nucleotide, Poisson regression was performed between ENC and the individual bases (A, T, G, and C). All other correlations reported are the non-parametric Spearman correlation. For comparing the mean of randomized energy profiles of coding sequences with original profiles, a Kolmogorov-Smirnov test was performed (to identify whether both profiles belong to same distribution) followed by the Wilcoxon test (to identify whether the folding energy profiles are significantly different from each other) for each window index. To quantify variation among the CAI values of homologous genes in the 6 *Leishmania* species, coefficient of variation was calculated for each gene in the homologous set.

2.2. Metabolic network reconstruction

Metabolic reconstruction essentially includes the identification and association of enzymes to crucial metabolic reactions and their assignment to appropriate subcellular locations (Pinney et al. 2007). The genes and corresponding enzymes/proteins considered for model reconstruction were obtained for *Leishmania infantum* JPCM5 genome (assembly ID: GCA_000002875.2 ASM287v2) updated as of 16/12/2011. Each enzyme was annotated for its molecular function and subcellular location within the cell. The complete workflow for reconstruction is given in Fig. 2.2.

2.2.1. Annotation of molecular function

The molecular function of each metabolic enzyme was annotated through primary sequence search against databases like NCBI GenBank (Benson et al. 2008), UniProt (UniProt Consortium 2014), and KEGG (Kanehisa et al. 2013); motif/domain search against domain databases, like Pfam (Finn et al. 2013), and Prosite ; and manual curation through literature survey. The reactions catalyzed by the metabolic enzymes and their corresponding Enzyme Classification numbers (E.C Nos.) were curated from databases like BRENDA (Schomburg et al. 2012), Expasy Enzyme (Bairoch 2000), KEGG (Kanehisa et al. 2013). Few ambiguous protein sequences were also annotated for substrate specificity by comparison with homologues of known function in other species.

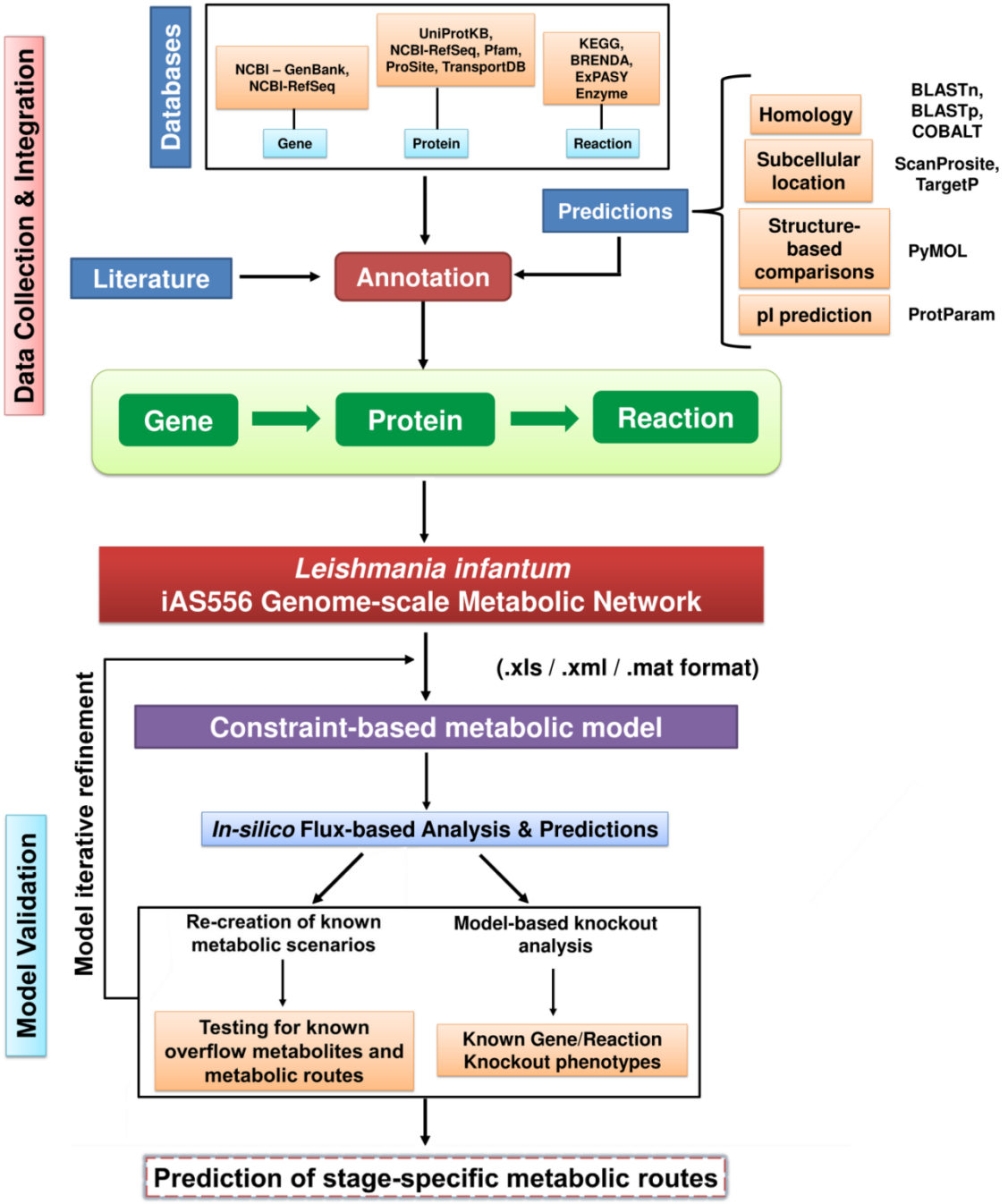

Figure 2.2. Strategy for reconstruction of the *Leishmania infantum* constraint-based models.

2.2.2. Assignment of metabolic enzymes to subcellular compartments

Most of the reactions were assigned to a particular cellular compartment on the basis of available evidence in *Leishmania* and other related Trypanosomatids. Alongwith literature evidence, the corresponding protein sequence was also perused for sequence-based subcellular targeting signals to affirm the subcellular location of an enzyme specific to *Leishmania infantum*. TargetP was used for predicting the presence of either a mitochondrial

target peptide or an endoplasmic reticular signal within a protein sequence (van Weelden et al. 2005; Emanuelsson et al. 2007). For transport of a protein to the glycosomes, presence of a peroxisomal targeting sequence (PTS-1 or PTS-2) is required (Borst 1986; Opperdoes and Szikora 2006). ScanProsite was used for predicting the presence of a PTS signal. PTS-1 can be detected by locating the PROSITE signature PS00342 within a sequence, which determines the requirement of presence of a particular tripeptide at the C-terminal of the protein sequence (De Castro et al. 2006; Opperdoes and Szikora 2006). PTS-2 was detected by finding a variable length PROSITE pattern <M-x(0,20)-[RK]-[LVI]-x(5)-[HKQR]-[LAIVFY] at the N-terminal of the protein sequence (Opperdoes and Szikora 2006). In order to further confirm the identified localization of proteins, isoelectric point for every protein was predicted *in silico* using the ProtParam Tool (Gasteiger et al. 2005). Almost all glycosomal proteins have isoelectric point (pI) in the range of 8.8-10.2 (Misset et al. 1986; Michels 1988; Guerra-Giraldez et al. 2002). For example, there are two gene copies encoding the asparaginase protein (UniProt IDs: A4HWE1, A4IDM2). The predicted pI of A4IDM2 was 9, whereas for A4HWE1 the predicted pI was 5.7. Based on this information, the two proteins, even though similar, were assigned glycosomal and cytosolic locations respectively.

2.2.3. Incorporation of transporters and exchanges

The transports and exchange reactions of previously reported carbon and nitrogen sources namely, carbohydrates, amino acids, fatty acids, amino sugars, purines and pyrimidines; fermentation products like pyruvate, lactate, acetate and succinate; vitamins folate, pantothenate, nicotinate, pyridoxine, protoheme and biopterin, other vital sources like iron, nitrates, nitrites, CO_2, H_2O, H_2S, HCO_3 and NH_3; and ions/protons were curated. Certain transport reactions could be associated with their corresponding genes/proteins in *L. infantum* by sequence comparison with annotated transporter genes of *Trypanosoma* and *Leishmania* available in the TransportDB database (Ren et al. 2006). Protein sequences with considerable sequence identity (> 40%) with known transporters were annotated accordingly. Some of the intracellular transport reactions occurring in membranes of organelles were curated as per reported observations (McConville et al. 2007; Saunders et al. 2011; Colasante et al. 2013; Saunders et al. 2014).

The glycosomal membranes of the trypanosomes are thought to be relatively impermeable to adenine nucleotides like ATP/ADP and NAD/NADH, thus maintaining their ratio inside the glycosome (Gualdron-López et al. 2012). Hence, their transport across

glycosomal membranes was not considered. In the iAS142 energy metabolic model, exchange and transport of acetyl-coA was assumed in the model to compensate for absence in uptake of fatty acids.

2.2.4. Confidence score

To pinpoint the reliability of every reaction considered in the model, a qualitative confidence score was proposed and assigned to every reaction. The confidence score is a five point score that unequally weighs the reliability of each model reaction, with respect to the subcellular location information obtained from literature and predictions from sequence analysis. This five point-scale is decided by three metrics, i) literature supporting the localization of a reaction, ii) confirmation through sequence analysis using ScanProsite (De Castro et al. 2006), and iii) confirmation through sequence analysis using TargetP (Emanuelsson et al. 2007). If and only if a literature finding clearly suggests about localization of a reaction, an arbitrary higher weight to the score (3) is given to that reaction. If the location of the enzyme and thus, the reaction was predicted solely through sequence analysis using ScanProsite and TargetP, then a score of 1 for each was assigned to that reaction. If the enzyme location could be identified with appropriate literature evidence and also through predictions by ScanProsite and TargetP, the confidence score of that reaction was considered to be the sum of the above individual scores [5 = 3 + 1 +1]. Thus, the highest score possible for a particular reaction would be 5 and the lowest score would be 1 depending upon the available information. For certain reactions, the subcellular location was not clearly identified through experiments, and is rationally assumed to occur in a particular subcellular location. One such reaction is glycosomal triose phosphate isomerase (TPIg), which is assumed from experiments to occur in the glycosome although the sequence-based subcellular location analysis does not indicate presence of any subcellular signals. For such reactions, a score of 2 was assigned so as to account for the uncertainty of its subcellular location through experiments.

2.2.5. Model iterative refinement for filling missing metabolic gaps

The primary metabolic networks formed in the previous step were iteratively refined to identify missing gaps by comparing with known observations in *Leishmania* literature. Applying constraints specific to *Leishmania* metabolism, the primary models were tested for predictions of known knockout phenotypes, prediction of overflow metabolites reported in *Leishmania* literature, matching predicted profiles with the experimental ^{13}C targeted metabolomics studies reported for related *L. infantum* or other *Leishmania* species (Saunders

et al. 2014), comparison with previously published *Leishmania* metabolic networks and qualitative mechanistic observations discretely reported in *Leishmania* literature. If there were discrepancies in predictions, the gaps were identified [either novel enzymes (maybe intracellular metabolic reaction or transport) or novel subcellular locations of existing enzymes] and filled. The modified network was again refined through the above procedure until the network is sufficiently able to reproduce the observations reported for *Leishmania* metabolism. Accordingly, reactions in appropriate subcellular reactions (for example, novel placement of fatty acid oxidation enzymes in glycosome for NAD redox coupling with glycolysis) and novel reactions (like threonine aldolase [UniProt ID: A4HRH1], which was previously annotated as an uncharacterized protein within the *L. infantum* JPCM5 genome) were identified.

2.2.6. Naming convention and integration of information

As per the established naming convention of curated metabolic models, the model name commences with an 'i' indicating *in silico*, followed by the first author's first and last initials ('AS') and followed by the number of genes that are part of the model. Hence, the energy metabolic reconstruction was named 'iAS142' and the genome-scale metabolic model of *L. infantum* was named 'iAS556'. The heterogeneous information related to the gene-protein-reaction (GPR) relationship as curated from above methodology, was organized in the rBioNet toolbox implemented in MATLAB R2012a to eventually obtain an organized model structure (Thorleifsson and Thiele 2011). The model was organized in a MATLAB structure (.mat) and also in the SBML (.xml) format (Subramanian and Sarkar 2017).

2.2.7. Flux Balance Analysis

The model structure contains a stoichiometric matrix *S*, a mathematical representation of the reconstructed network. The *S* matrix consists of *m* reactions and *n* metabolites. Each element in *Sij* represents the stoichiometric coefficient of metabolite *i* in reaction *j*. The coefficients are positive, if metabolite *i* is produced in reaction *j* and negative, if it is consumed. With respect to the metabolic information known for *Leishmania*, flux-based constraints can be applied to the reactions of the network. Apart from the applied exchange constraints, the inherent stoichiometry & reversibility of the reactions within the network and subcellular compartmentalization provide an important constraint to distribution of fluxes through reactions. A toy model and its stoichiometric representation is given in Fig. 2.3.

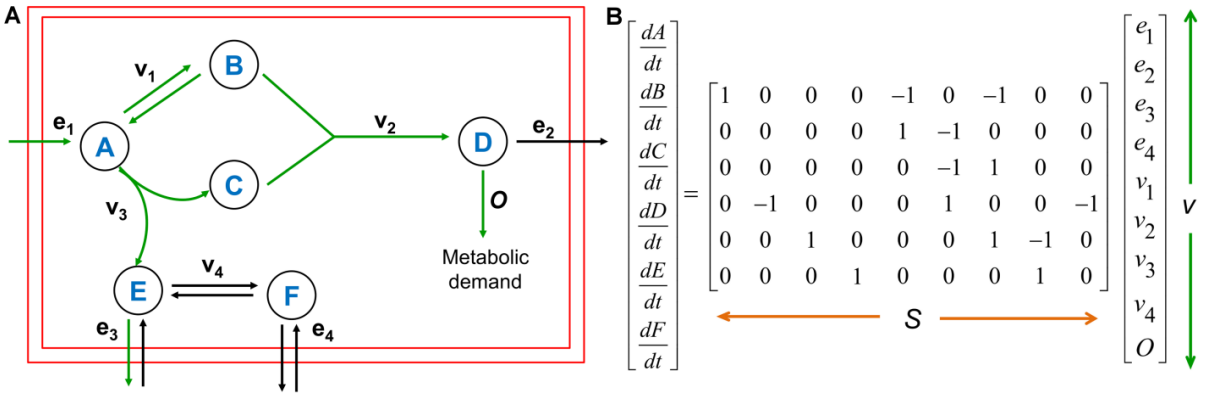

Figure 2.3. Stoichiometric representation of the metabolic network - A) A toy metabolic network demonstrating the inter-conversion of metabolites A, B, C, D, E and F through intracellular reactions. *O* represents the flux through the metabolic demand reaction, which represents the objective function; e_1, e_2, e_3 and e_4 represent the transport/exchange fluxes that either consume metabolite from environment or secrete metabolite into the environment; v_1, v_2, v_3 and v_4 represent the intracellular reaction fluxes. The green colored arrows represent the pathway of uptake and utilization of metabolite A through multiple reactions to satisfy the objective of maximizing metabolic demand; B) Decomposition of the rate of change of metabolites within the network into the *S* (stoichiometry) and *v* (flux) components. The stoichiometric matrix containing the participation of *m* metabolites (rows) in *n* reactions (columns) of the network.

The rate of change of concentration of every metabolite considered in the two *L. infantum* reconstructions can be represented as a function of individual reaction fluxes

$$\frac{dM}{dt} = S.v \qquad (7)$$

where, *M* = vector of metabolite concentrations; *S* = stoichiometric matrix of *m* rows of metabolites and *n* columns of reactions; *v* = reaction flux vector for *n* reactions (refer to Fig. 2.3 for an example).

Assuming the system to function at steady state, from Eq. (1),

$$\frac{dM}{dt} = 0 \text{ and } S.v = 0 \qquad (8)$$

As the nature of this system is underdetermined, FBA uses linear programming (LP) techniques to solve the above system of equations and identify a solution vector (flux distribution) in a particular constrained situation that would optimize a specific cellular objective. In both the iAS142 and iAS556 models, a maximization problem was formulated.

Note that each reaction flux v_i in the flux vector is constrained between bounds a_i and b_i such that, the LP optimization problem formulated for performing FBA was given by,

Maximize O, (9)

subject to constraints $S.v = 0$ and $a_i \leq v_i \leq b_i$,

where O = Objective function to be maximized,

and $O = c_1v_1 + c_2v_2 + \ldots + c_nv_n$

where $c_1, c_2, \ldots, c_n$ are stoichiometric coefficients of metabolites in demand reaction

and $v_1, v_2, \ldots v_n$ are the fluxes of model reactions that maximize the objective

a_i = lower bound of flux through every reaction i in the model and,

b_i = upper bound of flux through every reaction i in the model.

Uptake reactions were constrained irreversibly such that metabolites are consumed from the environment whereas secretions of metabolites are constrained such that metabolites are released into the environment. Exchanges for few metabolites that can be both utilized and produced (for example, alanine) were constrained reversibly. Uptake reactions in promastigote stage of the parasite were constrained such that all the uptakes of different carbon and nitrogen sources were sufficiently available whereas, the same uptake reactions in amastigotes were constrained to sub-optimal rates because uptake of glucose and amino acids is known to be reduced in this stage (Saunders et al. 2014). Also, in both stages, proline is typically formed from glutamate rather than uptake from the environment. The amastigotes are also characterized by hypoxic environmental conditions with reduced glucose, amino acid uptake and a unique fatty acid uptake chosen as an alternative to compensate for unavailability of glucose (Saunders et al. 2014). These constraints on the uptakes were applied so that fatty acids would be optimally selected to compensate for low glucose uptake in amastigotes and overflow metabolite secretion would be relatively high in promastigotes as compared to amastigotes. The exchange flux constraints applied to generate the two metabolic states (promastigote and amastigote) in the genome-scale iAS556 model and simulated uptake rates of different metabolites in both these stages are given in Table 2.1. As flux changes in the metabolic network are governed only by constraints on uptake of environmental metabolites, only the bounds on exchange fluxes were varied while fixing bounds of important internal reactions and transports to make them irreversible, according to the information obtained from *Leishmania* literature (Saunders et al. 2014). All reversible

reactions considered in the model were given bounds as a = -1000 and b = 1000. The bounds for the irreversible reactions considered in the model were a = 0 and b = 1000 or a = -1000 and b = 0. For metabolite exchange reactions or inter-compartmental transport reactions, exchange specifically considering release of a particular metabolite from the cell were bounded between a = 0 and b = 1000, whereas exchanges considering uptake of a metabolite were bounded between a = -1000 and b = 0.

Table 2.1. Stage-specific bounds and simulated optimal rates of external metabolite exchanges.

Exchange name	Promastigote		Amastigote	
	Lower bound	Optimal uptake rate*	Lower bound	Optimal uptake rate*
EX_o2(e)	-1000	-232.14	-65	-65
EX_glu-L(e)	-1000	0	-3	-3
EX_b-D-glucose(e)	-1000	-5.125	-0.2	-0.2
EX_asp-L(e)	-1000	-479.29	-3	-3
EX_pro-L(e)	0	0	-1000	-9.32
EX_hdca	-1000	0	-1000	-0.318
EX_ala-L(e)	-1000	137.75	-2	-2

Note: *calculated with respect to given exchange constraints, while maximizing metabolic demand. Upper bounds of given exchanges were constrained to zero except alanine, which is also produced as an overflow metabolite. Bounds of other internal reactions were constrained between -1000 (lower bound) and 1000 (upper bound) depending upon their reversibility.

The default medium considered for simulating the promastigote and the amastigote scenarios contains glucose, amino sugars, mannose, amino acids (both essential and non-essential), fatty acids, purines and pyrimidines like adenine, guanine, cytosine, uracil, hypoxanthine, inosine, deoxynucleotides, ions, vitamins, CO_2, H_2O, H_2S, HCO_3, NH_3, etc. This media scenario was created with respect to the evidence reported in *Leishmania* literature (McConville and Naderer 2011). The differences in promastigote and amastigote metabolic scenarios arise only due to reduction in the aforementioned uptake rates and the hypoxic condition observed in amastigotes. The medium considered for the iAS142 model involves only glucose and non-essential amino acids with exchanges for ions, CO_2, H_2O, H_2S, HCO_3, NH_3, etc.

2.2.7.1. The iAS142 objective function

Previous constraint-based models on Trypanosomatids have used biochemical data from distant organisms like *Bacillus* and *Tetrahymena* respectively, to construct the objective function (Chavali et al. 2008; Roberts et al. 2009). In comparison to these studies, a novel objective function representing a drain to specific metabolites present in the model was

formulated. The coefficient of the metabolites in the biomass reaction was derived from ^{13}C isotopic enrichment data quantitated from *Leishmania mexicana* log phase promastigotes grown in a completely defined medium containing ^{13}C labeled glucose (Saunders et al. 2014). Enrichment data for only those metabolites present in the iAS142 model were considered for creating biomass reaction. The enrichment values (given in mol percent) scaled between 0 and 1 were used as coefficients of the metabolites considered in the demand reaction. This was considered because the enrichment represents the fraction of that metabolite which is labeled by ^{13}C glucose, thereby, suggesting the catabolism of glucose into the metabolic demand.

The cultivation of the different developmental stages in *Leishmania* is performed in a completely defined medium (CDM) which is majorly composed of glucose when compared to the other carbon sources (Merlen et al. 1999). The above data have also used CDM for growth of *L. mexicana* promastigotes and axenic amastigotes. Glucose was also shown to be utilized in both the life stages of *Leishmania* (Hart and Coombs 1982). Additionally, most enzymes of the glucose catabolism have been shown to be essential for either survival or virulence of trypanosomatids (Coombs et al. 1982; Hart and Coombs 1982; Michels 1988; Barrett et al. 1999; Opperdoes and Michels 2001). Hence, it was assumed that, as glucose largely affects the metabolism of *Leishmania* as compared to other carbon sources, the objective function based on the ^{13}C isotopic enrichment data from CDM containing labeled glucose, would be as close to the actual metabolic demand reaction of *L. mexicana*. As a proof of concept, this derived biomass objective function can possibly be applied to *L. infantum*, considering it to be an evolutionarily close relative of *L. mexicana*.

The metabolic demand reaction thus created is given as follows:

0.9183pg[c] + 0.6634 asp_L[c] + 0.8192 glu_L[c] + 0.6667 ala_L[c] + 0.4743 g6p_B[c] + 0.8861 f6p[c] + 0.8915 s7p[c] + 0.8308 ru5p_D[c] + 0.8905 2pg[c] + 0.8483 pep[c] + 0.4481 g3p[c] + 0.7255 mal_L[c] + 0.6452 fum[c] + 0.7038 succ[c] + 0.4743 g6p_A[c] + 0.029 pro_L[c] + 0.0054 ser_L[c] → (10)

The objective of FBA was to maximize the flux through the metabolic demand, where every metabolite considered in eq. 10 is important for energy metabolism. It is important to note here that, none of the previous metabolic reconstructions have considered radio-labeled isotopic enrichment data for deriving the objective function. Further, it can be observed that the biomass reaction accounts for precursor metabolites of different pathways related to energy and non-essential amino acid metabolism in *Leishmania*. Further, other energy and

amino acid metabolic reconstruction studies of Trypanosomatids also considers precursor metabolites in their biomass reaction (Roberts et al. 2009). Also, ATP drain in the network, considered together with the biomass equation by previous models (Chavali et al. 2008; Roberts et al. 2009) was considered separately in iAS142. The terms for glutamate, alanine, and aspartate in the biomass reaction represent a drain for non-essential amino acids that can be primarily synthesized through the TCA cycle. The synthesis of non-essential amino acids through TCA requires a redox balance to be maintained within the mitochondrion through oxidative phosphorylation which drives cellular ATP production. The ATP produced is then drained through the cytosolic ATP maintenance that recycles ADP and Pi for ATP production. The considered biomass thus, accounts for maintenance of cellular energy requirements also, as it is coupled with the ATP maintenance in our model through redox balance.

2.2.7.2. The iAS556 objective function

To perform flux balance analysis (FBA) of this genome-scale metabolic network, a *Leishmania*-specific metabolic demand reaction was formulated using relative metabolite signal intensity data gathered from a published untargeted metabolomics experiment carried out in *L. donovani*, a visceral *Leishmania* species evolutionarily related to *L. infantum* (Silva et al. 2011). The metabolites considered for formulating the demand reaction were those that were included in the *L. major* iAC560 biomass objective function (Chavali et al. 2008), except for putrescine, spermidine, sphingomyelin, cardiolipin and phosphatidylinositol. Putrescine and spermidine are required to produce reduced trypanothione, which is a part of the demand reaction and hence, were not separately included. Sphingomyelin and cardiolipin were not detected in the above experiment and hence, not included in the demand reaction. Sphingolipids are not essential for growth in *Leishmania* and most of them are acquired after integration with host membranes (Zhang and Beverley 2010). Cardiolipin is found in very low concentrations in *Leishmania* and its functions in *Leishmania* are not very well understood (Zhang and Beverley 2010). Phosphatidylinositol is required to form diacylglycerol, which is included in the demand reaction and hence, was not included separately. Also, phosphatidylinositol is present in low detectable concentrations within the cell membranes of *Leishmania* (Zhang and Beverley 2010). It is also still debated whether phosphatidylserine is acquired from the host or *Leishmania* synthesizes it. The formulated objective function for the iAS556 model is given in Table 2.2.

Table 2.2. Metabolites chosen for metabolic demand and their stoichiometric coefficients.

Metabolite in demand reaction	**Metabolite abbreviation in the model**	**Identified metabolite in experiment** (Silva et al. 2011)	**Available metabolite fraction in total metabolite pool (Coefficient of metabolite in demand reaction)**
serine	ser_L	serine	0.01056
arginine	arg_L	arginine	0.00454
cysteine	cys_L	cysteine	0.00034
glutamic acid	glu_L	glutamic acid	0.03294
glutamine	gln_L	glutamine	0.01606
glycine	Gly	glycine	0.0001
histidine	his_L	histidine	0.00384
isoleucine	ile_L	isoleucine/leucine	0.05987
leucine	leu_L	isoleucine/leucine	0.05987
aspartic acid	asp_L	aspartic acid	0.00745
lysine	lys_L	lysine	0.00486
methionine	met_L	methionine	0.00882
phenylalanine	phe_L	phenylalanine	0.02738
proline	pro_L	proline	0.05032
threonine	thr_L	threonine	0.00642
asparagine	asn_L	asparagine	0.01606
trypanothione	Trprd	trypanothione	0.00036
alanine	ala_L	alanine	0.01467
tryptophan	trp_L	tryptophan	0.00092
tyrosine	tyr_L	tyrosine	0.00872
valine	val_L	valine	0.03742
AMP	AMP	AMP	0.0011
CMP	CMP	CMP	0.00004
GMP	GMP	GMP	0.00006
UMP	UMP	pseudouridine-5-phosphate, UMP	0.0001
dCMP	dCMP	assumed same as CMP	0.00004
dGMP	dGMP	assumed same as dCMP	0.00004
dTMP	dTMP	thymine	0.00022
dAMP	dAMP	assumed same as dTMP	0.00022
diacylglycerol	dag_LI	DAG(36:4)	0.0003
triacylglycerol	tag_LI	TAG(54:7)	0.00008
ergosterol	Ergst	ergosterol	0.00031
phosphatidylcholine	pc_LI	lysophosphatidylcholine(20:3/1)	0.00086
phosphatidylethanolamine	pe_LI	lysophosphatidylethanolamine(18:2/1)	0.0001
zymosterol	Zymst	zymosterol, cholestadienol, 7-dehydrocholesterol	0.00006
heme	hemeA	porphobilinogen	0.00065
pantothenol	pnto_R	pantothenol	0.00018
mannan	Mannan	glycogen, stachyose, maltotetraose	0.0002

Previous constraint-based models in Trypanosomatids have not used organism specific metabolomics data for this purpose (Roberts et al. 2009; Chavali et al. 2012). Maximizing the flux of this hypothetical reaction was considered as the objective for performing flux balance analysis (FBA).

2.2.8. Flux-coupling analysis

Flux coupling analysis (FCA) is a flux-based optimization procedure that calculates reaction subsets that are either coupled with each other via flux or represent a set of block reactions, given specific environmental exchange constraints (Burgard et al. 2004; Larhlimi et al. 2012). Let v_1 and v_2 be fluxes through reactions R_1 and R_2. Keeping either v_1 or v_2 as objective functions to be optimized, if a non-zero flux in v_1 imposes a non-zero flux in v_2 or vice versa, the two reaction fluxes are termed to be coupled with each other. If zeroing the flux of one reaction does not produce any effect on any other reaction within the metabolic network, then the reaction is termed to be uncoupled. If maximum or minimum of a particular reaction flux objective equals zero, then the reaction is termed to be blocked. Considering v_1 or v_2 to be objective functions, the coupled reactions can be classified into:

1) Fully coupled: If $v_1 = 0$ implies $v_2 = 0$ and if $v_2 = 0$ implies $v_1 = 0$, and $v_1 = v_2$, then the reaction pair is fully coupled.

2) Directionally coupled: If $v_1 = 0$ implies $v_2 = 0$ but if $v_2 = 0$ does not imply $v_1 = 0$, then the reaction pair is directionally coupled.

3) Partially coupled: If $v_1 = 0$ implies $v_2 = 0$ and if $v_2 = 0$ implies $v_1 = 0$, and $v_1 \neq v_2$, then the reaction pair is partially coupled.

4) Blocked reactions: if a reaction carries zero flux irrespective of a flux change in any other reaction within the network, it is classified to be blocked,

5) Uncoupled reactions: if a reaction pair does not fall into any of the above categories, it was classified as uncoupled.

FCA allows the computation of all possible set of coupled reaction subsets given the constraints and does not depend on a single hypothetical objective function like, FBA, which induces the artificial coupling of network reactions that can optimize the objective. This analysis was performed using the F2C2 tool that computes a flux coupling table (fctable) which contains the information related to flux-coupled reaction pairs (Larhlimi et al. 2012). To avoid hypothetical coupling of large set of reaction fluxes with the demand reaction, the metabolic demand reaction was removed from the network. Separate, independent reversible

drains for metabolites within the biomass were provided in the network allowing for both their uptake and release from/into the environment and then flux-coupling analysis was performed. With this addition, a revised network consisting of 1288 reactions and 1160 metabolites was subjected to flux coupling analysis. FCA within the *L. infantum* iAS556 metabolic network identified 2243 fully coupled pairs, 4 partially coupled pairs, 2460 directionally coupled pairs, 128 uncoupled reactions and 339 blocked reactions. The number of identified blocked reactions within the iAS556 network was also found to be comparable with the *L. major* iAC560 (Chavali et al. 2008) and *E. coli* iJO1366 (Orth et al. 2011) metabolic networks. Around 33.48% [865 of 2583] of total reactions within the *E .coli* network, 35.7% [398 of 1112] within the *L. major* network and 26.31% [339 of 1288] within the *L. infantum* iAS556 network were identified to be blocked. With more availability of information related to *Leishmania* metabolism, the number of blocked reactions can be expected to reduce.

2.2.8.1. Creation of a flux-coupled subnetwork

The reaction pairs that are coupled can be identified using the flux coupling table computed by the F2C2 tool. The above flux coupled pairs was represented as a mixed flux-coupled graph containing both directed and undirected edges that indicates the coupling nature between any two reactions (nodes in the flux coupled graph) (Subramanian and Sarkar 2016). The flux-coupled graph represents a subnetwork (subgraph) of the unipartite reaction projection of the bipartite metabolic network, where the nodes are the reactions and the edges represent whether the reactions are coupled (1) or not (0). Fig. 2.4 represents a typical FCA simulation on a toy model and the flux-coupled graph representation of the toy network.

F2C2 identified 949 unblocked reactions within the *L. infantum* iAS556 metabolic network. The flux-coupled graph was generated using the 949 x 949 adjacency matrix (obtained from the flux coupling table) that contains information of each pair of reactions to be either coupled (1) or uncoupled (0). Fully and partially coupled reaction pairs are reversibly coupled and hence within the flux-coupled graph, were connected by a single undirected edge. Directionally coupled reaction pairs were connected by a single directed edge from one node to another. Isolated nodes with zero connectivity represent the completely uncoupled reactions. The graph diagram for the flux-coupled reaction graph was generated in Gephi 0.8.2 using the Fruchterman-Reingold algorithm (Bastian et al. 2009).

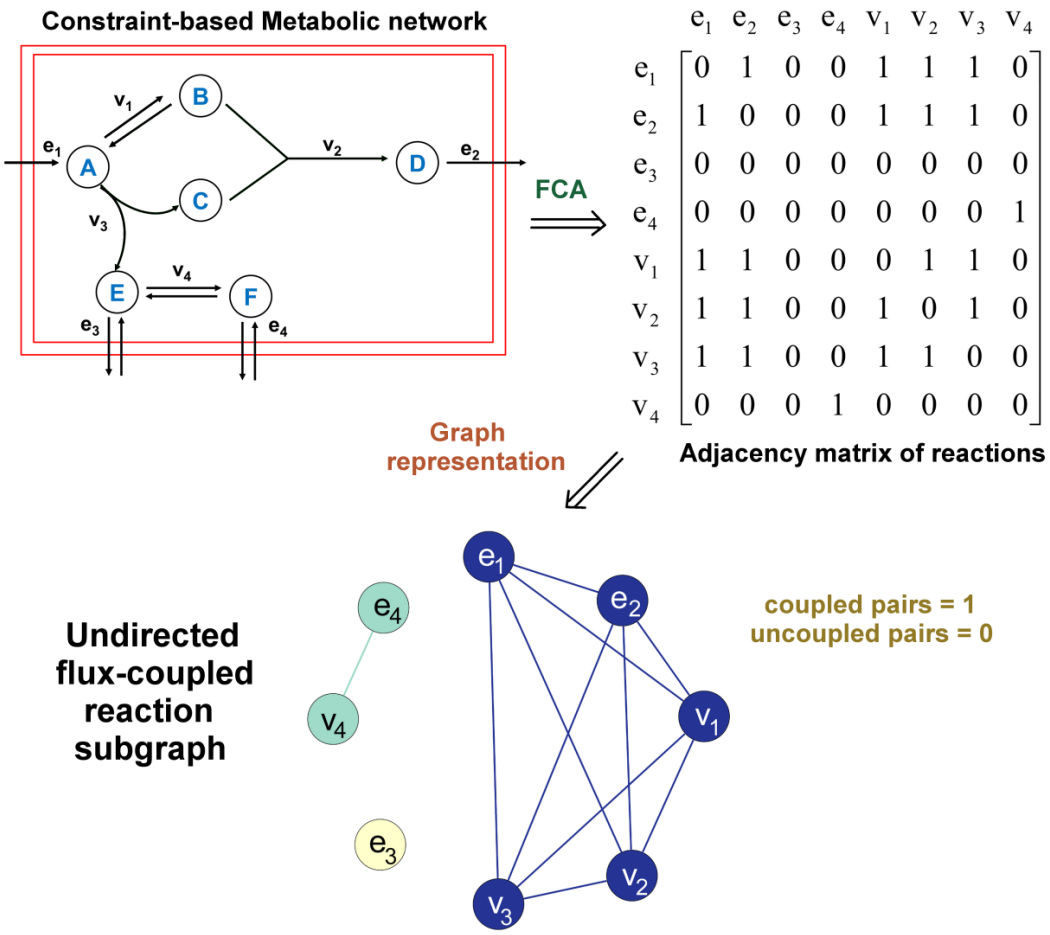

Figure 2.4. Creation of a flux-coupled subgraph from a given metabolic network - The toy model similar to Fig. 2.3 is given here. The model in Fig. 2.3 was modified to remove objective function before performing the FCA procedure. Performing FCA, an adjacency matrix of coupled reaction pairs is generated. FCA gives the total set of coupled reaction pairs classified as fully coupled, directionally coupled and partially coupled pairs. In the above example, the identified coupled reaction pairs are all fully coupled. The adjacency matrix is further modified to obtain reaction pairs that are either coupled (1) or not (0). Using this matrix, an undirected flux-coupled reaction subgraph is generated. In the above example, e_3 is uncoupled. The blue nodes and edges represent a complete flux-coupled module (that corresponds to the flux distribution path formed using FBA). The green nodes represent singular coupling nodes.

2.2.8.2. Topological analysis of the flux-coupled subnetwork

With the graphical representation of the flux-coupled subset of reactions, the flux-topological properties of enzymes within this graph were computed (Figure 2.4). The total number of reactions to which a given reaction can be coupled was calculated by computing the degree centrality of each node of the flux-coupled graph. Local clustering coefficient was computed for each node to predict the tendency of an enzyme to cluster together with other enzymes

with similar number of flux-couplings and thereby, the probability that a particular reaction would be a part of a strongly connected motif within the network.

2.2.9. Reaction knockout analysis

To simulate the effect of the single reaction deletions, the upper and lower bounds of all the reactions were constrained to be zero, sequentially one reaction at a time, and FBA was performed subsequent to each reaction knockout in a specified medium. A reaction was considered to be essential (lethal phenotype) when reaction flux through metabolic demand became zero after applying the aforementioned constraints; else it was deemed to be non-essential (non-lethal phenotype). In the iAS142 model, the reaction knockout analysis was performed in a medium containing glucose and non-essential amino acids, without stage-specificity. Knockouts for the iAS556 model were performed in an *in-silico* rich medium containing glucose; all essential amino acids; non-essential amino acids like cysteine, glycine, aspartate, alanine, glutamine and glutamate; nucleobases like adenine, guanine, hypoxanthine, xanthine, uracil; hexadecanoate, phosphatidic acid, pantothenate and protoheme separately for the promastigote and amastigotes. In both models, two separate transport reactions of glucose each considering the anomeric, inter-convertible forms of glucose are present within the network. In order to avoid redundant glucose uptake from both these transports, one of them was blocked while performing reaction knockouts. Lethality predictions remain the same for the model-presumed promastigote and amastigote scenarios, as these scenarios were re-created based only upon stage-specific environmental conditions experienced by the developmental stages of the parasite while maximizing for the same metabolic demand. In the iAS556 model, 188 reactions out of the total 1260 reactions were predicted to be lethal, under the above mentioned model constraints.

2.2.10. Utilization of carbon substrates

The fate of carbohydrates, like glucose (Glc) and mannose (Man), amino sugars like N-acetylglucosamine (Acgam), non-essential amino acids, like alanine (Ala), aspartate (Asp), glutamate (Glu) and hexadecanoate fatty acids (Hdca), was predicted for both the stages by computing the steady state fluxes for each reaction under uptake of each carbon source one at a time, while optimizing for the metabolic demand reaction. Exchanges of proline (Pro), glycine (Gly), essential amino acids, like serine (Ser), threonine (Thr), leucine (Leu), isoleucine (Ile), valine (Val), lysine (Lys), phenylalanine (Phe), were all provided simultaneously for all the other amino acids (AA) situation. The fate of the mentioned

metabolites was specifically studied in order to compare the model predictions with observations from experimental ^{13}C-isotope labeling profiles of the above metabolites generated in different developmental stages of *L. mexicana* (Saunders et al. 2014). The ^{13}C isotope labeling profiles in this experimental study were obtained by growing *L. mexicana* promastigotes and amastigotes within a chemically defined growth medium (CDM) containing specific ^{13}C -tagged carbon sources and analyzing the isotope enrichment of ^{13}C labeled intracellular metabolite pools using gas chromatography-mass spectrometry (GC-MS). ^{13}C isotope enrichment measures the amount of incorporation of ^{13}C tagged environmental metabolite (in mol percent) into its intracellular catabolic products thereby indicating the preference of utilization of a particular metabolite through specific pathways.

The *in-silico* stoichiometric-based utilization of each metabolite produced from previous reaction into subsequent reaction/s within the iAS556 network was traced by the contribution of previous reaction flux into its subsequent reaction flux. Statistical comparison of model predictions with experimental data was performed for the glucose-only situation as i) only glucose is catabolized by a large number of steps (sufficient sample size for comparison) each of which is labeled from ^{13}C isotope-resolved metabolomics and,
ii) because the normalized ^{13}C isotope enrichment values were reported for only labeled glucose.

Spearman rank correlation was computed to identify the strength of associativity between predicted reaction steady state fluxes (normalized to flux of hexokinase) and ^{13}C isotope enrichment values of metabolites (normalized to glucose-6-phosphate) grown on labeled ^{13}C glucose (Saunders et al. 2014). In all the mentioned conditions, uptake of glucose, essential amino acids, cofactors and ions were always kept unconstrained irrespective of change in environment, as they are absolutely essential for growth. Also, drains for overflow metabolites were provided in the model to release by-products.

2.2.11. Sensitivity analysis

To understand the role of non-essential amino acids, a sensitivity analysis was performed to understand the effect of glucose and uptakes of all the amino acids, while optimizing for the formation of glutamate, alanine, aspartate, glutamine, proline, glycine, myoinositol and mannogen, each formation considered as a separate objective function. The uptake of glucose and choice of myoinositol and mannogen synthesis as objectives is to understand and compare the coherent role of glucose with amino acids. Due to the role as intermediate by-products of reactions that are consumed for optimizing metabolic demand, for each

simulation, drain (secretion/release) for alanine, glutamate, glutamine and proline was considered. Each simulation was performed separately, considering one input at a time while constraining all other exchanges to zero. All remaining exchanges in the model were kept default (lower bound = -1000 and upper bound = 1000) (Ravikrishnan and Raman 2015).

The considered input uptakes can be given by x, where $x \in A$ = {Asn, Ala, Arg, Glc, Cys, Gln, Glu, His, Ile, Asp, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val}. Four different scenarios were considered for optimizing the considered objective functions namely,

i) only one input source (x) constrained to a non-zero flux value (lower bound = -1000) at a time and all other inputs (A\{x}) are constrained to zero;

ii) two inputs given at a time – glucose which is kept fixed and a variable second input (x) with non-zero flux given one at a time and all other inputs are constrained to zero;

iii) two inputs given at a time – one being isoleucine which is kept fixed and a variable second input (x) constrained to a non-zero flux value given one at a time and all other inputs are constrained to zero;

iv) three inputs given simultaneously – glucose, tyrosine which are fixed and a variable third input (x) with non-zero flux given one at a time and all other inputs are constrained to zero.

2.2.12. Effect of subcellular compartmentalization on flux distribution

For exploring the role of subcellular compartments in reinforcing coupling relationships, reactions occurring in specific subcellular compartments, like glycosome or mitochondrion, were shifted to the cytoplasm while deleting the corresponding compartment from the iAS556 model. Also, it was ensured that, reactions occurring in dual or multiple locations (for example, glycosome and cytoplasm or mitochondrion and cytoplasm) were considered as a single reaction in cytoplasm to avoid redundancy in the re-created model.

2.2.13. Reconstitution of flux relationships under random perturbations

Scenarios for perturbation of flux-coupled relationships due to random reaction malfunction were created by deleting random single or multiple reactions (0 to 20 simultaneous deletions) while re-performing FCA on the network for each case. 1000 random simulations were performed for each case. Each deletion represents the complete loss of function of a given gene or sets of genes absolutely eliminating the role of their corresponding reactions.

2.3. Multivariate analysis to identify confounding factors in metabolic enzyme evolution

In this part of the study, a total of 8 features representing the genotype and phenotype characteristics of the *Leishmania* parasite were computed. These features were chosen on the basis of their occurrence in previous literature related to *Leishmania* or other eukaryotes. Only those *Leishmania* species, for which a curated genome-scale metabolic reconstruction was available, were chosen for the multivariate analysis. Likewise, three datasets comprising of the 560 metabolic genes curated in the *L. major* iAC560 reconstruction, 604 metabolic genes curated in the *L. donovani* reconstruction and 556 genes curated in our *L. infantum* reconstruction were used in this study (Chavali et al. 2008; Sharma et al. 2017; Subramanian and Sarkar 2017).

2.3.1. Genomic features

The coding nucleotide sequences (CDS) of the metabolic genes curated within each metabolic reconstruction, obtained from the TriTrypDB database, v.8.1, release 32 (Aslett et al. 2010) were used for calculation of CAI, GC content and the gene length. CAI values for each gene were computed using the EMBOSS package (Rice et al. 2000), with respect to a reference set of ribosomal protein-coding genes in each species (Subramanian and Sarkar 2015). The length and GC content for each CDS was computed using an in-house PERL script.

2.3.2. Curation of gene expression features

In unicellular organisms, mRNA concentration of a gene collected during mid to late log phase of growth under rich media is presumed to represent an average concentration observed across cell cycle stages, which can be used for predicting evolutionary divergence (Zhang and Yang 2015). Hence, read counts for calculation of reads per million kilobases (RPKM) values of genes in *L. donovani* (GSE48475) and *L. infantum* (GSE48394) were calculated using the raw read data obtained from RNA-sequencing experiments carried out in control mid to late log phase promastigotes (Martin et al. 2014; Zhang et al. 2014), using the following formula,

$$\text{Reads Per Million } kbs \text{ (RPKM)} = \frac{\text{Total number of reads mapped to a gene sequence}}{\text{Gene Length in } kbs \times \text{Total number of reads in a sample per million}} \quad (11)$$

As there was no RNA-sequencing experiment conducted exclusively on *L. infantum*, RPKM values of genes were obtained from the RNA sequencing study in *L. donovani* where the reads were reported to be mapped onto the *L. infantum* JPCM5 genome (Martin et al. 2014). Fragments per million kilobases (FPKM) values of genes in *Leishmania major* were obtained from another independent RNA-sequencing experiment conducted in *L. major* promastigotes.

2.3.3. Functional constraint

Number of processes and functions: Predicted Gene Ontology (GO) processes and functions associated with each gene was extracted from the TriTrypDB database (Aslett et al. 2010). The number of predicted processes (NumProcs) and functions (NumFuncs) were extracted from this information using an in-house PERL code.

Flux-coupling potential of an enzyme: Initially, topological features like centrality of an enzyme (degree or number of flux-couplings, NCoup) and the tendency of an enzyme to cluster together with other enzymes with similar number of flux-couplings (local clustering coefficient, CCoFCA) were calculated for the flux-coupled subgraph of the genome-scale metabolic networks considered. The mapping of these topological properties with the enzymes of the network provides us with the flux-coupling potential of an enzyme.

As observed in the above *Leishmania* metabolic reconstructions, there is no one to one correspondence of metabolic enzymes with the metabolic reaction considered within the network. There are three possible categories of reaction-enzyme correspondence.

a) *Single enzyme – single reaction*: For a single enzyme catalyzing a single reaction, the degree per reaction associated with an enzyme is same as degree of the corresponding reaction within a flux-coupled graph.
b) *Multiple enzymes – single reaction*: In case of some reactions, multiple enzymes are associated with the same reaction. For example, in *L. infantum*, the E1 component of the pyruvate dehydrogenase complex (EC: 1. 2. 4. 1), which catalyzes a decarboxylation of pyruvate and a reductive acetylation of lipoic acid, is made of two subunits α and β, each encoded by two different genes LinJ.18.1360 and LinJ.25.1790 respectively. In such cases, the same degree (number of couplings within a flux-coupled graph) and clustering coefficient per reaction was associated to every enzyme involved in that reaction.
c) *Single enzyme – multiple reactions*: Also, in few cases, the same enzyme is used to catalyze conversion of multiple substrates or substrates in multiple subcellular

locations, into products (represented as separate reactions within the reconstruction). For example, in *L. infantum*, acetyl-coA synthetase catalyzes the conversion of two substrates acetate and propanoate to acetyl-coA and propionyl-coA while converting ATP to AMP. Hence, a gene is linked to two different network reactions.

FCA predicts the total number of possible flux coupled reactions that are associated with a given reaction. But, all the possible couplings might not always be activated in a given condition, as the substrate would be converted into only those products that are more required by the cell; which cannot be directly considered in FCA. Hence, a reaction that maybe important with respect to a required metabolite need not necessarily possess a large number of flux-couplings. To account for this effect, in such cases, the degree or clustering coefficient per reaction mapped onto a single enzyme was considered to represent the average of the degree/clustering coefficient of all the reactions catalyzed by that enzyme. A similar average-based weightage was considered previously for other metabolic flux and topological features in a previous study on *E. coli* (Plaimas et al. 2010).

2.3.4. Sequence-based evolutionary rates

For the estimation of the evolutionary rates, multiple sequence alignment of each gene in all the three species was performed with its orthologous sequences across five genomes, namely, *L. major*, *L. infantum*, *L. donovani*, *L. mexicana* and *L. braziliensis* species. The orthology information was available within the TriTrypDB database, v.8.1, release 32 (Aslett et al. 2010). The alignment was processed to remove sequence positions with gaps using a standalone version of the PAL2NAL program (Subramanian and Sarkar 2016). d_N, d_S and ω (d_N/d_S) were estimated using the one-ratio M0 branch model implemented in the 'codeml' subroutine of the PAML package version 4.8a (Yang 1998; Yang 2007).

2.3.5. Pre-processing the datasets for multivariate analysis

For each species, the dataset of metabolic genes was pre-processed to remove – a) genes with obsolete sequences, less than 200 codons, $d_S > 0.3$, duplicates and b) genes for which either of the targeted genomic, expression, or metabolic network-based features was unavailable. Finally, only 233 singletons common to the three species of *Leishmania* was considered for multivariate analyses.

Extraction of singleton enzymes (genes): Since the number of duplications with respect to each enzyme substantially varies across all the *Leishmania* species, the dataset in each

species was pre-processed to extract only singletons for comparison. A set of metabolic genes from a single species were considered to be duplicates if they belonged to the same orthologous group and were associated to the same reaction within the metabolic reconstruction. Further, all those genes with $d_S > 0.3$ were also removed from each data set as they might denote either misalignment or potential sequence saturation (Yang and Gaut 2011).

2.3.6. Multivariate analysis and clustering

Principal Component Regression: Principal Component Regression (PCR) analysis on metabolic counterpart of the three *Leishmania* species was used to identify the potential contribution of the genomic, gene expression and function-based features to the total variance in evolutionary rates among metabolic genes (Fig. 2.5). The 'pls' package version 2.6 implemented in R was used to perform PCR with d_N and d_S as the response and the aforementioned 8 parameters as the predictor variables. A subset of predictor variables with loadings of 0.45 or more was considered for interpretation of a principal component with respect to that subset (Tabachnick and Fidell 2007).

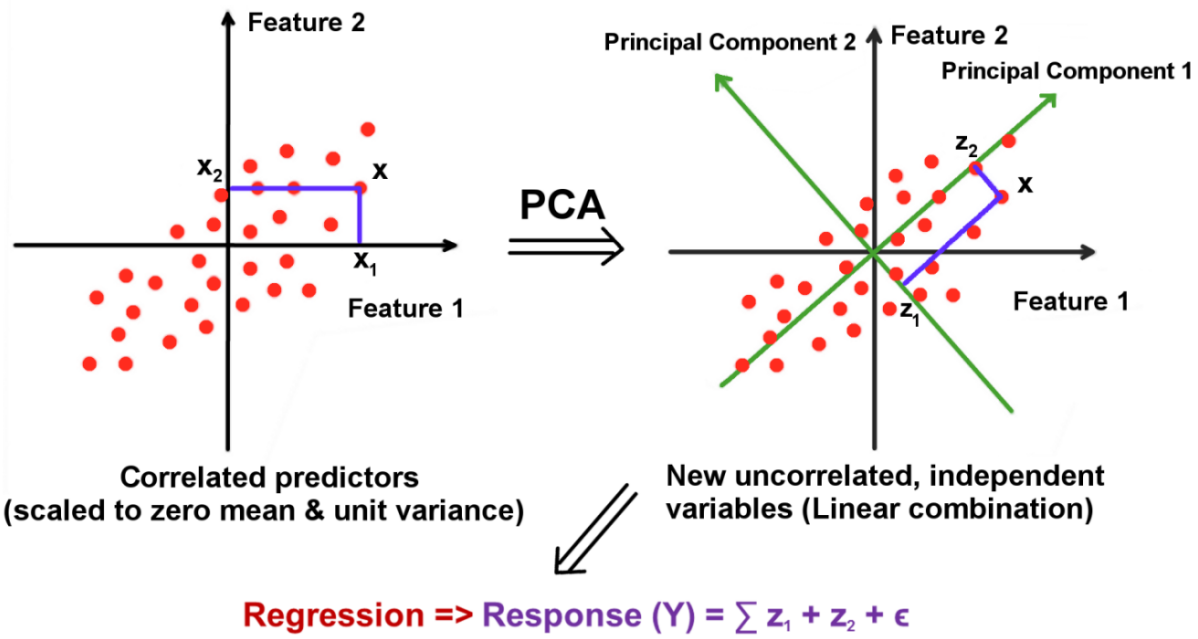

Figure 2.5. Principal Component Regression (PCR) Analysis – First step of the PCR is to scale each of the correlated predictors to zero mean and unit variance (standardization). Next, a principal component analysis (PCA) is performed to identify the principal components that explain maximum variation in the data. In above example, there are two features hence, two principal components are generated. These principal components are statistically independent of each other. In the ultimate step, these principal components are used for predicting the response by multiple linear regression. The advantage of PCR is that it allows predict the variation in the response by using the variance explained in the predictors.

Preparation of dataset for principal component regression: Before performing a PCR, variables were log transformed based on whether the log transformation led to a higher correlation coefficient between predictor and response (Drummond et al. 2006; Yang and Gaut 2011). Variables containing zero were added with a small constant before the log transformation. All the variables were scaled to zero mean and unit variance and then used for principal component analysis. Performing a PCR, the percentage variance in d_N, d_S and ω explained by each principal component was calculated.

Selection of minimum principal components for regression: A randomization test approach was used to check whether the squared prediction errors of regression models with fewer components are significantly ($P < 0.01$) larger than the reference model predicting absolute minimum prediction accuracy or not, by generating a distribution of prediction errors in each model for comparison using 1000 random permutations (van der Voet 1994). Out of these significant models, the model with least number of principal components was chosen as the best model to predict d_N and d_S in all three species. The randomization test approach is implemented within the 'pls' package.

K-means clustering: K-means clustering of genes was performed in an n-dimensional space, where n represents the selected number of principal components. Clustering was performed so as to identify the groups of genes, governed by a particular set of principal components and thereby a subset of predictors. The number of clusters represented in each dataset was determined by computing the Akaike's Information Criterion (Christopher et al. 2008) for every K clusters (AIC); where $K = 1$ to 100. The number of clusters corresponding to the model with least AIC was considered to be representative for each dataset.

Chapter 3 – Comparative codon usage analysis across *Leishmania* and other Trypanosomatids

3.1. Introduction

A distinct species-specific heterogeneity is observed in clinical manifestations of different *Leishmania* species within the human host. Among 7 completely sequenced *Leishmania* genomes published so far, six are of the species known to infect humans (El-Sayed et al. 2005; Ivens et al. 2005; Peacock et al. 2007; Downing et al. 2011; Raymond et al. 2011; Rogers et al. 2011; Real et al. 2013). The *Leishmania* genome is a highly intact genome and does not display major variations across *Leishmania* species, due to common factors such as lack of extended subtelomeric regions, maintenance of gene order and presence of degenerate transposable elements (Ghedin et al. 2004; Smith et al. 2007). This poses a complex problem in finding the probable reasons for species-specific differences in clinical manifestations observed within the same host. Numerous studies have indicated that species-specific differences in gene expression rather than specific genes themselves are responsible for differential clinical manifestation (McCall et al. 2013). In *Leishmania*, weak correlation is known to exist between protein and mRNA expression levels suggesting regulation of expression at translation level (Lahav et al. 2011). Even though proteomic studies can provide a better view in this respect, they are limited in *Leishmania* and are restricted to very few genes that show abundant expression at the protein levels (Santos et al. 2013). Large scale identification of codon usage patterns between the *Leishmania* species can immensely help to divulge these crucial differences and the evolutionary events responsible for them.

Codon Usage Bias (CUB) essentially depicts the unequal usage of synonymous codons in genes, where the preferred codon in the sequence primarily correlates with tRNA isoacceptor abundances. This preference or bias is a well-studied phenomenon across wide range of organisms ranging from prokaryotes to eukaryotes (Escalante et al. 1998; Das et al. 2006; Behura and Severson 2013). Bias in codon usage may arise due to several evolutionary factors like - translational selection, where the highly expressed genes tend to exhibit a high frequency of preferred codons (Hershberg and Petrov 2008; Plotkin and Kudla 2010), mutational pressure, where those codons are preferred in the genome that tend to demonstrate a bias towards particular nucleotide content (like GC or AT bias) at specific codon positions (like the wobble position) (Sueoka 1988) and amino acid composition bias, where the codons that code for highly frequent versus uncommon amino acids tend to be preferred in genes (Vicario et al. 2007; Behura and Severson 2012).

* The bulk of this chapter has appeared in *Genomics*, 106:232–241 (2015) and *Data in Brief*, 4:269–272 (2015), co-authored by A. Subramanian and R. R. Sarkar

During evolution of genomes, these diverse factors shape CUB, thereby using it as an instrument to control global regulation of protein expression (Sharp and Li 1987; Gustafsson et al. 2004), choice of gene function (Chiapello et al. 1998; Karlin et al. 1998) and formation of secondary structures within the mRNA (Gu et al. 2010; Tuller et al. 2010). Hence, large scale comparison of CUB between closely related organisms can help to unravel diverse patterns, further indicating the role of evolution in devising a mechanism for generating variation in coding regions of the genomes.

Previous studies on codon usage patterns in *Leishmania* have primarily attempted to uncover the evolutionary events that shape CUB (Alonso et al. 1992; Alvarez et al. 1994; Horn 2008; Gu et al. 2010; Tuller et al. 2010; Chauhan et al. 2011; Singh and Vidyarthi 2011; Rashmi and Swati 2013). Since these have been limited either by the low number of genes available or by the fewer number of species considered for comparison, a strong basis for CUB across Genus *Leishmania* was not established. These studies have mostly been restricted to understanding the causes of CUB and do not probe into the consequences of CUB on protein expression and hence, gene functions.

Overcoming these caveats, the codon usage patterns across *Leishmania* and other Trypanosomatids was compared to decipher traces of evolutionary events like mutational pressure, translational selection and amino acid composition bias. From phylogenetic studies using ribosomal RNAs and housekeeping proteins, *Crithidia* and *Trypanosoma* have been shown to be evolutionarily close genera to Genus *Leishmania* (Fernandes et al. 1993). Also, the sequenced genomes of species belonging to the *Crithidia* and *Trypanosoma* genera are publicly available. Hence, exploiting the available completely assembled genomes of 6 *Leishmania* species, 6 *Trypanosoma* species and a *Crithidia* species for studying the evolutionary causes of CUB, its possible relationship with relative protein abundance, pathway-level function and mRNA secondary structure formation was also evaluated across Trypanosomatids. The differences in codon usage observed in enzymes belonging to specific pathway/process between various *Leishmania* species might further provide the probable reasons for differential clinical manifestations observed between various *Leishmania* species. From this chapter, it can be hypothesized that for genomes with conserved gene organization like *Leishmania*, a comparative analysis of codon usage can be used to infer both evolutionary and functional relationships among species.

3.2. Results

3.2.1. Codon usage patterns across *Leishmania* and other Trypanosomatid genomes

To identify and understand the unequal usage of codons in *Leishmania* genomes in comparison with other Trypanosomatids, Relative Synonymous Codon Usages (RSCU) were computed for every codon in each Trypanosomatid genome. Performing hierarchical clustering of RSCU values between species (Fig. 3.1A), it can be observed that codons rich in G or C were more preferred in *Crithidia* and *Leishmania* species as compared to *Trypanosoma* species. Two distinct clusters of *Crithidia* + *Leishmania* (Leishmaniinae) and *Trypanosoma* were obtained from hierarchical clustering of RSCU values of each genome similar to the expected phylogenetic clustering of the genomes, suggesting RSCU to be an important species-specific feature for classification (Fig. 3.1B). Further, highly frequent codons across the *Crithidia* and *Leishmania* genomes commonly end with a G or C, and rare codons, with the exception of GGG (Gly) and AGG (Arg) end with an A or U. Also, it can be clearly observed that the frequencies of CUG (Leu) and CGC (Arg) are considerably higher in *Leishmania* and *Crithidia* as compared to *Trypanosoma* (Fig. 3.1A). Comparison of the RSCU values between the clusters indicated that specific codons are preferred by the Leishmaniinae genomes as compared to the *Trypanosoma* species (Table 3.1). It was found that 20 codons were significantly biased between the two clusters ($P < 0.05$). These codons are used as frequent codons in one cluster and rare codons in the other cluster or vice versa (Subramanian and Sarkar 2015). AUC (Ile), ACC (Thr), GCC (Ala), UUC (Phe), UCG (Ser), and GUC (Val) are consistently used as the frequent codons in *Crithidia* and *Leishmania*, whereas they are rare in *Trypanosoma* species. Further, specific codons (such as GUG – Val and CUG - Val) are preferred across all the 13 Trypanosomatid genomes and certain codons (such UUA - Leu and GUA – Val) are avoided (Fig. 3.1A).

3.2.2. Mutation pressure towards GC nucleotide composition affects codon usage

The preference of G or C in the third position of the codon in *Leishmania* observed in the RSCU comparative analysis could be due to the overall GC bias within the genome. In order to investigate this aspect, variations in GC content were captured and analyzed among the Trypanosomatids. Previously, variations in GC content among Trypanosomatids were studied for a small set of genes in a limited number of species (Alonso et al. 1992; Alvarez et al.

1994; Singh and Vidyarthi 2011). It was also predicted that there is a high GC mutational pressure for certain genes in certain *Crithidia* and *Leishmania* species in comparison to *Trypanosoma* species (Alonso et al. 1992). In order to examine this GC bias universally across genes between different known *Leishmania* genomes and to compare its behavior with other Trypanosomatids, a two tier analysis of the GC content variation was performed.

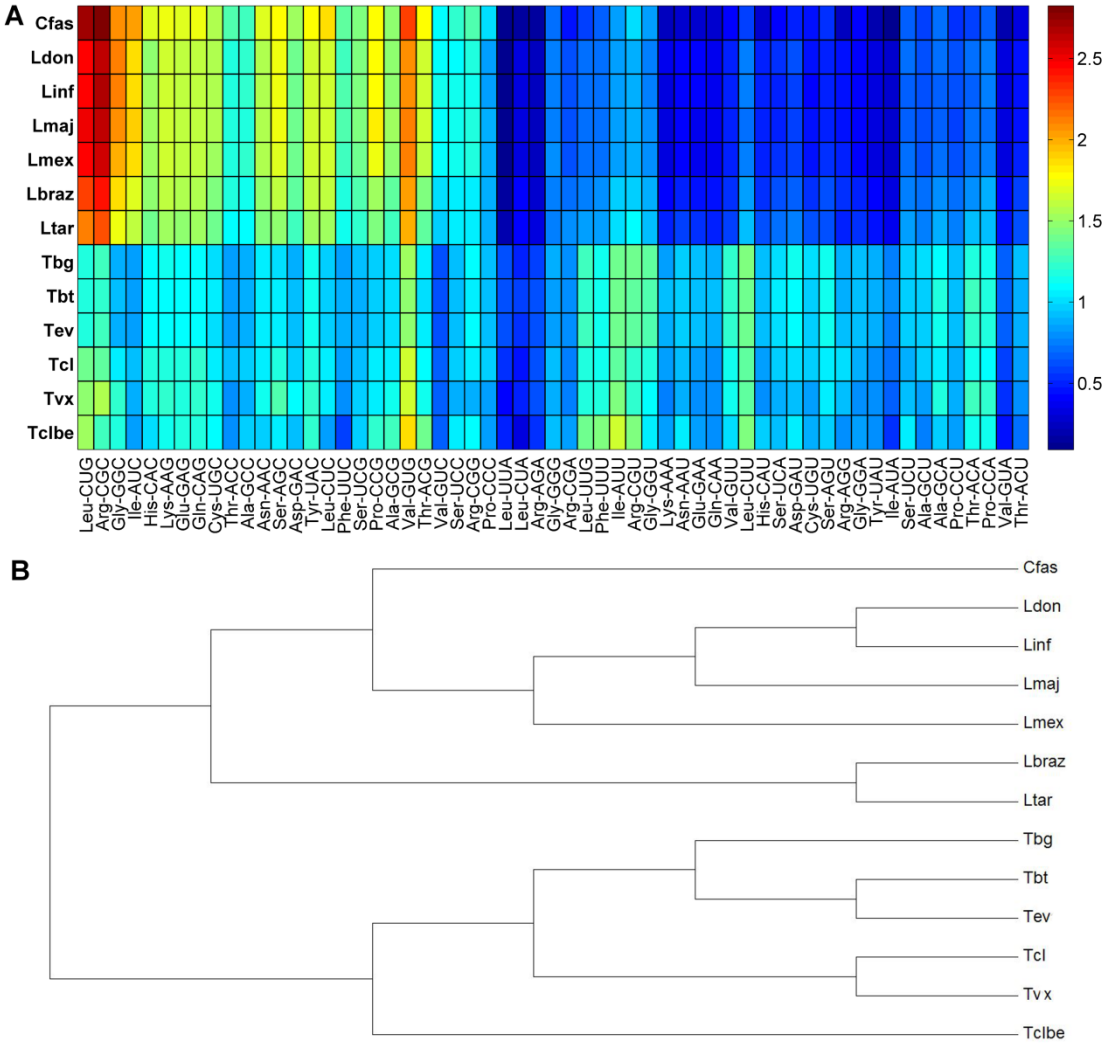

Figure 3.1. Comparison of RSCU between Trypanosomatids – A) The heat map shows the distribution of RSCUs for every codon between 13 different species of Trypanosomatid parasites. Higher the frequency of a codon, higher is the RSCU value (towards red color) and lower the frequency of a codon, lower is the RSCU value (towards blue color). In *Crithidia* and *Leishmania* genomes, codons ending in G or C (indicated by green to red colored values on the upper left of the heatmap) have higher RSCU values than codons ending in A or T. In *Trypanosoma* genomes, no such preference was observed; B) Tree representation of the hierarchical clustering performed for RSCU comparison between Trypanosomatid genomes. Two distinct clusters of *Crithidia* + *Leishmania* and *Trypanosoma* genomes can clearly be observed.

Table 3.1. Codons that are frequently used among the *Crithidia* + *Leishmania* and the *Trypanosoma* clusters.

Codon	*Crithidia* + *Leishmania* (RSCU>1)	*Trypanosoma* (RSCU>1)	*P* value
AUC - Ile	7	0	0.000583
ACC - Thr	7	0	0.000583
GCC - Ala	7	0	0.000583
UUC - Phe	7	0	0.000583
UCG - Ser	7	0	0.000583
GUC - Val	7	0	0.000583
GAC - Asp	7	1	0.004662
UCC - Ser	7	1	0.004662
AUU - Ile	1	6	0.004662
CGU - Arg	1	6	0.004662
UUG - Leu	0	6	0.000583
UUU- Phe	0	6	0.000583
GGU - Gly	0	6	0.000583
GUU - Val	0	6	0.000583
CUU - Leu	0	6	0.000583
ACA – Thr	0	6	0.000583
CCA - Pro	0	6	0.000583
GAU - Asp	0	5	0.004662
AGU - Ser	0	5	0.004662
GCA - Ala	0	5	0.004662

Note: The number of species (out of the *Crithidia* + *Leishmania* and *Trypanosoma* clusters) having frequently used codons (RSCU > 1) is shown. The p-value significance was calculated according to Fisher's exact test.

First level of the analysis was to identify bias in mutational patterns observed in coding sequences across the Trypanosomatid genomes chosen for the study. The mutationist hypothesis verifies this bias by determining the deviation of a genome from a genome having balanced GC and AT content (Alonso et al. 1992). This balance can be quantified for every pair of orthologous coding sequences (OCDS) between the genomes considered for comparison, by calculating the ratio of substitutions from G or C to A or T (p) and the substitutions in the opposite direction (q). A total of 1218 known single copy orthologous groups in the above Trypanosomatidae genomes were used for the analysis (Subramanian and Sarkar 2015). Performing a codon-based alignment between the OCDS, ratio of the (average number of pairwise G or C to A or T substitutions) / (average number of A or T to G or C pairwise substitutions) were calculated for every pair of genomes (Fig. 3.2A). In general, the high average p/q substitution ratios (>1) between Leishmaniinae genomes and *Trypanosoma*

genomes (green) specify a high mutational pressure towards maintaining A or T in *Trypanosoma* genomes. On the contrary, substitutions between *Trypanosoma* and Leishmaniinae genomes indicate low p/q ratios (<1) or a high mutational pressure towards maintaining G or C in Leishmaniinae genomes (pink). Further, high average p/q substitution ratio between all other *Leishmania* genomes and *L. tarentolae*, indicated that the later was the most AT-biased among all the *Leishmania* genomes (yellow). No clear distinction was observed in the substitutions between the *L. major*, *L. donovani* and *L. infantum* genomes signifying their evolutionary relatedness (cyan). Further, low p/q ratios between the New World *Leishmania* species (*L. braziliensis* and *L. mexicana*) and Old World *Leishmania* species (*L. major*, *L. donovani* and *L. infantum*) prove that the latter are comparatively more GC biased (purple).

The second level of analysis was to understand the effect of GC mutational pressure on the different codon positions. As genomes accumulate mutations during evolution, a comparison of the position-wise GC content of the codons with the genome GC content would delineate the effects of mutational pressure against purifying selection. A previous study reported this observation by comparing a certain set of known genes between few Trypanosomatid species (Alonso et al. 1992). This finding was validated universally by considering the 13 Trypanosomatid genomes (Fig. 3.2B). The average GC content at the synonymous wobble position (GC3s) follows the changes in the average genome GC content suggestive of a relatively weaker effect of purifying selection and hence, stronger effective mutational pressure. Additionally, *Crithidia* and *Leishmania* genomes have a higher GC mutation bias at 3rd position as compared to *Trypanosoma* species. The behavior of the average GC content at the 1st (GC1s) and 2nd codon position (GC2s) with genome GC content further represents a comparatively stronger selection pressure on the first two codon positions; selection pressure being higher on GC2s as compared to GC1s.

The above analysis reveals the existence of a mutational bias towards GC nucleotide content at the 3rd position but its role in affecting codon usage still needed to be investigated. Effective Number of Codons (ENC) is a non-directional measure of CUB that is known to be dependent upon nucleotide composition in genes (Wright 1990). In order to further identify the nucleotide bases governing codon usage bias, Poisson regression analysis was performed using base compositions at the 3rd codon positions as predictors to predict the ENC values of genes in each Trypanosomatid genome.

A

	Cfas	Lbraz	Ldon	Linf	Lmaj	Lmex	Ltar	Tbg	Tbt	Tcl	Tclbe	Tev	Tvx
Cfas	NA	1.6738	1.4019	1.4026	1.4136	1.444	1.8422	2.483	2.481	2.3769	2.4613	2.4911	2.423
Lbraz	0.634	NA	0.7428	0.7448	0.7556	0.7839	1.1591	1.7553	1.7589	1.6879	1.7106	1.7577	1.7079
Ldon	0.762	1.3978	NA	0.8008	1.0846	1.1576	1.8084	2.0109	2.0146	1.9214	1.961	2.0101	1.9449
Linf	0.7621	1.395	0.8384	NA	1.0875	1.1567	1.7999	2.0108	2.0146	1.9189	1.9562	2.0106	1.9429
Lmaj	0.7574	1.3713	1.0166	1.0167	NA	1.1196	1.7484	2.0001	2.0057	1.9038	1.9567	1.9995	1.9397
Lmex	0.7436	1.324	0.9373	0.939	0.9579	NA	1.6725	1.9704	1.9726	1.8809	1.9285	1.9801	1.9145
Ltar	0.5854	0.8887	0.5826	0.5875	0.6022	0.626	NA	1.6361	1.6346	1.5625	1.5794	1.6351	1.5803
Tbg	0.4278	0.5977	0.5233	0.5253	0.5255	0.5351	0.6418	NA	1.0769	0.9642	0.9771	0.2907	0.9893
Tbt	0.4264	0.5959	0.5221	0.5234	0.5248	0.5333	0.6414	1.1299	NA	0.9629	0.9783	1.1163	0.9877
Tcl	0.4455	0.624	0.5471	0.5466	0.5518	0.5589	0.6757	1.0812	1.0949	NA	1.0245	1.0903	1.0514
Tclbe	0.4432	0.6204	0.5444	0.5483	0.547	0.5563	0.6724	1.0935	1.0859	1.0328	NA	1.0797	1.0417
Tev	0.4245	0.5948	0.5219	0.5236	0.5242	0.5313	0.6403	0.3189	1.057	0.9631	0.9787	NA	0.987
Tvx	0.4361	0.611	0.5366	0.5379	0.539	0.5475	0.6603	1.0595	1.0625	1.0011	1.0177	1.0592	NA

B

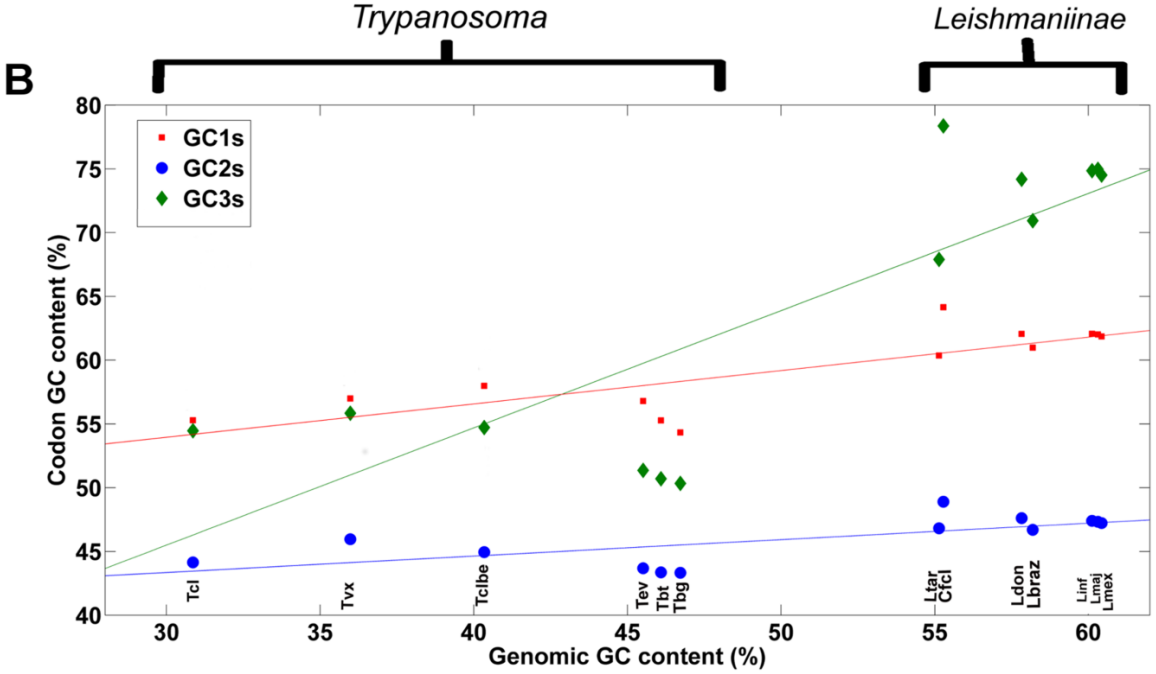

Figure 3.2. Mutational pressure and variations in GC content across Trypanosomatids - A) Mutation pressure matrix whose values represent the ratio of (average number of G or C to A or T substitutions (p))/ (average number of A or T to G or C substitutions (q)). Every row represents the p/q ratios of one genome with the other genomes that are represented in the columns. NA – indicates that substitution ratios between same pair of genomes cannot be calculated. Green, yellow and pink cells represent average p/q substitution ratio between *Crithidia*, *Leishmania* and *Trypanosoma* species, respectively. Purple and cyan cells represent average p/q substitution ratio between the 6 *Leishmania* species, comparing Old and New World species, and within Old World species, respectively; B) Total GC content of Trypanosomatidae and its relationship with the GC content at different codon positions –The red line indicates the least squares line fitted for GC1s vs. genomic GC content (R^2=0.6477), the blue line indicates the least squares line fitted for GC2s vs. genome GC content (R^2=0.4798) and the green line indicates the least squares line fitted for GC3s vs. genomic GC content (R^2=0.6823). R^2 indicates the coefficient of determination.

A linear regression model for ENC ~ A3/T3/G3/C3 was fitted with observed ENC values and the individual base frequencies at 3rd codon positions. It was observed that in *Leishmania* and *Crithidia* genomes, the ENC values for genes are negatively affected by the bases G and C at the 3rd position (indicated by negative Poisson regression coefficients) suggesting that the GC nucleotide composition explains CUB in *Crithidia* and *Leishmania* genes (Table 3.2). Whereas, the relationship of base compositions with ENC in *Trypanosoma* is rather complicated and does not represent a common pattern. Among *Trypanosoma* species, only *T. congolense* and *T. vivax* exhibit a significant negative effect of G and C biased compositions on ENC.

Table 3.2. Poisson regression coefficients of effective number of codons of genes within genome as a function of base compositions at the 3rd codon position.

Species Name	G3 (coefficient)	C3 (coefficient)	A3 (coefficient)	T3 (coefficient)
Cfas	-1.5074*	-1.6487*	2.8631*	2.7862*
Lmaj	-1.462*	-1.5252*	2.9602*	2.7055*
Lmex	-1.4114*	-1.5509*	2.863*	2.531*
Linf	-1.3729*	-1.5365*	2.879*	2.6425*
Lbraz	-1.2057*	-1.4308*	2.2393*	2.0207*
Ltar	-1.1803*	-1.3102*	2.1475*	1.8516*
Ldon	-1.3426*	-1.5397*	2.7757*	2.5066*
Tvx	-0.7881*	-1.2843*	-0.2164*	1.1616*
Tclbe	-0.002	0.0135	-0.0154	0.0098
Tcl	-0.2229*	-0.261*	0.2848*	0.1766*
Tev	0.235*	-0.0452	-0.1823*	0.0772*
Tbt	0.457*	-0.1431*	-0.2821*	0.1188*
Tbg	0.0867*	-0.075	0.1569*	-0.14*

***P* < 0.05**

3.2.3. GC bias at the synonymous position is a mechanism selected for efficient translation of a gene

The previous analyses verified the role of a GC mutational bias in affecting usage of codon across *Leishmania* genes. We asked the question whether translation selection has any role choice of codons within a gene and whether it is related to nucleotide composition at the synonymous position of a codon. Codon Adaptation Index (CAI) is a directional measure of CUB, which quantifies the degree of translation selection acting upon a gene (Sharp and Li 1987). A comparison between ENC and CAI values (Subramanian and Sarkar 2015) would

unveil the relationship between nucleotide composition and selection pressure acting on CUB.

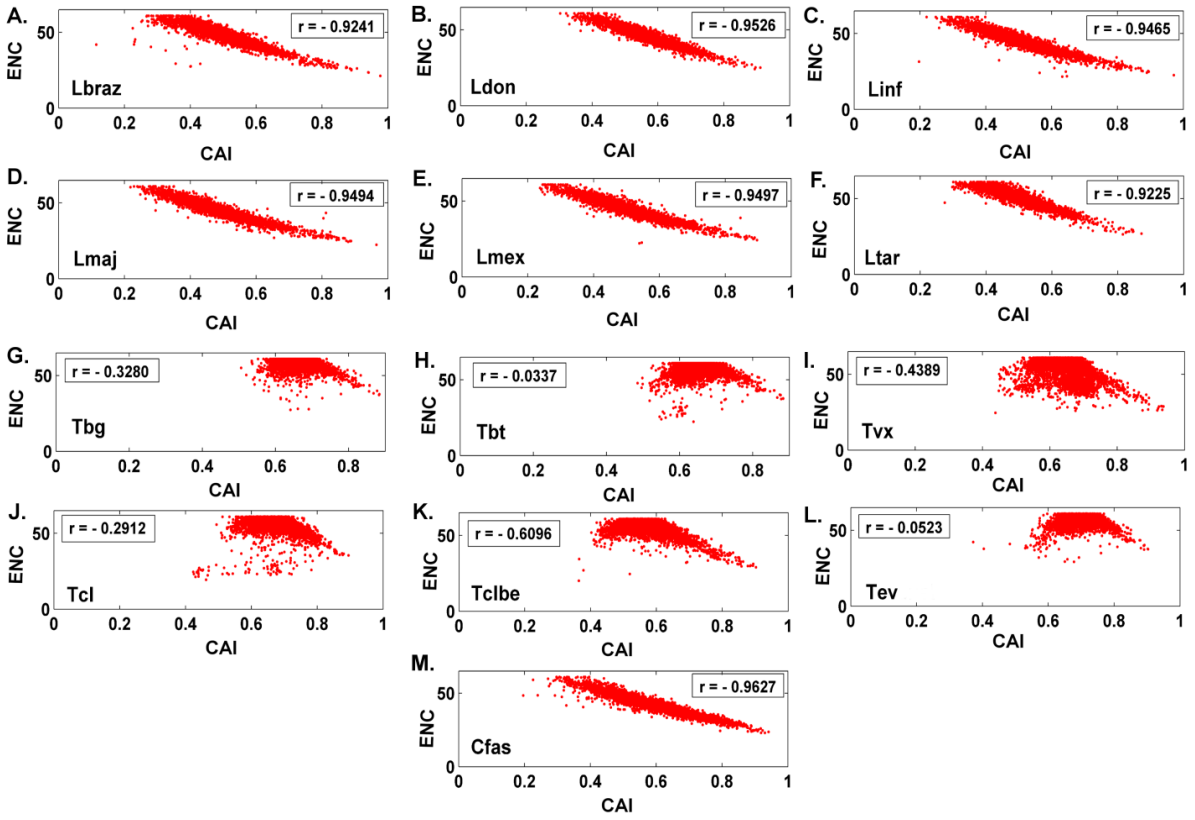

Figure 3.3. Scatter plot of CAI vs. ENC for each of the 13 Trypanosomatid species - The values indicated in the boxes in each plot are Pearson correlation coefficients (*r*) between ENC and CAI. The species abbreviations are also indicated in each plot. Each correlation value is statistically significant ($P < 0.01$).

When ENC was plotted against CAI in the chosen Trypanosomatid genomes, a negative association/correlation was obtained between the two in each species (Fig. 3.3). A negative correlation between ENC and CAI suggest a strong relationship between bias in a gene due to nucleotide composition and bias due to translation selection. A high negative Pearson correlation indicates that genes having biased codon usage due to biased nucleotide composition (lower ENC) tend to be more selected (higher CAI). The correlation values ranged from -0.0337 to -0.9627. The negative correlation reduced from *C. fasciculata* to *Leishmania* to *Trypanosoma.* Moreover, *L. donovani* (Fig. 3.3B) demonstrated the highest negative correlation between ENC and CAI in comparison to other *Leishmania* species, whereas *T. cruzi* (Fig. 3.3K) demonstrated the highest negative correlation between ENC and CAI in comparison to other *Trypanosoma* species. The negative correlation is also maintained throughout the Leishmaniinae genes (Fig. 3.3A-F and Fig. 3.3M). The correlation coefficient

is very close to zero in *T. brucei* TREU 927 (Fig. 3.3H) and *T. evansi* (Fig. 3.3L) suggesting loss of association between ENC and CAI in these genomes. In *Leishmania* species, ENC is highly negatively correlated with CAI suggesting that bias towards particular nucleotide composition in a gene is probably selected for efficient translation. Whereas, in *Trypanosoma* species, the correlation between ENC and CAI is lost, suggesting that bias in nucleotide composition is not governed by translation selection.

All the above analyses indicates that genes with low CAI tend to demonstrate low GC content as compared to genes with high CAI. In all the *Leishmania* species, between CAI of 0.2 and 0.4, genes related to DNA replication, DNA integration, protein phosphorylation, protein folding, and ubiquinone related oxidation-reduction and catabolic processes display a relatively low GC content (40-60% GC). Similarly, between CAI of 0.7 - 0.9, genes related to glycolysis, gluconeogenesis, translation, ATP biosynthesis and hydrolysis, arginine metabolic process and GTP catabolic process display a relatively high GC content (50-70% GC). Further, with respect to known proteomic datasets for the amastigote stage in two *Leishmania* species, CAI values positively correlate with relative protein abundance (measured as number of cognate mass spectra adjusted for gene length) demonstrating a log-linear association (*L. mexicana*; ρ=0.4563, $P < 0.01$, 810 genes and *L. major* ρ=0.2491, $P < 0.01$, 265 genes). This further suggests that CAI values from protein-coding sequences can predict relative steady state protein expression levels. While a large number of other parameters, such as, quality of the proteomic data, number of proteins detected, regulation and control of protein expression, and experimental sampling can also affect this relationship; these results do reinforce the fact that CUB plays a positive role in maintaining steady state levels of proteins.

3.2.4. Codon usage bias in *Leishmania* species is not affected by bias in amino acid composition

To quantify the relationship of CUB with amino acid composition bias, a modification of the ENC index, called as, $N_c(AA)$ was computed for finding bias on particular amino acid codons within a gene (Fuglsang 2003). Average $N_c(AA)$ values of all genes were compared between species and amino acids so as to understand variation in bias of specific amino acid codons (Fig. 3.4A and 3.4B). In 2- and 3-fold degenerate amino acids, isoleucine (Ile) codons display the least bias, whereas tyrosine (Tyr) codons display the highest bias in all Trypanosomatids (Fig. 3.4A). In 4- and 6-fold degenerate amino acids, serine (Ser) codons show the least bias as opposed to valine (Val) codons that tend to show a high bias among Trypanosomatids (Fig. 3.4B).

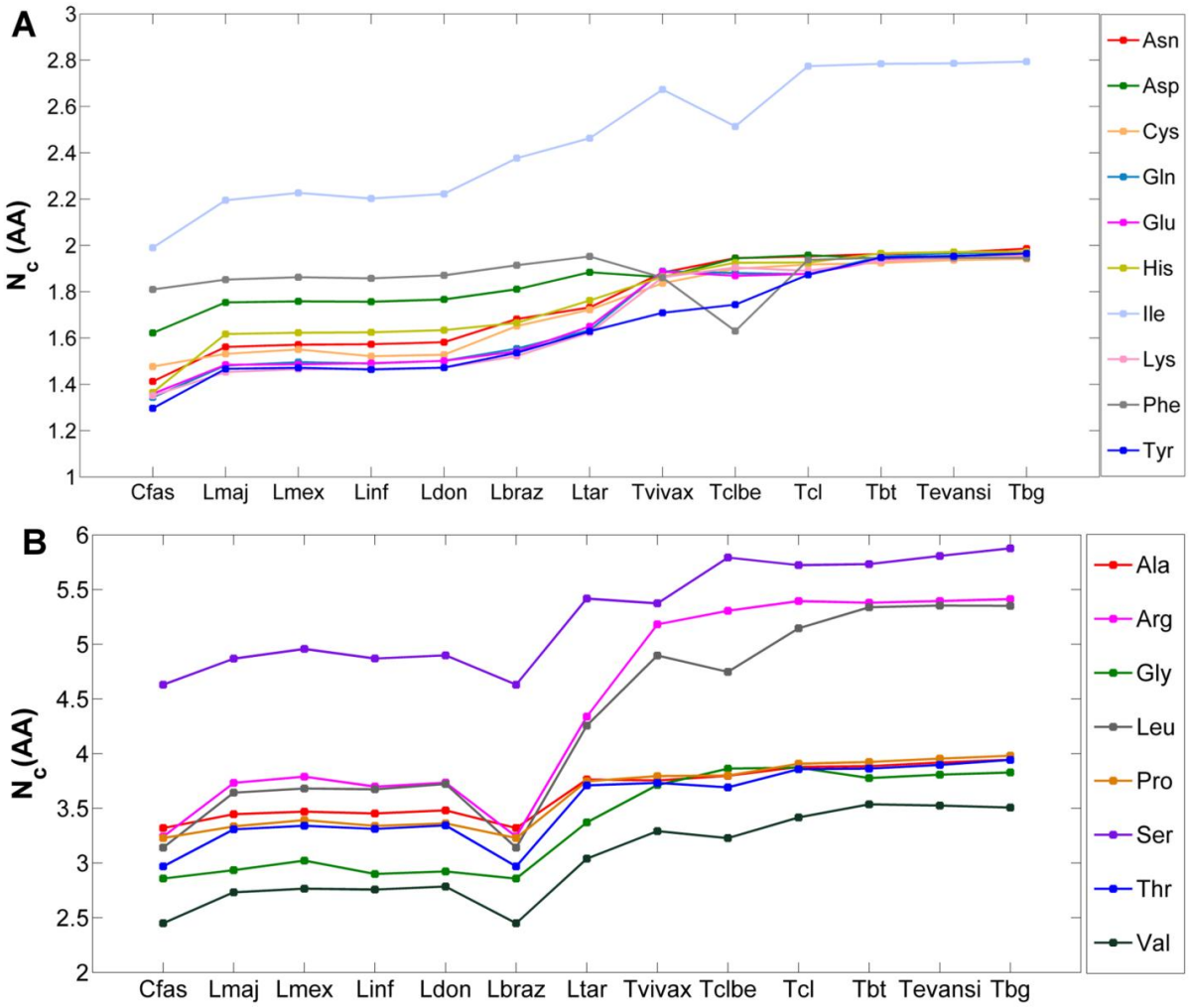

Figure 3.4. Codon usage bias and its relationship with amino acid composition bias and protein abundance - A) Comparison of $N_c(AA)$ in 2- or 3-fold degenerate amino acids among the 13 Trypanosomatid species; B) Comparison of $N_c(AA)$ in 4-or 6-fold degenerate amino acids. The Y axis suggests the average $N_c(AA)$ values of collective biasedness in codons for particular amino acids. For example, $N_c(AA)$ for Ile suggests the degree of bias on isoleucine codons. The amino acids in the figure are presented with respect to standard amino acid three letter codes.

A common observation from $N_c(AA)$ values for all amino acid codons is that bias decreases from *Crithidia* to *Leishmania* to *Trypanosoma* species. With respect to *Leishmania* species, average $N_c(AA)$ values of all amino acid codons are similar in *L. major*, *L. mexicana*, *L. infantum* and *L. donovani* as compared to *L. braziliensis* and *L. tarentolae*. For 2- and 3-fold degenerate amino acids, average $N_c(AA)$ values comparatively show higher values in *L. braziliensis* and *L. tarentolae* than other *Leishmania* species (Fig. 3.4A). For 4- and 6-fold degenerate amino acids, *L. tarentolae* maintains higher $N_c(AA)$ but *L. braziliensis* shows lowest $N_c(AA)$ values as compared to all other *Leishmania* species (Fig. 3.4B). Thus, in *L. braziliensis*, there is a very high bias towards 4- and 6- fold degenerate amino acid codons

whereas in *L. mexicana*, there is a least bias. Similarly, in *Trypanosoma* species, *T. congolense*, *T. brucei* TREU 927, *T. brucei gambiense*, and *T. evansi* tend to show very similar and higher $N_c(AA)$ values, whereas *T. vivax* and *T. cruzi* show lower $N_c(AA)$ values. Also, *T. cruzi* displays the highest bias for all the amino acid codons in comparison to other Trypanosomatids.

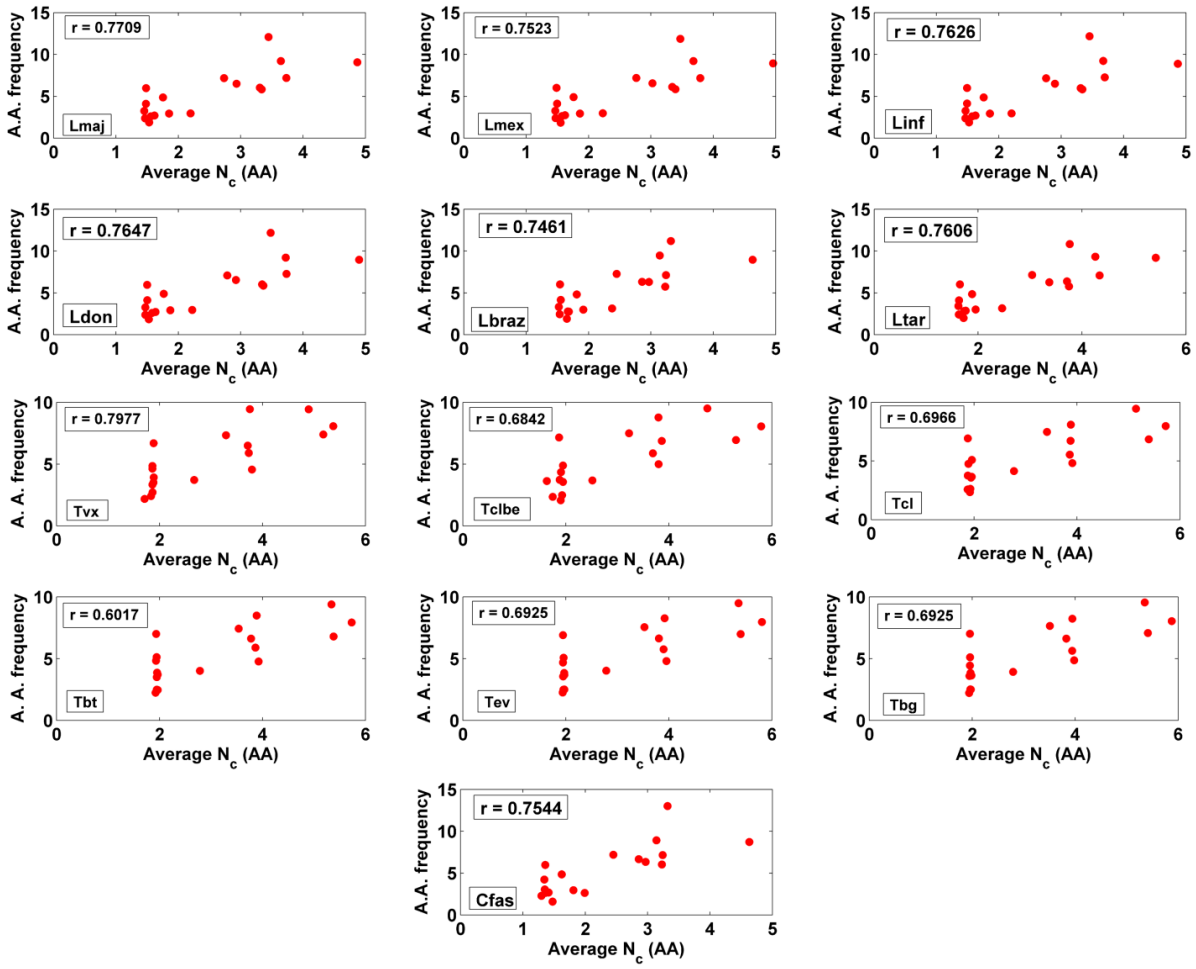

Figure 3.5. Relationship between amino acid frequencies and average N_c *(AA)* values in 13 Trypanosomatids - A positive Spearman correlation could be observed between amino acid frequency and average N_c *(AA)* in each Trypanosomatid species. Here, '*r*' denotes Spearman correlation coefficient. Each correlation value is statistically significant ($P < 0.01$).

To find out whether the average bias in codons of a particular amino acid is related with the frequency of that amino acid within the genome or not, the average $N_c(AA)$ was compared with amino acid frequency. In each of the Trypanosomatids, codons for the most frequent amino acids was found to experience a low degree of bias whereas codons for the least frequent amino acids experience a higher degree of bias (Fig. 3.5). This indicated that codon bias is associated with usage of specific amino acid codons that may not be necessarily

frequent. The relationship between $N_c(AA)$ and amino acid frequency could give an insight into the probable effect of amino acid frequency on codon usage bias. It could be observed that amino acid frequency demonstrates a positive correlation with average $N_c(AA)$ of amino acid codons within each genome (Fig. 3.5). This further suggests that codons for the most frequent amino acids experience a low degree of bias whereas codons for the least frequent amino acids experience a higher degree of bias.

3.2.5. Effect of codon usage bias in mRNA secondary structure formation - a mechanism of translation regulation

Secondary structures in mRNA have been experimentally shown to regulate translation initiation in many eukaryotes (Plotkin and Kudla 2010; Tuller et al. 2010). But, no study that globally inspects the influence of CUB on mRNA secondary structure in Trypanosomatids exists. Hence, to investigate whether CUB affects protein expression via formation of mRNA secondary structure at the 5' end of the open reading frame (ORF), the mean mRNA Folding Energy (MFE) profile of regions subsequent to the start codon was investigated. The MFEs were calculated considering the sliding windows of size 20, 30 and 40 nucleotides for each coding sequence in each species. For all the windows, similar average profiles of RNA folding energies can be observed. As a standard example, the folding energy profiles between the *Leishmania* and *Trypanosoma* species using the 40 nucleotide window was compared (Fig. 3.6). It can be clearly observed that the MFE of the coding sequences across the mRNA sequence in *Leishmania* species (-7.5 to -10.5 kcal/mol) (Fig. 3.6A-F) was comparatively lower than the *Trypanosoma* species (-6.4 to -7.8 kcal/mol) (Fig. 3.6G-L); a probable indication of a high GC bias observed within the *Leishmania* genome or a relative instability in the Trypanosome genomes observed due to high AT bias. Also, experiments in *L. tarentolae* and *Phytomonas serpens* suggest that probable secondary structures in the beginning of the open reading frame following the start codon may affect protein expression (Lukes et al. 2006). To investigate this, MFE of the first 40 nucleotides was compared with the MFEs of the rest of the windows. In *Leishmania* species, first 40 nucleotides demonstrated a significantly higher energy as compared to mean MFEs of initial 44-50 windows that cover around 84-90 nucleotides ($P < 0.05$). Further, in each of the species, MFEs of the initial 40 nucleotides were significantly higher ($P < 0.05$) in the actual genome as compared to the randomized sequence, suggesting that non-folding structures are selected for at the beginning of the ORF. Further, a significant drop in folding energy within the mRNA ($P < 0.05$) for a short window span (around the 18^{th} -26^{th} window for the *Leishmania*

species and the 15th – 39th window for the *Trypanosoma* species) encompassing nearly 48-64 nucleotides from the ORF in comparison to the other windows was predicted.

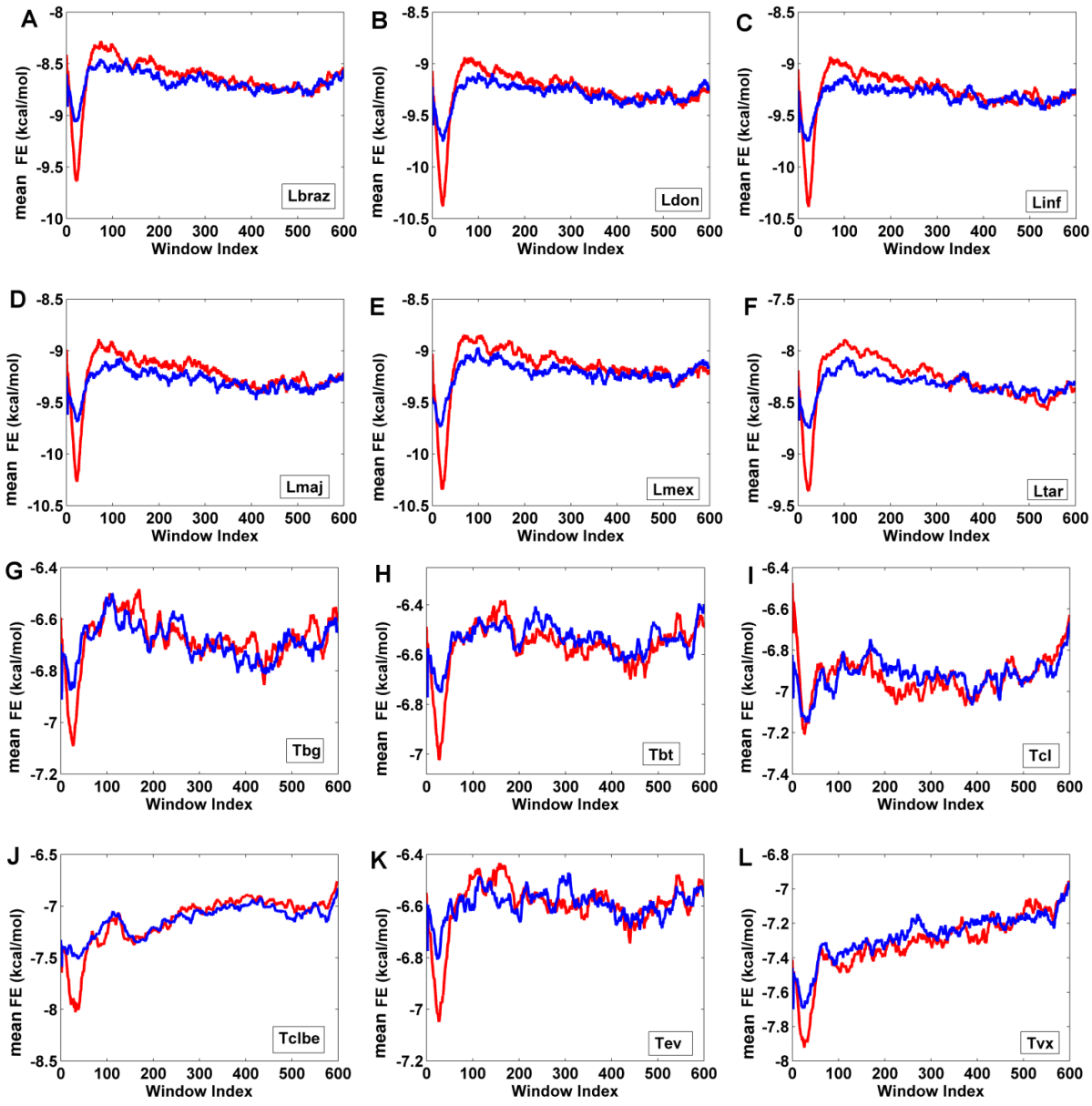

Figure 3.6. The mean folding energy (MFE) profiles of the whole coding set of sequences for each Trypanosomatid genome – A) *L. braziliensis;* B) *L. donovani*; C) *L. infantum*; D) *L. major*; E) *L. mexicana*; F) *L. tarentolae*; G) *T. brucei gambiense*; H) *T. brucei* TREU 927; I) *T. congolense*; J) *T. cruzi* CL Brener EL; K) *T. evansi*; L) *T. vivax*. The red curve indicates the mean folding energy profile for the actual set of sequences in each genome and blue curve indicates the folding energy profile for the randomized set of sequences. The mean folding energy (MFE) is given in kcal/mol. The window index denotes distance in nucleotides between the beginning of the ORF and beginning of the window.

To further identify if the folding profiles were affected by CUB, the actual profiles with MFEs of the randomized sequences were compared. The randomized sequences were

generated from the coding sequences in each Trypanosomatid species by replacing a codon within the coding sequence by a synonymous codon coding for the same amino acid. Comparing the MFE profile of actual set with randomized set in each *Leishmania* species, it was observed that, MFE of the actual coding set was higher as compared to the shuffled set of coding sequences up to the 400th window (444th nucleotide). This result further enhances the understanding that indeed, codons that avoid the formation of folding structures within the mRNA sequences are largely chosen within the *Leishmania* 5' coding regions. Trypanosome genomes as opposed to the *Leishmania* genomes do not demonstrate a common distinct pattern and show a high variability in MFE profiles when compared with the mean randomized energy profiles. Further, no clear distinction could be observed between the original profiles and the randomized energy profiles in *T. congolense*. The 18th – 26th window in *Leishmania* and 15th - 39th window in *Trypanosoma* distinctively demonstrate a lower MFE as compared to the randomized set of sequences possibly to minimize the formation of potentially deleterious structures in the region of the ribosome binding site.

3.2.6. Codon and amino acid contexts among Trypanosomatids

Another mechanism through which efficiency in translation might be mediated, is through the presence of defined paired codon contexts (Irwin et al. 1995; Behura and Severson 2012). Codon context patterns reveal a high variability among Trypanosomatid species. Homogenous codon contexts having high GC content are mostly preferred in *Leishmania* whereas homogenous contexts having bias for AT content are preferred in *Trypanosoma*. The frequency of all possible codon pairs for the 13 Trypanosomatids are provided in the supplementary results of the corresponding published article (Subramanian and Sarkar 2015). GAG-NNN, GCG-NNN, CAG-NNN, CUG-NNN, GUG-NNN, GCC-NNN (where N is any nucleotide) are some of the most preferred contexts across *Leishmania* species. Apart from the contexts considering stop codons, the AUA-AUA, NNN-UAU, NNN-AAU, NNN-AGG contexts have the lowest frequencies in *Leishmania* genome. GAG-NNN, GAU-NNN, GAA-NNN, GUG–NNN, AAG-NNN, etc. are the most preferred contexts across *Trypanosoma* species.

It was also investigated whether amino acid contexts had any effect on preferred codon context. The amino acid contexts demonstrate a striking conservation among Trypanosomatids. The frequencies of all possible amino acid pairs for the 13 Trypanosomatids are provided in the supplementary results of the corresponding published article (Subramanian and Sarkar 2015). Similar to homogenous codon contexts, homogenous

amino acid contexts are also preferred across the 13 Trypanosomatids. Amino acid contexts Arg-Ala and Ala-Pro are highly avoided in *C. fasciculata*. Homogenous contexts consisting of hydrophobic, neutral amino acid pairs like Ser-Ser, Pro-Pro, Gln-Gln, Ala-Ala, Gly-Gly, and charged amino acids like Arg-Arg, Glu-Glu, Arg-Gln are the most preferred contexts in all *Leishmania* species; although, the Arg-Gln context is highly avoided in *L. tarentolae*. The same sets of amino acid contexts are commonly frequent in *Trypanosoma* species as well. Similarly, codon contexts corresponding to amino acids like Ala, Glu, Val and Gln, Ser are highly frequent across *Leishmania* species whereas codon contexts related to amino acids like Glu, Asp, Val, Lys are highly frequent across *Trypanosoma* species. As an exception, in *T. cruzi*, there is a higher preference for lysine context (Lys-Lys and Lys-Glu). Although the role of tRNAs on codon context variation is still unknown in Trypanosomatids, this analysis further suggests that there is a probable high utilization of these amino acids and hence, the preference of the corresponding codon contexts so as to optimize translation efficiency.

3.2.7. Codon usage differences in biological processes across *Leishmania*

Assuming CAI to be a probable predictor of protein expression in *Leishmania*, the effect of CUB on pathway level functions was studied and compared among different species of *Leishmania*. A high CAI value (CAI > 0.5) suggests high relative protein abundance and a high degree of translation selection. Available predicted GO process functions for around 1500-2000 *Leishmania* genes/proteins were considered in this analysis. The remaining genes, annotated as hypothetical, were lumped together into a single 'no' or 'null' process. The percentage of genes belonging to a particular predicted GO function, having a CAI value greater than 0.5 was calculated. The number of genes with low CAI for a set of 19 different GO process-level functions was calculated for each species (Fig. 3.7). The functions that were chosen represent the subset of those pathways/processes to be crucial for parasite survival within the host as reported in literature. It can be observed for each process that a large percentage of genes in *L. donovani* demonstrate higher CAI values as compared to other *Leishmania* species. Also, a high percentage of genes (80-100%) belonging to glycolysis, translation elongation, fatty acid biosynthetic, fatty acid metabolic and glutathione metabolic processes in *L. infantum* and *L. donovani* have high CAI values as compared to other cutaneous *Leishmania* species. Almost 80-100% genes belonging to certain processes like ATP-dependent chromatin remodeling, heme-O-biosynthesis, translation, etc. commonly in all species show high CAI values. The percentage of genes belonging to other functions but having high CAI is listed separately in the supplementary results of the corresponding

published article (Subramanian and Sarkar 2015).

To further identify the genes having highly variable CAI values on a global scale, 4202 single copy orthologous genes common to all 6 species of *Leishmania* were extracted and variations in the CAI levels were analyzed. To get a comparative perspective between the different genes, the coefficient of variation for CAI values of the 4202 genes in the 6 different *Leishmania* species was calculated (Subramanian and Sarkar 2015). Enzymes of electron transport chain, ATP hydrolysis coupled proton transport, isoprenoid biosynthetic process, and purine nucleoside biosynthetic process demonstrate a low variance in codon adaptation. Whereas, enzymes of protein phosphorylation, proteolysis, ubiquitin-dependent protein catabolic process, RNA splicing, and protein folding tend to show a high variance in codon adaptation.

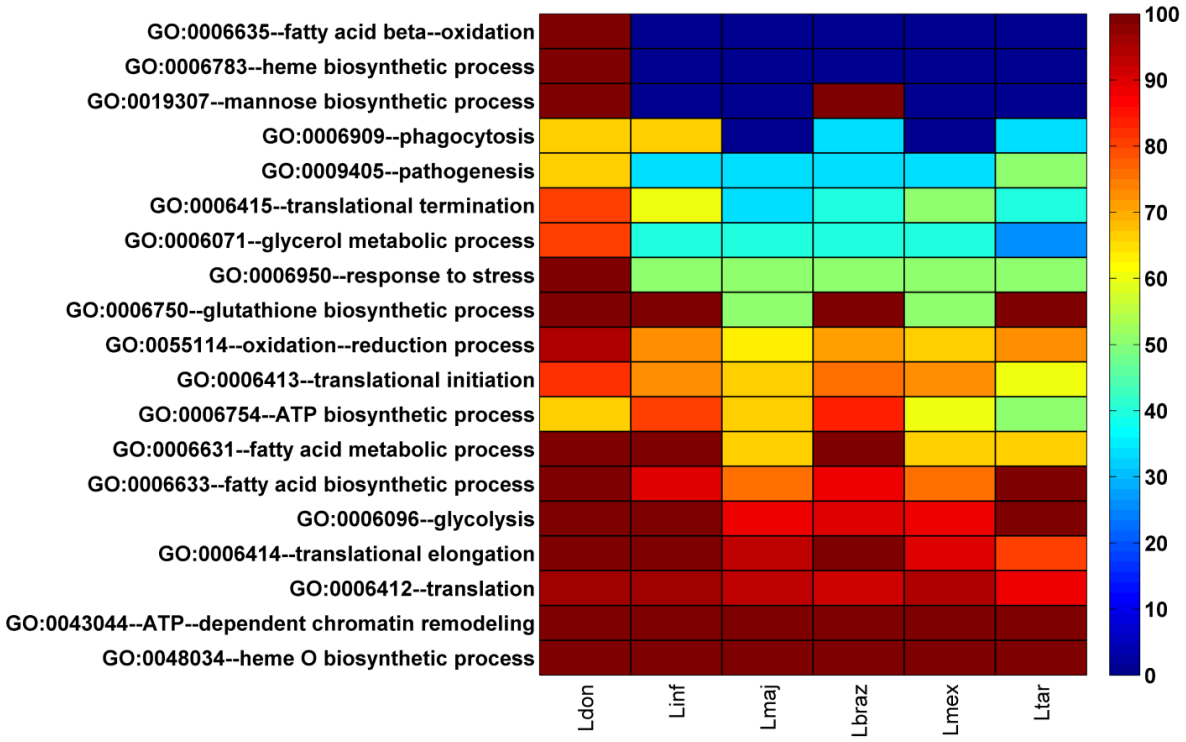

Figure 3.7. Heat map indicating the percentage of genes belonging to a particular GO function/process that have high CAI across *Leishmania* species - The colors in the heat map denote the percentage of genes of a particular function that have CAI > 0.5.

3.3. Discussion

Numerous studies previously used variation in codon usage across *Leishmania* species as a comparative genomics tool (Alvarez et al. 1994; Horn 2008; Chauhan et al. 2011; Singh and Vidyarthi 2011; Rashmi and Swati 2013). Former studies have largely focused on understanding the effect of GC distribution within codons due to evolutionary events like mutational pressure or purifying selection. Also, due to a scarcity of sequence and protein expression data in *Leishmania*, limited either by data for low number of known genes or availability of genomes only for a few species, a strict basis for codon bias across Genus *Leishmania* could not be established. From this large-scale comparative study, patterns of codon usage in 6 different *Leishmania* species was studied and compared with other Trypanosomatid species, for which the assembled genome sequences were available. Through this comparison, the probable causes and consequences of CUB that expose evolutionary and functional differences between *Leishmania* species were enumerated.

The study was largely divided into two phases. In the first phase, the causes of biased codon usage observed within each genome were studied. Codon bias was investigated at different levels to understand the influence of different factors like mutational pressure, translational selection and amino acid composition bias. Previous codon usage studies suggested a high mutational and selection pressure towards G or C in some *Leishmania* and *Crithidia* species (Alonso et al. 1992; Alvarez et al. 1994). Our results further suggest that both mutational pressure and translation selection act on the 3rd codon position to choose codons ending with G or C in all *Leishmania* species. Further, it could be observed that mutational pressure towards maintaining G or C in the genome was high in the Old world species (*L. major, L. infantum, L. donovani*), the highest being in *L. donovani;* as compared to the New world Species (*L. mexicana* and *L. braziliensis*). Among *Leishmania* species, *L. tarentolae* was the most A-T biased probably suggesting an evolutionary adaptation to its cold-blooded host. Also, codons coding for amino acids that were not necessarily frequent in the *Leishmania* proteome had the highest bias suggesting that bias due to amino acid usage does not affect codon usage across *Leishmania* species.

In the second phase, the consequences of CUB on protein expression, mRNA secondary structure formation, and gene function were studied. A positive correlation between codon adaptation and protein expression suggested that CUB was a useful predictor of gene expression in *Leishmania*. Further, it could also be observed that, on an average, codons that avoid secondary structure formation are preferred within the *Leishmania* coding

sequences. This also suggests presence of non-random genome-specific codon usage patterns within each *Leishmania* genome. Our codon context analysis suggests that homogenous codon contexts were more frequent in *Leishmania* and other Trypanosomatids as compared to non-homogenous contexts. This is probably because choice of homogenous codon contexts corresponding to the A and P sites of the ribosome maybe energetically less expensive during translation as the same tRNA can possibly be used twice for aminoacylation of codons with homogeneous contexts. Codon context patterns giving rise to biased amino acid contexts were found to be highly preferred within *Leishmania*. Considering CAI to be a predictor of relative protein abundance, it was identified that genes belonging to a particular predicted GO function had a high CAI. This could be interpreted to indicate that when a high number of genes belonging to a particular pathway have high CAI values, the pathway/process can be assumed to have higher activity as compared to other pathways/processes. House-keeping enzymes belonging to electron transport chain, ATP hydrolysis coupled proton transport, isoprenoid biosynthetic process, and purine nucleoside biosynthetic process demonstrate a low variance in codon adaptation suggesting a constitutive expression of these genes and associated pathways across *Leishmania*. Whereas, enzymes for protein phosphorylation, proteolysis, ubiquitin-dependent protein catabolic process, RNA splicing, and protein folding tend to show a high variance in codon adaptation representing putative differential expression of these pathways between species. A large percentage of genes in *L. donovani* belonging to various pathways demonstrated high codon adaptation suggestive of a putative global up-regulation of genes in *L. donovani* as compared to other species. Further, enzymes related to certain pathways required for adaptive parasite survival, like energy metabolism, demonstrate high CAI values in *L. donovani* and *L. infantum*, which could explain the greater degree of virulence exhibited by visceral species.

Chapter 4 – Genome-scale metabolic reconstruction and analysis of *Leishmania* metabolism

4.1. Introduction

The *Leishmania* parasite demonstrates a digenetic lifecycle that is accomplished within antagonistic host environments. The long-standing association with hosts has equipped the parasite with unique metabolic features that support their existence (Tuon et al. 2008). Mechanisms of metabolic adaptation of the *Leishmania* parasite to a variety of external metabolite sources and their stage-specific utilization still remains an intriguing, unsolved mystery (McConville et al. 2007; McConville and Naderer 2011). Recent ^{13}C-isotope-labelled tracing studies in *L. mexicana* indicate that glucose and amino acids, like aspartate, alanine, proline and glutamate, are catabolized through common metabolic routes employed by both the promastigote and amastigote stages of the parasite and propose stage-specific flux changes of pathways that are hard-wired to the differentiation signals within the parasite (Saunders et al. 2011; Saunders et al. 2014). But, the information related to these core metabolic routes and its implications on other pathways in the *Leishmania* metabolome is still fragmented, incomplete and requires a large scale unified understanding (Opperdoes and Coombs 2007). Information regarding stage-specific metabolic routes even for the core energy metabolism is meager for species other than *L. mexicana*. Moreover, the causal reasons for the choice of specific metabolites and their metabolic routes to satisfy the cellular demand of the parasite are still debated upon. One of the speculated reasons is that the differentiation of promastigote to amastigote brings about a lower uptake of non-essential amino acids, higher fatty acid uptake and hypoxic environment in amastigote metabolism as compared to the promastigote stages, that constrains the choice of metabolic routes (Saunders et al. 2011; Saunders et al. 2014); although the intrinsic adaptation of the metabolic network structure to the environment is less discussed in this context. Also, the metabolic enzyme repertoire across developmental stages in *Leishmania* (DNA microarrays in *L. major* and mass-spectrometry based quantitative proteomics in *L. infantum*) was speculated to be constitutively expressed (Leifso et al. 2007; Saunders et al. 2010), raising the question of how the flux changes between metabolic states for the two stages might be achieved. This is also supported by the observations from whole genome comparisons of codon usage among different species of *Leishmania*, which outline very few differences within metabolic genes between the species (see Chapter 3). All the above observations indirectly provided clues regarding the role of the underlying metabolic network structure in metabolic adaptation of *L. infantum* to the host environment.

*The bulk of this chapter has appeared in *PLoS ONE*, 10(9): e0137976 (2015) and *Scientific Reports*, 7(1):10262 (2017), co-authored by A. Subramanian and R. R. Sarkar

Genome-scale metabolic models of eukaryotic parasites provide a comprehensive overview of metabolic pathways and attempt to understand the stage-specific nature of parasite metabolism (Pinney et al. 2007; Chavali et al. 2008; Plata et al. 2010; Sharma et al. 2017). Sequencing of parasite genomes has paved way to the reconstruction of metabolic pathways from genome data. Metabolic pathway reconstruction essentially includes the identification of enzymes catalyzing metabolic reactions. These enzymes are associated to their appropriate subcellular locations by analyzing the information essentially obtained both from heterogeneous experiments and gene/protein sequence analysis. For large scale metabolic reconstructions where availability of kinetic data for numerous reactions is limiting, the reconstruction can be represented as a constraint-based model and used for predicting catabolic routes under defined biological constraints (Di Ventura et al. 2006; Bordbar et al. 2014). The entire metabolism of the *Leishmania major*, causative organism of cutaneous leishmaniasis was reconstructed using heterogeneous information and subjected to constraint-based analyses to understand whole cell metabolism and predict genes essential for growth (Chavali et al. 2008). But, no metabolic reconstruction was available for the species *Leishmania infantum*, which is presumed to cause infantile visceral leishmaniasis.

As mentioned above, glucose and certain non-essential amino acids have been shown to be utilized in both the life stages of *Leishmania* species and to be essential for either survival or virulence (Saunders et al. 2011; Saunders et al. 2014). Additionally, most enzymes of the glucose catabolism have been shown to be essential for either survival or virulence of trypanosomatids (Michels 1988; Barrett et al. 1999; Louassini et al. 1999; Verlinde et al. 2001; Croft and Coombs 2003). In order to study glucose, energy, non-essential amino acid metabolism and their interplay in within-host survival of *L. infantum* JPCM5 (a well-characterized strain of *L. infantum*), initially a constraint-based model iAS142 that accommodates the pathways related to these sources was reconstructed *de novo*. The iAS142 model encompasses a total of 237 total reactions, out of which 115 reactions could be associated with a total of 142 genes and are distributed across 5 different model compartments: four subcellular – the glycosome, the mitochondrion, the mitochondrial intermembrane space and the cytosol, and an extracellular compartment for exchange of metabolites. Validating this model through *in silico* reaction knockout and robustness analysis for simulating secretion of known overflow metabolites, the scenarios for the promastigote and amastigote metabolic states were created in the model, and the differences in preferred metabolic routes, required for enhanced adaptation in the variable conditions, were predicted.

As no genome-scale metabolic reconstruction was then available for the *L. infantum* parasite, the primary iAS142 reconstruction was expanded on a genome-scale to include a total of 1260 reactions and 1160 metabolites assembled within the iAS556 genome-scale metabolic reconstruction. In both the iAS142 and the iAS556 models, data from *Leishmania*-specific metabolomics experiments was used for simulating genome-scale metabolic adaptations. In the process of reconstructing the iAS142 and iAS556 models, the functions of a few metabolic enzymes and their subcellular locations were newly annotated. Analysis of these models not only captures observations previously reported by metabolomics studies in other *Leishmania* species, but also provides us with new observations on the basis of which, it can be hypothesized that the underlying reaction stoichiometry and reversibility present within the *L. infantum* metabolic network is adequate to explain the stage-specific catabolic route selections, which remain conserved across developmental stages. The corresponding changes in reaction fluxes are strong-armed only providing constraints on uptake of environmental metabolites. Auxiliary to the hypothesis, analyses of the proposed *L. infantum* genome-scale constraint-based model reveals the simplicity of metabolome organization within *Leishmania* and its utility to achieve complex metabolic phenotype traits for optimal usage/synthesis of essential metabolites under varying environments. Related to this, the factors that govern metabolome organization and their effects on distribution of metabolites were also investigated. Model analysis suggest that the coupling of specific reactions for reasons of mass, redox and energy balance, driven by subcellular compartmentalization of enzymes might be the most vital component for appropriate distribution of metabolites within the network. Comparison of the *L. infantum* metabolic network with reconstructions of other *Leishmania* species and existing experimental observations revealed the unique occurrence of enzymes in different subcellular compartments to govern the uniqueness of the *L. infantum* metabolic network structure. This metabolic organization ensures unaltered production of biomass metabolites despite random changes occurring within the parasite metabolome.

The results provided in this chapter highlight the importance of the comprehensive reconstruction studies in understanding the complete adaptations of the *Leishmania infantum* metabolism and in providing biologically feasible predictions. This understanding of the unique biology of *Leishmania* metabolism might further, lead to development of better treatment strategies and provide a means to eradicate the visceral infection.

4.2. Results

4.2.1. The core energy metabolic network model (iAS142)

The iAS142 metabolic network consists of 237 total reactions spanning 5 different subcellular compartments. Out of the 237, around 109 metabolic reactions, 99 transport, 27 exchange reactions, a metabolic demand reaction, and an ATP maintenance reaction constitute the iAS142 network (Fig. A.1). A total of 115 reactions could be associated with 142 genes from the *L.infantum* genome (Table 4.1). Notably, the metabolic demand reaction considered in this study is different from the biomass drain reactions employed by previous similar studies [as it was formulated using isotopic enrichment data acquired from ^{13}C isotope resolved metabolomics (Saunders et al. 2014), see Chapter 2].

Table 4.1. Properties of the iAS142 constraint-based model.

Property	Count
Genes	142
Reactions	237
Gene associated (intracellular)	108
Gene associated (transport)	7
Non-Gene associated (intracellular)	1
Non-Gene associated (transport)	92
Exchange	27
Demand (Metabolic)	1
ATP maintenance reaction	1
Metabolites	231
Compartments	5
Literature References (for reconstruction)	86
Databases Consulted (for reconstruction)	10

Fig. 4.1A depicts a pie chart that classifies reactions in the model according to the pathways in which they fall. Glycolysis, Citric acid cycle and Pentose Phosphate pathway constitute approximately 25% of the total reactions in the model. Amino acid metabolic pathways also represent a major portion of around 10% of the total, demonstrating the importance of integrating them with the carbohydrate metabolic pathways for studying energy metabolism. The metabolic network presented in this study does not encompass the full genome of *L. infantum*, by virtue of which the percentage of the reactions occurring in the cytosol, mitochondria and the cytosol appear to be nearly equal (Fig. 4.1B). Also, around 42% of the

total reactions are present in membranous compartments representing the intra-cellular and extracellular transport reactions for various metabolites. With respect to the pathways considered in the iAS142 network, glycosomes are linked primarily with glycolysis, pentose phosphate pathway and fermentative pathways of C4 dicarboxylic acids; while, mitochondria accounts for TCA Cycle, oxidative phosphorylation, pyruvate metabolism and few amino acid metabolic pathways (Fig. 4.1).

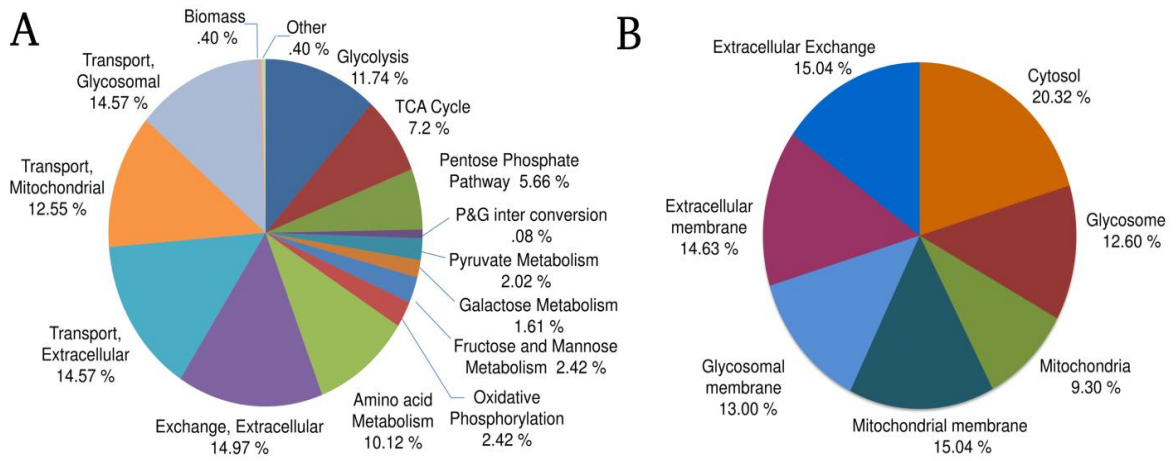

Figure 4.1. Reaction classification (iAS142 model) - A) Pie chart showing the pathways that comprise the iAS142 model; B) Pie chart showing the percentage of model reactions belonging to different subcellular locations.

Annotation of a putative glucose-6-phosphate isomerase

In case of aldose-1-epimerase, there was an ambiguity in assigning a substrate-level function to it. There are two copies of *L. infantum* aldose-1-epimerase protein - UniProt IDs: A4I082, A4IAA4. Both the protein sequences were analyzed for the presence of substrate specificity, if any. A glucose-6-phosphate epimerase (G6PE) was recently annotated in yeast through structural studies (PDB ID = 2CIR) (Graille et al. 2006). Using PyMol version 1.4.1 (DeLano 2002), the residues interacting with the phosphate moiety of glucose-6-phosphate in the 2CIR structure was identified. Structural analysis indicated that Arg-56 (RGGI stretch in protein sequence) and Arg-86 (RNST stretch in sequence) were in close proximity of the phosphate moiety of the substrate glucose-6-phosphate, providing an electrostatic interaction to the phosphate group. The sequence of the PDB structure of yeast G6PE was procured and a pairwise sequence alignment with the two aldose epimerase genes of *L. infantum* was performed. It was ffound that the aforementioned sequence stretches were conserved in the aldose-1-epimerase sequence A4IAA4 (R = 60, RGGV; R = 84, RIRS). Whereas, the other

sequence A4I082 did not show the presence of the amino acid stretches that were present in the yeast glucose-6-phosphate epimerase sequence. The conservation of the two arginine residues in one of the Aldose epimerase (A4IAA4) strongly indicated its substrate specificity towards glucose-6-phosphate. Hence, A4IAA4 was annotated as a "putative glucose-6-phosphate epimerase".

4.2.1.1. Model validations

A) Prediction of known reaction knockout phenotypes

Essential genes required for survival of an organism can be found by performing *in-silico* deletion of genes in constraint-based models by investigating its consequences on the metabolic demand reaction flux. Many reactions in the iAS142 model represent a set of genes associated with the enzyme catalyzing a corresponding reaction. Deletion of any single gene might predict that it is non-essential even though the reaction on the whole, might be essential for parasite growth. Thus, instead of gene deletions, an *in-silico* reaction knockout study can give a better perspective of checkpoints in the pathway that may lead to reduction in growth of the parasite (see Chapter 2). Both single and double reaction knockouts were performed to predict the subset of reactions that are essential for parasite growth (Table 4.2).

Validation of reaction knockouts and predictions

Reactions essential for parasite growth could be predicted by performing single reaction knockouts in the *L. infantum* iAS142 model. 61 reactions out of the total 142 reactions were predicted to be lethal (Table 4.2).

Table 4.2. Reaction lethality predictions (iAS412 model).

Reaction Deletion	Lethal	Trivial lethal	Non-trivial lethal	Non-lethal	Total cases
Single	61	NA	61	153	214
Double	10884	10829	55	33636	44520

Note: Trivial lethal – atleast one of the reactions in the double deletion pair is lethal in a single reaction deletion
Non-trivial lethal, single – reaction involved in the single deletion is lethal
Non-trivial lethal, double – reaction pair involved in the double deletion is lethal
NA – Not applicable

In addition to single reaction deletions, we also performed double reaction deletions of all possible paired combinations of reactions considered in the model (Table 4.2). A total of 10884 lethal combinations could be identified; out of which there were 10829 "trivial" cases

where at least one of the reactions in the pair was lethal in the single reaction knockout study. There were 55 "non-trivial" lethal combinations (see Table 4.2) where both genes involved were not lethal individually, but were lethal in a combination. These predicted novel lethal reaction combinations could be possible combinatorial therapeutic targets, each combination also providing an experimentally testable hypothesis.

Table 4.3. Reaction knockout validations of the iAS142 metabolic network.

Model reaction	Reaction (gene)	Model prediction	Expt. (Known)	Reference (for experiment)	Organism	Predicted wild type growth (%)
ACONTm	Aconitase	lethal	lethal	(Saunders et al. 2014)	*L. mexicana*	0 %
ATPSmm	ATP synthase (Mitochondrial membrane)	lethal	lethal	(Luque-Ortega et al. 2008)	*L. donovani*	0 %
CYOO6mm	Cytochrome-c-oxidase (Mitochondrial membrane)	lethal	lethal	(Luque-Ortega and Rivas 2007)	*L. donovani*	0 %
FBPg	Glycosomal fructose 1,6 bisphosphatase	nonlethal	nonlethal	(Naderer et al. 2006)	*L. major*	99.99 %
MAN6PI	Phosphomannose isomerase	nonlethal	nonlethal	(Garami and Ilg 2001a)	*L. mexicana*	99.99 %
PMANM	Phospho-mannomutase	nonlethal	nonlethal	(Garami et al. 2001)	*L. mexicana*	99.99 %
USPx	UDP-sugar pyrophosphorylase	nonlethal	nonlethal	(Lamerz et al. 2010)	*L. major*	99.99 %

Previously, in the *L. major* iAC560 model study, the growth phenotype information from knockout studies in closely related *Trypanosoma* species was considered for the validation of reaction knockouts in their model (Chavali et al. 2008). But, both bloodstream and insect forms of *Trypanosoma* live in an environment that is entirely different from the *Leishmania* species. Hence, it is inappropriate to explicitly validate the knockout phenotypes of *Leishmania* using *Trypanosoma* data. Experimentally determined growth phenotype information for only a few knockouts (7 knockout phenotypes) is explicitly available for any *Leishmania* species (Table 4.3). Although *Leishmania* specific knockout phenotypic information available to validate this analysis was limited to make a firm statement about model accuracy (Garami et al. 2001; Garami and Ilg 2001a; Naderer et al. 2006; Luque-Ortega and Rivas 2007; Luque-Ortega et al. 2008; Lamerz et al. 2010), the predicted lethality for model reaction knockouts exactly matched with the previously known growth phenotype

information for different *Leishmania* species suggesting the ability of the model in making biologically agreeable predictions (Table 4.3).

The model knockouts were also compared to knockouts predicted from the *L. major* iAC560 model (Table 4.4). The comparison suggests that the *L. infantum* iAS142 model performs better in predicting actual knockout phenotypes with respect to energy metabolism.

Table 4.4. Comparison of knockout phenotypes predicted from iAS142 and iAC560 models.

Reaction name	*L. major* iAC560 predictions	*L. infantum* iAS142 predictions	Experimental phenotype
ACONTm	Non-Lethal	Lethal	Lethal
ATPSmm	Lethal	Lethal	Lethal
CYOO6mm	Non-Lethal	Lethal	Lethal
FBPg	Non-Lethal	Non-Lethal	Non-Lethal
MAN6PI	Lethal	Non-Lethal	Non-Lethal
PMANM	Lethal	Non-Lethal	Non-Lethal
USPx	Not there in model	Non-Lethal	Non-Lethal

B) Model validation through prediction of known metabolic routes required for overflow metabolite secretion

Overflow metabolites are those metabolites which are secreted from the cell due to high uptake of external metabolites. *Leishmania* is known to exhibit an overflow metabolism under high glucose and low oxygen conditions, during which it secretes substantial amounts of overflow metabolites like succinate, acetate, pyruvate, CO_2 and small amounts of lactate (Keegan and Blum 1990; ter Kuile 1999; McConville and Naderer 2011). As a second level of model validation, uptake constraints of glucose and oxygen were varied to predict the secretion of these metabolites through the iAS142 model. At the optimum solution of oxygen uptake, clearly there was no secretion of overflow metabolites (Fig. 4.2A). At the optimum solution of oxygen uptake, glucose is catabolized completely for production of succinate which enters into the TCA cycle instead of being secreted out, so as to maximize biomass and ATP synthesis. This result indicates that the parasite demonstrates a glucose-dependent metabolism for its survival, where it is well adapted to maximize its metabolic demand and ATP synthesis by efficient utilization of glucose and oxygen from the environment. But, as the flux value of oxygen uptake decreases, acetate immediately starts to be secreted; followed by succinate under further oxygen decrement (Fig. 4.2A). The glucose uptake rate decreases in a piecewise linear fashion with respect to the decrease in oxygen uptake rate. Albeit, fixing the lower bound of glucose uptake to an arbitrary value of 200 consequently displayed the

secretion of pyruvate and succinate (Fig. 4.2B). The bounds of glucose uptake were fixed to 200, as pyruvate secretion is observed for glucose uptake between 190 and 210 under constant low oxygen uptake. Similarly, constraining both the upper and lower bounds of glucose uptake rate to 280 led to the secretion of lactate (Fig. 4.2C). This is because, glucose uptake between 0 and 280 under constant low oxygen uptake leads to lactic acid secretion. Under this constraint, glucose is distributed between the preferred succinate fermentation pathway and the D-lactate dehydrogenase reaction thereby leading to secretion of succinate and lactate. Under anaerobic constraints, high uptake of glucose leads to its conversion into succinate and lactate fermentation. The constraint on glucose uptake rate was removed when oxygen uptake rate decreased to a flux value of 40, as there is no solution of the simulation in that range, indicating infeasibility of such constraints in a real situation (Fig. 4.2C).

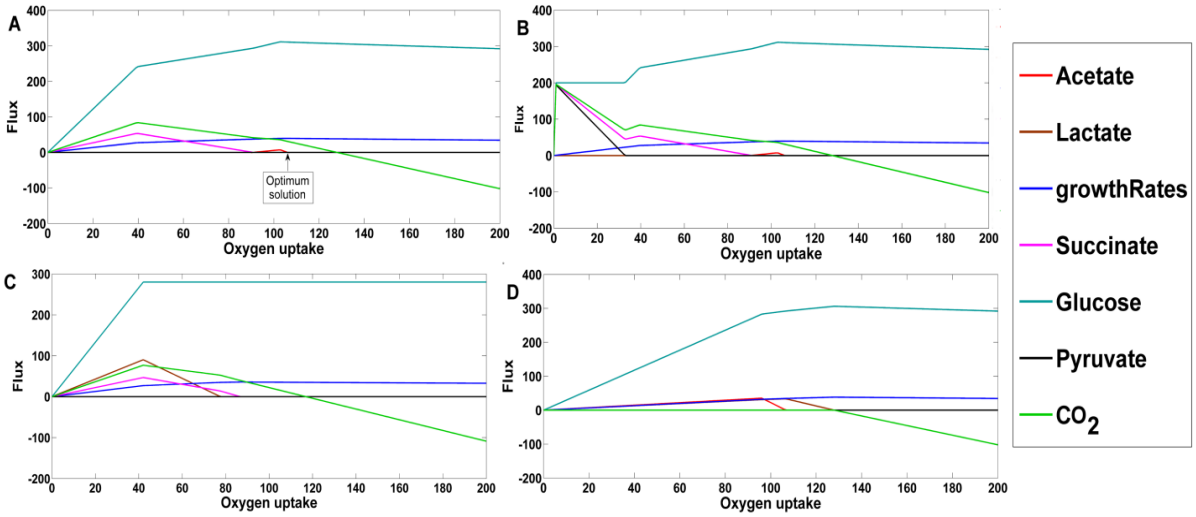

Figure 4.2. Robustness analysis with respect to oxygen uptake, to simulate overflow metabolite secretion - A) Secretion of different overflow metabolites and glucose uptake with variation of oxygen uptake; B) Secretion of different overflow metabolites and glucose uptake with variation of oxygen uptake by fixing lower bound of glucose uptake to a flux value of 200; C) Secretion of different overflow metabolites with variation of oxygen uptake by fixing upper and lower bounds of glucose uptake to a flux value of 280. The constraints for glucose uptake were changed to its default values at oxygen uptake rate <= 40 as the simulation gave no solution; D) Secretion of different overflow metabolites with variation of oxygen uptake and fixing lower bound of CO_2 to a value of -1000 and upper bound to 0.

For standard simulations, CO_2 transport was kept positive for flux values through oxygen uptake just below 130. To understand the importance of CO_2 uptake, CO_2 uptake was constrained so as to utilize CO_2 from the environment and the simulations were contined by varying the flux through oxygen uptake (Fig. 4.2D). From the figure, it can be observed that

secretion of succinate halted spontaneously and lactate and acetate secretion commenced profoundly. These results are allusive of the interplay between the glucose and oxygen uptakes, which may influence the final metabolites that are to be secreted. Moreover, CO_2 may be essential for the secretion of succinate during overflow metabolism.

4.2.1.2. Comparison of *L. infantum* iAS142 with other Trypanosomatid reconstructions

There are numerous differences observed between the energy metabolic reconstructions for *Leishmania* and other Trypanosomatids. To identify these differences, the iAS142 reconstruction has been compared with the *L. major* iAC560 and the *Trypanosoma cruzi* iSR215 model (Figs. 4.3A and 4.3B) (Chavali et al. 2008; Roberts et al. 2009). The differences in intracellular reaction subcellular location between models are listed in Table B.1. Around 17% of the model reactions (exchange reactions excluded) are unique to the iAS142 model and have been newly curated. The comparisons indicate that the iAS142 model accounts for 36 novel intracellular reactions (17 compartmental + 19 intracellular transport) and 9 metabolites updated for their subcellular locations. 16 distinct reactions are common to iAS142 and iAC560 and not to iSR215 and 2 distinct reactions are common to iAS142 and iSR215 and not to iAC560. 17 reactions are unique to the iAS142 model out of which 9 reactions have been newly added in the iAS142 model and 8 reactions have been updated for their reaction subcellular locations. These differences attribute the iAS142 model with a unique network structure that leads to prediction of biologically realistic scenarios and its role was further investigated in the expanded genome-scale metabolic network of *L. infantum*, discussed within this chapter.

Further, to establish the uniqueness of the iAS142 network in predicting biologically realistic scenarios, we simulated the iAS142 model and the energy metabolism subset of the iAC560 model and compared the metabolic routes selected for glucose catabolism in both the models. As it can be observed from Fig. 4.3C, the flux profiles of iAS142 and energy metabolism part of iAC560 models are similar (Spearman correlation coefficient (r) = 0.6672). Interestingly, flux through biomass reaction in iAC560 energy metabolism was relatively higher than flux through biomass reaction in iAS142 model. Further, the flux across individual reactions were considerably altered which was captured from the scatter plot. The *L. major* model displayed a high rate of lactate exchange from cytoplasm to extracellular and higher intake of pyruvate from extracellular to cytoplasm; whereas in the *L.*

infantum model, there is neither any secretion nor uptake of pyruvate or lactate (Fig. 4.3D). As there are significant differences in the occurrence of enzymes in different subcellular locations between the two models (Table B.1), the flux distribution also varied between the two models. Instead of succinate fermentation which was observed in the iAS142 model, an increased dependence of the iAC560 model on lactate fermentation and hence, pyruvate uptake was observed. This major difference was due to the presence of a cytoplasmic lactate dehydrogenase and a cytoplasmic alanine aminotransferase in the *L. major* iAC560 model, which was curated to occur in the glycosome and the mitochondrion within the *L. infantum* iAS142 model.

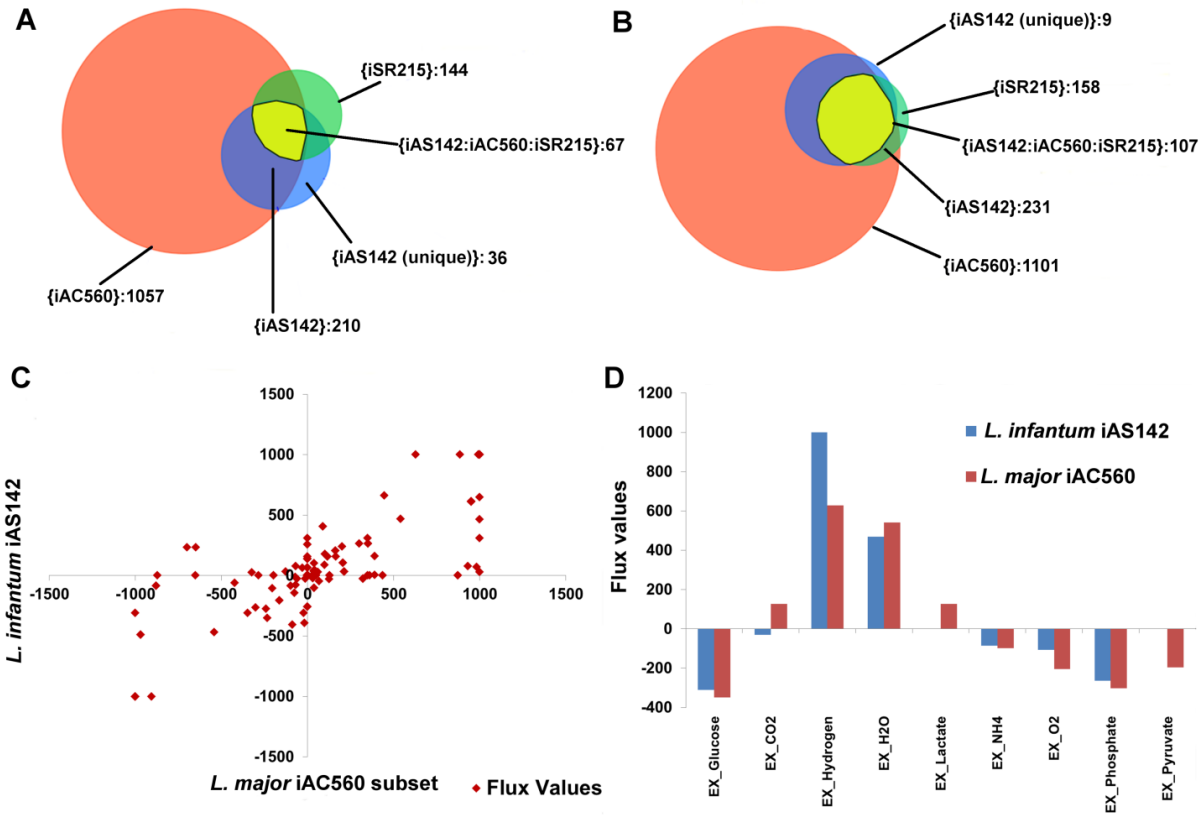

Figure 4.3. Comparison of the *L. infantum* iAS142 network with other Trypanosomatid reconstructions – A) Venn diagram showing the comparison between intracellular reactions (exchanges excluded) considered in the *L. infantum* iAS142, *L. major* iAC560 and *T. cruzi* iSR215 reconstructions; B) Venn diagram showing the comparison between the metabolites considered in the *L. infantum* iAS142, *L. major* iAC560 models, and *T. cruzi* iSR215 reconstructions [the brackets '{}' represent the set of elements constituting the particular area in the Venn Diagram]; C) Scatter plot showing comparison of the *L. infantum* iAS142 model with energy metabolism subset of *L. major* iAC560 model; D) Comparison of secretion of overflow metabolites between the *L. infantum* iAS142 and the energy metabolism subset of *L. major* iAC560 model – bar graph representing exchange fluxes from both the models.

Performing FBA on the *L. major* iAC560 energy metabolism subset, it was observed that cytosolic lactate dehydrogenase and alanine transaminase form a cycle along with their glycosomal (lactate dehydrogenase) and mitochondrial (alanine aminotransferase) counterparts respectively. This produces a sink for the pyruvate that is produced from mitochondrial alanine aminotransferase, pyruvate dehydrogenase reactions and extracellular pyruvate uptake. This cycle causes fermentation of lactate to be preferred over succinate. This requirement of lactate production leads to a corresponding higher activity of the pentose phosphate pathway (PPP) reactions to maintain redox balance within the glycosome in the iAC560 model. This leads to an over production of CO_2 from the 6-phopshogluconate dehydrogenase reaction which is released outside the cell to the extracellular (Fig. 4.3D). On the contrary, in the *L. infantum* iAS142 model, succinate fermentation is preferred to maintain redox balance within the glycosome, and thus, drives CO_2 uptake rather than release. Hence, the iAS142 model shows a minor uptake of CO_2 from extracellular to cytoplasm (Fig. 4.3D). The comparison between the flux profiles of the reactions from the two models leading to production of lactate revealed a significant difference ($P < 0.01$).

4.2.1.3. Effect of amino acids on metabolic fluxes when supplemented with glucose

L. mexicana developmental stages were shown to co-utilize glucose and a few non-essential amino acids like glutamate, aspartate, alanine and proline, when cultured in a completely defined medium consisting of a range of carbon sources (Saunders et al. 2010; Saunders et al. 2014). In order to understand the importance of these non-essential amino acids in energy metabolism of *L. infantum*, each amino acid was supplemented with glucose uptake separately. Performing FBA for each of these cases led to the generation of input-specific flux distributions, displayed in the form of a heat map (Fig. 4.4). The default simulation results where glucose is the only nutrient are also shown for comparison. It was observed that, when different non-essential amino acids supplement glucose as nutrients, fluxes of glycolysis, glutamate biosynthesis and glycine/serine biosynthesis reactions vary the most. Clearly, the fluxes through pentose-phosphate pathway and the citric acid cycle reactions are fairly constant, suggesting their housekeeping functions within the cell.

With respect to the default uptake bounds between -1000 to 1000, the reaction fluxes of TCA cycle reactions lie in the small range of -200 to 200, suggesting that molecular feed into the TCA cycle is quite low. At the same time, majority of glycolytic reactions retain higher flux values, between ranges 800 to 1000 and -800 to -1000 for all the scenarios. This

is a very important aspect of *Leishmania* metabolism, which suggest that utilization of glucose is always very prominent albeit the presence of other carbon sources. Also, growth was reduced to zero when the exchange of glucose was constrained to zero, in spite of the abundance of amino acids. This strongly suggested that glucose is the prime essential substrate for energy metabolism. Here, it is important to note that the iAS142 model does not take the fatty acid metabolism into consideration. To consider the effect of fatty acids, an acetyl-coA uptake was added to the model because acetyl-coA is the refined product of fatty acid catabolism. But even after this, there was no growth in the absence of glucose, arguing that the parasite probably cannot metabolize acetyl-coA under complete absence of glucose.

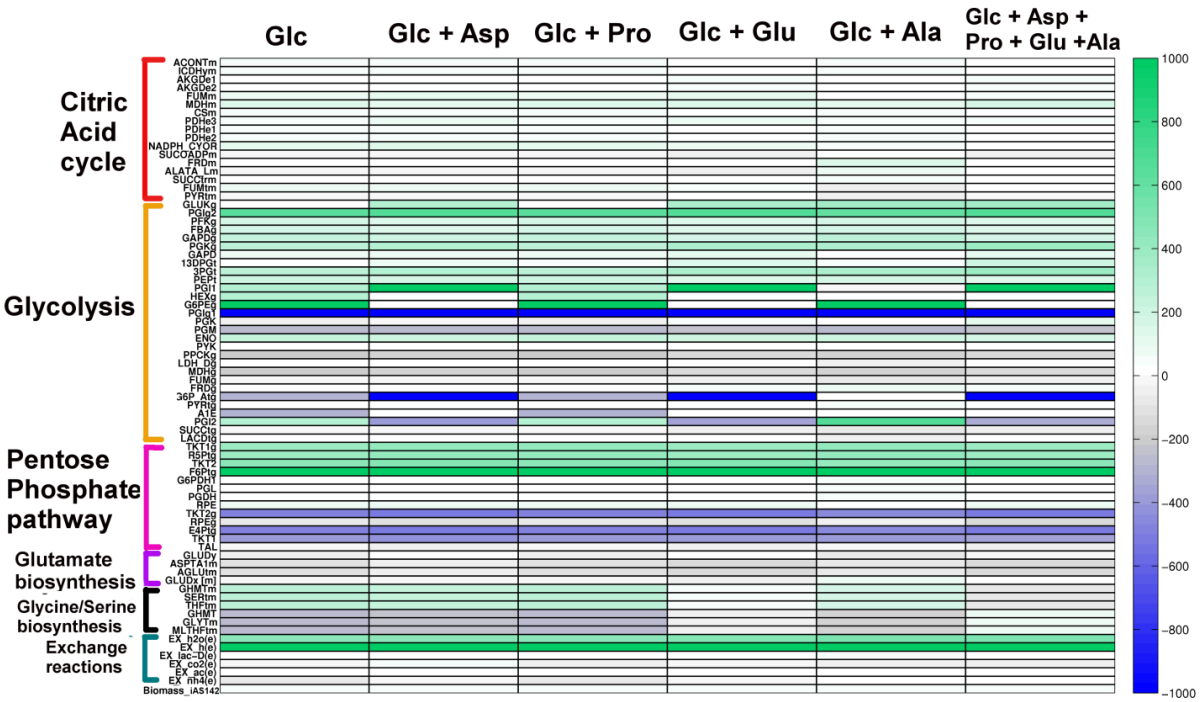

Figure 4.4. Effect of amino acids when coupled with glucose uptake - Heatmap to visualize differences in flux through crucial reactions for different scenarios. The amino acids were allowed to enter one at a time along with glucose to identify different metabolic routes operational in diverse conditions. Glc=Glucose, Asp=Aspartate, Glu=Glutamate, Ala=Alanine, Pro=Proline.

4.2.1.4. Choice of biomass objective function affects model flux distribution

In a previous energy metabolic model of *Trypanosoma cruzi* (iSR215), a Trypanosomatid related to *Leishmania*, an objective function (Tryp_biomass) derived from the biochemical data of *Bacillus subtilis* was used for performing flux balance analysis on the model (Roberts et al. 2009). Whereas, the objective function used for performing FBA in the iAS142 model was derived from the ^{13}C isotope enrichment data available for *L. mexicana* (Saunders et al. 2014). Therefore, both these biomass objective functions were chosen separately within the

iAS142 model for performing FBA and the results were compared. In Section 4.2.1.3, it was observed that glucose uptake is linked with the uptake of non-essential amino acids.

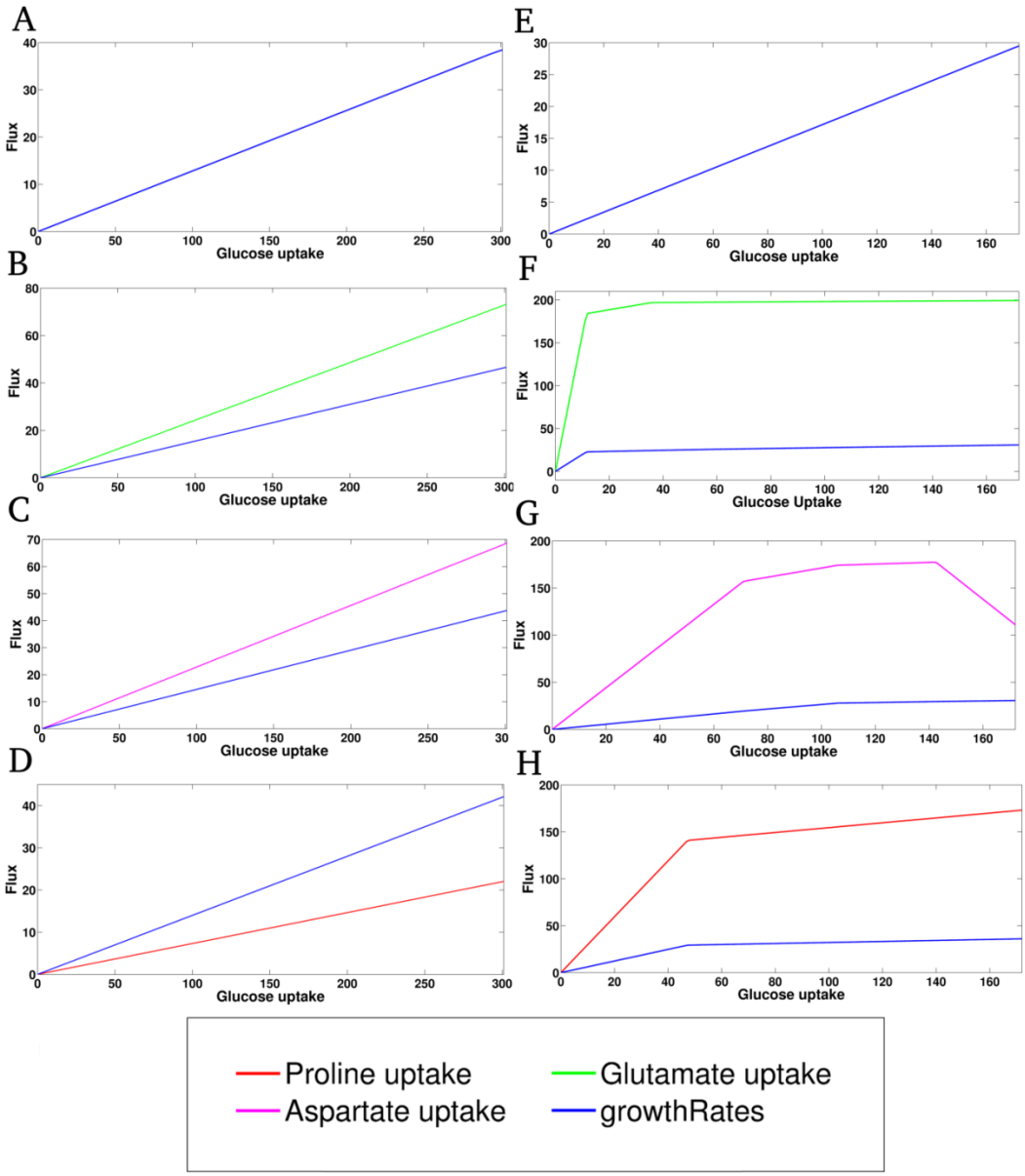

Figure 4.5. Comparison of the iAS142 metabolic demand reaction with the iSR215 biomass reaction – A) and E) represent situations where glucose is the sole substrate with respect to the two biomasses respectively. In all the other cases an amino acid supplements glucose uptake, which can be identified by their respective colors; B), C), and D) depict the effect of variation in glucose uptake on fluxes through glutamate, aspartate and proline uptake respectively, using the iAS142 biomass reaction. Effect on the objective function is also recorded in each case; F), G), and H) demonstrate the behaviour when iSR215 biomass reaction is used.

Thus, a robustness analysis of amino acid uptake and the metabolic demand reaction with respect to varying glucose uptake was performed to visualize any deviation resulting from the usage of different biomass objective functions (Fig. 4.5). The upper and lower bounds of

glucose uptake in the model for each biomass reaction was fixed to a value that was incrementally increased within the bounds of 0 to the optimal value of glucose uptake (optimal flux value = 310 for the default bounds between -1000 and 1000) obtained through the default FBA simulation. As it is very much evident from the figure, the *L. infantum* energy metabolic demand and amino acids uptake varies linearly with glucose uptake for all situations when the iAS142 metabolic demand is used (Figs. 4.5A-D). On the other hand, using the iSR215 biomass, linear relationship between growth rate and glucose uptake could be perceived only for the situation, where glucose was used as the sole carbon source (Fig. 4.5E). For rest of the other cases where amino acids supplemented glucose, both amino acid uptakes and biomass maintained a piece wise linear relationship with glucose uptake (Figs. 4.5F-H).

4.2.1.5. Stage specific energy metabolism of *Leishmania infantum*

As discussed before, the *Leishmania* parasite is able to persist in two extremely opposite environments, the sandfly gut and the human macrophage by switching between the promastigote and amastigote developmental stages, each stage adapted to survive in a particular environment. In order to identify the metabolic routes by which *L. infantum* achieves this feat, model conditions representing the promastigote and amastigote metabolic states were re-created. Details behind re-creation of these stage specific scenarios are discussed in Chapter 2.

Flux through around 90% of reactions in amastigote metabolism was substantially reduced when compared to the promastigotes (Fig. 4.6A). Around 113 reactions out of the total 237 reactions in the model completely shut down (no flux) in the amastigote flux profile (Fig. 4.6A). Around 11 reactions were uniquely functioning in amastigotes and absent in promastigotes. In the amastigote metabolism, glucose uptake reduced by 10 fold and glutamate uptake rate reduced by almost 15 fold when compared to the promastigote profile justifying the glucose-deficient conditions in which the amastigote survives (Fig. 4.6B). Also uptake of other non-essential amino acids, like aspartate and alanine, were considerably reduced. A similar qualitative reduction in glucose and non-essential amino acid uptake rate was previously reported in the *L. mexicana* amastigotes in comparison with the promastigote stage (McConville and Naderer 2011; Saunders et al. 2014). The metabolic flux through the glycolysis, tricarboxylic acid cycle (TCA), pentose phosphate pathway (PPP) and glutamate dehydrogenase in the amastigote metabolic state significantly drops (Figs. 4.6D-G). This leads to a reduced secretion of overflow metabolites (Fig. 4.6B).

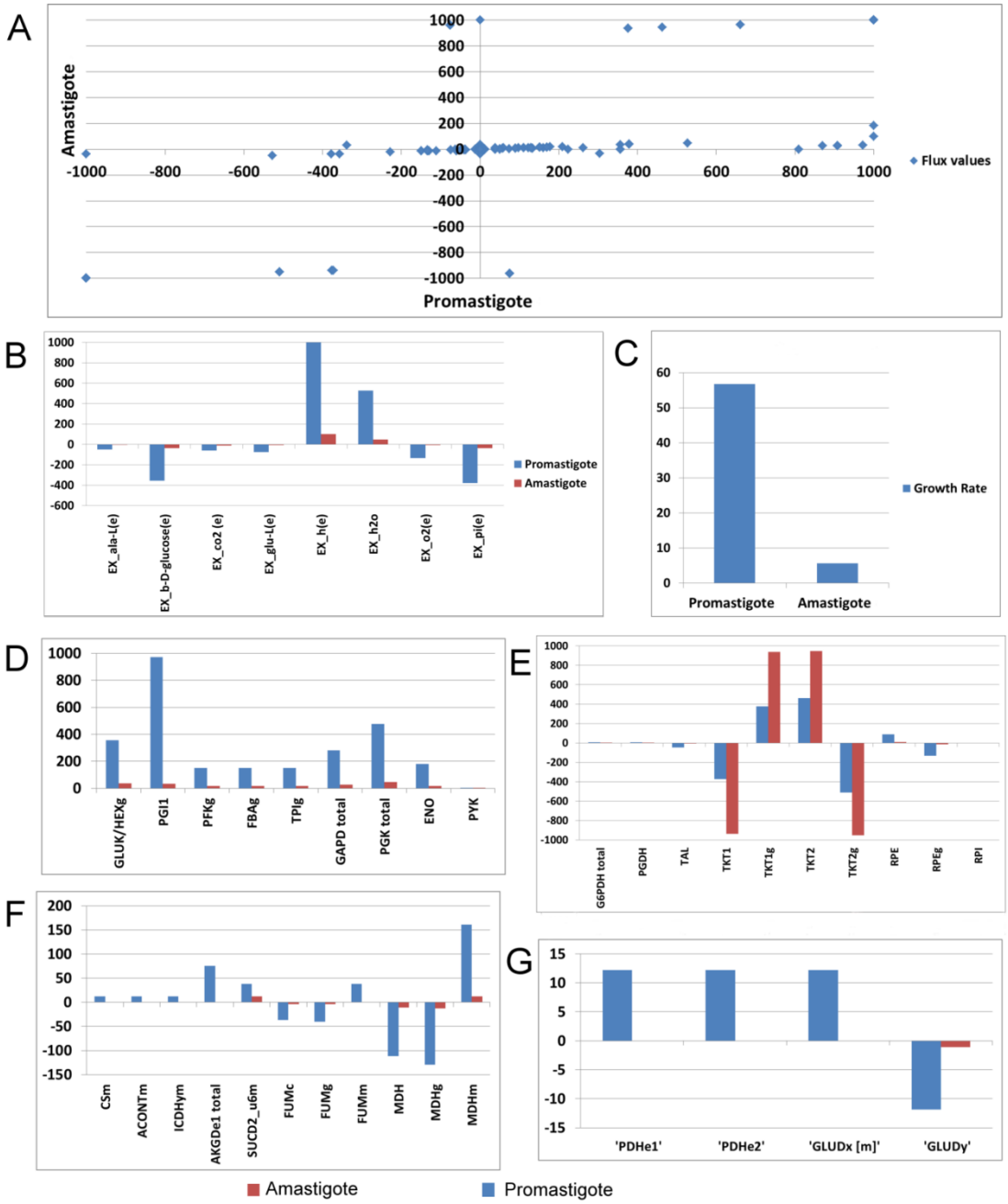

Figure 4.6. Comparison of flux distributions between the promastigote and amastigote scenarios – A) Scatter plot showing the variation of fluxes from the promastigote to amastigote forms; B) Bar plot showing differences between overflow metabolism in the two developmental stages of *L. infantum*; C) Bar plot showing the differences in promastigote and amastigote growth as predicted from the model; D) differences observed in the glycolytic reactions in the amastigote and promastigote stages; E) differences in the pentose phosphate pathway observed between the two stages; F) differences observed in the TCA cycle reactions in the two stages; H) differences between other important reactions in both the stages.

Since the real phenotype data specific to *L. infantum* was unavailable, the stage specific reaction fluxes observed from our model could be qualitatively compared to the behavior of enzymes of carbohydrate metabolism in the closely related visceral species *L. donovani*

(Meade et al. 1984). The change in the reaction fluxes of these pathways qualitatively relate to the changes observed in the specific activity of enzymes required for carbohydrate metabolism in the amastigote forms of the visceral *L. donovani* species (Meade et al. 1984). Even though the information used for the above comparison was not *L. infantum* specific, as they are closely related species, probably there would not be a drastic difference in the energy metabolism between these species.

Contrary to this belief, the optimal solution for oxygen exchange suggested a reduction in oxygen intake in the amastigote scenario as compared to the promastigote scenario, probably signifying the adaptation of amastigote metabolism to the hypoxic environment of the macrophage. As amastigotes reside in an acidic environment, it is quite intuitive that the amount of hydrogen transferred from the cell to the environment would be very low. Hence, amastigote metabolism was adapted to a high reduction in hydrogen ion release to the environment (Fig. 4.6B). The observed reductions in glucose and oxygen uptake perhaps led to rapid decrease in amastigote growth rate (Fig. 4.6C).

4.2.2. The genome-scale metabolic network model of *L. infantum* (iAS556)

The iAS142 model was expanded to integrate and include all the metabolic genes present within the *Leishmania infantum* genome. Known biological information and *in-silico* prediction of reaction subcellular locations and functions were integrated collectively to obtain a gene-protein-reaction (GPR) framework specific to the metabolism of *L. infantum.* The GPR framework was constructed with respect to reactions involved in production and transport of metabolites by following a stringent reconstruction strategy considering heterogeneous data, tools, supporting analysis and the assigned cellular compartments (see Chapter 2). A list of reactions, their corresponding genes, enzymes, and metabolites were compiled with appropriate literature support, in combination with sequence analysis to substantiate evidence related to reactions, its biological significance, and location. These metabolites and the corresponding reactions were distributed into 9 different model compartments i.e. glycosome, cytosol, mitochondrion, mitochondrial inter-membrane space endoplasmic reticulum, acidocalcisome, vacuole, nucleus and extracellular space. All information was organized in the rBioNet toolbox (Thorleifsson and Thiele 2011) to eventually obtain an extensive model reconstruction, iAS556 for *L. infantum* genome-scale metabolism.

Accordingly, the iAS556 metabolic network consists of 1260 total reactions spanning 9 different subcellular compartments. Out of the 1260, around 645 metabolic reactions, 478

transport, 136 exchange reactions, and a metabolic demand reaction constitute the iAS556 network. A total of 647 reactions could be associated with 556 genes from the *L. infantum* genome (Table 4.5). More importantly, the metabolic demand reaction considered in generating the present GPR network varied from the biomass drain reactions employed by previous similar studies since it was formulated using relative availability of metabolites in the log phase promastigotes as detected from metabolomics experiments (Silva et al. 2011) Refer Chapter 2 for further details. There are 136 exchange reactions included in the model, which is a high number compared to the previous genome-scale reconstructions in Trypanosomatids. This can be attributed to the presence of exchanges of comprehensive set of metabolites that are experimentally known to be salvaged by the parasite from the host environment.

Carbohydrate metabolism constitutes approximately 7% of total reactions in the model. Amino acid metabolic pathways represented a major portion of around 12% of the total, demonstrating their importance in metabolism of *Leishmania* (Fig. 4.7A). The iAS556 metabolic network encompasses the full genome of *L. infantum* majorly comprising of 25% cytosolic, 15% mitochondrial and 9% of glycosomal & endoplasmic reticulum reactions (Fig. 4.7B). Additionally, around 49% of the total reactions were present in membranous compartments representing the intra-cellular and extracellular transport reactions for various metabolites.

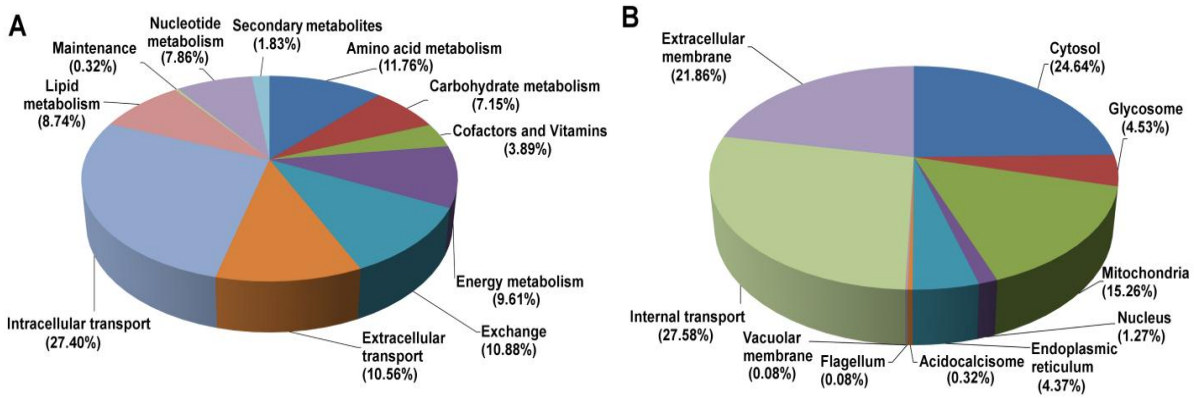

Figure 4.7. Reaction classification (iAS556 model) - A) Classification of model reactions into their metabolic pathways; B) Classification of model reactions into their compartments.

The iAS556 reconstruction was compared with the previously known *L. major* iAC560 constraint-based model (Chavali et al. 2008). Around 29% of the model reactions (excluding exchange reactions) are unique to the iAS556 model and have been newly curated (Figure

A.2). The comparisons indicate that the iAS556 model accounts for 223 novel intracellular reactions (35 compartmental + 188 intracellular transports). These differences arise due to inappropriate assignment of reaction subcellular locations, unavailability of appropriate information for certain reactions, absence/existence of multiple subcellular location of enzymes, absence of important intracellular transport reactions reported through literature, and absence of mitochondrial inter-membrane space compartment within the iAC560 model, when compared to iAS556 model. These major gaps have been filled in the iAS556 metabolic network. Also, one of the major updates within the *L. infantum* iAS556 model was the occurrence of few fatty acid enzymes (that convert NAD to NADH) within the glycosome, which was previously presumed in other reconstructions to occur within the cytoplasm. The reason for this placement was the NAD redox coupling that exists between fatty acid oxidation and glycolysis. This was also supported through proteomics experiments in *L. donovani* (Jamdhade et al. 2015). The differences in intracellular reaction subcellular location between the two models are enlisted in Table B.2. The confidence score for these reactions have also been indicated to emphasize the confidence with which their subcellular locations were determined (see Chapter 2).

Table 4.5. Properties of the *L. infantum* iAS556 network reconstruction.

Property	Count
Genes	556
Reactions	1260
(i) Gene associated (intracellular)	623
(ii) Gene associated (transport)	24
(iii) Non-Gene associated (intracellular)	22
(iv) Non-Gene associated (transport)	454
(v) Exchange	136
(vi) Demand (Metabolic)	1
Metabolites	1160
Compartments	9
Literature References (for reconstruction)	~160
Databases Consulted (for reconstruction) (Subramanian et al. 2015)	10

Uniqueness of the iAS556 model in comparison to the iAS142 model

An important aim of expanding the iAS142 to cover the genome-scale metabolism was to observe the relationship between the core energy metabolism and the peripheral metabolic

reactions. The iA556 model represents the whole genome-scale reconstruction of *L. infantum* metabolism that encompasses all the presently annotated metabolic pathways; while, the iAS142 model represents only the energy metabolism subset of the iAS556 model. In addition to the pathways covered by iAS142, the iAS556 model accounts for pathways related to all essential amino acids, fatty acid oxidation, fatty acid biosynthesis, lipid & phospholipid metabolism, sterol and isoprenoid biosynthesis, vitamin and cofactor biosynthesis, mannose metabolism (mannogen formation), purine, pyrimidine biosynthesis, their interconversions and secondary metabolite pathways. Exchange reactions that acted as proxies for fatty acid oxidation, removal of by-products of trypanothione metabolism, serine and threonine metabolism as considered within the iAS142 model were removed while building the iAS556 model, as their catabolism into other secondary pathways could be accounted.

With respect to these pathways, additional catabolic routes of intermediate metabolite conversions to form biomass metabolites were observed. Further, these reactions occur in distinct compartments thereby, changing the coupling between pathways for redox or mass balance, with respect to their requirement in biomass. Glucose is reported to distribute into catabolic routes which form mannogen, non-essential amino acids and precursors of phospholipids and sterols (Saunders et al. 2015). This observation was uniquely captured by the iAS556 model (see Section 4.2.2.2 for details). The coupling between glucose and the non-essential amino acid metabolism as observed in the iAS142 metabolism could also be captured from the iAS556 metabolism. But, this coupling was found to be essential in the iAS142 model for optimizing biomass, which was not the case in the iAS556 model. Glucose, tyrosine, valine, leucine and isoleucine are non-essentially linked to non-essential amino acid metabolism suggesting a supplementary role of these carbon sources in production of biomass (see Section 4.2.2.3 for details). Also, the amastigote-sepcific coupling of glycosomal fatty acid β-oxidation with lower glycolysis and succinate fermentation can be captured only from the iAS556 model (see Section 4.2.2.4 for details). This coupling provides the reason for utilization of fatty acids as an alternative for reduced glucose availability in the amastigote (Saunders et al. 2015).

4.2.2.1. Comparison of reaction knockout phenotypes predicted from the iAS556 model with experiments

Results from diverse experimental techniques that involved targeted gene deletion and inhibition of reactions by small molecule analogues were considered for validation purpose.

Experimentally determined growth phenotype information for only a few knockouts (43 knockout phenotypes) is explicitly available for any *Leishmania* species (Table B.3). In most cases (32 out of 43), knockout phenotypes are known and remain the same across both the stages (both-stage knockout phenotypes). Out of the 32, 20 phenotypes were lethal and 12 were non-lethal in both stages. For 11 reactions (one-stage knockout phenotypes), information is either available for promastigote and amastigote stages (10 phenotypes were lethal only in amastigote scenario and 1 phenotype was lethal only in promastigote). Considering the total number of known promastigote knockout phenotypes, 81% phenotypes were accurately predicted from the model. For the known knockout phenotypes of amastigotes, 84% phenotypes were accurately reproduced from the model. Percentage predicted wild type growth rates (standardized to maximum growth in a particular stage) remain the same between the promastigote and the amastigote situations; except for mitochondrial aconitase, which is non-lethal to the parasite but reduces parasite growth significantly (growth rate reduced by 77%) consistent with experiments (Saunders et al. 2014).

The model predictions can be classified as:

a) *Experimentally lethal/non-lethal in both stages and in-silico lethal/non-lethal (Perfect predictions)*: Our model was able to predict the experimentally known phenotypes for the 30 of the 32 both-stage knockout phenotypes perfectly. These knockouts demonstrate either lethal or non-lethal phenotypes from both experiments and predictions.

b) *Experimentally non-lethal but lethal in silico*: only adenine phosphoribosyltransferase (ADPTr) was predicted to be lethal *in silico* as compared to the experimentally known non-lethal phenotype in both the stages (observed in *L. donovani*). ADPTr is experimentally known to provide PRPP (Boitz and Ullman 2006; Carter et al. 2008). The reason for ADPTr becoming lethal in the model is due to the provision of phosphoribosyl pyrophosphate (prpp) to cytosolic phosphoribosylpyrophosphate synthetase (PRPPSic), which utilizes ribose-5-phosphate produced from pentose-phosphate shunt to form AMP (a very essential route). This might be a unique *L. infantum* JPCM5 specific feature. Rest of the inconsistencies is for the stage-specific knockout phenotypes, where the model was able to predict lethality for only either of the stages appropriately.

c) *Experimentally lethal but non-lethal in silico*: Only methylenetetrahydrofolate dehydrogenase (DH1), was predicted to be non-lethal *in silico* as compared to the experimentally lethal phenotype known in both the stages. The main aim of

methylenetetrahydrofolate dehydrogenase (DH1) is to produce glycine from serine through dihydrofolate metabolism. According to model predictions, the network structure restricts the essential glycine requirement only through direct uptake. Hence, there is no effect of this pathway on metabolic demand flux. Hence, DH1 and the other reactions of the dihydrofolate pathway (DHFR, TS) remain non-essential. Experimentally, the enzymes of dihydrofolate pathway are essential in nature (Nare et al. 1997), thereby supporting our prediction. Rest of the inconsistencies is for the stage-specific knockout phenotypes, where the model was able to predict lethality for only either of the stages appropriately.

For the one-stage knockout phenotypes, no information is available for the other stage. Also, one of the other reasons for the relatively inconsistent predictions is that the same objective function was used for optimization in the promastigote and amastigote states, which is a limitation in our model knockout analysis strategy. As sufficient data for the amastigote biomass requirements is unavailable, separate objective functions for promastigote and amastigote could not be defined. Once these become available, model predictions can be better compared with real phenotypes.

4.2.2.2. Prediction of stage-specific metabolic routes for catabolism of major carbon sources

Previous studies characterize a low uptake of non-essential amino acids, fatty acid uptake to compensate for the scarcity of glucose, reduced secretion of overflow metabolites, hypoxic and acidic environment as factors governing the metabolic state within the infective amastigote stage of the parasite (Rosenzweig et al. 2008; Saunders et al. 2014). Also, the possible strategic routes to utilize a variety of carbon sources within the two stages were also demonstrated (Saunders et al. 2014). Given these boundary (exchange) constraints (see Chapter 2), we asked the question whether the present constraint-based model demonstrates similar metabolic routes in the *L. infantum* network or not. Re-creating the above-mentioned scenarios, the fate of different carbon resources between promastigote and amastigote metabolic states was predicted and compared (Fig. 4.8A, see Chapter 2).

The predicted reaction steady state fluxes represent the stoichiometrically adjusted percentage conversion of a given substrate (fate) into a subsequent product. For example, glucose is entirely converted to glucose-6-phosphate (same stoichiometry). Hence, flux of glucokinase is equal to flux through glucose uptake and represents a 100% conversion of total input glucose into glucose-6-phosphate.

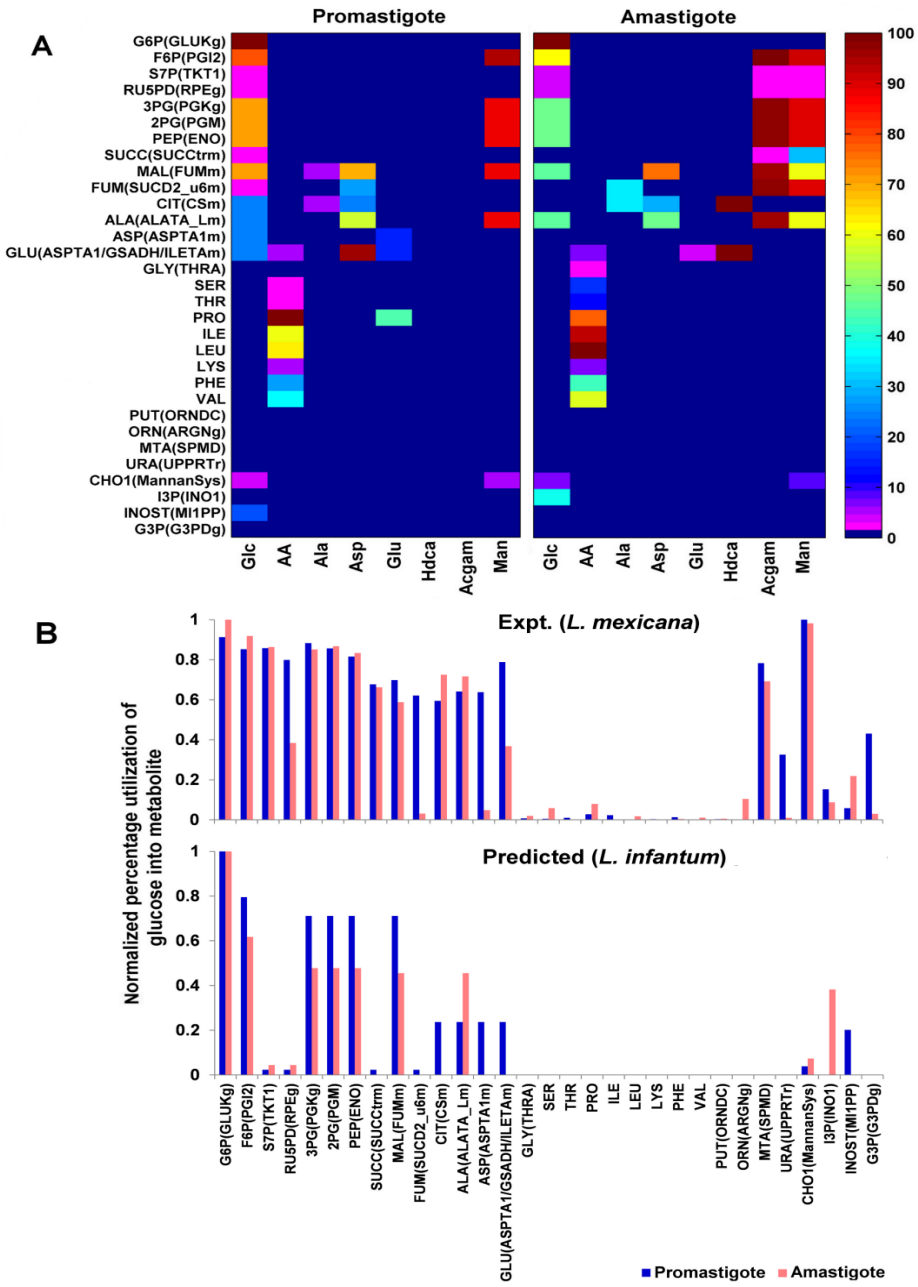

Figure 4.8. Fate of environmental metabolites within the *L. infantum* iAS556 metabolic network - (A) Metabolic utilization of carbon sources in developmental stages of *Leishmania infantum* – Each color in the heatmap indicates the percentage utilization of a given input metabolite (X-axis) for formation of a given internal metabolite when optimized for the metabolic demand reaction. In this case, the chosen metabolites represent the metabolites exclusively produced by the parasite to form biomass; (B) Comparison between promastigote and amastigote stages as given in the metabolomics experiment, where ^{13}C isotope enrichment of each metabolite was expressed in mole percentage & normalized to glucose-6-phosphate for parasites grown on ^{13}C labeled glucose and reaction fluxes (normalized to hexokinase/glucose uptake) as predicted for the glucose-only situation from the model. Blue colored bars indicate the promastigote metabolic state and red colored indicate the amastigote state. The metabolites in the X-axis and the order, in which they are arranged, were taken from the aforementioned ^{13}C isotope enrichment experiment performed in *L. mexicana* (Saunders et al. 2014).

With sufficient availability (promastigote), glucose was largely driven towards optimal production of overflow metabolites (2.34%), adenosine monophosphate (AMP) formation (2.34%), phospholipids (20.2%), glutamate (23.7%), aspartate (23.7%), alanine (23.7%) and mannan (3.9%); further, characterized by NAD redox coupling occurring between glycolysis and succinate fermentation, entry of pyruvate, succinate and glutamate into the tricarboxylic acid cycle (TCA), and regeneration of NAD reducing equivalents from C4-dicarboxylic acids by oxidative phosphorylation. In the amastigote, which experiences a decrease in glucose uptake, the quantity of glycolytic flux is largely compromised. The preference of glucose utilization is relatively more towards AMP formation (4.41%) and mannan synthesis (7.36%) when compared with the sufficient glucose situation. The remaining glucose is utilized towards alanine formation (45%) and phospholipid formation (38.27%), whereas asparagine, tyrosine and isoleucine are utilized for formation of glutamate, glutamine and proline.

As opposed to glucose, majority of alanine was utilized for satisfying the cellular demand; some of it (6.05%) was converted to pyruvate, which was either utilized for CO_2 fixation within glycosome or secreted as overflow in both abundant and constrained scenarios. Similarly, environmental aspartate in the promastigote is consumed and transaminated into glutamate (96.83%) and oxaloacetate in cytoplasm. Oxaloacetate forms malate, which enters mitochondria as fumarate, ultimately forming alanine (56.69%). In the amastigote, aspartate uptake reduces despite being catabolized via a similar route, the only difference being the sole formation of alanine (47.83%). Sufficient environmental glutamate is used for synthesis of glutamine (14.72%) via glycosomal GMP synthase, synthesis of proline (44.74%), aspartate by reverse de-amination of glutamate (14.72%), and glutamylcysteine to form trypanothione. In the amastigote, reduced glutamate uptake coerces its utilization into cellular metabolic demand and not as a precursor. Proline uptake (AA in Fig. 4.8) is preferred as the sole carbon source to produce glutamate, in order to satisfy the metabolic demand in absence of environmental glutamate (5.14%).

Serine, lysine, phenylalanine and valine (AA in Fig. 4.8) neither catabolize to give any intermediates nor are used as precursors and are directly utilized for biomass (Saunders et al. 2014). Under hypoxic conditions (amastigote condition), as biomass formation is sub-optimal, their utilization into biomass decreases. Glycine was produced solely from threonine. Leucine and isoleucine, both are utilized for provision of reduced FAD within mitochondrion. In the process, isoleucine converts into glutamate. As glutamate uptake largely reduces in amastigote, formation of glutamate and proline is largely compensated by

isoleucine and leucine. Fatty acids are preferred only under glucose deficient conditions (amastigote metabolic state) and primarily provide acetyl-coA to pyruvate dehydrogenase complex (McConville et al. 2015), thereby eventually forming glutamate. All the aforementioned predictions have also been experimentally observed from metabolomics studies (Saunders et al. 2014). In addition to the above carbon sources, the utilization of amino sugars (N-acetyl glucosamine) and mannose was also investigated. N-acetyl glucosamine is exclusively utilized under intracellular glucose-deficient, hypoxic conditions. N-acetyl glucosamine and mannose, both are utilized for provision of fructose-6-phosphate, which enters into mitochondria via succinate formation to produce alanine (98.3% and 87.76%, respectively). The uptake rates of both these sugars were predicted to decrease under hypoxic conditions of the amastigote; although, they do compensate the scarcity of glucose required for production of non-essential amino acids and mannan (Naderer et al. 2010; McConville and Naderer 2011). In addition, mannose also forms mannan (4.89%) and compensates the reduction in glucose uptake within the amastigote.

The stage-specific comparison between promastigote and amastigote flux-profiles as predicted by the model and ^{13}C isotope enrichment of metabolites, for the glucose-only situation is given in Fig. 4.8B. It is important to note that, as there was no such targeted ^{13}C metabolomics study available for *L. infantum* JPCM5 and because the metabolic microenvironment experienced by different *Leishmania* species is rather similar (McConville and Naderer 2011), the predicted reaction fluxes were compared with the ^{13}C isotope enrichment of metabolites reported in the metabolomics study available for *L. mexicana*. Normalization was performed for percentage values of fluxes and experimentally available ^{13}C isotope enrichment in a 0 to 1 scale, where 1 represents the highest normalized percentage flux value/isotope enrichment in mole percent and 0 the lowest. From Fig. 4.8B, it is clear that amastigotes demonstrate a lower synthesis of majority metabolites from glucose as compared to promastigotes owing to the reduced glucose uptake, in both experiments and model. Mannan production during the amastigote stage is larger than the promastigote in our *L. infantum* model, as compared to *L. mexicana*. The upregulation of mannan synthesis in amastigote was also observed in previous studies (McConville and Naderer 2011). A slight upregulation of the pentose-phosphate shunt was observed in the *L. infantum* amastigote indicating its role in satisfying the requirement of AMP under a reduced glucose uptake. It can also be observed that similar metabolites are produced from glucose in both *L. infantum* and *L. mexicana* but with significant quantitative differences in both stages.

4.2.2.3. Dynamic role of the non-essential amino acid motif in metabolic flux re-organizations

The observation of glucose being primarily diverted towards alanine under glucose-deficient conditions specifies that glucose can be substituted by a variety of other carbon sources to fulfill metabolite requirements other than alanine. The foremost question here is how this substitution can be achieved? To answer this, we introduce the notion of a "non-essential amino acid motif", which re-routes different metabolic inputs towards specific outputs under absence of preferred carbon sources, via implicit non-essential amino acid inter-conversions. The motif comprises of glutamate dehydrogenase, aspartate aminotransferase, tyrosine catabolic pathways, proline biosynthesis, glutamine biosynthesis, incomplete urea cycle, glycolytic and succinate fermentation pathways. To investigate their functionality, a sensitivity analysis was performed to understand the effect of glucose and uptakes of all amino acids, while optimizing for the formation of glutamate, alanine, aspartate, glutamine, proline, glycine, myoinositol and mannan, the formation of each considered as a separate objective function (see Chapter 2). Apparent optimal production of each amino acid due to self-uptake was not included in analysis. Further, the metabolic demand reaction was constrained to zero flux in each case.

Alanine is formed from a variety of resources and hence, tends to be an important overflow metabolite (Keegan and Blum 1990; ter Kuile 1999). Asparagine and aspartate, each of them when singly present in the environment can sub-optimally produce alanine (Fig. 4.9A). Presence of environmental glucose, isoleucine and tyrosine along with asparagine or aspartate amplifies production of alanine. Alanine is also produced when both proline and glucose are present in the environment, or when glutamate, glucose and tyrosine are made available. Glucose and tyrosine provide excess glutamate (Fig. 4.9B). Aspartate is optimally formed from asparagine, but sub-optimally formed from environmental proline and glucose or glutamate, glucose and tyrosine combinations. Similarly, glutamine is strictly formed when asparagine, aspartate and proline are part of the environment. Aspartate and isoleucine in the environment amplify flux towards production of glutamate, glutamine and proline (Fig. 4.9C). Glucose and tyrosine increase flux through proline degradation and glutamate formation, thereby producing excess glutamine (Fig. 4.9D). Glutamate biosynthesis follows the same route as glutamine. Proline, like any other non-essential amino acid is formed primarily from environmental asparagine and aspartate. Glutamate uptake from environment can also supplement glucose and tyrosine to form proline. The choice of various input

metabolite combinations to produce a precise repertoire of outputs is largely governed by their ability to generate internal glutamate as a precursor. The importance of glutamate has also been previously discussed in the context of energy metabolism from the iAS142 model. Also, these metabolite utilizations are typically due to presence of a redox and energy balance between the reactions utilizing these metabolic precursors.

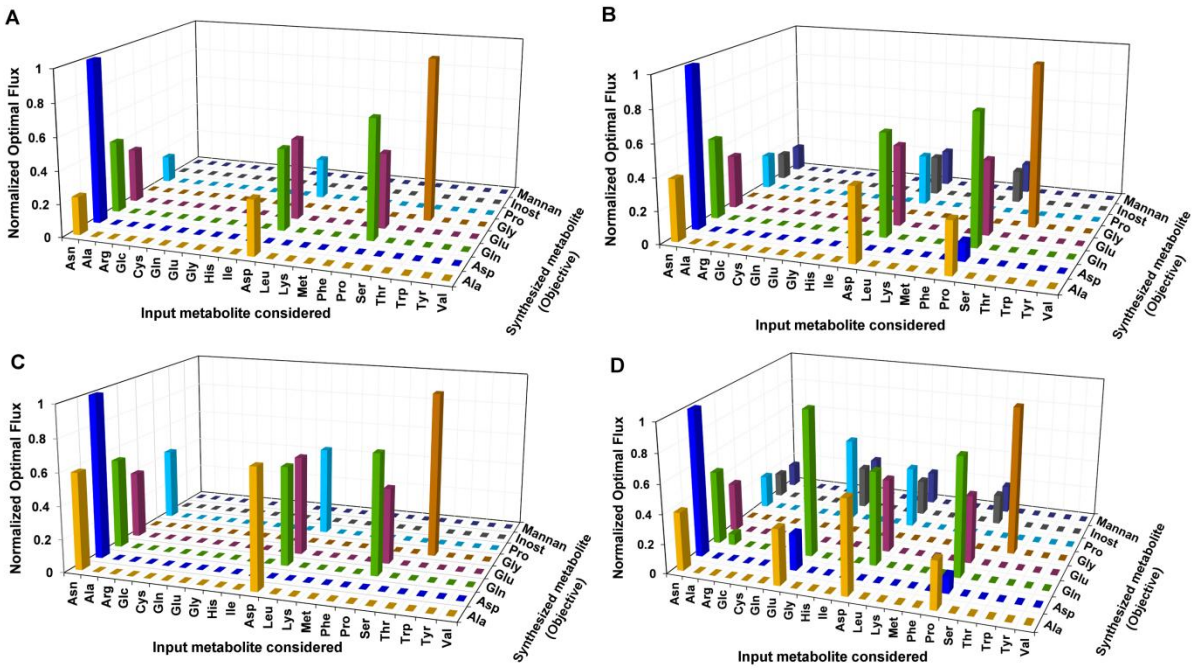

Figure 4.9. Sensitivity of flux towards synthesis of internal metabolites to specific input metabolite uptakes – Each bar in the plot represents the maximum flux through the synthesis of specific internal metabolites due to specific input metabolites, each synthesis considered as a separate metabolite drain (objective function) in the model. Four different scenarios were considered for optimizing the considered objective functions (synthesis of each separate output metabolite) namely, (A) only one given input at a time; (B) two inputs given at a time, glucose being fixed and a variable second input; (C) two inputs given at a time, one being isoleucine, which is kept fixed and a variable second input; and (D) Three inputs given at a time, glucose, tyrosine are fixed and a variable third input.

Environmental glucose and phosphatidic acid are catabolized through linear pathways for synthesis of mannan and phospholipids, respectively; although it is dependent on the non-essential amino acid motif via argininosuccinate synthetase for provision of ATP. Model-based knockout of argininosuccinate synthetase also predicts a lethal phenotype suggestive of its essential requirement (section 2.2.2). This further provides a novel insight into the possible essential role of argininosuccinate synthetase within the glycosome (Sardar et al. 2016).

4.2.2.4. Subcellular compartmentalization induces metabolite flux dependencies between distinct reactions

The utility and catabolism of environmental metabolites through defined metabolic routes is largely influenced by coupling between pathways for regaining electrons through reducing equivalents, ATP, small molecules, ions and other cofactors lost by one pathway from another pathway (Burgard et al. 2004; Saunders et al. 2010). The underlying network structure and its flux adaptation to the external sources are largely dictated by these dependencies. Intracellular boundaries created by subcellular compartments further reinforce these flux couplings (Allen et al. 2007; Go and Jones 2008). Glycosome and mitochondrion together harbor around 20% of metabolic enzymes (section 4.2.2.1.) that produce biomass metabolites. Formations of only few metabolites, like sterol and membrane lipids (around 4%), are restricted within the endoplasmic reticulum. Comparing scenarios for the presence and absence of the glycosome and mitochondrion, effect on pairing of reactions based on mass balance was studied. Simulations were performed in both the model-presumed promastigote and amastigote conditions for both the normal (subcellular compartment present) and perturbed (subcellular compartment absent) scenarios.

Glucose is catabolized for AMP synthesis, diversion into TCA for non-essential amino acid synthesis via succinate fermentation, myo-inositol and mannan formation (Fig. 4.10A). Given the absence of glycolytic regulation machinery in Trypanosomatids (Michels et al. 2006), glycosomes supervise this utilization of glucose towards appropriate outputs by promoting redox coupling between upper and lower part of glycolysis via NAD, glycosomal trypanothione reductase and pentose phosphate shunt via NADP, and cytosolic glutamate dehydrogenase with proline biosynthesis again via NADP, in glucose sufficient conditions. This dictates the use of precise combinations of external metabolites and their directed catabolism towards various internal metabolites via the non-essential amino acid motif. The glycosome also restricts the exchange of ATP/ADP and NAD/NADH generated from glycosomal reactions into the cytoplasm, thus, creating a balance between them (Gualdron-López et al. 2012). This balance is maintained by upper and lower part of glycolysis, argininosuccinate synthetase, and nucleotide salvage pathways within the glycosome. The above-mentioned reactions are coupled even in deficient glucose conditions (amastigote), although the drain of glucose into succinate fermentation for formation of alanine is largely compromised. This is compensated by the NAD redox coupling between succinate fermentation and glycosomal fatty acid β-oxidation, governed by the presence of glycosome.

The specificity of fatty acids as preferred carbon sources only in deficient glucose conditions is thus regulated by the glycosome. Removal of glycosome (Fig. 4.10B) gives rise to a distinct flux profile under glucose-deficient conditions when compared with normal flux scenarios ($P < 0.05$, Wilcoxon rank-sum test); although under glucose-sufficient conditions, the generated profiles seem to be similar.

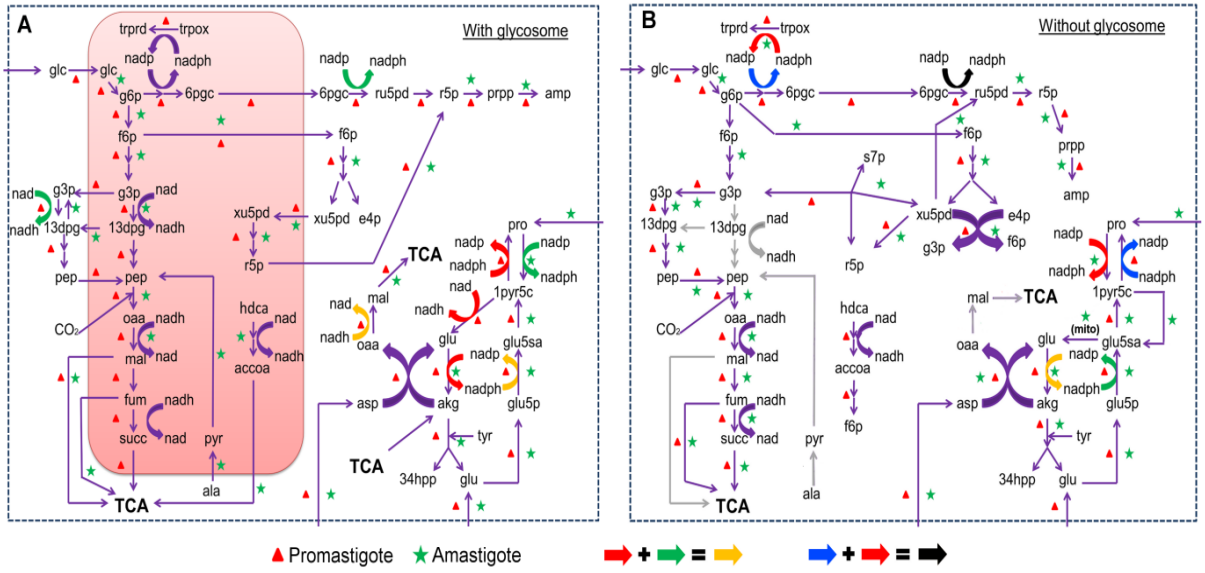

Figure 4.10. Comparison of flux profiles in glycosome-absent scenario - (A) Normal (with glycosome) scenario. Red colored box indicates the glycosome; (B) Perturbed (without glycosome) scenario – the reactions take place in cytoplasm (represented as a bounded white space). While arrows indicate conversions from one metabolite to another, the colors indicate the mass balance of redox equivalents between a set of reactions. Grey colored arrows indicate complete blockage of reactions that were present in the actual normal scenario. The cases where redox balance is maintained within more than one reaction are indicated by combination of colored arrows, as shown in the bottom of the figure.

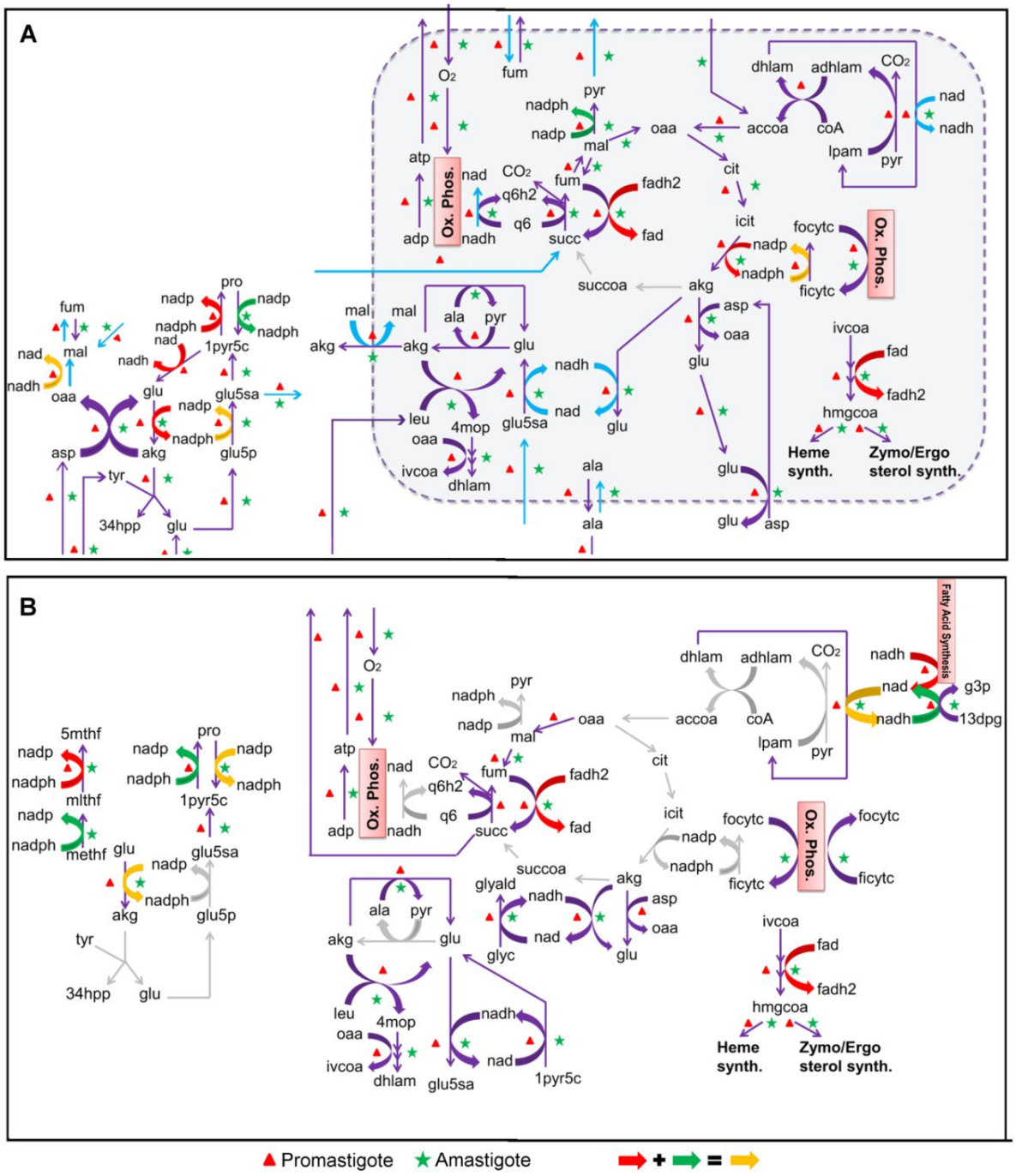

Figure 4.11. Comparison of flux profiles in mitochondrion-absent scenario - (A) Normal (with mitochondrion). Blue colored box indicates the mitochondrion; (B) Perturbed (without mitochondrion) – the reactions take place in cytoplasm (represented as a bounded white space). While arrows indicate conversions from one metabolite to another, the colors indicate the mass balance of redox equivalents between a set of reactions. Grey colored arrows indicate complete blockage of reactions that were present in the actual normal scenario. The cases where redox balance is maintained within more than one reaction are indicated by combination of colored arrows, as shown in the bottom of the figure.

A single mitochondrion exists in *Leishmania* that contains enzymes of TCA cycle, oxidative phosphorylation, amino acid catabolism, fatty acid biosynthesis and initial steps of sterol biosynthesis, each of which play an important role in the survival of the parasite (Fidalgo and Gille 2011). The presence of mitochondrion directs tyrosine and glucose into the non-essential amino acid motif to produce glutamate, glutamine, proline, aspartate and alanine. One of the most important functions of the mitochondrion is to redirect C4-dicarboxylic acids produced from glycolysis and non-essential amino acid motif to synthesize mitochondrial glutamate, thereby driving the TCA to produce reducing equivalents for oxidative phosphorylation (Fig. 4.11A). The mitochondrion also ensures utilization of tyrosine catabolism to provide excess glutamate for proline biosynthesis, protection of alanine for meeting cellular demand and restricts the excess drain of useful succinate into the environment to provide adequate FAD reducing equivalents for initial steps of sterol synthesis. Similar to the glycosome, removal of mitochondrion (Fig. 4.11B) gives rise to a distinct flux profile under glucose-deficient conditions when compared with normal flux scenarios ($P < 0.05$); although under glucose-sufficient conditions, the generated profiles seem to be similar.

4.2.2.5. Physiological flux coupling is robust against random reaction deletions

The above results demonstrate that structural constraints within the metabolic network confine the parasite to utilize specific environmental metabolites and direct it towards synthesis of various biomass metabolites. To answer the question, why the network might have been organized in a specific way, the flux-coupled subnetwork (Fig. 4.12) derived from the whole genome-scale network was subjected to random perturbations while monitoring the re-organization of physiological flux coupling relationships within the network (Chapter 2). These flux relationships represent a complete set of coupled reactions within the *L. infantum* metabolic network, with respect to all the exchanges considered in the network. The median number of flux-coupled pairs [both fully (FC) and directionally coupled (DC)] almost remains the same up to 5 simultaneous deletions (< 1% decrease in DC and < 0.2% decrease in FC pairs), after which the numbers gradually decrease with the largest change observed for 20 random deletions (25.6% decrease in DC and 10.7% decrease in FC pairs) (Figs. 4.13A, 4.13B). This suggests that reaction stoichiometry and reversibility impose constraints on the *L. infantum* metabolic network structure in such a way that only a few, specific reactions are tightly coupled to each other and rest of the reactions remain either isolated or coupled to fewer number of reactions. This lowers the chances of a coupled reaction set to be damaged

by random errors. This is also supported by the fact that the flux-coupled graph generated from the unperturbed bipartite *L. infantum* iAS556 metabolic network contains few numbers of reactions [256] that have greater than average connectivity and a large number of reactions [693] with less than average connectivity (Average connectivity = 14.67). Also, the median number of reactions represented within modules remains unchanged with the increase in number of simultaneous deletions suggesting that the network regains its modularity after every deletion and is relatively insensitive to random deletions (Fig. 4.13C).

The biological significance of these specific couplings towards parasite metabolism and its effect on the biomass metabolites was further investigated. For this, the fraction of biomass metabolites whose synthesis would be totally blocked after each sequential deletion was calculated. The synthesis of a biomass metabolite was considered to be completely obstructed, if all the reactions forming that metabolite were deemed to be blocked after a sequential deletion (see Chapter 2). The median percentage of metabolites expected to be blocked, also gradually increase with increasing number of reaction deletions; although only around 7% of biomass metabolites can be expected to be completely blocked after 20 sequential deletions (Fig. 4.14A). Also, the chance of obtaining more than 2% of metabolites to get completely blocked is higher (distribution skewed towards higher fractions) only if greater than 10 reactions are inhibited simultaneously (Fig. 4.14A). This suggests that a large number of less flux-coupled reactions observed within the network provide robustness to biomass-metabolite formation. Further, it seems that biomass-metabolite reactions are physiologically coupled to less number of reactions (median = 4) as compared to other non-terminal network reactions (median = 6), which on the other hand, tend to be coupled to a large number of other network reactions; although the fraction of variation explained by the difference between the biomass-forming vs. non-terminal reaction flux couplings is sufficiently small (Fig. 4.14B, $W = 54107$, $Z = -2.9627$, $r = -0.09$, $P = 0.003$, Wilcoxon rank-sum test). This probably suggests that biomass-forming reactions are not blocked any more than other non-terminal reactions and that any effect is due to a handful of biomass reactions which demonstrate a low flux-coupling, rather than being a general effect specific to all biomass reactions. Also, 37 of the 40 biomass metabolites are formed by more than one reaction groups (connected components) within the flux-coupled graph, probably for unhindered production of the important biomass metabolites.

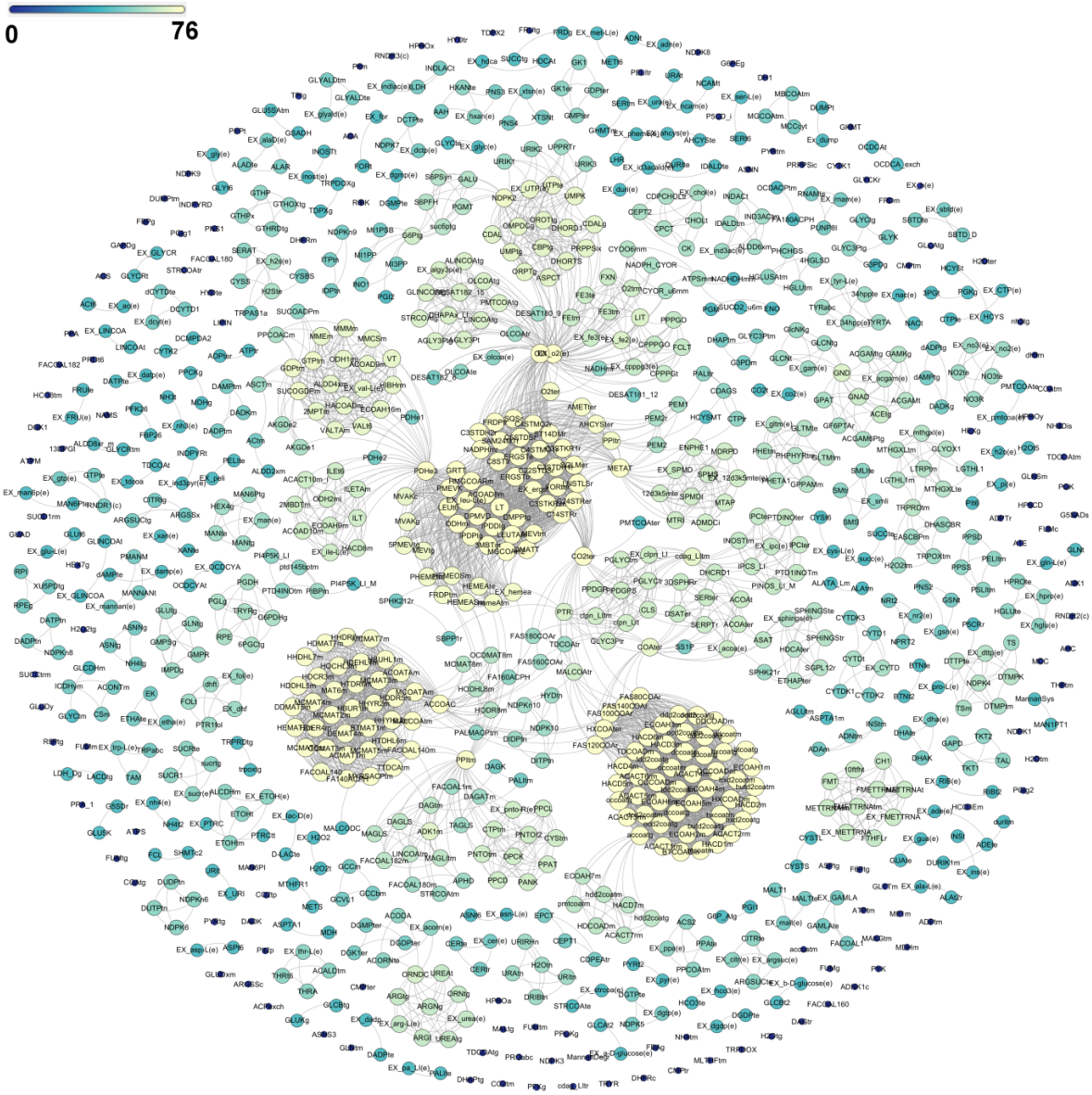

Figure 4.12. Network representation of the flux-coupled graph computed for the iAS556 metabolic network by flux coupling analysis - The network figure was created using the Fruchterman-Reingold algorithm implemented in Gephi version 0.8.2. Each node represents a reaction within the flux-coupled graph. Node sizes and colors were chosen according to the total number of reactions a particular reaction is coupled to, within the flux-coupled graph (degree of node within flux-coupled graph).

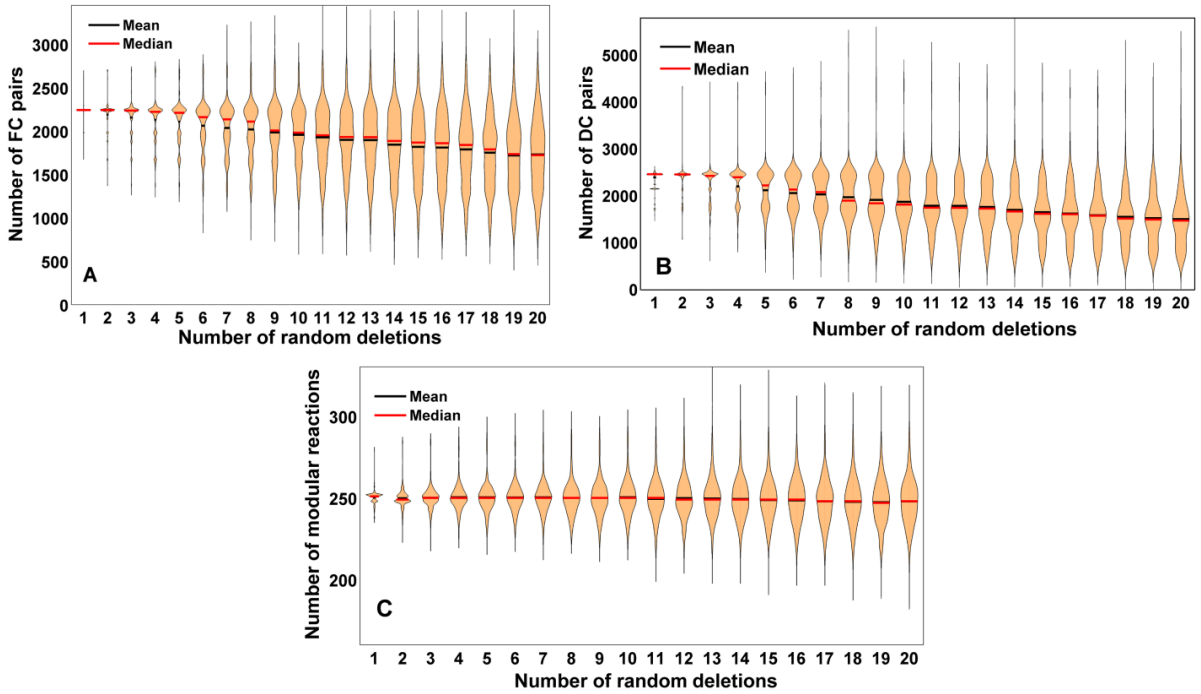

Figure 4.13. Effect of random deletions on physiological flux coupling relationships. Violin plots indicating - A) Number of FC pairs; B) Number of DC pairs; C) Number of modular reactions. A total of 1000 random sets of deletions were performed for each case.

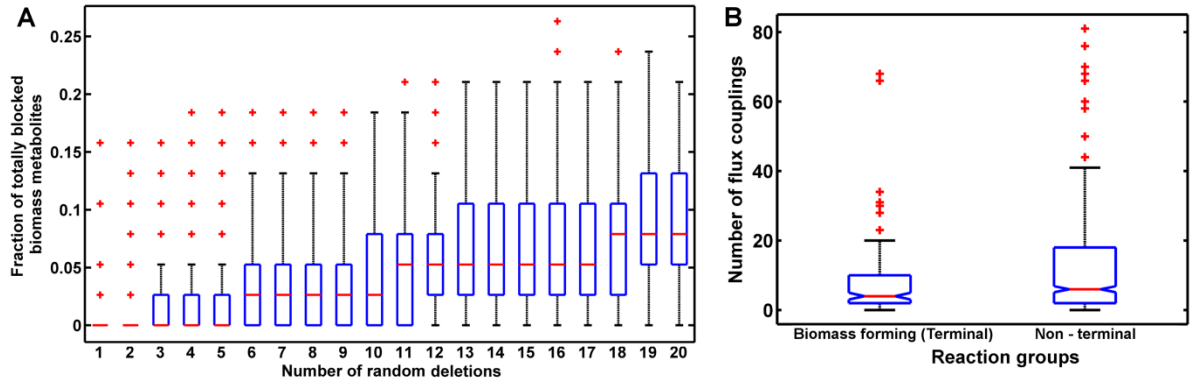

Figure 4.14. Robustness of physiologically coupled reactions within the *L. infantum* metabolic network to random deletions - (A) Boxplot indicating the distribution of the fraction of biomass metabolites that are totally blocked after each deletion or set of random deletions; (B) Notched boxplots for comparison of total number of flux couplings of biomass forming ultimate reactions (161 reactions) with all other non-terminal reactions (788 reactions) within the flux-coupled graph of the *L. infantum* JPCM5 metabolic network. The red colored points marked by ' + ', indicate the outliers in each distribution.

4.3. Discussion

In this chapter, two new reconstructions related to the *L. infantum* metabolism are introduced. Both the reconstructions are novel with respect to functional annotation & subcellular localizations of enzymes specific to *L. infantum* when compared to known reconstructions for other *Leishmania* species. The two reconstructions were validated with a plethora of observations reported for *Leishmania* metabolism. The validated reconstructions were used to predict and analyze the previously unknown aspects of *Leishmania* metabolism and its organization. The stand-out points related to the reconstruction strategy, multiple validations, metabolic routes essential for optimizing cellular demand in the two developmental stages and the organization of flux within *Leishmania* metabolism are discussed below in details.

4.3.1. Novelty in model development

A standardized strategy catered for model reconstruction in *Leishmania* metabolism was proposed (see Chapter 2). Both the iAS142 and the iAS556 models contain mitochondrial intermembrane space as a novel model compartment which none of the previous constraint-based models in Trypanosomatids have considered. The inclusion of this compartment was imperative, since the proton gradient established by oxidative phosphorylation is coupled to ATP synthesis between the mitochondrion and the intermembrane space, and not with the cytosol. Performing model iterative refinement, essential gaps in metabolism with respect to enzymes with novel function, known enzymes but in novel subcellular localization or missing transport reactions were identified. The iAS556 model accounts for 223 novel intracellular reactions (35 compartmental + 188 intracellular transports) in comparison to the previously known *L. major* genome-scale iAC560 model. Based on relevant proof from literature and *in silico* prediction of subcellular signal sequences, reactions have been assigned to specific subcellular locations in both the models, giving a high confidence to those reactions having strong literature proof. As a result, around 68 reactions within the iAS142 model and 250 reactions in the iAS556 model could be assigned to their appropriate locations with high confidence (confidence score > 3). Most importantly, *Leishmania*-specific novel metabolic demand reactions were proposed for both the iAS142 and iAS556 models (see Chapter 2).

4.3.2. Limitations to model validation by reaction knockout analysis and improvements

Constraint-based metabolic models are largely validated through comparison of model reaction knockouts with known gene knockout phenotypes. A total of 61 lethal reactions were

identified through single reaction deletions in the iAS142 model and 55 non trivial lethal reaction pairs were predicted to be essential through paired double reaction deletions. The 7 known knockout phenotypes were perfectly predicted through reaction knockout analysis in the iAS142 model. Performing a stage-specific knockout analysis in the iAS556 model, 81% of the total known promastigote phenotypes and 84% known phenotypes in amastigotes were accurately reproduced from the model.

Although, model validation based on knockouts is a popular method, there are three important shortcomings in validating a constraint-based model through reaction knockout studies (Roberts et al. 2009). First, due to a low availability of knockout phenotype information in *Leishmania* species, model accuracy with respect to comparison with less available information might be misleading. Second, many missing gaps are present in *Leishmania* metabolism, a feature that corresponds to a large number of non-annotated hypothetical proteins. Hence, it is possible that the model might predict a reaction to be non-lethal, even though in reality, it might be lethal or vice versa. This is a major difference between the reaction knockout phenotypes of the energy metabolic iAS142 model and the genome-scale metabolic model iAS556. Third, the experimental knockout data available for validation purposes are collected from heterogeneous studies that might have considered different experimental conditions of temperature, pH, or media to generate knockouts, which cannot be exactly implemented in our models. To overcome these issues, along with performing primary knockout analyses, validation with other types of biological data is required.

Leishmania is known to exhibit an overflow metabolism under varying glucose and oxygen uptake, during which it secretes substantial amount of succinate, acetate, pyruvate, CO_2 and small amounts of D-lactate (Keegan and Blum 1990; ter Kuile 1999). As the iAS142 model is an energy metabolic model, by performing a robustness analysis with respect to changes in glucose and oxygen uptake, model conditions required to discern the secretion of the mentioned overflow metabolites was identified. Results indicate pyruvate is secreted due to the increased flux through pyruvate kinase and succinate fermentation enzymes that helps maintaining the ATP/ADP ratio in glycosome and cytosol under highly anaerobic conditions. Succinate is a preferred overflow metabolite, generated through the fermentation of phosphoenolpyruvate within the glycosome and is secreted under excess glucose conditions (Saunders et al. 2011). Model robustness analysis for specific ranges of fixed oxygen uptake was able to predict lactate production via the D-lactate dehydrogenase reaction. This reaction is known to occur specifically for *Leishmania* (Darling et al. 1987). As observed from

experiments in *L. major* (Darling et al. 1989), CO_2 uptake is possible under anaerobic conditions and is fixed within the glycosome creating a reverse Pasteur effect and leading to increased succinate fermentation via a glycosomal phosphoenolpyruvate carboxykinase and other enzymes involved in glycosomal succinate fermentation.

The iAS556 genome-scale model was validated and compared with targeted ^{13}C isotope resolved metabolomics available for *L. mexicana*, an evolutionary relative of *L. infantum* (Saunders et al. 2014). Comparison of predicted steady state reaction fluxes in *L. infantum* JPCM5 with ^{13}C isotope enrichment of metabolites in *L. mexicana* indicated that similar metabolic routes of carbon source utilization are chosen across stages within *L. mexicana* and *L. infantum* metabolism, with only striking quantitative differences. The quantitative differences observed between the predicted reaction steady state fluxes in *L. infantum* JPCM5 and the ^{13}C isotope enrichment of metabolites available for *L. mexicana* might be due to –

a) species-specific genes unique to *L. mexicana* or *L. infantum*. For example, metabolites of the incomplete urea cycle, are largely independent of glucose utilization and are optimally produced from different carbon sources like arginine for utilization into the metabolic demand within the *L. infantum* network which might not be the case in *L. mexicana*, due to differences in underlying network structure,

b) multiple subcellular locations of different enzymes that might lead to alternative paths for distribution of intermediate metabolites produced from glucose within the pathways of glycolysis, pentose-phosphate-pathway and non-essential amino acid synthesis unique to the *L. infantum* JPCM5 metabolic network, and/or,

c) metabolic demand being different between *L. infantum* and *L. mexicana*, leads to an increased drain of intermediate metabolites like non-essential amino acids into the *L. infantum* metabolic demand as compared to *L. mexicana*. The aforementioned reasons are only suggestive of the possibilities underlying the observed differences.

Further inferences based on comparison of flux profiles in the two species can only be performed from an evolutionary perspective, given the degree of evolutionary divergence between *L. infantum* and *L. mexicana*, and the relatedness of *L. mexicana* with the Sauroleishmania (Harkins et al. 2016).

4.3.3. Reactions essential to the core energy metabolism across stages

Analysis of the iAS142 metabolic network indicated few pathways to be essential for core energy metabolism:

1) *Succinate fermentation*: Succinate fermentation takes place in the glycosomes and involves formation of succinate from phosphoenolpyruvate by virtue of four sequential reactions - PPCKg catalyzed by phosphoenol pyruvate carboxykinase, MDHg catalyzed by malate dehydrogenase, FUMg catalyzed by fumarate hydratase, and FRDg catalyzed by fumarate reductase. Through these reactions, NAD molecules consumed by the enzymes of upper glycolytic pathway are regenerated (Bringaud et al. 2006; Saunders et al. 2011). Hence, these reactions occurring in the glycosome govern a lethal phenotype. Interestingly, NAD could also be regenerated through lactate dehydrogenase reaction which also takes place in the glycosome. But the inability of the iAS142 model to produce non-essential amino acids like alanine through TCA cycle in the absence of succinate production, immediately points towards the importance of succinate.

2) *NADPH cytochrome P450 oxidoreductase* (NADPH_CYOR): NADPH_CYOR, an important reaction that was not considered in previous Trypanosomatid core energy metabolism reconstructions (Roberts et al. 2009) is required to maintain the NADPH redox balance in the mitochondrion by regeneration of NADP from NADPH, thus, rendering it to be essential. This role makes this enzyme an important therapeutic target as reported in *Trypanosoma cruzi*, a Trypanosomatid relative of *Leishmania* (Portal et al. 2008).

3) *Glutamate dehydrogenase* (GLUDy): Similarly, NADPH-dependent glutamate dehydrogenase (GLUDy) reaction was identified to be an important oxidoreductase which re-routes glutamate from cytoplasm to mitochondria via alpha-ketoglutarate and is also a major provider of non-reduced NADP within the cytosol. This re-routing along with production and conversion of other amino acid intermediates into C4-dicarboxylic acids within mitochondria is part of TCA anaplerosis in *Leishmania* (Saunders et al. 2011).

4) *ATP synthase*: Mitochondrial membrane ATP synthase (ATPSmm) that is majorly responsible for ATP production in the mitochondrion and for the entire cell also governs a lethal phenotype in the iAS142 model. Mitochondrial ATP synthase hence, is an important target in *Leishmania* metabolism (Luque-Ortega et al. 2008). The cytosolic ATPase reaction (ATPS) on the other hand, represents a continuous drain of ATP is not predicted to be lethal.

4.3.4. Differences between promastigote and amastigote metabolic states

As mentioned previously, the differentiation of promastigote to amastigote brings about a reduced glucose and non-essential amino acid uptake. Furthermore, the amastigotes reside in a relative hypoxic environment as compared to the promastigotes. With respect to the choice of metabolites in both these environmental conditions, glucose is the preferred substrate in promastigotes which is utilized along with coupled amino acid fluxes, as indicated in both the iAS142 and iAS556 models. In the amastigotes, a reduced glucose uptake leads to an increased uptake of fatty acids, amino sugars and mannose. Fatty acids and amino sugars are known alternatives for glucose within *Leishmania* amastigotes (Naderer et al. 2010; Saunders et al. 2014). Even though these conditions significantly constrain the choice of metabolites across stages, comparison of the promastigote and amastigote metabolic states, as performed for the iAS556 model, indicates a striking conservation of metabolite utilization routes with very few differences.

Aspartate aminotransferase and glutamate dehydrogenase occur along with enzymes of malate dehydrogenase, fumarate reductase and fumarate hydratase in the cytoplasm as well as mitochondrion. The malate-α-ketoglutarate shuttle and malate-aspartate shuttles present in the mitochondrial intermembrane space regulate the re-routing of aspartate and glutamate between the cytoplasm and mitochondrion, multiple localizations of aforementioned enzymes. This set of enzymes is further coupled with tyrosine catabolism, proline metabolism, glutamine biosynthesis and an incomplete urea cycle, thereby, creating a dynamic non-essential amino acid motif. This motif distributes the C4-dicarboxylic acids and non-essential amino acids produced from the glycosomal and cytoplasmic reactions towards the mitochondrion. Hence, in both the iAS142 and iAS556 models, energy metabolism concerned with the catabolism of carbohydrates and amino acids, is largely driven towards the synthesis of glutamate and alanine via TCA cycle, while also optimizing for ATP synthesis.

In the promastigotes, glucose is mainly used to produce overflow metabolites like pyruvate, succinate, fumarate, malate and mannan (a *Leishmania*-specific reserve storage material) within the glycosome, nucleotides like AMP via pentose-phosphate pathway within the mitochondria and cytosol which are important precursor intermediates produced via dynamic non-essential amino acid motif and myoinositol-1-phosphate that eventually synthesizes phospholipids in the endoplasmic reticulum. Pyruvate formed in the glycosome via the phosphoenolpyruvate carboxykinase is routed to the mitochondria for alanine

synthesis. Succinate, fumarate and malate are re-routed to form glutamate & aspartate. In the amastigotes, glycolytic flux reduces and is largely diverted towards mannan production with an accompanied increase in pentose-phosphate shunt to produce AMP. These phenomena were also reproduced by the energy metabolic iAS142 model suggesting differential requirements of energy metabolism between the two stages. Mannose is utilized through similar routes as glucose for the synthesis of mannan and alanine in both the stages. Amino sugars are specifically used in amastigotes for synthesis of alanine. The utilization routes of mannose and amino sugars are similar to glucose as they are catabolized within the glycosome and commonly enter glycolysis as fructose-6-phosphate. On the contrary, fatty acids typically enter into the TCA as acetyl-coA, which produces glutamate via citrate synthase within the amastigotes. Further, glutamate is used as an important precursor for synthesis of glutamine via glycosomal GMP synthase, synthesis of proline, aspartate by reverse de-amination of glutamate and glutamylcysteine to eventually form trypanothione and hence, is an important metabolite. Due to these constraints, glutamate can be formed by a number of metabolites like glucose, fatty acids, aspartate, proline, tyrosine, isoleucine and leucine within the promastigote. All other essential amino acids are utilized only for the synthesis of proteins across the two developmental stages.

4.3.5. Metabolic network organization in *Leishmania*

The organization of *L. infantum* metabolism, though simple, can be dynamic enough to re-route external metabolites towards synthesis of specific internal biomass metabolites. This dynamic nature is due to the non-essential amino acid motif-based reaction sets within the metabolic network that get coupled with respect to various combinations of environmental metabolites. The dynamic non-essential amino acid motif diverts specific input metabolites towards specific outputs and constrains the utilization of diverse resources simultaneously. This analysis also suggests that glucose and tyrosine alone cannot produce other biomass metabolites, and their role is to only supplement the quantities of output metabolites formed through non-essential amino acids, indicating a tight coupling between glucose, tyrosine and non-essential amino acids. This observation is also supported by the iAS142 energy metabolic model where glucose and non-essential amino acids were found to be catabolized together. This constrained distribution is largely influenced by the presence of subcellular compartments that restrict coupling of specific reactions to achieve redox or ATP balance. Glycosome restricts the transfer of ADP between the glycosomal reactions and the cytoplasmic reactions, thereby inducing a redox coupling between trypanothione reductase

and pentose phosphate shunt, and upper and lower part of glycolysis in the promastigote. Similarly, due to reduced glucose uptake observed in the amastigote, NAD redox coupling between fatty acid β-oxidation and lower part of glycolysis is established and hence, its role as an alternative of glucose. The role of mitochondrion on the other hand, is to redirect C4 dicarboxylic acids produced from glycolysis and non-essential amino acid motif towards TCA cycle for formation of mitochondrial glutamate, which regenerates reducing equivalents for oxidative phosphorylation.

The *L. infantum* flux-coupled network structure is highly modular and assortative, thereby remaining robust to a random node failure (Hase et al. 2010; Subramanian and Sarkar 2016). This also suggests that the modular nature of the network and presence of enzymes in dual/multiple subcellular locations are important characteristics of *L. infantum* network that provide adaptation to the changing environment. The network robustness ensures an unobstructed production of biomass metabolites. Furthermore, the benefit of the network structure being robust to random mutations is not only gained by biomass-forming reactions but also equally by all the other network reactions. Few biomass metabolites like ergosterol, heme A, isoleucine, leucine, valine, pantothenate, and zymosterol, are produced only from a set of specific enzymes, which form a highly connected, single connected component (central within the flux coupled subnetwork), which might be useful as drug targets. Enzymes of heme and sterol metabolism in *Leishmania* have already been indicated as important targets (Roberts et al. 2003; Koreny et al. 2013).

Chapter 5 - Identification of confounding genotype-phenotype features that constrain metabolic enzyme evolution

5.1. Introduction

Metabolism is one of the primary biological processes that underlie the survival of an organism within a given environment, due of its fundamental role in synthesis of biomass and energy generation. Even though the individual metabolic enzymes *per se* are highly conserved across species, adaptation to diverse environments brings about novel innovations in metabolic pathway function (Szappanos et al. 2016). Numerous features like horizontal gene transfer, gene expression, gene dispensability, gene duplications and metabolic network structure are responsible for changes in metabolic function (Yamada and Bork 2009; Papp et al. 2011). The dominance of one feature over another largely depends on the nature, variations in the environment and the effective contribution of a factor towards successful adaptation to that particular environment. In general, the change in metabolic function due to changes in a feature can either be selected in a population for its usefulness in adaptation or else it can be purged, if deleterious. This change in function is reflected within the coding sequence of a gene and is conventionally measured by assessing the number of non-synonymous substitutions per non-synonymous site relative to the number of synonymous substitutions per synonymous site, commonly referred as the evolutionary rates (Yang 1998). However, the knowledge of potential determinants of evolutionary rates of a metabolic enzyme within an organism still remains to be an open, unsolved problem.

Members of the *Leishmania* genus cause the widespread neglected tropical disease leishmaniasis in humans. Biologically, the *Leishmania* parasite exhibits a digenetic lifecycle, where the promastigote stages thrive within the midgut of the sandfly vector, and the amastigotes persists in the macrophage phagolysozome of the human host; the environments being largely antagonistic with respect to pH, temperature and availability of carbon sources (Zilberstein and Shapira 1994; McConville and Naderer 2011). To ensure maximal survival, the parasites need to selectively adapt to these dual environmental constraints. This controlled biological setup provides us with a unique platform for investigating the contributory role of different genotypic and phenotypic factors in metabolic enzyme evolution. Numerous genotype and phenotype factors are known to contribute to evolutionary rate variation in eukaryotes. The factors that are known to have a probable effect on protein evolution largely falls into two categories, namely, translation selection and functional constraint.

*The bulk of this chapter is communicated and some part has also appeared in the *Proceedings of BIOMAT-2015 Symposium*

Translation selection refers to the evolutionary selection of features that can increase efficiency of translation, whereas, functional constraint of an enzyme refers to the degree at which random mutations are removed from the population by natural selection so as to avoid their deleterious effect on protein function (Zhang and Yang 2015). With respect to features explaining translation selection, gene expression, mRNA transcript length (or length of a coding sequence) and codon usage were demonstrated as important factors that explain the evolution of protein-coding genes in yeast and *Arabidopsis* (Kawaguchi and Bailey-Serres 2005; Drummond et al. 2006; Zhang and Yang 2015). With respect to features explaining functional constraint, pleiotropy of a gene due to multiple functional domains, involvement of enzymes in multiple biological processes, and multiple gene duplications can contribute to enzyme evolution thereby providing a dynamism to the metabolic network structure (Salathé et al. 2005; Warringer and Blomberg 2006; Chu et al. 2014; Chesmore et al. 2016). Another less studied functional constraint that affects the evolution of a metabolic enzyme is the role of an enzyme in the context of other enzymes within a metabolic network. As metabolic function is a result of stepwise transformation and utilization of different environmental metabolites through multiple pathways, it is not the effect of a single enzyme. Hence, more central proteins within a metabolic network are also resistant to functional change (Vitkup et al. 2006). Previous studies in yeast and human erythrocytes have also demonstrated that enzymes bearing higher metabolic flux tend to evolve slowly (Vitkup et al. 2006; Colombo et al. 2014). It was also demonstrated that co-regulation in metabolic genes is largely explained by flux-coupling within a metabolic network (Notebaart et al. 2008) suggesting it to be an important factor constraining metabolic function and hence, enzyme evolution.

Similar to other organisms, a few studies in *Leishmania* species also provide indirect hints towards the roles of translation selection and functional constraints on metabolic enzyme evolution. Stage-specific transcriptomics and proteomics studies identify variations in transcriptome and proteome abundances of metabolic genes across stages and species in *Leishmania* (Lahav et al. 2011; Nirujogi et al. 2014). Mutation pressure and translation selection is shown to preserve GC usage at the synonymous codon position and frequent codons within a gene as probable modes of translation regulation within genes (Subramanian and Sarkar 2015). Chromosomal aneuploidy is another well-known mechanism that causes variations in gene copy numbers across *Leishmania* species (Mannaert et al. 2012). Recent computational predictions of metabolic flux for different input metabolites and targeted ^{13}C-based metabolomics studies have identified that the *Leishmania* metabolome adapts to changing host environments through common metabolic routes, which are largely constrained

by the metabolic network structure (Saunders et al. 2014; Subramanian and Sarkar 2017). The network structure also constrains enzyme evolution in *L. major* metabolism (Subramanian and Sarkar 2016).

The aforementioned studies in *Leishmania* have largely explored the genotype and phenotype complements of metabolism independently. The combined effects of these features on the disparate forces of conservation and divergence in enzyme evolution are yet to be tested. To establish their effects on evolutionary rates among metabolic enzymes, a comprehensive comparative strategy that can examine the relative effects of the different genotype and phenotype features simultaneously is required. In this study, we estimate the rate of non-synonymous substitutions per non-synonymous site (d_N), rate of synonymous substitutions per synonymous site (d_S), and their ratio ($\omega = d_N / d_S$) and for the first time, identify the potential determinants of d_N, d_S and ω among orthologous singleton metabolic genes in three *Leishmania* species using a principal component regression (PCR) based analysis (Drummond et al. 2006; Jovelin and Phillips 2009; Alvarez-Ponce and Fares 2012; Alvarez-Ponce et al. 2017). We introduce the flux-coupling potential of an enzyme within a metabolic network (Subramanian and Sarkar 2016), as a potential feature for regression along with other available features for *Leishmania* metabolism. Additionally, a group of genes with evolutionary rates explained by a diverse set of independent features was also identified for the three species. Comparisons of gene clusters across the three species identified the differential constraints experienced by a given gene across species and also important species-specific genes constrained by multiple features. The results provided in this article thus, highlight the importance of comprehensive multivariate studies in understanding the various causes and consequences of evolutionary divergence and conservation in *Leishmania* metabolism; thereby, identifying targetable mechanisms and genes, which can be further perused for designing novel strategies against parasite persistence.

5.2. Results

5.2.1. Frequency distribution of evolutionary rates in singleton metabolic genes

Comparing the ranges of evolutionary accounting for 90% of the genes and the coefficient of variation with respect to evolutionary rates of genes, it can be observed that the d_N and ω rates, when compared to the d_S rates demonstrate a larger variance (Fig. 5.1, Table 5.1.). The average and median values of d_S rates are very close to each other. A similar distribution of

d_N, d_S and ω values is also observed in complete proteome of *Arabidopsis* (Yang and Gaut 2011).

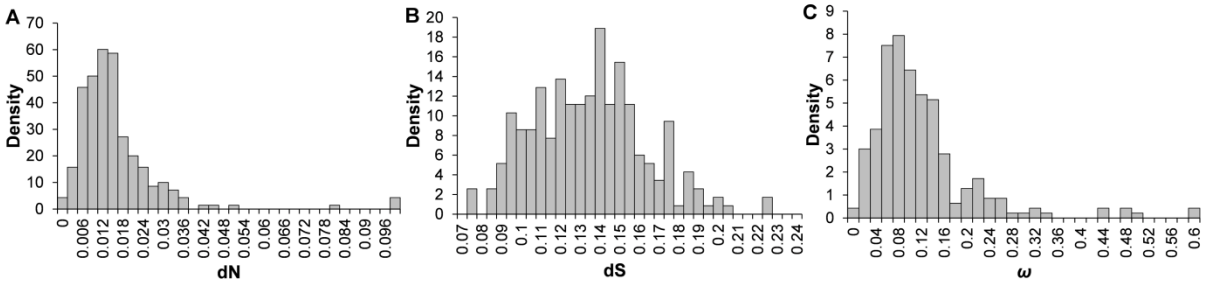

Figure 5.1. Frequency distributions of evolutionary rates across *Leishmania* species - A) d_N; B) d_S; C) d_N/d_S (ω).

As indicated by average ω estimates (Table 5.1), a 4-fold difference is observed between d_N and d_S suggesting a high evolutionary conservation (a relatively large number of synonymous substitutions compared to non-synonymous substitutions) in singleton orthologous genes of *Leishmania* metabolism.

Table 5.1. Statistics of evolutionary rates for the considered singleton orthologues.

	d_N	d_S	ω
Mean	0.0161	0.137	0.125
SD	0.0093	0.0286	0.082
CV (%)	57.52	20.85	65.62
Range (90%)	0.0057 - 0.0322	0.095 - 0.188	0.037 - 0.268
Median	0.0145	0.138	0.104

5.2.2. Features associated with evolutionary rates are also inter-correlated in *Leishmania* species

Performing a pairwise correlation analysis for the orthologous metabolic genes in *Leishmania major* Friedlin, *Leishmania donovani* BPK282A1 and *Leishmania infantum* JPCM5 (Fig. 5.2), it was identified that there is no significant correlation obtained between d_N and d_S, whereas ω is significantly correlated with both d_N (in all three species, r = 0.9053; P = 7.3 x 10^{-88}) and d_S (in all species, r = -0.4195; P = 2.4 x 10^{-11}); with d_N having a stronger association with ω. A significant correlation is observed between the codon adaptation index (CAI) and evolutionary rates in all species, suggesting an obvious association of translation selection and enzyme evolution (Fig. 5.2). In comparison, features representing functional constraints demonstrate relatively weak species-specific associations with d_N, d_S and ω. In *L.*

major (Fig. 5.2A), d_N and ω are negatively correlated with number of processes in which a gene is involved (NumProcs), indicative of a weak functional constraint (d_N: r = -0.161; P = 0.014, ω: r = -0.172, P = 0.008). Similarly, the number of flux couplings per reaction associated with a gene (NCoup) significantly associates with ω (r = -0.152; P = 0.02). In *L. donovani* (Fig. 5.2B), d_N seems to weakly correlate with NumProcs (r = -0.16592; P = 0.011). In *L. infantum* (Fig. 5.2C), ω demonstrates a weak positive association with a gene's tendency to occur in a flux-coupled module (CCoFCA) (r = 0.165; P = 0.011).

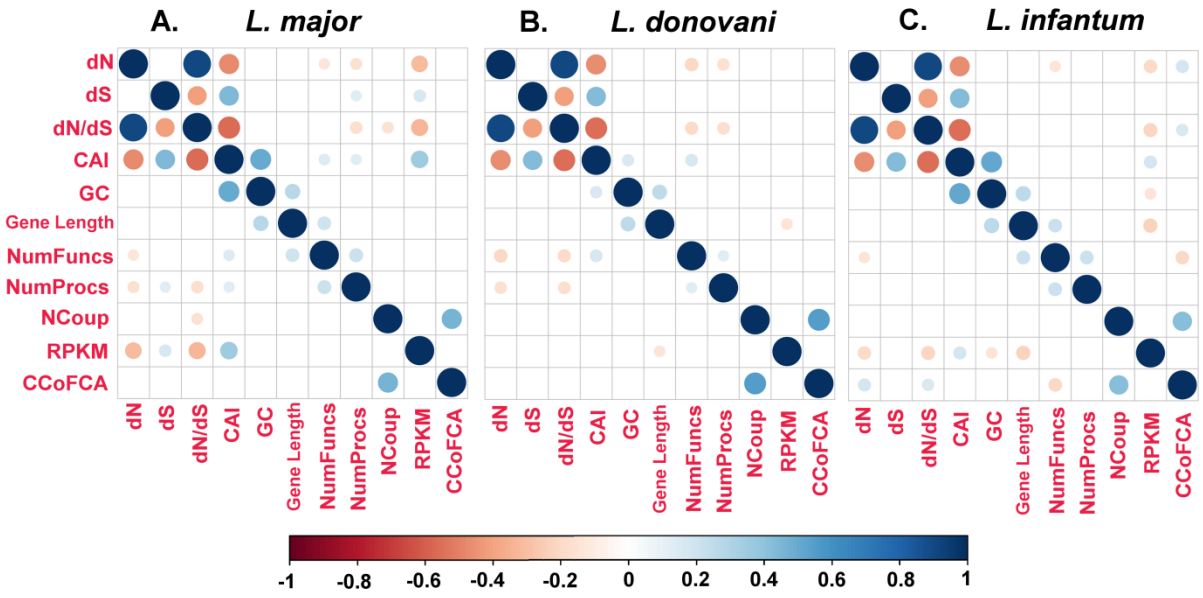

Figure 5.2. Correlation dot plot demonstrating inter-correlations between the eight predictors and evolutionary rates for the three *Leishmania* species - A) *Leishmania major;* B) *Leishmania donovani*; C) *Leishmania infantum*. This plot displays correlated pairs of features having significant correlation at $P < 0.05$. Dots represent significant positive or negative correlations. Colors and size of the dots represent the nature and degree of the association, as measured by Pearson's correlation.

Apart from associations of the predictors with evolutionary rates, inter-correlations between predictors were also observed. As observed in a previous study (Subramanian and Sarkar 2015), CAI also correlates positively with GC content with varying strengths of associations in each species. GC content of a gene increases with larger gene lengths as indicated by their significant association across species (Fig. 5.2). In *L. major* and *L. donovani* (Figs. 5.2A, 5.2B), CAI of a gene is positively associated with NumFuncs (*L. major*: r = 0.145, P = 0.026; *L. donovani*: r = 0.173, P = 0.008) suggestive of multifunctional genes to contain more frequent codons. As popularly known, CAI correlates with mRNA abundance (measured in reads per million kilobases, RPKM) in *L. major* and *L. infantum* (Figs. 5.2A, 5.2C). In *L. donovani* and *L. infantum*, gene length and RPKM are negatively

correlated suggesting expression of metabolic genes is probably limited by gene length in these species (Figs. 5.2B, 5.2C). Specifically in *L. infantum*, the number of functions associated with a gene (NumFuncs) demonstrates a weak negative association (r = -0.20133; P = 2 x 10^{-3}) with the tendency of a gene to cluster with genes demonstrating similar physiological fluxes (CCoFCA) hinting the role of multifunctional genes in routing fluxes within functional flux modules.

5.2.3. Contribution of features to the variation observed in enzyme evolutionary rates

As indicated in Fig. 5.2, although many features are independently correlated with the evolutionary rates, some of them are also inter-correlated with each other. Hence, it is difficult to identify the potential contribution of each individual features to evolutionary rates. For this purpose, PCR was performed to identify independent principal components, which represent a linear combination of features; the coefficients representing the weight of a particular feature in explaining the variation in d_N, d_S or ω (Drummond et al. 2006). PCR analysis with d_N and d_S in all three species indicates that the amount of variation explained by the principal components in the response variables (d_N and d_S) need not always be in descending order of the principal components (Jolliffe 1982). Additionally, it can also be observed that in most of the cases, a 90% variation in d_N and d_S cannot be explained by considering only the first few components suggesting that no single factor dominates enzyme evolutionary rates. Another important observation suggests that though the flux-topological features explain a low variance in d_N, their occurrence within the 1st principal component suggest that these features explain a majority of variation observed for metabolic genes in all three species.

With respect to d_N, it can be observed that the first two components (principal components 2, 3 of *L. major*, 2, 3 of *L. donovani* and 3, 7 of *L. infantum*), which cumulatively represent around 28.01% variance in *L. major* (Fig. 5.3A), 21.18% variance in *L. donovani* (Fig. 5.3D) and 23.41% variance in *L. infantum* (Fig. 5.3G) are dominated by genomic and gene expression features like CAI, GC, RPKM and gene length. In all the cases (Fig. 5.3A, 5.3D, 5.3G), the components dominated by flux-coupling potential and functional constraints explain a relatively small amount of variance in evolutionary rates. In *L. infantum* (Fig. 5.3G), a comparable amount of variation (7.92%) in d_N is explained by the principal components (1 and 8), which is dominated by metabolic flux-coupling potential of an enzyme, where the total variance explained in d_N by all the principal components is 38.21%.

With respect to d_S, it can be observed that the first two components (principal components 1, 8 of *L. major*, 2, 3 of *L. donovani* and 3, 8 of *L. infantum*), which cumulatively represent around 13.96% variance in *L. major* (Fig. 5.3B), 14.24% variance in *L. donovani* (Fig. 5.3E) and 21.37% variance in *L. infantum* (Fig. 5.3H) are dominated by genomic and gene expression features like CAI, GC, RPKM and gene length. A relatively large amount of variance (7.2%) is also explained in d_S rate of enzymes in *L. major* by two principal components governed by the flux-coupling potential, where the total variance explained in d_S by all the principal components is 25.6% (Fig. 5.3B).

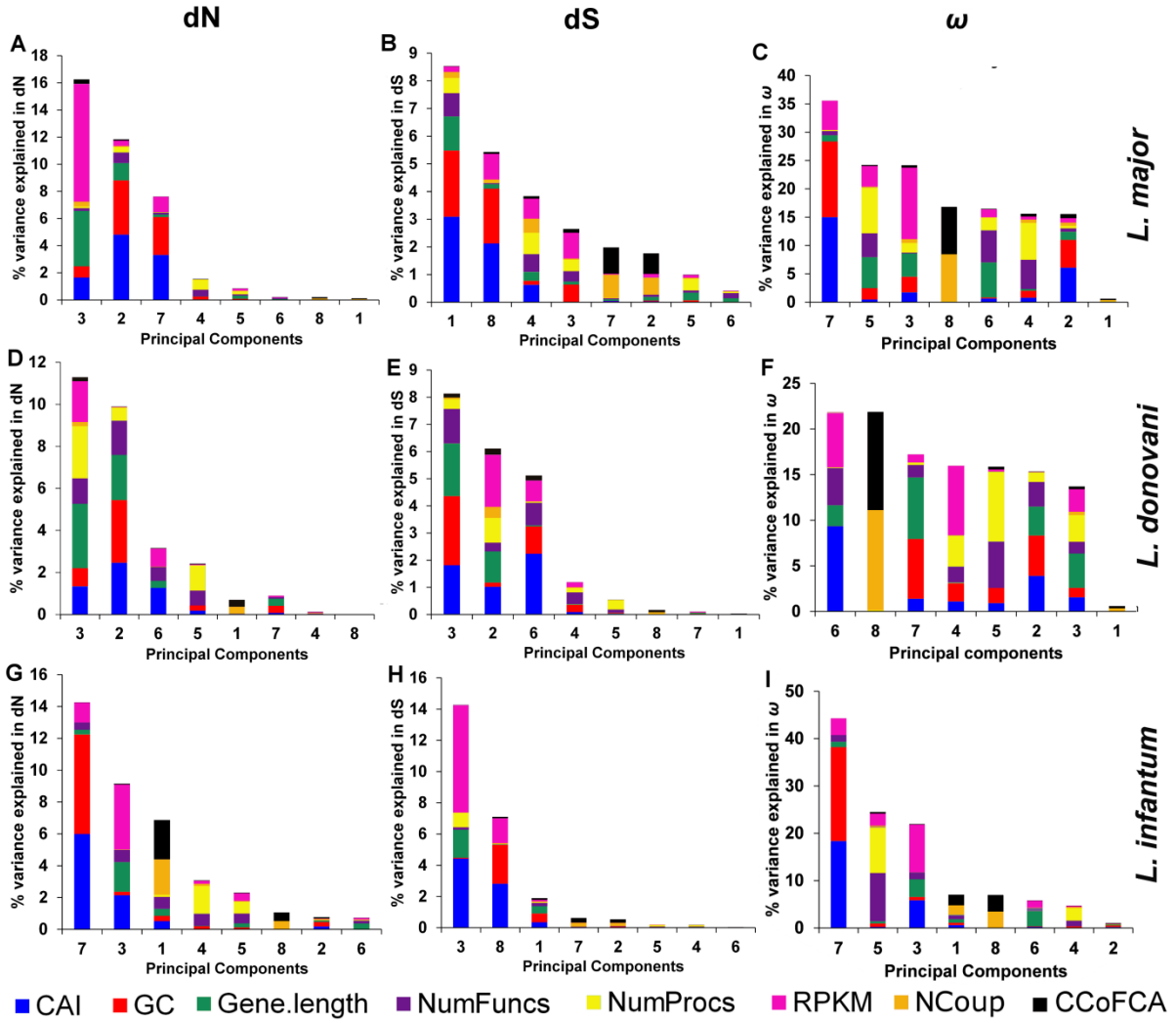

Figure 5.3. Principal component regression of evolutionary rates using 8 different features in the three *Leishmania* species - d_N (A, D, G), d_S (B, E, H) and ω (C, F, I) rates of 233 singleton orthologous metabolic genes in *L. major*, *L. donovani* and *L. infantum*. Each principal component represents a linear combination of the eight predictors, dominated by components that demonstrate a large variation in d_N and d_S. The colors correspond to the percentage variance explained by a particular feature, with respect to that principal component.

As observed for the d_N and d_S rates, the largest percentage of the total variation in ω is explained by features related to translation selection (CAI and GC content), as indicated by the 6th or 7th principal components in all the three species (variations - 35.5% in *L. major*, 21.88% in *L. donovani* and 44.33% in *L. infantum;* Figs. 5.3C, 5.3F, 5.3I). But, as observed in all the three species, multifunctionality and flux topology features explain larger variations in ω as compared to the individual d_N and d_S rates. An almost equal variation is explained by flux-coupled features (NCoup and CCoFCA) in *L. donovani* (8th principal component - 21.86%, Fig. 5.3F). The second largest percentage of variance is explained by the variable related to multi-functionality in *L. major* (24.19%, Fig. 5.3C) and *L. infantum* (24.51%, Fig. 5.3I). Similar to the d_N and d_S rates, no single component is alone enough to explain more than 90% of the variation in ω.

5.2.4. Selection of components for predicting enzyme evolutionary rates

A set of principal components were shortlisted for predicting evolutionary rates using a randomization test approach (see Chapter 2). Principal components 1 to 5 in *L. major* (Fig. 5.4A), 1 to 3 in *L. donovani* (Fig. 5.4D) and 1 to 7 in *L. infantum* (Fig. 5.4.G) are minimally required to explain significant amount of variation in d_N. Principal components 1 and 2 in *L. major* (Fig. 5.4B), 1 to 3 in *L. donovani* (Fig. 5.4F), and 1 to 3 in *L. infantum* (Fig. 5.4H) are minimally required to explain significant amount of variation in d_S. Principal components 1 to 7 in *L. major* (Fig. 5.4C), 1 to 3 in *L. donovani* (Fig. 5.4F) and 1 to 7 in *L. infantum* (Fig. 5.4I) are minimally required to explain significant amount of variation in ω.

Features with loadings greater than 0.45, were considered for interpreting a principal component (Table B.4). Most of the principal components explaining any variation in d_N or d_S, can be interpreted on the basis of three distinct classes of features – *a*. codon usage (CAI) and GC content, *b*. multi-functionality (NumProcs, NumFuncs) and *c*. flux phenotypic features (NCoup, CCoFCA). Most importantly, in all species (except *L. infantum*), effect of CAI and GC content of a gene on evolutionary rates can be interpreted by the same principal component suggesting their combinatorial effect in constraining d_N and d_S. To explain d_N rate of a gene, two principal components (2 and 7) involving CAI and GC content as principle features can be observed in *L. infantum*, where GC content negatively contributes to d_N in the 2nd principal component and positively contributes to d_N in the 7th principal component. Additionally, the 7th component has a relatively large role in explaining d_N as compared to the 2nd component. In all species, CAI negatively relates to d_N and positively relates to d_S. In all species, number of processes associated with a gene (NumProcs) negatively contributes to d_N.

Further, no principal component can be interpreted solely on the basis of gene length, to explain both d_N and d_S. Gene expression (RPKM) positively contributes to d_S rate in *L. donovani* and *L. infantum* and negatively contributes to d_N rate in *L. major* and *L. infantum*. In case of *L. major*, it can be seen that, distinct principal components (2 & 3) can be interpreted using CAI and RPKM respectively, suggesting weak associations with each other and their independent associations with d_N (Table B.4).

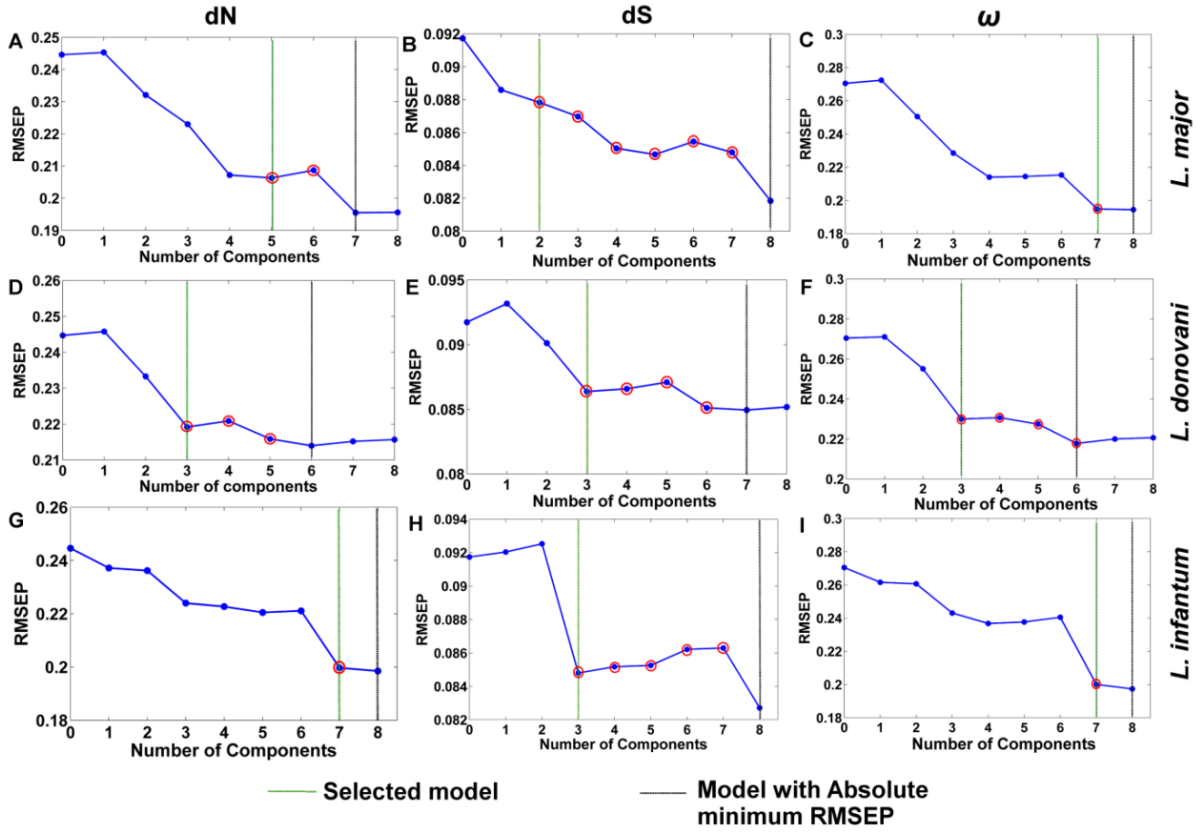

Figure 5.4. Selection of components for predicting evolutionary rates using a randomization test - d_N (A, D, G), d_S (B, E, H) and ω (C, F, I). Red circles indicate that the model performance with respect to a minimal subset of components is similar to the model displaying the absolute minimum RMSEP as predicted by the randomization test. Green line indicates the best selected model demonstrating optimal performance (minimal RMSEP) with a subset of all components. Black line indicates the model with highest performance (absolute minimum RMSEP).

To explain d_S in *L. donovani*, it can be seen that, distinct principal components (3 & 2) can be interpreted using CAI and RPKM respectively, suggesting their independent relationships with d_S and no association with each other (Table B.4). On the contrary, in *L. infantum*, principal component 3 can be interpreted by both CAI and RPKM suggesting their inter-relatedness. Interestingly, an important observation points out that synonymous substitution rates are not constrained by the multifunctional potential of a gene (NumFuncs, NumProcs).

Flux topological features significantly contribute to d_N rates of genes in *L. infantum* and d_S rates of genes in *L. major*.

Patterns common to both d_N and d_S are observed with respect to the ω rate across the three *Leishmania* species (Table B.4). Features related to translation selection (CAI, GC, RPKM, gene length) demonstrate a significant association with ω in all the three species. In *L. major*, translation selection is the only factor affecting ω. Multifunctionality (NumFuncs, NumProcs) is significantly associated negatively with ω in *L. donovani* and *L. infantum*. Further, in *L. infantum*, the flux topological features (NCoup, CCoFCA) are also significantly associated with the ω rate.

5.2.5. Relationship between physiological flux coupling and enzyme evolutionary rates

The pairwise correlation analysis indicated a weak correlation between flux-coupling features and evolutionary rates in *L. major* and *L. infantum* (Fig. 5.2). But, in the above analysis, it was found that across *Leishmania* species, physiological flux coupling potential seems to be a poor predictor of evolutionary rates (Table B.4). This relationship between evolutionary rates and flux-coupling potential can be affected because certain enzymes demonstrate no flux coupling with other reactions within the network. Apart from explaining variations, PCR analysis also allows us to classify genes into two clusters, with respect to the contribution of the predictor features of the genes (interpreted through a principal component) to a response. It was observed that the potential of an enzyme to be physiologically coupled to other enzymes within metabolism or not, can be classified only using scores of enzymes loaded on the first principal component (PC1) associated with the three evolutionary rates in all the species (Insets, Figs. 5.5A-5.5I).

With respect to this coupled set of enzymes (cluster 1 in Insets, 5.5A-5.5I), a negative relationship is observed between d_N or ω and the number of couplings associated with an enzyme with varying strengths (Figs. 5.5). In all three species, no association was observed between d_S and number of couplings (Figs. 5.5B, 5.5E, and 5.5H). With respect to the number of couplings, the association between d_N or ω and NCoup decreases as *L. major* < *L. donovani* < *L. infantum*. In *L. major* (Figs. 5.5A, 5.5C, 5.5D, 5.5F, 5.5G, 5.5I), the association, although weak, is statistically significant at $P < 0.01$ (d_N: $r = -0.252$, $P = 0.007$; ω: $r = -0.291$, $P = 0.002$). In *L. donovani* (Figs. 5.5D, 5.5F), the association is weaker than *L. major* (d_N: $r = -0.159$, $P = 0.094$, ω: $r = -0.198$, $P = 0.036$). In *L. infantum* (Figs. 5.5E and 5.5F), the association is the weakest and seems to be a purely chance phenomenon (d_N: r = -

0.019, P = 0.83; ω: r = -0.094; P = 0.29). The associations become weaker from *L. major* to *L. infantum* due to the gain or loss of flux couplings by enzymes across species. This gain or loss is affected by the coupling between duplicated and singleton genes in unique subcellular locations across species. Furthermore, the number of flux couplings observed for duplicated genes is much higher as compared to singletons.

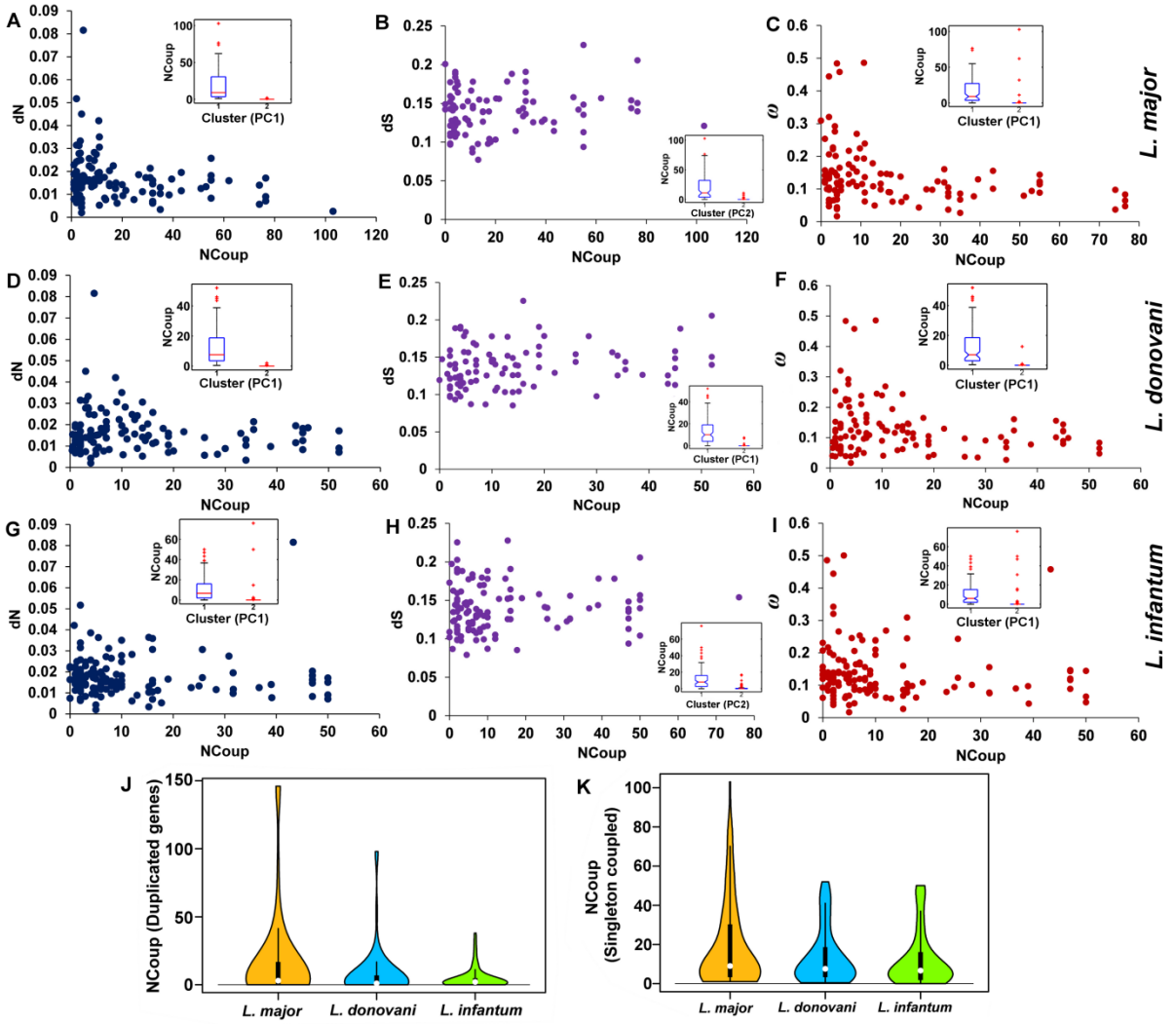

Figure 5.5. Association between rates of protein evolution and number of couplings (NCoup) is affected by gene duplications - Relationship between d_N rates and NCoup of flux-coupled set of enzymes is given for A) *L. major*; D) *L. donovani;* and G) *L. infantum*. Relationship between d_S rates and NCoup of flux-coupled set of enzymes is given for B) *L. major*; E) *L. donovani;* and H) *L. infantum*. Relationship between ω and NCoup is given for C) *L. major*; F) *L. donovani*; and I) *L. infantum*. J) Violin plot demonstrating the differences in the variance of number of couplings associated with duplicated genes between *L. major* (median = 3), *L. donovani* (median = 1.03) and *L. infantum* (median = 2); K) Violin plot demonstrating the differences in variance of singleton genes between *L. major* (median = 9), *L. donovani* (median = 7.55) and *L. infantum* (median = 6.64). Insets represent the two clusters of metabolic enzymes that are flux-coupled (1) and uncoupled (2).

Hence, we asked the question whether gene duplications affect the relationship between d_N or ω and number of couplings associated with an enzyme or not? Comparing the distributions of number of couplings associated with duplicated enzymes in the three species revealed that most of the duplicated enzymes are coupled to a less number of other enzymes within the metabolic network of *Leishmania* species (Fig. 5.5J). But, the variance in the number of couplings of duplicated enzymes is notably higher in *L. major* with some duplicated enzymes displaying a large number of couplings. On the contrary, the variance drastically reduces in *L. donovani* and *L. infantum* as compared to *L. major*. Similar to duplicated enzymes, comparing the distributions of number of couplings associated with coupled set of singleton enzymes in the three species also revealed that most of the singleton enzymes are coupled to a less number of other enzymes within the metabolic network of *Leishmania* species (Fig. 5.5K), with decreasing variance from *L. major* to *L. infantum*. Comparing the variance in the number of flux couplings across species in both the duplicated and singleton cases using Levene's test of homogeneity of variances (Martin and Bridgmon 2012) indicated that the variance in number of couplings are significantly different between species at $P < 0.001$ (duplicated: $F = 10.968$, $P = 3.25 \times 10^{-5}$, singletons: $F = 8.54$, $P = 2.6 \times 10^{-4}$). The similarity in distributions of number of couplings between duplicated enzymes and singletons indicates that more gene duplications might indirectly create new flux coupling associations with singletons, under stoichiometry, reversibility and environmental constraints, thereby promoting the association of the evolutionary rate with number of couplings associated with singleton genes. Furthermore, variance in number of couplings from *L. major* to *L. infantum* decreases at a slower rate in singletons as compared to duplicated enzymes indicating that the association between evolutionary rates and number of couplings in singletons is not promoted equally by all gene duplication events across species.

5.2.6. Identification of metabolic genes constrained by translation selection, multi-functionality and flux-topology

To identify genes whose evolutionary rates are governed by multiple classes of features, a cluster analysis was performed (see Chapter 2).

Identification of gene clusters with respect to d_N and d_S

The selected principal components representing an *n*-dimensional feature subspace of the entire 8-dimensional feature space were further subjected to *K*-means clustering for identifying clusters of metabolic genes, whose evolutionary rates can be explained by a combination of independent principal components. Akaike's Information Criterion (AIC)

values for every possible number of clusters were computed and the number corresponding to the minimum AIC was considered to be the possible number of clusters in the *n*-dimensional feature space (Fig. 5.6).

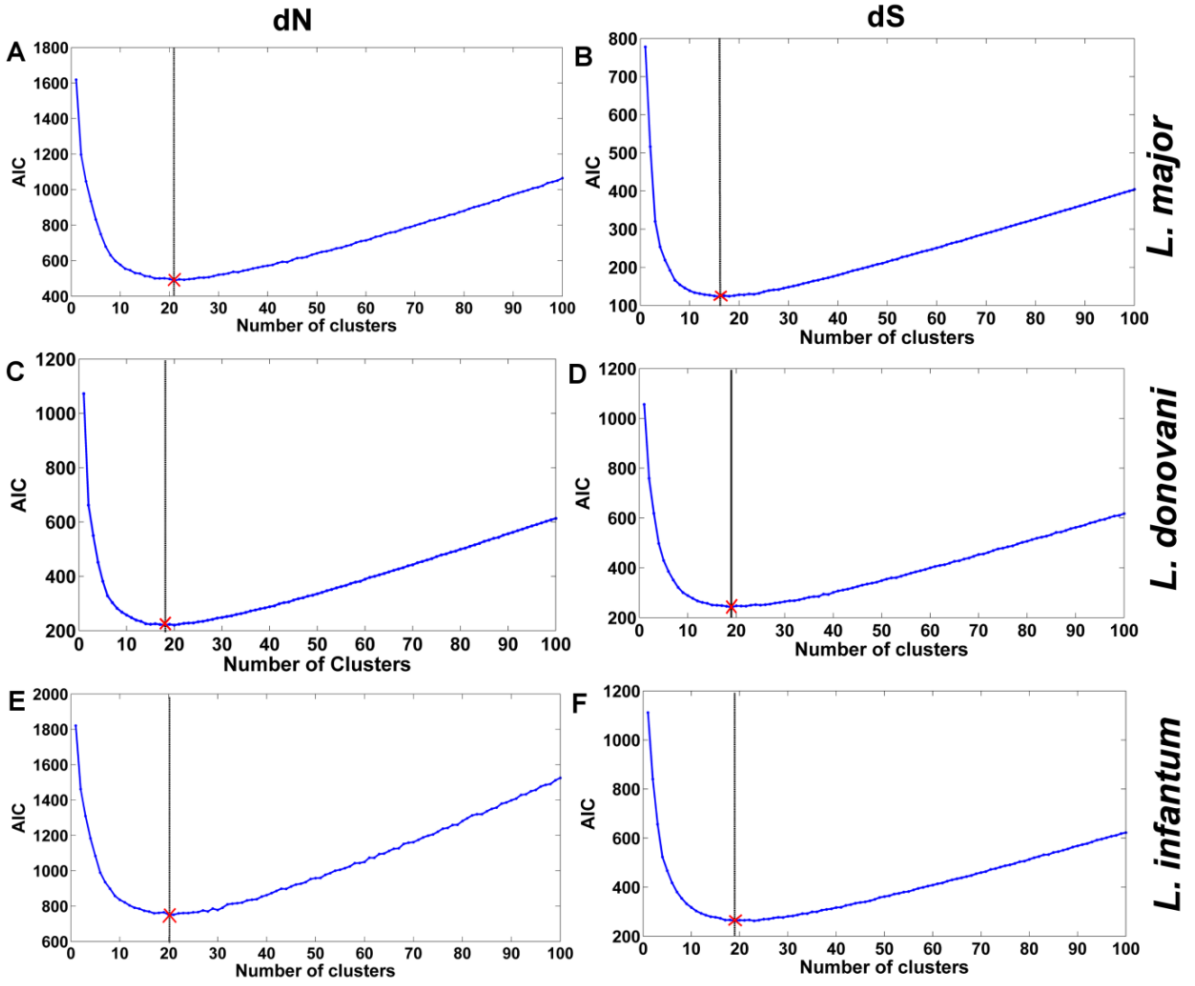

Figure 5.6. Number of gene clusters in the 8-dimensional feature space regressed with respect to evolutionary rates - d_N (A, C, E) and d_S (B, D, F). The plot demonstrates the Akaike's Information Criterion (AIC) for every *K* clusters.

Likewise, the d_N rate of 21 gene clusters in *L. major*, 18 clusters in *L. donovani* and 20 clusters in *L. infantum* were identified to be governed by a combination of features (Figs. 5.6A, 5.6C, 5.6E). Similarly, the d_S rate of 16 clusters in *L. major*, 19 clusters in *L. donovani* and 19 clusters in *L. infantum* can be explained by a specific subset of features (Figs. 5.6B, 5.6D, 5.6F). The centroid of each cluster within the *n*-dimensional feature space provides us with the coordinates corresponding to the scores of principal components, which represent the properties of gene clusters. This helps to identify the subset of genes whose evolutionary rates are dominated by a combination of features.

From Table B.4, it is possible to identify principal components that can be interpreted by the independent features namely, CAI, Number of processes (NumProcs) and number of flux couplings (NCoup) associated with a gene and the nature of their contributions to the evolutionary rates. Each of these features explains the role of translation selection, multifunctionality and flux topology respectively on evolutionary rates of metabolic genes. Observing the centroids of the clusters, the gene clusters that are associated with contributions of the above multiple principal components can be identified. Likewise, the d_N rate of genes in cluster numbers 4 and 14 in *L. major*, 9, 12, 13, 17 in *L. donovani* and 8, 18 in *L. infantum* are dominated by non-zero values of NCoup (positive scores on respective principal component in *L. major*, *L. donovani* and negative scores on respective principal component in *L. infantum*) and low values of NProcs (positive scores on principal component in *L. major*, *L. donovani* and negative scores on principal component in *L. infantum*) and CAI (negative scores on respective principal component in *L. major*, *L. donovani* and positive scores on respective principal component in *L. infantum*). Multi-functionality, which is represented by NumProcs or NumFuncs does not appear to be a dominant predictor in explaining the d_S rate and hence, does not occur as a major contributor in any of the selected components (Table B.4). Hence, those gene clusters whose evolutionary rates can be interpreted by CAI and flux-topology alone were identified. Likewise, the d_S rate of genes in cluster numbers 2, 3, 4, 7, 8, 11 in *L. major*, 1, 2, 3 10, 15, 18 in *L. donovani* and 2, 3, 10, 18 in *L. infantum* are associated with high values of CAI (positive scores on respective principal component in all three species) and NCoup (positive scores on respective principal component in *L. major* and *L. infantum* and negative scores on respective principal component in *L. donovani*). Comparison of chosen genes between the species indicates 5 genes in all species, whose evolutionary rates are dominated by all the three factors – translation selection, multi-functionality and flux-topology; whereas, 13 genes whose evolutionary rates are governed by translation selection and flux-topology (Fig. 5.7). There is a larger overlap of genes between the *L. major* and *L. donovani* species with respect to d_N as compared to d_S. Further, the overlap between *L. donovani* and *L. infantum* is restricted with respect to d_N as compared to d_S. In all species, there are also a unique set of genes whose evolutionary rates are specifically explained by the identified independent features (Tables B.5, B.6, B.7).

The following sets of genes were identified to be unique across the three species and the functions of the genes were identified through the eggNOG classification (Table B.5). The species-specific genes whose d_N rates can be explained by CAI, NumProcs and NCoup

are given in Table B.6. The species-specific genes whose d_S rates can be explained by CAI and NCoup are given in Table B.7.

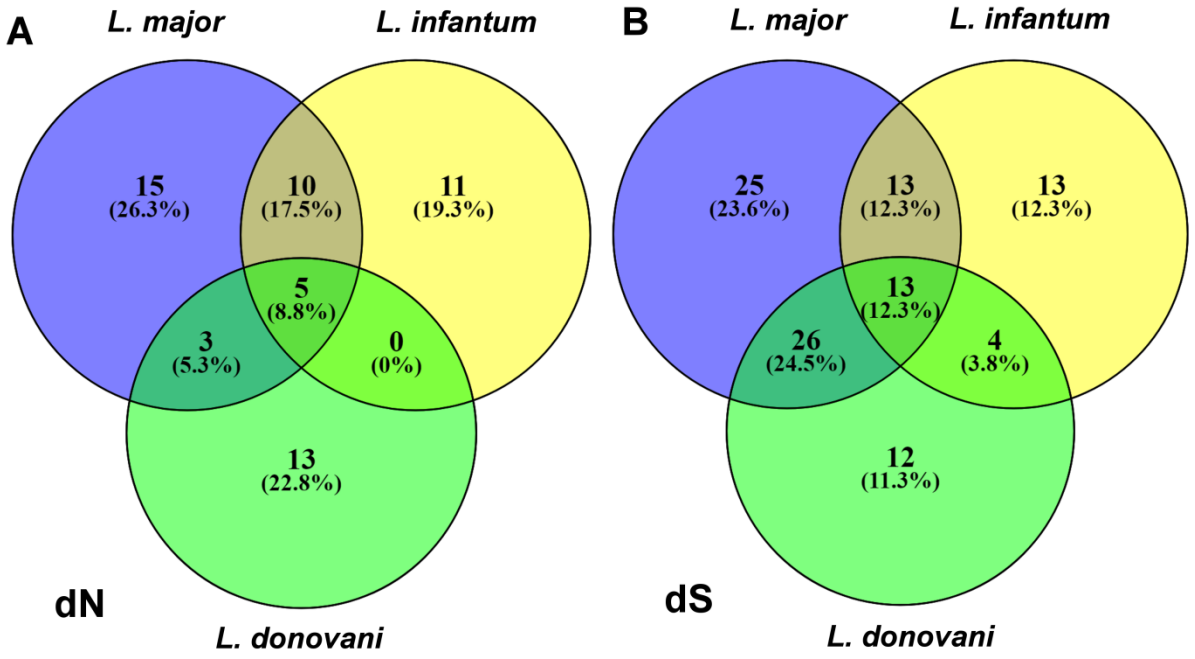

Figure 5.7. Comparison of genes demonstrating high values of independent dominant factors between species (namely, codon adaptation, number of biological processes and number of flux-coupling associations) with respect to - A) d_N and B) d_S rates.

5.3. Discussion

Owing to its parasitic nature and the long-standing evolutionary association with hosts, *Leishmania* species experience a largely constrained metabolic environment. For efficient adaptation within the host, both translation selection and functional constraint might constrain evolution of enzymes within *Leishmania* metabolism. To our knowledge, there is no study available till date in *Leishmania* parasites, that compares these heterogeneous potential determinants in predicting non-synonymous (d_N) and synonymous (d_S) substitution rates in metabolic enzymes simultaneously, on a single platform. Also, the inter-relationship between these factors and their differences across species is seldom explored. As used in other eukaryotes (Drummond et al. 2006; Yang and Gaut 2011; Alvarez-Ponce et al. 2017), the present study integrates the available, potential features of metabolic enzymes into a principal component-based regression model to identify the unknown confounding factors that explain observed variation in the evolutionary rates and compares them across three *Leishmania* species.

As observed in other eukaryotes (Drummond et al. 2006), codon usage negatively correlates with d_N, ω and positively correlates with d_S in all species, signifying translation selection to be an important constraint in *Leishmania* metabolic enzyme evolution. This can also be observed from the highest percentage of variation explained by the principal component dominated by CAI. Furthermore, GC content also occurs as a dominating factor of the same principal component as CAI, indicating their relatedness, supporting previous observations (Subramanian and Sarkar 2015). But, as observed in all the three *Leishmania* species, neither a single principal component is enough to explain a significant proportion of variation among evolutionary rates nor does a single set of similar features explain sufficient variation across principal components, indicating that multiple features potentially contribute to enzyme evolution in *Leishmania* species. Hence, more than one principal component was observed to be selected for regression (van der Voet 1994). Although with an exception in *L. infantum*, results indicate that gene expression (RPKM) does not always occur in the same principal component as CAI, suggesting their independent roles in governing evolutionary rates of enzymes. This is contrary to the observations in yeast and *E. coli*, where gene expression complements CAI as a dominant factor governing evolutionary rates (Drummond et al. 2006). This might be due to the weak association observed between mRNA and protein abundances in *Leishmania* species (Lahav et al. 2011); CAI being an important predictor of protein abundance (Subramanian and Sarkar 2015). Similarly, the occurrence of CAI, multi-functionality and flux-coupling features as dominant features on distinct principal components suggests that these features affect evolutionary rates independently. Further, the multi-functionality of a gene (NumProcs, NumFuncs) contributes only to the non-synonymous substitution rate (d_N) and is negatively associated with d_N. Hence, as observed in yeast (Salathé et al. 2005), genes (enzymes) with multiple processes or functions evolve slowly as compared to genes associated with low number of functions in the *Leishmania* species as well.

For the first time, we introduce the notion of the flux-coupling potential of an enzyme within its metabolic network and investigate whether it is an important determinant of evolutionary rate in *Leishmania* species or not. Although the associations of the flux-coupling features with evolutionary rates are weak, unlike multi-functionality, the occurrence of flux topological features in the first principal component and the selection of their associated principal component for regression against evolutionary rates explains their important contribution to variation in both d_N and d_S rates. Supporting this factor a significant amount of variation in the d_S rate of enzymes in *L. major* and d_N rate of enzymes in *L. infantum* is also

sufficiently explained by these features. Considering only the flux-coupled set of enzymes in all three species, a weak negative association can be observed between d_N, ω and number of couplings associated with an enzyme (NCoup). Flux-coupling reaction subsets capture the total number of paths of metabolite distribution under defined uptake constraints, as they can explain co-regulation between metabolic genes (Notebaart et al. 2008). A negative association was observed between ω and metabolic flux through an enzyme in yeast, human RBCs and *L. major* (Vitkup et al. 2006; Colombo et al. 2014; Subramanian and Sarkar 2016). This suggests that an enzyme is slow-evolving if it is coupled to large number of other enzymes by flux (hubs) within the flux-coupled network when compared to enzymes with low number of couplings. Further, few numbers of enzymes with high number of flux-couplings are observed as compared to enzymes with low number of flux couplings. This indicates that a hierarchical organization of fluxes within *Leishmania* metabolism is largely constrained during evolution.

Chromosomal aneuploidy in *Leishmania* gives rise to significant variations in copy numbers of genes across species that might increase genomic plasticity, gene dosage, and rescue of essential functions from deleterious mutations (Mannaert et al. 2012). In addition to the aforementioned roles, for the first time, we document an observation indicating a possible species-specific involvement of duplicated metabolic enzymes in increasing the evolutionary constraints on other metabolic enzymes within a network, through re-wiring of physiological flux dependencies within the metabolism. This is typically indicated by a higher variance in the number of couplings associated with singleton & duplicated enzymes and relatively stronger associations between number of couplings associated with singletons & evolutionary rates. With decrease in the variance of number of couplings of duplicated enzymes from *L. major -> L. donovani -> L. infantum*, the strength of associations between number of couplings and evolutionary rates also reduces. A similar rewiring of fluxes due to cross-compartmentalized metabolism was also hypothesized for glycolysis and isoprenoid biosynthesis in other Trypanosomatids (close evolutionary relatives of *Leishmania*) and other protists (Ginger et al. 2010). Interestingly, not all gene duplications are highly flux-coupled with other enzymes in the network, suggesting that the species-specific metabolic network structure dynamically constrains the choice of unique gene duplications occurring at multiple subcellular locations for flux rewiring, thereby, imposing evolutionary constraints on other singletons associated with them.

Previously, codon bias, pleiotropy and centrality within a biomolecular network were implicated to impose relatively strong evolutionary constraints on enzymes that are important

pharmacological targets for a disease (Searls 2003; Pál et al. 2006; Gladki et al. 2013; Lv et al. 2016). As mentioned above, codon adaptation, multi-functionality and flux-topological constraints independently affect evolutionary rates; each of these features being negatively associated with d_N. Comparison of genes with the d_N rate dominated by these factors lead to the identification of both common and species-specific enzymes, which are evolutionarily constrained by multiple genotype-phenotype factors, reckoning them to be important enzymes. Likewise, this analysis was able to identify enzymes like trypanothione reductase, aspartate carbamoyltransferase, orotidine-5-phosphate decarboxylase and dihydrolipoamide dehydrogenase common to all three species. Among the enzymes common to the three *Leishmania* species, trypanothione reductase, the sole enzyme in the *Leishmania* parasite to combat oxidative stress (Tovar et al. 1998), aspartate carbamoyltransferase and orotidine-5-phosphate decarboxylase, involved in production of pyrimidines, like ump and cmp, (Mukherjee et al. 1988; Bello et al. 2007) are previously speculated pharmacological targets in *Leishmania* and other eukaryotes. On the other hand, unique enzymes majorly belonging to energy metabolism and conservation (C), Carbohydrate transport and metabolism (G), Amino acid transport and metabolism (E) and Nucleotide transport and metabolism (F) were also identified for each species. Among these unique enzymes, known virulence factors like trypanothione synthetase, phosphomannose isomerase and GDP-mannose pyrophosphorylase were specifically identified for *L. major;* dihydrofolate-reductase/thymidylate synthase, pyrroline-5-carboxylate reductase and phosphomannomutase were identified for *L. infantum* and tyrosine aminotransferase for *L. donovani* (Mukherjee et al. 1988; Titus et al. 1995; Tovar et al. 1998; Garami and Ilg 2001a; Garami and Ilg 2001b; Scott et al. 2008; Moreno et al. 2014; Mantilla et al. 2015). Their role in virulence probably makes them more resistant to change. From this analysis, few more novel species-specific enzymes were also predicted. Their biological role in virulence or survival of the parasite needs to be experimentally investigated.

Chapter 6 – Conclusion and Future directions

6.1. Conclusion

Leishmaniasis is a complex, multi-faceted disease that still remains the second largest parasitic killer worldwide. The widespread prevalence of the disease depends upon the successful survival mechanisms of the *Leishmania* parasite within the sandfly vector and human hosts. This is largely governed by genotype-phenotype factors expressed by the parasite that either increase proliferation, vitality, virulence within the hosts or evade host immune mechanisms. With respect to the varying capacities of the above factors, *Leishmania* species are also known to demonstrate differential clinical manifestations in the human host (McCall et al. 2013). Hence, it is necessary to identify and segregate the stage and species-specific factors and mechanisms that contribute to this broad spectrum of complications. The work performed in this thesis aims towards identification of these factors that might provide new insights into the adaptive strategies that helps the *Leishmania* parasite to endure host stresses thereby, ensuring viability.

Comparisons of sequenced genomes of different *Leishmania* species identified that very few genes are putatively identified to be specific to a particular *Leishmania* genome (Peacock et al. 2007; Rogers et al. 2011). A similar perspective was also discovered from transcriptomic experiments that identified a very few stage-specific or species-specific differences in transcriptome abundances, which is unique in *Leishmania* as opposed to other prokaryotes and eukaryotes (Leifso et al. 2007; Depledge et al. 2009). Furthermore, the mRNA and protein levels did not correlate with each other in *Leishmania* questioning the contribution of mRNA level variations to the parasite phenotype (Lahav et al. 2011). As opposed to the transcriptome, more stage-specific differences were identified in proteome abundances for different *Leishmania* species (Rosenzweig et al. 2008; Nirujogi et al. 2014).

But, an important limitation of mass-spectrometry approaches remains the inability to detect relatively small but significant changes in proteome abundances that are most likely to occur in case of *Leishmania* parasites (Gygi et al. 2000; Avila-levy 2014). In other prokaryotes and eukaryotes, codon usage was demonstrated as an important predictor of protein abundance (Sharp and Li 1987; Horn 2008; Plotkin and Kudla 2010). As discussed in Chapter 3, a comparative analysis across *Leishmania* and other related Trypanosomatids was performed to study the causes and consequences of codon usage in *Leishmania* species (Subramanian and Sarkar 2015). Directional mutation pressure and translation selection both act towards maintaining a G or C at the synonymous position of the codon, thereby, preserving frequent codons in genes.

An additional codon context analysis indicated that homogenous codon contexts were more frequent in *Leishmania* as compared to non-homogenous contexts. This is probably because choice of homogenous codon contexts corresponding to the A and P sites of the ribosome maybe energetically less expensive during translation as the same tRNA can possibly be used twice for aminoacylation of codons with homogeneous contexts. Codon usage significantly correlates with protein abundance, implying it to be a predictor of proteome abundance across *Leishmania*. Codons that avoid secondary structure formation in mRNA are preferred within the *Leishmania* coding sequences with non-random codon usage patterns unique to each *Leishmania* species. Considering CAI to be a predictor of relative protein abundance, genes belonging to specific predicted GO functions demonstrated a biased codon usage. Enzymes for protein phosphorylation, proteolysis, ubiquitin-dependent protein catabolic process, RNA splicing, and protein folding tend to show a high variance in codon adaptation representing putative differential expression of these pathways between species. A large percentage of genes in *L. donovani* belonging to pathways related to pathogenesis, phagocytosis, response to oxidative stress, heme, fatty acid oxidation, glycerol metabolism, mannose metabolism and translation demonstrated a high codon adaptation suggestive of a putative global up-regulation of genes in *L. donovani* as compared to other species. Further, enzymes related to essential pathways required for parasite survival like energy metabolism, demonstrated high CAI values in both *L. donovani* and *L. infantum* as opposed to *L. major*, *L. mexicana* and *L. braziliensis*, demonstrating evolutionarily related codon usage. All these results conclude that a large-scale comparative codon usage analysis might further help to uncover probable evolutionary and functional differences among *Leishmania* species which is otherwise, at a raw sequence level conserved and not visible. This also helps to understand the role of evolution in shaping the otherwise conserved *Leishmania* genome to demonstrate differential expression and function-level differences for efficient survival of the parasite within the host.

As opposed to other prokaryotes and eukaryotes, the number of differences identified in the transcriptome and proteome across stages was very modest, suggesting the *Leishmania* genome to be constitutively expressed (Leifso et al. 2007). Isotope-resolved metabolomics experiments in *L. mexicana* recently discovered that the stage-specific changes in metabolism are hardwired in the uptake rates of certain metabolites that relative changes in metabolic adaptations, as opposed to relative changes in enzyme levels (Saunders et al. 2014). The changes observed across the two developmental stages are quantitative with respect to variable uptake rates while following largely similar metabolic routes. This further questions

whether the phagolysozome represents a stressed or a minimal environment for the parasite. If one considers the environment to be sub-optimal, it is more intriguing to understand the underlying organization of *Leishmania* metabolism that contributes to this adaptation. As discussed in Chapter 4, for the first time, two new model genome-scale metabolic reconstructions for the *Leishmania infantum* species are proposed (Subramanian et al. 2015; Subramanian and Sarkar 2017). The iAS142 model is an energy metabolic model whereas the iAS556 model is a comprehensive genome-scale reconstruction. Through the analysis of the energy metabolic model (Subramanian et al. 2015), glycosomal succinate fermentation due to a novel set of reductases related to malate, fumarate and succinate occurring in glycosome, a novel NADPH cytochrome P450 oxidoreductase, a dual cytosolic/glycosomal glutamate dehydrogenase and the mitochondrial ATP synthase were predicted to be essential enzymes representing the core energy metabolism. Moreover, it was also predicted that glucose and non-essential amino acid catabolic routes are largely coupled with each other. The iAS556 genome-scale metabolic network and flux-based analysis yielded a similar set of results detecting similar catabolic routes for multiple carbon sources across the two developmental stages (Subramanian and Sarkar 2017). With respect to the analysis of the iAS556 model, it was hypothesized that the parasite possesses a unique network of metabolic reactions whose functioning can be explained with respect to the stoichiometric and reversibility constraints imposed by the underlying metabolic network structure alone; the network structure being largely governed by the species-specific occurrence of enzymes in multiple subcellular locations. This creates a dynamic non-essential amino acid network motif that provides a restricted re-distribution of resources within metabolism. This further reinforces the role of glucose and tyrosine as sources to only supplement the quantities of output metabolites formed through non-essential amino acids. This also ensures that fatty acids, amino sugars and mannose can compensate for reduced glucose uptake within the amastigote. Fatty acids produce intracellular glutamate, whereas amino sugars and mannose satisfies the production of intracellular aspartate and mannan, respectively. The *L. infantum* flux-coupled network structure is highly modular and assortative (Hase et al. 2010; Subramanian and Sarkar 2016), thereby remaining robust to a random node failure. The modular nature created by the species-specific presence of enzymes in dual/multiple subcellular locations imply adaptive characteristics of *Leishmania* metabolism to the changing host environments and optimal utilization of resources. Many of these observations can pave ways to design suitable experiments that can explore the different possibilities of metabolic adaptation of *L. infantum* within the host.

Chapters 3 and 4 have indirectly indicated the important translational and functional constraints of codon usage and metabolic organization on the evolution of enzymes and hence, the survival strategies of the *Leishmania* parasites. So as to test the contribution of these broad constraints along with other related features on the evolution of metabolic genes, as part of Chapter 5, a principal component-based regression model that can tease apart these important confounding factors governing evolutionary divergence of genes was proposed for three different *Leishmania* species. This analysis reveals that codon adaptation, multi-functionality and flux-coupling potential of an enzyme are independent contributors of enzyme evolutionary rates, which can together explain a large variation in enzyme evolutionary rates across species. Also, as opposed to other prokaryotes and eukaryotes, codon usage and transcriptome abundance were predicted to be independent of each other, owing to the weak correlation between protein and transcriptome abundances (Lahav et al. 2011). Also, for the first time, it was observed that a species-specific occurrence of duplicated genes in novel subcellular locations can create new flux routes through certain singleton flux-coupled enzymes, thereby constraining their evolution. A cross-species comparison revealed both common and species-specific genes whose evolutionary divergence was constrained by multiple independent factors. Out of these, previously known pharmacological targets and virulence factors in *Leishmania* were identified, suggesting their evolutionary reasons for being important survival factors to the parasite. The identification of multiple factors in constraining evolutionary divergence within metabolic enzymes suggests that the survival and adaptation of the parasite within the host is a complex problem. This emphasizes the need for systems-level wet-lab experiments to identify other features like UTR length, recombination rate, gene essentiality, protein-protein interactions features, etc. unavailable at an organismal level for *Leishmania* species and to analyze their integrated effect. The integration of these diverse features can thus provide the complete knowledge of the strategies employed by the parasite for survival and virulence, which can help the community to combat this largely neglected tropical parasitic infection.

6.2. Future directions

Leishmaniasis is a complex disease that is likely to be governed by a number of underlying factors. Although many new mechanisms were uncovered in this work, it still remains a challenging problem to bridge heterogeneous data sets, scrutinize them and arrive at a logical conclusion. The observations in each chapter of the thesis provide new insights that can be

further explored in details and many testable hypotheses for researchers and systems biologists in this field.

In all the aforementioned chapters, numerous important genes and mechanisms have been identified. These provide testable hypotheses for gene essentiality experiments. The predicted essential reactions can be used as novel targets for inhibitor prediction and prioritization. From this work, the influence of codon usage on mRNA secondary structures was investigated. The role of mRNA secondary structures at different locations of genes needs to be tested. This can also be verified by designing experiments involving ribosomal footprinting approaches. Also, with the availability of proteomics experiments with whole proteome coverage for multiple *Leishmania* species, their relationship with protein abundances will validate mRNA secondary structure formation as a codon usage-based mechanism for global translation regulation across *Leishmania*.

A species-specific role of enzyme organization was observed while analyzing the genome-scale reconstructions of *Leishmania infantum* metabolism. This difference across species was due to differences in gene duplications and their role in creating new flux routes that are optimal across environments given biological/biochemical constraints of stoichiometry, reversibility and subcellular compartmentalization. Solving genome-scale metabolic networks of other *Leishmania* species and other closely related Trypanosomatid species might help to reveal the pattern of duplications across species and identify other novel mechanisms by which metabolic organization has evolved in the Kinetoplastid lineage. To detect flux-based changes experimentally across species, there is also a necessity of comparative ^{13}C isotope-resolved metabolomics experiments in a variety of *Leishmania* species.

The evolution of metabolic enzymes across *Leishmania* species being a multi-variable problem is governed by many genotype-phenotype factors. Measurement of features like mRNA degradation rates, UTR length, recombination rate, gene essentiality, protein-protein interaction features, etc. for *Leishmania* species is required. Once the parasite frontier is explored sufficiently, the interplay between the parasite adaptations on the host and the resistance created by the host on the parasite can be the next level of scrutiny, thereby, covering ground on all aspects of clinical manifestations.

The work performed in this thesis has proven to be instrumental in understanding the factors playing roles at multiple levels of complexities in the parasite. These leads may further help the scientific community to combat this neglected but deadly tropical parasitic infection.

Appendix A

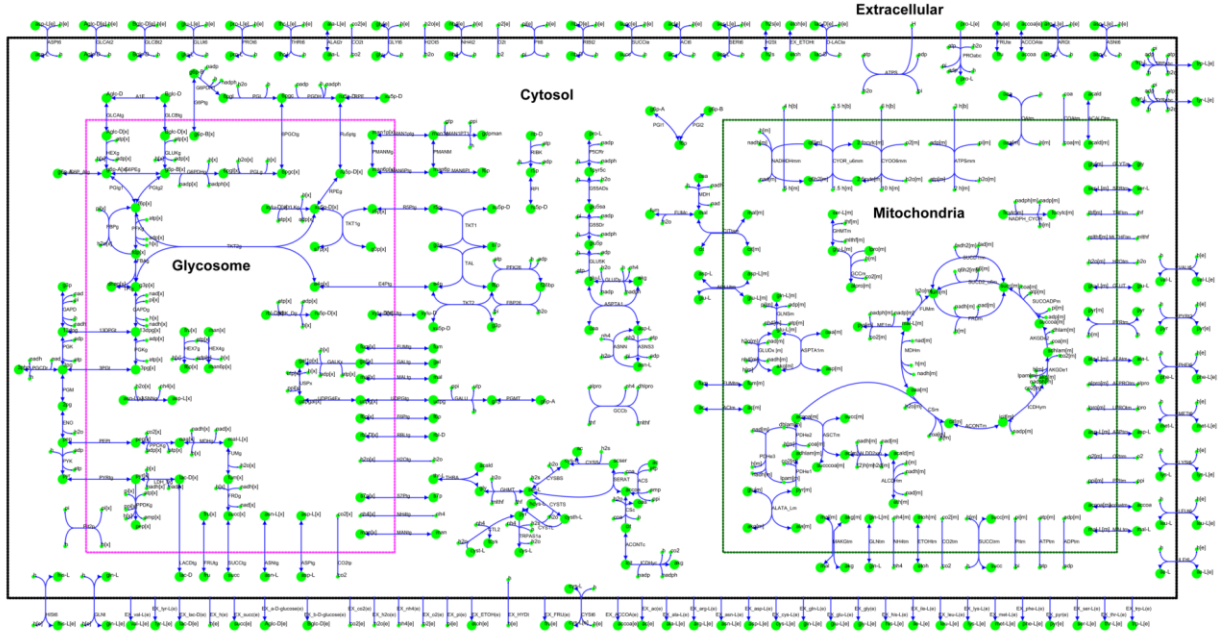

Figure A.1. The map of the iAS142 metabolic network – The iAS142 network comprises of 237 reactions that occur in 5 major model compartments: 4 cellular compartments - the glycosome, cytoplasm, the mitochondrion, the mitochondrial inter-membrane space and 1 extracellular compartment as shown in the figure. The mitochondrial inter-membrane space though included in the model, is not explicitly shown in the figure. The reactions of oxidative phosphorylation occur from the mitochondrial compartment (m) to the mitochondrial inter-membrane space (mm) and vice versa. The reactions occurring along the borders of the compartments are transport reactions. The metabolite exchanges have been shown at the bottom of the figure.

Appendix A

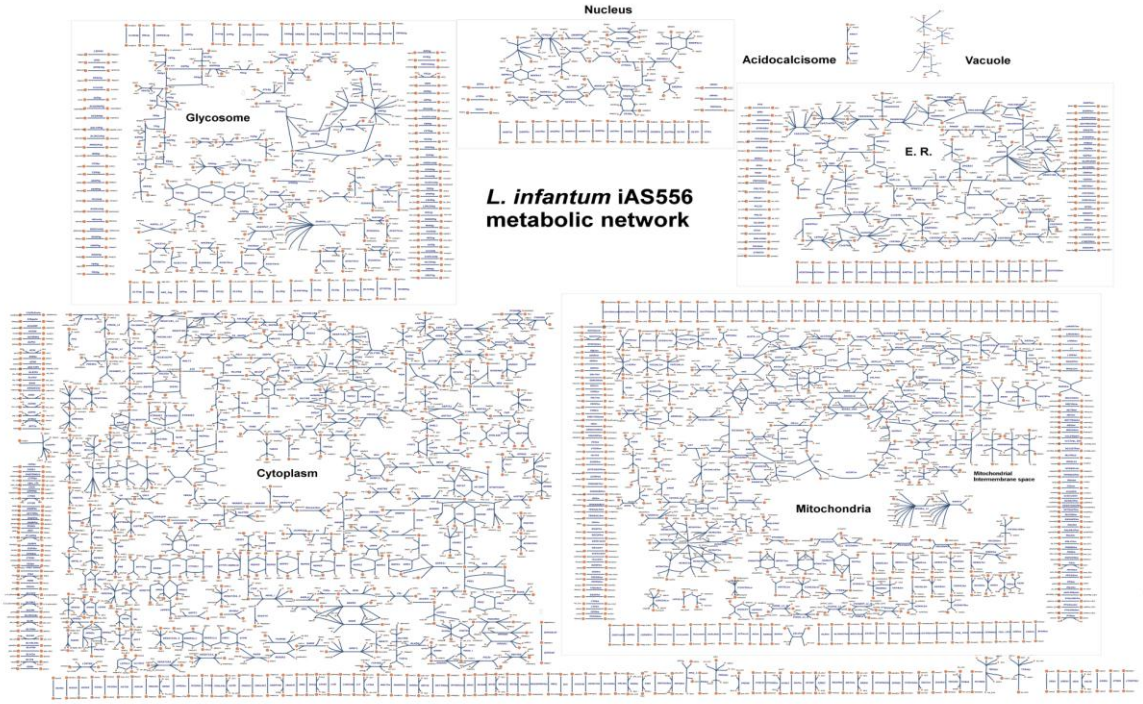

Figure A.2. Schematic figure of the complete genome-scale metabolic network of *L. infantum* (iAS556) - The iAS556 metabolic network consists of 1260 total reactions spanning the 9 different subcellular compartments - cytosol, extracellular, glycosome, mitochondrion, nucleus, endoplasmic reticulum, flagellum, acidocalcisome and vacuole.

Appendix B

Table B.1. Comparison of reaction subcellular locations of the iAS142 model with the iAC560 and iSR215 models.

Reaction (subcellular location)	Reaction abbreviation (as per iAS142)	iAS142	iAC560	iSR215	Confidence score in iAS142
Glucose-6-phosphate isomerase	PGIg1/PGIg2	Glycosome	Glycosome	Glycosome	5
	PGI1/PGI2[ǂ]	Cytoplasm		Cytoplasm	4
D-Lactate dehydrogenase	LDH_Dg[€,£]	Glycosome	Glycosome	Not considered	2
			Cytoplasm		
Citrate lyase	CSc*	Cytoplasm	Not considered	Not considered	5
Aconitase	ACONTm	Mitochondria	Mitochondria	Mitochondria	5
	ACONTc*	Cytoplasm			4
Isocitrate dehydrogenase	ICDHym	Mitochondria	Mitochondria	Mitochondria	5
	ICDHyc*	Cytoplasm			5
Fumarate hydratase	FUMm	Mitochondria	Mitochondria	Mitochondria	4
	FUMg	Glycosome	Glycosome	Glycosome	5
	FUMc[ǂ]	Cytoplasm		Cytoplasm	5
Glutamine synthetase	GLNSm[¥]	Mitochondria	Cytoplasm	Not considered	2
Asparaginase	ASNNg	Glycosome	Glycosome	Not considered	2
	ASNN*	Cytoplasm			1
glycine/serine hydroxymethyltransferase	GHMT	Cytoplasm	Cytoplasm	Not considered	5
	GHMTm*	Mitochondria			5

(continued in the next page)

Appendix B

pyrroline-5-carboxylate reductase	P5CRr[€,£]	Cytoplasm	Mitochondria Cytoplasm	Mitochondria	2
Glycine cleavage complex	GCCb[€]	Cytoplasm	Cytoplasm	Not considered	5
	GCCm*	Mitochondria			4
Ribokinase	RIBK[*]	Cytoplasm	Not considered	Glycosome	2
UDP sugar phosphorylase	USPx*	Glycosome	Not considered	Not considered	2
Aldose 1 epimerase	A1E[€,£]	Cytoplasm	Glycosome Cytoplasm	Glycosome	2
Alanine aminotransferase	ALATA_Lm[€,£]	Mitochondria	Cytoplasm Mitochondria	Cytoplasm Mitochondria	2
Hexokinase	HEXg, HEX4g, HEX7g[€]	Glycosome	Glycosome	Cytoplasm	5
Fumarate reductase	FRDg[€]	Glycosome	Glycosome	Glycosome	5
	FRDm[€]	Mitochondria	Mitochondria		5
Ribose phosphate isomerase	RPI[€]	Cytoplasm	Cytoplasm	Glycosome Cytoplasm	2
Ribulokinase	RBK_Dg[€]	Glycosome	Glycosome	Absent	2
pyrroline-5-carboxylate synthetase	G5SADs[€]	Cytoplasm	Cytoplasm	Mitochondria	--
glutamate-5-semialdehyde dehydrogenase	G5SDr[€,£]	Cytoplasm	Cytoplasm Mitochondria	Mitochondria	2
Malic enzyme	ME1m[¥]	Mitochondria	Cytoplasm	Cytoplasm	2

(continued in the next page)

Appendix B

				Mitochondria	
Acetyl-coA synthetase	ACS¥	Cytoplasm	Mitochondria	Not considered	5
Alcohol dehydrogenase	ALCDHm¥	Mitochondria	Cytoplasm	Not considered	2
NADPH cytochrome oxidoreductase	NADPH_CYOR*	Mitochondria	Not considered	Not considered	4
NADH:ubiquinone oxidoreductase	NADHDHmm¥,#	Mitochondrial membrane	Mitochondria	Mitochondria Cytoplasm	5
ubiquinol-6 cytochrome c reductase	CYOR_u6mm¥,#	Mitochondrial membrane	Mitochondria	Mitochondria	5
cytochrome c oxidase	CYOO6m¥,#	Mitochondrial membrane	Mitochondria	Mitochondria	5
V-type H+-transporting ATPase subunit A	ATPSmm¥,#	Mitochondrial membrane	Mitochondria	Mitochondria	5
	ATPS	Cytoplasm	Acido-calcisome Cytoplasm	Cytoplasm	5
phosphoglucomutase	PGMT€	Cytoplasm	Cytoplasm	Not considered	5
Aldehyde dehydrogenase	ALDD2Xm€	Mitochondria	Mitochondria	Not considered	5
6-phosphofructo-2-kinase	PFK26€	Cytosol	Cytosol	Not considered	5
Fructose-2,6-bisphosphate 2-phosphatase	FBP26€	Cytosol	Cytosol	Not considered	5

Note - ǂ : Reactions common to iAS142 and iSR215 but not iAC560

€ : Reactions common to iAS142 and iAC560 but not iSR215

£ : Reactions having single subcellular locations in iAS142 when compared to other models

¥ : Reactions that have been updated for subcellular locations in iAS142

* : Reactions that have been newly added to iAS142

#: Reactions considered in a separate mitochondrial membrane compartment (b) uniquely within iAS142 model. The mitochondrial membrane is not considered as a separate compartment in iSR215 and iAC560 models.

Appendix B

Table B.2. Comparison of reaction subcellular locations between the iAS556 and iAC560 models

Reaction	Abbreviation	iAC560	iAS556	Confidence Score in iAS556
4-Coumarate:CoA ligase (AMP-forming)	4COUCOAL, 4COUCOALm	cytosol	cytosol	2
		--	mitochondria	2
L-4-hydroxyglutamate semialdehyde dehydrogenase	4HGLSD	mitochondria	cytosol	2
aldose 1-epimerase	A1E	cytosol	cytosol	2
		glycosome	--	
acetyl-CoA carboxylase	ACCOAC	mitochondria	cytosol	2
aconitate hydratase	ACONTc, ACONTm	--	cytosol	4
		mitochondria	mitochondria	5
acetyl-CoA synthetase	ACS, ACS2	mitochondria	cytosol	5
		mitochondria	cytosol	2
adenosine kinase	ADNK1c	cytosol	cytosol	5
		endoplasmic reticulum	--	
		mitochondria	--	
		glycosome	--	
1-acylglycerol-3-phosphate O-acyltransferase	AGPATr	mitochondria	endoplasmic reticulum	2
acetyl-CoA:1-alkyl-sn-glycero-3-phosphate 2-O-acetyltransferase	AKG3PAT_LI	cytosol	glycosome	2
aldehyde reductase	ALDR	cytosol	mitochondria	2
argininosuccinate synthase	ARGSSc,	cytosol	cytosol	2

(continued in the next page)

	ARGSSx	--	glycosome	
alanine aminotransferase	ALATA_Lm	cytosol	--	5
		mitochondria	mitochondria	
carbon-dioxide:ammonia ligase (ADP-forming,carbamate-phosphorylating)	CDAL, CDALg	cytosol	cytosol	5
		--	glycosome	5
choline/ethanolamine phosphotransferase	CEPT1, CEPT2	cytosol	endoplasmic reticulum	1
		cytosol	endoplasmic reticulum	1
citrate Synthase/lyase	CSc, CSm	--	cytosol	5
		mitochondria	mitochondria	5
cytidylate kinase (dCMP)	CYTK2n	nucleus	cytosol	1
dephospho-CoA kinase	DPCK	cytosol	mitochondria	2
diphosphomevalonate decarboxylase	DPMVD	glycosome	--	5
		cytosol	cytosol	
deoxyuridine triphosphatase	DUTPDPn	mitochondria	--	2
		nucleus	nucleus	
cis-2-Methyl-5-isopropylhexa-2,5-dienoyl-CoA hydro-lyase	ECOAH13m	glycosome	--	2
		mitochondria	mitochondria	
trans-2-Methyl-5-isopropylhexa-2,5-dienoyl-CoA hydro-lyase	ECOAH14m	glycosome	--	2
		mitochondria	mitochondria	
ferulate:CoA ligase (AMP-forming)	FERCOAL, FERCOALm	cytosol	cytosol	2
		--	mitochondria	2

(continued in the next page)

fumarase	FUMc, FUMg, FUMm	--	cytosol	5
		glycosome	glycosome	3
		mitochondria	mitochondria	4
fatty-acid--CoA ligase	FACOAL140m, FACOAL160m, FACOAL180m, FACOAL182m, FACOAL1m, FACOAL2m	cytosol	mitochondria	2
L-glutamate 5-semialdehyde dehydratase (spontaneous)	G5SADs	cytosol	cytosol	2
		mitochondria	--	
glucose 6-phosphate dehydrogenase	G6PDHg	cytosol	--	5
		glycosome	glycosome	
glycine cleavage complex	GCCbm	cytosol	--	4
		mitochondria	mitochondria	
glycine hydroxymethyltransferase	GHMT, GHMTm	cytosol	cytosol	5
		--	mitochondria	5
alcohol dehydrogenase	GLCDHm	cytosol	mitochondria	1
glutamine synthetase	GLNSm	cytosol	mitochondria	2
GMP synthase (glutamine-hydrolysing)	GMPS2, GMPSg	cytosol	cytosol	4
		--	glycosome	4
hydroxymethylglutaryl-CoA reductase	HMGCOArm	glycosome	--	5

(continued in the next page)

		mitochondria	mitochondria	
N-acetylglucosamine-6-phosphate deacetylase	GNAD	cytosol	glycosome	5
Guanine phosphoribosyltransferase	GPRTg	cytosol	glycosome	5
L-glutamate 5-semialdehyde dehydrogenase	GSADH	mitochondria	cytosol	2
ubiquinone biosynthesis monooxygenase Coq7	H3MS2	mitochondria	cytosol	2
4-Hydroxy-2-ketopimelate aldolase	HKA	mitochondria	cytosol	2
3-hydroxy-1-pyrroline-5-carboxylate dehydrogenase	HP5CD	mitochondria	cytosol	2
L-hydroxyproline reductase (NAD)	HPROa	cytosol	cytosol	2
		mitochondria	--	
L-hydroxyproline dehydrogenase (NAD)	HPROx, HPROy	mitochondria	cytosol	2
		mitochondria	cytosol	2
isocitrate dehydrogenase (NADP+)	ICDHyc, ICDHym	--	cytosol	5
		mitochondria	mitochondria	5
isopentenyl-diphosphate D-isomerase	IPDDIg	cytosol	--	2
		glycosome	glycosome	
di-trans,poly-cis-Decaprenyl-diphosphate:isopentenyl-diphosphate undecaprenylcistransferase	IPDPUPT	cytosol	mitochondria	2
inositol-1,3,4,5-trisphosphate 5-phosphatase	IPP5P2m, IPP5Pm	cytosol	mitochondria	2
		cytosol	mitochondria	2
D-lactate dehydrogenase	LDH_Dg	cytosol	--	
		glycosome	glycosome	2

(continued in the next page)

methylcrotonoyl-CoA carboxylase	MCC, MCCcyt	mitochondria	mitochondria	2
		--	cytosol	2
malic enzyme (NADP)	ME1m	cytosol	mitochondria	2
methylglutaconyl-CoA hydratase	MGCOAH2	cytosol	mitochondria	2
pyrroline-5-carboxylate reductase	P5CRr	cytosol	cytosol	2
		mitochondria	--	
methylene-fatty-acyl-phospholipid synthase	PEM1	cytosol	endoplasmic reticulum	3
phosphatidyl-N-methylethanolamine N-methyltransferase	PEM2,	cytosol	endoplasmic reticulum	1
	PEM2r	cytosol	endoplasmic reticulum	1
glucose-6-phosphate isomerase	PGI1, PGI2, PGIg1,PGIg2	cytosol	cytosol	4
		cytosol	cytosol	4
		--	glycosome	5
		--	glycosome	5
L-1-Pyrroline-3-hydroxy-5-carboxylate (spontaneous conversion to L-4-hydroxyglutamate semialdehyde)	PHCHGS	cytosol	cytosol	2
		mitochondria	--	
phenylalanine transaminase	PHETA1	cytosol	cytosol	5
		mitochondria	--	
1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase	PI45BPP_LI, PI45BPP_LI_M	cytosol	cytosol	2
		--	mitochondria	2
phosphatidylinositol 4-phosphate 5-kinase	PI4P5K_LI,	cytosol	cytosol	2

(continued in the next page)

	PI4P5K_LI_M	--	mitochondria	2
1-phosphatidylinositol 3-kinase	PIN3K_LI, PIN3K_LI_M	cytosol	cytosol	1
		--	mitochondria	2
phosphatidylinositol synthase	PINOS_LI_M	cytosol	mitochondria	1
phosphomevalonate kinase	PMEVK	cytosol	cytosol	5
		glycosome	--	
phosphopantothenoylcysteine decarboxylase(hypothetical protein)	PPCD	cytosol	mitochondria	1
phosphopantothenate--cysteine ligase	PPCL	cytosol	mitochondria	2
propionyl-CoA carboxylase	PPCOAC, PPCOACm	--	cytosol	2
		mitochondria	mitochondria	2
phosphatidylserine decarboxylase	PPSD	cytosol	mitochondria	5
phosphatidylserine synthase	PPSS	cytosol	cytosol	
		mitochondria	mitochondria	2
phosphomannomutase	PMANM	cytosol	cytosol	5
		glycosome	--	
ribonucleoside-diphosphate reductase	RNDR1(c), RNDR1(n), RNDR2(c), RNDR2(n), RNDR3(c), RNDR3(n), RNDR4(c),	--	cytosol	5
		nucleus	nucleus	5
		--	cytosol	5
		nucleus	nucleus	5
		--	cytosol	5
		nucleus	nucleus	5

(continued in the next page)

	RNDR4(n)	--	cytosol	5
		nucleus	nucleus	5
sinapate:CoA ligase (AMP-forming)	SINCOAL, SINCOALm	--	cytosol	2
		mitochondria	mitochondria	2
squalene synthase	SQSm, SQSr	glycosome	mitochondria	4
		cytosol	endoplasmic reticulum	4
3,4-Dihydroxy-trans-cinnamate:CoA ligase (AMP-forming)	TCAFCOAL, TCAFCOALm	cytosol	cytosol	2
		--	mitochondria	2
trans-Cinnamate:CoA ligase (AMP-forming)	TCINCOAL, TCINCOALm	cytosol	cytosol	2
		--	mitochondria	2
tetrahydrofolate:L-glutamate gamma-ligase (ADP-forming)	THFGLUS	cytosol	mitochondria	2
uridine nucleosidase	URIRHn	nucleus	cytosol	5
trypanothione reductase	TRYR, TRYRg, TRYRn,	--	cytosol	5
		glycosome	glycosome	5
		--	nucleus	5
undecaprenyl-diphosphate synthase	UDPDPS	cytosol	mitochondria	2

Appendix B

Table B.3. Comparison of predicted reaction knockout phenotypes with experimentally determined phenotypes

Reaction (Gene) Name	Abbreviation	iAS556	Expt.	Organism	Reference	Stages	Predicted wild type growth (%)	
							P	A
aconitase	ACONTm	NL	NL	*L. mexicana*	(Saunders et al. 2014)	Both	100%	23%
adenosine deaminase	ADA, ADAer, ADAg, ADAm	NL	NL	*L. donovani*	(Boitz et al. 2012)	Both	100%	100%
adenosine kinase	ADNK1c	NL	NL	*L. donovani*	(Boitz et al. 2012)	Both	100%	100%
1-(5-Phosphoribosyl)-5-amino-4-imidazolecarboxamide:pyrophosphate phosphoribosyltransferase	ADPT2	NL	NL	*L. donovani*	(Boitz and Ullman 2006; Carter et al. 2008)	Both	100%	100%
adenine phosphoribosyltransferase	ADPTr	L	NL	*L. donovani*	(Boitz and Ullman 2006; Carter et al. 2008)	Both	0%	0%
arginase	ARGNg	NL	NLa	*L. donovani*	(Boitz et al. 2016)	A	100%	100%
argininosuccinate synthase	ARGSSx	L	La	*L. donovani*	(Lakhal-Naouar et al. 2012)	A	0%	0%

(continued in the next page)

Appendix B

L-Arginine transport	ARGt	L	L	*L. donovani*	(Shaked-Mishan et al. 2006)	Both	0%	0%
ATP synthase	ATPSmm	L	L	*L. donovani and L. pifanoi*	(Luque-Ortega et al. 2008)	Both	0%	0%
cytochrome-c-oxidase	CYOO6mm	L	L	*L. donovani and L. pifanoi*	(Luque-Ortega et al. 2008)	Both	0%	0%
methylenetetrahydrofolate dehydrogenase	DH1	NL	L	*L. major*	(Murta et al. 2009)	Both	100%	100%
dihydroxyacetonephosphate acyltransferase	DHAPAx_LI	NL	La	*L. major*	(McConville et al. 2007)	A	100%	100%
dihydrofolate reductase-thymidine synthase	DHFRc, DHFRm, TS, TSm	NL	NL	*L. major*	(Nare et al. 1997)	Both	100%	100%
fructose-1,6-bisphosphatase	FBPg	NL	NL	*L. major*	(Naderer et al. 2006)	Both	100%	100%
glucose transporter	GLCBt2	L	La	*L. mexicana*	(McConville et al. 2007)	A	0%	0%

(continued in the next page)

Appendix B

glutamylcysteine ligase	GLUCYSL	L	L	*L. infantum*	(Mukherjee et al. 2009)	Both	0%	0%
N-acetylglucosamine-6-phosphate deacetylase	GNAD	NL	NL	*L. major*	(Naderer et al. 2010)	Both	100%	100%
glucose-6-phosphate N-acetyltransferase	GPAT	NL	NL	*L. major*	(Naderer et al. 2015)	Both	100%	100%
guanine phosphoribosyltransferase	GPRTg	NL	NL	*L. donovani*	(Boitz and Ullman 2006; Carter et al. 2008)	Both	100%	100%
glutathionylspermidine synthase	GSS	NL	NL	*L. infantum*	(Sousa et al. 2014)	Both	100%	100%
hypoxanthine phosphoribosyltransferase (hypoxanthine)	HXPRTg	NL	NL	*L. donovani*	(Boitz and Ullman 2006; Carter et al. 2008)	Both	100%	100%
myo-inositol-3-phosphate synthase	INO1	NL	La	*L. mexicana*	(McConville et al. 2007)	A	100%	100%
trypanothione-dependent glyoxalase I	LGTHL1, LGTHL1m	NL	NL	*L. donovani*	(Chauhan and Madhubala 2009)	Both	100%	100%
LIT iron transporter	LIT	L	L	*L. amazonensis*	(Jacques et al. 2010)	Both	0%	0%

(continued in the next page)

mannose-1-phosphate guanylyltransferase	MAN1PT1	L	La	*L. mexicana*	(Garami and Ilg 2001b)	A	0%	0%
mannose-6-phosphate isomerase	MAN6PI	L	La	*L. mexicana*	(Garami et al. 2001; Garami and Ilg 2001c)	A	0%	0%
myo-Inositol-1-phosphate synthase	MI1PSB	NL	NL	*L. mexicana*	(Ilg 2002)	Both	100%	100%
5,10-methylenetetrahydrofolate reductase (NADP)	MTHFR1	NL	NL	*L. major*	(Vickers et al. 2006)	Both	100%	100%
NAD-dependent SIR2	NDSIR2	NL	La	*L. infantum*	(Gazanion et al. 2011)	A	100%	100%
ornithine decarboxylase	ORNDC	NL	La	*L. donovani*	(Gilroy et al. 2011)	Both	100%	100%
phenylalanine-4-monooxygenase	PHE4MOi	NL	NL	*L. major*	(Lye et al. 2011)	Both	100%	100%
phosphomannomutase	PMANM	L	La	*L. mexicana*	(Garami et al. 2001)	A	0%	0%
pteridine reductase 1	PTR1	NL	NL	*L. major*	(Nare et al. 1997)	Both	100%	100%
serine-C-palmitoyltransferase	SERPTr	NL	NL	*L. major*	(Zhang et al. 2003)	Both	100%	100%
serine hydroxymethyltranferase (SHMT-S)	SHMTc2	NL	NL	*L. major*	(Roy and Ouellette 2015)	Both	100%	100%
spermidine synthase	SPMS	L	L	*L. donovani*	(Gilroy et al. 2011)	Both	0%	0%

(continued in the next page)

Appendix B

squalene epoxidase	SQLMer	L	L	*L. amazonensis*	(Vannier-Santos et al. 1995; Chawla and Madhubala 2010)	Both	0%	0%
squalene synthase	SQSr, SQSm	L	L	*L. chagasi*	(Granthon et al. 2007)	Both	0%	0%
sterol 14-demethylase	ST14DMr	L	L	*L.donovani*	(McCall et al. 2015)	Both	0%	0%
trypanothione reductase	TRYR	NL	La	*L. donovani*	(Tovar et al. 1998)	A	100%	100%
trypanothione synthetase	TRYS	L	L	*L. infantum*	(Sousa et al. 2014)	Both	0%	0%
UDP sugar pyrophosphorylase	USPx	NL	NL	*L. major*	(Lamerz et al. 2010)	Both	100%	100%
xanthine phosphoribosyltransferase	XPRTgr	NL	NL	*L. donovani*	(Boitz and Ullman 2006; Carter et al. 2008)	Both	100%	100%

Note - Both: Promastigote & Amastigote; P: Promastigote; A: Amastigote

L – Lethal, NL – Non-lethal, La- Lethal in Amastigote, NLa – Non-lethal in Amastigote,

% Predicted wild type growth – predicted percentage of maximum growth after a single gene deletion

Appendix B

Table B.4. Contribution of the eight predictors to the selected principal components (loading cut-off > 0.45) and hence, the d_N, d_S & ω rates in *L. major*, *L. donovani* and *L. infantum*.

PC	$\log_{10}(d_N)$			$\log_{10}(d_S)$			$\log_{10}(\omega)$		
	L. major	***L. donovani***	***L. infantum***	***L. major***	***L. donovani***	***L. infantum***	***L. major***	***L. donovani***	***L. infantum***
1	NCoup (+), CCoFCA (+)	NCoup (+), CCoFCA (+)	NCoup (-), CCoFCA (-)	CAI (+), GC (+)	NCoup (+), CCoFCA (+)	GC (+), GeneLength(+)	NCoup (-), CCoFCA (-)	NCoup (+), CCoFCA (+)	NCoup (+), CCoFCA (+)
	0.0059	0.0146	-0.046***	0.02***	0.0013	0.0094*	-0.015124	0.015192	0.050955***
2	CAI (-), GC (-)	CAI (-), GC(-), GeneLength (-)	CAI (+), GC (+)	NCoup (-), CCoFCA (-)	RPKM (-)	NCoup (-), CCoFCA (-)	CAI (-), GC(-)	CAI (-), GC(-), GeneLength(-)	GC (+)
	0.0641***	0.065***	-0.017	-0.0097*	-0.0185***	-0.0053	0.080628***	0.090106***	-0.020855*
3	GeneLength(-), RPKM (+)	GeneLength(-), NumProcs (+)	CAI (-), GeneLength(+), RPKM (-)	--	CAI (+), GC (+), GeneLength(+)	CAI (+), RPKM (+)	RPKM (+)	GeneLength(-), NumProcs (+)	CAI (-), RPKM (-)
	-0.088***	-0.073***	0.065***		0.0223***	0.029***	-0.119743***	-0.086660***	0.091368***
4	NumFuncs (+), NumProcs (+)	--	NumFuncs (-), NumProcs (-)	--	--	--	NumFuncs (+), NumProcs (-)	NumProcs (+), RPKM (-)	NumFuncs (-), NumProcs (-)

(continued in the next page)

	-0.0281*		0.04**				0.005839	0.020986	0.048809***
5	GeneLength(+), NumProcs (-)	--	NumFuncs (+), NumProcs (-), RPKM (+)	--	--	--	GeneLength (-), NumProcs (+)	NumFuncs (-), NumProcs (+)	NumFuncs (-), NumProcs (+)
	-0.0244		-0.042**				0.006488	0.044021**	0.049803***
6	--	--	GeneLength(-), NumFuncs (+)	--	--	--	GeneLength (+), NumFuncs (-)	CAI (-), RPKM (+)	GeneLength (+), RPKM (+)
			-0.025				0.030698*	0.073631***	0.035479*
7	--	--	CAI (-), GC (+)	--	--	--	CAI (+), GC (-)	--	CAI (-), GC (+)
			0.156***				-0.170591***		0.204003***
8	--	--	--	--	--	--	--	--	--

Note: The positive and negative signs in brackets indicate the nature of their contributions to the principal component as demonstrated by the principal component loadings. The numbers below each combination of features indicates the regression coefficients associated with that principal component. The regression coefficients corresponding to each principal component were obtained after regressing the chosen principal components to the response evolutionary rates. Genes with positive or negative scores with respect to a principal component correspond to the positive or negative contribution of features of those genes as indicated by the loadings on that component. The dashes (--) in the table indicate that the corresponding principal component was not selected for regression, as identified by the randomization test approach. P-values of regression coefficients: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$.

Appendix B

Table B.5. eggNOG functional categories of the species-specific metabolic genes

	eggNOG functional categories
C	Energy production and conversion
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
G	Carbohydrate transport and metabolism
H	Coenzyme transport and metabolism
I	Lipid transport and metabolism
M	Cell wall/membrane/envelope biogenesis
S	Function unknown

Appendix B

Table B.6. The species-specific set of genes (d_N rates) explained by codon adaptation index, multi-functionality and number of flux-couplings associated with an enzyme (gene)

Gene	Protein name	GO Process
L. major		
LmjF.23.0360	NADP-dependent alcohol dehydrogenase, putative	C
LmjF.24.1630	Succinate dehydrogenase [ubiquinone] flavoprotein subunit, mitochondrial	C
LmjF.35.5330	isopentenyl-diphosphate delta-isomerase (type II), putative	C
LmjF.36.5910	2,3-diketo-5-methylthio-1-phosphopentane phosphatase, putative	C
LmjF.13.1620	squalene monooxygenase-like protein	C, H
LmjF.27.1870	trypanothione synthetase	E
LmjF.22.0110	GMP synthase [glutamine-hydrolysing]	F
LmjF.33.1090	guanylate kinase, putative	F
LmjF.27.0420	ribokinase, putative	G
LmjF.32.1580	phosphomannose isomerase	G
LmjF.34.0080	glucose-6-phosphate 1-dehydrogenase, putative	G
LmjF.14.0910	glutathione synthetase, putative	I
LmjF.14.1200	phosphatidylserine synthase, putative	I
LmjF.23.0110	GDP-mannose pyrophosphorylase	M
LmjF.20.0970	1,2-Dihydroxy-3-keto-5-methylthiopentene dioxygenase, putative	S
L. donovani		
LdBPK_070210	cytochrome c1, heme protein, mitochondrial, putative	C
LdBPK_100310	isocitrate dehydrogenase [NADP], mitochondrial precursor, putative	C
LdBPK_351540	rieske iron-sulfur protein, mitochondrial precursor, putative	C
LdBPK_070240	cobalamin-dependent methionine synthase, putative	E
LdBPK_260790	asparagine synthetase a, putative	E
LdBPK_330530	d-xylulose reductase, putative	E
LdBPK_362490	tyrosine aminotransferase	E
LdBPK_160590	carbamoyl-phosphate synthase, putative	F
LdBPK_211450	thymidine kinase, putative	F
LdBPK_280980	ribonucleoside-diphosphate reductase large chain, putative	F
LdBPK_313080	acetyl-CoA carboxylase	I
L. infantum		
LinJ.05.0180	dihydrolipoamide branched chain transacylase, putative	C
LinJ.10.0560	glycerol-3-phosphate dehydrogenase [NAD+], glycosomal	C
LinJ.28.0240	glycerol-3-phosphate dehydrogenase (FAD-dependent), mitochondrial	C
LinJ.29.1950	dihydrolipoamide dehydrogenase, putative	C
LinJ.13.1420	pyrroline-5-carboxylate reductase	E
LinJ.26.0040	pyridoxal phosphate containing glycine decarboxylase, putative	E
LinJ.06.0890	dihydrofolate reductase-thymidylate synthase	F

Appendix B

Table B.7. The species-specific set of genes (d_S rates) explained by codon adaptation index, multi-functionality and number of flux-couplings associated with an enzyme (gene)

Gene	Protein name	GO Process
L. major		
LmjF.20.0560	cytidine triphosphate synthase, putative	F
LmjF.24.1630	Succinate dehydrogenase [ubiquinone] flavoprotein subunit, mitochondrial	C
LmjF.33.2720	3-oxoacyl-acyl carrier protein synthase II, putative	I
LmjF.24.0850	triosephosphate isomerase	G
LmjF.36.3100	ATP synthase, putative	C
LmjF.34.0080	glucose-6-phosphate 1-dehydrogenase, putative	G
LmjF.36.1260	fructose-1,6-bisphosphate aldolase	G
LmjF.22.0110	GMP synthase [glutamine-hydrolysing]	F
LmjF.27.0420	ribokinase, putative	G
LmjF.32.1580	phosphomannose isomerase	G
LmjF.05.1140	V-type proton ATPase subunit D, putative	C
LmjF.32.3140	pyrroline-5-carboxylate synthetase-like protein	E
LmjF.06.0650	lanosterol synthase, putative	I
LmjF.14.0910	glutathione synthetase, putative	I
LmjF.35.1480	arginase	E
LmjF.21.1430	2-oxoisovalerate dehydrogenase alpha subunit, putative	C
LmjF.30.0880	adenosine kinase, putative	G
LmjF.14.1200	phosphatidylserine synthase, putative	I
LmjF.36.2360	tyrosine aminotransferase	E
LmjF.35.2740	galactokinase-like protein	G
LmjF.13.1620	squalene monooxygenase-like protein	C, H
LmjF.35.3870	nucleoside diphosphate kinase, putative	F
LmjF.25.2010	2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase, putative	G
LmjF.07.0730	malonyl-coa decarboxylase-like protein	G
LmjF.27.1870	trypanothione synthetase	E
L. donovani		
LdBPK_280140	pantothenate kinase subunit, putative	H
LdBPK_044320	phosphatidylinositol phosphate kinase alpha	T
LdBPK_221380	dephospho-CoA kinase, putative	H
LdBPK_060370	glutamine synthetase, putative	E
LdBPK_271330	ethanolamine kinase, putative	M
LdBPK_323830	enoyl-CoA hydratase/isomerase family protein, putative	I
LdBPK_070240	cobalamin-dependent methionine synthase, putative	E
LdBPK_341170	dihydroxyacetone phosphate acyltransferase, putative	I
LdBPK_351480	choline/ethanolamine kinase, putative	T
LdBPK_130300	long-chain-fatty-acid-CoA ligase, putative	I

(continued in the next page)

LdBPK_060580	deoxyuridine triphosphatase, putative	F
LdBPK_081040	phosphoribosylpyrophosphate synthetase	F
LdBPK_010550	long chain fatty acid CoA ligase, putative	I
L. infantum		
LinJ.03.0190	delta-1-pyrroline-5-carboxylate dehydrogenase, putative	C
LinJ.06.0890	dihydrofolate reductase-thymidylate synthase	F
LinJ.06.1330	coproporphyrinogen III oxidase	H
LinJ.12.0200	glyoxalase II	S
LinJ.17.1510	myo-inositol-1(or 4)-monophosphatase 1, putative	G
LinJ.29.2910	inosine-adenosine-guanosine-nucleosidehydrolase, putative	F
LinJ.31.1150	monoglyceride lipase, putative	I
LinJ.31.2360	Phosphatidylethanolamine n-methyltransferase-like protein	I
LinJ.31.3250	Phosphatidylethanolamine n-methyltransferase-like protein	I
LinJ.33.1660	ribulose-5-phosphate 3-epimerase, putative	G
LinJ.34.0130	adenylate kinase 2	F
LinJ.36.6960	2,3-bisphosphoglycerate-independent phosphoglycerate mutase	G

References

Abu-Dayyeh I, Shio MT, Sato S, Akira S, Cousineau B, Olivier M. 2008. *Leishmania*-induced IRAK-1 inactivation is mediated by SHP-1 interacting with an evolutionarily conserved KTIM motif. PLoS Negl. Trop. Dis. 2:e305.

Allen DK, Shachar-Hill Y, Ohlrogge JB. 2007. Compartment-specific labeling information in 13C metabolic flux analysis of plants. Phytochemistry 68:2197–2210.

Alonso G, Guevara P, Ramirez JL. 1992. Trypanosomatidae codon usage and GC distribution. Mem{ó}rias do Inst. Oswaldo Cruz 87:517–523.

Alvarez-Ponce D, Fares MA. 2012. Evolutionary rate and duplicability in the *Arabidopsis thaliana* protein-protein interaction network. Genome Biol. Evol. 4:1263–1274.

Alvarez-Ponce D, Feyertag F, Chakraborty S. 2017. Position matters: Network centrality considerably impacts rates of protein evolution in the human protein–protein interaction Network. Genome Biol. Evol. 9:1742–1756.

Alvarez F, Robello C, Vignali M. 1994. Evolution of codon usage and base contents in kinetoplastid protozoans. Mol. Biol. Evol. 11:790–802.

Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X, et al. 2010. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. 38:D457-D462.

Avila-levy CM. 2014. Proteins and Proteomics of *Leishmania* and Trypanosoma. (Santos ALS, Branquinha MH, d'Avila-Levy CM, Kneipp LF, Sodré CL, editors.), 2013. Springer Science & Business Medi, 74

Bairoch A. 2000. The ENZYME database in 2000. Nucleic Acids Res. 28:304–305.

Barrett MP, Coombs GH, Mottram JC. 1999. Recent advances in identifying and validating drug targets in trypanosomes and leishmanias. Trends Microbiol. 7:82–88.

Bastian M, Heymann S, Jacomy M. 2009. International AAAI conference on weblogs and social media. San Jos{é} California, USA.

Behura SK, Severson DW. 2012. Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. PLoS One 7:e43111.

Behura SK, Severson DW. 2013. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. Biol. Rev. 88:49–61.

Bello AM, Poduch E, Fujihashi M, Amani M, Li Y, Crandall I, Hui R, Lee PI, Kain KC, Pai EF, et al. 2007. A potent, covalent inhibitor of orotidine 5'-monophosphate decarboxylase with antimalarial activity. J. Med. Chem. 50:915–921.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2008. GenBank. Nucleic Acids Res. 36:D25-D30.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. Trends Genet. 21:219–226.

Biswas S, Subramanian A, ELMojtaba IM, Chattopadhyay J, Sarkar RR. 2017. Optimal combinations of control strategies and cost-effective analysis for visceral leishmaniasis disease transmission. PLoS One 12:e0172465.

Boitz JM, Gilroy CA, Olenyik TD, Paradis D, Perdeh J, Dearman K, Davis MJ, Yates PA, Li Y, Riscoe MK, et al. 2016. Arginase is essential for Survival of *Leishmania donovani* promastigotes but not intracellular amastigotes. Infect. Immun.:IAI - 00554.

Boitz JM, Strasser R, Hartman CU, Jardim A, Ullman B. 2012. Adenine aminohydrolase from *Leishmania donovani* unique enzyme in parasite purine metabolism. J. Biol. Chem. 287:7626–7639.

Boitz JM, Ullman B. 2006. *Leishmania donovani* singly deficient in HGPRT, APRT or XPRT are viable in vitro and within mammalian macrophages. Mol. Biochem. Parasitol. 148:24–30.

Bordbar A, Monk JM, King Z, Palsson BO. 2014. Constraint-based models predict metabolic and associated cellular functions. Nat. Rev. Genet. 15:107–120.

Borst P. 1986. How proteins get into microbodies (peroxisomes, glyoxysomes, glycosomes). Biochim. Biophys. Acta (BBA)-Gene Struct. Expr. 866:179–203.

Bringaud F, Müller M, Cerqueira GC, Smith M, Rochette A, El-Sayed NM a, Papadopoulou B, Ghedin E. 2007. Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*. PLoS Pathog. 3:1291–1307.

Bringaud F, Rivière L, Coustou V. 2006. Energy metabolism of trypanosomatids: adaptation to available carbon sources. Mol. Biochem. Parasitol. 149:1–9.

Brotherton MC, Racine G, Foucher AL, Drummelsmith J, Papadopoulou B, Ouellette M. 2010. Analysis of stage-specific expression of basic proteins in *Leishmania infantum*. J. Proteome Res. 9:3842–3853.

Burgard AP, Nikolaev E V, Schilling CH, Maranas CD. 2004. Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Res. 14:301–312.

Carter NS, Yates P, Arendt CS, Boitz JM, Ullman B. 2008. Purine and pyrimidine metabolism in *Leishmania*. In: Drug targets in kinetoplastid parasites. Springer. p. 141–154.

De Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 34:W362-W365.

Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Structural approaches to sequence evolution. Springer. p. 207–232.

Chauhan N, Vidyarthi AS, Poddar R. 2011. Comparative multivariate analysis of codon and amino acid usage in three *Leishmania* genomes. Gen. Prot. Bioinform. 9:218–228.

Chauhan SC, Madhubala R. 2009. Glyoxalase I gene deletion mutants of *Leishmania donovani* exhibit reduced methylglyoxal detoxification. PLoS One 4:e6805.

Chavali AK, Blazier AS, Tlaxca JL, Jensen PA, Pearson RD, Papin JA. 2012. Metabolic network analysis predicts efficacy of FDA-approved drugs targeting the causative agent of a neglected tropical disease. BMC Syst. Biol. 6:27.

Chavali AK, Whittemore JD, Eddy JA, Williams KT, Papin JA. 2008. Systems analysis of metabolism in the pathogenic trypanosomatid *Leishmania major*. Mol. Syst. Biol. 4:177.

Chawla B, Madhubala R. 2010. Drug targets in *Leishmania*. J. Parasit. Dis. 34:1–13.

Chesmore KN, Bartlett J, Cheng C, Williams SM. 2016. Complex patterns of association between pleiotropy and transcription factor evolution. Genome Biol. Evol. 8:3159–3170.

Chiapello H, Lisacek F, Caboche M, Hénaut A. 1998. Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. Gene 209:GC1-GC38.

Christopher DM, Prabhakar R, Hinrich S. 2008. Introduction to information retrieval. An introd. to inf. retr. 151:177.

Chu S, Wang J, Cheng H, Yang Q, Yu D. 2014. Evolutionary study of the isoflavonoid pathway based on multiple copies analysis in soybean. BMC genetics, 15:76.

Cloutier S, Laverdière M, Chou M-N, Boilard N, Chow C, Papadopoulou B. 2012. Translational control through eIF2alpha phosphorylation during the *Leishmania* differentiation process. PLoS One 7:e35085.

Cohen-Freue G, Holzer TR, Forney JD, McMaster WR. 2007. Global gene expression in *Leishmania*. Int. J. Parasitol. 37:1077–1086.

Colasante C, Voncken F, Manful T, Ruppert T, Tielens AGM, van Hellemond JJ, Clayton C. 2013. Proteins and lipids of glycosomal membranes from *Leishmania tarentolae* and *Trypanosoma brucei*. F1000Research 2.

Colombo M, Laayouni H, Invergo BM, Bertranpetit J, Montanucci L. 2014. Metabolic flux is a detereminant of the evolutionary rates of enzyme-encoding genes. Evolution. 68:605–613.

Coombs GH, Craft JA, Hart DT. 1982. A comparative study of *Leishmania mexicana* amastigotes and promastigotes, enzyme activities and subcellular locations. Mol. Biochem. Parasitol. 5:199–211.

Copley SD. 2012. Toward a systems biology perspective on enzyme evolution. J. Biol. Chem. 287:3–10.

Crick F. 1970. Central dogma of molecular biology. Nature 227:561–563.

Croft SL, Coombs GH. 2003. Leishmaniasis--current chemotherapy and recent advances in the search for novel drugs. Trends Parasitol. 19:502–508.

Croft SL, Sundar S, Fairlamb AH. 2006. Drug resistance in leishmaniasis. Clin. Microbiol. Rev. 19:111–126.

Darling TN, Davis DG, London RE, Blum JJ. 1987. Products of *Leishmania braziliensis* glucose catabolism : Release of D-lactate and under anaerobic conditions, glycerol. 84:7129–7133.

Darling TN, Davis DG, London RE, Blum JJ. 1989. Carbon dioxide abolishes the reverse Pasteur effect in *Leishmania major* promastigotes. Mol. Biochem. Parasitol. 33:191–202.

Das S, Paul S, Dutta C. 2006. Synonymous codon usage in adenoviruses: influence of mutation, selection and protein hydropathy. Virus Res. 117:227–236.

DeLano WL. 2002. The PyMOL user's manual. DeLano Sci. San Carlos, CA 452.

Depledge DP, Evans KJ, Ivens AC, Aziz N, Maroof A, Kaye PM, Smith DF. 2009. Comparative expression profiling of *Leishmania*: Modulation in gene expression between species and in different host genetic backgrounds. PLoS Negl. Trop. Dis. 3:e476.

Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton J a, Hilley JD, de Doncker S, Maes I, Mottram JC, et al. 2011. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. Genome Res. 21:2143–2156.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. 23:327–337.

Duncan RC, Salotra P, Goyal N, Akopyants NS, Beverley SM, Nakhasi HL. 2004. The application of gene expression microarray technology to kinetoplastid research. Curr. Mol. Med. 4:611–621.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 95:14863–14868.

El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey E a, Hertz-Fowler C, et al. 2005. Comparative genomics of trypanosomatid parasitic protozoa. Science 309:404–409.

Emanuelsson O, Brunak S, Von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. 2:953–971.

Escalante AA, Lal AA, Ayala FJ. 1998. Genetic polymorphism and natural selection in the malaria parasite *Plasmodium falciparum*. Genetics 149:189–202.

Fernandes AP, Nelson K, Beverley SM. 1993. Evolution of nuclear ribosomal RNAs in kinetoplastid protozoa: perspectives on the age and origins of parasitism. Proc. Natl. Acad. Sci. 90:11608–11612.

Fidalgo LM, Gille L. 2011. Mitochondria and trypanosomatids: targets and drugs. Pharm. Res. 28:2758–2770.

Fiebig M, Kelly S, Gluenz E. 2015. Comparative life cycle transcriptomics revises *Leishmania mexicana* genome annotation and links a chromosome duplication with parasitism of Vertebrates. PLoS Pathog. 11:1–28.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2013. Pfam: the protein families database. Nucleic Acids Res. 42:D222-D230.

Flórez AF, Park D, Bhak J, Kim B-C, Kuchinsky A, Morris JH, Espinosa J, Muskus C. 2010. Protein network prediction and topological analysis in *Leishmania major* as a tool for drug target selection. BMC Bioinformatics 11:484.

Fuglsang A. 2003. The effective number of codons for individual amino acids: some codons are more optimal than others. Gene 320:185–190.

Tabachnick BG, Fidell LS. 2007. Using Multivariate Statistics.

Garami A, Ilg T. 2001a. The role of phosphomannose isomerase in *Leishmania mexicana* glycoconjugate synthesis and virulence. J. Biol. Chem. 276:6566–6575.

Garami A, Ilg T. 2001b. Disruption of mannose activation in *Leishmania mexicana*: GDP-mannose pyrophosphorylase is required for virulence, but not for viability. EMBO J. 20:3657–3666.

Garami A, Ilg T. 2001c. The role of phosphomannose isomerase in *Leishmania mexicana* glycoconjugate synthesis and virulence. J. Biol. Chem. 276:6566–6575.

Garami A, Mehlert A, Ilg T. 2001. Glycosylation defects and virulence phenotypes of *Leishmania mexicana* phosphomannomutase and dolicholphosphate-mannose synthase gene deletion mutants. Mol. Cell. Biol. 21:8168–8183.

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A. 2005. Protein identification and analysis tools on the ExPASy server. Springer

Gazanion E, Garcia D, Silvestre R, Gerard C, Guichou JF, Labesse G, Seveno M, Cordeiro-Da-Silva A, Ouaissi A, Sereno D, et al. 2011. The *Leishmania* nicotinamidase is essential for NAD+ production and parasite proliferation. Mol. Microbiol. 82:21–38.

Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. Nature 425:737–741.

Ghedin E, Bringaud F, Peterson J, Myler P, Berriman M, Ivens A, Andersson B, Bontempi E, Eisen J, Angiuoli S, et al. 2004. Gene synteny and evolution of genome architecture in trypanosomatids. Mol. Biochem. Parasitol. 134:183–191.

Gilroy C, Olenyik T, Roberts SC, Ullman B. 2011. Spermidine synthase is required for virulence of *Leishmania donovani*. Infect. Immun. 79:2764–2769.

Ginger ML, McFadden GI, Michels PAM. 2010. Rewiring and regulation of cross-compartmentalized metabolism in protists. Philos. Trans. R. Soc. B Biol. Sci. 365:831–845.

Gladki A, Kaczanowski S, Szczesny P, Zielenkiewicz P. 2013. The evolutionary rate of antibacterial drug targets. BMC Bioinformatics 14.

Go Y-M, Jones DP. 2008. Redox compartmentalization in eukaryotic cells. Biochim. Biophys. Acta (BBA)-General Subj. 1780:1273–1290.

Graille M, Baltaze J-P, Leulliot N, Liger D, Quevillon-Cheruel S, van Tilbeurgh H. 2006. Structure-based functional fnnotation yeast ymr099c codes for a d-hexose-6-phosphate mutarotase. J. Biol. Chem. 281:30175–30185.

Granthon AC, Braga M V, Rodrigues JCF, Cammerer S, Lorente SO, Gilbert IH, Urbina JA, de Souza W. 2007. Alterations on the growth and ultrastructure of *Leishmania chagasi* induced by squalene synthase inhibitors. Vet. Parasitol. 146:25–34.

Gu W, Zhou T, Wilke CO. 2010. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput. Biol. 6:e1000664.

Gualdron-López M, Vapola MH, Miinalainen IJ, Hiltunen JK, Michels PAM, Antonenkov VD. 2012. Channel-forming activities in the glycosomal fraction from the bloodstream form of *Trypanosoma brucei*. PLoS One 7:e34530.

Guerra-Giraldez C, Quijada L, Clayton CE. 2002. Compartmentation of enzymes in a microbody, the glycosome, is essential in *Trypanosoma brucei*. J. Cell Sci. 115:2651–2658.

Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. Trends Biotechnol. 22:346–353.

Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R. 2000. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc. Natl. Acad. Sci. 97:9390–9395.

Haile S, Papadopoulou B. 2007. Developmental regulation of gene expression in trypanosomatid parasitic protozoa. Curr. Opin. Microbiol. 10:569–577.

Harkins KM, Schwartz RS, Cartwright RA, Stone AC. 2016. Phylogenomic reconstruction supports supercontinent origins for *Leishmania*. Infect. Genet. Evol. 38:101–109.

Hart DT, Coombs GH. 1982. *Leishmania mexicana*: Energy metabolism of amastigotes and promastigotes. Exp. Parasitol. 54:397–409.

Hase T, Niimura Y, Tanaka H. 2010. Difference in gene duplicability may explain the difference in overall structure of protein-protein interaction networks among eukaryotes. BMC Evol. Biol. 10:358.

Hershberg R, Petrov DA. 2008. Selection on codon bias. Annu. Rev. Genet. 42:287–299.

Holzer TR, McMaster WR, Forney JD. 2006. Expression profiling by whole-genome interspecies microarray hybridization reveals differential gene expression in procyclic promastigotes, lesion-derived amastigotes, and axenic amastigotes in *Leishmania mexicana*. Mol. Biochem. Parasitol. 146:198–218.

Horn D. 2008. Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids. BMC Genomics 9:2.

Ilg T. 2002. Generation of myo-inositol-auxotrophic *Leishmania mexicana* mutants by targeted replacement of the myo-inositol-1-phosphate synthase gene. Mol. Biochem. Parasitol. 120:151–156.

Irwin B, Heck JD, Hatfield GW. 1995. Codon pair utilization biases influence translational elongation step times. J. Biol. Chem. 270:22801–22806.

Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream M-A, Adlem E, Aert R, et al. 2005. The genome of the kinetoplastid parasite, *Leishmania major*. Science 309:436–442.

Jacques I, Andrews NW, Huynh C. 2010. Functional characterization of LIT1, the *Leishmania amazonensis* ferrous iron transporter. Mol. Biochem. Parasitol. 170:28–36.

Jamdhade MD, Pawar H, Chavan S, Sathe G, Umasankar PK, Mahale KN, Dixit T, Madugundu AK, Prasad TSK, Gowda H, et al. 2015. Comprehensive proteomics analysis of glycosomes from *Leishmania donovani*. Omics. J. Integr. Biol. 19:157–170.

Jolliffe IT. 1982. A note on the use of principal components in regression. Appl. Stat.:300–303.

Jovelin R, Phillips PC. 2009. Evolutionary rates and centrality in the yeast gene regulatory network. Genome Biol. 10:R35.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2013. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42:D199-D205.

Karlin S, Mrázek J, Campbell AM. 1998. Codon usages in different gene classes of the Escherichia coli genome. Mol. Microbiol. 29:1341–1355.

Kawaguchi R, Bailey-Serres J. 2005. mRNA sequence features that contribute to translational regulation in *Arabidopsis*. Nucleic Acids Res. 33:955–965.

Keegan F, Blum JJ. 1990. Effects of oxygen concentration on the intermediary metabolism of *Leishmania major* promastigotes. Mol. Biochem. Parasitol. 39:235–245.

Koonin E V, Wolf YI. 2006. Evolutionary systems biology: links between gene evolution and function. Curr. Opin. Biotechnol. 17:481–487.

Koreny L, Obornik M, Lukeš J. 2013. Make it, take it, or leave it: heme metabolism of parasites. PLoS Pathog 9:e1003088.

ter Kuile BH. 1999. Regulation and adaptation of glucose metabolism of the parasitic protist *Leishmania donovani* at the enzyme and mRNA levels. J. Bacteriol. 181:4863–4872.

Kumar R, Engwerda C. 2014. Vaccines to prevent leishmaniasis. Clin. Transl. Immunol. 3:e13.

Lahav T, Sivam D, Volpin H, Ronen M, Tsigankov P, Green A, Holland N, Kuzyk M, Borchers C, Zilberstein D, et al. 2011. Multiple levels of gene regulation mediate differentiation of the intracellular pathogen *Leishmania*. FASEB J. 25:515–525.

Lakhal-Naouar I, Jardim A, Strasser R, Luo S, Kozakai Y, Nakhasi HL, Duncan RC. 2012. *Leishmania donovani* argininosuccinate synthase is an active enzyme associated with parasite pathogenesis. PLOS Negl Trop Dis 6:e1849.

Lamerz A-C, Damerow S, Kleczka B, Wiese M, Van Zandbergen G, Lamerz J, Wenzel A, Hsu F-F, Turk J, Beverley SM, et al. 2010. Deletion of UDP-glucose pyrophosphorylase reveals a UDP-glucose independent UDP-galactose salvage pathway in *Leishmania major*. Glycobiology 20:872–882.

Larhlimi A, David L, Selbig J, Bockmayr A. 2012. F2C2: a fast tool for the computation of flux coupling in genome-scale metabolic networks. BMC Bioinformatics 13:57.

Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: Detection of (Co-) orthologs in large-scale analysis. BMC Bioinformatics 12:124.

Leifso K, Cohen-Freue G, Dogra N, Murray A, McMaster WR. 2007. Genomic and proteomic expression analysis of *Leishmania* promastigote and amastigote life stages: The *Leishmania* genome is constitutively expressed. Mol. Biochem. Parasitol. 152:35–46.

Liang X, Haritan A, Uliel S. 2003. trans and cis Splicing in Trypanosomatids : Mechanism , Factors , and Regulation. Eukaryot. Cell 2.

Liu W, Lin W, Davis AJ, Jordán F, Hwang M. 2007. A network perspective on the topological importance of enzymes and their phylogenetic conservation. BMC Bioinformatics 8:121.

Lorenz R, Bernhart SHF, Zu Siederdissen CH, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. Algorithms Mol. Biol. 6:26.

Louassini M, Foulquié MR, Ben'\itez R, Adroher FJ. 1999. Activity of key enzymes in glucose catabolism during the growth and metacyclogenesis of *Leishmania infantum*. Parasitol. Res. 85:300–306.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 320:1632–1635.

Lu C, Zhang Z, Leach L, Kearsey MJ, Luo ZW. 2007. Impacts of yeast metabolic network structure on enzyme evolution. Genome Biol. 8:407.

Lukes J, Paris Z, Regmi S, Breitling R, Mureev S, Kushnir S, Pyatkov K, Jirků M, Alexandrov K a. 2006. Translational initiation in *Leishmania tarentolae* and *Phytomonas serpens* (Kinetoplastida) is strongly influenced by pre-ATG triplet and its 5' sequence context. Mol. Biochem. Parasitol. 148:125–132.

Luque-Ortega JR, Rivas L. 2007. Miltefosine (hexadecylphosphocholine) inhibits cytochrome c oxidase in *Leishmania donovani* promastigotes. Antimicrob. Agents Chemother. 51:1327–1332.

Luque-Ortega JR, van't Hof W, Veerman ECI, Saugar JM, Rivas L. 2008. Human antimicrobial peptide histatin 5 is a cell-penetrating peptide targeting mitochondrial ATP synthesis in *Leishmania*. FASEB J. 22:1817–1828.

Lv W, Xu Y, Guo Y, Yu Z, Feng G, Liu P, Luan M, Zhu H, Liu G, Zhang M, et al. 2016. The drug target genes show higher evolutionary conservation than non-target genes. Oncotarget 7:4961–4971.

Lye L-F, Kang SO, Nosanchuk JD, Casadevall A, Beverley SM. 2011. Phenylalanine hydroxylase (PAH) from the lower eukaryote *Leishmania major*. Mol. Biochem. Parasitol. 175:58–67.

Mannaert A, Downing T, Imamura H, Dujardin J-C. 2012. Adaptive mechanisms in pathogens: universal aneuploidy in *Leishmania*. Trends Parasitol. 28:370–376.

Mantilla BS, Paes LS, Pral EMF, Martil DE, Thiemann OH, Fernández-Silva P, Bastos EL, Silber AM. 2015. Role of Δ1-pyrroline-5-carboxylate dehydrogenase supports mitochondrial metabolism and host-cell invasion of *Trypanosoma cruzi*. J. Biol. Chem. 290:7767–7790.

Martin JL, Yates PA, Soysa R, Alfaro JF, Yang F, Burnum-Johnson KE, Petyuk VA, Weitz KK, Camp II DG, Smith RD, et al. 2014. Metabolic reprogramming during purine stress in the protozoan pathogen *Leishmania donovani*. PLoS Pathog. 10:e1003938.

Martin WE, Bridgmon KD. 2012. Quantitative and statistical research methods: From hypothesis to results. John Wiley & Sons

McCall L-I, El Aroussi A, Choi JY, Vieira DF, De Muylder G, Johnston JB, Chen S, Kellar D, Siqueira-Neto JL, Roush WR, et al. 2015. Targeting ergosterol biosynthesis in *Leishmania donovani*: essentiality of sterol 14alpha-demethylase. PLoS Negl Trop Dis 9:e0003588.

McCall L-I, Matlashewski G. 2010. Localization and induction of the A2 virulence factor in *Leishmania*: evidence that A2 is a stress response protein. Mol. Microbiol. 77:518–530.

McCall L-I, Zhang W-W, Matlashewski G. 2013. Determinants for the development of visceral leishmaniasis disease. PLoS Pathog. 9:e1003053.

McConville MJ, Naderer T. 2011. Metabolic pathways required for the intracellular survival of *Leishmania*. Annu. Rev. Microbiol. 6:543–561.

McConville MJ, Saunders EC, Kloehn J, Dagley MJ. 2015. *Leishmania* carbon metabolism in the macrophage phagolysosome-feast or famine? F1000Research 4.

McConville MJ, de Souza D, Saunders E, Likic VA, Naderer T. 2007. Living in a phagolysosome; metabolism of *Leishmania* amastigotes. Trends Parasitol. 23:368–375.

Meade JC, Glaser TA, Bonventre PF, Mukkada AJ. 1984. Enzymes of carbohydrate metabolism in *Leishmania donovani* amastigotes. J. Protozool. 31:156–161.

Merlen T, Sereno D, Brajon N, Rostand F, Lemesre J-L. 1999. *Leishmania* spp: completely defined medium without serum and macromolecules (CDM/LP) for the continuous in vitro cultivation of infective promastigote forms. Am. J. Trop. Med. Hyg. 60:41–50.

Michels PAM. 1988. Compartmentation of glycolysis in trypanosomes: a potential target for new trypanocidal drugs. Biol. Cell 64:157–164.

Michels PAM, Bringaud F, Herman M, Hannaert V. 2006. Metabolic functions of glycosomes in trypanosomatids. Biochim. Biophys. Acta (BBA)-Molecular Cell Res. 1763:1463–1477.

Misset O, Bos OJM, Opperdoes FR. 1986. Glycolytic enzymes of *Trypanosoma brucei*. Eur. J. Biochem. 157:441–453.

Monzote L. 2009. Current Treatment of Leishmaniasis : A Review. Open Antimicrob. Agents J. 1:9–19.

Moreno MA, Alonso A, Alcolea PJ, Abramov A, de Lacoba MG, Abendroth J, Zhang S, Edwards T, Lorimer D, Myler PJ, et al. 2014. Tyrosine aminotransferase from *Leishmania infantum*: A new drug target candidate. Int. J. Parasitol. Drugs Drug Resist. 4:347–354.

Mukherjee A, Roy G, Guimond C, Ouellette M. 2009. The γ-glutamylcysteine synthetase gene of *Leishmania* is essential and involved in response to oxidants. Mol. Microbiol. 74:914–927.

Mukherjee T, Ray M, Bhaduri A. 1988. Aspartate transcarbamylase from *Leishmania donovani*. A discrete, nonregulatory enzyme as a potential chemotherapeutic site. J. Biol. Chem. 263:708–713.

Murta SMF, Vickers TJ, Scott DA, Beverley SM. 2009. Methylene tetrahydrofolate dehydrogenase/cyclohydrolase and the synthesis of 10-CHO-THF are essential in *Leishmania major*. Mol. Microbiol. 71:1386–1401.

Naderer T, Ellis MA, Sernee MF, De Souza DP, Curtis J, Handman E, McConville MJ. 2006. Virulence of *Leishmania major* in macrophages and mice requires the gluconeogenic enzyme fructose-1, 6-bisphosphatase. Proc. Natl. Acad. Sci. 103:5502–5507.

Naderer T, Heng J, McConville MJ. 2010. Evidence that intracellular stages of *Leishmania major* utilize amino sugars as a major carbon source. PLoS Pathog 6:e1001245.

Naderer T, Heng J, Saunders EC, Kloehn J, Rupasinghe TW, Brown TJ, McConville MJ. 2015. Intracellular survival of *Leishmania major* depends on uptake and degradation of extracellular matrix glycosaminoglycans by macrophages. PLoS Pathog 11:e1005136.

Nare B, Hardy LW, Beverley SM. 1997. The roles of pteridine reductase 1 and dihydrofolate reductase-thymidylate synthase in pteridine metabolism in the protozoan parasite *Leishmania major*. J. Biol. Chem. 272:13883–13891.

Nirujogi RS, Pawar H, Renuse S, Kumar P, Chavan S, Sathe G, Sharma J, Khobragade S, Pande J, Modak B, et al. 2014. Moving from unsequenced to sequenced genome: reanalysis of the proteome of *Leishmania donovani*. J. Proteomics 97:48–61.

Notebaart RA, Teusink B, Siezen RJ, Papp B. 2008. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. PLoS Comput. Biol. 4:e26.

Novoa EM, de Pouplana L. 2012. Speeding with control: codon usage, tRNAs, and ribosomes. Trends Genet. 28:574–581.

Olson-Manning CF, Lee C-R, Rausher MD, Mitchell-Olds T. 2012. Evolution of flux control in the glucosinolate pathway in *Arabidopsis thaliana*. Mol. Biol. Evol. 30:14–23.

Opperdoes FR, Coombs GH. 2007. Metabolism of *Leishmania*: proven and predicted. Trends Parasitol. 23:149–158.

Opperdoes FR, Michels PAM. 2001. Enzymes of carbohydrate metabolism as potential drug targets. Int. J. Parasitol. 31:482–490.

Opperdoes FR, Szikora J-P. 2006. *In silico* prediction of the glycosomal enzymes of *Leishmania major* and trypanosomes. Mol. Biochem. Parasitol. 147:193–206.

Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BØ. 2011. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism—2011. Mol. Syst. Biol. 7:535.

Paape D, Barrios-Llerena ME, Le Bihan T, Mackay L, Aebischer T. 2010. Gel free analysis of the proteome of intracellular *Leishmania mexicana*. Mol. Biochem. Parasitol. 169:108–114.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat. Rev. Genet. 7:337–348.

Papp B, Notebaart RA, Pál C. 2011. Systems-biology approaches for predicting genomic evolution. Nat. Rev. Genet. 12:591–602.

Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail M a, Peters N, Adlem E, Tivey A, Aslett M, et al. 2007. Comparative genomic analysis of three Leishmania species that cause diverse human disease. Nat. Genet. 39:839–847.

Peters C, Kawakami M, Kaul M, Ilg T, Overath P, Aebischer T. 1997. Secreted proteophosphoglycan of *Leishmania mexicana* amastigotes activates complement by triggering the mannan binding lectin pathway. Eur. J. Immunol. 27:2666–2672.

Pinney JW, Papp B, Hyland C, Wambua L, Westhead DR, McConkey GA. 2007. Metabolic reconstruction and analysis for parasite genomes. Trends Parasitol. 23:548–554.

Plaimas K, Eils R, König R. 2010. Identifying essential genes in bacterial metabolic networks with machine learning methods. BMC Syst. Biol. 4:1.

Plata G, Hsiao T-L, Olszewski KL, Llinás M, Vitkup D. 2010. Reconstruction and flux-balance analysis of the *Plasmodium falciparum* metabolic network. Mol. Syst. Biol. 6.

Plotkin JB, Kudla G. 2010. Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. 12:32–42.

Portal P, Villamil SF, Alonso GD, De Vas MG, Flawiá MM, Torres HN, Paveto C. 2008. Multiple NADPH-cytochrome P450 reductases from *Trypanosoma cruzi*: Suggested role on drug resistance. Mol. Biochem. Parasitol. 160:42–51.

Rao VS, Srinivas K, Sujini GN, Kumar GNS. 2014. Protein-Protein Interaction Detection : Methods and Analysis. 2014.

Rashmi M, Swati D. 2013. Comparative Genomics of Trypanosomatid Pathogens using Codon Usage Bias. Bioinformation 9:912.

Rastrojo A, Carrasco-Ramiro F, Mart'\in D, Crespillo A, Reguera RM, Aguado B, Requena JM. 2013. The transcriptome of *Leishmania major* in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq. BMC Genomics 14:223.

Ravikrishnan A, Raman K. 2015. Critical assessment of genome-scale metabolic networks: the need for a unified standard. Brief. Bioinform. 16:1057–1068.

Raymond F, Boisvert S, Roy G, Ritt J-F, Légaré D, Isnard A, Stanke M, Olivier M, Tremblay MJ, Papadopoulou B. 2011. Genome sequencing of the lizard parasite *Leishmania tarentolae* reveals loss of genes associated to the intracellular stage of human pathogenic species. Nucleic Acids Res. 40:1131–1147.

Real F, Vidal RO, Carazzolle MF, Mondego JM, Costa GGL, Herai RH, Würtele M, de Carvalho LM, e Ferreira RC, Mortara RA, et al. 2013. The genome sequence of *Leishmania* (*Leishmania*) amazonensis: functional annotation and extended analysis of gene models. DNA Res. 20:567–581.

Ren Q, Chen K, Paulsen IT. 2006. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res. 35:D274--D279.

Rezende AM, Folador EL, Resende D de M, Ruiz JC. 2012. Computational prediction of protein-protein interactions in *Leishmania* predicted proteomes. PLoS One 7:e51304.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16:276–277.

Roberts CW, McLeod R, Rice DW, Ginger M, Chance ML, Goad LJ. 2003. Fatty acid and sterol metabolism: potential antimicrobial targets in apicomplexan and trypanosomatid parasitic protozoa. Mol. Biochem. Parasitol. 126:129–142.

Roberts SB, Robichaux JL, Chavali AK, Manque PA, Lee V, Lara AM, Papin JA, Buck GA. 2009. Proteomic and network analysis characterize stage-specific metabolism in *Trypanosoma cruzi*. BMC Syst. Biol. 3:1.

Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, Depledge DP, Harris D, Her Y, Herzyk P, Imamura H, Otto TD, et al. 2011. Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. 21:2129–2142.

Rosenzweig D, Smith D, Myler PJ, Olafson RW, Zilberstein D. 2008. Post-translational modification of cellular proteins during *Leishmania donovani* differentiation. Proteomics 8:1843–1850.

Rosenzweig D, Smith D, Opperdoes F, Stern S, Olafson RW, Zilberstein D. 2008. Retooling *Leishmania* metabolism: from sand fly gut to human macrophage. FASEB J. 22:590–602.

Roy G, Ouellette M. 2015. Inactivation of the cytosolic and mitochondrial serine hydroxymethyl transferase genes in *Leishmania major*. Mol. Biochem. Parasitol. 204:106–110.

Salathé M, Ackermann M, Bonhoeffer S. 2005. The effect of multifunctionality on the rate of evolution in yeast. Mol. Biol. Evol. 23:721–722.

Santos A, Branquinha M, d'Avila-Levy C, Kneipp L, Sodré C. 2013. Proteins and Proteomics of *Leishmania and Trypanosoma*. Springer Science & Business Media

Sardar AH, Jardim A, Ghosh AK, Mandal A, Das S, Saini S, Abhishek K, Singh R, Verma S, Kumar A, et al. 2016. Genetic manipulation of *Leishmania donovani* to explore the involvement of argininosuccinate synthase in oxidative stress management. PLoS Negl Trop Dis 10:e0004308.

Saunders EC, Ng WW, Chambers JM, Ng M, Naderer T, Krömer JO, Likic V a, McConville MJ. 2011. Isotopomer profiling of *Leishmania mexicana* promastigotes reveals important roles for succinate fermentation and aspartate uptake in tricarboxylic acid cycle (TCA) anaplerosis, glutamate synthesis, and growth. J. Biol. Chem. 286:27706–27717.

Saunders EC, Ng WW, Kloehn J, Chambers JM, Ng M, McConville MJ. 2014. Induction of a stringent metabolic response in intracellular stages of *Leishmania mexicana* leads to increased dependence on mitochondrial metabolism. PLoS Pathog. 10:e1003888.

Saunders EC, de Souza DP, Naderer T, Sernee MF, Ralton JE, Doyle MA, MacRae JI, Chambers JL, Heng J, Nahid A, et al. 2010. Central carbon metabolism of *Leishmania* parasites. Parasitology 137:1303–1313.

Schomburg I, Chang A, Placzek S, Söhngen C, Rother M, Lang M, Munaretto C, Ulas S, Stelzer M, Grote A, et al. 2012. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res.:gks1049.

Schroeder J, Aebischer T. 2014. Vaccines for Leishmaniasis: From proteome to vaccine candidates. Hum. Vaccin. 7:10–15.

Scott DA, Hickerson SM, Vickers TJ, Beverley SM. 2008. The role of the mitochondrial glycine cleavage complex in the metabolism and virulence of the protozoan parasite *Leishmania major*. J. Biol. Chem. 283:155–165.

Searls DB. 2003. Pharmacophylogenomics: genes, evolution and drug targets. Nat. Rev. Drug Discov. 2:613.

Shah P, Gilchrist MA. 2011. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proc. Natl. Acad. Sci. 108:10231–10236.

Shaked-Mishan P, Suter-Grotemeyer M, Yoel-Almagor T, Holland N, Zilberstein D, Rentsch D. 2006. A novel high-affinity arginine transporter from the human parasitic protozoan *Leishmania donovani*. Mol. Microbiol. 60:30–38.

Sharma M, Shaikh N, Yadav S, Singh S, Garg P. 2017. A systematic reconstruction and constraint-based analysis of *Leishmania donovani* metabolic network: identification of potential antileishmanial drug targets. Mol. Biosyst. 13:955–969.

Sharp PM, Li W-H. 1987. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.

Silva AM, Cordeiro-da-Silva A, Coombs GH. 2011. Metabolic variation during development in culture of *Leishmania donovani* promastigotes. PLoS Negl Trop Dis 5:e1451.

Singh AK, Vidyarthi AS. 2011. Codon usage pattern of differentially expressed genes in *Leishmania* species. 10:188–195.

Singh VP, Ranjan A, Topno RK, Verma RB, Siddique NA, Ravidas VN, Kumar N, Pandey K, Das P. 2010. Estimation of under-reporting of visceral leishmaniasis cases in Bihar, India. Am. J. Trop. Med. Hyg. 82:9–11.

Smith DF, Peacock CS, Cruz AK. 2007. Comparative genomics: from genotype to disease phenotype in the leishmaniases. Int. J. Parasitol. 37:1173–1186.

Smith M, Blanchette M, Papadopoulou B. 2008. Improving the prediction of mRNA extremities in the parasitic protozoan *Leishmania*. BMC Bioinformatics 9:158.

Sousa AF, Gomes-Alves AG, Ben'itez D, Comini MA, Flohé L, Jaeger T, Passos J, Stuhlmann F, Tomás AM, Castro H. 2014. Genetic and chemical analyses reveal that trypanothione synthetase but not glutathionylspermidine synthetase is essential for *Leishmania infantum*. Free Radic. Biol. Med. 73:229–238.

Subramanian A, Jhawar J, Sarkar RR. 2015. Dissecting *Leishmania infantum* energy metabolism - A systems perspective. PLoS One 10.

Subramanian A, Sarkar RR. 2015. Data in support of large scale comparative codon usage analysis in *Leishmania* and Trypanosomatids. Data in Brief. 4:269–272.

Subramanian A, Sarkar RR. 2015. Comparison of codon usage bias across *Leishmania* and Trypanosomatids to understand mRNA secondary structure, relative protein abundance and pathway functions. Genomics 106:232–241.

Subramanian A, Sarkar RR. 2016. Network structure and enzymatic evolution in *Leishmania* metabolism: a computational study. In: BIOMAT 2015: *Proceedings of the International Symposium on Mathematical and Computational Biology - BIOMAT 2015*., *World Scientific*, ISBN: 978-981-3141-90-2, 1 – 20, 11/2015, DOI:10.1142/9789813141919_0001

Subramanian A, Sarkar RR. 2017. Revealing the mystery of metabolic adaptations using a genome scale model of *Leishmania infantum*. Sci. Rep. 7.

Sueoka N. 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. 85:2653–2657.

Szappanos B, Fritzemeier J, Csörg Ho B, Lázár V, Lu X, Fekete G, Bálint B, Herczeg R, Nagy I, Notebaart RA, et al. 2016. Adaptive evolution of complex innovations through stepwise metabolic niche expansion. Nat. Commun. 7:11607.

Thorleifsson SG, Thiele I. 2011. rBioNet: A COBRA toolbox extension for reconstructing high-quality biochemical networks. Bioinformatics 27:2009–2010.

Titus RG, Gueiros-Filho FJ, de Freitas LA, Beverley SM. 1995. Development of a safe live *Leishmania* vaccine line by gene replacement. Proc. Natl. Acad. Sci. 92:10267–10271.

Tovar J, Wilkinson S, Mottram JC, Fairlamb AH. 1998. Evidence that trypanothione reductase is an essential enzyme in *Leishmania* by targeted replacement of the tryA gene locus. Mol. Microbiol. 29:653–660.

Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. Proc. Natl. Acad. Sci. 107:3645–3650.

Tuon FF, Neto VA, Amato VS. 2008. *Leishmania*: origin, evolution and future since the Precambrian. FEMS Immunol. Med. Microbiol. 54:158–166.

UniProt Consortium. 2014. Activities at the universal protein resource (UniProt). Nucleic Acids Res. 42:D191-D198.

Vannier-Santos MA, Urbina JA, Martiny A, Neves A, Souza W. 1995. Alterations induced by the antifungal compounds ketoconazole and terbinafine in *Leishmania*. J. Eukaryot. Microbiol. 42:337–346.

Vasconcelos EJR, Terrão MC, Ruiz JC, Vêncio RZN, Cruz K. 2012. In silico identification of conserved intercoding sequences in *Leishmania* genomes: unraveling putative cis-regulatory elements. Mol. Biochem. Parasitol. 183:140–150.

Veitia RA, Bottani S, Birchler JA. 2008. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. Trends Genet. 24:390–397.

Di Ventura B, Lemerle C, Michalodimitrakis K, Serrano L. 2006. From in vivo to in silico biology and back. Nature 443:527–533.

Verlinde CLMJ, Hannaert V, Blonski C, Willson M, Périé JJ, Fothergill-Gilmore LA, Opperdoes FR, Gelb MH, Hol WGJ, Michels PAM. 2001. Glycolysis as a target for the design of new anti-trypanosome drugs. Drug Resist. Updat. 4:50–65.

Vicario S, Moriyama EN, Powell JR. 2007. Codon usage in twelve species of *Drosophila*. BMC Evol. Biol. 7:226.

Vickers TJ, Orsomando G, de la Garza RD, Scott DA, Kang SO, Hanson AD, Beverley SM. 2006. Biochemical and genetic analysis of methylenetetrahydrofolate reductase in *Leishmania* metabolism and virulence. J. Biol. Chem. 281:38150–38158.

Vitkup D, Kharchenko P, Wagner A. 2006. Influence of metabolic network structure and function on enzyme evolution. Genome Biol. 7:R39.

van der Voet H. 1994. Comparing the predictive accuracy of models using a simple randomization test. Chemom. Intell. Lab. Syst. 25:313–323.

Warringer J, Blomberg A. 2006. Evolutionary constraints on yeast protein size. BMC Evol. Biol. 6:61.

van Weelden SWH, van Hellemond JJ, Opperdoes FR, Tielens AGM. 2005. New functions for parts of the Krebs cycle in procyclic *Trypanosoma brucei*, a cycle not operating as a cycle. J. Biol. Chem. 280:12451–12460.

Westrop GD, Williams RAM, Wang L, Zhang T, Watson DG, Silva AM, Coombs GH. 2015. Metabolomic analyses of *Leishmania* reveal multiple species differences and large differences in Amino Acid Metabolism. PLoS One 10.

Wright F. 1990. The "effective number of codons" used in a gene. Gene 87:23–29.

Yamada T, Bork P. 2009. Evolution of biomolecular networks — lessons from metabolic and protein interactions. Nat. Rev. Mol. Cell Biol. 10:791–803.

Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. Mol. Biol. Evol. 28:2359–2369.

Yang Z. 1998. Synonymous and nonsynonimous rate variation in nuclear genes of mammals. J Mol Evol 46:409–418.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Zand M, Narasu ML. 2013. Vaccination against leishmaniasis. Ann. Biol. Res. 4:170–174.

Zar JH. 1996. Biostatistics. Prentice Hall. Englewood Cliffs, New Jersey, USA.

Zhang J, Yang J-R. 2015. Determinants of the rate of protein sequence evolution. Nat. Rev. Genet. 16:409.

Zhang K, Beverley SM. 2010. Phospholipid and sphingolipid metabolism in *Leishmania*. Mol. Biochem. Parasitol. 170:55–64.

Zhang K, Showalter M, Revollo J, Hsu F-F, Turk J, Beverley SM. 2003. Sphingolipids are essential for differentiation but not growth in *Leishmania*. EMBO J. 22:6016–6026.

Zhang W-W, Matlashewski G. 2001. Characterization of the A2-A2rel gene cluster in *Leishmania donovani*: involvement of A2 in visceralization during infection. Mol. Microbiol. 39:935–948.

Zhang W-W, Matlashewski G. 2010. Screening *Leishmania donovani*-specific genes required for visceral infection. Mol. Microbiol. 77:505–517.

Zhang WW, Ramasamy G, McCall L-I, Haydock A, Ranasinghe S, Abeygunasekara P, Sirimanna G, Wickremasinghe R, Myler P, Matlashewski G. 2014. Genetic analysis of *Leishmania donovani* tropism using a naturally attenuated cutaneous strain. PLoS Pathog. 10:e1004244.

Zilberstein D, Shapira M. 1994. The role of pH and temperature in the development of *Leishmania* parasites. Annu. Rev. Microbiol. 48:449–470.

Zuker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 9:133–148.