Characterisation of the Transcriptomes of *Leishmania mexicana* Promastigotes and Amastigotes

A thesis submitted for the degree of Doctor of Philosophy at the University of Oxford

Trinity Term 2014

Michael Fiebig
University College

Sir William Dunn School of Pathology

" ‘Why don't you just make 10 louder and make
10 be the top number and make that a little louder?’
‘These go to eleven!’ "

(Spinal Tap)

„Da steh ich nun, ich armer Tor!
Und bin so klug als wie zuvor“

(Faust, J. W. Goethe)

Abstract

Leishmania spp. undergo substantial adaptations from being promastigotes, found in sandflies, to being amastigotes, residing in parasitophorous vacuoles within mammalian macrophages. In the past, microarray studies have sought to elucidate these adaptations using axenic amastigote systems or amastigotes purified from host-cells, raising the question whether the observed transcriptomic signatures were a true reflection of intracellular amastigotes. Moreover, with ever-improving genome annotations being available, it is clear that these studies failed to address the transcriptomic behaviour of a considerable number of transcripts.

In the work presented herein, I employed RNA-sequencing to obtain transcriptomic profiles of *Leishmania mexicana* axenic promastigotes (PRO), axenic amastigotes (AXA) and intracellular amastigotes (AMA) in murine bone-marrow derived macrophages. The intracellular amastigotes were not purified from host cells, but instead sequencing reads assigned to a hybrid *L. mexicana* - *Mus musculus* genome and the transcriptomes separated *in silico.* We were able to map pre-mRNA processing sites, thereby defining transcript boundaries, proposing 184 truncations and 1253 extensions of existing gene models as well as discovering 936 novel genes. Mass-spectrometric evidence was obtained for both proposed extended and novel proteins. Using this improved genome annotation, we generated gene expression profiles for AMA, AXA and PRO, identifying 3832 differentially expressed transcripts between PRO and AMA as well as 2176 between PRO and AXA and 1234 between AXA and AMA. Transcripts differentially expressed between AMA and PRO correlated well with previous reports, were enriched for novel transcripts identified in this study and contained an unprecedented wealth of yet uncharacterised transcripts.

Guided by these data, I performed a GFP-tagging screen identifying two proteins which may play an important role in *L. mexicana* biology, LmxM.16.0500, a member of a small, divergent, amastin-derived gene family, which appears to be released from the cell body of PRO, and LmxM.09.1330 a specific marker of the amastigote flagellar pocket.

Acknowledgements

First of all I would like to thank my supervisors. Eva, thank you very much for all the discussions and advice and above all thank you for the patience you had with me. Keith, thank you for knowing how to find the right balance between challenge and encouragement to keep me going.

I would also like to express my gratitude to Steve Kelly for all those hours of bioinformatics and for making SLaP Mapper possible…it IS catchier than PASSTA. Then I would like to thank all of the members of the Gluenz and Gull labs. Whilst you have often made me "lose the will to live", you have been a great source of advice and motivation and I won't forget this.

I must also acknowledge my parents, my grandparents and my sister, for instilling in me an interest in figuring out "how things work" and the stubbornness to pull through with this.

Thank you Clive for always listening to me moan about work, Jammy, Simon and everyone else from "Stores" for bailing me out whenever my planning was non-existent, as well as Svenja and Ben for their help with proteomics.

Furthermore, cannot thank my two Dutchies enough: My housemate Joram, not only for dealing with my usual messy madness, distillery and amplifier parts at home, but also for making sure I actually survived this…then again…what could possibly go wrong?

And my love Anne, for your caring support, love and understanding, for letting me scribble in your calendar when "I have an idea, its genius!" and for at least trying to keep me from wearing "that shirt" to work again.

Finally, I would like to thank Louie, the hound, the one and only four-legged, ferociously fluffy fur ball, for always bringing the silliest possible smile to my face.

Declarations

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other university or institute of learning.

Excluding references the text of this thesis consists of 43,474 words.

Preface

This thesis is arranged in three results chapters (Chapter 2, 3 &4) with a general introduction (Chapter 1), discussion (Chapter 5) and materials and methods (Chapter 6).

Figures and tables are each shown on a separate page and inserted close to their initial reference. All tables spanning more than two pages are provided on a separate disk as are computer scripts and data bases.

Contents

Chapter 1 - Introduction

1.1 *Leishmania spp*. and Leishmaniasis

Parasites of the order of kinetoplastida are a group of protozoan organisms famed for being the etiological agents of three important human diseases, namely African sleeping sickness (*Trypanosoma brucei*), Chagas disease (*Trypanosoma cruzi*) and Leishmaniasis (*Leishmania spp.*). The latter, depending on the particular species *Leishmania* contracted, is a spectrum of diseases ranging from self-healing local cutaneous ulcers (cutaneous leishmaniasis) (*L. major, L. mexicana*) over utterly disfiguring mucocutaneous disease (*L. braziliensis*, *L. amazonensis*) to visceral leishmaniasis (*L. infantum, L. donovani*), which, if left untreated, is fatal (WHO Report 2013). Additionally, variant disease manifestations exists, such as a dermal leishmaniasis occurring after treatment of Old-World visceral leishmaniasis as a result of an immunological reaction to parasites persisting within in the skin (Post kala-azar dermal leishmaniasis) (Desjeux et al. 2013) or diffuse, disseminating cutaneous leishmaniases found in Latin America, influenced by cellular immune responses (Turetz et al. 2002; Paniz Mondolfi et al. 2013). Cutaneous leishmaniases may take 1- 5 years to resolve, with the potential of leaving atrophic, acne-like scars or wholly disfiguring and disabling scars, whilst visceral leishmaniasis may only manifest after >1 year post-infection, but will progress to a severe wasting disease and ultimately death within 2-3 years (Pace 2014).

Traditionally distributed across tropical climates, South America and the Mediterranean basin, climate change and increased global travel have made Leishmaniasis a diagnosis many medical practitioners may have to entertain at one

Table 1.1 Disease manifestations and transmission of predominant *Leishmania* species found in the New and Old Worlds Table summarising the Sandfly vector, geographical distribution, main mammalian reservoir and disease manifestation of diverse *Leishmania spp.* Table taken from (Pace, 2014)

***Leishmania* species**	**Sandfly vector (*P.* = *Phlebotomus* , *L.* = *Lutzomyia*)**	**Main affected areas**	**Reservoir**	**Disease manifestations**
L. aethiopica	*P. longipes, P. pedifer*	Ethiopia, Kenya	Hyraxes	Cutaneous, diffuse, mucosal
L. amazonensis	*L. flaviscutellata*	East Andes	Rodents	Cutaneous, disseminated
L. braziliensis	*L. ovallesi, L. wellcomei, L. neivai, L. whitmani*	East and West Andes	Rodents, marsupials, dog	Cutaneous, mucosal
L. donovani	*P. argentipes, P. martini P. orientalis*	India, Bangladesh, Nepal, Bhutan	Human	Visceral
L. guyanensis	*L. umbratilis*	East Andes	Arboreal edentate mammals	Cutaneous, mucosal
L. infantum (same as L. chagasi in the New World)	*P. ariasi, P. perniciosus, L. longipalpis*	Mediterranean region, Latin America	Dog	Visceral, cutaneous
L. major	*P. duboscqi, P. papatasi*	Sub-Saharan Aftrica, North Africa, Middle East, Iran, Pakistan, India	Gerbils, Rodents	Cutaneous
L. mexicana	*L. olmeca olmeca*	West Andes	Rodents, marsupials	Cutaneous, diffuse, mucosal
L. panamensis	*None proven*	West Andes	Arboreal edentate mammals	Cutaneous, mucosal
L. peruviana	*None proven*	Peru	Rodents, marsupials, dog	Cutaneous, mucosal
L. tropica	*P. sergenti, P. arabicus, P. guggisbergi*	North Africa, Middle East, Iran, Afghanistan, North and sub-Saharan Africa	Human, Hyraxes	Cutaneous

Table 1.2 Drugs used to treat leishmaniasis Table summarising drugs used to treat diverse manifestations of leishmaniasis, their mode of action, route of administration as well as adverse effects and practical advatages and disadvantages. Table taken from (Pace, 2014)

Drug	Mode of action on Leishmania parasite	Route & main indication	Adverse effects	Advantages and disadvantages
Pentavalent antimonials -sodium stibogluconate -meglumine antimoniate	Inhibition of glycolysis and fatty acid oxidation Dose dependent inhibition of ATP and GTP formation	im/ iv: VL, CL, MCL, PKDL Intralesional: CL	Systemic: Pancreatitis, thrombocytopenia, leucopenia, cardiac arrhythmia, deranged liver enzymes	Cheapest formulations Development of resistance is problematic
Pentamidine isethionate	Inhibition of polyamine biosynthesis and disruption of mitochondrial membrane potential	im: CL, MCL Intralesional: CL	Pain at injection site Hypoglycaemia Hypotension Diabetes Renal dysfunction	Development of resistance is problematic Adverse effects limit its use
Amphotericin B and lipid formulations	Inhibition of cell membrane synthesis by binding to ergosterol Pore formation in cell membrane	iv: VL, CL, MCL, PKDL	Fever, chills, bone pain, Hypokalaemia, nephrotoxicity	Lipid formulations less toxic than amphotericin B Effective total dose varies with geographical region More expensive than antimonials
Paromomycin	Possible interference with RNA synthesis and membrane permeability	im: VL Topical: CL	Topical: Pain, erythema, blistering Systemic: hepatotoxicity, reversible VIII nerve damage	Combination with antimonials results in higher cure rates of VL in India, but not in Africa
Allopurinol	Interference with protein synthesis (purine salvage cycle)	Oral: VL, CL	Rash	Ineffective as monotherapy: used in combination with sodium stibogluconate for VL
Azole derivatives: -fluconazole, ketoconazole, itraconazole	Inhibition of 14a-lanosterol demethylase required for ergosterol biosynthesis	Oral: CL	Hepatotoxicity	Inconsistent success between species
Alkylphosphocoline analogues: Miltefosine (hexadecylphosphocholine)	Alteration of glycosylphosphatidylinositol anchor synthesis, ether lipid metabolism, signal transduction and alkyl-specific acyl-coenzyme A acyl-transferase	Oral: VL, CL, PKDL	Gastrointestinal disturbances Hepatorenal toxicity Teratogenic: contra-indicated in pregnancy	Lack of compliance results in emergence of resistance especially in anthroponotic transmission Higher cure rates when used concurrently with paromomycin or sequentially after liposomal amphotericin B for VL in India

VL: visceral leishmaniasis; CL: cutaneous leishmaniasis, MCL: Muco-cutaneous leishmaniasis, PKDL: Post Kala-azar Dermal Leishmaniasis, im: intramuscular, iv: intravenous.

point in their career (Maguire et al. 1998; Manfredi et al. 2001; P. D. Ready 2010; Demers, Forrest, and Weichert 2013).

Leishmaniases are transmitted during the bloodmeal of sandflies (family *Phlenotominae*, genera *Phlebotomus* (Old World) and *Lutzomyia* (New World)). Recent estimates show 93 of around 800 known sandfly species to transmit leishmaniasis (World Health Organization 2010). Different species of these 2-3 mm long insects are known to feed both at night and during the day as well as indoors and outdoors, complicating insecticide-spraying prevention approaches (Killick-Kendrick 1999). Additionally, sandflies infected with *Leishmania spp.* are more persistent feeders and therefore better transmitters, as the parasite alters their feeding behaviour leading to multiple probing of the same host (Paul D. Ready 2013).

For the majority of *Leishmania* species, humans are only a secondary host, and often not conducive to further transmission. Rodents and dogs are thought to represent the main reservoirs of *Leishmania spp.*, as summarised in Table 1.1 (Reviewed in (Pace 2014)), again with complicating implications for disease prevention and eradication.

Whilst treatments based on antimony-compounds, amphotericin B, paromomycin or miltefosine are available, these are often harsh on, inaccessible to or unaffordable for the patient (H. W. Murray et al. 2005). Table 1.2 summarises a range of treatment regimens commonly used. Moreover, it shows the diverse modes of action of different drugs, their route of administration as well as the adverse effects these drugs may have on a treated patient. Moreover, resistance to existing treatments is arising (Singh, Kumar, and Singh 2012). In the light of this,

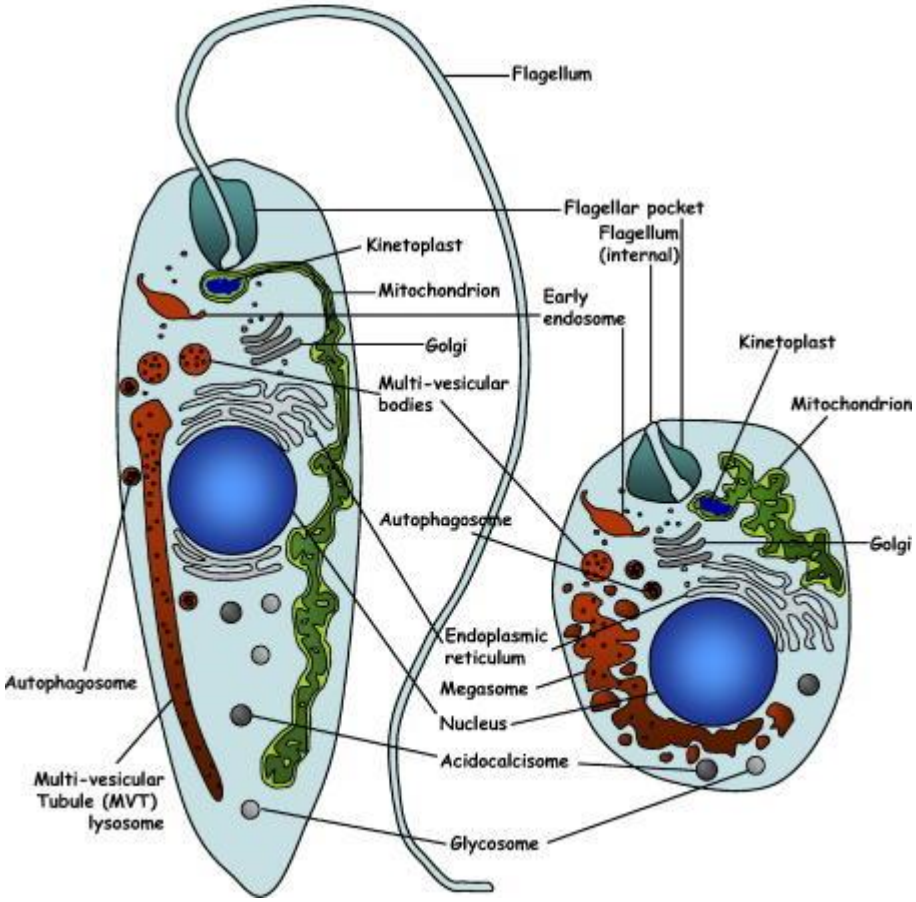


Figure 1.1 Subcellular organisation of *Leishmania spp.* promastigotes and amastigotes
Cartoon, taken from Besteiro *et al.* 2007, showing the organisation of subcellular structures in promastigotes (left) and amastigotes (right).

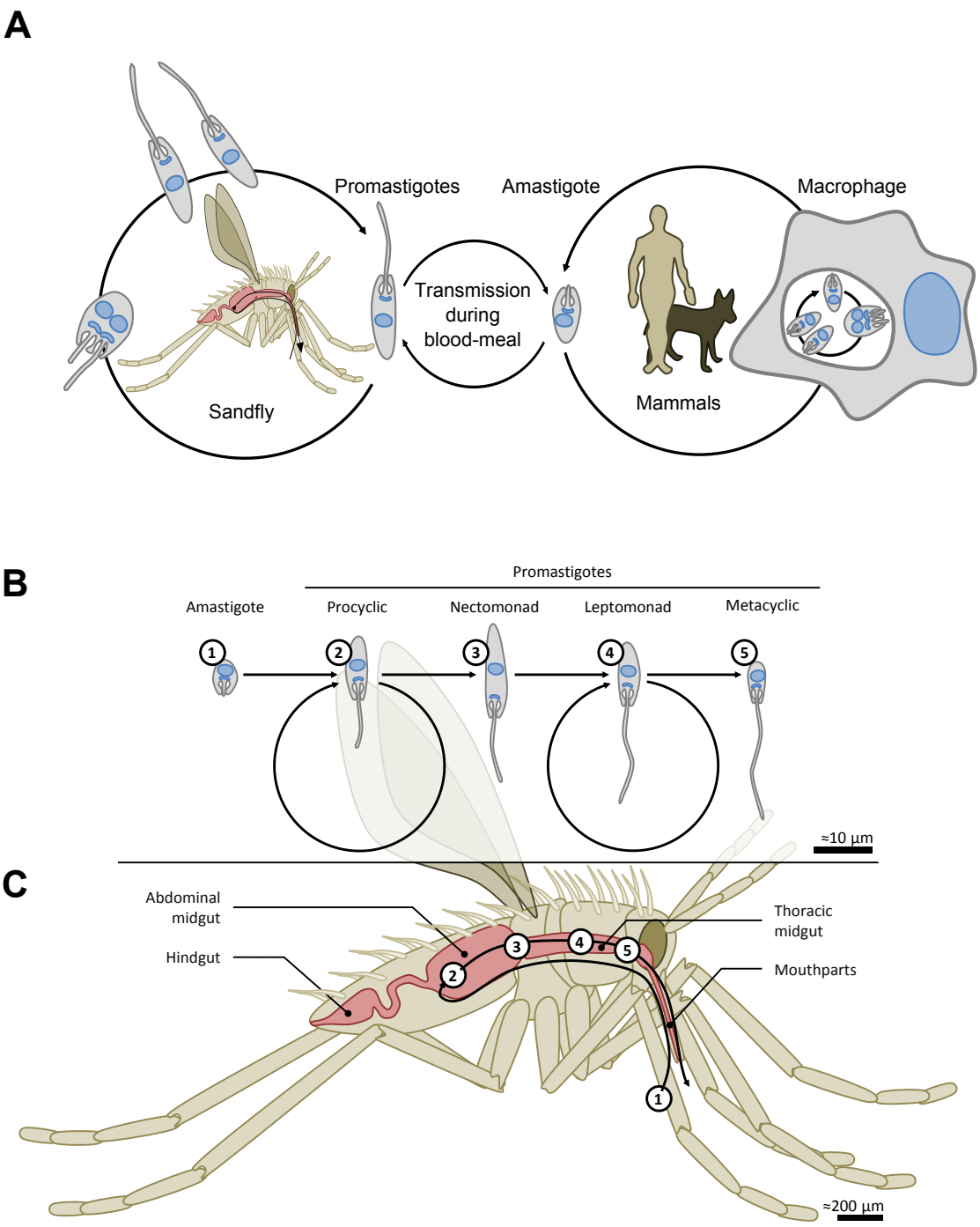


Figure 1.2 The life-cycle of *Leishmania spp.* Illustrations of the life-cycle of *Leishmania*. (**A**) *Leishmania* exist within sandflies as promastigotes and within vertebrate (generally mammalian) macrophages as amastigotes. (B) *Leishmania* transition through different morphologies, including proliferative (2 & 4) and non-proliferative (3 & 5) promastigote forms found in distinct compartments (**C**) within the sandlfy (2 - 5). Haptomonad promastigotes are not shown. (Adapted with kind permission from R. Wheeler, 2012, DPhil Thesis, University of Oxford).

research into the unique biology of *Leishmania spp.* with the ultimate aim of finding fulcrums for new medicinal interventions are as topical as ever.

Leishmaniasis has been described as a disease of poverty (Alvar, Yactayo, and Bern 2006), and it is easy to see how factors such as malnutrition, additional infectious diseases and lack of medical intervention contribute to this. Furthermore, basic, street-level housing (if housing is available at all), also contributes to disease risk as sandflies struggle to cover vertical distances of >1 m, making higher floors relatively safer than lower floors (Hewitt et al. 1998).

Leishmania parasites undergo drastic biochemical and morphological (see Figure 1.1) changes during their digenic life-cycle, from promastigotes in the sandfly to amastigotes inside phagocytes (Figure 1.2 A)

Biochemically, subcellular fractionation and enzymatic activity assays indicated that glycolysis over fatty-acid oxidation dominates in promastigotes with the inverse observed for amastigotes (Coombs, Craft, and Hart 1982). Moreover, promastigotes express high levels of Lipophosphoglycan (LPG) on their surface. LPG is a polymer of polymer of repeating Gal(β1,4)Man(α1-PO_4→6) units, linked to a 1-*O*-alkyl-2-*lyso*-phosphatidyl*(myo)*inositol anchor via a glycan core (Descoteaux and Turco 1999). Amastigotes, do not express LPG, instead express high levels of glycoinositol phospholipids (Winter et al. 1994). (Further surface markers of the promastigotes and amastigotes will be discussed in Chapter 4.)

Morphologically, promastigotes are 15-20 μm (M. E. Rogers, Chance, and Bates 2002) and amastigotes 3-5 μm long. Crucially, promastigotes feature a long, motile flagellum with a 9+2 microtubule architecture, in contrast to the 9+0-architecture

of the amastigote flagellum barely protruding from the flagellar pocket and closely apposed to the membrane of the parasitophorous vacuole (Gluenz et al. 2010). A range of promastigote forms exist in different compartments within the sandfly, varying in their cell-to-flagellum size-ratio as well as proliferative potential (Figure 1.2 B &C).

Taken up as amastigotes during the blood-meal of sandflies, *Leishmania* differentiate first procyclic promastigotes in the abdominal mid-gut of the sandfly (Figure 1.2 B & C). Subsequently, non-proliferative nectomonad promastigotes form which migrate to the anterior midgut. Here leptomonad promastigotes proliferate resulting in haptomonad promastigotes, attached to the cuticle-lined stomodeal valve and non-proliferative metacyclic promastigotes in the anterior thoracic midgut and posterior mouthpart components (Killick Kendrick, Molyneux, and Ashford 1974; M. E. Rogers, Chance, and Bates 2002; Gossage, Rogers, and Bates 2003; Paul A. Bates 2007).

Metacyclic promastigotes constitute the virulent promastigote form, displaying resistance to complement lysis (Howard, Sayers, and Miles 1987) and featuring a variant, thickened LPG coat (Turco and Descoteaux 1992).

Leishmania parasites are able to promote transmission during the blood-meal of the sandfly. Secretion of high levels of chitinase, causing damage to the stomodeal valve (Matthew E. Rogers et al. 2008), and secretion of proteophosphoglycan, a filamentous, mucin-like protein forming a gel blocking the anterior mid-gut (Matthew E. Rogers et al. 2004), promote regurgitation of sandfly saliva and thereby egestion of metacyclic promastigotes into the mammalian host.

Upon transfer into the mammal, parasites are taken up by macrophages, although

uptake by other phagocytic cells is also reported and may play a key role during infection. These cells include neutrophils (Ritter, Frischknecht, and van Zandbergen 2009), which, if not activated by infection with *Leishmania* promastigotes, will undergo apoptosis. The infected apoptotic neutrophils act as a "Trojan Horse" as they are phagocytosed by macrophages without activation of an inflammatory response, thereby leading to secondary infection of macrophages. Dendritic cells (DCs) , too, play a crucial role in infection, as amastigote-infected DCs are major source of Interleukin 12, responsible for inducing a protective adaptive Th1 and cytotoxic T-cell response (Kautz-Neu et al. 2012; Ashok and Acha-Orbea 2014), affecting the activation of adaptive immune-responses.

Uptake of metacyclic promastigotes into macrophages via the phagocytic pathway has been shown to be mediated by a range of host receptors, such as the complement receptor (CR) 3 and CR 1 interacting with complement opsonised promastigote surfaces, or the mannose receptor (CD206) interacting with the mannan-capped LPG backbone (Ueno and Wilson 2012; Polando et al. 2013). These interactions differ between *Leishmania spp.* (extensively reviewed in (Ueno and Wilson 2012)). The increase in temperature from 26-28 °C to 32-37 °C (Shapira, McEwen, and Jaffe 1988; Alcolea et al. 2010) and acidification of the phagosomal compartment (pH 4.7 – 5.3 (Antoine et al. 1990)) containing the initially metacyclic parasites leads to differentiation to amastigotes, which are able to withstand and proliferate in the acidic and lytic environment of the phagolysosome. This process is reproducible in axenic culture systems, exposing promastigotes to low pH (around pH5.5) and elevated temperature (32-37 °C) to generate axenic amastigotes. These systems exist for a range of *Leishmania* species,

e.g. *L. mexicana, L. amazonensis, L. braziliensis, L. chagasi, L. donovani* (P. A. Bates et al. 1992; Hodgkinson et al. 1996; Balanco et al. 1998; Sereno et al. 1998; Somanna, Mundodi, and Gedamu 2002; Debrabant et al. 2004), with, to date, the notable exception of *L.* major. Inside the macrophage, *Leishmania* cells interact with the cellular machinery of the host cell in a variety of ways. LPG, delays the maturation of the phagolysosome by promoting accumulation of F-actin around the phagosome preventing recruitment of lysosomal marker LAMP-1 and PKCα (Holm et al. 2001). Additionally, LPG disrupts the lipid microdomains of the parasite harbouring vacuole, thereby preventing recruitment and assembly of the NADPH oxidase (responsible for anti-microbial respiratory bursts) and v-ATPase (responsible for vacuolar acidification). Moreover, the parasite inhibits the JAK/STAT signalling pathway (Forget, Gregory, and Olivier 2005) by inducing degradation of STAT1α by the macrophage proteasome. Furthermore, the parasite metalloprotease GP63 cleaves a variety of host transcription factors such as NF-κB, CREB and AP-1 (Gregory et al. 2008; Gomez et al. 2009). A result of these interactions is inhibition of inflammatory cytokine (e.g. TNF-α, IL-12) and nitric oxide production as well as reduced sensitivity to INF-γ (Contreras et al. 2010), promoting parasite survival. (These and other virulence factors are discussed in more detail in Chapter 4.)

Equally, the evolutionary position of kinetoplastids in the eukaryotic lineage (Figure 1.3) means that understanding of kinetoplastid biology may provide general insights into the requirements to and evolution of eukaryotic cell biology: In the phylum euglenozoa, to which kinetoplastids belong, many core processes differ compared to other eukaryotes. For example, cytochrome C structure and

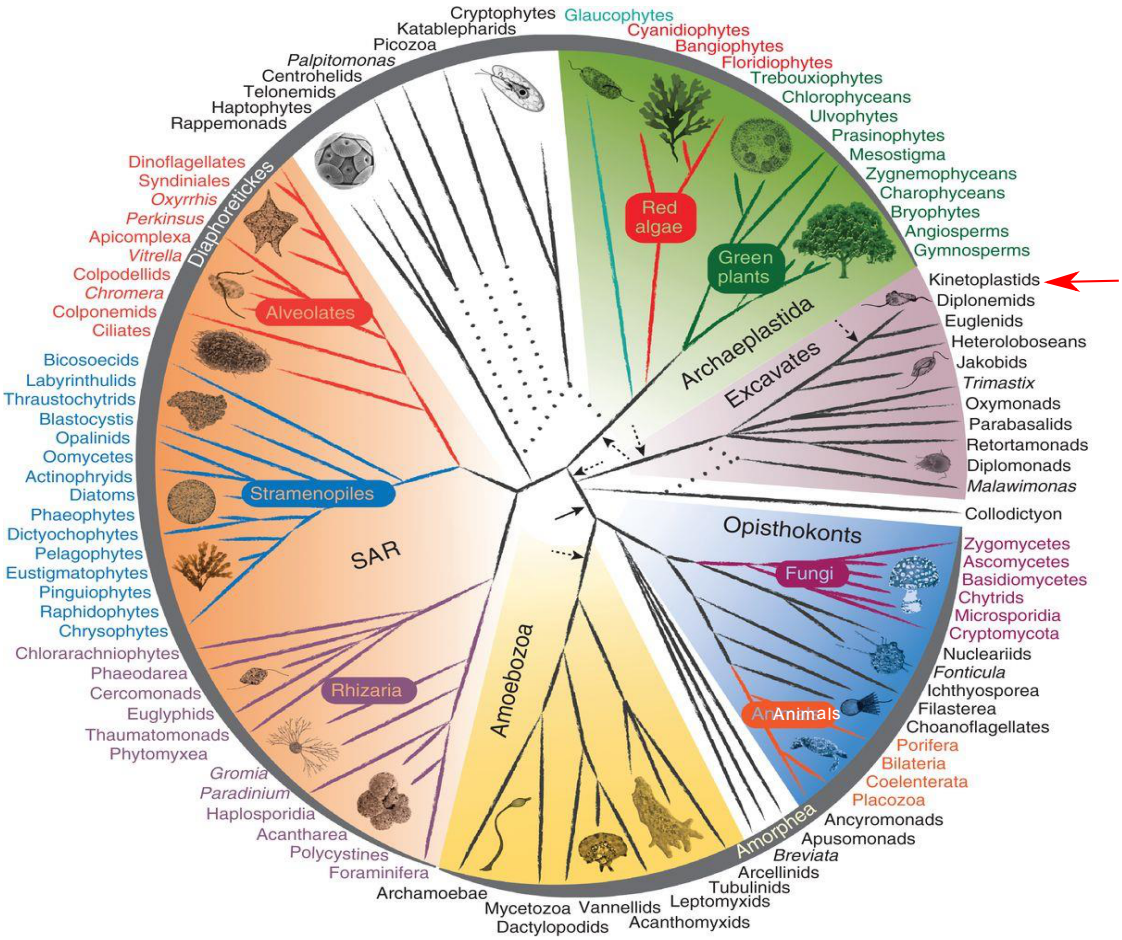


Figure 1.3 The Tree of Eukaryotes Cartoon illustrating the diversity of eukaryotic life taken from Burki, 2014. The major eukaryotic groups are shown. Solid lines represent certain, dotted lines uncertain relationships. The red arrow indicates the position of kinetoplastids. This tree is unrooted, with black arrows indicating proposed possible roots for the eukaryotic lineage. Note the possible root of eukaryotes close to the Kinetoplastid branch.

biosynthesis differs, as only a single-cysteine haem-binding motif is found and none of the genes of the three known Cytochrome C-biosynthetic pathways found in kinetoplastids (Allen et al. 2008). Another example is found in DNA-segregation. None of the conventional kinetochore proteins found in other eukaryotes were found in kinetoplastids (Akiyoshi and Gull 2013). Only recently did biochemical approaches by Akiyoshi (Akiyoshi and Gull 2013) reveal the identity of kinetochore components in *T. brucei*, which bore no detectable homology to conventional kinetochore proteins. Equally, the biology of the mitochondrion is highly unusual in kinetoplastids. The mitochondrion in only present as a single copy , with all of its DNA condensed into a single mass, the kinetoplast (Figure 1.1), which is attached to the flagellar basal body via a structure called the tripartite attachment complex (TAC) (Ogbadoyi, Robinson, and Gull 2003). The TAC is composed of filamentous structures between the basal body and the adjacent outer mitochondrial membrane, filaments linking the kinetoplast to the inner mitochondrial membrane as well as a differentiated, linear mitochondrial membrane devoid of cristae (Ogbadoyi, Robinson, and Gull 2003). Transmembrane proteins localising to the TAC and required for kinetoplast DNA segregation (Z. Zhao et al. 2008; Schnarwiler et al. 2014) have been identified across the inner (Ochsenreiter et al. 2008) and outer (Schnarwiler et al. 2014) mitochondrial membrane. Also, mitochondrial protein translocation differs in kinetoplastids compared to other eukaryotes as exemplified by the bacteria-like characteristics of the outer mitochondrial membrane protein translocase (Harsman et al. 2012).

As a result of such differences, evolutionary models placing euglenozoa at the base of the eukaryotic tree have been proposed (Cavalier-Smith 2010), albeit other

models are also under review (Figure 1.3), such as a root between amorphea (which includes amoebozoa and ophistokonts) and all other eukaryotes (Burki 2014).

Studies of the morphological adaptations of the parasite and its diverse molecular interactions with its hosts have proven invaluable to our understanding of *Leishmania spp.* biology, however holistic molecular studies promised yet more detailed insight into the mechanisms employed by the parasite to establish infection. Around ten years ago, the first kinetoplastid genomes were published (Berriman et al. 2005; El-Sayed, Myler, Bartholomeu, et al. 2005; Ivens et al. 2005), permitting comparative analyses of these very different, yet closely related organisms (Figure 1.4). Especially the publication of the first *Leishmania spp.* genome (*L. major*) (Ivens et al. 2005) opened the door to a new era of scientific investigation into *Leishmania spp.* biology. In the following I will review advances in our understanding of *Leishmania spp.* gene content, mechanisms of transcription and differentiation in the post-genomic era and which methodologies have principally contributed to this. Finally, I will try to indicate which advances in these fields lie immediately ahead of us.

1.2 Our understanding of the *Leishmania* genome before the sequence

Even in the early days of the discovery of *Leishmania spp.* in India independently by Leishman and Donovan in 1903 (Leishman 1903; Donovan 1903) the genome of *Leishmania* has played important roles in our understanding of the biology of the parasite: In a post-mortem preparation from the spleen of a soldier succumbed to a feverish disease, Leishman observed a large and small chromatin mass (nuclear and kinetoplast/mitochondrial DNA) in the suspected pathogen by light

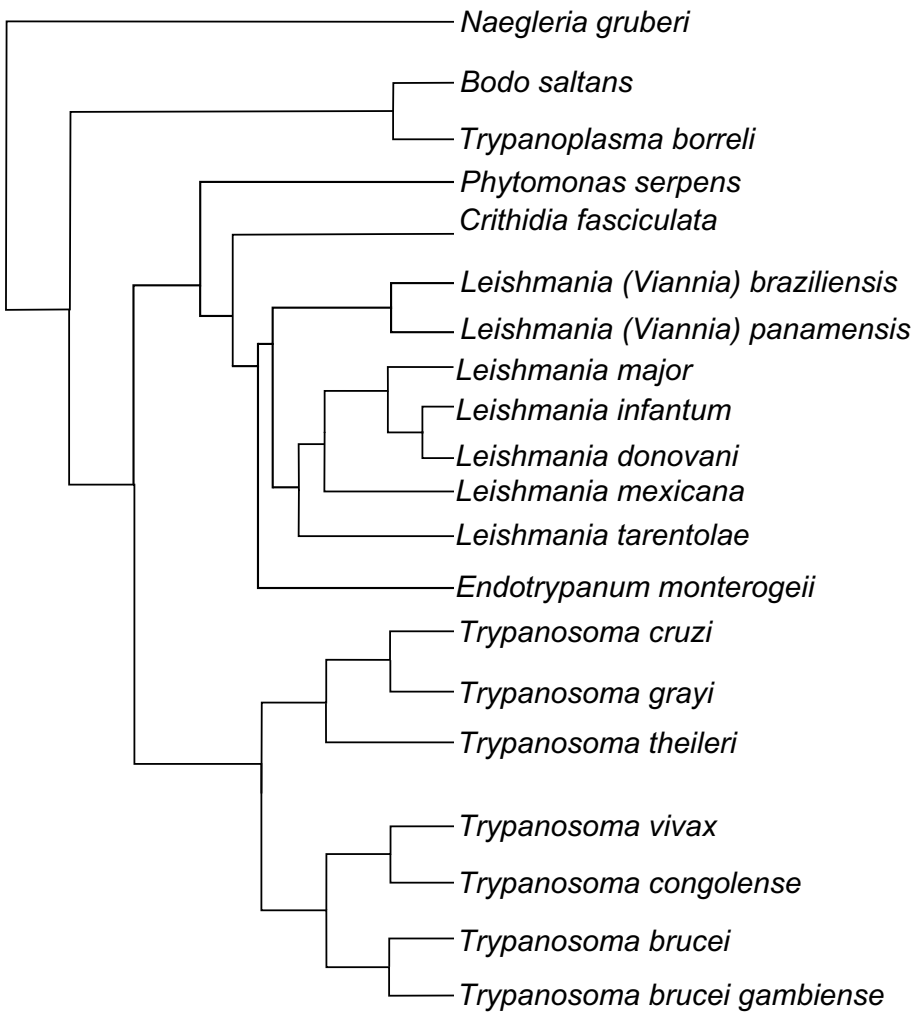


Figure 1.4 The evolutionary relationship of kinetoplastid species Cladogram of Kinetoplastid species rooted at *Naegleria gruberi,* a non-kinetoplastid flagellated excavate (see Fig. 1.3) The cladogram is adapted from Manna, Kelly and Field, 2013, with additional information from Van der Auwera *et al*. 2013 and Marcili *et al.* 2014.

microscopy that led him to draw links between the organisms in his samples and the trypanosomes that cause the disease Nagana in cattle in Africa discovered by Bruce (Bruce 1895), thereby predicting an evolutionary relationship validated by later discoveries and guiding research today ((Lukes et al. 1997; El-Sayed, Myler, Blandin, et al. 2005) and large sections of this thesis).

Following this brilliant insight however, the genome of *Leishmania spp.* posed many challenges and contained many surprises. Already the determination of the chromosome number in *Leishmania spp.* by light microscopy proved difficult as chromosomes did not condense during mitosis, making them difficult to count. As a result, only following development of pulse-field gel electrophoresis technology combined with oligonucleotide-probing was it possible to identify 36 chromosomes in *L. infantum* (Wincker et al. 1996).

Leishmania spp. are considered diploid organisms, although observations of aneuploidy were reported even before determination of the full number of chromosomes (Bastien, Blaineau, and Pages 1992; Cruz, Titus, and Beverley 1993; Sunkin et al. 2000). Indeed, attempts to delete dihydrofolate reductase-thymidylate synthase (DHFR-TS) from *L. major* (Cruz, Titus, and Beverley 1993) proved unsuccessful, instead resulting in genomic tetraploids containing two wild-type and two chromosomes with replaced DHFR-TS-loci or aneuploid trisomic cell lines with chromosomes bearing two replaced and one wild-type DHFR-TS locus. Similarly, attempts at replacing the protein kinase LmjF.01.0750 led to ploidy changes in *L. major* (Martínez-Calvillo, Stuart, and Myler 2005).

Comparisons of available gene sequences from different *Leishmania* species to other kinetoplastid species permitted establishment of evolutionary relationships

amongst these organisms (Lukes et al. 1997). Importantly, this allowed the discernment of two main groups: *Leishmania Leishmania spp.* (including *L. major, L. infantum, L. donovani, L. mexicana*) and *Leishmania Viannia spp.* (including *L. braziliensis, L. panamensis*) (Cupolillo, Grimaldi, and Momen 1994) (Figure 1.3) The divergence of *Leishmania spp.* and *Trypanosoma spp.* was placed in the region of 400-600 million years ago (Ghedin et al. 2004).

The advent of whole-genome sequencing projects of eukaryotic organisms in the 1990s, releasing the yeast genome in 1996 (Goffeau et al. 1996) and the human genome in 2001, did not pass trypanosomatid-research and sequencing projects were completed for the genomes of three trypanosomatid genomes, *T. brucei, T. cruzi* and *L. major* in 2005 (Berriman et al. 2005; El-Sayed, Myler, Bartholomeu, et al. 2005; Ivens et al. 2005). A full list of kinetoplastid genomes sequenced to date (September 2014) is shown in Table 1.3.

The completion of whole genome sequences for *Leishmania spp.* and related trypanosomatids was hoped to provide further insight into the gene content of these organisms and into their unusual mechanisms of gene expression. Leading on from these insights, predictions were made with regards to other aspects of kinetoplastid cell biology that would benefit from the completion of the genome projects (Gull 2001). Particularly the characterisation of gene-expression patterns within different hosts was viewed with anticipation, not only from the view point of the results, but also from the viewpoint of which methods will be the most informative considering the wealth of post-transcriptional processes in kinetoplastids (Gull, 2001).

Table 1.3 Sequenced kinetoplastid genomes A full list of kinetoplastid genomes sequences to date (September 2014), showing species, sequenced strain, size of the genome and, where available, the currently predicted gene number. The associated publication or laboratory providing the sequencing data is also given. Data source: TriTrypDB.com and Pubmed.gov

Species	Strain	Haploid genome size (Mbp)	Current number of predicted genes per haploid genome (incl. non-protein coding genes)	Associated publication / Data source
Crithidia fasciculata	Cf-Cl	40.29	11950	S.M. Beverly Lab
Crithidia mellificae	SF	32.8	9979	Runckel et al. 2014
Endotrypanum monterogeii	LV88	32.52	N/A	S.M. Beverly Lab
Leishmania aethiopica	L147	31.99	N/A	S.M. Beverly Lab
Leishmania amazonensis	MHOM/BR/71973/M2269	29.6	8100	Real et al. 2013
Leishmania arabica strain	LEM1108	31.44	N/A	S.M. Beverly Lab
Leishmania braziliensis	MHOM/BR/75/M2903	35.21	8966	S.M. Beverly Lab
Leishmania braziliensis	MHOM/BR/75/M2904	32.09	8505	Peacock et al. 2007
Leishmania donovani	BPK282A1	32.44	8195	Downing et al.
Leishmania enriettii	LEM3045	30.78	N/A	S.M. Beverly Lab
Leishmania gerbilli strain	LEM452	31.4	N/A	S.M. Beverly Lab
Leishmania infantum	JPCM5	32.13	8381	Peacock et al. 2007
Leishmania major	Friedlin	32.86	9378	Ivens et al. 2005
Leishmania major strain	LV39c5	32.33	N/A	S.M. Beverly Lab
Leishmania major strain	SD 75.1	31.24	N/A	S.M. Beverly Lab
Leishmania mexicana	MHOM/GT/2001/U1103	32.11	9063	Rogers et al.
Leishmania panamensis	MHOM/COL/81/L13	31.26	N/A	S.M. Beverly Lab
Leishmania sp.	MAR LEM2494	30.87	N/A	S.M. Beverly Lab
Leishmania tarentolae	Parrot-TarII	31.63	8530	Raymond et al. 2012
Leishmania tropica	L590	32.99	N/A	S.M. Beverly Lab
Leishmania turanica	LEM423	32.32	N/A	S.M. Beverly Lab
Phytomonas	HART1	18.1	6381	Porcel et al. 2014
Phytomonas	EM1	17.8	6451	Porcel et al. 2014
Trypanosoma brucei gambiense	DAL972	22.15	10000	Jackson et al. 2010
Trypanosoma brucei	Lister strain 427	26.75	9302	Becker et al. 2004
Trypanosoma brucei	TREU927	35.83	12094	Berriman et al. 2005
Trypanosoma congolense	IL3000	41.37	13358	Jackson et al. 2012
Trypanosoma cruz	JR cl. 4	41.48	N/A	G.A. Buck
Trypanosoma cruzi	CL Brener Esmeraldo-like	32.53	10600	El-Sayed et al. 2005 & Weatherley 2009
Trypanosoma cruzi	CL Brener Non-Esmeraldo-like	32.53	11109	El-Sayed et al. 2005 & Weatherley 2009
Trypanosoma cruzi	marinkellei strain B7	38.65	10282	Franzen et al. 2012
Trypanosoma cruzi	CL Brener	36.03	3397	El-Sayed et al. 2005
Trypanosoma cruzi	Sylvio X10/1	38.59	10947	Franzen et al. 2011 & 2012
Trypanosoma cruzi	Esmeraldo	38.08	N/A	G.A. Buck
Trypanosoma cruzi	Tula cl2	83.51	N/A	Hamilton et al. 2011
Trypanosoma evansi	strain STIB 805	25.43	10176	A. Schnaufer
Trypanosoma grayi	ANR4	20.95	10686	Manna et al. 2013
Trypanosoma vivax	Y486	47.5	12581	Pathogen Sequencing Unit - Wellcome Trust Sanger Institute

1.3 Gene Content of Leishmania

The completion and release of the *Leishmania major* genome (Ivens et al. 2005) as well as of *Trypanosoma brucei* (Berriman et al. 2005) and *Trypanosoma cruzi* (the "TriTryps" (El-Sayed, Myler, Blandin, et al. 2005)) made available a wealth of data allowing characterisation of each species on its own and in a comparative manner.

As indicated by older studies, the genomes of the TriTryps are characterised by a high degree of synteny (El-sayed 2005, Ghedin 2004). *L. major* and *T. brucei* were predicted to have a similar number of protein coding genes (8311 vs 9068 nt) respectively but protein coding sequences (mean 1731 vs 1511 nt) and especially intergenic sequences (mean 1431 vs 721 nt) in *L. major* were found to be larger than in *T. brucei* (El-Sayed, Myler, Blandin, et al. 2005) (Table 1.4).

The first *Leishmania* genome*, L. major,* was annotated using a variety of gene finding algorithms using e.g. codon usage (e.g. TESTCODE (Fickett 1982)) or Hidden-Markov models (e.g. GLIMMER (Salzberg et al. 1998)) and by identifying sequences with homologies in other organisms (e.g. using BLAST (Altschul et al. 1990)), most notably in *Trypanosoma brucei* (Ivens et al. 2005; Berriman et al. 2005).

Despite the strong synteny many differences between the TriTryps were identified, which for *L. major* consisted of a large expansion of amastins surface glycoprotein, ATP-binding cassette transporters and heat-shock proteins 90 (HSP 90) (El-Sayed, Myler, Blandin, et al. 2005). Conversely, *L. major* (and *T. cruzi*) e.g. lacked candidates for components of the RNA-interference (RNAi) pathway, most notably *Dicer* homologues, correlating with RNAi activity observed in *T. brucei* but not in the other TriTryps (El-Sayed, Myler, Blandin, et al. 2005).

Table 1.4 Comparison of TriTryp genomes Table showing a comparison of genome metrics from the original publication of the *T. brucei* , *T. cruzi* and *L. major* genome. This table was taken from El-Sayed, 2005.

	T. brucei	*T. cruzi*	*L. major*
Haploid genome size (Mbp)	25*	55	33
No. of chromosomes (per haploid genome)	11*	~28†	36
No. of genes (per haploid genome)	9068‡	~12000§	8311¥
Total regions with synteny blocks (Mbp)	19.9	NC	30.7
Mean CDS size (bp) in syntenic three-way clusters of othologous genes	1511	1457	1731
Mean inter-CDS size (bp) between syntenic three-way clusters of orthologous genes	721	561	1431

*Excluding ~100 mini- and intermediate-sized choromosomes (totalling ~10 Mb). †Exact number is not known and homologues can differ substantially in size. ‡ Includes 904 pseudogenes. § The exact number of haploid genes has not been determined in T. cruzi. ¥ Included 34 pseudogenes.

The completion of the *L. infantum* and *L. braziliensis* genome revealed that relatively few (~200) genes differed between species (Peacock et al. 2007), in turn suggesting that only very few genes may contribute to disease tropism observed between different species. Intriguingly, components of the RNAi pathway were discovered in *L. braziliensis* (Ives et al. 2011) which has been linked to the presence of an RNA-virus in *L. braziliensis* with important implications for disease severity (Lye et al. 2010).

The expansion of delta-amastins in *Leishmania spp.* (Jackson 2010) was implicated in interaction of the parasite with the mammalian host, not only by their amastigote-enriched expression (Wu et al. 2000) but also by their reduced number in the genome of *L. tarentolae* (Raymond et al. 2011), a species of *Leishmania* infecting lizards (Elwasila 1988), non-pathogenic to mammals (Breton et al. 2005).

Between TriTryp genomes evolutionary hotspots are concentrated at chromosome ends and strand-switch regions. Transcription start-sites are found, albeit not exclusively (Thomas et al. 2009), at divergent strand-switch regions, suggesting a link between transcription, DNA-replication and generation of genetic diversity (El-Sayed, Myler, Blandin, et al. 2005). Gene duplication, an important mechanism of generation of genetic diversity by allowing one copy to acquire new function due to relaxed functional constraints (Jackson 2007), has led to expansion of entire gene families in *Leishmania spp.*, and are common features at genomic sites where synteny between trypanosomatids species is broken (El-Sayed, Myler, Blandin, et al. 2005).

Mechanistically, abundant repeat sequences in *Leishmania spp.* genomes have been implicated to mediate duplications (and deletions) upon selective pressure in a Rad51-dependent manner (Ubeda 2014). Along with gene duplication, aneuploidy has now been proposed as a key feature for diversity between *Leishmania* species (M. B. Rogers et al. 2011). Investigation of the apparent mosaicism of ploidy lead to the identification of predominantly supernumerary chromosomes (Sterkers et al. 2011; Sterkers et al. 2012; Mannaert et al. 2012), most notably orthologues of LmjF.31, which was found to be supernumerary in every species and isolate tested (M. B. Rogers et al. 2011; Mannaert et al. 2012). Recently, aneuploidy of 8 chromosomes, along with amplification of sub-chromosomal regions, was shown in antimony-resistant *L. infantum* mutant (Brotherton et al. 2013). These changes may therefore also underlie drug-resistance in *Leishmania spp.*

The availability of whole-genome sequences has allowed further identification of genes without the use of homology searches or prediction algorithms. Moreover, some approaches have also permitted refinement of existing annotations, either by extending or truncating gene models.

Proteomic studies have sought to use unassigned mass-spectra in combination with genome sequences to confirm predicted or identify novel proteins. Proteomic analysis of *L. braziliensis* (Cuervo et al. 2007) identified 38 predicted or previously not annotated proteins by *de novo* sequencing of unassignable 2D-electrophoresis gel spots and Blast search against *L. braziliensis* data bases. By searching mass-spectrometric data against 6-frame translations of the *L. donovani* genome Nirujogi *et al.* (Nirujogi et al. 2014) were able to identify 20 proteins absent reference annotation and obtained peptide evidence for 40 protein extensions.

Recent developments in RNA-sequencing (RNA-seq) technology (Mortazavi et al. 2008) have allowed the investigation of transcriptomes at a single-nucleotide level. For kinetoplastids, this methodology was first employed for *T. brucei* (Siegel et al. 2010; Kolev et al. 2010; Nilsson et al. 2010) and allowed the identification of 1114 novel genes by mapping the positions of spliced-leader acceptor sites (SLAS) and polyadenylation sites (PAS) (Kolev et al. 2010).

In 2013 Rastrojo and co-workers (Rastrojo et al. 2013) used RNA-seq to determine the gene expression profile of *L. major* axenic promastigotes and from these data undertook an annotation of transcript boundaries in *L. major.* By combining sequencing read-coverage with trans-splicing and poly-adenylation data, the authors identified 1884 novel transcripts and proposed truncations to 410 genes due to splice-sites within annotated coding sequences. The resulting total number of genes in *L. major* and *T. brucei* is therefore now thought to lie at around 10 000 (c.f. (Rastrojo et al. 2013) and Table 1.3). Such transcriptomic studies offer genome-wide insight into the possible landscape of mRNAs, however unlike proteomic studies they do not offer evidence for translation of these transcripts. Proteomic studies however are limited by the drawbacks of detection of small or rare proteins. This is where the wide dynamic range of RNA-seq is powerful (5-6 orders of magnitude (Mortazavi et al. 2008), theoretically only limited by sequencing depth).

One way of overcoming drawbacks from both approaches may lie in ribosome-profiling as illustrated by a study on *T. brucei* (Vasquez et al. 2014). Here, segments of mRNA protected from RNase digestion by cycloheximide-stalled ribosomes are sequenced and mapped to the genome. The resulting single-nucleotide resolution data is able to inform about whether particular stretches of mRNA are being

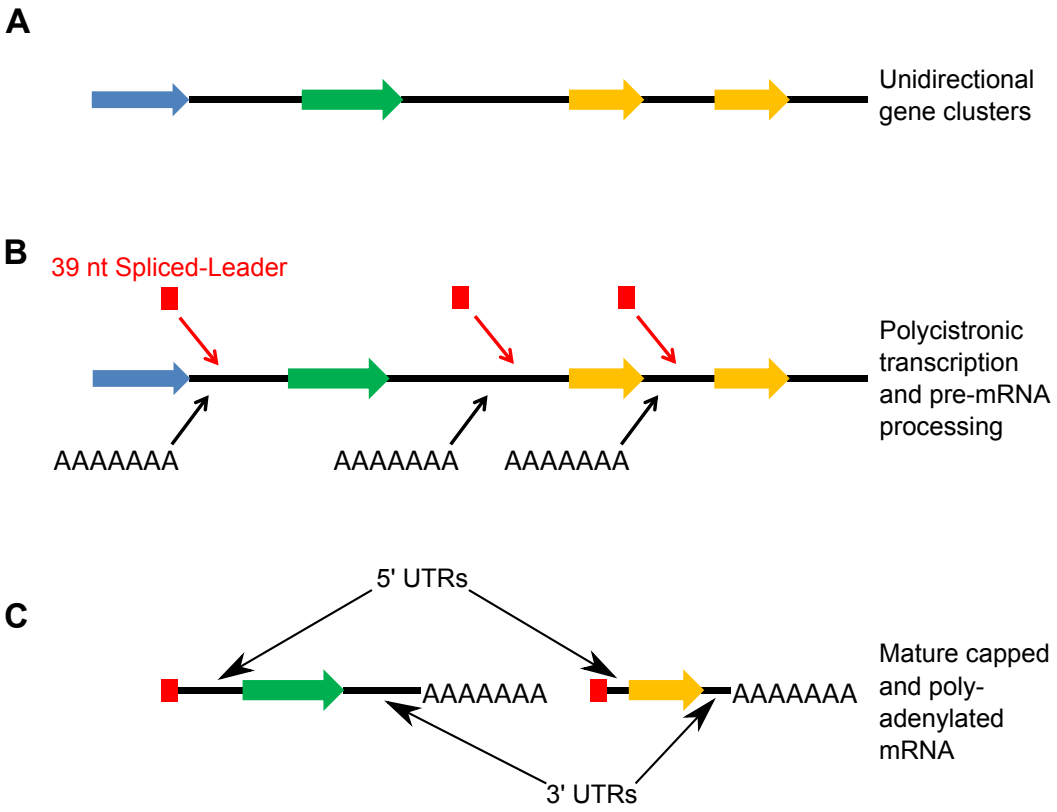


Figure 1.5 Transcription of protein coding transcripts in Kinetoplastids Schematic of the processed during transcription of mRNA in Kinetoplastids. Protein coding genes are organised in an unidirectional gene cluster (**A**) and are transcribed polycistronically. Trans-splicing of a capped 39 nt Spliced-Leader and poly-adenylation (**B**) leads to mature protein coding trancripts with 5' and 3' untranslated regions (UTRs) (**C**).

translated, whilst being able to detect such events over a RNA-seq-like dynamic range. Moreover, these data suggested extensive translation of small, upstream open reading frames (uORFs) (Vasquez et al. 2014) which may further contribute to the diversity of proteins within a cell. Without doubt, application of RNA-sequencing to other *Leishmania spp.* and cell types and ribosome-profiling to any *Leishmania spp.* will provide further insight into the gene content of these organisms.

1.4 Mechanisms of gene expression

Investigations in the 1980s and 1990s in *Trypanosoma brucei* and *Leishmania spp.* yielded insight into the, unusual transcriptomic processes in kinetoplastids (Figure 1.5). It was found that protein coding mRNAs are generated co-transcriptionally (Huang and Ploeg 1991) from poly-cistronic RNA-polymerase II-transcribed primary transcripts (Van der Ploeg 1986; Muhich and Boothroyd 1988) by 5' trans-splicing of a capped 39 nt spliced-leader (Lenardo, Dorfman, and Donelson 1985; Sutton and Boothroyd 1986) and 3' poly-adenylation to yield mono-cistronic mature mRNA. Splicing and poly-adenylation were found to be mechanistically and spatially coupled (LeBowitz et al. 1993), with the splicing reaction occurring first, and poly-adenylation of the message upstream occurring second at 400-500 nt distance (for *L. major*) to the spliced leader acceptor site (SLAS) (LeBowitz et al. 1993). An AG-dinucleotide was identified as a consensus SLAS (Sutton and Boothroyd 1986), whilst apparent redundancy of sequences forming the poly-adenylation site (PAS) was observed (LeBowitz et al. 1993). Unlike in other eukaryotic systems (Wickens 1990), no poly-adenylation sequence motif was found in kinetoplastids (Schürch et al. 1994) and only a poly-pyrimidine rich

element determined to lie between PAS and SLAS affecting site choice (Huang and Van der Ploeg 1991; Schürch et al. 1994; Matthews, Tschudi, and Ullu 1994).

The availability of the first chromosome-wide sequencing data for *L. major* revealed a dense genomic arrangement of genes, predicting around 9800 genes for the entire genome (Myler et al. 1999). For long stretches intron-less genes are arranged in unidirectional clusters (Donelson, Gardner, and El-Sayed 1999; Myler et al. 2000) e.g. amongst the initially reported 79 Genes of chromosome 1 of *L. major* only one strand-switch region is found. It was, amongst other functions, proposed that these strand-switch regions act as transcription start sites (Monnerat et al. 2004). Strikingly, chromosomal segments in *Leishamania spp.* and *Trypanosoma spp.* were found have highly conserved gene arrangements (synteny) (Ghedin et al. 2004).

Release of whole-genome sequences made feasible a range of investigations into mechanisms underlying gene expression in *Leishmania spp.* Importantly, genome wide information about chromatin modifications and RNA-processing sites have shed light onto fundamental processes such as transcription initiation, mechanisms underlying splicing and poly-adenylation as well as post-transcriptional regulation of gene expression as discussed below.

1.4.1 Transcription initiation

Whilst it was known that transcription in *Leishmania spp.* only initiates at very few sites in the genome (Monnerat et al. 2004), it was not known where these lie. Strand-switch regions had been implicated as transcription-start sites (Martínez-Calvillo et al. 2003), but it was only by mapping of acetylated H3-histone marks as well as binding patterns of the TATA-binding protein (essential for recruiting

transcription factors) and Small Nuclear Activating Protein complex (involved in initiating transcription of small nuclear RNAs) that a comprehensive view of transcription initiation sites in *L. major* was obtained (Thomas et al. 2009). As expected, these marks localised to divergent strand-switch regions. However, they were also found at chromosome ends and within poly-cistronic transcription units (Thomas et al. 2009). More recently, DNA curvature and secondary structure have been suggested to contribute to transcription initiation in *L. major* (Smircich et al. 2013). In turn, transcription termination sites have been found to be marked by the modified nucleotide glucosylated hydroxymethyluracil ("Base J"), which acts to prevent transcriptional read-through (van Luenen et al. 2012; Reynolds et al. 2014).

Investigations in *T. brucei* identified 4 histone variants enriched at both transcription start and termination sites (Siegel et al. 2009) and tri-methylated H3K4 at transcription start sites (Wright, Siegel, and Cross 2010), which may be molecular patterns that may be present in *Leishmania spp.* as well.

1.4.2 Mechanisms underlying splicing and polyadenylation

Traditionally, SLASs and PASs were determined painstakingly by sequencing reverse-transcribed splice-junctions (e.g. (Moore, Santrich, and LeBowitz 1996)) . The availability of single-nucleotide resolution transcriptomic data made it possible to determine SLAS and PAS on a genome wide level, first in *T. brucei* (Siegel et al. 2010; Kolev et al. 2010; Nilsson et al. 2010) and later in *L. major* (Rastrojo et al. 2013). This led to the definition of transcript dimensions in *T. brucei* and *L. major* and the definition of 5' and 3' UTRs in *T. brucei*. The mean 5' UTR dimensions in *T. brucei* determined in these studies agreed well having been

found to be 128 nt (Siegel et al. 2010), 130 nt (Kolev et al. 2010) and 140 nt (Nilsson et al. 2010) in size.

Between these studies, 16-22% of transcripts feature a potential uORF, based on presence of at least one ATG codon (Siegel et al. 2011), which have been shown to influence translational efficiency of the main CDS in trypanosomatids and are translated themselves (Vasquez et al. 2014). Which biological role proteins encoded by uORFs play, remains to be seen.

Furthermore, substantial alternative splicing was detected, some of which may be stage-regulated (Nilsson et al. 2010). The importance of such processes is exemplified by the differential splicing of the *T. brucei* isoleucyl-tRNA synthetase transcript, which has been shown to mediate dual localisation to the mitochondrion or cytosol (Rettig et al. 2012).

Mean 3' UTR sizes in *T. brucei* were determined as 388 nt (Kolev et al. 2010) and 400 nt (Siegel et al. 2010). Extensive heterogeneity exists amongst PAS and it is not inconceivable that alternative PAS choice may affect transcript stability in different life- or cell-cycle stages (Siegel et al. 2011).

Whilst SLAS and PAS data is available for *L. major* (Rastrojo et al. 2013), these data still require a comprehensive analysis to determined UTR dimensions. Curiously, as a result of SLAS and PAS mapping, Rastrojo *et al*. (Rastrojo et al. 2013) report the presence of poly-cistronic mature mRNA. These arose through assembly of transcripts based on read-coverage spanning several CDS which were subsequently not split by mapping of SLAS and PAS. It will be interesting to see whether other lines of evidence will be able to confirm the existence of these

mRNA species and what biological role polycistronic molecules may play in kinetoplastids.

The generation of SLAS and PAS position data bears promises for the identification of mechanisms that underlie processing site-choice. And whilst to date no features resembling the AAUAAA polyadenylation signal found in many eukaryotic systems have been identified in kinetoplastids, these data have been used to support predictions made by statistical methods implicating secondary RNA-structure elements in SLAS and PAS delineation (Kelly et al. 2011).

With more transcriptomic data sets emerging and methods for mapping and quantifying SLAS and PAS becoming more user-friendly (Fiebig et al. 2014), I expect future comparative studies of SLAS and PAS patterns between species to shed further light onto the mechanisms underlying and affected by splicing and polyadenylation.

1.4.3 Post-transcriptional regulation of gene expression

Polycistronic transcription in kinetoplastids precludes regulation of gene expression at the level of transcription initiation. The availability of genome sequences provided insights into gene order in kinetoplastids. In *T.* brucei, and unlike in prokaryotes (Monod et al. 1960), genes within an single polycistronic transcription clusters are not thought to be functionally linked (S Kelly et al. 2012), and very little to no co-regulation of adjacent genes have been observed for *Leishmania spp*. genes (A. Saxena et al. 2007). As a result, post-transcriptional mechanisms are thought to lie at the core of gene-expression regulation in

kinetoplastids, most importantly differential RNA-stability/degradation and translational efficiency (extensively reviewed in (Clayton 2014)).

In the pre-genomic era, some insights into mechanisms of differential gene expression regulation were gained by identification of sequences in the 3' (Charest, Zhang, and Matlashewski 1996; Brooks et al. 2001; Boucher et al. 2002; Mishra et al. 2003) or 5' UTRs (Teixeira, Kirchhoff, and Donelson 1999) that conferred stage-specific expression of transcripts. Moreover, both transcript abundance and translational efficiency were found to be regulated by elements in the 3' UTR of heat-shock protein (HSP) 83 transcripts.

A genome-wide search of the *L. major* genome (Holzer et al. 2008) revealed that the PRE-element, initially identified in the 3' UTRs of paraflagellar rod (PFR) 2 gene in *L. mexicana* (Mishra et al. 2003), was not only distributed amongst other PFR genes, but also found amongst other promastigote-enriched transcripts (Holzer et al. 2008). Also sequences mediating amastigote-enriched transcript abundances have been identified such as in the 3' UTR of the A600-4 transcript (A. Murray et al. 2007).

More recently, SIDER 2 retrotransposon elements have been implicated in wide-scale post-transcriptional regulation of gene expression in *Leishmania spp*. SIDER elements are heavily enriched in *Leishmania* compared to *T. brucei* (Bringaud et al. 2007) and almost exclusively localise to 3' UTRs. In particular SIDER 2 elements were shown to mediate mRNA instability (Bringaud et al. 2007) via an unusual de-adenylation independent pathway (Müller et al. 2010). Furthermore, protein factors like the Piwi-like protein (Padmanabhan et al. 2012) or members of the Alba protein group (Dupé, Dumas, and Papadopoulou 2014) have emerged as

important genome-wide post-transcriptional effectors of stage-regulated mRNA abundance. Intriguingly, translation has been shown to stabilise transcripts, as shown for the A2-transcript (Cloutier et al. 2012), where transcripts are stabilised through association with polysomes.

In the 3' UTRs of amastin transcripts, a ~100 nt and a ~450 nt element were identified, which upon heat-shock was able to stimulate translation of a reporter gene (McNicoll et al. 2005). The accumulation of amastin transcripts was shown to be independent of said sequence elements, showing separation of transcript abundance- and translational control (McNicoll et al. 2005).

The mechanism by which HSP83 transcripts undergo translational regulation was further elucidated by identification of a thermosensitive secondary-structure element in the 3'UTR encompassing a poly-pyrimidine rich tract (David et al. 2010). Translational control on a global scale has received much attention following findings that during differentiation of promastigotes to amastigotes, global translation is suppressed (Rosenzweig, Smith, Opperdoes, et al. 2008; Lahav et al. 2011). This has been linked to the phosphorylation state of the translational initiation factor eIF2alpha (Cloutier et al. 2012) mediated by the PERK eIF2alpha kinase (Chow et al. 2011). On a transcript-by-transcript basis, findings in *T. brucei* show that translational efficiency is strongly affected by the presence of uORF in transcripts (Vasquez et al. 2014) and it is probable that these mechanisms will extend to *Leishmania spp.* as well.

1.5 Insights into Differentiation

The transition of *Leishmania spp.* promastigotes to amastigotes is not only a crucial step during the completion of the parasite's life-cycle, but also a truly fascinating

process that allows the parasite to establish infection within macrophages and persist within the hostile environment of the phagolysosome. The gross morphological differences between promastigotes and amastigotes, i.e. presence and absence of a motile flagellum respectively, have been known for over a century (L. Rogers 1904), but only a few molecular adaptations *Leishmania spp.* undergo during promastigote-to-amastigote differentiation. These included elucidation of metabolic differences, such as the shift from glycolysis in promastigotes to fatty-acid metabolism in amastigotes (Coombs, Craft, and Hart 1982), differential expression of cytoskeletal components (Fong et al. 1984; Bellatin et al. 2002) and loss of expression flagellar components in amastigotes such as the paraflagellar rod (Moore, Santrich, and LeBowitz 1996) as well as differential expression of surface markers between promastigotes and amastigotes (Chang and Fong 1982; Medina-Acosta et al. 1989; Charest, Zhang, and Matlashewski 1996; Wu et al. 2000). Examples of these studies will be discussed in more detail in Section 1.5.1. These investigations focussed on single genes or gene-families, however global molecular adaptations permitting the amastigote life-style had largely remained elusive. The availability of whole genome sequences made possible advances in this field, primarily spearheaded by array technologies and proteomic approaches. In recent years however, studies on *T.* brucei, have exemplified the contribution RNA-sequencing technologies can make to our understanding of different kinetoplastids life-cycle stages (Siegel et al. 2010; Nilsson et al. 2010; Kolev et al. 2012) and, with variations such as ribosome-profiling (Vasquez et al. 2014), are likely to shed unprecedented insight into the differentiation processes in *Leishmania spp.* as well.

1.5.1 Pre-genomic transcriptomics

The identification of differentially expressed transcripts was a challenge overcome by a variety of different approaches. Differential expression of β-tubulin variants between promastigotes, expressing a single species, and amastigotes, which express three species, was discovered using tubulin probes derived from chicken cDNA on Northern Blots (Fong et al. 1984). Similarly, the differentially expressed ATPase 1b (Meade et al. 1989) was discovered using such an approach. Prompted by investigations into glucose and amino-acid metabolism of amastigotes, the authors (Meade et al. 1989) used a conserved sequence found in ATPases of lower and higher eukaryotes (Meade et al. 1987) to design Northern Blot probes and identify differentially and constitutively expressed ATPases (Meade et al. 1989). Whilst elegant, the distant evolutionary relationship of common eukaryotic model organisms and *Leishmania spp.* places limits on the fields of research that may be investigated in such a way.

Antibodies had proven to be an effective way of identifying differentially expressed proteins, so methods were devised permitting the identification of the gene encoding the epitope recognised by particular antibodies. Moore *et al.* (Moore, Santrich, and LeBowitz 1996) generated poly-clonal antibody-sera in mice using flagellar protein preparations, depleted these of antibodies recognising epitopes common to promastigotes and amastigotes by negatively selecting using amastigote protein samples. Subsequently, these antibodies were used to screen a phage-library prepared from promastigote-derived cDNA. This permitting identification of nucleotide sequences encoding the recognised epitopes, allowing cloning and sequencing of the whole genes along with the design of Northern Blot

probes (Moore, Santrich, and LeBowitz 1996). This work showed the promastigote specific expression of PFR-2 in *L. mexicana.*

Genome wide screens using differential cDNA hybridisation methodology permitted identification of amastigote specific genes such as the A2-genes in *L. donovani* (Charest and Matlashewski 1994), which served as amastigote-specific markers in subsequent studies (e.g. (Saar et al. 1998; Barak et al. 2005)).

Bellatin *et al.* (Bellatin et al. 2002) employed a selective-suppression PCR method, allowing hybridisation of promastigote- and amastigote-derived single-stranded cDNA with each other, and subsequent selective PCR-amplification of remaining single-stranded species to identify the A600 gene and the amastigote specific A-850 β-tubulin.

Genome-wide investigations into the transcriptome of *Leishmania spp.* using micro-array technology commenced in the early 2000s, even before the completion of the *L. major* genome (Ivens et al. 2005). The lack of genome sequences to design single gene-based hybridisation platforms from, required researchers to find alternative solutions.

In the beginning, researchers created arrays using random mechanically sheared genomic fragments of *L. major* constituting the Genome Survey Sequences (GSS) (Akopyants et al. 2001), calculated to cover around 2/3 of the *L. major* genome. These sequences were first used in studies to compare the transcriptomes of *L. major* promastigotes and metacyclics (Saxena et al. 2003) and soon after in the first comparison of *L. major* axenic promastigotes (PRO), metacyclics (META) and intracellular amastigotes (AMA). In pairwise comparisons, transcripts hybridising with array spots were deemed differentially expressed if they showed a >2-fold

changed signal in at least one of the three replicates. Using these criteria, 3.1 % of transcripts were differentially abundant between PRO and AMA, and 3.2% between PRO and META. These GSS sequences however, originated from randomly sheared DNA and therefore did not *per se* correspond to any particular transcript and may indeed correspond to fragments of several transcripts. To counteract this, other arrays were designed using randomly selected cDNA-probes obtained from *L. major* PRO and AMA mRNA as well as PCR products from known open-reading frames (Almeida et al. 2004). These probes corresponded to 1001 and 842 unique genes for cDNA- and ORF-based probes respectively. Comparing PRO and AMA transcriptomes, 35 % of genes were found to be differentially expressed based on a statistical cut-offs ($p < 0.05$, NB: not multiple-testing corrected).

1.5.2 Micro-arrays in the post-genomic era

The release of the *L. major* genome (Ivens et al. 2005) paved the way for genome-annotation based probe design. In 2006 Holzer and co-workers (Holzer, McMaster, and Forney 2006) utilised 60-mer nucleotide-probes based on *L. major* genes (11 per gene) for their microarrays. Interestingly, the authors did not investigate the transcriptome of *L. major*, but investigated PRO, AMA and 3-4 d axenic amastigotes (AXA) of *L. mexicana*, relying on the close evolutionary relationship between *Leishmania Leishmania spp.* (e.g. (as argued by the authors) 91-96 % nucleotide identity for proteins coding genes between *L. major* and *L. donovani* (Myler et al. 2001)). This was also the first study to address the similarity of AMA and AXA, latter ones having been a popular surrogate model for AMA since the early 1990s (P. A. Bates et al. 1992). By defining differentially expressed genes based on both statistical cut-offs ($p <= 0.05$, multiple-testing corrected (M.T.C)) and z-ratio (+/- 1.96), 3.5 % of genes were predicted to be differentially expressed between PRO

and AMA, 2.1 % between AXA and AMA and 0.2% between PRO and AXA. The discrepancy to Almeida *et al.* (Almeida et al. 2004) was in parts put down to the lack of multiple-testing correction in the differential expression-calling and in parts to the possibility of skewing of the hybridisation probes during cDNA preparation and selection towards highly and differentially expressed genes (Holzer, McMaster, and Forney 2006). The design of the array also permitted functional analysis of the observed transcriptomic changes. As expected, components of the motile flagellum were preferentially expressed in PRO. Perhaps surprisingly, components of the translational machinery and histones were also found preferentially expressed in PRO.

Strikingly, AXA and AMA differed substantially and the transcriptome of AXA rather resembled that of PRO, casting doubt over the utility of axenic amastigote forms in amastigote-research.

Taking the GSS-probe based approach to the post-genomic era, Saxena *et al.* (A. Saxena et al. 2007) used *L. major* GSS-sequences to investigate *L. donovani* transcriptomes during axenic differentiation from PRO to AXA, and assigned GSS sequences to genes in the *L major* genome to obtain functional information about the differentiation process. Defining differentially expressed genes based on statistical cut-offs ($p<=0.05$, M.T.C.), the authors found 1.6% of genes to be differentially expressed after 24 h, albeit throughout the differentiation time-course (5,10, 24 h post-differentiation initiation) more than twice as many genes were likely differentially expressed. 12 different expression profiles were detected, indicating ordered progression of transcriptomic changes as well as transient roles for certain transcripts, such as a heat-shock proteins, protein kinases and a histone deacetylase. The authors moreover concede that this may still be underestimating

the total number of differentially expressed genes as only around half of all *L. major* genes were covered by the probe-library and probes may still correspond to multiple genes. Gene Ontology analyses pointed to over-representation of genes associated with cell growth and motility amongst promastigote enriched transcripts, whilst pointing to under-representation of genes associated with protein metabolic functions and enrichment of transporters amongst AXA-enriched transcripts.

Using arrays based on *L. major* genes (eleven 24-mers per gene) Leifso *et al.* (Leifso et al. 2007) compared *L. major* PRO and AMA, and, using the definition of differentially expressed genes changing expression level by at least 2-fold and satisfying a statistical cut-off ($p<=0.05$, M.T.C.), 1.4 % and 1.5 % of genes were identified as preferentially expressed in AMA and PRO respectively. Due to a lack of an axenic amastigote system for *L. major*, *L. infantum* cells were chosen for the proteomic study. Employing ICAT-isotope labelling (Gygi et al. 1999) and using 2D-liquid chromatography fractionated samples following detection on two different mass-spectrometers, 91 proteins were identified and their differential expression quantified. Of these 91 proteins, 8% and 20% were determined as differentially expressed in AXA and PRO respectively. Sadly, no comparisons between transcriptomic and proteomic results were performed. Moreover, from these findings, the authors postulated that the *Leishmania* genome is constitutively expressed and constitutively adapted to utilising the nutrients available in either sandfly or host.

When Rochette *et al*. (Rochette et al. 2008) compared PRO and AMA of *L. infantum* and *L. major* on a mixed-species chip comprised of genome-based 70-mer probes,

7 % and 9.3 % of genes were differentially expressed in *L. infantum* and *L. major* respectively (min. 1.7-fold change, p<=0.05) . Remarkably, only 10-12 % of differentially expressed genes were orthologues between the two species, pointing to substantial differences between species. Whilst it could be argued that differences between amastigote cell types could originate from different culture systems (lesion vs. THP-1 cell line derived), even the promastigote cells, cultured under the same conditions, differed substantially. That notwithstanding, functional analyses of the differentially expressed genes revealed enrichment of transcripts coding for components of the motile flagellum as well as proteins involved in glucose metabolism (enzymes, glucose transporters) in PRO. In AMA, particular enrichment for membrane transporters was seen.

Using the same approach, Rochette and co-workers (Rochette et al. 2009) furthermore showed that 12.5% and 7.1 % of *L. infantum* genes are differentially expressed between AXA versus PRO and AMA versus PRO, respectively. Not only did this unexpectedly suggest that AXA were more different to PRO than AMA, but it also showed that there is little overlap (90 genes) between the genes preferentially expressed in AXA (518) and AMA (309) compared to PRO. Moreover, fairly little overlap with previous studies of genes preferentially expressed in *L. major* AMA (Rochette et al. 2008) and *L. donovani* AXA (A. Saxena et al. 2007) was seen and none with *L. mexicana* (Holzer, McMaster, and Forney 2006), albeit the authors do concede that comparisons between studies are technically difficult (Rochette et al. 2009). Functional analysis of genes preferentially expressed in AXA and AMA showed that, even though both are enriched for genes involved in beta-oxidation, enrichment for genes involved in fatty-acid elongation is only seen in AXA. This was put down to the differential availability of fatty acids in the growth

medium and raised the notion that AXA may be "stuck" in an early stage of differentiation.

The apparent differences between *Leishmania spp.* received further attention in a study by Depledge and co-workers (Depledge et al. 2009), who used a micro-array comprising genes differentially distributed between *L. major*, *L. infantum* and *L. braziliensis*, as well as a selection of genes with predicted amino-acid repeats, predicted N-myristoylation sites and a cohort of control genes. PRO, META and AMA samples of the three species were compared. Remarkably, genes differentially distributed between species are predominantly constitutively expressed, whilst conserved genes were more likely to be differentially expressed. In addition, comparison of *L. major* AMA derived from foot-pad lesions of wild-type and Rag $2^{-/-}$ $\gamma_c^{-/-}$ mice showed only 3 of 700 investigated genes to be differentially expressed, suggesting that the immunological state of the host has only minor effect on the transcriptome of the parasite.

To discern which of the two triggers employed in axenic systems to induce amastigogenesis (low pH, elevated temperature) plays a dominant role, Alcolea *et al.* (Alcolea et al. 2010) compared temperature- or pH-shocked *L. infantum* AXA to cells exposed to both signals. The gene expression profile obtained from temperature-shock resembled that of cells exposed to both signals whilst pH-shock alone was a poor differentiation signal.

Microarray studies have permitted genome-wide insight into the gene-expression changes during promastigote-to-amastigote differentiation. These changes follow a variety of patterns, but appear to follow a strict chronology (A. Saxena et al. 2007). The precise number of genes differentially expressed between promastigotes and

amastigotes is a contentious topic and may lie anywhere between 3.1 % (Saxena et al. 2003) and 35 % (Almeida et al. 2004). Differential transcript abundances point to metabolic differences between promastigotes and amastigotes, with amastigotes favouring beta-oxidation over glycolysis (Rochette et al. 2009). Substantial differences between axenic and intracellular amastigotes (Holzer, McMaster, and Forney 2006; Rochette et al. 2009) put in question the utility of axenic amastigotes as a robust model for amastigote biology. Temperature shock has emerged as a dominant signal in the differentiation of promastigotes to amastigotes (Alcolea et al. 2010), whilst properties of the host-organism, such as its immunological state, appeared to have little effect on gene expression patterns of amastigotes (Depledge et al. 2009). For list a of post-genomic transcriptomic studies (restricted to microarrays) focussing on different *Leishmania spp.* cell types please refer to Table 1.5.

1.5.3 Proteomics in the pre-genomics era

Proteomic studies open the possibility of investigating the protein composition of a cell on a large scale. In particular differentially expressed proteins have been at the focus of proteomic studies. Using sera of immunised animals and monoclonal antibodies, a range of stage-specific proteins were identified in the 1980s and 1990s (Chang and Fong 1982; Handman, Jarvis, and Mitchell 1984; Pan 1986; Pan and McMahon-Pratt 1988; Eperon and McMahon-Pratt 1989; Jaffe and Rachamim 1989). Whilst these studies did not elucidate the identity of the antigen recognised by the antibodies used, monoclonal antibodies permitted detection of expression of amastigote specific markers during culture of *L. major* and *L. mexicana* promastigotes under acidic conditions, thereby giving first insight into the signal leading to differentiation of promastigotes to amastigotes (Zilberstein et al. 1991).

The growing number of genomic and proteomic databases in the 1990s and early 2000s made possible the application of mass-spectrometry to the investigation of *Leishmania spp.* proteomes. Even before the release of complete *Leishmania spp.* genomes, mass-spectrometry-based proteomic studies were undertaken, albeit often hampered by the inability to determine the identity of differentially expressed protein. The first study comparing on *L. infantum* PRO and AXA (El Fakhry, Ouellette, and Papadopoulou 2002) proteomes, was able to detect around 2000 individual protein spots on 2D-electrophoretic gels, of which 62 were differentially expressed. However, only 2 amastigote specific proteins were identifiable in a search against protein sequence data bases. These were both carbon-metabolic proteins, the TCA-cycle component isocitrate dehydrogenase and the glycolytic enzyme triosephosphate dehydrogenase. Based on the identification of the latter, the authors drew metabolic parallels between *Leishmania* amastigotes and the blood-stream form of *T. brucei*, which had been shown to be reliant on glycolysis for energy production (Bursell et al. 1973; Balogun 1974; Wurst et al. 2012).

Bente and co-workers (Bente et al. 2003) used a similar approach as El Fakhry *et al.* (El Fakhry, Ouellette, and Papadopoulou 2002), however searched peptide matches against a data-base including protein sequences from *L. major* and other kinetoplastids to improve identification of protein sequences. Notably, similar protein expression patterns were observed when *L. donovani* AXA cells were compared to PRO treated with an inhibitor of heat shock protein (HSP) 90, suggesting a role for HSP90 in differentiation.

A subsequent study on *L. mexicana* PRO, META and AXA (Nugent et al. 2004) also

Table 1.5 Transcriptomic and proteomic studies of *Leishmania spp.* cell types and differentiation Table summarising post-genomic transcriptomic (light grey highlight) and proteomic studies (darker grey) of *Leishmania spp.* cell types and differentiation processes. For each study the species and cell types analysed are shown, along with the method used to generate these comparisons, i.e. for microarray studies the composition of the array, for proteomic studies the quantitation method for protein abundances (label-free if blank). GSS= Genome Survey Sequences (Akopyants*et al.* 2001), iTRAQ= Isobaric tag for realtive and absolute quantitation (Ross *et al.* 2004)

	Authors	Date	Species	Cell Type	Comparison method	Additional note
TRANSCRIPTOMIC	Akopyants *et al.*	2001	*L. major*		GSS-chips	
	Saxena *et al.*	2003	*L. major*	PRO, META	GSS-chips	
	Akopyants *et al.*	2004	*L. major*	PRO, AMA, META	GSS-chips	
	Almeida *et al.*	2004	*L. major*	PRO, AMA, META	cDNAs, selected ORFs	
	Holzer *et al.*	2006	*L. mexicana*	PRO, AMA, AXA	*L. major* 60-mer probes (11/gene)	
	Saxena *et al.*	2006	*L. donovani*	PRO to AXA differentiation	Random *L. donovani* DNA fragments	
	Leifso *et al.*	2006	*L. major*	PRO, AMA	*L. major* 24-mer probes (11/gene)	
	Srividya *et al.*	2007	*L. donovani*	AMA to PRO differentiation	Random *L. donovani* DNA fragments	
	Rochette *et al.*	2008	*L. donovani* *L. major*	PRO, AMA	70-mer genome array of both *L. donovani* and *L. major*	
	Rochette *et al.*	2009	*L. infantum*	PRO, AMA, AXA	70-mer genome array of both *L. donovani* and *L. major*	
	Alcolea *et al.*	2009	*L. infantum*	PRO, META	*L. infantum* shotgun sequences	
	Depledge *et al.*	2009	*L. major,* *L. infantum,* *L. braziliensis*	PRO, AMA, AXA, META	Selected genes from three species	
PROTEOMIC	El Fakhry *et al.*	2002	*L. infantum*	PRO, AXA	2D-gel	
	Bente *et al.*	2003	*L. donovani*	PRO, AXA	2D-gel	
	Nugent *et al.*	2004	*L. mexicana*	PRO, AXA, META	2D-gel	
	McNicoll *et al.*	2006	*L. infantum*	PRO, AXA	2D-gel	also microarray
	Morales *et al.*	2007				Phosphoproteomics
	Rosensweig *et al.*	2008	*L. donovani*	PRO to AXA differentiation	iTRAQ	
	Rosenzweig et al.	2008	*L. donovani*	PRO to AXA differentiation		Phosphorylation, methylation, acetylation and glycosylation proteomics
	Paape *et al.*	2008	*L. mexicana*	PRO, AMA	2D-Gel	
	Paape *et al.*	2010	*L. mexicana*	PRO, AMA		
	Hem *et al.*	2010	*L. donovani*	PRO to AXA differentiation		Phosphoproteomics
	Lahav *et al.*	2011	*L. donovani*	PRO to AXA differentiation	iTRAQ	also microarray
	Tsigankov *et al.*	2014	*L.donovani*	PRO to AXA differentiation		Phosphoproteomics

detected around 2000 individual 2D-gel protein spots, of which 147 were differentially abundant. The proteins preferentially expressed in AXA featured components of the translational machinery, chaperones and HSPs, glycolytic enzymes, cysteine proteinases and various tubulins.

1.5.4 Proteomics in the post-genomic era

Following the publication of the genome and advances isotope-based quantification methods permitting gel-free sample preparation along with improving mass-spectrometers, more holistic insights into the proteomic changes of the parasite during differentiation could be gained. (For list of post-genomic proteomic studies (restricted to mass-spectrometric studies) focussing on different *Leishmania spp.* cell types please refer to Table 1.5.)

By using iTRAQ isobaric labelling (Ross et al. 2004) and 2D liquid chromatography to pre-fractionate samples, Rosenzweig *et al.* (Rosenzweig, Smith, Opperdoes, et al. 2008) followed *L. donovani* axenic differentiation over 7 time-points and were able to obtain protein expression data for 21 % of the annotated proteome (1713 genes). Apart from undergoing well-characterised morphological changes during differentiation, parasite cells shifted from a metabolic signature characterised by glycolysis, to one characterised by beta-oxidation and amino-acid metabolism. The TCA-cycle appears more active in amastigotes, who are furthermore characterised by a down-regulation of translation. These results stand at odds with previous reports (El Fakhry, Ouellette, and Papadopoulou 2002), who proposed that amastigotes were relying on glycolysis for energy production. However the glucose-poor and environment of the phagolysosome and potentially glucose-rich

mid-gut of the sandfly following a blood-meal, would speak for a shift away from glycolysis in PRO to beta-oxidation in AXA.

All the above mentioned proteomic investigations of amastigotes had focussed on axenically differentiated amastigotes. Through the development of an amastigote purification method, in which host-cells were mechanically disrupted and fluorescent transgenic *L. mexicana* amastigotes sorted in a flow-cytometer, Paape and co-workers (Paape et al. 2008) were at first able to detect only around 6% (509) of the predicted *L. mexicana* proteome using 2D protein gels and the *L. major* genomic data base (NB: The *L. mexicana* genome was not published until 2011 (M. B. Rogers et al. 2011)). Of these 509 identified proteins, 34 were preferentially expressed in AMA. Subsequently, using a gel-free sample-fractionation method, label-free quantification, and a combined *L. major, L. infantum* and *L. braziliensis* data base, 1765 proteins were identified (Paape et al. 2010). Functional analysis of differentially expressed proteins supported the switch of parasites towards beta-oxidation in the amastigote stage.

With both proteomic and transcriptomic approaches providing insight in to the biology of differentiation in *Leishmania spp.*, the question posed to what degree these approaches agree and what further biological insights this may provide. Nicoll *et al.* (McNicoll et al. 2006) combined a 2D-protein gel electrophoresis and micro-array approach to compare PRO and AXA of *L. infantum.* 145 differentially expressed proteins were identified. For proteins preferentially expressed in AXA, the direction of transcript regulation agreed in around 50 % of cases, however the fold changes of the transcript were more subtle than for proteins. For PRO, correlation of transcript and protein abundance was particularly poor. These

findings pointed to extensive and significant translational regulation of protein expression. Using iTRAQ isobaric-labelling and 2D liquid-chromatographic sample fractionation, Lahav and co-workers (Lahav et al. 2011) followed protein and transcript levels of 902 genes during axenic *L. donovani* differentiation. During differentiation, around 1/3 of all mRNAs and proteins changed, with the best correlation of changes of protein and mRNA levels seen at 5-10 h into differentiation. At later timepoints, the correlation becomes poorer. This suggests that regulation of transcript abundances is a main regulatory factor in the early stages of differentiation, whilst translational regulation dominates at later stages.

Functionally, an initial upregulation of translational activity is seen, followed by down-regulation on both mRNA and protein level. The proposed reduction of protein-production was furthermore supported by ^{35}S-methionine incorporation assays and polysome-size assays as well as detection of phosphorylation of the translational initiation factor eIF2alpha. Curiously, especially trypanosomatid-specific genes showed weak correlation between mRNA and protein levels (Lahav et al. 2011).

Proteomic studies have also permitted insight into the landscape of post-translational modifications and the roles these play during differentiation. Morales *et al.* (Morales et al. 2008) performed a first survey of changes of protein phosphorylation from PRO to AXA in *L. donovani* using a 2D-protein gel system, identifying particularly HSPs and ribosomal proteins as targets for stage-specific phosphorylation. The phosphorylation of the former was followed up (Morales et al. 2010) leading to the identification of the functional importance of dynamic phosphorylation sites in the molecular cochaperone STI1 (Webb et al. 1997),

bearing an 8-fold difference in phosphorylation between *L. donovani* PRO and AXA. Two of the three phosphorylation sites are essential to the parasite, illustrating the importance of protein chaperone-phosphorylation to parasite biology.

Further advances in the breadth of insight into post-translational modifications were gained by employing liquid-chromatographic prior to mass-spectrometric analysis. Phosphorylation was found to be a dynamic process during differentiation, investigated in *L. donovani* PRO to AXA differentiation (Rosenzweig, Smith, Myler, et al. 2008). Moreover, recent studies (Tsigankov et al. 2014) have sought to discern the contributions the main differentiation triggers (low pH and elevated temperature) individually and in combination make to the spectrum of phosphorylations observed. Phosphorylation, however, was not the sole post-translational modification identified. Methylation, acetylation and glycosylation have all been shown to be dynamically affected during differentiation (Rosenzweig, Smith, Myler, et al. 2008) and it will be interesting to see which contributions these make to the biology of the amastigote.

Proteomic studies have permitted insight into global proteomic changes during promastigote-to-amastigote differentiation (Rosenzweig, Smith, Opperdoes, et al. 2008). These may be in agreement with, but often have very little correlation with transcriptomic changes, due to extensive translational control of gene expression during differentiation (Lahav et al. 2011). Protein phosphorylation plays a crucial role in control of translation (Lahav et al. 2011), but, along with other post-translational modifications, is likely to form a network controlling other cellular activities during differentiation as well (Morales et al. 2008; Morales et al. 2010; Rosenzweig, Smith, Myler, et al. 2008; Tsigankov et al. 2014).

1.5.5 The contribution of sequencing-based methods

Recent advances in RNA-sequencing technology, such as the application of massive parallel nucleotide sequencing technology to cDNA libraries (Mortazavi et al. 2008), allow quantification of transcript levels over a wider dynamic range than microarrays are able to, with RNAseq with >5 orders of magnitude compared to microarrays at around 3 orders of magnitude (c.f. (A. Saxena et al. 2007; Siegel et al. 2010; Kolev et al. 2010; S. Zhao et al. 2014)). This promises the identification of further differentially expressed genes, which, along with ever-improving genome annotation, will permit further functional insight into *Leishmania spp.* biology. Moreover, the single-nucleotide resolution of these methods, will permit delineation of transcripts and together with differential expression data may lead to discovery of further regulatory elements contained within transcripts.

The first studies employing RNA-sequencing to understand *Leishmania spp.* differentiation have been published (Mittra et al. 2013; Martin et al. 2014). Iron depletion was shown to trigger differentiation of *L. amazonensis* promastigotes into amastigotes in a process dependent on the ferrous iron transporter LIT1 and reactive oxygen species produced by iron superoxide dismutase (Mittra et al. 2013). To determine whether the observed amastigote-like morphological changes corresponded to adaptation of amastigote-transcriptomic pattern, the authors utilised RNA-sequencing of a cDNA library generated using random hexamer primers in the first round, and in the second round primers complementary to the spliced-leader sequences, to enrich for spliced-leader-containing reads (SL-RNAseq) thereby confirming expression of control genes. Using this method similarities between pH and temperature versus iron starvation mediated axenic differentiation were uncovered, albeit differences persist which may point to

distinct pathways inducing differentiation. Using a combination of liquid-chromatography mass-spectrometry and SL-RNAseq, purine starvation in *L. donovani* was found to trigger profound remodelling of the cell-proteome, in a process dependent on transcriptomic regulation, affected by both UTRs and coding sequences of purine permeases and components of the purine pathway, as well as translational regulation (Martin et al. 2014). Moreover, findings in this study suggest that purine starvation may be a partial trigger for metacyclogenesis.

1.6 Conclusions

Fundamentally, the completion of *Leishmania spp.,* to date, genomes has permitted systematic correlation of transcriptomic or proteomic data with genes in the genome. Prior to this, e.g. microarrays could elucidate transcriptomic patterns, but correlating a particular spot with a specific gene could prove difficult or indeed impossible. Similarly for proteins, differential abundance of a protein could be detected electrophoresis gels, but the identity of the protein could elude researchers. The availability of genome sequences have made possible a wealth of transcriptomic and proteomic studies (Table 1.5), where findings are linked to standardised genomic identifiers, allowing effective correlation of studies between laboratories.

Microarray studies have proven to be able to give wide-ranging insights into transcriptomic patters, and with these data permit enrichment analyses to characterise cell types. Whilst proteomic studies overcome the uncertainty of how transcriptomic patterns truly correlate with the protein composition of a cell, the low number of proteins identified using 2D-electrophoresis methods may mislead researchers trying to integrate their findings into a wider biological picture (c.f. glycolysis in amastigotes (El Fakhry, Ouellette, and Papadopoulou 2002)).

Advances through development isotope-labelling, and label-free quantification methods, permitting gel-free sample comparison increased the number of proteins identifiable by mass-spectrometry from hundreds (McNicoll et al. 2006; Paape et al. 2008) to thousands (Paape et al. 2008; Rosenzweig, Smith, Opperdoes, et al. 2008) and yielded results comparable and complementary to micro-array results (Lahav et al. 2011). Notable is the high degree of translational regulation, and even when good correlation between mRNA and protein changes are observed, the degree of mRNA change is more subtle than that of the corresponding protein (McNicoll et al. 2006; Lahav et al. 2011).

It has become clear that *Leishmania* undergo a switch during promastigote-to-amastigote differentiation from relying on glycolysis to utilising beta-oxidation and amino-acid catabolism to obtain energy and cellular building blocks (Rosenzweig, Smith, Opperdoes, et al. 2008; Lahav et al. 2011). The notion that the parasite is constitutively adapted to life in both hosts, with very little changes in transcript and protein levels, and differential availability of nutrients promotes the cellular differences observed between promastigotes and amastigotes (Leifso et al. 2007) is difficult to reconcile with the ordered progression through differentiation in axenic systems (Figure 1.6) (A. Saxena et al. 2007; Rosenzweig, Smith, Opperdoes, et al. 2008; Lahav et al. 2011). As Tsigankov *et al.* (Tsigankov et al. 2012) point out, supply of differentiation signals whilst maintaining cells in an otherwise unchanged medium results in ordered differentiation and metabolic shift, showing that a fixed program is being executed that is, at least in some aspects, irrespective of nutrient availability.

Temperature shock is a crucial trigger for differentiation, albeit acidification is a contributing factor for full differentiation (Alcolea et al. 2010; Tsigankov et al.

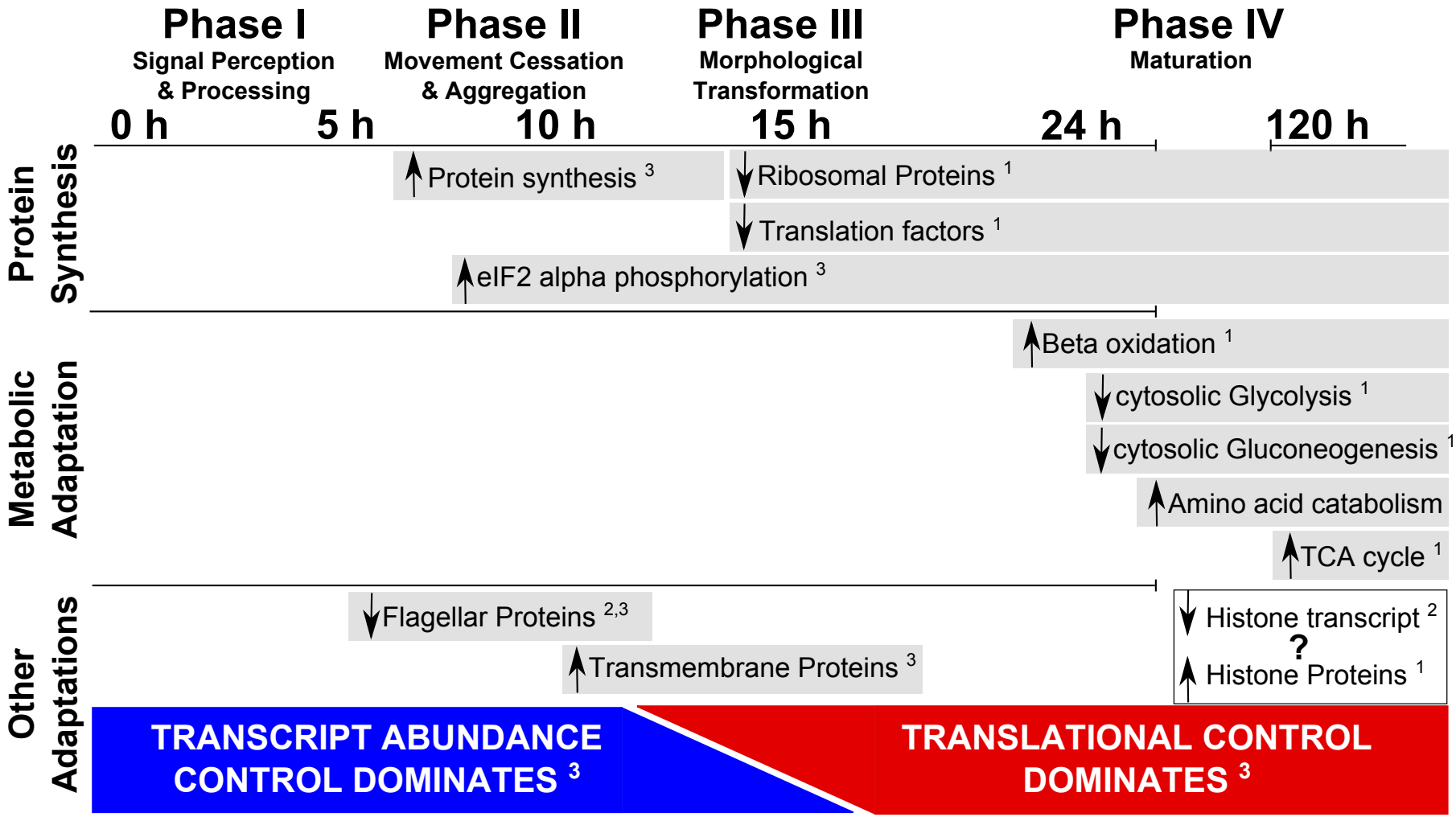


Figure 1.6 Chronology of events during promastigote to amastigote differentiation Schematic showing the chronology of gene expression changes during promastigote to amastigote differentiation based on transcriptomic and proteomic data. Arrows indicate up- or downregulation whilst grey bars indicate approximate duration of events. Classification into phases adopted from Rosenzweig *et al.* 2008.
References: [1] = Rosenzweig *et al.* 2008, [2] = Saxena *et al.* 2007, [3] = Lahav *et al.* 2011.

2014). Moreover, iron starvation is emerging as a trigger and may, *in vivo,* have functions in inducing a full amastigote cell type.

Translational activity in amastigotes is reduced compared to promastigotes, which has been shown by a range of methods as discussed above. Notably, this also correlates with post-translational control of the translation initiation factor eIF2-alpha through phosphorylation (Lahav et al. 2011). The wide-ranging and dynamic patterns of protein phosphorylation (and other post-translational modifications) will certainly turn out to include further levels of control utilised by *Leishmania spp.* to mediate differentiation (Rosenzweig, Smith, Myler, et al. 2008; Tsigankov et al. 2014). The investigation of post-translational modifications is where proteomic studies are the approach of choice. In the light of these, one may ask what lies in stall for transcriptomic approaches?

The promises borne by the advent of sequencing-based transcriptomic methods are starting to fulfil with the availability of SLAS and PAS positions as well as transcript dimensions for various kinetoplastids (Siegel et al. 2010; Kolev et al. 2010; Nilsson et al. 2010; Rastrojo et al. 2013). Moreover, the wide dynamic range of transcript detection possible in these methods will permit identification of more differentially expressed genes and provide further insight into the biology of these organisms.

At the moment however, RNA-seq studies on *Leishmania spp.* (Rastrojo et al. 2013; Mittra et al. 2013; Martin et al. 2014), show how RNA-sequencing may be employed, but on their own have not tapped into the full potential of what is technically possible. This is either for the reason of having used SL-RNAseq (Mittra et al. 2013; Martin et al. 2014), which only yields reads derived from the spliced leader and no information about other poly-A sites and thereby transcript

dimensions. Or, because only one cell type (PRO) was analysed (Rastrojo et al. 2013), making functional interpretation of transcript abundances extremely difficult.

1.7 Aims of the following work

I therefore decided to employ RNA-sequencing technology to investigate the transcriptomes of *L. mexicana* PRO as well as AMA and AXA 24 h into differentiation. First, sequencing data was used to define transcript boundaries. I aimed to determine SLAS and PAS for each gene, and define novel genes where appropriate. These findings were supplemented by proteomic investigations. Subsequently, using the generated transcript models, the transcriptomes of PRO, AMA and AXA were quantified and differentially expressed genes identified, not only to elucidate differences between promastigote and amastigote cell-types, but also to identify similarities and dissimilarities between axenic and intracellular amastigotes. Finally, a selection of candidate genes was chosen for more in-depth analyses consisting of bioinformatic characterisation and sub-cellular localisation of GFP-fusion proteins.

Chapter 2 – Prediction of gene models in *L. mexicana* using RNA-sequencing guided definition of transcript boundaries

2.1 Introduction

One of the primary aims of this thesis was to generate and accurately quantify gene expression profiles for *Leishmania mexicana* promastigotes and amastigotes using RNA-sequencing. The accurate quantification of gene expression profiles is dependent on the accuracy of gene models, both in terms of completeness of annotation of all genes in the genome and in terms of knowing the boundaries of transcripts.

The 32 megabase genome of *Leishmania mexicana* comprises 8333 genes of which 8250 are predicted to be protein coding according to the annotation used for this study (TriTrypDB v4.2). Gene models were largely transferred by orthology from the annotation of *Leishmania major* (Rogers et al. 2011). As discussed in section 1.3 the *L. major* genome was annotated by a range of gene-prediction tools and using homology-based strategies. Very little empirical evidence was used to support these gene models, and even though the current annotations have proven accurate enough to facilitate hundreds of research projects into *Leishmania spp.* biology in the last 10 years, there may be sets of genes without homologies in other organisms or eluding gene finding algorithms because they are unusual in their physical properties such as size or nucleotide composition. This begs the question of how many genes can still be found in the genome of *L. mexicana*, and what biological role they might play.

Moreover, for protein coding genes, only coding sequences are defined in *L. mexicana* with no genome-wide information about untranslated regions (UTRs) and transcript dimensions available. Definition of the entire mRNA transcript as opposed to only the coding sequence (CDS), will increase the precision with which transcript abundances can be quantified in e.g. RNA-sequencing studies. This is because sequencing-read distribution

is not homogenous over a given transcript body and local coverage biases can be compensated for by quantifying read-coverage over a larger area of a genome (Sendler, Johnson, and Krawetz 2011).

Definition of UTRs will give important insight into the biology of *L. mexicana* and related organisms. Kinetoplastid genomes are organised into polycistronic transcription units (PTUs) (Johnson, Kooter, and Borst 1987; Ivens et al. 2005) that are constitutively transcribed. Pre-mRNA processing, generates mature mRNA transcripts (Sutton and Boothroyd 1986; LeBowitz et al. 1993). Genes within a single PTU are not functionally linked (Kelly et al. 2012) and since mRNA abundance cannot be regulated at the level of transcription initiation, mRNA degradation is a key process in regulating mRNA abundance (Clayton and Shapira 2007). Sequence elements in the 3' and 5' untranslated regions have been shown to mediate mRNA stability within the cell cycle (Abanades et al. 2009; Archer et al. 2011) and between life-cycle stages (Mishra et al. 2003; Holzer et al. 2008; Murray et al. 2007; Dupé, Dumas, and Papadopoulou 2014) (see Section 1.4.3)

Experimentally, only very few UTRs and transcript dimensions have been defined for *L. mexicana.* (e.g. PFR-2 (Moore, Santrich, and LeBowitz 1996)). For *L. major*, transcript boundaries have been published (Rastrojo et al. 2013), however transcript sizes were not compared to coding-sequence sizes to define UTRs in *L. major*. Definition of UTRs will benefit the further identification of regulatory sequences that may play a role in developmental regulation of gene expression and give additional insight into the organisation of the genome.

Considering the pre-mRNA processing sites on their own, namely the 5' trans-splice site and 3' poly-adenylation site, the precise mechanisms underlying site-selection are still not fully elucidated. Furthermore, recent genome-wide evidence has pointed to wide-spread heterogeneity of processing sites (Siegel et al. 2010; Kolev et al. 2010; Rastrojo et al. 2013)

leading to the question whether or not this heterogeneity has any function in generating different transcripts between life-cycle stages.

2.2 Aims

The aim of the following work was to employ RNA-sequencing of *L. mexicana* exponentially growing axenic promastigotes, amastigotes 24 h into axenic differentiation and intracellular amastigotes 24 h post-infection to define and quantify pre-mRNA processing sites. From this, transcript boundaries were mapped, which lead to the identification of novel genes and refinement of existing gene models. Using these data, the dimensions of UTRs and the sequences surrounding pre-mRNA processing sites could be analysed. Subsequently, mass-spectrometric studies were used to confirm refinement of existing and predictions of novel genes.

2.3 Results

2.3.1 Isolation of RNA from AMA, AXA and PRO

To sequence the transcriptome, three *L. mexicana* cell types were grown *in vitro*: Axenic promastigotes (PRO), axenic amastigotes 24 h after transfer into differentiation medium (AXA) and intracellular amastigotes 24 h after infection of bone-marrow derived macrophages (AMA). Furthermore, RNA was extracted from murine bone-marrow derived macrophages (BMDM). For each cell type, RNA samples of three biological replicates were obtained. For the generation of intracellular amastigotes, murine bone-marrow derived macrophages, which were harvested as described in Materials and Methods, were used as host-cells. To confirm the macrophage phenotype prior to infection with parasites, the murine cells were stained for the murine macrophage markers F4/80 (Austyn and Gordon 1981), MAC-1 (Springer et al. 1979), and the granulocyte marker GR-1. Stained cells were analysed in a BD FACSCalibur flow cytometer (see Materials and Methods) (Figure 2.1). The fluorescence shift in Fig. 2.1 A following anti-F4/80 and anti-MAC1 stain relative to the isotype control indicates an F4/80 - and MAC1-positive phenotype. Coincidence of the anti-GR1 signal with the isotype control indicates a GR-1 negative phenotype. The bone-

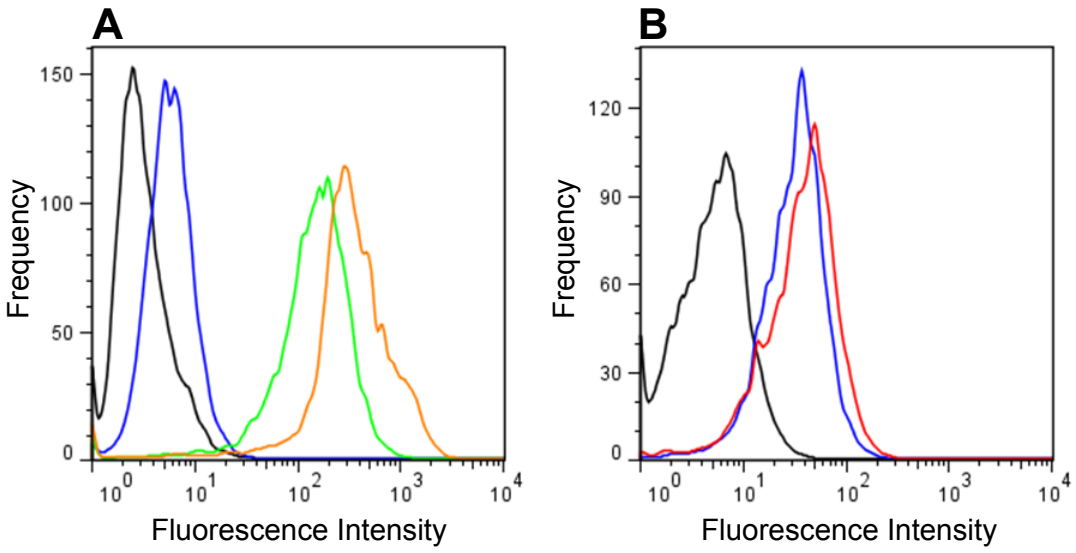


Figure 2.1 Flow-cytometric characterisation of BMDMs Fluorescence intensity histograms of matured bone-marrow derived cells (black trace) stained with isotype-controls (blue) and (A) anti-F4/80 (green), anti-MAC1 (orange) and (B) anti-GR-1 antibodies (red) as described in Materials and Methods. 5000 events acquired for each sample.

marrow derived cells were therefore considered bone-marrow derived macrophages.

Stationary growth phase *L. mexicana* promastigotes were used to infect BMDMs for 24 h. The growth history of the *L. mexicana* cells used in this study is shown in Figure 2.2, and the infection protocol is outlined in Materials and Methods. 24 h post-infection intracellular amastigotes can be detected in infected macrophages by light-microscopy (Figure 2.3). Between replicates, different infection levels were observed ranging from 207 to 748 amastigotes per 100 macrophages and 70.4 to 94.6% infected macrophages. Infection levels and distributions of amastigote-counts in infected macrophages are summarised in and Figure 2.4 A-D.

PRO cells were kept in exponential growth, whilst AXA cells were obtained by 24 h *in vitro* differentiation of stationary growth phase *L. mexicana* promastigote cells in Schneider's pH 5.5 medium (see Materials and Methods).

RNA was extracted from AMA, AXA, PRO and an uninfected culture of BMDMs as described in Materials and Methods and the yields summarised in Table 2.1. Parasite cells were not purified from host cells in the AMA samples, resulting in an RNA sample composed of both parasite and host cell material.

For all samples other than BMDM2 & BMDM3 >10 μg total RNA were obtained at concentrations above 100 ng/μl. Measurement of optical density (OD) ratios at different wavelengths were used to assess RNA over DNA composition (260/280 nm) of the sample as well as to check for possible carbohydrate contamination (260/230 nm) (see Materials and Methods). Results are shown in Table 2.1. OD 260/280 nm ratios range from 1.96 – 2.14, with only 3 measurements (AMA2, 3 and PRO3) below 2.0, indicating that the samples had absorption characteristics typical of pure RNA samples. OD 260/230 ratios range from 1.72 – 2.13 with only one measurement below 2.0 indicating no considerable carbohydrate contamination.

RNA integrity, was assessed by analysis of ribosomal RNA peaks using an Agilent Bioanalyser (see Materials and Methods). Analogous to the RNA Integrity Number (RIN)

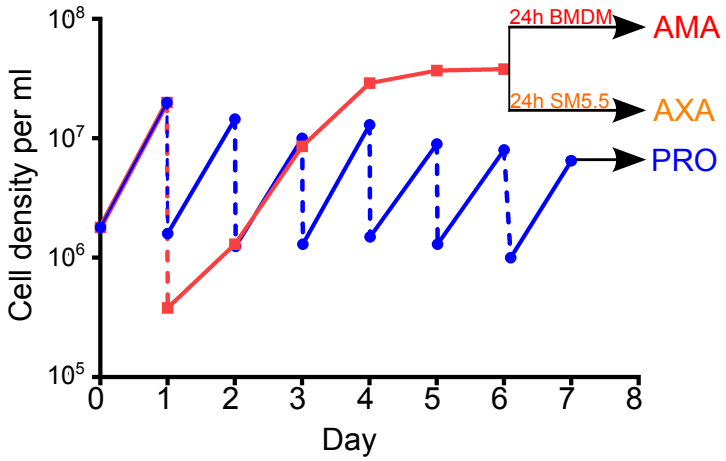


Figure 2.2 History of *Leishmania* cells used for RNA-preparation Cell density measurements of *Leishmania mexicana* cell cultures used to generate AMA, AXA and PRO samples. The data shown is from cultures used to generate AMA1, AXA1 and PRO1. The same approaches were used to generate replicates.

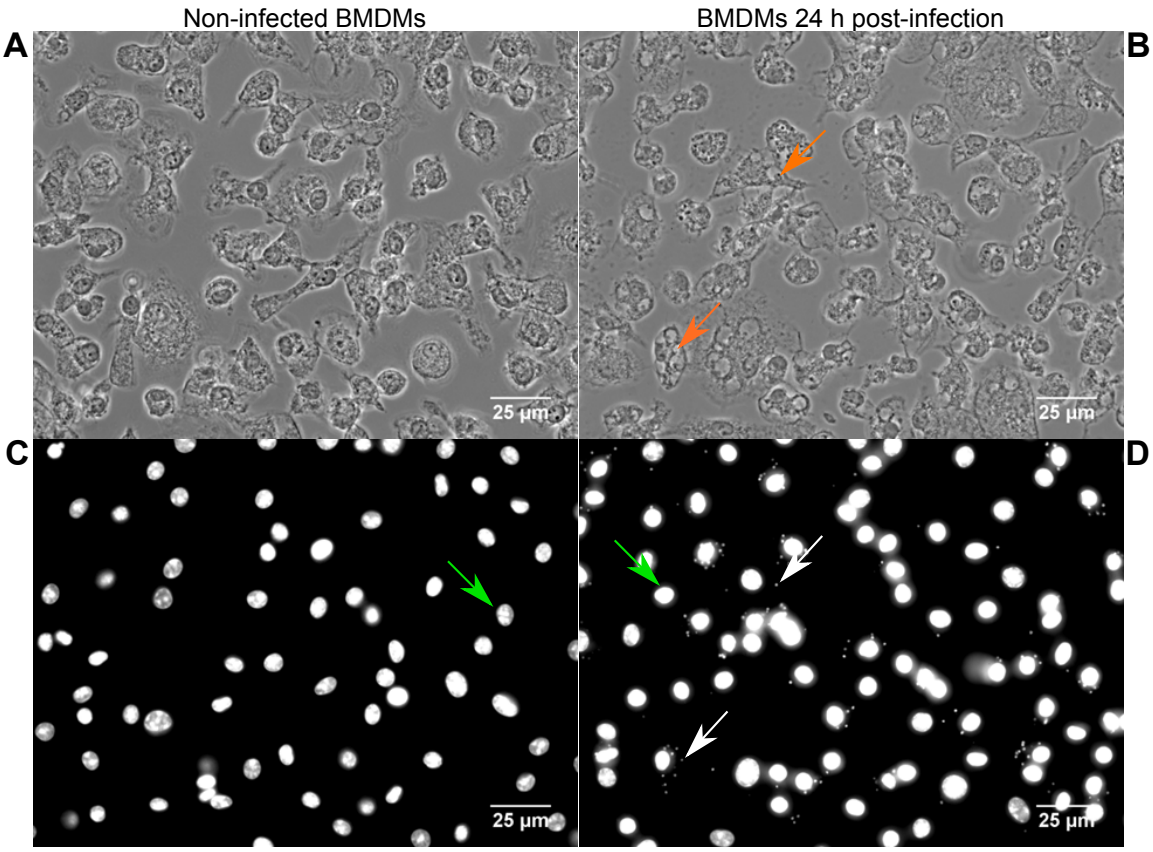
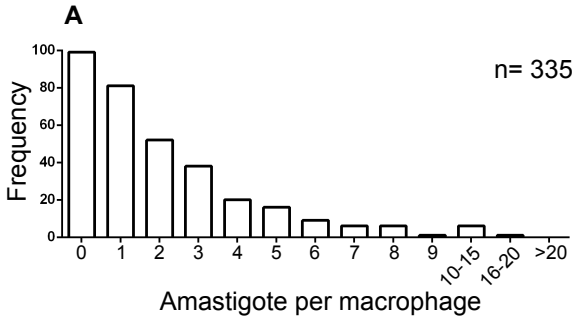
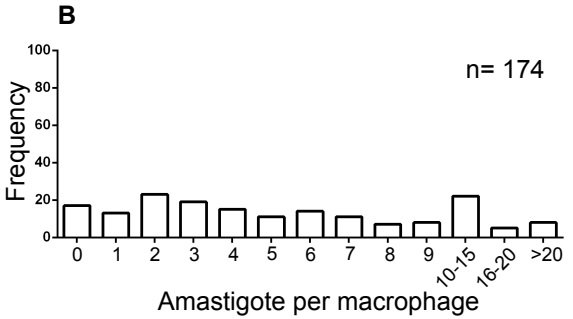
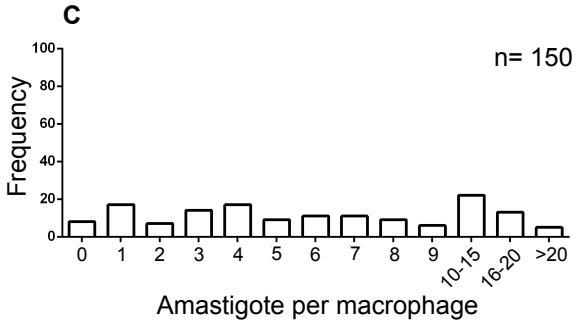


Figure 2.3 Microscopic analysis of AMA sample Panels A and B show phase contrast images of (A) uninfected BMDMs and (B) BMDMs 24 h post-infection (AMA1 sample used as example). Vacuoles containing *Leishmania* cells highlighted with orange arrows. Panels C and D show fluorescence microscopy images of propidium iodide staining. Macrophage nuclei are highlighted with green arrows, *L. mexicana* DNA with white arrows.







D

Sample Name	Percentage infected macrophages	Amastigotes per 100 macrophages	Macrophages counted
AMA1	70.4%	207	335
AMA2	90.2%	636	174
AMA3	94.6%	748	150

Figure 2.4 Infection levels in AMA samples Panels A-C show histograms of the distribution of infection levels in AMA1 (A), AMA2 (B) and AMA3 (C) as assessed my microscopic analyses. Panel D shows a table summarising these data.

Table 2.1 RNA-sample information Summary of concentration, amount and purity based on optical density ratios for AMA, AXA, PRO and BMDM.

Sample	Concentration (ng/µl)	Volume (µl)	Total amount (ug)	OD260/280	OD260/230
AMA1	730	30	21.9	2.09	2.22
AMA2	308	37	11.4	1.96	2.13
AMA3	299	37	11.1	1.97	2.01
AXA1	1340	15	20.1	2.14	2.37
AXA2	315	120	38.0	2.01	2.26
AXA3	230	120	28.0	2.03	2.29
PRO1	1330	15	20.0	2.14	2.37
PRO2	330	120	40.0	2.00	2.04
PRO3	434	120	52.0	1.96	2.19
BMDM1	579	30	17.4	2.06	2.21
BMDM2	148	37	5.5	2.00	1.72
BMDM3	128	37	4.8	2.03	2.09

(Schroeder et al. 2006), which could only be determined for BMDMs, but not for *L. mexicana* or mixed RNA-samples, ribosomal RNA peaks were assessed for peak-broadening, an indicator of RNA degradation. The RINs, determined on a scale of 0 to 10, for the BMDM samples were 9.6, 9.7 and 9.8 for BMDM1-3 respectively. Currently, the Bioanalyser software is unable to determine RINs for kinetoplastid rRNAs or mixed-species samples (AMA), as in kinetoplastids the large ribosomal transcript is cleaved resulting a total of three main rRNA peaks, namely the Large Subunit α, Large Subunit β and Small Subunit (Figure 2.6, PRO1 panel) (Villalba, Dorta, and Ramírez 1985) which interfere with RIN quantification. However, comparison of the peak-shapes seen in AMA, AXA and PRO to the peak-shapes seen in the BMDM samples, determined to be non-degraded, together with low background signal, suggested that no considerable degradation of RNA had taken place during RNA extraction. (Figure 2.5 & 2.6).

From the electropherograms in Figure 2.5 for AMA samples relative RNA contributions from each species in AMA were determined by first calculating absolute RNA using a modified Beer-Lambert law (Lambert 1760; Beer 1852): With constant path-length in each well of the bioanalyzer, quantifying the area under the peak of the Large ribosomal Subunit β (LSUβ) (*L. mexicana*) and the 4.7 kb 28 s ribosomal subunit (murine) and dividing the result by the size of the respective rRNA transcript provides absolute measures of the RNA species contributing, permitting calculation of relative amounts. Between 21.2 % and 58.9 % *L. mexicana* derived rRNA was thereby detected in the sample, correlating with different infection levels in AMA samples (Table 2.2).

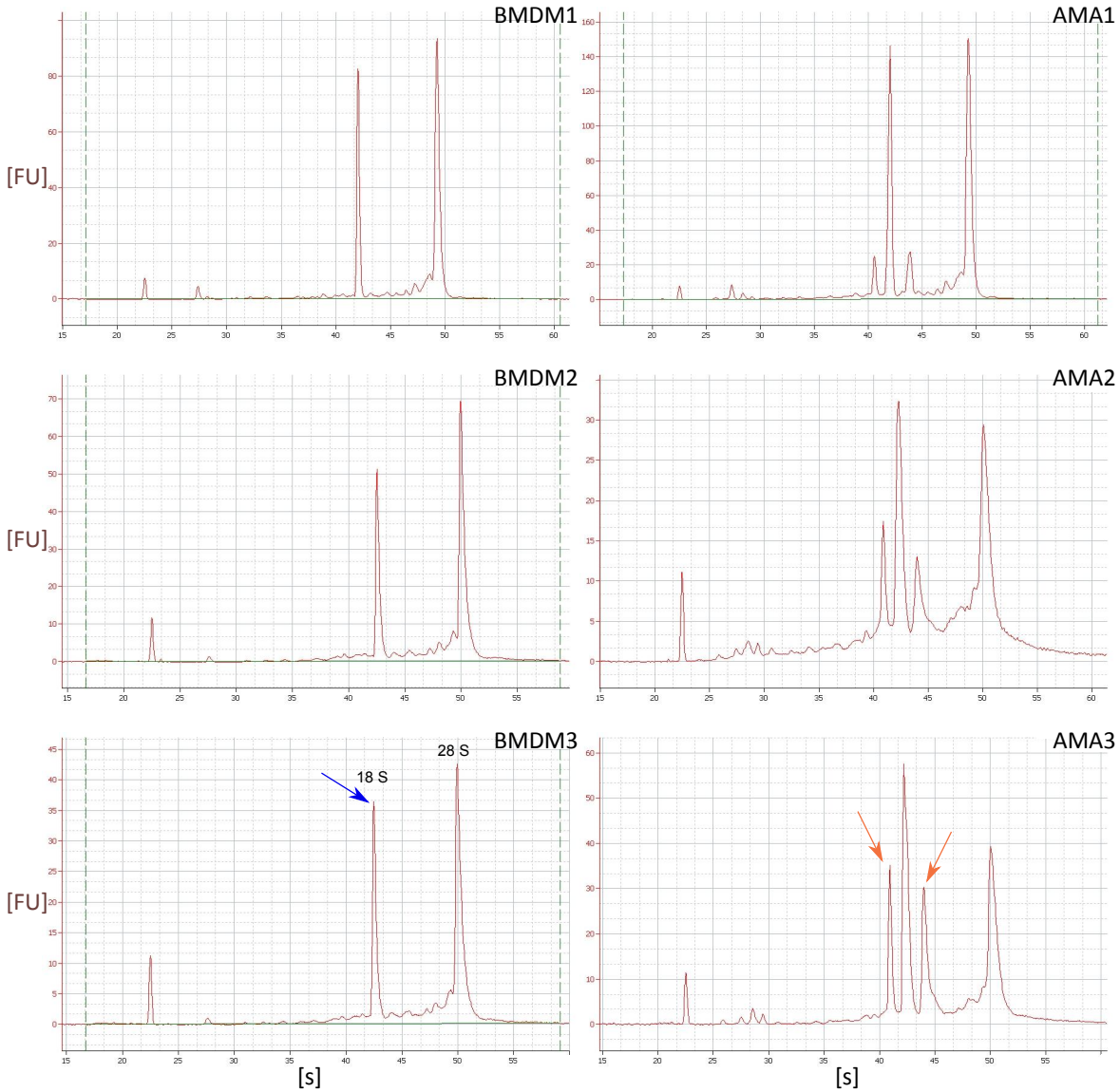


Figure 2.5 Electropherograms of BMDM and AMA samples Electropherograms of BMDM and AMA samples plotting arbitrary fluorescence units [FU] against capillary gel migration time [s]. Only a single major peak is seen between 40-45 seconds migration time in BMDM samples, the 18 S peak (blue arrow). In AMA samples additional peaks are seen (orange arrows), which are of leishmanial origin.

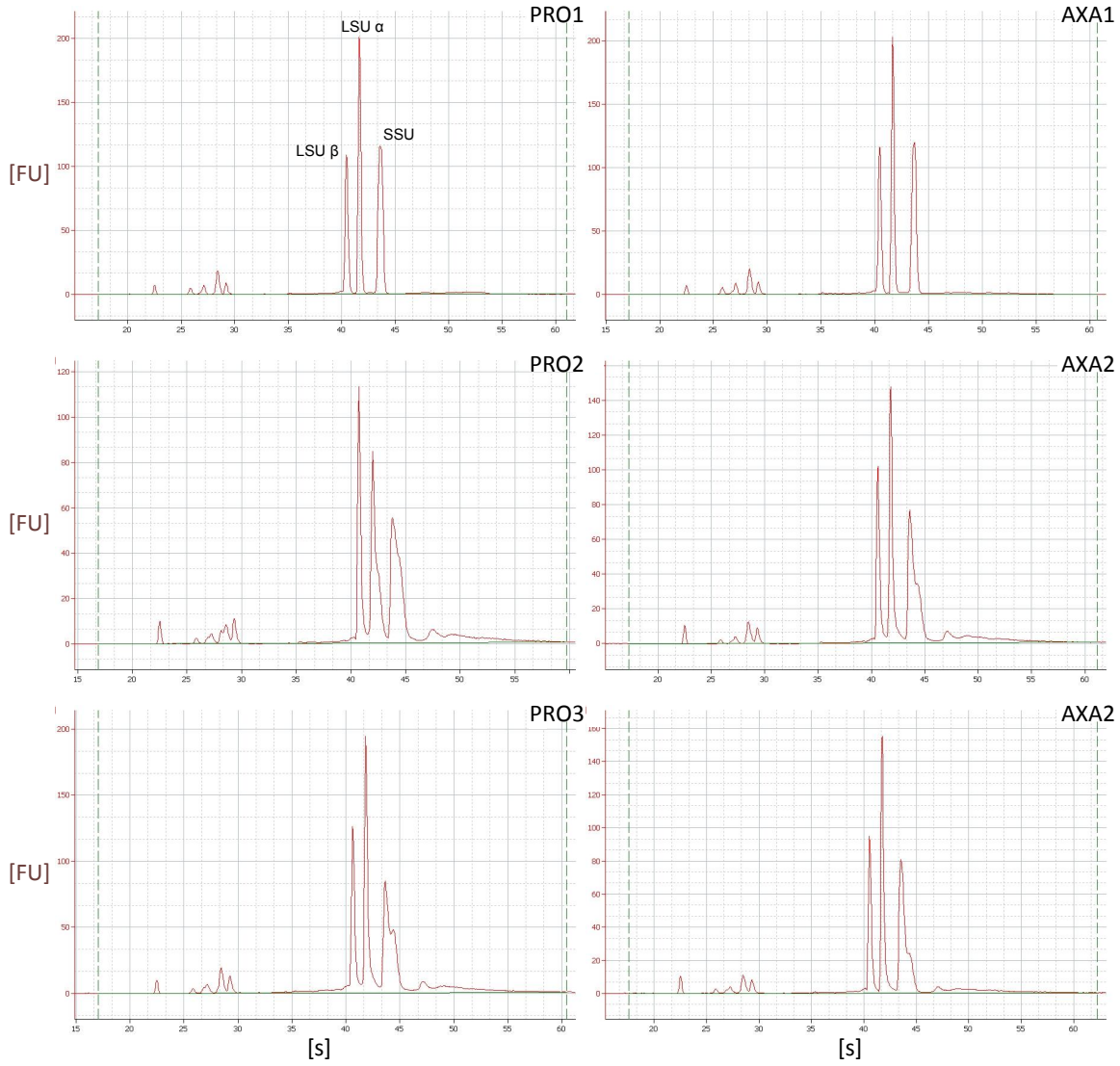


Figure 2.6 Electropherograms of PRO and AXA samples Electropherograms of *L. mexicana* promastigotes and axenic amastigotes plotting arbitrary fluorescence units [FU] against capillary gel migration time [s]. The major peaks are, from left to right, Large rRNA Subunit β (LSUβ), Large rRNA Subunit α (LSUα) and Small rRNA Subunit (SSU).

Table 2.2 Mixed species sample composition Determination of relative amounts of *L. mexicana* RNA in AMA samples using rRNA-peak sizes. Infection levels are shown for comparison.

Sample Name	Peak Size LSU β, 1.6 kb (A.F.U.)	Peak Size 28 S, 4.7 kb (A.F.U.)	Relative *L. mexicana* rRNA amount	Relative murine rRNA amount	*L.mexicana* r RNA in sample	Infection level as Amastigotes per 100 macrophages
AMA1	3.4	36.6	2.1	7.8	21.2%	207
AMA2	4.8	16.5	3.0	3.5	46.2%	636
AMA3	8.4	17.5	5.3	3.7	58.9%	748

2.3.2 Paired-end sequencing of RNA-samples yielded high quality sequencing data with a nucleotide composition reflecting the proportion of RNA originating from *L. mexicana* and *Mus musculus*

The extracted RNA was selected for poly-adenylated transcripts and sequenced in a paired-end manner on an Illumina HiSeq 2100 platform. Importantly, two types of cDNA library were sequenced. Library 1 was used for the mapping of spliced-leader acceptor sites and quantification of transcript abundances, library 2 for the mapping of poly-adenylation sites (see Material and Methods).

To generate library 1, random hexamer-primed first- and second-strand cDNA synthesis was performed for all replicates of AMA, AXA and PRO. The technical results, including the number of reads, read length, Q20 percentage (i.e. the percentage of bases with a probability of being called incorrectly lower that 1:100) and the GC-content of the sequencing reads are summarised in Table 2.3. Around 26 million reads were obtained for all samples with read lengths being 90 nt. The Q20 (Ewing and Green 1998) percentage was above 95.9% across all sample, i.e. for over 95.9% of the 90 nt reads there is at worst a 1 in 100 likelihood of miscalling a nucleotide, indicating high-quality read data. The GC-content was around 59% for all *L. mexicana*-only samples, which is consistent with the GC-content of the *L. mexicana* genome reported to be 59.7% (Rogers et al. 2011). The mouse genome has a considerably lower GC-content of 42 % (*Nature* 2002), albeit higher in gene-rich regions. The lower GC-content of the mouse genome accounts for the slightly lower GC-contents being observed in AMA compared to AXA or PRO. Indeed the lowest GC-content in AMA is found in AMA1 (52.0%) which has the lowest proportion of *L. mexicana* RNA. AMA2 and AMA3 have 56.3 and 56.8% GC-content reflective of their higher parasite load compared to AMA1 and reflective of AMA3 having slightly more parasites than AMA2 (c.f. Figure 2.4 D).

Analysis of the average nucleotide frequency along each base of random hexamer primed sequencing reads revealed a heavy compositional bias in the first 12-15 bases. Figure 2.7

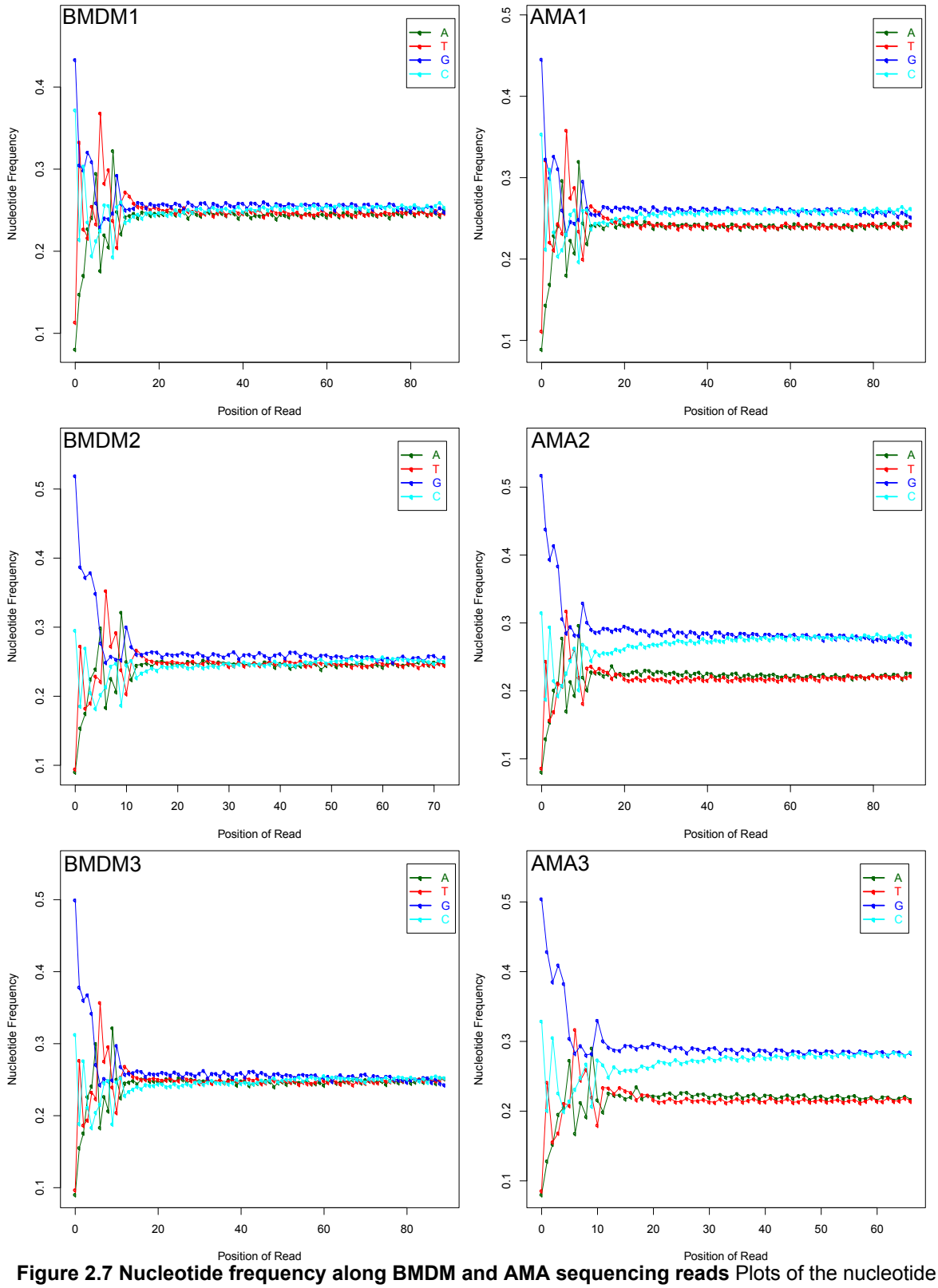


Figure 2.7 Nucleotide frequency along BMDM and AMA sequencing reads Plots of the nucleotide frequency for every position of BMDM and AMA sample sequencing reads. The 5' nucleotide bias disappears after the first 12-20 nucleotides and remains evenly distributed, albeit reflecting the higher GC-content of samples containing *L. mexicana* material. Due to the higher infection levels in AMA2 and AMA3 versus AMA1, former show a much higher GC-content.

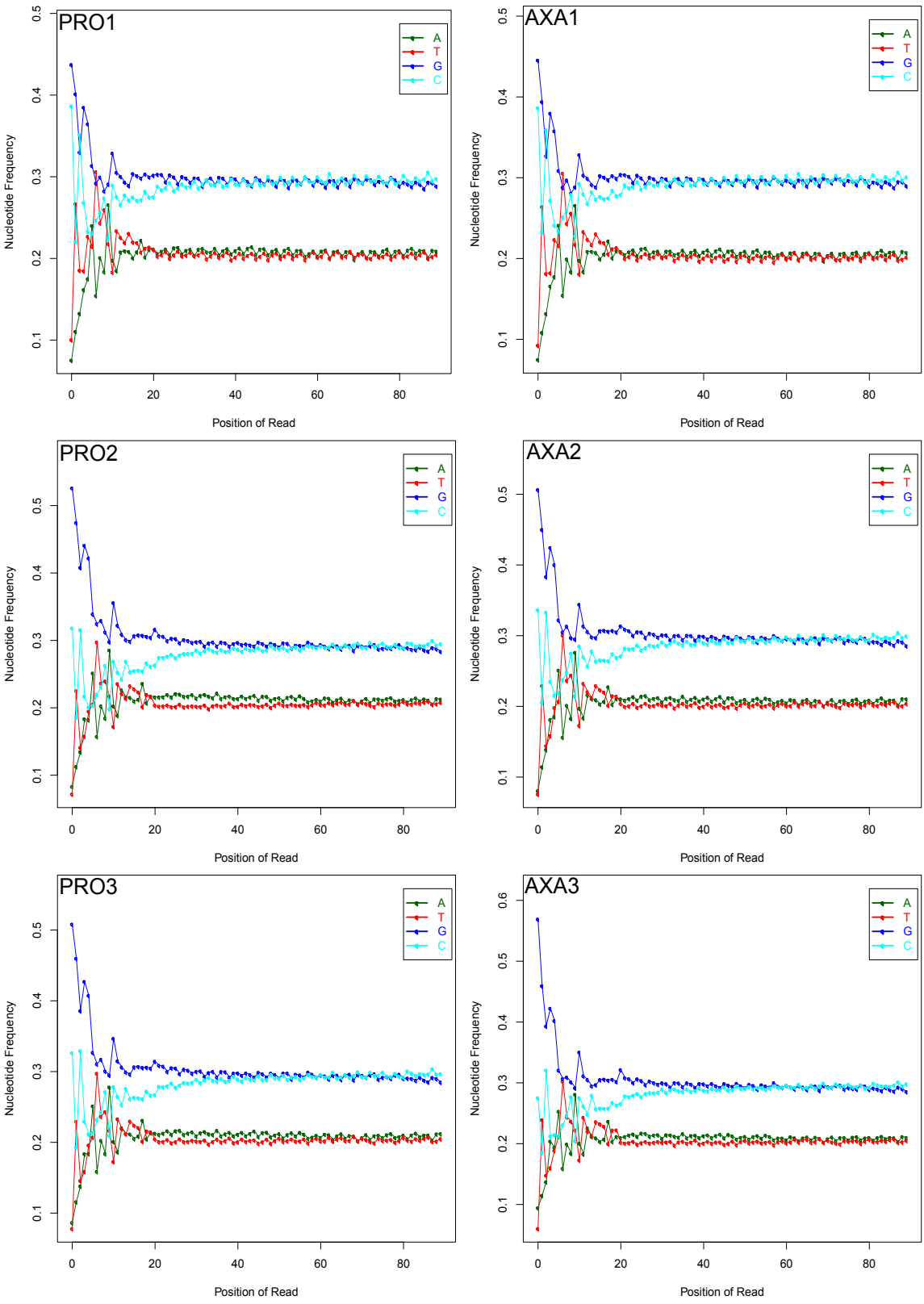


Figure 2.8 Nucleotide frequency along PRO and AXA sequencing reads Plots of the nucleotide frequency for every position of PRO and AXA sample sequencing reads. The 5' nucleotide bias disappears after the first 12-20 nucleotides and remains evenly distributed, albeit reflecting the approximately 60% GC-content of *L. mexicana* transcripts.

and 2.8 shows the average nucleotide composition along all reads in AMA, AXA and PRO. For comparison the plots are also shown for BMDM samples. All traces started off with drastically changing average nucleotide frequencies, which stabilise by about read-position 15. This is a known artefact of random hexamer priming and Illumina sequencing (Hansen, Brenner, and Dudoit 2010). The different heights of traces of G and C compared to A and T are reflective of the GC-content of the samples as previously discussed.

Library 2 library was prepared from AMA2, AMA3, AXA2, AXA3, PRO2 and PRO3 RNA using 5'-T15VN-3'-primed first-strand synthesis , where V is any base but T and N being any base, followed by random hexamer primed second strand synthesis. This procedure is designed to enrich for sequencing reads starting on the poly-A tail of mRNAs to allow mapping of poly-adenylation sites (Kolev et al. 2010).

2.3.3 Mapping of sequencing reads to genome reflects proportion of RNA originating from *L. mexicana* and mouse

As we have seen in section 2.3.1 and 2.3.2, the composition of the mixed-species AMA samples in terms of relative rRNA contribution and GC-content reflects the infection levels of the samples as determined by microscopy. In a next step I sought to investigate whether or not sequencing reads could be mapped to the mouse and *L. mexicana* genomes and whether or not the proportion mapping to either reflected the relative amounts of murine and leishmanial RNA in the input material. To test this, reads from the random hexamer primed cDNA libraries were first quality-trimmed using FASTX-Toolkit and mapped to a hybrid mouse-*Leishmania mexicana* genome using Stampy software (Lunter and Goodson 2011).

Reads could be mapped to the hybrid genome and importantly, the percentage of reads mapping to the *L. mexicana*-part of the hybrid genome correlates with the sample composition as determined by quantification of rRNA abundances in section 2.3.1 (Figure 2.9), with an R^2-value for a linear correlation of 0.985. This shows that the sequencing data is of high-enough quality to reflect the input material upon *in silico* separation of

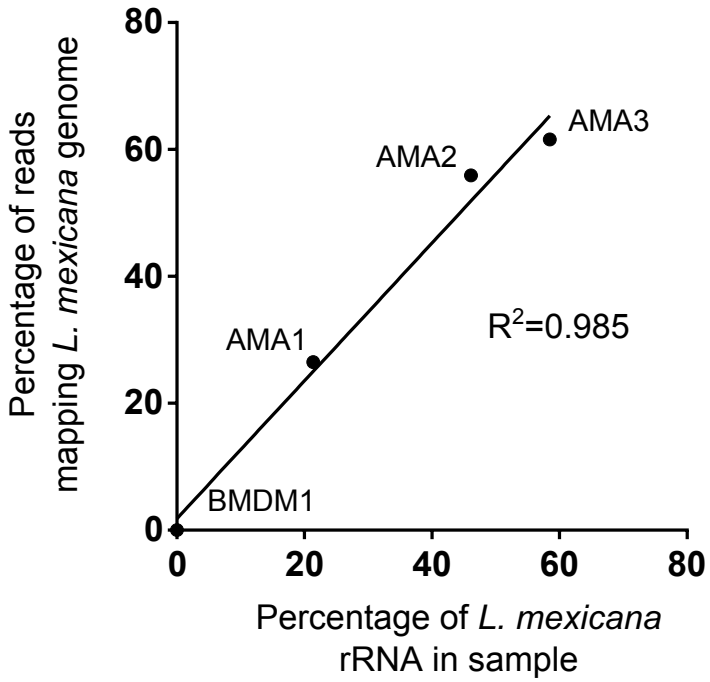


Figure 2.9 Correlation of read-mapping with sample composition Plot of the precentage of reads mapping to *L. mexicana* genome in AMA samples against the precetage of *L. mexicana* rRNA in AMA samples. Instead of artificially rooting the linear regression line at the origin, data for BMDM1 is shown, in which 0.03% of reads mapped to the *L. mexicana* genome.

transcriptomic data originating from the two species present in the mixed species samples.

2.3.4 Mapping of spliced-leader acceptor sites and polyadenylation sites generated transcript models, revealed a cohort of novel genes and allowed refinement of gene models

In order to generate transcript boundaries, the spliced-leader acceptor sites (SLAS) and polyadenylation sites (PAS) were mapped and assigned to coding sequences. To map SLAS, sequencing reads containing the last 12 nt of the 39 nt spliced-leader (SL) (Full sequence: AACTAACGC TATATAAGTA TCAGTTTCTG TACTTTATTG, (Agami and Shapira 1992)) closest to the splice site were extracted from random-hexamer-primed sequencing libraries, the SL sequence trimmed off and the remainder mapped to the genome. The 5' position of this mapped read is the SLAS. PAS were identified by extracting all sequencing reads from sequencing libraries generated by 5'-T15VN-3'-primed first-strand cDNA synthesis with five or more As or Ts at the start or end of the read. The runs of As and Ts were trimmed off and the remainder mapped to the genome. Wherever no run of As or Ts corresponding to the trimmed-off sequence was present in the genome sequence, the 3' end of the mapped remainder was considered a PAS (see Materials and Methods).

Examples of genes with their respective SLAS and PAS are shown in Figures 2.10, 2.11. Closer inspection of the reference gene models in Figure 2.11 revealed a locus between LmxM.23.0480 and LmxM.23.0490 with SLAS, PAS and clear read-coverage, however no gene was annotated in that locus. Furthermore, a start-codon was found in-frame with the annotated start codon but downstream of the SLAS for LmxM.23.0490, suggesting a potential 5' extensions of the CDS.

Since this indicated that there were novel transcripts from non-annotated genes, I decided to use the SLAS and PAS positions to guide a new annotation of CDSs in *L. mexicana*, and only upon completing this assign SLAS and PAS to CDSs. To this aim, all SLAS were mapped (see Materials and Methods). Only SLAS detected at least 9 times across all nine

Figure 2.10 Examples of SLAS and PAS positions defining transcript boundaries Example of four genes from chromosome LmxM.01 with the positions of their SLAS and PAS delineating transcript boundaries.

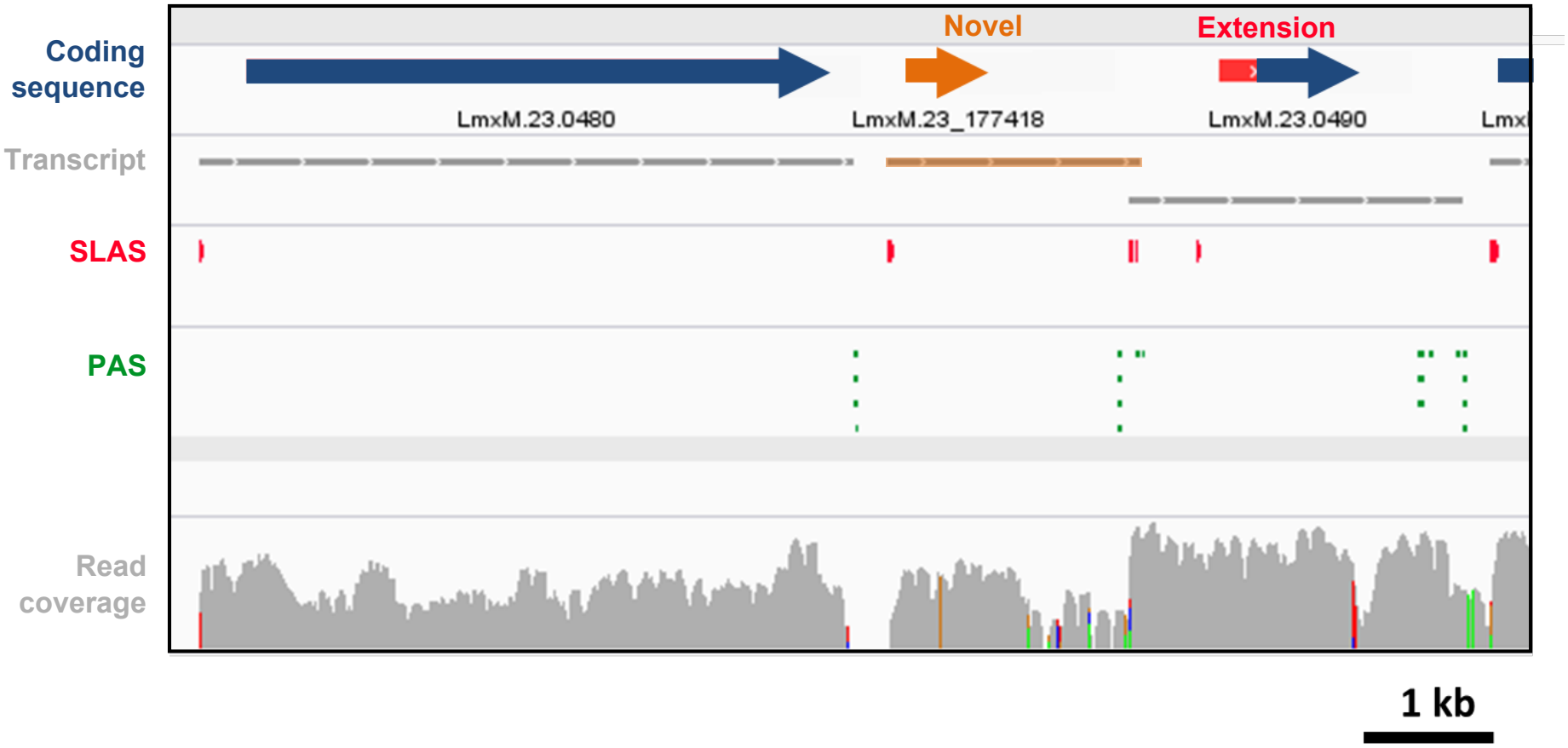


Figure 2.11 Example of a novel transcript and putative CDS extensions A genomic locus on chromosome LmxM.23 showing coding sequences transcript dimensions, SLAS and PAS sites as well as RNA-seq read-coverage. The novel CDS and transcript are shown in orange and the proposed 5'-extension of LmxM.23.0490 is shown in red.

RNA samples were considered for downstream processing. In total, 21243 SLAS were identified with an AG-dinucleotide forming the splice-acceptor in 96.0% of cases (Figure 2.12). Table 2.4 A shows the number of paired end reads within the random hexamer primed library containing the SL-sequence (as defined above) and the number of reads amongst those that were mapped to identify SLAS positions.

Each mapped SLAS was considered a potential start of a protein-coding transcript. From each SLAS, the directionally appropriate downstream sequence was scanned for the first open reading frame (ORF) encoding a protein ≥25 AA without stretching over the next SLAS. This ORF was recorded (R-script used for this provided in Supplementary Materials). Complementation of this using non-conflicting entries of reference annotation (Logan-Klumpler et al. 2012) (version 4.2) and subsequent comparison showed that in 6796 cases the same CDS as annotated in the reference annotation is found in my annotation (Figure 2.13). In 184 cases the position of the SLAS suggests that the reference ORF should be truncated, i.e. remain in the same translational frame but use a downstream start ATG (Figure 2.13) (for detailed explanation of truncation rules see Materials and Methods). For 1253 genes an in-frame ATG exists between the annotated start codon and the SLAS, suggesting that an upstream start codon may be used (Figure 2.13). For 1122 genes a non-overlapping upstream ORF (uORF) was identified, i.e. and upstream ORF exists in the same inter-SLAS –space as an annotated gene (Figure 2.13).

Finally, around 2000 putative novel ORFs were identified which were subjected to further curation. In particular, integration of PAS-data into this annotation allowed removal of potential false positives (see Materials and Methods). The number of reads in the 5'-T15VN-3'-primed cDNA libraries (library 2) containing PAS-sequences (as defined above) is shown in Table 2.4 B. Only PAS detected 6 times or more across all six samples used to generate cDNA library 2 were considered for this and all downstream steps. Elimination of

Table 2.4 Summary of reads used for determination of SLAS and PAS The number of reads from the (A) random-hexamer primed cDNA library containing the last 12 nt of the Spliced-Leader sequence and (B) the number of reads from the 5'T15VN3'-primed cDNA library containing a poly(A) tail-sequence as described in the text.

A

	AMA1	AMA2	AMA3	AXA1	AXA2	AXA3	PRO1	PRO2	PRO3
Total paired-end reads	13,477,640	13,357,246	13,044,810	13,124,114	13,748,422	12,778,984	12,835,963	13,087,624	12,858,646
Low-quality paired-end reads	9,434	13,357	11,740	17,061	34,371	17,891	20,538	23,558	29,575
Total reads with SL sequence (last 12nt)	262,651	499,028	572,112	786,722	1,009,393	1,225,884	673,297	1,067,867	937,353
Mapped reads with SL sequence (last 12nt)	259,961	493,948	566,435	778,432	994,815	1,208,732	665,046	1,051,560	923,254

B

		AMA2	AMA3		AXA2	AXA3		PRO2	PRO3
Total paired-end reads		11,978,297	12,848,473		10,967,895	10,225,192		12,978,183	11,427,219
Low-quality paired-end reads		4,806,622	4,498,233		3,455,035	2,718,125		4,006,487	4,308,920
Total reads with poly(A) tail (≥5 A)		1,793,186	1,476,286		2,200,926	2,149,944		2,097,561	2,088,542
Mapped reads with poly(A) tail (≥5 A)		424,491	373,912		805,680	812,005		756,459	767,004

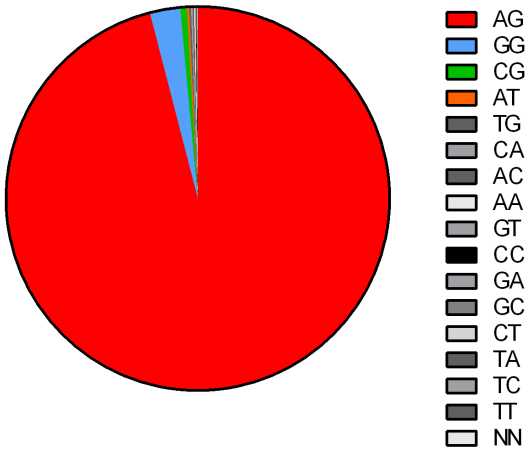


Figure 2.12 Dinucleotide useage at SLAS Piechart showing the proportion of different dinucleotides that form the accetor site for the spliced leader. The cannonical AG-dinucleotide is used in 96.0% of cases, with GG-dinucleotides forming the second most common acceptor site (2.6%).

putative novel transcripts using PAS resulted in a final number of 936 novel transcripts bringing the total number of predicted protein coding genes to 9169 (Figure 2.13).

After completion of annotation of all CDSs, SLAS and PAS were assigned to these CDSs (Figure 2.14). Importantly, SLAS and PAS were assigned in such a way as to allow for CDS-internal SLAS and PAS. Therefore, all SLAS on the same strand as a given CDS (CDS-1) and upstream of the stop codon, but downstream of the most proximate coordinate of the neighbouring upstream CDS (CDS-0) situated 5' direction are assigned to a CDS-1. Similarly, all PAS on the same strand as a CDS (CDS-1) and downstream of the start codon, but upstream of the most proximate coordinate of the CDS (CDS-2) situated in 3' direction were assigned to CDS-1.

The distance between the most distant SLAS and PAS constitutes the dimensions of the transcript. Where no SLAS or PAS was available, the start or end of the CDS was used respectively. It was possible to detect SLASs for 8882 of the total 9169 CDSs and PASs for 8796 CDSs, with 8540 CDSs having both PAS and SLAS. For only 58 CDSs neither a SLAS nor a PAS was detected. It has to be noted that none of the latter CDSs are part of the novel transcripts identified in this study due to the requirement for SLASs in the prediction and PAS in the curation of novel transcripts. The genomic coordinates of the SLAS, PAS, CDS- and transcript-dimensions are available in General Feature Format (.gff) in the supplementary material to this thesis (Supplementary Table 2.1)

Our method of mapping SLAS and PAS and assigning these to CDS models was devised in collaboration with Dr. Steven Kelly, Dept. of Plant Sciences, University of Oxford, and has been integrated in a freely available web-server and published in Fiebig *et al.* (Fiebig et al. 2014).

The mean number of SLASs per gene was found to be 2.4 (median: 2). Greater heterogeneity was found amongst PAS, where the mean number per gene was 10.9

	In agreement with existing annotation	Adjusted ORFs	New ORFs	uORFs
Number of ORFs	6796	1437	936	1122 (excl. from total)

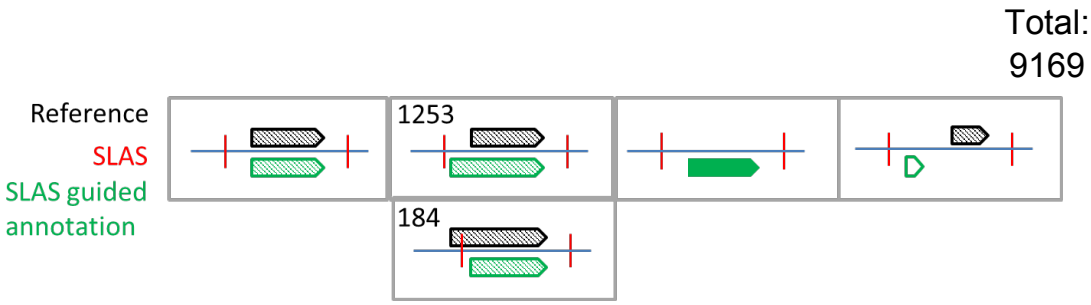


Figure 2.13 RNA-sequencing guided annotation of CDS in *L. mexicana* Table and schematic summarising the numbers of genes where the reference annotation agreed with the RNA-sequencing guided annotation and where extensions or truncations of existing CDSs have been proposed. The number of genes with non-overlapping uORFs is given. It has to be noted that amongst latter genes may also be genes where extentions of CDS have been proposed. The resulting number of protein coding senquences in *L. mexicana* is 9169.

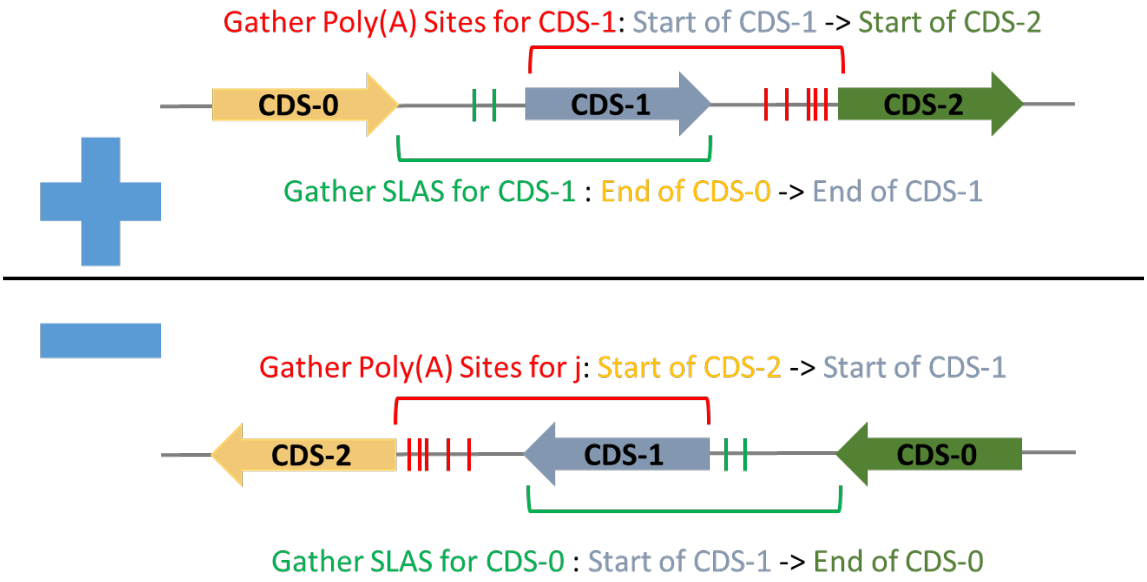


Figure 2.14 SLAS and PAS assignement rules Schematic showing how SLASs and PASs are assigned to plus- and minus-strand orientated CDS. SLAS and PAS are only assigned to a CDS if they lie on the same strand as said CDS. Furthermore CDS-2 and CDS-0 do not have to lie on the same strand as CDS-1, the most proximal coordinate of CDS-2 or CDS-0 are used in those cases. Therefore, these rules will also hold true at strand-switch regions.

(median: 9) (Figure 2.15).

Examples of the generated transcript models and their respective SLAS and PAS are shown in Figures 2.10 and 2.11. Figure 2.10 shows a typical example of four genes from chromosome LmxM.01 and illustrates the problem of defining transcript dimensions based solely on read-coverage: Whilst in the case of LmxM.01.0550, read-coverage and SLAS-PAS-defined transcript dimension agree very clearly, for LmxM.01.0560 and LmxM.01.0570 the junction between transcripts is less apparent and difficult to define without the positions of the SLAS and PAS. The precise reason for this coverage is unknown, but they could represent un-processed pre-mRNA intermediates, of which small amount was found to be present in *T. brucei* cells, owing either to a lag in assembly of the spliceosome or due to protection of the splice site by the transcription complex (Ullu, Matthews, and Tschudi 1993). It may therefore be possible that such short-lived intermediates may also exist in *L. mexicana*.

To validate the transcript sizes determined based on SLAS and PAS mapping we compared these with published data. Figure 2.16 shows a comparison of the transcript size for the glucose transporter LmGT2 as determined by RNA-sequencing (Fig. 2.16 A) to published Northern-Blot (Fig. 2.16 B) analysis. In this example, defining the correct transcript dimensions based solely on read-coverage would have been difficult as the read coverage of LmGT2 is interrupted by a sharp dip (which does not correlate with a gap in the genome assembly). Moreover, Figure 2.17 A shows a comparison of published transcript sizes determined by Northern Blot with sizes determined by SLAS and PAS mapping, which yielded an R^2-value to the linear regression of 0.73. Comparing this analysis to a correlation of reference CDS-sizes of the same genes compared to their published transcript sizes (Figure 2.17 B), resulting in an R^2-value to the linear regression of 0.44, shows that the transcript dimensions determined by SLAS and PAS mapping correlate much better with previously published results that the size of the CDS alone would. Moreover, we see that the difference between the transcript size and the CDS size is not

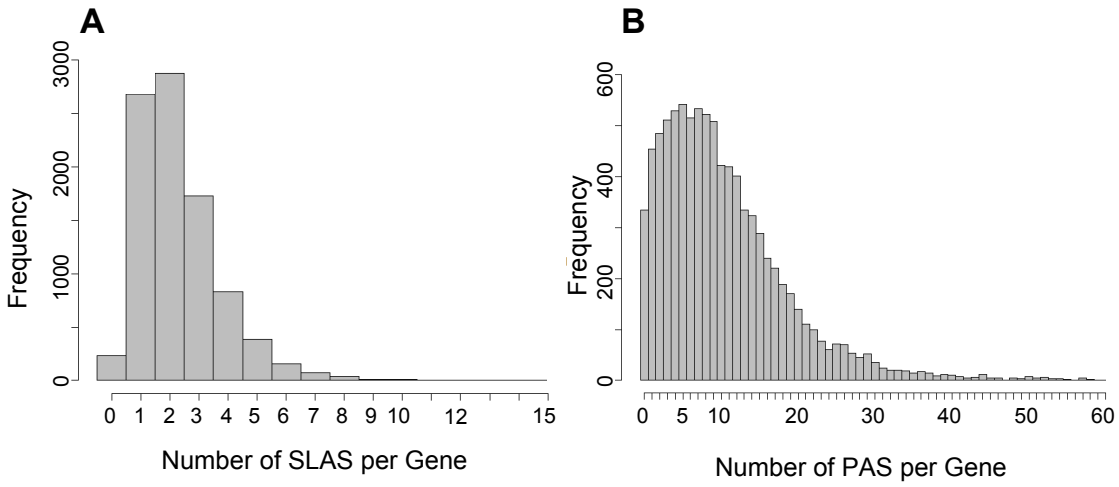


Figure 2.15 SLAS and PAS hetereogeneity Histograms of (A) the number of SLAS per gene and (B) the number of PAS per gene. The mean number of SLAS is 2.4 with a median of 2 (n=20812). More heterogeneity is seen with PAS, where the mean number per gene is 10.9 with a median of 9 (n = 95097).

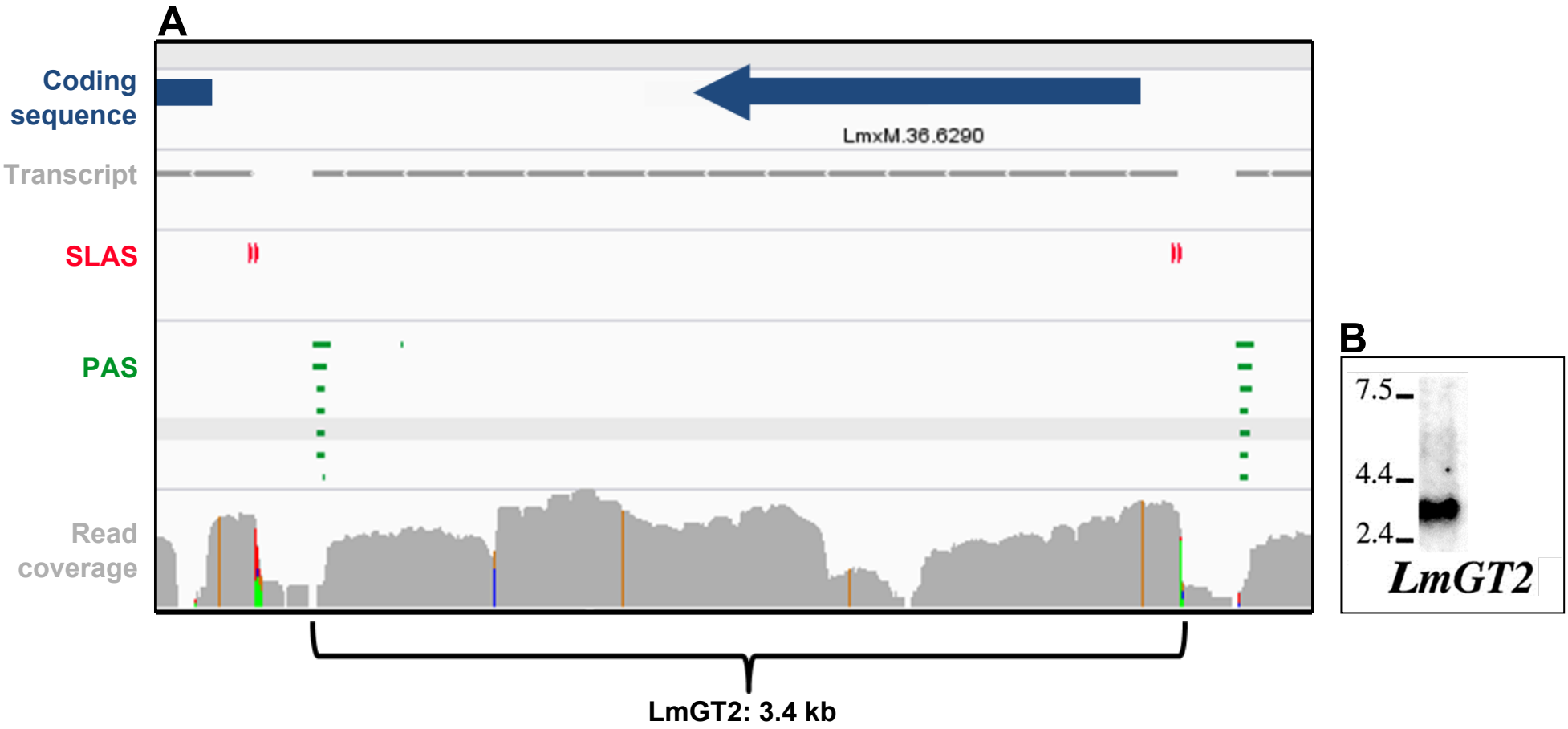


Figure 2.16 Example of transcript dimensions defined by SLAS and PAS correlating with Northern Blot analysis The genomic locus of the *Leishmania mexicana* glucose transporter 2 (LmGT2) (**A**) showing its coding sequence and transcript dimensions along with SLAS and PAS sites. **B** shows the size of the transcript as determined by Northern Blot (Burchmore and Landfear,1998).

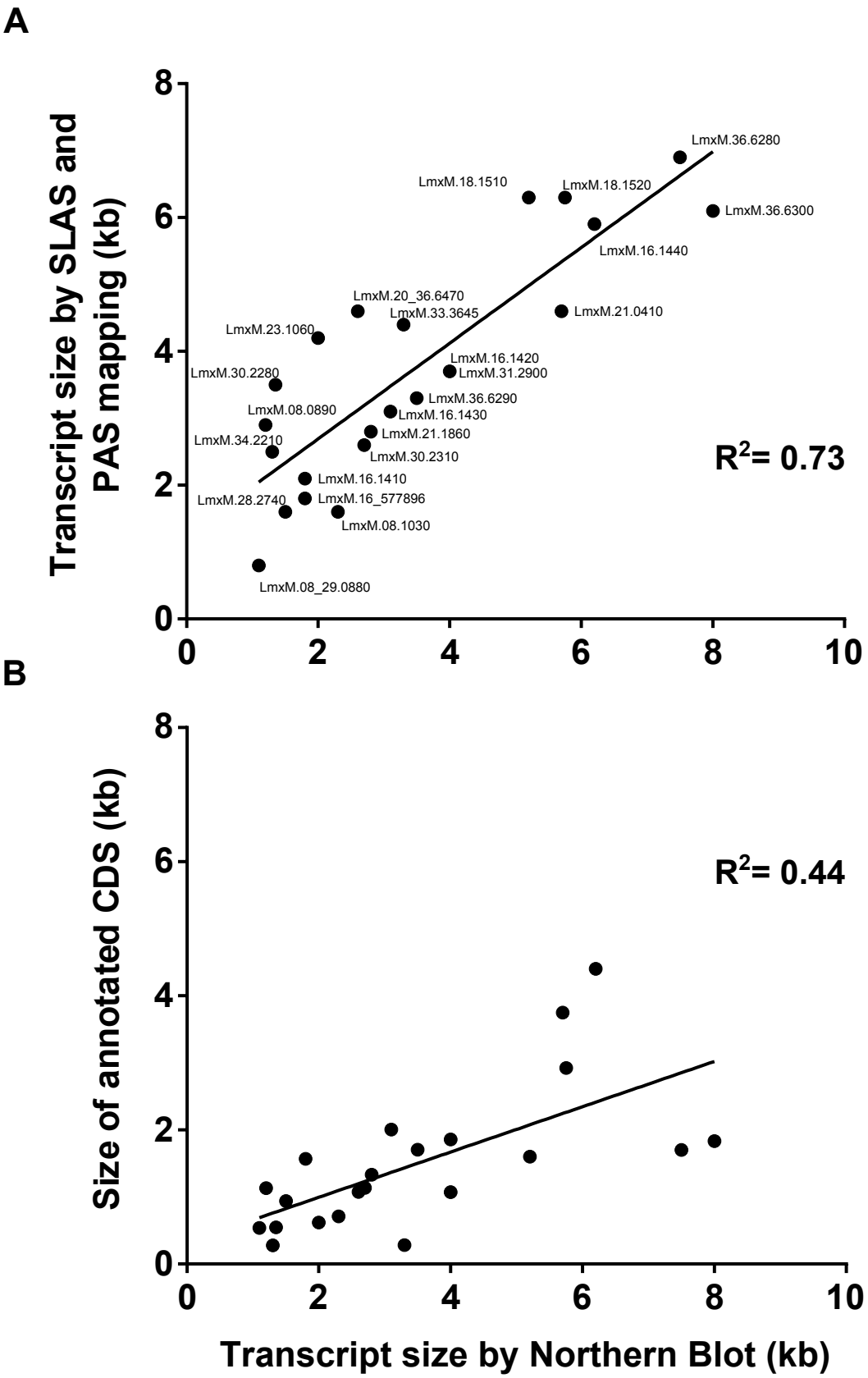


Figure 2.17 Comparison of the correlations of transcript sizes determined by SLAS and PAS mapping and reference CDS-sizes with published transcript sizes Plots correlating (**A**) transcript sizes obtained through mapping of SLAS and PAS and (**B**) reference CDS-sizes for the same genes (where available) with published transcript sizes determined by Northern Blots. Gene names in (**B**) omitted for display purposes. The R^2-value to the linear regression line is given in both plots. References for transcript sizes can be found in Table 3.5.

simply a systematic-offset, e.g. a relatively constant length of 3' and 5' UTR added to the CDS size, but instead the size-relationships between the UTRs and CDS are more variable. Bearing in mind the technical difficulty of accurately determining absolute transcript-sizes from Northern Blots, the variable dimensions of poly(A)-tails and the fact that the experimental determinations of transcript dimensions were performed in diverse *Leishamania*-species, the improved correlation of the transcript-dimensions, over CDS-size, showed that the transcript dimensions determined by SLAS-PAS-mapping made reliable predictions with regards to transcript sizes in *L. mexicana*.

Following the definition of transcript dimensions, it was possible to determine precise coordinates for 3' and 5' UTRs. The size distributions are shown in Figure 2.18 A & B. The mean 5' UTR size is 459 nt (median: 242 nt) whilst the mean 3' UTR size 912 nt (median: 584 nt).

I compared these findings with available data for *T. brucei,* where UTR sizes are considerably smaller, with the median length of 5' UTR reported to be 128 nt (Siegel et al. 2010) and 130 nt (Kolev et al. 2010), whilst 3' UTRs are reported to have median lengths of 388 nt (Siegel et al. 2010) and 400 nt (Kolev et al. 2010). To exclude biases originating from the way I and the other authors integrated multiple SLAS and PAS sites into transcript models, I extracted the available SLAS and PAS data from (Kolev et al. 2010) and processed it in exactly the same way as my own data (see Materials and Methods) to obtain mean 5' UTR sizes of 138 nt (median: 83 nt) and mean 3'UTR sizes of 675 nt (median: 416 nt) (Figure 2.18 C & D). The differences observed between the calculations of median UTR lengths reported in (Kolev et al. 2010) and obtained by me upon re-processing of the data are likely to be due to the inclusion of all SLAS and PAS for each gene in my calculations as opposed to just using the dominant SLAS or PAS as Kolev *et al*. did. Importantly, using the same processing procedure, distributions of UTR sizes observed in *L. mexicana* and *T. brucei* are significantly different based on Kolmogorov-Smirnov testing ($p < 2.2x10^{-16}$).

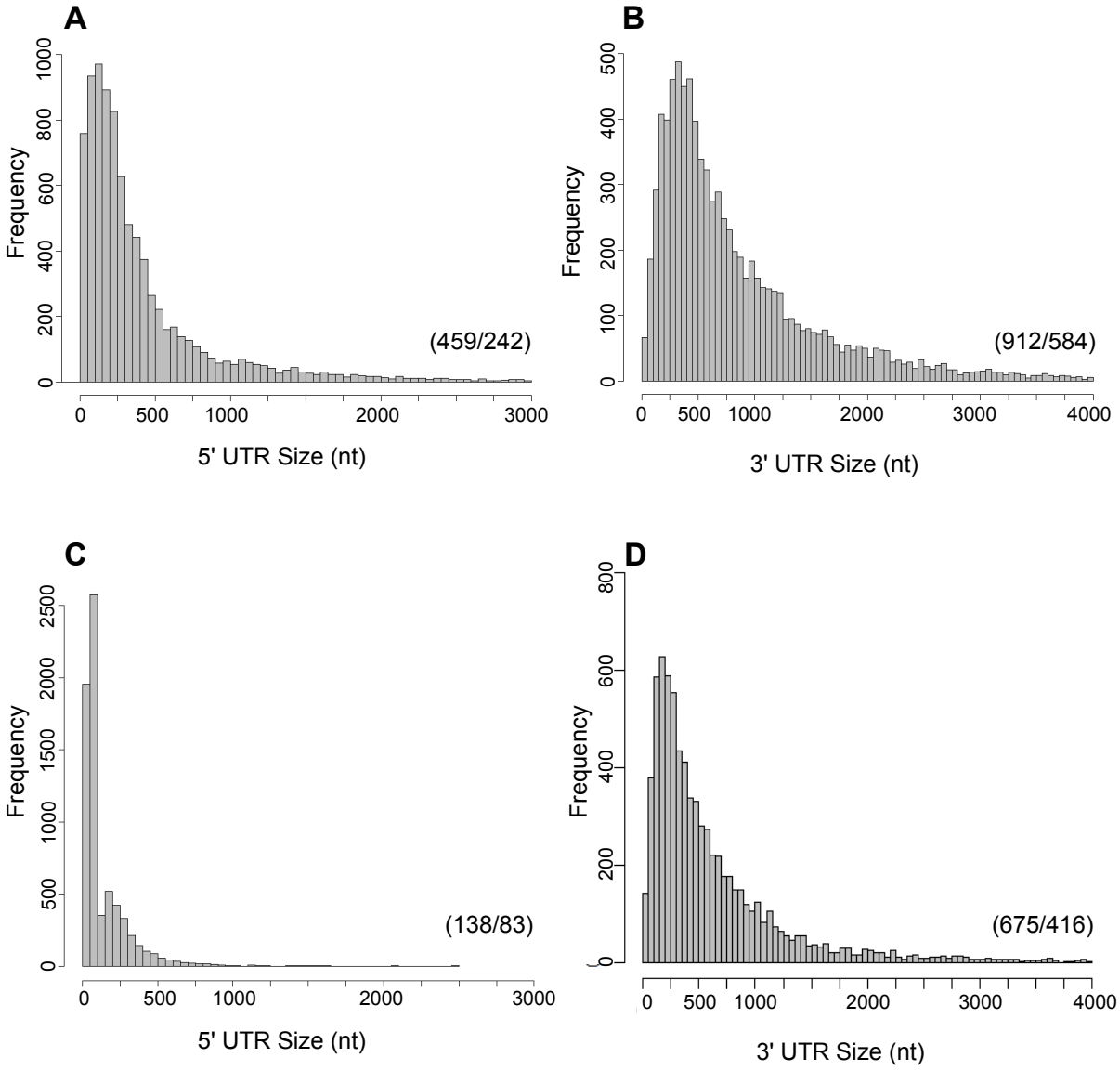


Figure 2.18 Differing UTR sizes in *L. mexicana* and *T. brucei* Histograms of (A) 5' and (B) 3' UTR sizes (both n=9029) in *L. mexicana* as well as (C) 5' and (D) 3' UTR sizes (n=6954 and n=7636 respectively) in *T. brucei* as calculated from data published by Kolev *et al*. 2010. Mean and median UTR sizes are displayed on each plot (Mean/Median).

As a result of the proposed changes to existing CDS models, changes occur to the number of predicted transmembrane domains (TMDs) (Table 2.5), signal peptides (SPs) (Table 2.6) and PFAM-domains (Tables 2.7 & 2.8). Extensions result in 96 changes to the TMD numbers (out of 1253 CDSs), ranging from loss of a single TMD (15 occurrences) to gain of 5 TMDs (one occurrence) with the dominant change being gain of a single TMD (68 occurrences). Similarly, 293 predicted SPs are gained whilst 143 are lost upon extension. Truncation leads to 18 changes of TMD numbers (out 184 CDSs), ranging from loss of 5 TMDs (one occurrence) to gain of one TMD (four occurrences), with the dominant change being loss of a single TMD (10 occurrences). Similarly, 21 predicted SPs are gained and 61 lost. Tables 2.7 & 2.8 summarise the loss and gain of PFAM-domains following extension and truncation of CDSs respectively. For details about the prediction of TMDs, SPs and PFAM-domains refer to Materials and Methods.

2.3.5 Nucleotide composition around RNA processing sites in *L. mexicana* differs to *T. brucei* and may contribute to differing UTR sizes

The sites of pre-mRNA processing in kinetoplastids were ill-defined prior to genome wide transcriptomic studies that allowed empirical determination of these (Siegel et al. 2010; Kolev et al. 2010; Nilsson et al. 2010; Rastrojo et al. 2013). No canonical motif is known for PAS and neither is the AAUAAA motif found near poly-adenylation sites (Wickens 1990) of most other eukaryotes present in kinetoplastids (Schürch et al. 1994).

SLAS are formed by an AG-dinucleotide reminiscent of cis-splice sites in other Eukaryotes (Sutton and Boothroyd 1986). A polypyrimidine rich element (or polypyrimidine tranct, PPT) lies between the PAS and SLAS and is known to affect PAS and SLAS choice (Huang and Van der Ploeg 1991; Schürch et al. 1994; Matthews, Tschudi, and Ullu 1994), however the precise mechanisms underlying splice-site choice remain elusive. Studies in *L. major* showed that trans-splicing precedes poly-adenylation and that both processes are spatially linked (LeBowitz et al. 1993).

Table 2.5 TMD gain and loss following adjustment of CDS dimensions Table summarising the number of genes with changing TMD numbers following extension and truncation of CDS based on SLAS positions.

TMD Difference	-5	-3	-2	-1	+1	+2	+3	+4	+5
Number of Genes with TMD change upon extension	0	0	0	15	68	6	5	1	1
Number of Genes with TMD change upon truncation	1	1	2	10	4	0	0	0	0

Table 2.6 SP gain and loss following adjustment of CDS dimensions Table shows the number of gained or lost SP following extension and truncation of CDS based on SLAS positions.

CDS Modification	Gained SP	Lost SP
Extension	294	143
Truncation	21	61

Table 2.7 PFAM-domain gain or loss upon extension of CDS Summary of PFAM domains gained or lost upon extension of CDS based of SLAS positions. Bit-scores and e-values for the predictions are given. PFAM domains are were precited with a Gathering Threshold cut-off.

Gain/Loss	Gene Accession	HMM Accession	HMM Name	Type	Bit-score	e-value
Gained	LmxM.03.0640	PB001461	Pfam-B_1461	Pfam-B	17.4	8.60E-04
Gained	LmxM.04.0780	PF00226.26	DnaJ	Domain	63.5	1.10E-17
Gained	LmxM.06.0290	PF01693.11	Cauli_VI	Family	60.7	9.70E-17
Gained	LmxM.07.0340	PB006794	Pfam-B_6794	Pfam-B	23.2	1.20E-05
Gained	LmxM.07.0340	PB011478	Pfam-B_11478	Pfam-B	21.2	3.30E-04
Gained	LmxM.07.0750	PF00226.26	DnaJ	Domain	28	1.30E-06
Gained	LmxM.07.0780	PF00226.26	DnaJ	Domain	34.1	1.60E-08
Gained	LmxM.09.0450	PB001728	Pfam-B_1728	Pfam-B	20	3.40E-04
Gained	LmxM.10.0330	PF01398.16	JAB	Family	22.3	7.80E-05
Gained	LmxM.11.0060	PF05773.17	RWD	Domain	28.4	1.10E-06
Gained	LmxM.12.0110	PF01363.16	FYVE	Domain	31.7	1.00E-07
Gained	LmxM.12.0850	PB002910	Pfam-B_2910	Pfam-B	28.4	7.90E-07
Gained	LmxM.12.0850	PB002673	Pfam-B_2673	Pfam-B	41.3	8.60E-11
Gained	LmxM.12.0850	PB002910	Pfam-B_2910	Pfam-B	30.4	1.90E-07
Gained	LmxM.12.0850	PB002910	Pfam-B_2910	Pfam-B	26.5	2.90E-06
Gained	LmxM.13.0120	PB008660	Pfam-B_8660	Pfam-B	19.7	8.40E-04
Gained	LmxM.13.0480	PB005973	Pfam-B_5973	Pfam-B	19.5	2.90E-04
Gained	LmxM.13.0670	PB001028	Pfam-B_1028	Pfam-B	22.2	7.40E-05
Gained	LmxM.13.0670	PB000193	Pfam-B_193	Pfam-B	23.5	1.10E-05
Gained	LmxM.13.0670	PB003777	Pfam-B_3777	Pfam-B	18	9.20E-04
Gained	LmxM.13.0670	PB000983	Pfam-B_983	Pfam-B	20.5	2.20E-04
Gained	LmxM.13.0670	PB000182	Pfam-B_182	Pfam-B	22.8	3.10E-05
Gained	LmxM.13.0670	PB002418	Pfam-B_2418	Pfam-B	20.5	1.50E-04
Gained	LmxM.13.0670	PB004388	Pfam-B_4388	Pfam-B	21.8	1.80E-04
Gained	LmxM.17.0084	PF00009.22	GTP_EFTU	Domain	181.2	1.30E-53
Gained	LmxM.17.0084	PB014832	Pfam-B_14832	Pfam-B	31.5	1.20E-07
Gained	LmxM.17.0084	PB004266	Pfam-B_4266	Pfam-B	34	2.00E-08
Gained	LmxM.21.1850	PB006430	Pfam-B_6430	Pfam-B	28.3	1.30E-06
Gained	LmxM.22.1500	PF01424.17	R3H	Domain	41.6	7.00E-11
Gained	LmxM.23.0630	PF12595.3	Rhomboid_SP	Family	79.2	3.10E-22
Gained	LmxM.26.1000	PF12999.2	PRKCSH-like	Family	115.6	1.70E-33
Gained	LmxM.26.1260	PF01363.16	FYVE	Domain	46.8	1.90E-12
Gained	LmxM.26.1670	PF08557.5	Lipid_DES	Domain	46.3	1.70E-12
Gained	LmxM.26.1810	PF13465.1	zf-H2C2_2	Domain	32.5	6.10E-08
Gained	LmxM.28.0330	PF13638.1	PIN_4	Domain	30.8	2.50E-07
Gained	LmxM.28.1095	PF01202.17	SKI	Domain	46.2	4.10E-12
Gained	LmxM.28.1165	PF13516.1	LRR_6	Repeat	9.1	1.50E+00
Gained	LmxM.28.1165	PF13516.1	LRR_6	Repeat	11.4	2.70E-01
Gained	LmxM.29.1380	PB017848	Pfam-B_17848	Pfam-B	28.1	1.50E-06
Gained	LmxM.29.1380	PB004279	Pfam-B_4279	Pfam-B	38.9	7.50E-10
Gained	LmxM.29.1760	PB010636	Pfam-B_10636	Pfam-B	26.4	7.00E-06
Gained	LmxM.29.2260	PB001083	Pfam-B_1083	Pfam-B	17.8	9.80E-04
Gained	LmxM.29.2730	PF07792.7	Afi1	Domain	26.2	6.30E-06
Gained	LmxM.30.0250	PB002827	Pfam-B_2827	Pfam-B	20	1.50E-04
Gained	LmxM.31.1360	PB011619	Pfam-B_11619	Pfam-B	25.8	1.10E-05
Gained	LmxM.31.3431	PB007543	Pfam-B_7543	Pfam-B	28.8	8.00E-07
Gained	LmxM.32.0160	PB000203	Pfam-B_203	Pfam-B	20.9	2.00E-04
Gained	LmxM.32.2360	PB013513	Pfam-B_13513	Pfam-B	48.8	9.30E-13
Gained	LmxM.32.2960	PB004658	Pfam-B_4658	Pfam-B	18.7	5.70E-04
Gained	LmxM.32.2960	PB002827	Pfam-B_2827	Pfam-B	18.7	3.60E-04
Gained	LmxM.33.1260	PB001559	Pfam-B_1559	Pfam-B	31.2	1.40E-07
Gained	LmxM.34.4040	PB016458	Pfam-B_16458	Pfam-B	19.4	9.30E-04
Gained	LmxM.36.0610	PB007186	Pfam-B_7186	Pfam-B	19.8	5.20E-04

Gained	LmxM.36.1660	PB000181	Pfam-B_181	Pfam-B	18.9	4.40E-04
Gained	LmxM.36.3070	PB013374	Pfam-B_13374	Pfam-B	33.9	1.90E-08
Gained	LmxM.36.3070	PB018400	Pfam-B_18400	Pfam-B	42.9	5.40E-11
Gained	LmxM.36.3950	PB002892	Pfam-B_2892	Pfam-B	20.2	2.90E-04
Gained	LmxM.36.3950	PB006829	Pfam-B_6829	Pfam-B	30.3	2.90E-07
Gained	LmxM.36.6160	PF08911.6	NUP50	Domain	30.5	3.30E-07
Gained	LmxM.36.6520	PB005483	Pfam-B_5483	Pfam-B	19	3.50E-04
Lost	LmxM.01.0610	PF14555.1	UBA_4	Domain	22.1	7.50E-05
Lost	LmxM.04.1080	PB008914	Pfam-B_8914	Pfam-B	21.7	1.20E-04
Lost	LmxM.04.1220	PB001290	Pfam-B_1290	Pfam-B	18.5	7.90E-04
Lost	LmxM.04.1220	PB011362	Pfam-B_11362	Pfam-B	17.9	4.50E-04
Lost	LmxM.10.0180	PF00560.28	LRR_1	Repeat	10.3	5.80E-01
Lost	LmxM.12.0260	PF00270.24	DEAD	Domain	20.8	2.10E-04
Lost	LmxM.12.0850	PF00560.28	LRR_1	Repeat	10	7.60E-01
Lost	LmxM.13.0670	PB003635	Pfam-B_3635	Pfam-B	23	9.10E-05
Lost	LmxM.13.0670	PB000152	Pfam-B_152	Pfam-B	21.7	1.20E-04
Lost	LmxM.13.0670	PB005087	Pfam-B_5087	Pfam-B	18.2	1.00E-03
Lost	LmxM.13.0670	PB006063	Pfam-B_6063	Pfam-B	18.9	1.00E-03
Lost	LmxM.13.0670	PB004962	Pfam-B_4962	Pfam-B	18.1	8.60E-04
Lost	LmxM.15.1030	PB017515	Pfam-B_17515	Pfam-B	18.9	6.80E-04
Lost	LmxM.17.0490	PB003458	Pfam-B_3458	Pfam-B	19.6	7.50E-04
Lost	LmxM.20.1080	PB001105	Pfam-B_1105	Pfam-B	27.6	3.00E-06
Lost	LmxM.26.2660	PB001252	Pfam-B_1252	Pfam-B	18.3	9.20E-04
Lost	LmxM.30.1450	PB002673	Pfam-B_2673	Pfam-B	17.9	1.00E-03
Lost	LmxM.34.4960	PB007669	Pfam-B_7669	Pfam-B	19	5.60E-04

Table 2.8 PFAM-domain gain or loss upon truncation of CDS Summary of PFAM domains gained or lost upon truncation of CDS based of SLAS positions. Bit-scores and e-values for the predictions are given. PFAM domains are were precited with a Gathering Threshold cut-off.

Gain/Loss	Gene Accession	HMM Accession	HMM Name	Type	Bit-score	e-value
Gained	LmxM.13.1220	PF00467.24	KOW	Family	21.1	1.70E-04
Gained	LmxM.17.0350	PB009580	Pfam-B_9580	Pfam-B	19	6.30E-04
Gained	LmxM.25.2235	PB002230	Pfam-B_2230	Pfam-B	17.8	9.90E-04
Gained	LmxM.36.5790	PF00560.28	LRR_1	Repeat	11	3.50E-01
Lost	LmxM.07.1070	PF13894.1	zf-C2H2_4	Domain	12.8	1.20E-01
Lost	LmxM.09.0180	PB009282	Pfam-B_9282	Pfam-B	59.8	2.50E-16
Lost	LmxM.25.0450	PB003013	Pfam-B_3013	Pfam-B	21.9	1.20E-04
Lost	LmxM.29.2030	PB012530	Pfam-B_12530	Pfam-B	34	2.10E-08
Lost	LmxM.29.2030	PB002700	Pfam-B_2700	Pfam-B	19.9	4.80E-04
Lost	LmxM.30.0221	PB005800	Pfam-B_5800	Pfam-B	20.1	3.10E-04
Lost	LmxM.34.1310	PB008261	Pfam-B_8261	Pfam-B	20	8.40E-04

To investigate whether any insight can be gained into the mechanisms underlying PAS and SLAS choice from my data set I determined the average nucleotide frequency 1000 nt up- and downstream of each SLAS in the genome of *L. mexicana* and *T. brucei* (genome-wide SLAS data of *T. brucei* kindly provided by Dr. Steven Kelly). Plotting the C:U-content (i.e. pyrimidine-base content) clearly identifies the PPT before the SLAS (Figure 2.19). By averaging the predicted propensity for secondary RNA structure (mean base-pairing probability) (Lorenz et al. 2011) of all sequences around SLAS (see Materials and Methods), we can see drastically changing mean base-pairing probability around the SLAS. Indeed, the A of the AG-dinucleotide is the least-likely base to be base-paired and the G of the AG-dinucleotide is the most likely to be base-paired. This suggests a model with a strict secondary structure requirement for splicing.

Furthermore, we can see that the PPT introduces a region of low predicted secondary structure around the SLAS, which may play a role in permitting access of splicing factor to the splice-site.

In *T. brucei*, the PPT appears slightly smaller using this type of analysis (Figure 2.19), which has a remarkable effect on the mean base-pairing probability around the SLAS. First, the region of low predicted secondary structure preceding the SLAS is shorter than in *L. mexicana*. Second, the region of low secondary structure downstream of the SLAS is also shorter compared *L. mexicana*. As I showed in Section 2.3.4, the 5'UTRs in *L. mexicana* are longer than in *T. brucei* and one could speculate that there could be a link between the longer region of low secondary structure downstream of the SLAS and the longer absence of coding sequence downstream of the SLAS (i.e. the longer 5' UTR).

If we consider the longer stretch of low secondary structure upstream of the SLAS in *L. mexicana* compared to *T. brucei*, one may wonder how this affects the distance between PASs and the downstream SLAS. Comparison of *T. brucei* (Kolev et al. 2010) and *L. mexicana* (Figure 2.20 A & B) shows that the PAS-to-SLAS distance is larger in *L. mexicana*

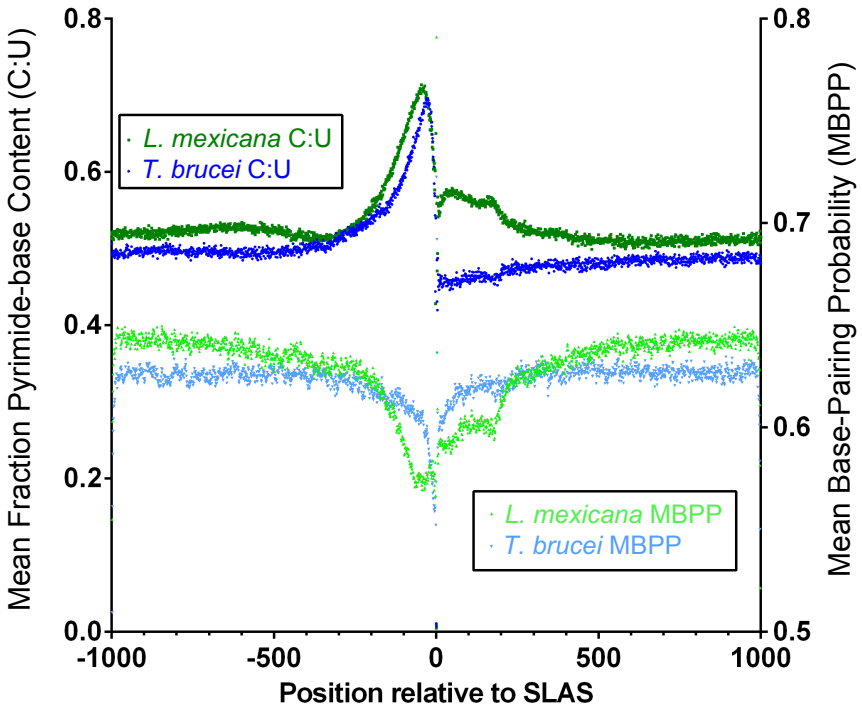


Figure 2.19 Secondary RNA structure analysis around SLAS Plots showing the mean fraction of pyrimidine-base content in positions 1000 nt upstream and downstream of each SLAS in the genome. From these averaged sequences the mean base-pairing probability for each position was calculated and plotted here.

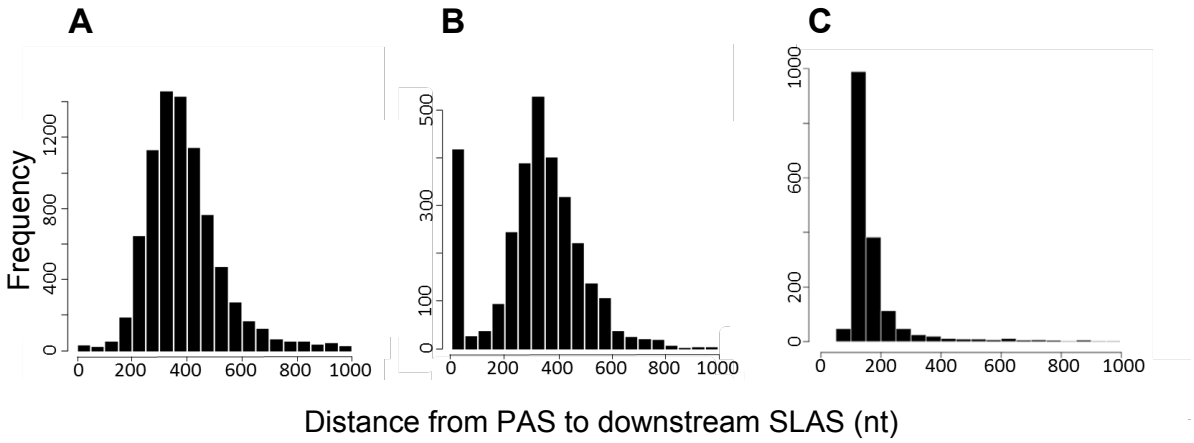


Figure 2.20 PAS-to-SLAS distances comparison across selected trypanosomatids Histograms of PAS-to-SLAS distances in (A) *L. mexicana* (n=8034), (B) *L. major* (n= 3087) (Rastrojo *et al.* 2013) and (C) *T. brucei* (n= 1759) (from Kolev *et al.* 2010). A and B were generated by me using dominant SLAS and PAS, C was taken from the original publication due to inconsistent information regarding the genome version utilised by the authors (Kolev et al. 2010) and the resulting ambiguity in recreating the gene order along chromosomes required for this analyis. The median distances observed in *L. mexicana* and *L. major* are 368 nt and 329 nt respectively, whilst a median distance of 142 nt is reported for *T. brucei* (Kolev *et al.* 2010).

than in *T. brucei* with a median distance of 142 nt and 368 nt respectively. This observation is consistent with analysis of the PAS-to-SLAS distance in *L. major* (Rastrojo et al. 2013) (Figure 2.20 C) which, at 329 nt median PAS-to-SLAS distance, is very similar to *L. mexicana.* Additionally this reflects the larger reported mean size of intergenic sequences in *L. major* (1731 nt) compared to *T. brucei* (1511 nt) (El-Sayed et al. 2005).

2.3.6 Predicted Novel genes encode predominantly small proteins

Considering the detection of uORFs for annotated genes, it is clear that a subset of novel genes will also contain these and I will address this in section 2.3.9. For the moment however, we will assume that the first ORF encoding a protein ≥25 AA is the CDS in each novel transcript. The proteins predicted to be encoded by the novel transcripts have a mean length of 80 AA (median: 58 AA) making them much smaller than proteins in the reference annotation (mean: 628 AA, median: 465 AA) (Figure 2.21 A-C).

Discovery of predicted small proteins in novel transcripts is a common feature of RNA-sequencing studies performed not only in kinetoplastids (Kolev et al. 2010; Rastrojo et al. 2013) but also in mouse, humans, zebrafish, fruit fly, Arabidopsis and yeast (Ruiz-Orera et al. 2014). Especially the availability of the data for other kinetoplastids opens up the possibility of investigating whether orthologues of the same genes were identified in RNA-sequencing studies of *L. major* (Rastrojo et al. 2013) , *T. brucei* (Kolev et al. 2010) and *L. mexicana* (this study). Figure 2.22 shows a Venn-diagram with the overlap of novel genes detected between these studies. Comparison was performed using a reciprocal best tBlastx method (see Materials and Methods) (Hirsh and Fraser 2001; Jordan et al. 2002) to exclude the possibility of not identifying an orthologue due to definition of the wrong ORF within a transcript. 29 novel transcripts are shared between all three studies. 46 transcripts identified as novel in *T. brucei* (Kolev et al. 2010) were also identified in this study. The biggest overlap exists between the novel genes identified in *L. major* and in this study with 598 being shared. Over 1000 genes appear unique to *T. brucei* and *L. major*,

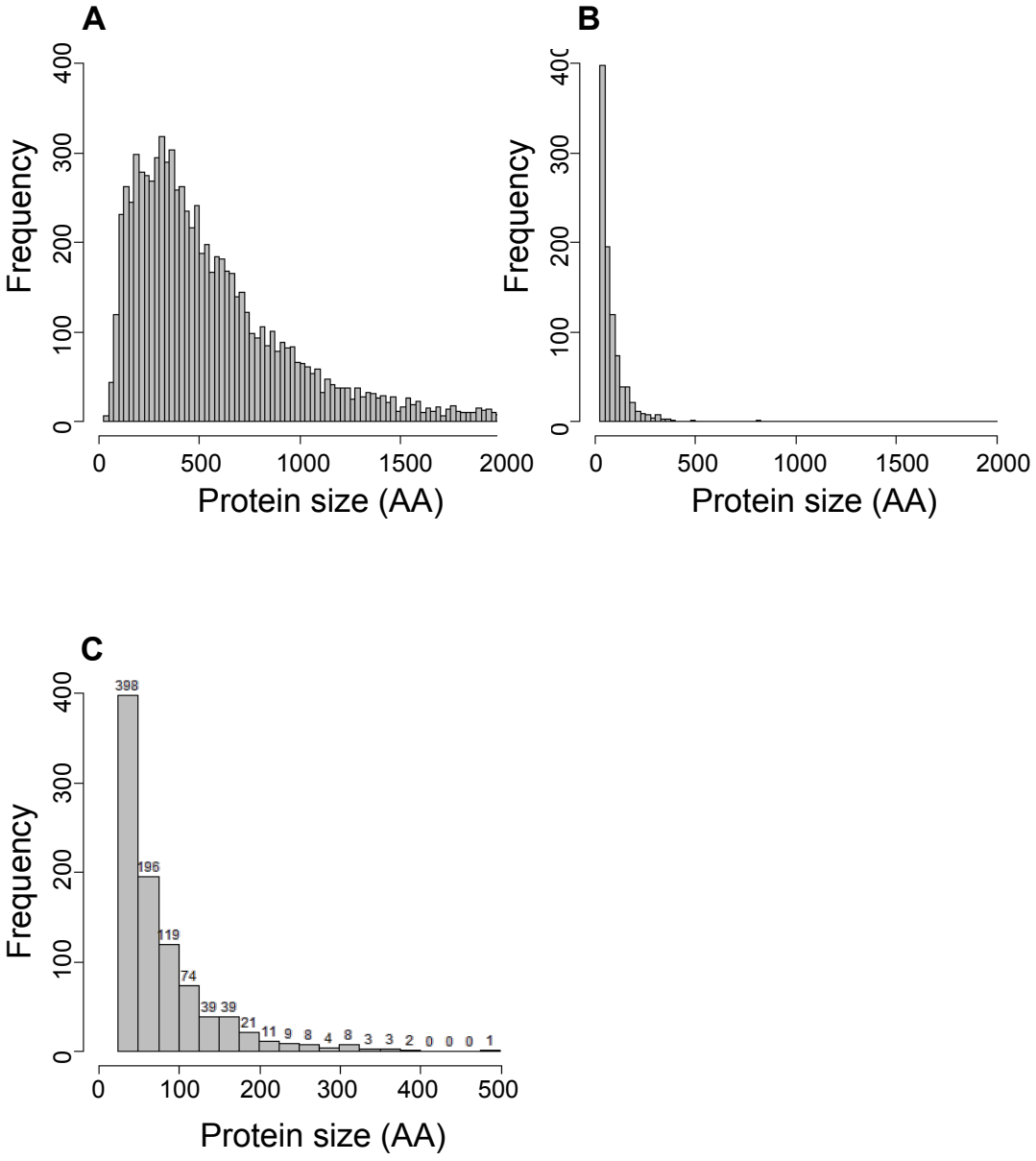


Figure 2.21 Size distribution of proteins from reference annotation versus predicted novel proteins Histograms showing (A) the distribution of protein sizes in the reference annotation of the *L. mexicana* genome (v7) (n=8250), (B) the sizes of proteins predicted to be encoded by novel transcripts identified by RNA-sequencing on the same scale as (A), and (C) a close-up view (n= 936) (For display purposes 235 proteins ranging between 2004 and 6737 AA omitted in (A) and one protein >500 AA in (C).

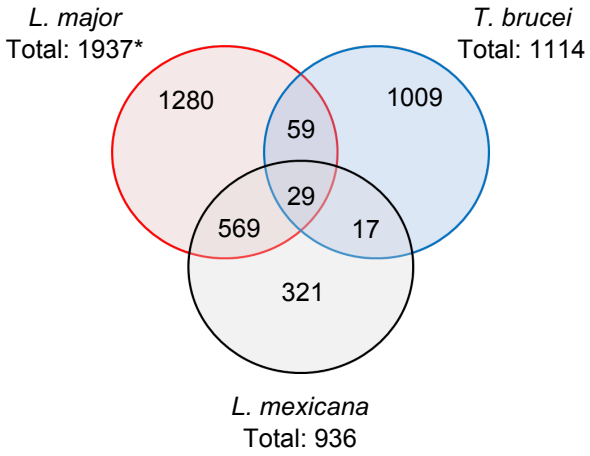


Figure 2.22 Overlap between novel genes discovered in different RNA-sequencing studies of trypanosomatids Venn diagram showing the degree of overlap between novel genes discovered in *T. brucei* by Kolev et al (2010), in *L. major* by Rastrojo et al. (2013) and in *L. mexicana* in this study. Overlap was determined by Reciprocal Best tBlastx (Materials and Methods).

*The number of novel genes discovered in *L. major* is stated as being 1884 in Rastrojo et al. (2013) but genomic coordinates are provided for 1937.

suggesting that the novel genes discovered in these studies represent an important contribution to inter-species differences.

To characterise further the proteins predicted to be encoded by the novel transcripts, TMDs and SPs were predicted for the first ORF encoding a protein >25 AA in the novel transcripts. 95 proteins are predicted to have TMDs, with TMD numbers ranging between 1 and 4. 224 proteins are predicted to have a SP and 52 proteins both TMDs and a SP (Table 2.9) .

PFAM-domains were also predicted for the novel proteins. Here however, the entire transcript was translated in 3 frames and all resulting sequences searched for PFAM domains (see Materials and Methods). This, again, is to exclude failing to identify a PFAM domain due to definition of the wrong ORF within a transcript.

Due to the size of the table the results of the prediction are shown in Supplementary Table 2.2. 58 transcripts are predicted to encode for a protein with a PFAM-domain, with 35 having PFAM-A domains. Amongst those are 6 zinc-finger proteins, 3 ribosomal proteins, 2 histones and 2 amastins. Curiously, not all transcripts have predicted PFAM domains in only one frame of translation, whilst the identity of the predicted PFAM domain can be the same between translational frames. These may primarily be due to the repetitive nature of some predicted PFAM-B domains e.g. the repeating Cys-Val motif (which may on the DNA level be encoded by “TGT” and “GTG”, respectively) corresponding to PFAM-B_11478 and PFAM-B_9129. It is possible that these repetitive motifs may indicate false positives in the detection of functional protein domains in the proteins encoded by these novel transcripts and have to be treated with care.

Table 2.9 TMD and SP numbers of novel proteins Summary of the numbers of TMDs and SPs predicted for novel proteins.

Predicted to have:	1 TMD	2 TMD	3 TMD	4 TMD	SP	SP & TMD
Number of Genes	71	19	3	2	224	52

2.3.7 Evidence for novel genes and extended gene models found by mass spectrometry promastigotes and axenic amastigotes

To validate the predicted novel proteins and extended gene models, I sought to gather proteomic-evidence for these using mass-spectrometry. Whole-cell lysates of PRO and AXA were made and analysed by mass-spectrometry (see Materials and Methods). The raw results were then pooled with other mass-spectrometric data sets from the Gluenz laboratory courtesy of François Demay and Tom Beneke (see Materials and Methods). The combined data sets were then searched against a data-base containing the annotated *L. mexicana* proteome, however with N-terminally extended protein sequences corresponding to 5' extended gene-models (data base supplied in Supplementary Material) Furthermore, the database contained three frame translations of novel transcripts, uORFs found upstream of reference genes.

Using this combined data set, 3904 proteins were identified with a p-value ≥ 0.95 for their correct identification. For 42 novel proteins unique peptides were detected, with 40 of these proteins having a probability of being correctly identified p ≥ 0.95 (Supplementary Table 2.3). Indeed, one identification based on a unique peptide, LmxM.26_815658, was disregarded due to poor confidence on the unique peptide. The shortest of these proteins, LmxM.09_94875, is 39 AA long. The mean length of the detected novel proteins is 141.8 AA with a median of 129 AA.

Previously, Paape and co-workers (Paape et al. 2010) published a proteomic analysis of intracellular amastigotes. We obtained the raw data from this study and re-analysed using our novel gene models to obtain further mass-spectrometric evidence for novel proteins Supplementary Table 2.4 shows the results of this analysis (see Materials and Methods for analysis parameters). In total, unique peptides for 20 novel genes were detected (all p-value ≥0.95), of which 5 had not been not detected in the PRO and AXA data-sets from our laboratory. The mean length of the novel proteins detected was 150.7 AA with a median of

138 AA. In total we have obtained peptide-evidence for 47 novel proteins (40 p-value >= 0.95).

To find evidence for predicted N-terminal extensions, all unique peptides from proteins predicted to have N-terminal extensions were identified as described in Materials and Methods. For this analysis only data from PRO and AXA samples prepared in our laboratory were used. Of the 1253 proteins predicted to be extended, 433 were detected with peptides corresponding to any part of the proteins. For 116 proteins peptides originating from predicted N-terminal extensions were identified. Proteins with proven N-terminal extensions along with the peptide evidence for these are shown in Table 2.10. An example of peptides mapping an extension is shown in Figure 2.23.

To address whether any more extensions could in theory be detected, an *in silico* tryptic digest (allowing up to 2 missed cleavages) of the 433 detected proteins was performed (see Materials and Methods). All peptides larger than 6 AA and smaller than 29 AA (95^{th} size-percentile of peptides detected in my mass-spectrometric data set) were extracted and mapped back onto the protein sequences. In this analysis tryptic peptides could in theory be obtained from 411 of the 433 mass-spectrometrically detected proteins with proposed extensions.

This shows that, N-terminal protein extensions proposed through mapping of SLAS can be confirmed by mass-spectrometry, however computational prediction does indicate that more extensions could in theory be identified. This may be a technical artefact or it may indicate the first in-frame start ATG-codon does not always form the start codon. The latter option will certainly apply to some of these proteins as findings in *T. brucei* gained through profiling of ribosome binding patterns showed that ATG-codons may be skipped in favour of a further downstream translation initiation site (Vasquez et al. 2014).

No peptide-evidence was detected for the translation of uORFs.

Table 2.10 Peptide-evidence for predicted N-terminal protein extensions Table showing the reference and extended coordinates of genes for which unique peptides were identified that map to the extensions. The genomic coordinates corresponding to the peptides are also shown. Genomic coordinates are given in GFF format.

Accession	Chromosome	Strand	Ref. GFF start	Ref. GFF stop	Ext. GFF start	Ext. GFF end	Peptide GFF start	Peptide GFF end	Peptide Sequence
LmxM.01.0320	LmxM.01	+	78587	79510	78491	79510	78494	78565	TQSVSIGADHRLEATTSDCVSSVR
LmxM.01.0320	LmxM.01	+	78587	79510	78491	79510	78494	78526	TQSVSIGADHR
LmxM.03.0080	LmxM.03	+	20144	21826	19673	21826	20078	20104	VNLVEEVSR
LmxM.03.0200	LmxM.03	+	49414	51030	49348	51030	49390	49449	SPPVLNEPMDNFEPGSASAK
LmxM.03.0210	LmxM.03	+	52958	53821	52682	53821	52931	52996	SEEKDVYVAMDYIPADLSSVIK
LmxM.03.0230	LmxM.03	+	56984	58954	56837	58954	56909	56947	EYSSYGAQSVAVK
LmxM.03.0230	LmxM.03	+	56984	58954	56837	58954	56900	56947	AYREYSSYGAQSVAVK
LmxM.03.0430	LmxM.03	+	135335	135628	135296	135628	135299	135391	SAETLACTYAALMLSDAGLPTSAENIAAAVK
LmxM.03.0430	LmxM.03	+	135335	135628	135296	135628	135299	135439	SAETLACTYAALMLSDAGLPTSAENIAAAVKAAGVEMRPTLPIIFAR
LmxM.03.0690	LmxM.03	-	250685	256870	250685	257320	256907	256957	FLASFAGQPAAIVCIQK
LmxM.03.0690	LmxM.03	-	250685	256870	250685	257320	256856	256906	ALTEVAVDFHYAMARPR
LmxM.04.0050	LmxM.04	+	23094	24668	23073	24668	23079	23195	SSELIMEQGPSRLEAHSMTNESPSNQAHEQSSLEPQLEK
LmxM.04.0050	LmxM.04	+	23094	24668	23073	24668	23079	23114	SSELIMEQGPSR
LmxM.04.0580	LmxM.04	-	221584	222435	221584	222486	222406	222483	PGPGLLSDGWFREESTMWPGQAQGLK
LmxM.04.0580	LmxM.04	-	221584	222435	221584	222486	222448	222483	PGPGLLSDGWFR
LmxM.04.0630	LmxM.04	-	245957	246553	245957	246559	246461	246556	TMNFGNMTLGGAMATFGGQSNPMCNYTSPLAK
LmxM.04.1000	LmxM.04	-	353933	357340	353933	357457	357398	357457	MQLSPGDFEGAVAGPPLHYR
LmxM.04.1160	LmxM.04	-	395875	396837	395875	396930	396811	396855	GEFTLLMMAIQTSVK
LmxM.04.1160	LmxM.04	-	395875	396837	395875	396930	396877	396915	TPTPTTLTQYIIK
LmxM.04.1160	LmxM.04	-	395875	396837	395875	396930	396877	396918	RTPTPTTLTQYIIK
LmxM.06.0350	LmxM.06	-	118082	119173	118082	119305	119258	119293	ELNPLLSPEQKR
LmxM.06.0350	LmxM.06	-	118082	119173	118082	119305	119159	119182	LVEMLVER
LmxM.06.0350	LmxM.06	-	118082	119173	118082	119305	119183	119221	CVVTGGTGFVGTR
LmxM.06.0540	LmxM.06	+	186354	188474	185883	188474	186267	186377	APPYADATEVASHAHSAAAPEDDAGPVESMNSSAATR
LmxM.07.0340	LmxM.07	-	152493	153740	152493	154172	153783	153812	SEEEIATWLR
LmxM.07.0340	LmxM.07	-	152493	153740	152493	154172	153753	153782	ENSITIYGDR
LmxM.08.0080	LmxM.08	+	1177310	1179367	1176767	1179367	1177019	1177078	AQAGDATSATAQNDPAAFLR
LmxM.08.0560	LmxM.08	+	1380147	1380860	1380030	1380860	1380108	1380158	SPAGQFVNVVPEGMQPR
LmxM.09.0400	LmxM.09	+	146140	147408	145999	147408	145999	146046	MDSSSAAADAGGAAAR
LmxM.09.1220	LmxM.09	-	449224	453048	449224	453333	453211	453294	SSVGPPAEPLQGTPTHATAAAAAATLPR
LmxM.09.1320	LmxM.09	-	485712	487529	485712	487532	487488	487532	MMSSSAIVLPTYPSR
LmxM.09.1320	LmxM.09	-	485712	487529	485712	487532	487461	487532	MMSSSAIVLPTYPSRDVQNVENHR
LmxM.09.1430	LmxM.09	-	512621	513184	512621	513337	513290	513337	MQWVPFIATGAALAHR
LmxM.09.1430	LmxM.09	-	512621	513184	512621	513337	513155	513259	AIRDPTDVDAVSAVGELSSLDSLENMKNCMMADQR
LmxM.09.1510	LmxM.09	-	530862	535496	530862	535790	535653	535676	IAVIYEAR
LmxM.09.1510	LmxM.09	-	530862	535496	530862	535790	535608	535652	KGAQPSAEEIESQLR
LmxM.09.1510	LmxM.09	-	530862	535496	530862	535790	535581	535607	VAQYPLAPR
LmxM.09.1540	LmxM.09	-	538884	540707	538884	540875	540828	540872	SGTPAIGAHENLASR
LmxM.10.0160	LmxM.10	+	54587	56083	53969	56083	54284	54331	MNATGMNVSALSPHGR
LmxM.10.0270	LmxM.10	+	102295	105075	102154	105075	102157	102222	PSSSVQGTSSSALDGEGAHVAR
LmxM.10.0370	LmxM.10	+	150495	152303	150216	152303	150372	150416	RVPVFGEAVEGYGPK
LmxM.10.0370	LmxM.10	+	150495	152303	150216	152303	150375	150416	VPVFGEAVEGYGPK
LmxM.10.0370	LmxM.10	+	150495	152303	150216	152303	150456	150479	GVADQILR
LmxM.10.0370	LmxM.10	+	150495	152303	150216	152303	150321	150356	YIHPEAASLFAR
LmxM.10.0370	LmxM.10	+	150495	152303	150216	152303	150480	150530	GQTYAMMMDRYGIDVAR
LmxM.10.0370	LmxM.10	+	150495	152303	150216	152303	150480	150509	GQTYAMMMDR
LmxM.10.0700	LmxM.10	-	300384	302843	300384	303260	303135	303176	EAPLLEITDVQFIR
LmxM.11.0090	LmxM.11	+	25197	26777	24777	26777	24855	24893	TITTVDSYVGGTR
LmxM.11.0400	LmxM.11	+	129157	132495	128608	132495	128866	128898	RVTADLQAAHK
LmxM.11.0400	LmxM.11	+	129157	132495	128608	132495	128929	128955	ALSVQLESR
LmxM.11.0590	LmxM.11	+	210036	211640	209565	211640	209802	209900	SAEEPRVEATPTPSEATAAPAAAATPAPATLIK
LmxM.11.0590	LmxM.11	+	210036	211640	209565	211640	209958	210035	STSQSPNILRFPANTSDPTFQENMAR
LmxM.11.0590	LmxM.11	+	210036	211640	209565	211640	209988	210035	FPANTSDPTFQENMAR
LmxM.11.0590	LmxM.11	+	210036	211640	209565	211640	209958	209987	STSQSPNILR
LmxM.11.0590	LmxM.11	+	210036	211640	209565	211640	209925	209957	LYAHHPIDYER
LmxM.11.0820	LmxM.11	+	302167	303180	301753	303180	301825	301875	HTPAKEEASVSTGPQVK
LmxM.11.0820	LmxM.11	+	302167	303180	301753	303180	301915	301974	SFLDALRSGKKPAAPIAAAR
LmxM.11.0820	LmxM.11	+	302167	303180	301753	303180	301876	301935	SQISNSAPAPAGKSFLDALR
LmxM.11.0820	LmxM.11	+	302167	303180	301753	303180	301975	302067	VTVPAAAPAEDTQAAAAAAPPAPAAVVTPAK
LmxM.11.0820	LmxM.11	+	302167	303180	301753	303180	301975	302103	VTVPAAAPAEDTQAAAAAAPPAPAAVVTPAKESPAPEPPKEER
LmxM.11.0820	LmxM.11	+	302167	303180	301753	303180	301936	301974	SGKKPAAPIAAAR
LmxM.11.0820	LmxM.11	+	302167	303180	301753	303180	301876	301914	SQISNSAPAPAGK
LmxM.11.0890	LmxM.11	+	345909	348932	345891	348932	345894	345977	GDTAAMTPLPQQPPTPPVASPYITPEPK
LmxM.12.0640	LmxM.12	+	327210	328841	326595	328841	326820	326900	TPAQLLANHSGGDSQGAVSDVVSPVER
LmxM.12.0850	LmxM.12	+	359820	360620	357480	360620	357612	357644	NTLTVLQAFAR
LmxM.12.0850	LmxM.12	+	359820	360620	357480	360620	358314	358337	GLTIVILR
LmxM.12.0850	LmxM.12	+	359820	360620	357480	360620	357948	357974	SLTSLTVEK
LmxM.13.0040	LmxM.13	-	9231	10115	9231	10298	10116	10148	LLVAQEAALVR
LmxM.13.0120	LmxM.13	-	33874	34233	33874	34344	34210	34320	VPSAAAAGSTAAPTDAGSASSSPTLTAEEMNEFAELR
LmxM.13.0140	LmxM.13	-	42045	42605	42045	42809	42657	42707	SAPWVTPEVAPAYPSAR
LmxM.13.0480	LmxM.13	+	129460	131574	129298	131574	129298	129333	MWASFLDQIQQR
LmxM.13.1110	LmxM.13	-	362212	363762	362212	363813	363748	363795	TVAAAAAATTSMSSTR
LmxM.14.0220	LmxM.14	+	56506	57870	56155	57870	56155	56199	MLSNNANPFEEEYQR
LmxM.14.0290	LmxM.14	+	79726	80568	79537	80568	79549	79599	LLLFGGTGFVGSLVAQK
LmxM.14.1440	LmxM.14	+	590416	593316	590332	593316	590416	590490	MYAQMPEIYLCPCCSEFSKQEERER
LmxM.14.1440	LmxM.14	+	590416	593316	590332	593316	590335	590391	SLFGEALPFEDDTATSLIR
LmxM.14.1440	LmxM.14	+	590416	593316	590332	593316	590335	590394	SLFGEALPFEDDTATSLIRR
LmxM.15.0130	LmxM.15	-	40165	42114	40165	42498	42118	42177	TAGAAGAAATIEGSEQTQEK
LmxM.16.0180	LmxM.16	-	62736	63188	62736	63500	63171	63233	RPFNAAWSNSGSTSLMEDAPR
LmxM.16.0180	LmxM.16	-	62736	63188	62736	63500	63441	63497	SDKRPGVPNSGGYVAPNAR
LmxM.16.0980	LmxM.16	+	349236	351431	347955	351431	348906	348965	SLVHYVAPPGDTAIPSYGCR
LmxM.16.0980	LmxM.16	+	349236	351431	347955	351431	348966	349007	DIATGSDAGYSPSR
LmxM.16.0980	LmxM.16	+	349236	351431	347955	351431	349155	349205	YADEAAAAREEHEAQTR
LmxM.16.1420	LmxM.16	-	567662	568732	567662	569242	568868	568909	AASSTSGASASQAR
LmxM.16.1420	LmxM.16	-	567662	568732	567662	569242	569072	569119	CQSIAHDDLTVAEGPR
LmxM.16.1420	LmxM.16	-	567662	568732	567662	569242	569150	569230	EQSPSIMIDSDANSLTSSPSYSEHNIR
LmxM.16.1420	LmxM.16	-	567662	568732	567662	569242	568724	568768	FATTVSQTPSASMPR
LmxM.16.1420	LmxM.16	-	567662	568732	567662	569242	568769	568846	LSTASAATPASPAAGATRPLSVPAAR
LmxM.16.1420	LmxM.16	-	567662	568732	567662	569242	568997	569050	NGSRVPIQTNSIDAEKPR
LmxM.16.1420	LmxM.16	-	567662	568732	567662	569242	568910	568996	TASMNSNGALQPTVTVPNLHLSGLQAQLR
LmxM.16.1420	LmxM.16	-	567662	568732	567662	569242	568997	569038	VPIQTNSIDAEKPR
LmxM.16.1460	LmxM.16	-	587206	594904	587206	595222	595037	595102	GAPTTPAAAGPAGSPAPPLPPR
LmxM.16.1460	LmxM.16	-	587206	594904	587206	595222	594962	595000	ESPRPQSTTAFER
LmxM.16.1460	LmxM.16	-	587206	594904	587206	595222	595013	595036	TPEPPSAR
LmxM.17.0150	LmxM.17	-	103723	104895	103723	105852	105103	105135	AEEAVAELKPR
LmxM.17.0150	LmxM.17	-	103723	104895	103723	105852	105316	105384	AGSGSSGTSHSETLFAVSAALPR
LmxM.17.0150	LmxM.17	-	103723	104895	103723	105852	105673	105717	DRFDDAGDVAASVLR
LmxM.17.0150	LmxM.17	-	103723	104895	103723	105852	105739	105792	DRPMGSAVTVDLAALLTR
LmxM.17.0150	LmxM.17	-	103723	104895	103723	105852	105628	105672	IEALADDVLDYLLHR
LmxM.17.0150	LmxM.17	-	103723	104895	103723	105852	105811	105852	MDVIPLVAQEIVAR
LmxM.17.0150	LmxM.17	-	103723	104895	103723	105852	105136	105204	TAGVDLMSLSTTVDTPLAELEAK
LmxM.17.0150	LmxM.17	-	103723	104895	103723	105852	104902	104922	LLLAELR
LmxM.17.0150	LmxM.17	-	103723	104895	103723	105852	105292	105315	AQVTAFLR
LmxM.17.0300	LmxM.17	-	139937	140545	139937	140749	140579	140632	GVAAATGGGSASLDSEKR
LmxM.17.0300	LmxM.17	-	139937	140545	139937	140749	140633	140671	GTADESSTGTDAR
LmxM.17.0760	LmxM.17	-	311322	312296	311322	312575	312384	312440	EGVGGVPPPPAPPVNTAER
LmxM.17.0810	LmxM.17	-	382511	383650	382511	383965	383798	383839	GAVSASHLATEGTR
LmxM.18.0350	LmxM.18	-	142215	143129	142215	143414	143142	143198	NVPSEPGAHAELSTAELER
LmxM.20.0230	LmxM.20	-	2714061	2714510	2714061	2714636	2714529	2714576	AAGSSSPASASSLDAR
LmxM.20.0300	LmxM.20	+	2745692	2746624	2745392	2746624	2745623	2745685	SASTNIYGAEDPNLTCPAELR
LmxM.20.1520	LmxM.20	-	3280506	3285257	3280506	3285506	3285276	3285314	SFNSDVDAAQQQR
LmxM.20.1520	LmxM.20	-	3280506	3285257	3280506	3285506	3285366	3285395	AAGGGTLHER
LmxM.20.1650	LmxM.20	-	3296285	3296710	3296285	3296806	3296744	3296803	SVAPQYYHAHAVDATIQHEK
LmxM.20.1650	LmxM.20	-	3296285	3296710	3296285	3296806	3296732	3296803	SVAPQYYHAHAVDATIQHEKAYQR
LmxM.20.1650	LmxM.20	-	3296285	3296710	3296285	3296806	3296693	3296731	QTAVNENMHRPSR
LmxM.22.0480	LmxM.22	-	183333	184667	183333	185126	184761	184838	AGDGSGASAALPSLSSASSHAAAQAR
LmxM.22.0980	LmxM.22	+	391817	393943	391169	393943	391523	391585	SAATLLAEATGAPPPPPAPSR

LmxM.22.0980	LmxM.22	+	391817	393943	391169	393943	391172	391213	SNSYSPFSTAPVAR
LmxM.22.1500	LmxM.22	+	605330	608593	604934	608593	604976	605041	RGDHSQQGYLAAQGINDQLSEK
LmxM.22.1500	LmxM.22	+	605330	608593	604934	608593	605261	605326	AFVDDKDSNELSFPSTLTPAQR
LmxM.23.0020	LmxM.23	+	5959	7110	5809	7110	5848	5898	LLDALFDNTSSFGILSK
LmxM.23.0020	LmxM.23	+	5959	7110	5809	7110	5848	5901	LLDALFDNTSSFGILSKR
LmxM.23.0020	LmxM.23	+	5959	7110	5809	7110	5812	5841	SVDYLNAVER
LmxM.23.0020	LmxM.23	+	5959	7110	5809	7110	5812	5847	SVDYLNAVERSR
LmxM.23.0020	LmxM.23	+	5959	7110	5809	7110	5842	5898	SRLLDALFDNTSSFGILSK
LmxM.23.0560	LmxM.23	-	249592	251562	249592	252054	251638	251682	NITSEDAAEGTEAFK
LmxM.23.0560	LmxM.23	-	249592	251562	249592	252054	251632	251682	NITSEDAAEGTEAFKNR
LmxM.23.0560	LmxM.23	-	249592	251562	249592	252054	251542	251595	VRPFIKEELAEMEGQERR
LmxM.23.0560	LmxM.23	-	249592	251562	249592	252054	251608	251631	VSEITSFR
LmxM.23.0560	LmxM.23	-	249592	251562	249592	252054	251545	251577	EELAEMEGQER
LmxM.23.1110	LmxM.23	-	507639	509732	507639	510011	509949	510008	SSSGGPSTLLGSSPSKDVAR
LmxM.26.0850	LmxM.26	-	232893	233858	232893	234146	233844	233903	AGDGGEDPFAFEFDPMAGVR
LmxM.26.1000	LmxM.26	-	287461	288474	287461	288963	288454	288480	AVMHSAAVR
LmxM.26.1020	LmxM.26	+	293630	305767	293411	305767	293546	293599	FQSPSSFIFHGSSAASEK
LmxM.26.1020	LmxM.26	+	293630	305767	293411	305767	293411	293437	MNPDNSSFR
LmxM.26.1020	LmxM.26	+	293630	305767	293411	305767	293411	293440	MNPDNSSFRK
LmxM.26.1020	LmxM.26	+	293630	305767	293411	305767	293519	293545	TQNAPQLYK
LmxM.26.1550	LmxM.26	+	534509	536689	534245	536689	534431	534463	NRLWQTHQAGR
LmxM.26.1550	LmxM.26	+	534509	536689	534245	536689	534494	534517	IAWIVMNR
LmxM.26.2490	LmxM.26	+	943347	947780	943095	947780	943095	943169	MESHMDPVGEVGSVSPEHQPSAGVR
LmxM.26.2600	LmxM.26	+	987077	989158	986711	989158	986792	986860	GGLSELNESSDNFDDVVAHVTSR
LmxM.26.2600	LmxM.26	+	987077	989158	986711	989158	986861	986914	AAAPPSSSAGVTEVGVPR
LmxM.26.2600	LmxM.26	+	987077	989158	986711	989158	986987	987064	AQTDTQTDPGGGPSAADVDGAASLHR
LmxM.27.0160	LmxM.27	-	37680	38465	37680	38675	38433	38471	IDMNASFGGSFAR
LmxM.27.2130	LmxM.27	-	888045	893126	888045	893600	893133	893174	SLLLGYGGGPSLAR
LmxM.27.2180	LmxM.27	-	914474	915337	914474	915652	915413	915463	GAVSSSESPSASSSMTR
LmxM.27.2250	LmxM.27	-	939144	939884	939144	940331	939858	939905	AGILLGNMGLVDAYDR
LmxM.28.0330	LmxM.28	+	98141	100903	97925	100903	98015	98056	DIGCGIVDCPLCVK
LmxM.28.0610	LmxM.28	-	211032	223727	211032	223820	223737	223820	MLQSFLRPQSGATTVATGDDEEPTAPVR
LmxM.28.0740	LmxM.28	-	262470	263654	262470	263771	263664	263714	STDSVPADEDALKEFLR
LmxM.28.0740	LmxM.28	-	262470	263654	262470	263771	263715	263747	NTNGALAYSHR
LmxM.28.0890	LmxM.28	+	312911	315310	312902	315310	312902	312943	METMSDSAPMIIKR
LmxM.28.1720	LmxM.28	-	648270	649670	648270	650138	649977	650030	NAAQQSNAAAGSASATAK
LmxM.28.2570	LmxM.28	+	987553	990633	987238	990633	987553	987618	MGSSGVTAGGSTPLFASGSTPR
LmxM.29.0150	LmxM.29	-	45520	46170	45520	46464	46159	46197	GLPTLCDGVMACR
LmxM.29.0150	LmxM.29	-	45520	46170	45520	46464	46378	46416	MTNLESLDLKDNK
LmxM.29.0450	LmxM.29	-	145989	146915	145989	147338	147273	147335	STPLDVRAAEYSPSFAVTVKR
LmxM.29.0450	LmxM.29	-	145989	146915	145989	147338	147156	147209	TGINTTYPMPPPPPMPEK
LmxM.29.0580	LmxM.29	-	183785	185224	183785	185362	185201	185266	SAPGAATAASSGPYMPAYSEVK
LmxM.29.0760	LmxM.29	+	235530	238706	235293	238706	235353	235412	GVDTPEYTIICAQAPDAARR
LmxM.29.0760	LmxM.29	+	235530	238706	235293	238706	235434	235487	LQTSTESSAAQNVTLLLR
LmxM.29.0760	LmxM.29	+	235530	238706	235293	238706	235293	235352	MEDIFDQMVCGSPMQQLNLK
LmxM.29.0760	LmxM.29	+	235530	238706	235293	238706	235353	235409	GVDTPEYTIICAQAPDAAR
LmxM.29.0890	LmxM.29	+	281550	282380	281289	282380	281427	281456	TLFETILTDR
LmxM.29.1110	LmxM.29	+	366812	368671	366089	368671	366350	366469	YPMPPGHSGGGAPLPPLGGGPPAPPQQPAPSGAPPSNAPR
LmxM.29.1380	LmxM.29	+	470462	471433	470231	471433	470375	470476	SFGCPALEIKHPVPTAVVADLTGSAGAGPMMALR
LmxM.29.2440	LmxM.29	-	896050	898374	896050	898734	898573	898629	GGDGAPAAAADLFSAALQR
LmxM.29.3130	LmxM.29	-	1115722	1118625	1115722	1118646	1118566	1118640	SLTPEMAATYDPAAVEADWYPWWEK
LmxM.29.3170	LmxM.29	-	1138056	1139291	1138056	1139396	1139286	1139318	QKEEENFAAMR
LmxM.29.3170	LmxM.29	-	1138056	1139291	1138056	1139396	1139319	1139348	YKAVQAELAR
LmxM.31.0050	LmxM.31	-	17980	18981	17980	19023	18991	19017	AQTFPQQPR
LmxM.31.0470	LmxM.31	-	172072	174726	172072	175260	175018	175044	FDAAIVCSR
LmxM.31.0470	LmxM.31	-	172072	174726	172072	175260	175057	175080	ELTASFQR
LmxM.31.1360	LmxM.31	+	520045	522099	518770	522099	519793	519849	GASPNSLSPSEAGNAAPVR
LmxM.31.1360	LmxM.31	+	520045	522099	518770	522099	519346	519432	RPPLTSPSGSPLSLQQQQQQQQLQNTGAR
LmxM.31.1360	LmxM.31	+	520045	522099	518770	522099	519904	519984	SETVSQQHQPQQQQEQSVAGSTSLGPR
LmxM.31.1360	LmxM.31	+	520045	522099	518770	522099	519433	519489	YTAPQPPPLSQSLPPAIPR
LmxM.31.1360	LmxM.31	+	520045	522099	518770	522099	518968	519051	TSKPVATAPAPSAENYETPNTIAAAGAK
LmxM.31.1360	LmxM.31	+	520045	522099	518770	522099	518905	518967	SLAGAGPASNDDVPLLSGFKR
LmxM.31.2020	LmxM.31	-	782071	782553	782071	782736	782554	782595	EVCFYEEAPGVGVK
LmxM.31.3050	LmxM.31	+	1190172	1192649	1189560	1192649	1190112	1190168	NAGASGDAPQLSDLQESLR
LmxM.31.3320	LmxM.31	+	1317380	1318294	1316993	1318294	1316996	1317055	AALREVSNLASNLGADIKQR
LmxM.31.3450	LmxM.31	+	1379064	1380743	1378848	1380743	1378866	1378913	STEPTGLSSNAVSVPR
LmxM.32.0620	LmxM.32	+	210730	211638	210385	211638	210412	210468	KQPVEDATGPTNALGVDDR
LmxM.32.0620	LmxM.32	+	210730	211638	210385	211638	210499	210528	MLNYETSFCR
LmxM.32.1530	LmxM.32	+	610528	611073	610357	611073	610435	610554	ADVSTDTATAHSTPPAYPGSDEPLPSYVSIVMSDIPIAAK
LmxM.32.1890	LmxM.32	+	720920	722590	720680	722590	720758	720811	AVAVSPSSTTTATAAAAK
LmxM.32.2040	LmxM.32	+	789077	791440	788708	791440	788906	788998	TYAEWYLPSASGSNAFPSPGSVSNAATAAVR
LmxM.32.2830	LmxM.32	+	1152140	1152850	1151678	1152850	1152005	1152043	SADLGTASSGPVR
LmxM.32.2830	LmxM.32	+	1152140	1152850	1151678	1152850	1151921	1151983	SHGDATAEPNNSSNMYAESAR
LmxM.32.3090	LmxM.32	+	1356480	1360700	1356366	1360700	1356432	1356464	DLTDASLAAER
LmxM.33.0850	LmxM.33	+	359676	360566	359295	360566	359508	359552	NSSVSSTSSSSSATK
LmxM.33.1030	LmxM.33	+	441126	441890	440724	441890	440775	440828	TSVVNTSSTTSVVASPAR
LmxM.33.1060	LmxM.33	+	452558	454429	452063	454429	452447	452506	EKPTPPAQGENVNDTFAKDR
LmxM.33.1060	LmxM.33	+	452558	454429	452063	454429	452207	452245	TPSEGAQAAPAAR
LmxM.33.1060	LmxM.33	+	452558	454429	452063	454429	452519	452542	GFEAFYAK
LmxM.33.1520	LmxM.33	-	664127	664525	664127	664588	664529	664588	MDNFQTTFEAFASFGSAPSK
LmxM.33.1520	LmxM.33	-	664127	664525	664127	664588	664502	664588	MDNFQTTFEAFASFGSAPSKEMDNSHFSK
LmxM.33.4120	LmxM.33	+	1557289	1558113	1557175	1558113	1557223	1557282	TSPTIVESPAAISYALASFR
LmxM.34.0320	LmxM.34	+	76691	78271	76640	78271	76643	76714	STAAMIHDSLQAASTPMSEAALKR
LmxM.34.1080	LmxM.34	+	413454	414980	412446	414980	412629	412685	FGAESGDGGGDAASVASSR
LmxM.34.1360	LmxM.34	+	506475	507824	506280	507824	506280	506321	MESGHADAGASGHR
LmxM.34.2270	LmxM.34	+	843343	846795	841903	846795	841906	841965	SGHNPNRAEEPGSGGGAPAR
LmxM.34.2270	LmxM.34	+	843343	846795	841903	846795	842911	843003	SQLAGMPAHLSQSSGAPLPSAAAPFNAPSPK
LmxM.34.2270	LmxM.34	+	843343	846795	841903	846795	842023	842070	GQASPPVPSGEGGPIR
LmxM.34.3890	LmxM.34	-	1439666	1440085	1439666	1440103	1440056	1440094	STLMKEAFELLQR
LmxM.36.0950	LmxM.20	+	352463	365074	352262	365074	352385	352423	EVTTAAPATLHSR
LmxM.36.3070	LmxM.20	-	1218947	1219651	1218947	1219837	1219622	1219669	VIVEPHMLHPGVFISK
LmxM.36.3950	LmxM.20	+	1485347	1486099	1484663	1486099	1484906	1484953	NASGSVASSGGAASAK
LmxM.36.6480	LmxM.20	-	2463720	2467562	2463720	2467646	2467542	2467562	MVQVVHR
LmxM.36.6730	LmxM.20	-	2544544	2547051	2544544	2548626	2547559	2547603	EPLPYDAVLLEGVVR
LmxM.36.6730	LmxM.20	-	2544544	2547051	2544544	2548626	2547235	2547312	KVDLAALPPAELPEVEDTEAAQEAER
LmxM.36.6730	LmxM.20	-	2544544	2547051	2544544	2548626	2547328	2547381	LPDHLSSEYDMPAKPEPK
LmxM.36.6730	LmxM.20	-	2544544	2547051	2544544	2548626	2547559	2547612	NAREPLPYDAVLLEGVVR
LmxM.36.6730	LmxM.20	-	2544544	2547051	2544544	2548626	2548156	2548224	YAAQLLTDRNAAVQATLNDMMER
LmxM.36.6730	LmxM.20	-	2544544	2547051	2544544	2548626	2547526	2547558	SVDAYKTLAQR
LmxM.36.6730	LmxM.20	-	2544544	2547051	2544544	2548626	2548198	2548224	YAAQLLTDR
LmxM.36.6730	LmxM.20	-	2544544	2547051	2544544	2548626	2548156	2548197	NAAVQATLNDMMER
LmxM.36.6730	LmxM.20	-	2544544	2547051	2544544	2548626	2548021	2548062	RLSSVDDGAVSAVR
LmxM.36.6730	LmxM.20	-	2544544	2547051	2544544	2548626	2548582	2548623	SVEDDYRASIAQAR
LmxM.36.6730	LmxM.20	-	2544544	2547051	2544544	2548626	2547700	2547774	ALQAASPDAMEKEEEGADEWAELGR
LmxM.36.6730	LmxM.20	-	2544544	2547051	2544544	2548626	2547493	2547525	LSNQLLHPHLK

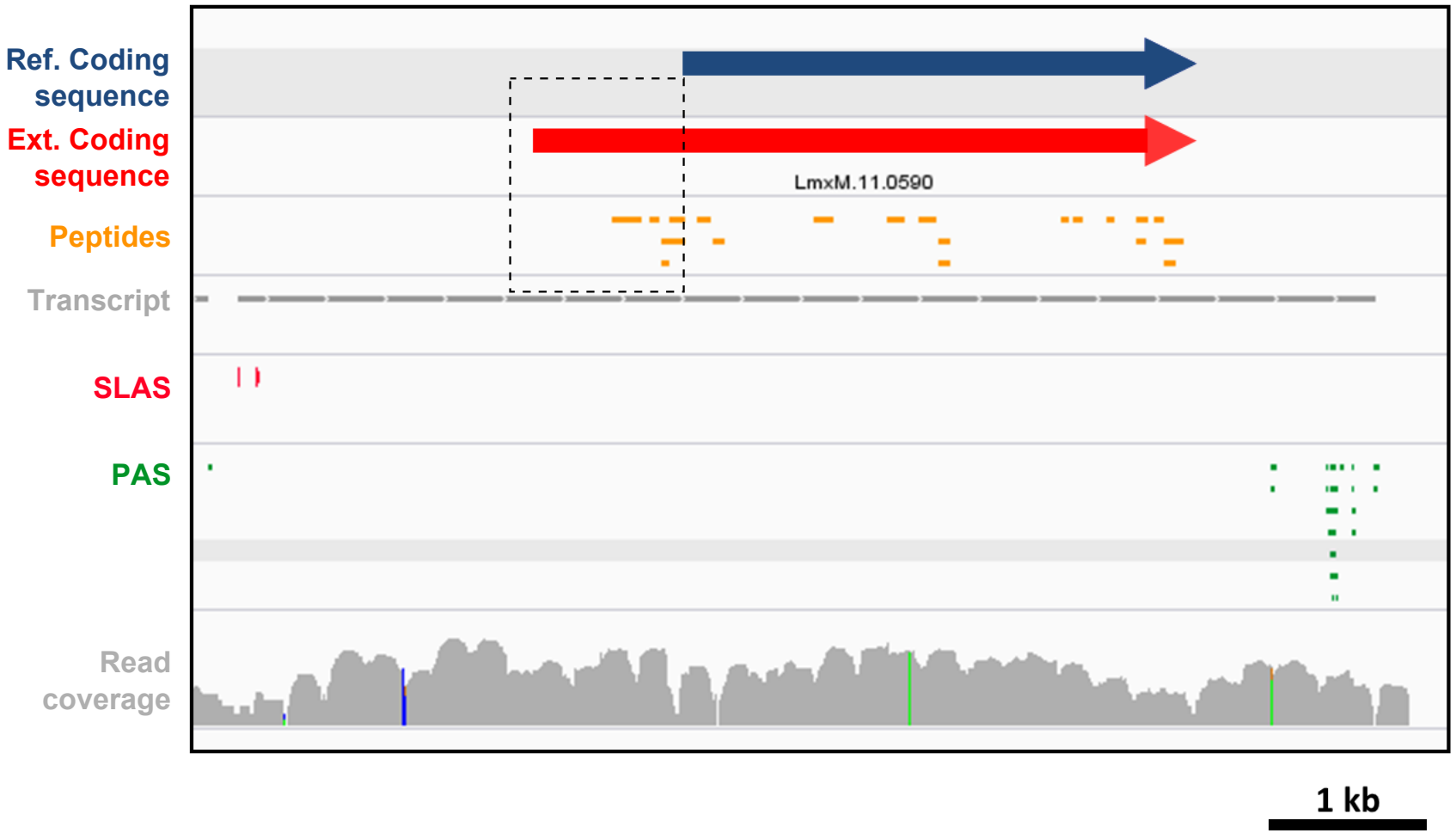


Figure 2.23 Example of peptides mapping to proposed extensions of a coding sequence The locus of LmxM.11.0590, showing the reference and proposed extended coding sequence along with the transcript dimension, SLAS and PAS positions as well as read coverage. Unique peptides detected for this protein were converted into genomic coordinates and are displayed in orange. Peptides mapping to the proposed extension are highlighted in the dashed box.

2.3.8 Novel transcript sequences are absent from annotated proteomes, but are highly conserved amongst *Leishmania* *Leishmania spp.* and to a lesser degree amongst other kinetoplastids

The availability of many kinetoplastid genomes (Table 1.3) opens up the possibility to explore the conservation of the novel transcripts amongst other kinetoplastids. This may not only provide insight into their evolutionary history, but may also act as further evidence for these being *bona fide* genes: Protein coding DNA sequences are under more specific evolutionary constraints with respect to their sequence (Cargill et al. 1999) compared to a stretch of non-coding DNA, with the exception maybe of nucleoprotein-binding sites.

A common way of identifying orthologues of genes is the Reciprocal Best Blast (R.B.B.) analysis (Hirsh and Fraser 2001; Jordan et al. 2002). In a first instance, the R.B.B. method was adapted to search for orthologues of the novel transcripts amongst annotated genes of other kinetoplastids. To this aim the transcript nucleotide sequences were queried against the amino-acid sequences of the proteomes of 9 kinetoplastid species, including *L. mexicana* itself (Table 2.11). 18 R.B.Bs were found in the annotated *L. mexicana* proteome which indicate gene-duplications (6 cases) and artefacts of annotated genes replaced by a gene in our new annotation that is more likely base of SLAS and PAS pattern (6 cases) or where it is difficult to discern which gene-model should be given preference (4 cases). Moreover there are two cases where my gene prediction appears less likely than the reference gene model. In other *Leishmania* species between 20 and 31 R.B.B. are identified with the exception of *L. tarentolae* where 93 are found. In *T. brucei* and *T. cruzi* 60 and 53 R.B.B. are found in the annotated proteome, respectively. *C. fasciculata* has a remarkable 182 R.B.B. in its annotated proteome. The high number of R.B.B. in *L. tarentolae* and *C. fasciculata*, both of which are not mammalian infective kinetoplastids, may be surprising but is very likely to be due to the quality of their only very recently released genome annotation (2011 for *L. tarentolae* (Raymond et al. 2011), 2013-14 for *C.*

Table 2.11 Presence of novel genes amongst annotated kinetoplastid genomes Tables showing the number of Reciprocal Best Blastx (R.B.B) between the 936 novel genes identified in *L. mexicana* and the annotated proteomes of different kinetoplastid species.

Species	Number of R.B.B
Leishmania mexicana	18
Leishmania major	20
Leishmania infantum	25
Leishmania donovani	25
Leishmania tarentolae	93
Leishmania braziliensis	31
Crithidia fasciculata	182
Trypanosoma brucei	60
T. cruzi Brenner Esmeraldo-like	53

fasciculata, source: TriTrypDB(Aslett et al. 2010)). For example, the annotation for *L. tarentolae* not only used employed homology searches against the *L. infantum, L. major, L. braziliensis, T. cruzi* and *T. brucei*, albeit using a minimum size of *L. tarentolae* ORFs of 100 AA, but also employed more advanced gene prediction software (Augustus (Stanke et al. 2004; Stanke, Tzvetkova, and Morgenstern 2006)) trained on existing protein data sets from trypanosomatids, leading identification of proteins missed in other trypanosomatid genome annotations (Raymond et al. 2011).

Moving beyond existing genome annotation, the R.B.B. approach was adapted to utilise a tBlastx method (see Materials and Methods) to determine if orthologues of the novel transcripts can be found in other kinetoplastid genomes irrespective of existing annotation. To test the method, it was applied to a set of 10 control genes and their downstream intergenic sequence (IGS), expected to show less conservation than the respective CDS (Figure 2.24). The 10 control genes were chosen to include proteins conserved, or expected to be conserved, amongst kinetoplastids i.e. the paraflagellar rod protein 2 (PFR2, LmxM.16.1430) (Bastin, Matthews, and Gull 1996), glycosomal glyceraldeylde 3-phosphate dehydrogenase (GAPDH, LmxM.29.2980), γ-tubulin (LmxM.25.0960), SAS-6 (LmxM.34.4280) (Hodges et al. 2010). Furthermore, small proteins (expected to be have orthologues amongst other kinetoplastids based on annotation an TriTrypDB (Aslett et al. 2010)) were chosen to reflect the small predicted size of proteins encoded by novel transcripts such as RNA-polymerase subunit 12 (9kDa, RPB12, LmxM.20.0490) and the small nuclear ribonucleoprotein (11.7 kDa, SmD2, LmxM.32.3190). In addition, ascorbate peroxidase (APX, LmxM.33.0070) was chosen, which is known to be absent from African trypanosomes but present in other kinetoplastids (Castro and Tomás 2008). The small myristoylated protein-2 (SMP-2, LmxM.20.1300) was also added to the list of control genes due to the presence of paralogues in the genome of *L. mexicana* (Tull et al. 2004), as were two genes with homologues in the genome of *L. mexicana* (Histone H3 (LmxM.10.0970) and a δ-amastin

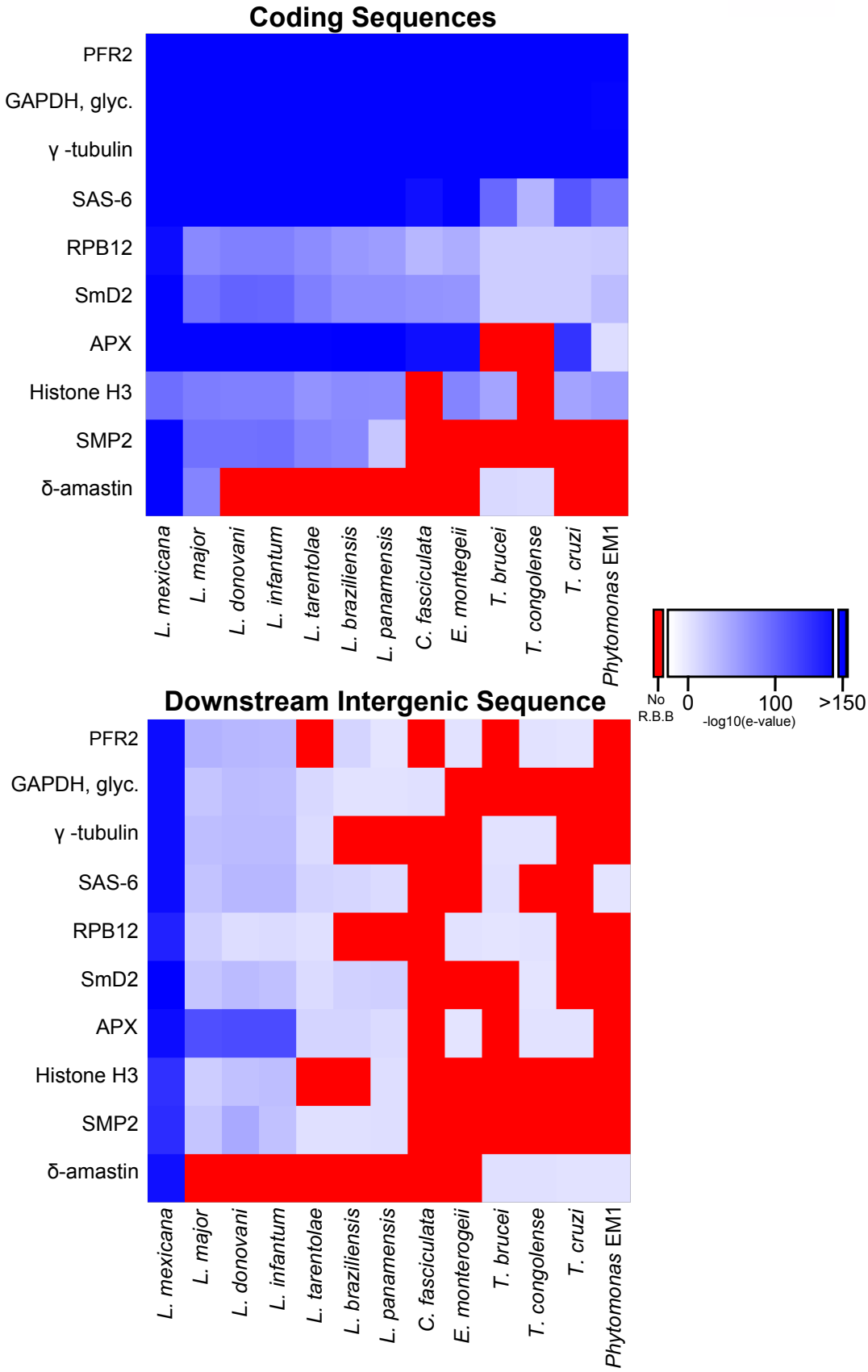


Figure 2.24 Reciprocal Best tBlastx analysis of control coding sequences and corresponding downstream intergenic sequences Reciprocal Best tBlastx analysis of the CDS and downstream IGS of single- and multi-copy genes with known conservation amongst kinetoplastids. Amongst *Leishmania spp*. convervation of downstream IGS is high enough to allow detetion by R.B.B., albeit with lower scores than observed for coding sequences. The R.B.B. score, where available, is the -$\log_{10}$(e-value) of the returning tBlastx. For versions of genomes utlised in this analyses see Materials and Methods. PFR2= LmxM.16.1430, GADPH, glyc. (glycosomal)= LmxM.29.2980, γ-tubulin= LmxM.25.0960, SAS-6= LxmM.34.4280, RPB12= LmxM.20.0490, SmD2 (small nuclear ribonucleoprotein)= LmxM.32.3190, APX (ascorbate peroxidase) = LmxM.33.0070, Histone H3= LmxM.10.0970, δ-amastin= LmxM.08.0760.

(LmxM.08.0760)) to investigate the power of the R.B.B. method at discriminating paralogues and homologues. In each case the entire corresponding 3' IGS was used in a second tBlastx search for comparison with the result of the CDS.

Genes like PFR2 or glycosomal GAPDH are very highly conserved at the level of the coding sequence by R.B.B. but their downstream IGS generally only finds R.B.B. amongst *Leishmania spp.* and even then with considerably poorer e-value scores than the coding sequences. For example, for the R.B.B. analysis of glycosomal GAPDH against *L. major, L. donovani, L. infantum* and *L. tarentolae,* we obtain infinite $-\log_{10}$(e-values) for the CDS (i.e. e=0, identification with exceptional confidence) , whilst for the IGS $-\log_{10}$(e-values) 26.2, 32.5 31.0 and 11.7 for *L. major, L. donovani, L. infantum* and *L. tarentolae,* respectively. The range of non-zero e-values for the R.B.B. values reaches down to 10^{-170}, i.e. $-\log_{10}(e)= 170$ (e.g. R.B.B. to APX in *L. dononvani*). This shows that the employed R.B.B. method is sensitive to the difference between the similarity of coding sequences as opposed to non-coding sequences, albeit in closely related species even IGS may be very similar. This is particularly apparent in the case of the ascorbate peroxidase (APX). Known to be absent in African trypanosomes (Castro and Tomás 2008), which is reflected in the R.B.B. results, the downstream IGS amongst mammalian infective *Leishmania Leishmania spp.* is particularly high, which may indicate presence of a conserved regulatory element and will be an interesting observation to follow up on.

What is furthermore apparent is that the R.B.B. method struggles with multi-gene family genes such as δ-amastins. These are present in other *Leishmania* (Jackson 2010), however identification of the reciprocal-best blast is upstaged by the abundance of homologous or paralogous proteins in the queried *Leishmania* genomes, a common failure of R.B.B. approaches (Wall, Fraser, and Hirsh 2003).

The detection of R.B.B. for IGS in *T. brucei* and *T. cruzi* was surprising and indicated that the chosen R.B.B. method may be prone to false positives arising from the high likelihood

of finding a conserved amino-acid sequence given the relatively close evolutionary relationship between these organisms. Closer inspection of the reverse tBlastx scores showed that these were fairly poor, e-values above $e=10^{-5}$. I therefore decided to apply a quality threshold to the identification of R.B.B. To determine a threshold, I calculated the mean reverse tBlastx e-value for R.B.B. of IGS in *L. tarentolae*. This species was chosen as it is the evolutionarily closest species to *L. mexicana* showing distinctly lower reverse tBlastx e-values compared to the mammalian infective *Leishamania Leishmania spp.* (*L. mexicana*, *L. major, L. infantum, L. donovani*) (Figure 2.24) and might therefore represent a species at the threshold of evolutionary divergence where non-coding sequences are still conserved, i.e. one would expect some motifs to be conserved but sufficient evolutionary drift to have occurred to make others unidentifiable. The mean tBlastx e-value for R.B.B. of IGS in *L. tarentolae* was calculated to be $10^{-9.84}$, which was rounded to 10^{-10}. Applying this threshold to all R.B.B. (i.e. an R.B.B. with an e-value higher than that was excluded from the list of R.B.B.) the heatmaps were redrawn (Figure 2.25). Amongst the IGS, only R.B.B. between *Leishmania spp.* are found, and expectedly few amongst the lizard-infective *L. tarentolae*. The only changes observed as a consequence of setting an e-value threshold in this analysis for R.B.B. of CDS are the loss of the R.B.B. for APX in *Phytomonas* EM1, which has independently been found to be absent from *Phytomonas* (personal communication with Dr. Steven Kelly), and absence of R.B.B. for the δ-amastin from *T. brucei* and *T. congolense*. The latter gene is a member of a group that is thought to have expanded after the divergence of *Leishmania spp.* and *Trypanosoma spp.* and the proto- δ-amastin lost in *T. brucei* (Jackson 2010), so identification of an R.B.B. would have been surprising. The chosen threshold therefore appears to have been able to remove false positives from the identification of R.B.B.

After addressing the power and potential drawbacks of this method, I used it to investigate conservation of the novel genes amongst other kinetoplastids. The nucleotide sequences of the 936 novel transcripts were searched against a set of 12 kinetoplastid genomes using

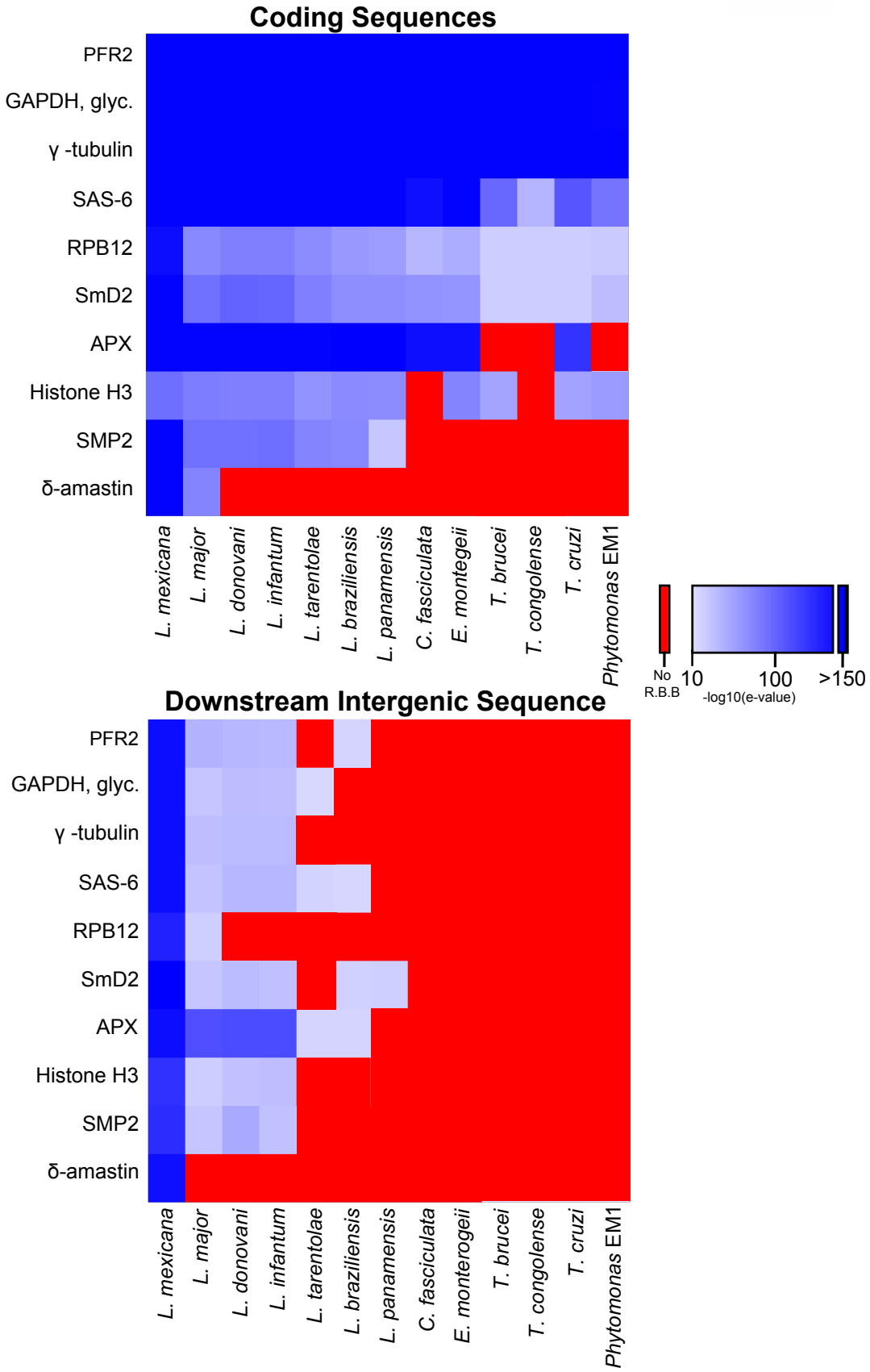


Figure 2.25 Reciprocal Best tBlastx analysis of control coding sequences and corresponding downstream intergenic sequences with e-value threshold Reciprocal Best tBlastx analysis of the CDS and downstream IGS of single- and multi-copy genes with known conservation amongst kinetoplastids. The R.B.B. score, where available, is the -$\log_{10}$(e-value) of the returning tBlastx. Only R.B.B. with return tBlastx e-values below 10^{-10} are shown as R.B.B.. For versions of genomes utlised in this analyses see Materials and Methods. PFR2= LmxM.16.1430, GADPH, glyc. (glycosomal)= LmxM.29.2980, γ -tubulin= LmxM.25.0960, SAS-6= LxmM.34.4280, RPB12= LmxM.20.0490, SmD2 (small nuclear ribonucleoprotein)= LmxM.32.3190, APX (ascorbate peroxidase) = LmxM.33.0070, Histone H3= LmxM.10.0970, Small myristoylated protein-2, (SMP-2, LmxM.20.1300), δ-amastin= LmxM.08.0760.

the above mentioned reciprocal best tBlastx method. When a reciprocal best blast (RBB) was identified, the negative decadic logarithm of the reverse blast e-value is reported (– $\log_{10}$(e-value)). The results are visualised as a heat-map in Figure 2.26 and summarised in brief in Table 2.12. Both R.B.B. with e-value cut-off at $e=10^{-10}$ and without cut-off are shown Table 2.12.

The novel transcripts identified in this study are highly conserved amongst *Leishmania Leishmania spp.* (831 – 823 R.B.B. with e-value cut-off) and *Sauroleishmania* (508 R.B.B. with e-value cut-off) and less amongst *Leishmania Viannia* 313 – 298 R.B.Bs.). *Crithidia fasciculata* and *Endotrypanum monterogeii* have 196 and 185 R.B.B. with e-value cut-off respectively. Amongst *Trypanosoma spp.* between 38 and 34 R.B.B. are found and the plant-infective *Phyotmonas* EM1 has 45 R.B.B (Table 2.12).

The dominant pattern of conservation is amongst *Leishmania Leishmania spp.* and *Sauroleishmania*, a trend furthermore substantiated by analyses of the number of R.B.B. with e-value cut-off per transcript shown in Figure 2.27. In total, 231 novel transcripts have R.B.B. (in all cases with e-value cut-off) at least amongst all *Leishmania spp.* analysed, 231 have R.B.B. amongst at least all *Leishmania spp.* but not *L. tarentolae* (447 including *L. tarentolae*) and 761 have R.B.B. in at least all mammalian infective *Leishmania Leishmania spp.* 12 genes are conserved between all species analysed.

These analyses have shown that orthologues of the novel genes discovered in this chapter are conserved between kinetoplastid species, particularly amongst *Leishmania spp.* which may have implications for their evolutionary history and the biology they may be contributing to.

2.3.9 Reciprocal Best Blast analyses allow prediction of the coding sequences within transcripts

Without proteomic evidence or identification of functional domains, it is difficult to predict which ORF constitutes the CDS within a transcript. However, I showed in Figure 2.24, CDSs

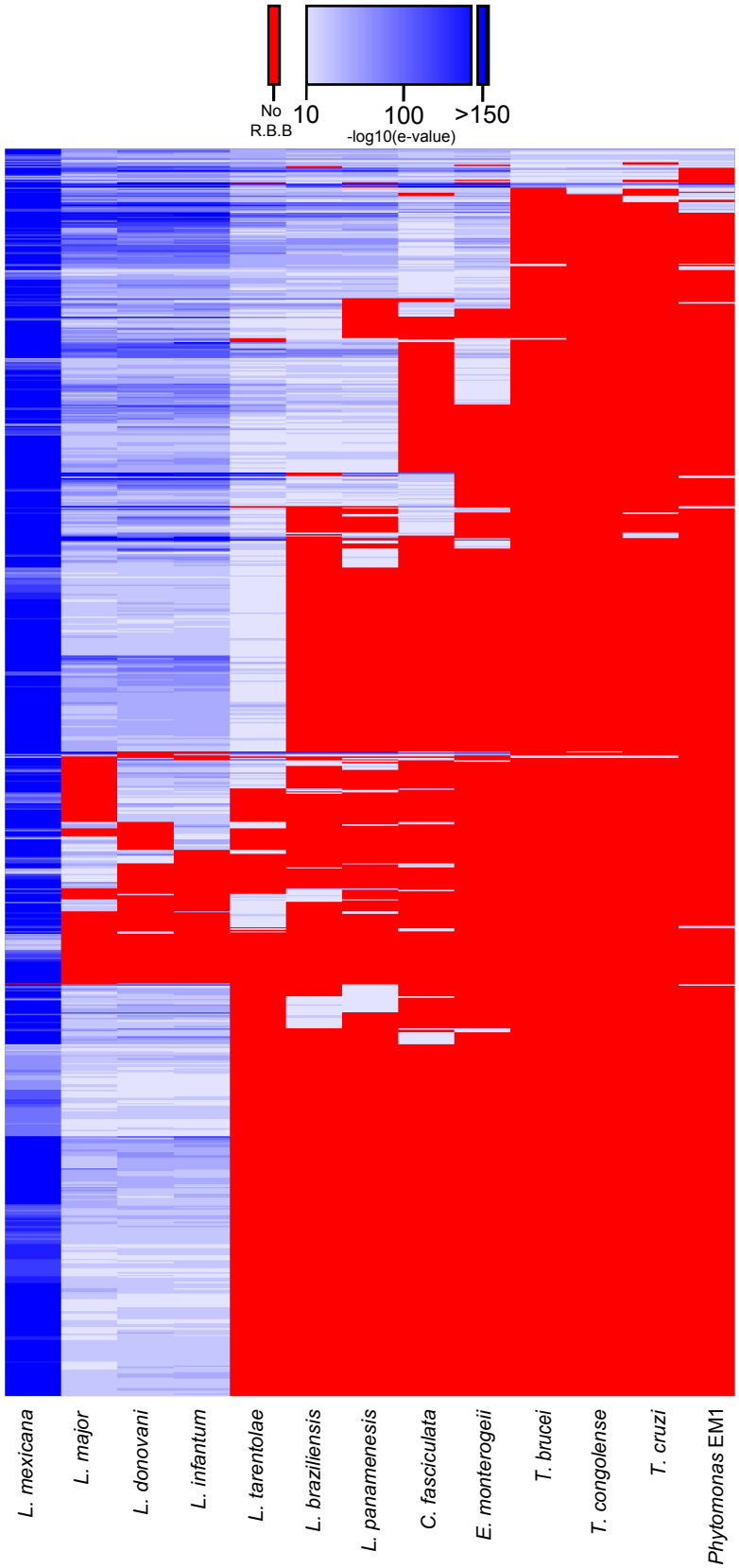


Figure 2.26 Conservation of novel genes amongst kinetoplastids Heatmap showing the conservation of novel transcripts idenified in *L. mexicana* amongst other kinetoplastids as determined by Reciprocal Best tBlastx (R.B.B). Blue indicates presence of an R.B.B.with the shading being the -$\log_{10}$(e-value) of the reverse Blast. Red indicates that no R.B.B. was found or that the reverse tBlastx e-value of the R.B.B. was above 10^{-10}.

Table 2.12 Conservation of novel genes amongst kinetoplastids Tables showing the number of Reciprocal Best Blastx (R.B.B), without and with e-value cut-off at e< $10^{-10,}$ between the 936 novel genes identified in *L. mexicana* and the genomes of different kinetoplastid species.

Species	Number of R.B.B.	Number of R.B.B. with e-value < 10^{-10}
Leishmania major	839	817
Leishmania infantum	854	831
Leishmania donovani	846	823
Leishmania tarentolae	751	508
Leishmania braziliensis	530	313
Leishmania panamensis	545	298
Crithidia fasciculata	374	196
Endotrypanum monterogeii	350	185
Trypanosoma brucei	359	34
Trypanosoma congolense	437	37
Trypanosoma cruzi Brenner Esmeraldo-like	300	38
Phytomonas EM1	231	45

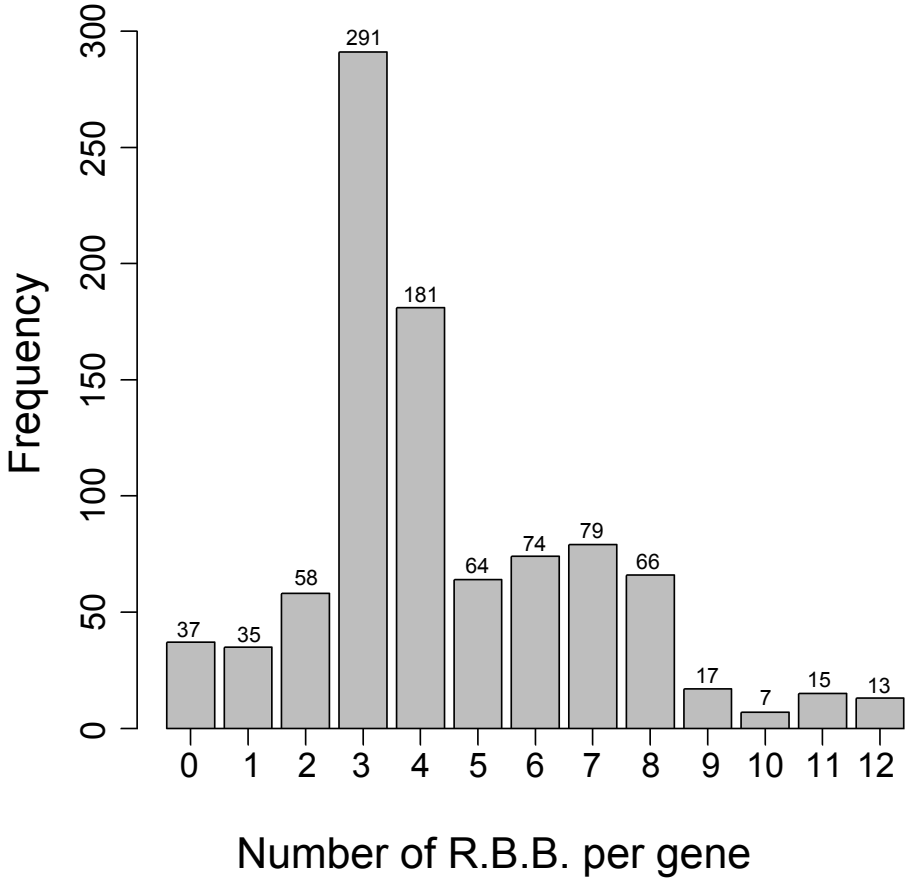


Figure 2.27 Distribution of novel transcripts with different numbers of R.B.B. per transcript across 12 kinetoplastid genomes Histogram showing the distribution of transcripts with different numbers of R.B.B. with an reverse tBlastx e-value $<10^{-10}$ per transcript across the genomes of *L. major*, *L. infantum*, *L. donovani*, *L. tarentolae*, *L. braziliensis*, *L. panamensis*, *C. fasciculata*, *E. monterogeii*, *T. brucei*, *T. cruzi*, *Phytomonas* EM1. The number of transcripts contained in each category are given above each column.

show a higher degree of conservation between species than non-coding sequences. Using the wealth of reciprocal best blast data I sought to identify the most conserved ORF within each transcript, which should allow me to predict the ORF that is the most likely to constitute the coding sequence (see Materials and Methods).

To this aim, all R.B.Bs from section 2.3.8 for each novel transcript were collected and the stretch of nucleotides found in at least 80% of reverse blasts (i.e. the stretch of DNA in *L. mexicana* genome retrieved by the best blast in another kinetoplastids) recorded as the "Consensus Reverse Blast" (CRB). The CRB was then used as a query in a Blastx against a data base consisting of all ORFs >25 AA from three-frame translations of the transcript in question. The highest ranking ORF in the Blastx was earmarked as being the most likely CDS based on conservation.

To test if this Best-CRB-Blast method of predicting CDS was valid, I asked if the correct CDS had been identified for those transcripts where proteomic evidence exists for a CDS.

The Best-CRB-Blast method correctly identified the CDS in 40/42 cases based on supporting peptide evidence (Table 2.13). Moreover, in only 5 out of 42 the CDS identified is not the first ORF within the transcript (encoding a protein larger than 25 AA), i.e. in 11.9 % of cases a uORF is present, consistent with the genome average of 13.6 % of genes (1122 out of 8233 (all genes excluding novel transcripts) having uORFs. In the light of these findings, it appears that predicting CDS within a transcript by Best-CRB-Blast is an effective tool when no other information is available. In turn, when no R.B.Bs are available, the first ORF within a transcript >25 AA is still likely to be correct in around 86 % of cases. All ORFs predicted to be the CDS within the novel transcripts are shown in (Supplementary Table 2.5).

Table 2.13 Agreement of CDS predicted by conservation and mass-spectrometrically identified CDS Table showing for novel protein detected by mass-spectrometry, which was ORF predicted to be the CDS based on conservation amongest other kinetoplastids and whether or not this prediction is supported by peptide evidence. Furthermore, it is indicated whether or not the identified CDS is the first ORF in the transcript or not. In cases of disagreement between prediction and peptide-evidence, it is indicated that the peptide evidence originates from the first ORF of the transcript, whilst the Blast-based prediction did not match this finding.

Transcript	Blast predicted CDS	Is prediction supported by peptide evidence?	Is detected ORF first in transcript
LmxM.01_240110	LmxM.01_240110_239930_240110_+	Peptide Supported	Yes
LmxM.04_422972	LmxM.04_422972_423229_422971_-	Peptide Supported	Yes
LmxM.05_426918	LmxM.05_426918_427409_426917_-	Peptide Supported	Yes
LmxM.06_68747	LmxM.06_68747_68980_68746_-	Peptide Supported	Yes
LmxM.07_533696	LmxM.07_533696_533387_533696_+	Peptide Supported	Yes
LmxM.08_1048727	LmxM.08_1048727_1048004_1048727_+	Peptide Supported	Yes
LmxM.09_94875	LmxM.09_94875_94940_95057_+	Peptide Supported	No
LmxM.10_478397	LmxM.10_478397_478798_478396_-	Peptide Supported	Yes
LmxM.15_314945	LmxM.15_314945_316063_314944_-	Peptide Supported	Yes
LmxM.16_270128	LmxM.16_270128_270325_270127_-	Peptide Supported	Yes
LmxM.17_574236	LmxM.17_574236_573903_574236_+	Peptide Supported	Yes
LmxM.18_241026	LmxM.18_241026_240768_241026_+	Peptide Supported	Yes
LmxM.19_375604	LmxM.19_375604_375763_376174_+	Not Peptide Supported	Peptides yes, Blast no
LmxM.20_2177943	LmxM.20_2177943_2178083_2177942_-	Peptide Supported	No
LmxM.20_344708	LmxM.20_344708_344619_345015_+	Peptide Supported	No
LmxM.21_369741	LmxM.21_369741_369345_369741_+	Peptide Supported	Yes
LmxM.22_389830	LmxM.22_389830_389994_390183_+	Not Peptide Supported	Peptides yes, Blast no
LmxM.23_177418	LmxM.23_177418_176818_177418_+	Peptide Supported	Yes
LmxM.23_565432	LmxM.23_565432_564406_565432_+	Peptide Supported	Yes
LmxM.23_701025	LmxM.23_701025_700578_701025_+	Peptide Supported	Yes
LmxM.25_253118	LmxM.25_253118_253302_252840_-	Peptide Supported	No
LmxM.26_512878	LmxM.26_512878_512416_512878_+	Peptide Supported	Yes
LmxM.26_835753	LmxM.26_835753_835730_836168_+	Peptide Supported	No
LmxM.26_971146	LmxM.26_971146_970888_971146_+	Peptide Supported	Yes
LmxM.27_626028	LmxM.27_626028_625860_626028_+	Peptide Supported	Yes
LmxM.27_628647	LmxM.27_628647_628461_628647_+	Peptide Supported	Yes
LmxM.28_347210	LmxM.28_347210_346766_347210_+	Peptide Supported	Yes
LmxM.28_601219	LmxM.28_601219_601539_601218_-	Peptide Supported	Yes
LmxM.29_1311888	LmxM.29_1311888_1311645_1311888_+	Peptide Supported	Yes
LmxM.30_873749	LmxM.30_873749_874648_873748_-	Peptide Supported	Yes
LmxM.31_1168386	LmxM.31_1168386_1167831_1168386_+	Peptide Supported	Yes
LmxM.31_1377213	LmxM.31_1377213_1376787_1377213_+	Peptide Supported	Yes
LmxM.31_340171	LmxM.31_340171_339799_340171_+	Peptide Supported	Yes
LmxM.31_492479	LmxM.31_492479_492197_492479_+	Peptide Supported	Yes
LmxM.31_833519	LmxM.31_833519_833893_833518_-	Peptide Supported	Yes
LmxM.31_961883	LmxM.31_961883_962104_961882_-	Peptide Supported	Yes
LmxM.32_787928	LmxM.32_787928_787376_787928_+	Peptide Supported	Yes
LmxM.32_947772	LmxM.32_947772_946884_947772_+	Peptide Supported	Yes
LmxM.33_497824	LmxM.33_497824_498357_497823_-	Peptide Supported	Yes
LmxM.34_103118	LmxM.34_103118_102197_103118_+	Peptide Supported	Yes
LmxM.34_1814992	LmxM.34_1814992_1814680_1814992_+	Peptide Supported	Yes
LmxM.34_473668	LmxM.34_473668_473113_473668_+	Peptide Supported	Yes

2.4 Discussion and Conclusions

Employing a cell culture systems we used *L. mexicana* promastigotes to generate axenic and intracellular amastigotes 24 h into differentiation. I was able to assess the degree of infection of murine bone-marrow derived macrophages by *L. mexicana* cells, showing that the infection protocol used had indeed worked. However, I did observed striking differences in the infection levels between biological replicates. What the precise underlying cause for this was is unclear, but it is plausible that the promastigotes, which were grown into stationary growth phase to mimic the infective metacyclic-promastigote stage (Bates 1994), varied between experiments. Without a marker of the metacyclic stage in *L. mexicana* it is not possible to quantify the homogeneity of the cultures used for infection. As a result, it may be possible that more infective cells were present in the cultures used to generate AMA 2 and AMA 3 than in the culture used to generate AMA 1.

Following RNA-extraction I was able to detect rRNA from both species in the mixed-species samples using an Agilent Bioanalyzer to visualise rRNA peaks. From these I determined relative proportions of RNA present in each sample, which correlated well with the microscopically determined infection levels. This was however using rRNA as a marker for total RNA, without actually measuring total RNA form each speices. Mapping of the sequencing reads from the mixed-species samples to both *L. mexicana* and mouse genome revealed that the proportions of RNA predicted to be present by rRNA-quantification were reflected in the proportion of reads mapping to either genome.

Using RNA-sequencing data we mapped SLAS and PAS in *L. mexicana*. From this I was able to identify 936 novel genes, propose 1437 changes to CDS models and define transcript dimensions and therewith the dimensions of UTRs. We opted for a CDS-centric approach to defining transcript, i.e. SLAS and PAS were assigned to reference (and newly predicted) CDS. It would have been possible to use a different approach, where transcripts are first defined as spaces between SLAS and PAS and then CDS identified within, in what I would

call a transcript-centric approach. The latter approach would have prevented the occurrence of some overlapping transcript boundaries that can be observed in the annotation generated by the approach we chose, but would in turn have struggled with genes where no SLAS and/or PAS were detected as is the case for 629 genes in our annotation (2.3.3). As we were not trying to *de novo* generate a genome annotation, but rather improve an existing one containing a wealth of annotated and verified CDSs, the CDS-centric approach chosen herein is favourable. In a completely uncharacterised genome with similar splicing and polyadenylation mechanisms, a transcript-centric approach would be the method to choose.

Comparison of UTRs to and the sequences surrounding pre-mRNA processing sites show considerable differences to *T. brucei* and suggest that the PPT determines the distance between PAS and downstream SLAS in trypanosomatids.

Mass-spectrometric evidence was found for 47 novel genes as well as for 116 predicted extensions of CDSs. Bearing in mind the chemical diversity of possible peptides that could emanate from these novel protein and extensions and thereby the varying detectability by mass-spectrometry, it is difficult to say how complete this data set is. Repeating the preparation of protein samples for mass-spectrometry using a different enzyme other than trypsin to perform the digest, will result in different peptides being generated that may be more readily detectable. This could complement the existing data set, either expanding it, showing that more is yet to be discovered, or quite inversely, only yield similar results, showing that either no more of the novel peptides or extensions are present in the cells or can be detected using such a fairly crude proteomic analysis.

The 936 novel genes identified currently constitute around 10 % of the total 9169 predicted protein coding genes of *L. mexicana* and appear to encode a hitherto neglected group of small proteins particularly conserved amongst *Leishmania Leishmania spp*., albeit some novel transcripts are conserved between all kinetoplastids analysed. The R.B.B. method chosen to identify conserved genes was not initially able to discriminate between

control CDS and intergenic sequences, raising the possibility that sequences found to be conserved by this method could represent non-coding elements. However, addition of a threshold for the reverse blast score (e-value), permitted, at least for the chosen control CDS and IGS, discrimination of conserved coding and non-coding sequences. As a result, the R.B.B. analysis investigating conservation of the novel genes across a wide range of kinetoplastids genomes should provide a truthful insight into the evolutionary origin of the novel genes discovered in this chapter.

The fairly specific conservation of the novel transcripts amongst *Leishmanias spp.* may indicate that the novel transcripts play a role specific to the life-style of these species and investigation of their transcriptomic profiles may provide further support for this notion.

In this chapter I have presented the first genome-wide description of transcript boundaries in *L. mexicana* which will improve accuracy with which transcript abundances can be quantified in the following chapter.

Chapter 3 – Transcriptomic Characterisation of Promastigotes, Axenic Amastigotes and Intracellular Amastigotes

3.1 Introduction

Since the publication of the first *Leishmania* genome (Ivens et al. 2005), genome wide transcriptomic studies have sought to elucidate the differences between promastigotes and amastigotes, and thereby the adaptations amastigotes undergo to persist within the parasitophorous vacuole. As discussed in Section 1.5, pre-genomic studies had identified and characterised genes preferentially expressed in both promastigotes and amastigotes (e.g. (Moore, Santrich, and LeBowitz 1996; Bellatin et al. 2002)), but the availability of whole-genome sequences combined with microarray technology promised the identification of differentially expressed genes on a much larger scale. The first *Leishmania* microarray chips designed were based on random genomic sequences from *L. major*, from which spots could be sequenced and assigned to coding sequences in the genome (Akopyants et al. 2001; Saxena et al. 2003; Akopyants et al. 2004; A. Saxena et al. 2007). Later on, microarray chips were devised using custom synthesised probes derived from coding sequences in the genome (Holzer, McMaster, and Forney 2006; Rochette et al. 2009).

Generally, the number of genes identified to be differentially expressed between promastigotes and amastigotes was modest, ranging between 2.9-12.5% (Leifso et al. 2007; Rochette et al. 2009). The genes that were found differentially expressed reflected the morphological differences between promastigotes and amastigotes, correlating with pre-genomic studies (e.g. expression of flagellar components such as the paraflagellar rod proteins in promastigotes (Moore, Santrich, and LeBowitz 1996; Holzer, McMaster, and Forney 2006)). Equally, previously proposed metabolic differences (Jeremy C. Mottram

and Coombs 1985), with glycolysis dominating in promastigotes compared to fatty-acid oxidation in amastigotes were reflected in microarray data (Rochette et al. 2009).

Timecourse experiments of *L. donovani* differentiating axenically over 120 h (A. Saxena et al. 2007) showed that gene expression changes can be of a transient nature and a range of expression patterns are observed. By combining proteomic and transcriptomic data over a similar timecourse (Lahav et al. 2011), it was not only shown that, whilst protein synthesis spikes in the first 5-10 h of differentiation, global protein synthesis is reduced after 10-15 h. At the same time, an enriched expression of transmembrane proteins was observed. Moreover, a gradual shift from gene expression regulation at the transcript level, to translational control could be seen at 15-24 h into differentiation.

Most of these studies used axenic amastigotes to study amastigote biology as these are easily cultured (Bates et al. 1992) and readily obtained in large quantities. How these axenic amastigotes might compare to intracellular amastigotes on a genome-wide transcriptomic was unclear. Holzer *et al.* (Holzer, McMaster, and Forney 2006) addressed this in *L. mexicana* whilst Rochette *et al.* (Rochette et al. 2009) addressed this in *L. infantum*. However, for their comparisons of axenic and intracellular amastigotes both purified the parasite cells from the host-cell material either by syringe-passaging of 6-8 wk infected lesions or 4 d infected THP-1 cell-line macrophages ((Holzer, McMaster, and Forney 2006) and (Rochette et al. 2009) respectively). This procedure exposes the parasites to a very different environment as is found in the phagolysosome and one may wonder what artefacts could be introduced by this.

Therefore the question remains what the transcriptomic differences between intracellular amastigotes, actually residing within parasitophorous vacuoles, and promastigotes as well axenic amastigotes are. Moreover, what these differences are at an early timepoint, where the extensive translational control has not fully started dominating control of gene expression. Furthermore, constant improvement of genome annotation, notably the

identification of novel genes in Chapter 2 of this work, begs the question how the novel transcript are expressed and what contribution they could be making to *Leishmania spp.* and particularly amastigote biology.

Recent advances in RNA-sequencing technology (Mortazavi et al. 2008; Nagalakshmi et al. 2008; Wang, Gerstein, and Snyder 2009) now not only allow quantification of transcript abundances over a wider dynamic range, and thereby more accuracy, than possible with microarrays (Zhao et al. 2014), but also permit acquisition of data irrespective of a pre-determined hybridisation platform. Consequently, acquired data may be repeatedly analysed using different gene models without the requirement for re-acquisition of the experimental data. With these advances in mind, we decided to embark on an RNA-sequencing based transcriptomic profiling of *L. mexicana* promastigotes and amastigotes.

3.2 Aims

The aims of this chapter were to generate gene expression profiles of promastigotes (PRO), intracellular amastigotes 24 h post-infection (AMA) and axenic amastigotes 24 h after transfer into differentiation medium (AXA). For the AMA samples, the amastigotes were not extracted from their host cells, but instead mixed-species samples generated directly from both cells to minimise handling-artefacts. Gene expression profiles were generated for each cell type on their own, and pairwise differential expression testing performed. Identified differentially expressed genes analysed for enrichment of annotated functions such as GO-terms or participation in an annotated biological pathway, as well as for enrichment of PFAM domains and transmembrane domains and signal peptides, all in an attempt to discern biological patterns that differ between the cell types analysed. All transcriptomic analysis was performed using the transcripts dimensions generated in Chapter 2 and included expression analysis of the novel genes identified in Chapter 2, with an aim of gaining more insight into possible biological functions of these novel genes.

Finally, the distribution of differentially expressed genes in the genome was analysed to assess whether this may provide further insight into mechanisms affecting gene expression in *Leishmania spp.*

3.3 Results

3.3.1 RNA-sequencing generated description of relative transcript abundances within three Leishmania mexicana cell types

3.3.1.1 Biological replicates show a degree of heterogeneity

Transcript abundances were quantified from the sequencing data obtained from the random hexamer primed cDNA library 1 (see Materials and Methods). First, low-quality reads were filtered from the sequencing data using Trimmomatic software (Bolger, Lohse, and Usadel 2014) using settings given in Materials and Methods. Then reads were aligned to a hybrid-transcriptome (hybrid cDNA library) of *L. mexicana* and *Mus musculus* using RSEM (Li and Dewey 2011) (see Materials and Methods). Reads mapping to multiple sites are assigned according to the statistical model devised by (Li et al. 2010). Table 3.1 shows the results of the quality filtering and mapping steps. In *L. mexicana* between 9112 and 9133 out 9169 protein coding genes are detected. Of all the read-counts from *L. mexicana* 5.62 – 7.41% originate from the novel transcripts identified in the previous chapter (Table 3.2), showing that these transcripts form a substantial part of the total transcriptome.

From the obtained read counts, relative abundances of transcripts were determined in the form of Fragments per Kilobase per million mapped reads (FPKM) using RSEM (Li and Dewey 2011). The full data set is given in the Supplementary Table 3.1. The distributions of FPKM values are shown in Figure 3.1. Mean and median FPKM values for AXA and PRO samples are narrowly spread between a mean 51.1-56.5 FPKM and a median 21.8- 27.7 FPKM (Figure 3.2). For AMA samples mean FPKM values lie between 13.4- 32.9 FPKM and median 7.1-16.3 FPKM. The lower FPKM values obtained for AMA samples are due to the

Table 3.1 Summary of read mapping Table showing the Number of sequencing reads in each sample, number of low-quality reads (see Materials and Methods) and the number of reads mapping either to the *L. mexicana* or mouse genome. Furthermore, the number of *L. mexicana* genes with non-zero read counts upon mapping of sequecing reads to hybrid-transcriptome are shown.

	AMA1	AMA2	AMA3	AXA1	AXA2	AXA3	PRO1	PRO2	PRO3
Total paired-end reads	13,477,640	13,357,246	13,044,810	13,124,114	13,748,422	12,778,984	12,835,963	13,087,624	12,858,646
Low-quality reads	9,434	13,357	11,740	17,061	34,371	17,891	20,538	23,558	29,575
Paired-end reads mapping to *L. mexicana* genome	3,437,982	7,224,190	7,767,651	12,525,089	12,950,160	11,893,323	12,220,790	12,301,132	12,109,395
Reads mapping to *Mus musculus* genome	8,346,622	5,005,449	4,183,661	9,175	4,114	3,828	11,534	5,225	7,697
L. mexicana Genes with non zero read counts	9,112	9,123	9,123	9,133	9,129	9,125	9,129	9,119	9,119

Table 3.2 Read-counts from novel transcripts Summary of number of read-counts originating from novel transcripts upon mapping of sequencing reads to hyrbid-transcriptome.

	AMA1	AMA2	AMA3	AXA1	AXA2	AXA3	PRO1	PRO2	PRO3
Total read counts	3019060	6421148	6927016	11273116	11395635	10249450	11015929	10700961	10687120
Read counts from novel transcript	216969	475942	455849	781995	727423	677664	619213	642194	631037
Percentage of count read counts from novel transcripts	7.19%	7.41%	6.58%	6.94%	6.38%	6.61%	5.62%	6.00%	5.90%

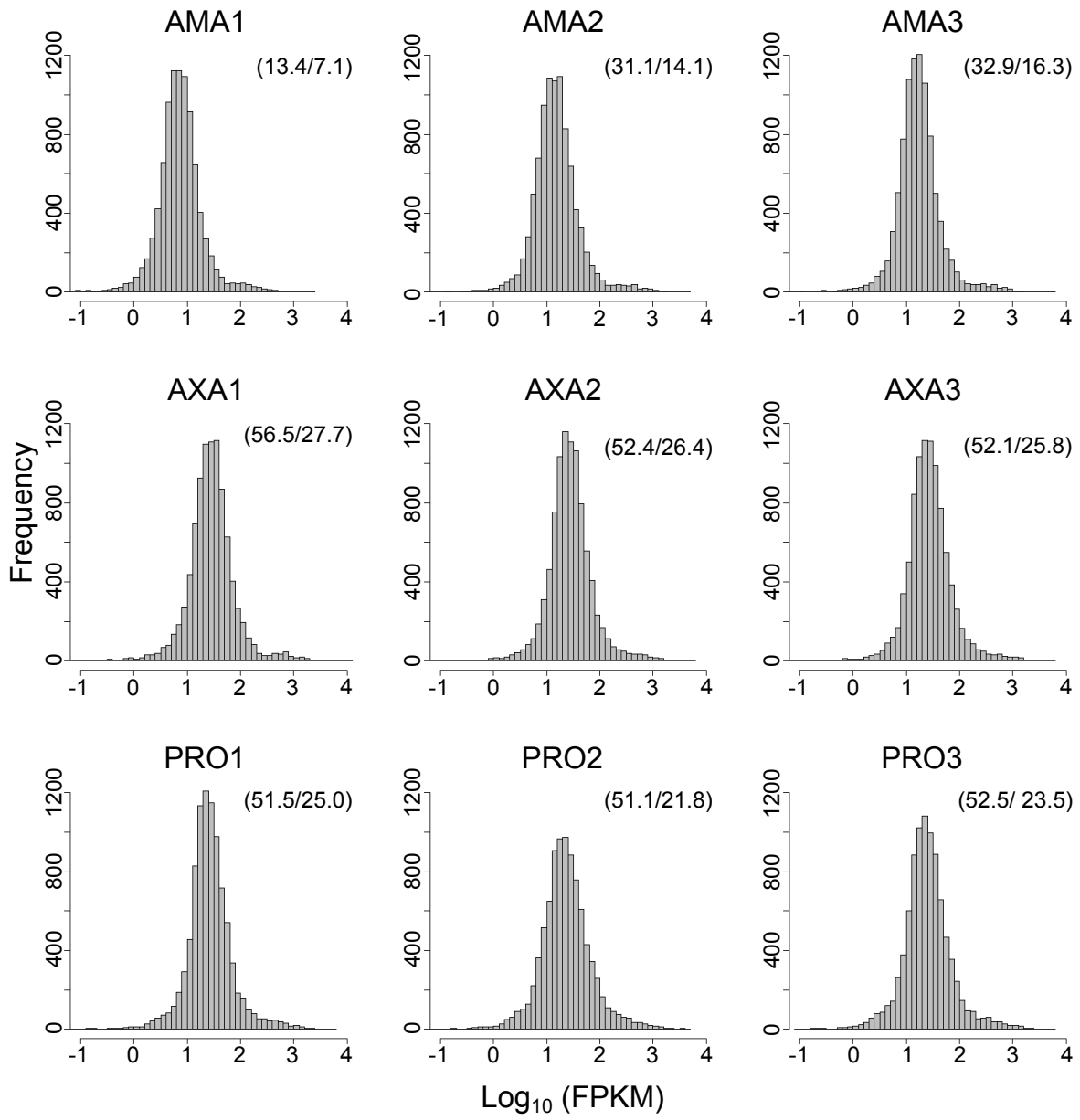


Figure 3.1 Distribution FPKM values accross AMA, AXA & PRO Histograms of FPKM values accross AMA , AXA and PRO. Mean and median FPKM values are given in brackets as (Mean/Median).

proportion of reads mapping to the mouse-part of the hybrid genome used in quantification, thereby reducing the number of reads that will, per million mapped reads, map to any *L. mexicana* gene. In all samples FPKM values are detected over 4-5 orders of magnitude.

Comparison of samples against each other using Pearson's correlation co-efficient, a measure of linear correlation, shows that biological replicates have R^2-values >0.9 relative to each other, with the exception of PRO1 and PRO2 where an R^2= 0.87 is observed (Table 3.3). The first samples of each cell type (i.e. AMA1, AXA1, PRO1) are consistently more different to the other two replicates than latter ones are to each other. I attribute this to the first samples having been extracted and sequenced separately from the next two replicates, which were processed in parallel.

Comparison to an RNA-sequencing study of different life-cycle stages in *T. brucei* (Siegel et al. 2010), where the two biological replicates for procyclic- and bloodstream-forms have R^2-values of 0.97 and 0.99, respectively, indicates that we have more heterogeneity between our replicates, but this was deemed to be acceptable.

3.3.1.2 Most abundant transcripts show overlap between cell types

Sorting genes by their mean FPKM in each cell type and comparing the list of the highest expressed percentile shows that there is considerable overlap between cell types (Figure 3.2). All gene accessions are given in Supplementary Table 3.5. The majority of the shared genes are histones, heat-shock proteins, elongation factors and ribosomal proteins. Indeed, even amongst the genes not shared between the highest expressed percentiles ribosomal proteins and histones are abundant and one may expect these to rank highly in the other cell-types as well. There are however also hypothetical and uncharacterised proteins amongst the highest expressed genes, including an uncharacterised protein (LmxM.16.0500) in AMA that will be investigated further experimentally in Chapter 4. 12 Novel transcripts identified in Chapter 2 are also amongst the highest expressed genes.

Table 3.3 *Leishmania* -sample correlations Pearson correlation coefficients for AMA, AXA and PRO samples based on FPKM of *Leishmania mexicana* genes only.

	AMA1	AMA2	AMA3	AXA1	AXA2	AXA3	PRO1	PRO2	PRO3
AMA1	1.00	0.91	0.94	0.93	0.85	0.83	0.82	0.67	0.75
AMA2	0.91	1.00	0.99	0.92	0.92	0.92	0.86	0.81	0.86
AMA3	0.94	0.99	1.00	0.95	0.93	0.91	0.88	0.78	0.86
AXA1	0.93	0.92	0.95	1.00	0.93	0.91	0.93	0.76	0.86
AXA2	0.85	0.92	0.93	0.93	1.00	0.99	0.94	0.89	0.95
AXA3	0.83	0.92	0.91	0.91	0.99	1.00	0.92	0.91	0.95
PRO1	0.82	0.86	0.88	0.93	0.94	0.92	1.00	0.87	0.95
PRO2	0.67	0.81	0.78	0.76	0.89	0.91	0.87	1.00	0.96
PRO3	0.75	0.86	0.86	0.86	0.95	0.95	0.95	0.96	1.00

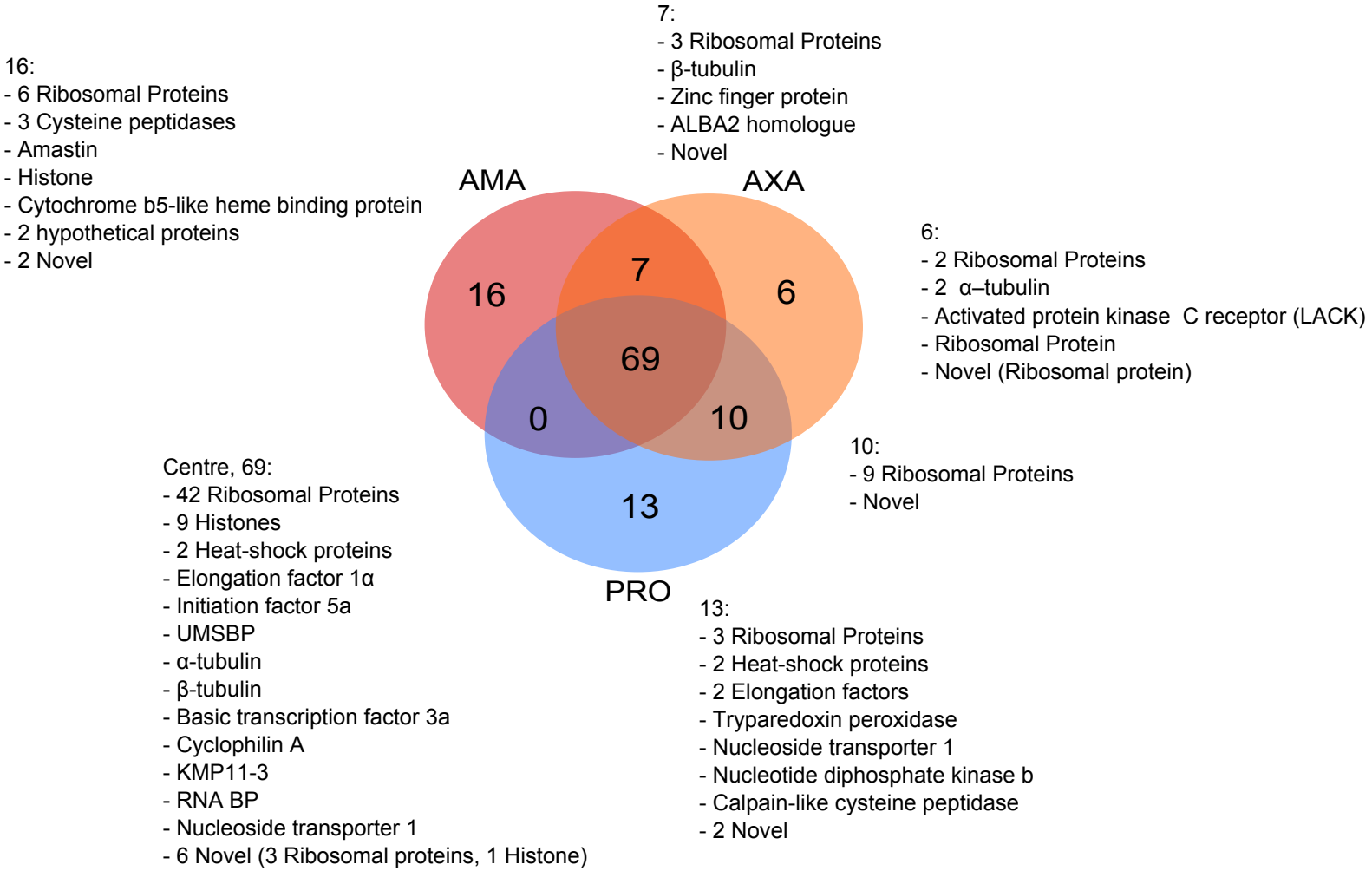


Figure 3.2 Top percentile of transcripts expressed in AMA, AXA and PRO Venn diagram showing the overlap of the top mean-FPKM percentile of transcripts in AMA, AXA and PRO. Where possible, annotated or predicted (by PFAM domain) functions are shown. Novel genes are genes discovered in Chapter 2 of this thesis. For novel genes putative functions based on predicted PFAM domains are shown where possible.

Whilst identification of some of these may not be surprising as 5 of these novel transcripts are predicted to encode ribosomal proteins and histones based on presence of PFAM-domains, there are still 8 highly expressed novel transcripts that cannot be assigned a function by presence of PFAM-domains.

The findings that histones, heat-shock proteins and ribosomal proteins are the highest expressed transcripts agrees well with finding by (Rastrojo et al. 2013) in *L. major*. It does however lie in stark contrast to findings in *T. brucei*, more specifically in bloodstream forms, where a single member of the VSG surface protein-family is the highest expressed transcript (Siegel et al. 2010). Analogous to this we found that summation of the FPKMs of all amastins (see Materials and Methods for identification of amastins) in AMA would make these higher expressed than any other single transcript. However, this is not by orders of magnitude as seen for VSG in *T. brucei* (Siegel et al. 2010), instead the combined FPKMs of amastins amount to 5770 FPKM compared to the highest expressed gene (LxmM.13.0280, α-tubulin) which is expressed at 4004 FPKM.

3.3.1.3 Not all annotated genes are detected on a transcript level

On the other end of the expression-spectrum, not all annotated genes were detected at a transcript level (below the minimal 0.01 FPKM value quantified by RSEM). Figure 3.3 shows a Venn diagram of genes not detected to be expressed in any of the three cell types analysed and the overlap between them (Gene accessions are given in Supplementary Table 3.6). In total 45 genes, of which 19 are on unassigned contigs (LmxM.00) and 12 are novel, are found not to be expressed in one or multiple cell types, with 17 not expressed in any cell type, of which 6 are on unassigned contigs and 6 are novel. This indicates that the majority of genes not found expressed may for one be false positives from the detection of novel transcripts based on SLAS and PAS, and second be genes on unassigned contigs that suffer from other irregularities such as undefined surrounding regions with low read count that may in turn favour assignment of reads to other loci (c.f. RSEM in Materials and Methods).

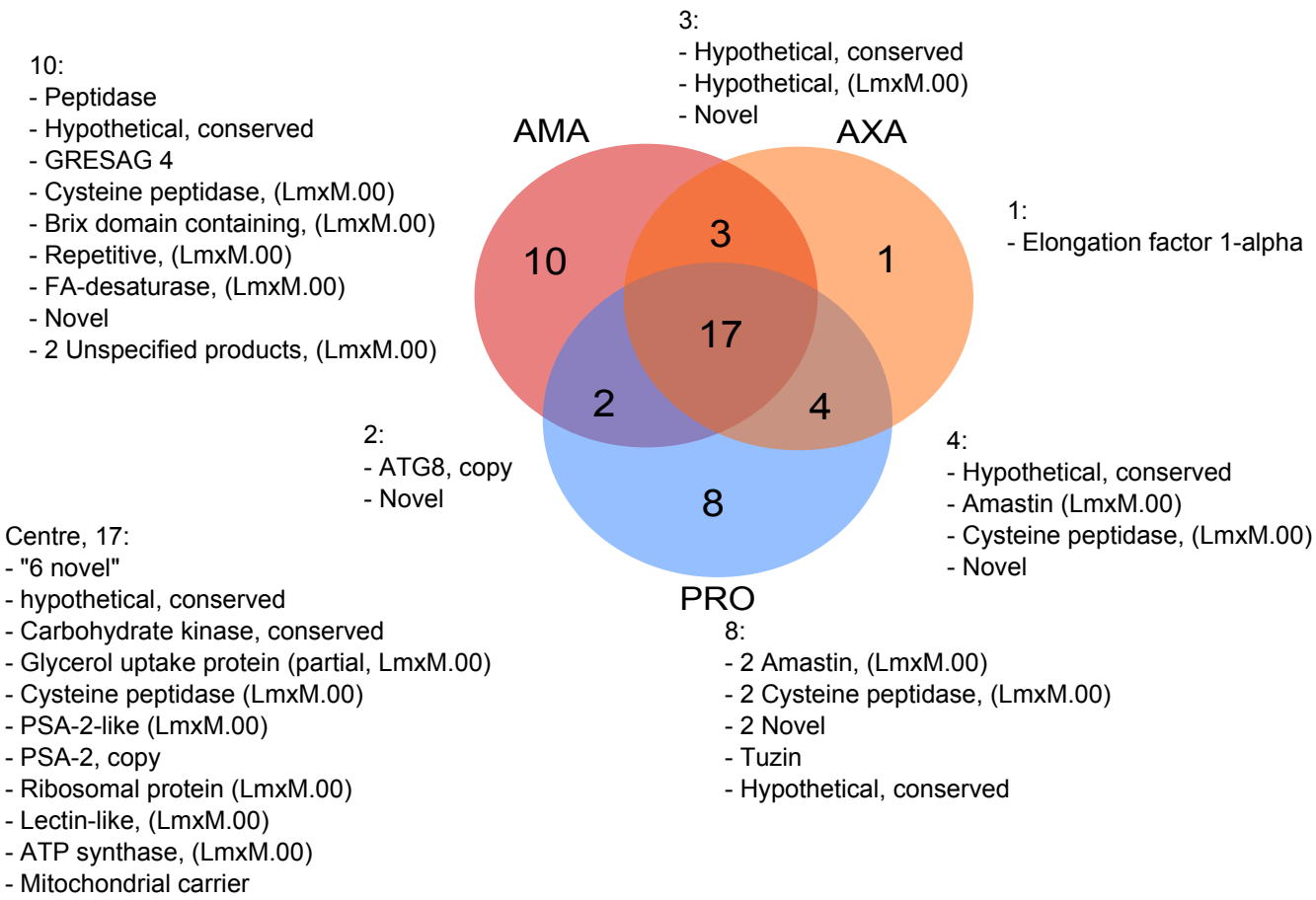


Figure 3.3 Overlap of genes not found to expressed in AMA, AXA and PRO Venn diagram showing the overlap of genes not found to be expressed in AMA, AXA and PRO. For each set of genes, predicted gene functions or classes are shown. LmxM.00 indicates that the gene is located on an unassigned contig. Novel indicates that these are genes discovered in Chapter 2 of this thesis.

A different group of barely detected transcripts are ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs). Prior to sequencing, samples were enriched for poly-adenylated transcripts, and therefore depleted for non-poly-adenylated transcripts like rRNAs and tRNAs. Table 3.4 shows the FPKM values for all annotated tRNAs and rRNAs. In these analyses no tRNAs are detected. No rRNA genes are detected apart from reads from a 18S ribosomal transcript (LmxM.27:rRNA:rfamscan:982402-983034), which was present in all samples with FPKM values ranging between 6039 and 19282. Low FPKM values (between 4 and 12) are also assigned to an adjacent copy of the 18S gene (LmxM.27:rRNA:rfamscan:990836-991467) in AMA1, PRO1 and PRO2. Additionally, reads corresponding to the internal transcribed spacer region (LmxM.00:rRNA:rfamscan:576565-576732 and LmxM.00:rRNA:rfamscan:655615-655782) are detected in AXA2, AXA3 and PRO3 at levels of 9576, 7257 and 8531 FPKM respectively. The absence of detected tRNAs and low number of rRNA genes (4/16) detected shows that selection of poly-adenylated transcripts prior to sequencing was successful in the sense that we observed no tRNA a low diversity of rRNA and transcripts. However, the magnitude of the FPKM values for the detected rRNA is higher than of protein coding genes, where the highest detected FPKM value is for LmxM.13.0280 in AXA1 (FPKM = 10091). The rRNA and tRNA transcripts were absent from the transcript-sequences used in the quantification of read abundances for protein coding transcripts (see Materials and Methods for details on RSEM-based read-quantification). Therefore, the high but variable abundance of sequencing reads from rRNA genes should not skew quantification of FPKM values for protein coding genes by affecting the relative distribution of read-abundances within the total library of mapped reads (Dillies et al. 2013).

Table 3.4 FPKM values of rRNAs and tRNAs Summary of all the FPKM values detected for rRNAs and tRNAs in AMA, AXA and PRO samples

transcript_id	AMA1	AMA2	AMA3	AXA1	AXA2	AXA3	PRO1	PRO2	PRO3
LmxM.00:rRNA:rfamscan:1128121-1128229	0	0	0	0	0	0	0	0	0
LmxM.00:rRNA:rfamscan:576565-576732	0	0	0	0	9576	7257	0	0	8531
LmxM.00:rRNA:rfamscan:655615-655782	0	0	0	0	9576	7257	0	0	8531
LmxM.03:tRNA:rfamscan:249999-250071	0	0	0	0	0	0	0	0	0
LmxM.05:rRNA:rfamscan:361991-362109	0	0	0	0	0	0	0	0	0
LmxM.05:tRNA:rfamscan:361857-361928	0	0	0	0	0	0	0	0	0
LmxM.07:tRNA:rfamscan:362192-362263	0	0	0	0	0	0	0	0	0
LmxM.08:tRNA:rfamscan:371234-371314	0	0	0	0	0	0	0	0	0
LmxM.08:tRNA:rfamscan:371370-371451	0	0	0	0	0	0	0	0	0
LmxM.09:rRNA:rfamscan:251967-252085	0	0	0	0	0	0	0	0	0
LmxM.09:rRNA:rfamscan:380921-381039	0	0	0	0	0	0	0	0	0
LmxM.09:rRNA:rfamscan:391554-391672	0	0	0	0	0	0	0	0	0
LmxM.09:tRNA:rfamscan:252160-252231	0	0	0	0	0	0	0	0	0
LmxM.09:tRNA:rfamscan:252293-252375	0	0	0	0	0	0	0	0	0
LmxM.09:tRNA:rfamscan:380785-380857	0	0	0	0	0	0	0	0	0
LmxM.09:tRNA:rfamscan:381102-381172	0	0	0	0	0	0	0	0	0
LmxM.09:tRNA:rfamscan:381364-381436	0	0	0	0	0	0	0	0	0
LmxM.09:tRNA:rfamscan:391428-391499	0	0	0	0	0	0	0	0	0
LmxM.09:tRNA:rfamscan:391736-391808	0	0	0	0	0	0	0	0	0
LmxM.09:tRNA:rfamscan:391883-391954	0	0	0	0	0	0	0	0	0
LmxM.09:tRNA:rfamscan:392019-392089	0	0	0	0	0	0	0	0	0
LmxM.10:tRNA:rfamscan:480568-480640	0	0	0	0	0	0	0	0	0
LmxM.10:tRNA:rfamscan:480678-480748	0	0	0	0	0	0	0	0	0
LmxM.10:tRNA:rfamscan:480805-480877	0	0	0	0	0	0	0	0	0
LmxM.11:rRNA:rfamscan:157979-158097	0	0	0	0	0	0	0	0	0
LmxM.11:rRNA:rfamscan:163755-163873	0	0	0	0	0	0	0	0	0
LmxM.11:rRNA:rfamscan:361629-361747	0	0	0	0	0	0	0	0	0
LmxM.11:tRNA:rfamscan:158151-158232	0	0	0	0	0	0	0	0	0
LmxM.11:tRNA:rfamscan:158306-158377	0	0	0	0	0	0	0	0	0
LmxM.11:tRNA:rfamscan:163344-163414	0	0	0	0	0	0	0	0	0
LmxM.11:tRNA:rfamscan:163475-163546	0	0	0	0	0	0	0	0	0
LmxM.11:tRNA:rfamscan:163620-163701	0	0	0	0	0	0	0	0	0
LmxM.11:tRNA:rfamscan:361348-361420	0	0	0	0	0	0	0	0	0
LmxM.11:tRNA:rfamscan:361496-361567	0	0	0	0	0	0	0	0	0
LmxM.15:rRNA:rfamscan:288997-289115	0	0	0	0	0	0	0	0	0
LmxM.15:tRNA:rfamscan:288738-288809	0	0	0	0	0	0	0	0	0
LmxM.15:tRNA:rfamscan:288850-288921	0	0	0	0	0	0	0	0	0
LmxM.16:tRNA:rfamscan:439263-439334	0	0	0	0	0	0	0	0	0
LmxM.17:tRNA:rfamscan:330767-330838	0	0	0	0	0	0	0	0	0
LmxM.17:tRNA:rfamscan:330915-330995	0	0	0	0	0	0	0	0	0
LmxM.17:tRNA:rfamscan:331056-331128	0	0	0	0	0	0	0	0	0
LmxM.20:tRNA:rfamscan:1018965-1019035	0	0	0	0	0	0	0	0	0
LmxM.20:tRNA:rfamscan:1019332-1019403	0	0	0	0	0	0	0	0	0
LmxM.20:tRNA:rfamscan:1019467-1019537	0	0	0	0	0	0	0	0	0
LmxM.20:tRNA:rfamscan:1019611-1019682	0	0	0	0	0	0	0	0	0
LmxM.20:tRNA:rfamscan:1587125-1587196	0	0	0	0	0	0	0	0	0
LmxM.20:tRNA:rfamscan:2433914-2433985	0	0	0	0	0	0	0	0	0
LmxM.20:tRNA:rfamscan:2434749-2434830	0	0	0	0	0	0	0	0	0
LmxM.20:tRNA:rfamscan:2434894-2434965	0	0	0	0	0	0	0	0	0
LmxM.20:tRNA:rfamscan:2435038-2435109	0	0	0	0	0	0	0	0	0
LmxM.20:tRNA:rfamscan:484609-484692	0	0	0	0	0	0	0	0	0
LmxM.20:tRNA:rfamscan:484784-484855	0	0	0	0	0	0	0	0	0
LmxM.20:tRNA:rfamscan:485635-485706	0	0	0	0	0	0	0	0	0
LmxM.20:tRNA:rfamscan:485798-485881	0	0	0	0	0	0	0	0	0
LmxM.21:rRNA:rfamscan:156060-156178	0	0	0	0	0	0	0	0	0
LmxM.21:rRNA:rfamscan:430811-430929	0	0	0	0	0	0	0	0	0
LmxM.21:tRNA:rfamscan:156239-156319	0	0	0	0	0	0	0	0	0
LmxM.21:tRNA:rfamscan:431007-431078	0	0	0	0	0	0	0	0	0
LmxM.21:tRNA:rfamscan:431226-431299	0	0	0	0	0	0	0	0	0
LmxM.21:tRNA:rfamscan:431365-431437	0	0	0	0	0	0	0	0	0
LmxM.23:rRNA:rfamscan:217667-217785	0	0	0	0	0	0	0	0	0
LmxM.23:tRNA:rfamscan:217271-217344	0	0	0	0	0	0	0	0	0
LmxM.23:tRNA:rfamscan:217397-217469	0	0	0	0	0	0	0	0	0
LmxM.23:tRNA:rfamscan:217537-217608	0	0	0	0	0	0	0	0	0

LmxM.23:tRNA:rfamscan:217851-217922	0	0	0	0	0	0	0	0	0
LmxM.23:tRNA:rfamscan:218142-218224	0	0	0	0	0	0	0	0	0
LmxM.23:tRNA:rfamscan:218284-218355	0	0	0	0	0	0	0	0	0
LmxM.23:tRNA:rfamscan:218428-218499	0	0	0	0	0	0	0	0	0
LmxM.23:tRNA:rfamscan:218816-218888	0	0	0	0	0	0	0	0	0
LmxM.23:tRNA:rfamscan:218929-219000	0	0	0	0	0	0	0	0	0
LmxM.23:tRNA:rfamscan:219076-219147	0	0	0	0	0	0	0	0	0
LmxM.24:tRNA:rfamscan:206541-206612	0	0	0	0	0	0	0	0	0
LmxM.24:tRNA:rfamscan:626468-626539	0	0	0	0	0	0	0	0	0
LmxM.24:tRNA:rfamscan:626779-626860	0	0	0	0	0	0	0	0	0
LmxM.24:tRNA:rfamscan:626942-627015	0	0	0	0	0	0	0	0	0
LmxM.24:tRNA:rfamscan:628063-628136	0	0	0	0	0	0	0	0	0
LmxM.24:tRNA:rfamscan:684993-685064	0	0	0	0	0	0	0	0	0
LmxM.27:rRNA:rfamscan:982402-983034	13818	6039	6391	17066	15562	15073	15450	12867	19282
LmxM.27:rRNA:rfamscan:990836-991467	12	0	0	0	0	0	4	6	0
LmxM.29:tRNA:rfamscan:736547-736618	0	0	0	0	0	0	0	0	0
LmxM.30:tRNA:rfamscan:1278048-1278120	0	0	0	0	0	0	0	0	0
LmxM.30:tRNA:rfamscan:1278163-1278243	0	0	0	0	0	0	0	0	0
LmxM.30:tRNA:rfamscan:219043-219115	0	0	0	0	0	0	0	0	0
LmxM.30:tRNA:rfamscan:487888-487958	0	0	0	0	0	0	0	0	0
LmxM.30:tRNA:rfamscan:488008-488079	0	0	0	0	0	0	0	0	0
LmxM.32:tRNA:rfamscan:101520-101592	0	0	0	0	0	0	0	0	0
LmxM.32:tRNA:rfamscan:101652-101724	0	0	0	0	0	0	0	0	0
LmxM.32:tRNA:rfamscan:101800-101871	0	0	0	0	0	0	0	0	0
LmxM.32:tRNA:rfamscan:485268-485340	0	0	0	0	0	0	0	0	0
LmxM.33:tRNA:rfamscan:1338044-1338116	0	0	0	0	0	0	0	0	0
LmxM.33:tRNA:rfamscan:1338176-1338259	0	0	0	0	0	0	0	0	0
LmxM.33:tRNA:rfamscan:1338326-1338397	0	0	0	0	0	0	0	0	0
LmxM.33:tRNA:rfamscan:1338482-1338553	0	0	0	0	0	0	0	0	0
LmxM.33:tRNA:rfamscan:1338614-1338686	0	0	0	0	0	0	0	0	0
LmxM.33:tRNA:rfamscan:454797-454870	0	0	0	0	0	0	0	0	0
LmxM.33:tRNA:rfamscan:455103-455176	0	0	0	0	0	0	0	0	0
LmxM.33:tRNA:rfamscan:455237-455309	0	0	0	0	0	0	0	0	0
LmxM.33:tRNA:rfamscan:455343-455422	0	0	0	0	0	0	0	0	0
LmxM.33:tRNA:rfamscan:455497-455577	0	0	0	0	0	0	0	0	0

3.3.1.4 Sequencing depth was insufficient to reliably quantify abundance of transcripts within the lowest percentiles of absolute expression

The total number of read-counts originating from *L. mexicana* is lowest in the AMA samples due to the presence of mouse RNA in the sample. Therefore, the sequencing depth achieved using the same sequencing-library size will be lower, affecting the confidence with which transcript abundances can be quantified compared to AXA and PRO samples. To address the extent of this difference, I sought to determine the FPKM-saturation for each sample, which gauges, at subsets of the total RNA-library, the fraction of genes close to their final FPKM (I used ± 10 %), in dependence of their final absolute expression level (Mortazavi et al. 2008). Existing packages performing this analysis (e.g. RSeQC (L. Wang, Wang, and Li 2012)) do not employ the same read-aligner and transcript-level quantifier combination as we employed (Bowtie2 and RSEM (see Materials and Methods). To make the analysis of FPKM-saturation consistent with the quantification of FPKM in this this study, I wrote a custom script employing RSEM and Bowtie2 to determine FPKM-saturation.

To this aim the sequencing read libraries were randomly sampled at the level of the raw data at different fractions of the total library size. FPKM-quantifications from these sub-libraries as determined by RSEM (running Bowtie2) were compared to the FPKM-values obtained at full library size to obtain a measure of FPKM-saturation (see Materials and Methods, scripts provided in Supplementary Material). The plots of these analyses are shown in Figure 3.4. The higher expressed genes reach FPKM-saturation at smaller sub-library sizes than lower expressed genes. In AXA and PRO FPKM-saturation is reached above the 5th abundance percentile, whilst for AMA2-3 FPKM saturation is only reached above the 10th and for AMA1 only above the 25th abundance percentile. This shows that the mixed species sample has a lower sequencing depth which will affect accurate quantification of transcript abundances amongst at least the 10th lowest abundance percentile of transcripts.

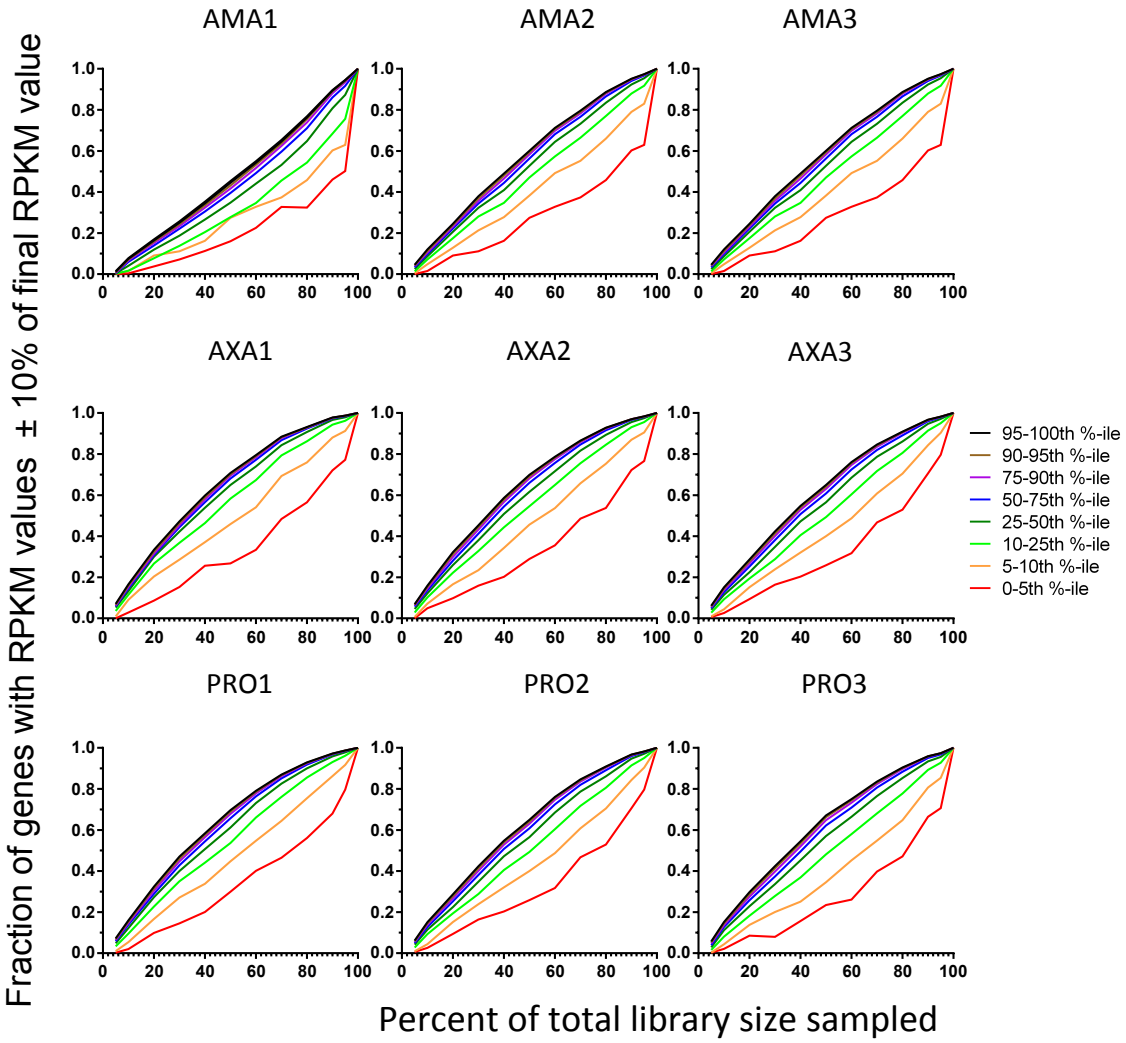


Figure 3.4 FPKM saturation of transcripts Plots showing the fraction of genes with RPKM values within ± 10% of final FPKM value at different sample sizes of final library for AMA, AXA and PRO samples. Samples from each library were obtained by random sampling of raw data and quantification of FPKM valuesusing RSEM. In each plot, genes were analysed in groups based on percentile bins of the global FPKM distributionat maximum library size.

3.3.1.5 There is very little correlation between transcript abundance and mass-spectrometric protein level quantification

With the quantification of transcript abundances it is now possible to correlate transcriptomic data with mass-spectrometric quantifications. The questions this is meant to address are how well transcript abundances correlate with protein abundances and over what range of transcript abundances the mass-spectrometry experiments performed were able to detect peptides from. The former question relates to the widely observed and substantial degree of translational control in trypanosomatids (Lahav et al. 2011; Vasquez et al. 2014). The latter question will give insight into the power of mass-spectrometric experiment at sampling proteins from a spectrum of transcript abundances.

For these correlations, the whole cell mass-spectrometric data for AXA and PRO (Section 2.3.7) were quantified using the label-fee SINQ method (Trudgian et al. 2011). SINQ estimates relative protein abundances by summing the spectral intensities of all fragments of peptides belonging to a given protein, normalising these by the length of the protein and dividing by the sum of all spectral intensities to obtain SINQ-scores. The SINQ-scores obtained for the whole cell lysates were then plotted against mean FPKM values for these genes in AXA and PRO (Figure 3.5 A & B). For SINQ-scores between 10^{-10} and 10^{-6} no correlation between FPKM and SINQ-scores is observed, with FPKM values predominantly ranging between 10 and 100. SINQ scores above 10^{-6} tended to correlate with higher FPKM scores, mainly between 100 and 1000 (red boxes on Figure 3.5).

In Figure 3.6 A proteins detected only in AXA or PRO compared to all other detected proteins are plotted as a function of the mean FPKM in AXA or PRO of the corresponding transcript. We can see that mass-spectrometric detection ranged over 4 orders of FPKM-magnitude. There is considerable overlap of the FPKM-distribution of proteins only detected in AXA or PRO, which are shown separately in Figure 3.6 B & C to avoid misleading interpretations due to superimposition of the AXA-only proteins over the PRO-

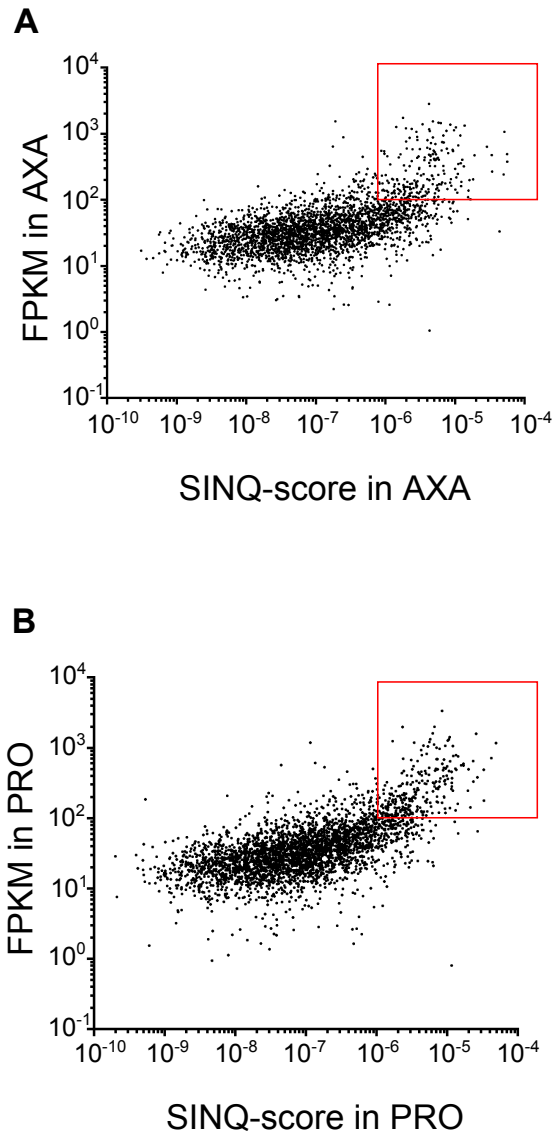


Figure 3.5 Correlation of FPKM and SINQ-scores Plots of FPKM against SINQ-scores in (A) AXA and (B) PRO. Red box indicates region above 10^{-6} SINQ-score and 10^{2} FPKM.

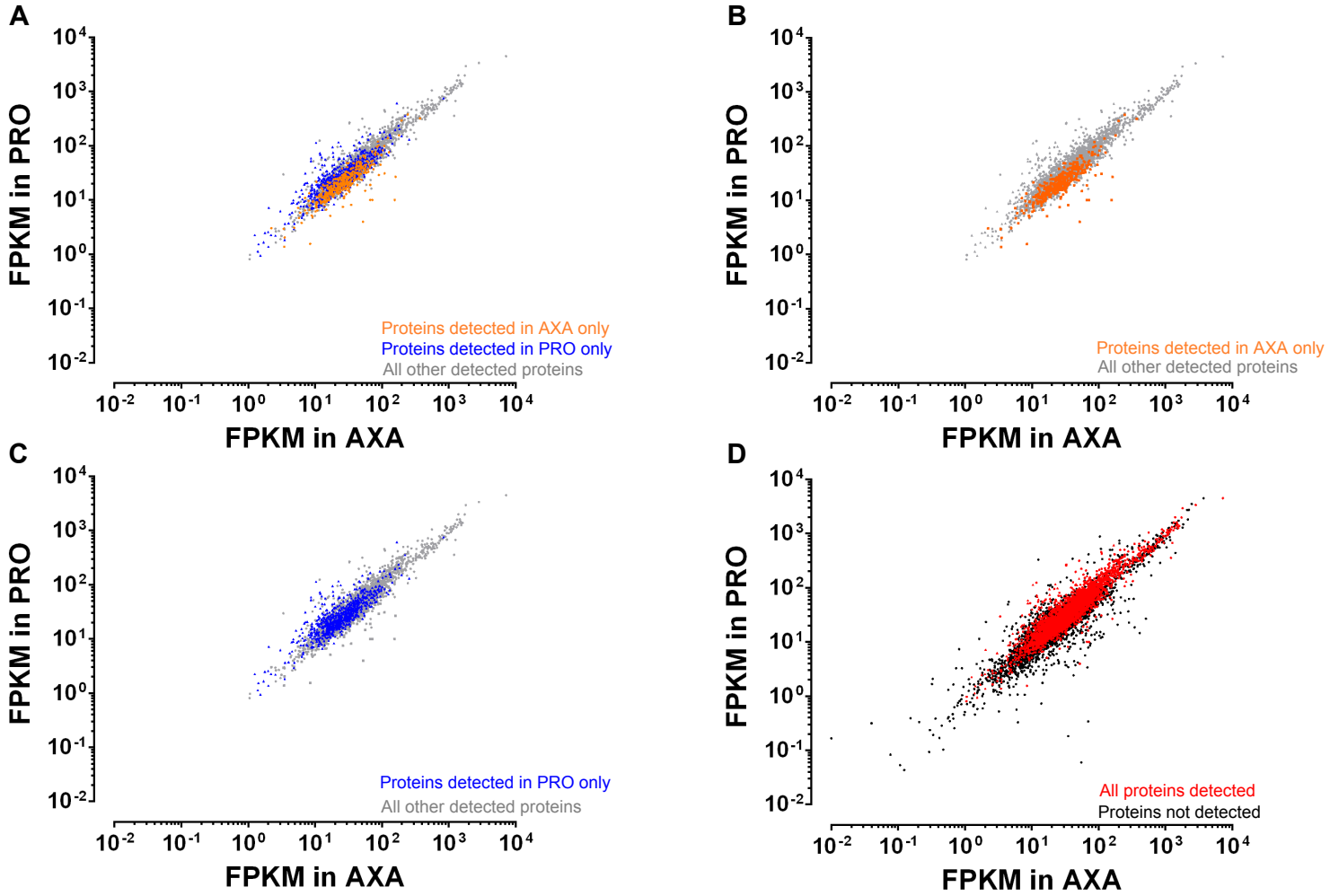


Figure 3.6 Comparison of mass-spetrometric detection of proteins in AXA and PRO with transcriptomic detection of the corresponding transcript in AXA and PRO Plots showing proteins as a function of their mean FPKM values of the corresponding transcript in AXA and PRO (FPKM-distribution) with a colour code indicating mass-spectrometric detection in (A) either AXA or PRO and separately only in (B) AXA and (C) PRO relative to all other detected proteins. D shows the corresponding FPKM-distribution of proteins detected in either or both AXA and PRO compared to proteins not detected by mass-spectrometry.

only proteins. A slight bias of PRO-only proteins towards higher FPKM-values in PRO vs. AXA and conversely of AXA-only proteins towards higher FPKM-values in AXA vs. PRO is seen. In Figure 3.6 D we can see that mass-spectrometric detection of proteins, although possible over 4 orders of FPKM-magnitude of the corresponding transcripts, predominantly happens above 10 FPKM in both AXA and PRO, with many proteins with corresponding FPKM-values below 10 not being detected by mass-spectrometry. Amongst the highest expressed transcripts however, too, not all corresponding proteins are detected.

This means that transcript abundance has only limited predictive power with regards to detection of the corresponding protein by mass-spectrometry and quantification of relative protein abundances by SINQ, consistent with previous reports of different translational efficiencies observed for different mRNAs in trypanosomatids (Vasquez et al. 2014).

3.3.2 Over 40% of genes are differentially expressed between AMA, AXA and PRO based on statistical thresholds

3.3.2.1 Comparison of transcript abundances between life stages identified differentially expressed genes

On their own, transcriptomic profiles have only limited potential of informing about specific biology of a cell-type, especially considering the poor correlation of transcript abundances with protein levels within a cell-type as addressed in Section 3.3.4.5. Differential abundance of transcripts between cell-types is more powerful at informing us about differential biology of two cell types: If we (admittedly naively) assume similar translational efficiency of single transcripts in different cell types, changes in transcript levels will result in changes in protein levels.

Differential expression (DE) analyses were performed using DESeq2 (Anders and Huber 2010). The input for this are read counts per gene (RCPG), and not FPKM. FPKM values are normalised by library size (Fragments per Kilobase per Million fragments mapped), which can lead to skews of the transcriptomic data introduced by outliers e.g. a very highly but differentially expressed gene (Dillies et al. 2013).

Due to the different library sizes between AMA samples compared to AXA and PRO because of the mouse RNA present in AMA samples, normalisation was however necessary. To this aim, RCPG in all samples were normalised by a scaling factor calculated from their median RCPG abundance (Dillies et al. 2013) (see Materials and Methods). The distribution of RCPG before and after median-normalisation are shown in Figures 3.7 and 3.8. Whilst median RCPG values before normalisation still ranged between 193 and 727, after median normalisation these ranged only between 514 and 529, making RCPG values across samples comparable.

Following pairwise DE-testing (see Materials and Methods), all genes were displayed as a function of their fold differential expression between compared cell types and the probability of the observed difference in counts between cell types having occurred by chance (Figure 3.9). We set a multiple-testing corrected p-value (Benjamini et al. 2001) threshold at 0.05, i.e. p-values <0.05 irrespective of their fold change. Other studies comparing different trypanosomatid cell-types have chosen to report differentially expressed gene number only above a fold-change threshold e.g. a 2-fold cut-off (Siegel et al. 2010). We decided to opt against a fold-change threshold. Whilst we may therefore be including transcripts with very small fold-changes, these may nevertheless have a significant biological role, by either reaching a functional threshold-abundance within a cell or be transcript with a low fold change as a function of transient regulation (Saxena et al. 2007).

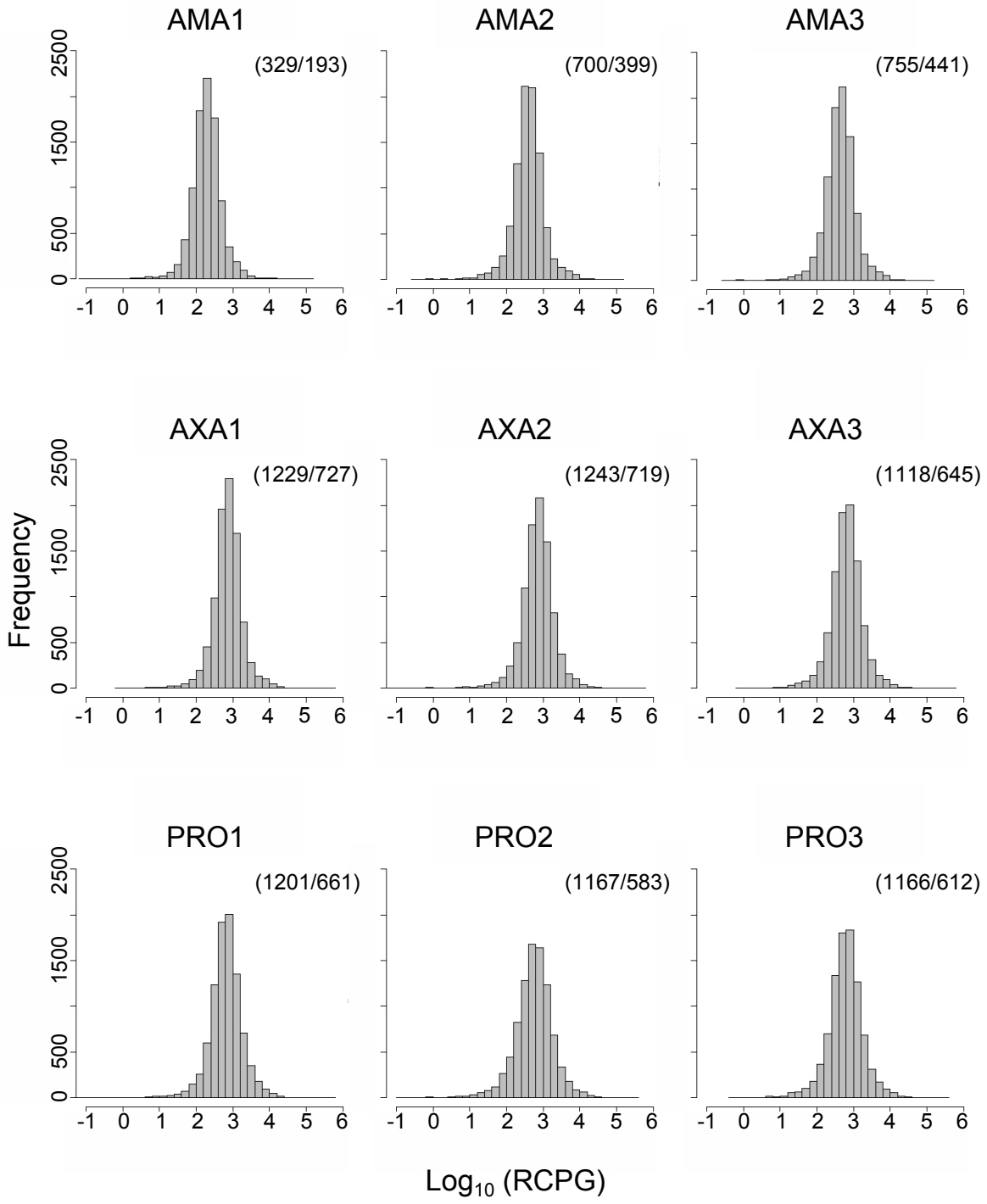


Figure 3.7 Distribution of read-counts per gene Histograms showing distribution of read-counts per gene (RCPG) before median normalisation of AMA, AXA and PRO samples. Mean and median count vlues showin in brackets (mean/median).

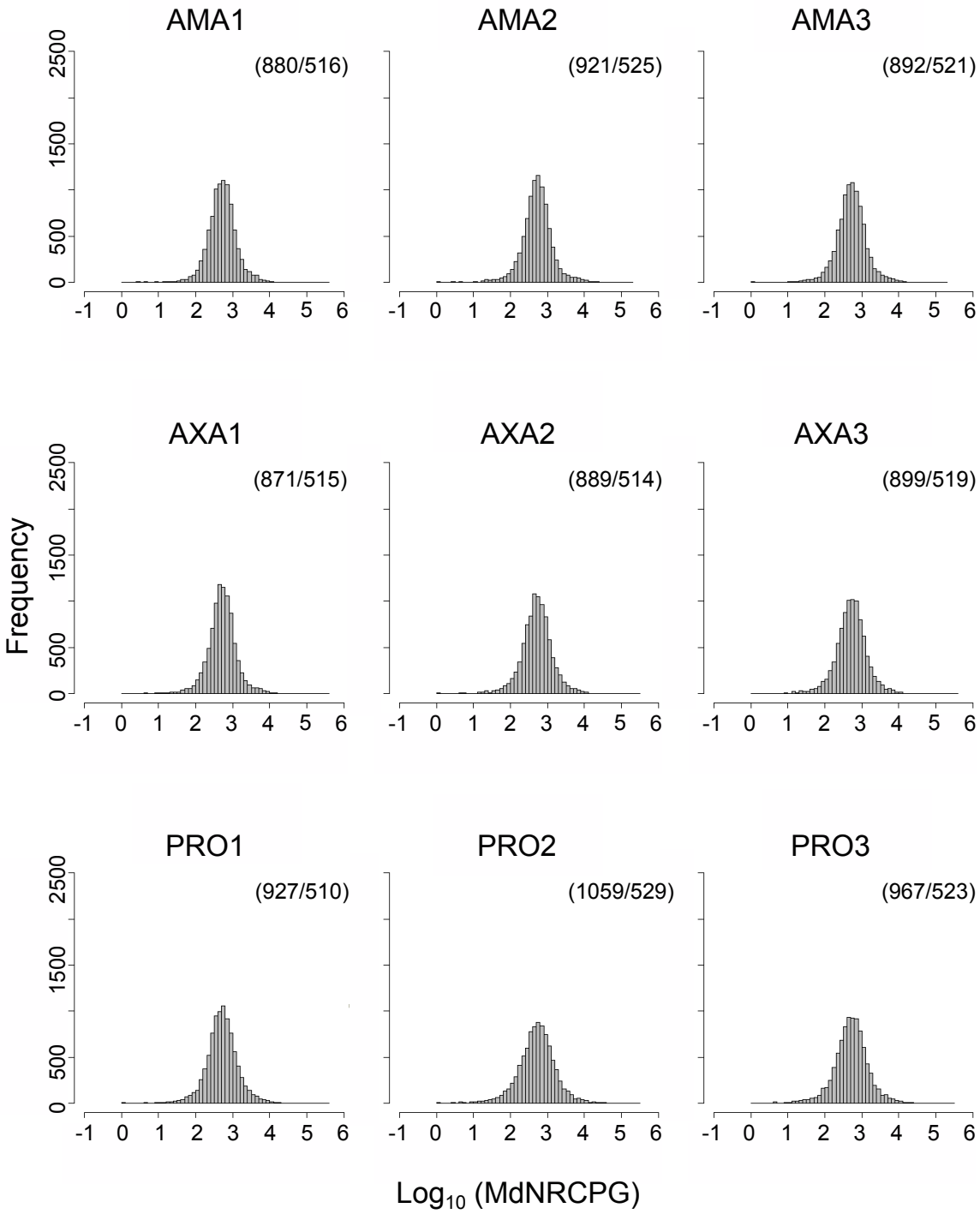


Figure 3.8 Distribution of read-counts per gene: Histograms showing distribution of median normalised read-counts per gene (MdNRCPG) after median normalisation of AMA, AXA and PRO samples. Mean and median count values shown in in brackets (mean/median).

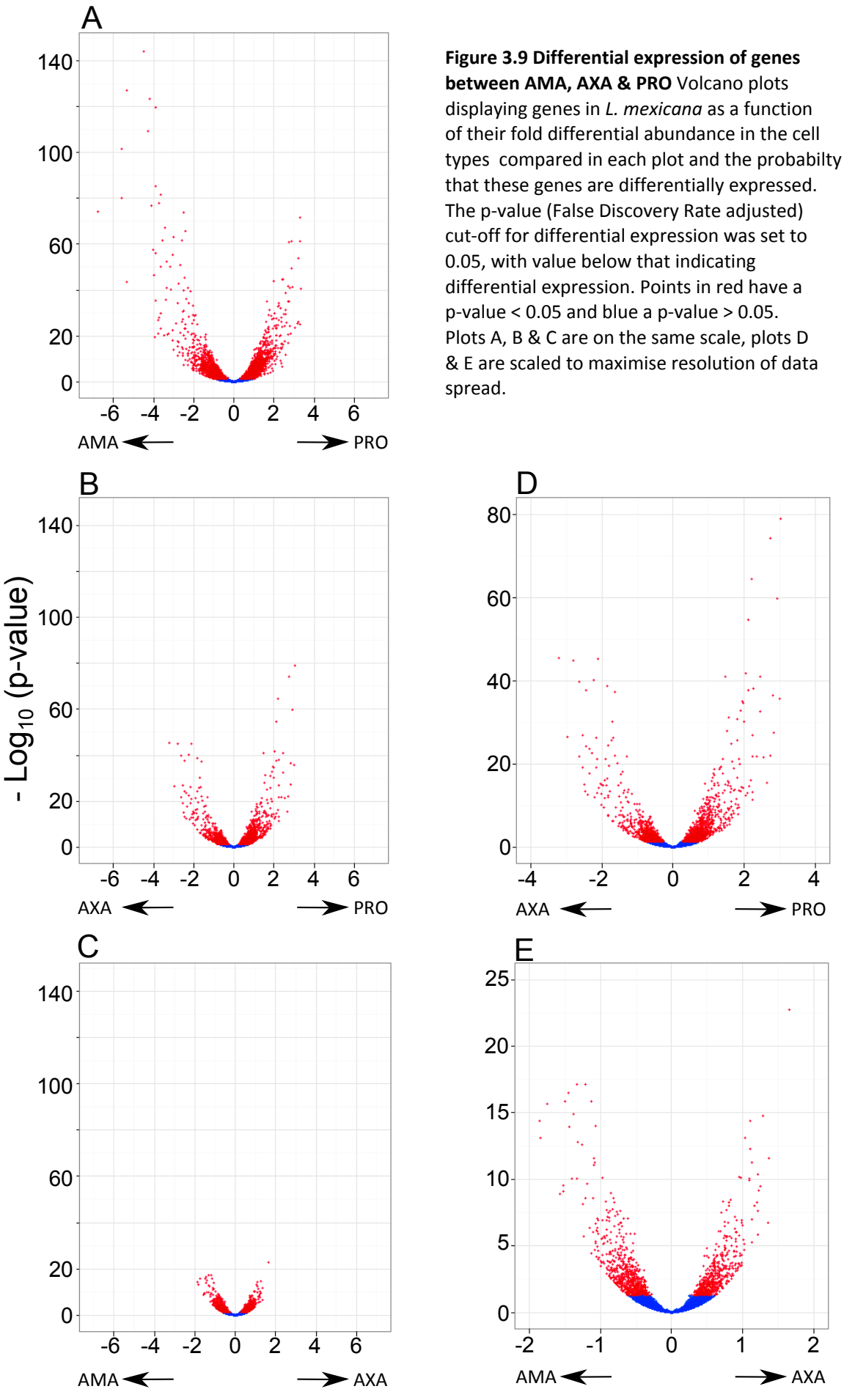


Figure 3.9 Differential expression of genes between AMA, AXA & PRO Volcano plots displaying genes in *L. mexicana* as a function of their fold differential abundance in the cell types compared in each plot and the probabilty that these genes are differentially expressed. The p-value (False Discovery Rate adjusted) cut-off for differential expression was set to 0.05, with value below that indicating differential expression. Points in red have a p-value < 0.05 and blue a p-value > 0.05. Plots A, B & C are on the same scale, plots D & E are scaled to maximise resolution of data spread.

Of the 9169 predicted protein coding transcripts, in the comparison of AMA and PRO 1853 transcripts were preferentially expressed in AMA (625 ≥2-fold) compared to 1979 in PRO (675 ≥=2-fold), indicating that that 41.8% of the genome is differentially expressed at 24 h post-infection. Comparison of AXA against PRO revealed 23.7% of all transcripts being differentially expressed, with 951 preferentially expressed in AXA (138 ≥2-fold) and 1225 in PRO (228 ≥2-fold). Analysis of AXA against AMA shows 563 transcripts to be preferentially expressed in AXA (24 ≥2-fold) and 671 in AMA (46≥2-fold). This means that even between the two amastigote cell-types 13.5 % of genes are differentially expressed. This supports previous findings by Holzer and colleagues (Holzer, McMaster, and Forney 2006; Annie Rochette et al. 2009) showing that there are considerable differences between *bona-fide* and axenically differentiated amastigotes. Moreover, we see that PRO and AMA are more different from each other than PRO and AXA, positioning AXA as an intermediate between PRO and AMA. All DE-analyses are given in Supplementary Tables 3.2-3.4.

3.3.2.2 Differential-expression data is consistent with previous studies

With the wealth of DE data obtained in this study, it is impossible to verify the expression pattern of each gene by other methods. Therefore, results from the differential expression analyses were compared to previous reports of differential expression based on quantitative PCR or Northern Blot. Table 3.5 summarises these comparisons.

A colour code indicates cases in which measurements in this study agree or disagree with previous reports or where the comparison is inconclusive. Inconclusive cases may be due to differential expression being reported in a different *Leishmania* species or relative to one amastigote model (e.g. promastigotes vs. axenic amastigotes) but being only detected for PRO vs. AMA in this study. Further complicating is the fact that many differentially expressed genes are found in multiple copies in the genome (e.g. paraflagellar rod

Table 3.5 Comparison of differential expression analyses with other reports from literature Table comparing differential expression data from the literature with data obtained in this study. A colour code in the "Stage-specificity shown in paper" -column indicates whether findings agree (green), disagree (red) or comparisons are inconclusive (orange).

L. mexicana Gene ID	Gene name	Function	Species (for which RNA comparison data is shown in study)	Comparisons done in study	Method used for comparison	Reference	Stage-specificity shown in paper	PROvAMA log_2 (fold change)	PROvAMA padj	PROvAXA log_2 (fold change)	PROvAXA padj	AXAvAMA log_2 (fold change)	AXAvAMA padj
LmxM.08.0890	Polβ	DNA polymerase beta; putative mitochondrial; DNA replication, DNA repair and base-excision DNA repair	*L. infantum*	PRO, AMA	Northern Blot, qRT	(Ramiro et al. 2002)	AMA	1.00	1.39E-06	0.52	1.49E-02	4.45E-01	9.28E-03
LmxM.08.1030	CPB2.8	cathepsin L-like cysteine proteinase; virulence factor	*L. mexicana*	PRO, lesion AMA (AXA)	Northern Blot	(Souza et al. 1992; Mottram et al. 1997)	lesion AMA (AXA)	-2.49	6.37E-08	-0.63	NA	-4.22E-01	5.22E-02
LmxM.08.1040	CPB2.8							-1.11	NA	NA	NA	-7.01E-01	NA
LmxM.08.1070	CPB2.8							-9.87	0.00E+00	-1.06	NA	-3.78E-01	1.33E-01
LmxM.08.1070partial	CPB2.8							-0.73	NA	-0.02	NA	-4.56E-02	NA
LmxM.08_29.0880	LdARL-3A	ADP-ribosylation factor 3; likely role in flagellum biogenesis	*L. amazonensis*	PRO, lesion AMA	Northern Blot	(Cuvillier et al. 2000)	PRO	0.50	8.26E-02	0.14	6.68E-01	3.46E-01	1.33E-01
LmxM.08_29.1750 and 1760	PFR1	major component of paraflagellar rod	*L. mexicana*	PRO, AXA, AMA	Northern Blot	(Burchmore and Landfear 1998)	PRO	1.96	1.28E-11	2.22	1.61E-22	-2.72E-01	4.30E-01
LmxM.08_29.1760	PFR1							3.29	2.09E-72	3.02	1.09E-79	1.95E-01	3.33E-01
LmxM.08_29.1830	GCVL-1	subunit of the mitochondrial glycine cleavage complex	*L. infantum*	PRO, AXA	Northern Blot	(Müller and Papadopoulou 2010)	constitutive	-0.38	0.09	-0.08	8.18E-01	-2.48E-01	3.53E-01
LmxM.13.0280	alpha tubulin	cytoskeleton	*L. mexicana*	PRO, AXA, AMA	Northern Blot	(Burchmore and Landfear 1998)	constitutive	-0.74	1.97E-03	-0.43	1.93E-01	-1.94E-01	5.72E-01
LmxM.13.0390	alpha tubulin							-0.41	3.36E-01	-0.36	3.00E-01	-2.56E-02	9.60E-01
LmxM.14.1320	SHMT-S	Serine hydroxymethyltransferase; Folate metabolism	*L. infantum*	PRO, AXA, (AMA)	qRT-PCR	(Gagnon et al. 2006)	AXA, (AMA)	1.41	5.21E-10	0.74	4.83E-04	5.81E-01	9.66E-03
LmxM.16.0390	VG7 A5	unknown	*L. mexicana*	PRO, meta, AXA	Notrhern	(Liu et al. 2000)	AXA	-0.07	8.00E-01	0.10	7.16E-01	-1.60E-01	4.67E-01
LmxM.16.1410	PFR2 downsream gene 1	unknown	*L. mexicana*	PRO, AXA	Northern Blot	(Moore, Santrich, and LeBowitz 1996)	PRO	0.35	1.39E-01	0.15	6.62E-01	1.54E-01	6.69E-01
LmxM.16.1420	PFR2 downstream gene 1	unknown	*L. mexicana*	PRO, AXA	Northern Blot	(Moore, Santrich, and LeBowitz 1996)	PRO	1.95	3.18E-15	0.50	1.04E-01	1.25E+00	3.42E-10
LmxM.16.1430	PFR 2C	major component of paraflagellar rod	*L. mexicana*	PRO, AXA	Northern Blot	(Moore, Santrich, and LeBowitz 1996)	PRO	2.73	1.04E-61	2.11	2.23E-55	5.77E-01	1.70E-04
LmxM.16.1440	PFR2 upstream gene 1	unknown	*L. mexicana*	PRO, AXA	Northern Blot	(Moore, Santrich, and LeBowitz 1996)	constitutive	-0.03	9.10E-01	-0.33	1.55E-02	3.00E-01	8.32E-02
LmxM.16.1450	PFR2 upstream gene 2	unknown	*L. mexicana*	PRO, AXA	Northern Blot	(Moore, Santrich, and LeBowitz 1996)	constitutive	-1.07	5.70E-06	-0.60	3.62E-02	-3.46E-01	1.82E-01
LmxM.18.1080	PKAC1	protein kinase A catalytic subunit isoform 1	*L. tropica (L. major)*	PRO, AXA, AMA (PRO, lesion AMA)	RT-PCR; Northern Blot	(Siman-Tov et al. 1996)	PRO	2.23	8.89E-34	2.03	1.45E-42	1.62E-01	6.05E-01
LmxM.18.1510	ATPase 1a	P-type H -ATPase, putative	*L. donovani (L. pifanoi)*	PRO, lesion AMA (PRO, AXA)	Northern Blot	(Meade et al. 1989)	lesion AMA (AXA)	1.87	1.78E-29	0.94	4.86E-05	6.82E-01	4.79E-03
LmxM.18.1520	ATPase 1b	P-type H -ATPase, putative	*L. donovani (L. pifanoi)*	PRO, lesion AMA (PRO, AXA)	Northern Blot	(Meade et al. 1989)	lesion AMA (AXA)	-0.56	5.37E-02	-0.13	7.10E-01	-3.90E-01	7.00E-02
LmxM.19.1440	MAPK4	Map kinase	*L. panamensis*	PRO, AXA	RT-PCR	(Gutiérrez et al. 2001)	AXA	-0.18	6.24E-01	0.07	8.74E-01	-1.76E-01	6.83E-01
LmxM.20_36.6470	LmxMPK1 (lmpk)	Map kinase	*L. mexicana*	PRO, AMA	Northern Blot	(Wiese 1998)	AMA	-0.45	1.16E-02	-0.09	6.33E-01	-3.41E-01	3.97E-02
LmxM.21.0410	PIWI-like protein	unclear	*L. infantum*	PRO, AXA	Northern Blot	(Padmanabhan et al. 2012)	AXA	-0.55	3.86E-02	-0.33	1.56E-01	-1.85E-01	5.58E-01
LmxM.21.1860	A850 (beta-tubulin)	cytoskeleon	*L. mexicana*	PRO, lesion AMA; PRO, AXA	Northern Blot	(Bellatin et al. 2002)	leison AMA; AXA	-2.52	2.11E-74	-1.45	2.90E-16	-8.48E-01	6.65E-06
LmxM.22.0691	A2	possibly stress/heat shock resistance	*L. donovani*	PRO, lesion AMA; AXA	Northern Blot	(Charest and Matlashewski 1994)	lesion AMA; AXA	-3.30	3.15E-21	-1.24	5.18E-04	-7.29E-01	1.33E-02
LmxM.22.0692	A2							-3.01	8.35E-15	-1.39	4.60E-05	-7.50E-01	1.05E-02

LmxM.23.1050	SHERP1	unclear	*L. donovani*	PRO, AXA	Northern Blot	(Saxena et al. 2007)	AXA	0.67	2.04E-01	0.14	7.83E-01	4.77E-01	1.57E-01
LmxM.23.1060	HASPB	unclear	*L. mexicana; L .amazonensis*	PRO, meta, AXA	Northern Blot	(Depledge et al. 2010)	AXA	-2.54	3.77E-28	-1.35	2.20E-05	-4.85E-01	1.41E-01
LmxM.23.1061	SHERP1							-0.59	1.09E-02	0.64	2.99E-02	-9.64E-01	1.58E-04
LmxM.26.0030	GCVP	subunit of the mitochondrial glycine cleavage complex	*L. infantum*	PRO, AXA, AMA	Northern Blot; qRT	(Müller and Papadopoulou 2010)	AXA, AMA	0.84	4.82E-04	0.60	4.28E-03	2.07E-01	5.11E-01
LmxM.28.0980	Ldp27	mitochondrial membrane protein	*L. donovani*	PRO, AXA	Northern Blot	(Dey et al. 2010)	AXA	-3.90	9.59E-57	-2.15	7.77E-20	-1.43E+00	1.07E-14
LmxM.28.2370	SHMT-L	Serine hydroxymethyltransferase; Folate metabolism	*L. infantum*	PRO, AXA, (AMA)	qRT-PCR	(Gagnon et al. 2006)	AXA, (AMA)	-0.81	9.93E-04	0.31	NA	-1.08E+00	8.38E-12
LmxM.28.2740	LACK	ADP-ribosylation factor 3; likely role in flagellum biogenesis	*L. donovani (L. amazonensis)*	PRO, lesion AMA	Northern Blot	(Cuvillier et al. 2000)	constitutive	-0.03	9.47E-01	-0.12	7.71E-01	9.31E-02	8.36E-01
LmxM.29.1500	LmaC1N	3' nucleotidase/nuclease	*L. major*	PRO, lesion AMA	RT-PCR	(Farajnia et al. 2004)	lesion AMA	-1.61	1.29E-12	-1.36	2.79E-14	-1.74E-01	6.22E-01
LmxM.29.1510	LmaC1N							-1.90	9.75E-30	-1.38	6.04E-08	-2.16E-01	5.57E-01
LmxM.30.2280	LdARF1	Adenosine diphosphate ribosylation factor I	*L. donovani*	PRO, AXA	Northern Blot	(Porter-Kelley et al. 2004)	AXA	0.70	8.01E-04	0.36	1.88E-01	2.90E-01	2.66E-01
LmxM.30.2310	*L . mexicana* 3'-nucleotidase/nuclease	purine uptake	*L. mexicana*	PRO, lesion AMA ("cultured axenically")	Northern Blot	(Sopwith et al. 2002)	PRO	1.80	4.59E-18	0.64	5.74E-02	7.12E-01	1.11E-02
LmxM.31.2900	DC2	outer dynein arm docking complex component; flagellar motility	*L. donovani*	PRO, AXA	Northern Blot	(Harder et al. 2010)	PRO	0.91	7.78E-08	0.71	7.10E-07	1.96E-01	2.57E-01
LmxM.31.3310	GCVL-2	subunit of the mitochondrial glycine cleavage complex	*L. infantum*	PRO, AXA	Northern Blot	(Müller and Papadopoulou 2010)	constitutive	1.77	4.02E-32	0.74	6.32E-05	9.09E-01	1.10E-07
LmxM.33.0070	Ascorbate peroxidase	oxidative stress protection	*L. donovani*	PRO, AXA	Northern Blot	(Saxena et al. 2007)	AXA	-2.66	1.29E-51	-1.32	1.00E-12	-1.20E+00	7.16E-18
LmxM.33.3645	A600-4	unknown	*L. mexicana*	PRO, lesion AMA; PRO, AXA	Northern Blot	(Bellatin et al. 2002)	leison AMA; AXA	-5.60	2.24E-102	-3.21	3.67E-46	-1.53E+00	8.17E-10
LmxM.34.2210	KMP-11	unclear	*L. infantum*	PRO, lesion AMA	Northern Blot	(Berberich et al. 1998)	PRO	-2.51	1.76E-40	-0.67	2.35E-02	-1.40E+00	9.09E-11
LmxM.34.2220	KMP-11							0.18	2.94E-01	-0.21	1.77E-01	4.02E-01	1.82E-03
LmxM.34.2221	KMP-11							0.34	1.04E-01	0.16	5.10E-01	1.82E-01	3.21E-01
LmxM.34.4720	GCVH	subunit of the mitochondrial glycine cleavage complex	*L. infantum*	PRO, AXA	Northern Blot	(Müller and Papadopoulou 2010)	constitutive	0.01	9.66E-01	0.36	8.34E-03	-3.10E-01	1.54E-01
LmxM.36.3800	GCVT-1	mitochondrial glycine cleavage complex (T-protein component)	*L. infantum*	PRO, AXA, AMA	Northern Blot; qRT	(Müller and Papadopoulou 2010)	AXA, AMA (RNA, PRO protein))	0.35	7.70E-02	0.27	1.20E-01	8.58E-02	7.74E-01
LmxM.36.3810	GCVT-2	mitochondrial glycine cleavage complex (T-protein component)	*L. infantum*	PRO, AXA, AMA	Northern Blot; qRT	(Müller and Papadopoulou 2010)	AXA, AMA (RNA and protein)	-0.42	2.71E-02	-0.01	9.57E-01	-3.84E-01	5.10E-03
LmxM.36.6280	LmGT3	Glucose transporter 3	*L. mexicana*	PRO, AXA	Northern Blot	(Burchmore and Landfear 1998)	constitutive	-1.95	1.09E-13	-0.55	9.40E-03	-1.15E+00	4.62E-07
LmxM.36.6290	LmGT2	Glucose transporter 2	*L. mexicana*	PRO, AXA		(Burchmore and Landfear 1998)	PRO	2.43	5.21E-17	1.01	2.63E-03	9.34E-01	2.17E-04
LmxM.36.6300	LmGT1	Glucose transporter 1	*L. mexicana*	PRO, AXA	Northern Blot	(Burchmore and Landfear 1998)	constitutive	1.19	2.02E-06	0.36	2.15E-01	6.70E-01	8.00E-03

proteins, CPB2.8-cysteine peptidases, KMP-11), which was not possible to be addressed in the studies due to promiscuity of probes (e.g. KMP-11) (see references within Table).

38 genes were compared (excl. multiple copies), of which 7 comparisons proved inconclusive, 7 disagreed and for 24 at least one copy of the gene showed the same differential expression profile in our study as in the published study. As a general trend, values in PRO vs AXA followed the pattern seen in the PRO vs AMA comparison, but at a smaller magnitude in terms of fold change of transcript abundance and p-value.

Particularly well-characterised differentially expressed gene in *L. mexicana* agree with our findings, such as PFR2 (Moore, Santrich, and LeBowitz 1996), the cathepsin L-like cysteine proteinases (J. C. Mottram et al. 1997), A600 genes and A850 β-tubulin (Bellatin et al. 2002) as well as the glucose transporters LmGT1 and LmGT2 (Burchmore and Landfear 1998). Disagreement exists for LmGT3, having been reported as constitutively expressed (Burchmore and Landfear 1998), but being detected as preferentially expressed in AMA and AXA in our data-set. The significance of this is not clear. LmGT3, the PFR2 upstream gene 2 (LmxM.16.1450) (Moore, Santrich, and LeBowitz 1996) and the alpha-tubulin LmxM.13.0280 (Burchmore and Landfear 1998) are the only examples where an expression pattern established in *L. mexicana* has been published that disagrees with our findings. All other cases were established in other *Leishmania spp.* raising the possibility that these disagreements are due to inter-species differences.

Overall, we may conclude that our findings are consistent with published data. Consequently, all other transcripts determined to be differentially expressed ought to be considered developmentally regulated.

3.3.2.3 Novel Genes are over-represented amongst AMA-upregulated genes

Analysis of genes differentially expressed between AMA and PRO reveals significant over-representation of novel transcripts identified in Chapter 2 amongst DE genes when compared to the genome-wide occurrence of differentially expressed genes with 95 being preferentially expressed in PRO (Chi-squared p-value = 1.3×10^{-3}) and 293 in AMA (Chi-

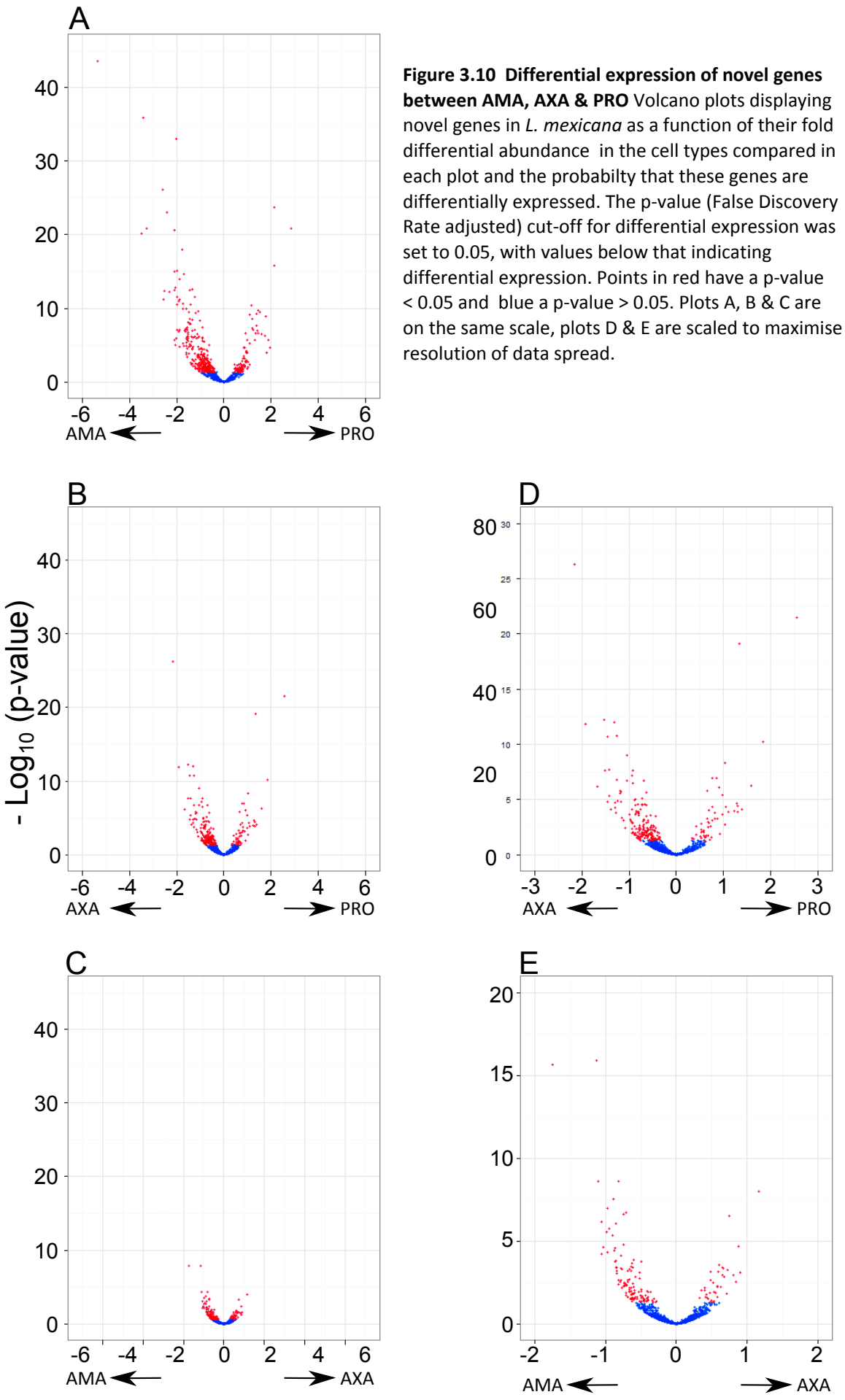


Figure 3.10 Differential expression of novel genes between AMA, AXA & PRO Volcano plots displaying novel genes in *L. mexicana* as a function of their fold differential abundance in the cell types compared in each plot and the probabilty that these genes are differentially expressed. The p-value (False Discovery Rate adjusted) cut-off for differential expression was set to 0.05, with values below that indicating differential expression. Points in red have a p-value < 0.05 and blue a p-value > 0.05. Plots A, B & C are on the same scale, plots D & E are scaled to maximise resolution of data spread.

squared p-value = $1.15x10^{-10}$). 232 novel transcripts are differentially expressed between PRO and AXA, with 52 having preferential expression in PRO and 180 in AXA. Comparison of AMA and AXA reveals a total of 119 novel transcripts with differential expression with 23 preferentially expressed in AXA and 96 in AMA. Volcano plots for the differential expression analysis of the novel genes are shown in Figure 3.10. The enrichment of preferential expression of the novel genes in AMA and AXA compared to PRO point to a role for disproportionally high number of the novel genes in amastigotes, particularly intracellular amastigotes. This role may be furthermore supported by the findings presented in Chapter 2, showing a predominantly *Leishmania*-specific conservation of the novel genes, which could again point to a particular role in biology specifically acquired upon divergence of *Leishmania spp.* from other trypanosomatids such as their intracellular parasitism (Fernandes, Nelson, and Beverley 1993).

3.3.2.4 Enrichment analyses reflect morphological, metabolic changes as well as adaptations to the milieu inhabited by parasite

To test for enrichment of known and predicted functions amongst differentially expressed genes we used the GOseq R-package (Young et al. 2010). For the length bias (Oshlack and Wakefield 2009) correction we used the transcript lengths obtained in Chapter 2 and controlled the false-discovery rate according to Benjamini-Hochberg (Benjamini et al. 2001) setting a p-value cut-off for significance at $p < 0.05$. I performed Gene Ontology (GO)-term, Pathway and PFAM-domain enrichment analyses. Moreover, I searched for over-representation of proteins predicted to have transmembrane domains and signal peptides. For the sake of legibility, transcripts preferentially expressed in AMA, AXA and PRO will be referred to as DE-AMA, DE-AXA and DE-PRO respectively.

Results of GO-term and Pathway (Goto et al. 1997; Doyle et al. 2009) enrichment analyses are shown in Tables 3.6-3.8 and 3.9-10. In comparisons of DE-PRO to either DE-AMA or DE-AXA the dominating features relate to tRNA charging, microtubule based motility and metabolism, in particular central carbon metabolism and proton-gradient driven ATP-

Table 3.6 GO-term enrichment amongst genes preferentially expressed in PRO vs AMA Table summarising GO-term enrichment results for PRO vs AMA.

GO	Description	p-val	Genes in GO-term group	Genes in DE List	GO terms in List
GO:0051082	unfolded protein binding	7.65E-05	51	1979	29
GO:0006457	protein folding	7.65E-05	59	1979	31
GO:0004812	aminoacyl-tRNA ligase activity	7.65E-05	24	1979	18
GO:0006418	tRNA aminoacylation for protein translation	7.65E-05	24	1979	18
GO:0003777	microtubule motor activity	9.74E-05	78	1979	40
GO:0030286	dynein complex	1.69E-04	13	1979	12
GO:0005737	cytoplasm	2.10E-04	112	1979	48
GO:0006334	nucleosome assembly	1.18E-03	31	1979	15
GO:0008152	metabolic process	1.28E-03	269	1979	92
GO:0005516	calmodulin binding	1.28E-03	8	1979	8
GO:0009434	microtubule-based flagellum	1.28E-03	8	1979	8
GO:0007018	microtubule-based movement	2.11E-03	81	1979	38
GO:0000786	nucleosome	2.49E-03	27	1979	13
GO:0044267	cellular protein metabolic process	2.60E-03	12	1979	10
GO:0003824	catalytic activity	5.25E-03	304	1979	98
GO:0005509	calcium ion binding	5.47E-03	29	1979	15
GO:0015986	ATP synthesis coupled proton transport	5.47E-03	28	1979	15
GO:0003746	translation elongation factor activity	7.58E-03	13	1979	9
GO:0005634	nucleus	9.37E-03	142	1979	49
GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	1.30E-02	6	1979	6
GO:0004298	threonine-type endopeptidase activity	1.30E-02	15	1979	9
GO:0005839	proteasome core complex	1.30E-02	15	1979	9
GO:0051603	proteolysis involved in cellular protein catabolic process	1.30E-02	15	1979	9

Table 3.7 GO-term enrichment amongst genes preferentially expressed in PRO vs AXA Table summarising GO-term enrichment results for PRO vs AXA.

GO	Description	p-val	Genes in GO-term group	Genes in DE List	GO terms in List
GO:0005737	cytoplasm	3.51E-05	112	1225	39
GO:0005516	calmodulin binding	1.35E-04	8	1225	8
GO:0009434	microtubule-based flagellum	1.35E-04	8	1225	8
GO:0006457	protein folding	2.87E-04	59	1225	23
GO:0004812	aminoacyl-tRNA ligase activity	3.48E-04	24	1225	14
GO:0006418	tRNA aminoacylation for protein translation	3.48E-04	24	1225	14
GO:0051082	unfolded protein binding	3.48E-04	51	1225	21
GO:0044267	cellular protein metabolic process	1.21E-03	12	1225	9
GO:0009116	nucleoside metabolic process	4.10E-03	9	1225	7
GO:0045261	proton-transporting ATP synthase complex, catalytic core F(1)	4.23E-03	7	1225	6
GO:0046933	hydrogen ion transporting ATP synthase activity, rotational mechanism	1.01E-02	11	1225	7
GO:0005875	microtubule associated complex	1.59E-02	8	1225	5
GO:0003746	translation elongation factor activity	2.96E-02	13	1225	7
GO:0006364	rRNA processing	3.77E-02	13	1225	7
GO:0009190	cyclic nucleotide biosynthetic process	4.21E-02	11	1225	7
GO:0016849	phosphorus-oxygen lyase activity	4.21E-02	11	1225	7

Table 3.8 GO-term enrichment amongst genes preferentially expressed in AXA vs AMA Table summarising GO-term enrichment results for AXA vs AMA.

GO	Description	p-val	Genes in GO-term group	Genes in DE List	GO terms in List
GO:0000786	nucleosome	1.55E-05	27	563	11
GO:0006334	nucleosome assembly	5.06E-05	31	563	11
GO:0003677	DNA binding	9.23E-05	154	563	28
GO:0006260	DNA replication	1.63E-04	33	563	12
GO:0005634	nucleus	9.01E-04	142	563	24
GO:0004298	threonine-type endopeptidase activity	1.43E-02	15	563	6
GO:0005839	proteasome core complex	1.43E-02	15	563	6
GO:0051603	proteolysis involved in cellular protein catabolic process	1.43E-02	15	563	6
GO:0004175	endopeptidase activity	1.43E-02	10	563	5

Table 3.9 Pathway enrichment analysis for PRO vs AMA Table showing the results of enrichment analyses of pathways associated with genes prefrentially expressed in PRO vs AMA.

Pathway	Description	p-val	Genes in pathway	Genes in enrichment list	Genes from enrichment list in pathway
TRNA-CHARGING-PWY	tRNA charging pathway	6.78E-03	25	1979	17
PWY3IU-61	superpathway of glycolysis, pyruvate dehydrogenase and TCA cycle	9.05E-03	45	1979	26
PWY3IU-93	superpathway of sterol biosynthesis	1.41E-02	21	1979	13
PWY3IU-99	superpathway of central carbon metabolism	1.58E-02	61	1979	32
ANARESP1-PWY	respiration (anaerobic)	1.58E-02	19	1979	12
TCA	TCA cycle	3.90E-02	21	1979	15

Table 3.10 Pathway enrichment analysis for PRO vs AXA Table showing the results of enrichment analyses of pathways associated with genes prefrentially expressed in PRO vs AXA.

Pathway	Description	p-value	Genes in pathway	Genes in enrichment list	Genes from enrichment list in pathway
TRNA-CHARGING-PWY	tRNA charging pathway	9.20E-04	25	1225	14
PWY3IU-61	superpathway of glycolysis, pyruvate dehydrogenase and TCA cycle	8.65E-03	45	1225	20
PYRUVDEHYD-PWY	acetyl-CoA biosynthesis (from pyruvate)	2.54E-02	6	1225	5

Table 3.11 PFAM-A&B domain enrichment summary Table summarising enrichement analyses of PFAM-domains, defined by gathering threshold, amongest differentially expressed genes.

Category	p-value	Number amongst DE	Numer in Genome	p-value	Description	Enrichment group
PF07344.6	2.56E-22	54	75	8.50E-19	Amastin surface glycoprotein	AMA in PROvAMA
PF07344.6	7.94E-24	38	75	2.63E-20	Amastin surface glycoprotein	AMA in AXAvAMA
PF07344.6	9.69E-31	49	75	3.21E-27	Amastin surface glycoprotein	AXA in PROvAXA
PF00125.19	7.74E-10	12	27	2.56E-06	Core histone H2A/H2B/H3/H4	AXA in AXAvAMA
PF00012.15	1.51E-05	7	14	2.51E-02	Hsp70 protein	AXA in AXAvAMA
PF03028.10	2.09E-07	13	14	3.73E-04	Dynein heavy chain and region D6 of dynein	PRO in PROvAMA
PF08393.8	2.84E-07	13	14	3.73E-04	Dynein heavy chain, N-terminal region 2	PRO in PROvAMA
PB001262	1.01E-06	12	13	3.73E-04	Pfam-B_1262	PRO in PROvAMA
PB001985	1.01E-06	12	13	3.73E-04	Pfam-B_1985	PRO in PROvAMA
PB004112	1.01E-06	12	13	3.73E-04	Pfam-B_4112	PRO in PROvAMA
PF12774.2	1.01E-06	12	13	3.73E-04	Hydrolytic ATP binding site of dynein motor region	PRO in PROvAMA
PF12775.2	1.01E-06	12	13	3.73E-04	P-loop containing dynein motor region D3	PRO in PROvAMA
PF12780.2	1.01E-06	12	13	3.73E-04	P-loop containing dynein motor region D4	PRO in PROvAMA
PF12781.2	1.01E-06	12	13	3.73E-04	ATP-binding dynein motor region D5	PRO in PROvAMA
PF13833.1	1.29E-06	11	14	4.27E-04	EF-hand domain pair	PRO in PROvAMA
PB002892	3.27E-06	11	12	9.86E-04	Pfam-B_2892	PRO in PROvAMA
PB006187	3.59E-06	11	12	9.92E-04	Pfam-B_6187	PRO in PROvAMA
PF00125.19	4.25E-06	14	27	1.08E-03	Core histone H2A/H2B/H3/H4	PRO in PROvAMA
PF05149.7	1.40E-05	8	8	3.33E-03	Paraflagellar rod protein	PRO in PROvAMA
PB001252	2.04E-05	12	15	4.24E-03	Pfam-B_1252	PRO in PROvAMA
PF12777.2	2.05E-05	12	15	4.24E-03	Microtubule-binding stalk of dynein motor	PRO in PROvAMA
PF00118.19	3.64E-05	10	12	7.11E-03	TCP-1/cpn60 chaperonin family	PRO in PROvAMA
PB010967	8.43E-05	7	7	1.55E-02	Pfam-B_10967	PRO in PROvAMA
PF00137.16	0.000139071	7	9	2.43E-02	ATP synthase subunit C	PRO in PROvAMA
PB001679	0.000152332	8	9	2.52E-02	Pfam-B_1679	PRO in PROvAMA
PF00227.21	0.000300214	9	15	4.74E-02	Proteasome subunit	PRO in PROvAMA
PF05149.7	4.05E-07	8	8	1.34E-03	Paraflagellar rod protein	PRO in PROvAXA
PF13833.1	1.84E-06	9	14	3.05E-03	EF-hand domain pair	PRO in PROvAXA
PF00118.19	9.65E-06	9	12	1.07E-02	TCP-1/cpn60 chaperonin family	PRO in PROvAXA
PF01145.20	4.08E-05	5	5	3.38E-02	SPFH domain / Band 7 family	PRO in PROvAXA
PF00251.15	6.96E-05	6	7	4.30E-02	Glycosyl hydrolases family 32 N-terminal domain	PRO in PROvAXA
PF00612.22	7.78E-05	7	9	4.30E-02	IQ calmodulin-binding motif	PRO in PROvAXA

Table 3.12 TMD enrichment analyses Enrichment of genes predicted to encode TMD-containing proteins amongst DE genes. Prediction performed using TMHMM 2.0.

Description	Proteins with TMDs	Genes in List	TMD containing proteins in list	p-val	Enrichment group
Enriched for TMDs	1615	1853	417	4.29E-10	AMA in PROvAMA
Enriched for TMDs	1615	671	197	9.41E-15	AXA in AXAvAMA
Enriched for TMDs	1615	951	264	1.13E-16	AXA in PROvAXA

Table 3.13 SP-enrichment analyses Enrichment of genes predicted to encode proteins with predicted signal peptides amongst DE genes. Prediction using SignalP3.0 using both HMM and/or NN.

Category	Number amongst DE	Numer in Genome	padj	Enrichment Group
Enriched for SP	154	1763	1.24E-02	AMA in AXAvAMA
Enriched for SP	400	1763	7.30E-03	AMA in PROvAMA
Enriched for SP	244	1763	2.84E-07	AXA in PROvAXA
Depleted of SP	1020	7401	2.29E-02	PRO in PROvAXA

synthesis (Tables 3.6 -7 & 3.9-10). Comparing DE-AXA against DE-AMA shows an overrepresentation of GO-terms relating to nucleosome assembly and DNA replication as well as proteolytic/proteosomal activity (Table 3.12).

PFAM-domain enrichment (Table 3.11) in the comparison of DE-PRO to DE-AMA and DE-AXA analyses particularly show enrichment of PFAM-domains associated with microtubule based motility such as dyneins (Wickstead and Gull 2007). Equally, Paraflagellar rod proteins, found in the motile flagellum of kinetoplastids (Hyams 1982), including trypanosomatids (Russell et al. 1983), are overrepresented amongst DE-PRO genes compared to DE-AMA and DE-AXA.

Only the PFAM domain Amastin surface glycoprotein is enriched for amongst DE-AMA and DE-AXA when compared to DE-PRO, with the expression of these proteins being higher in DE-AMA versus DE-AXA (Table 3.11). Strikingly, in both DE-AXA and DE-PRO we find overrepresentation of core histones H2A/H2B/H3/H4 compared to DE-AMA, which may suggest differing growth rates between the three cell types analysed.

Transcripts predicted to encode for proteins with TMDs (Table 3.12) and SPs (Table 3.13) are significantly enriched in DE-AMA and DE-AXA compared to DE-PRO. Indeed proteins with TMDs are significantly depleted in DE-PRO compared to DE-AXA. Transcripts predicted to encode proteins with SPs are also enriched for in DE-AXA compared to DE-AMA. This is indicative of wide-ranging adaptation of the population of surface-proteins from PRO to amastigote cells-types, of course bearing in mind that not all transmembrane proteins localise to the cell surface.

3.3.2.5 Correlation of differential expression analysis with SINQ-proteomics

In section 3.3.4.5 it was established that there is very little correlation of transcript abundances as quantified by FPKM and protein abundances as quantified by SINQ scores. Using the SINQ scores obtained for AXA and PRO I sought to correlate fold-change in SINQ (fcSINQ) score with fold differences in transcript abundances (fcDESeq) (Figure 3.11).

fcSINQ and fcDESeq show very little linear correlation with an R^2-value of 0.09 (Figure 3.11 A). If only genes are compared that satisfy the statistical cut-off for differential transcript expression expression ($p < 0.05$), the R^2-value for linear correlation increases to 0.18 (Figure 3.11 B). If ever more stringent p-value cut-offs are applied, the correlation of fcSINQ and fcDESeq continues to improve (Figure 3.11 C), but even at a p-value $<1x10x^{-25}$ an R^2-value to the linear regression of only 0.6 is reached. From this it may be said that correlation between fcSINQ and fcDESeq in this data set is not good, albeit the changes of the most consistently differentially expressed genes on a transcript level (i.e. lowest p-values) are also reflected on a proteomic level.

The main drawback contributing to the low correlation of changing transcript- with changing protein-levels may be entirely technical: No replicate experiments were conducted for the mass-spectrometric measurements as these were not initially aimed at being used in a quantitative manner, but just for the detection of peptides for novel genes and extensions. It will be very interesting to see whether correlation of fcSINQ and fcDESeq would improve if a statistical threshold based on replicate data could be applied to both data sets. Moreover, RNA and proteins were not extracted from the same sample which could also contribute to the lack of correlation.

3.3.3 Distribution of DE-genes across chromosomes reveals functional biases of chromosomes and may point to key role of *L. mexicana* chromosome 30 in amastigote biology

Core transcriptional mechanisms are conserved between *Trypanosoma spp.* and *Leishmania spp.* Unlike in Prokaryotes, genes within a single polycistronic transcription unit are not functionally linked in *T. brucei* (Kelly et al. 2012), with the exception of duplicated genes or gene arrays, and it is very likely that this is the case in *Leishmania spp.* too. Nevertheless, I wondered whether the differential expression data could help identify

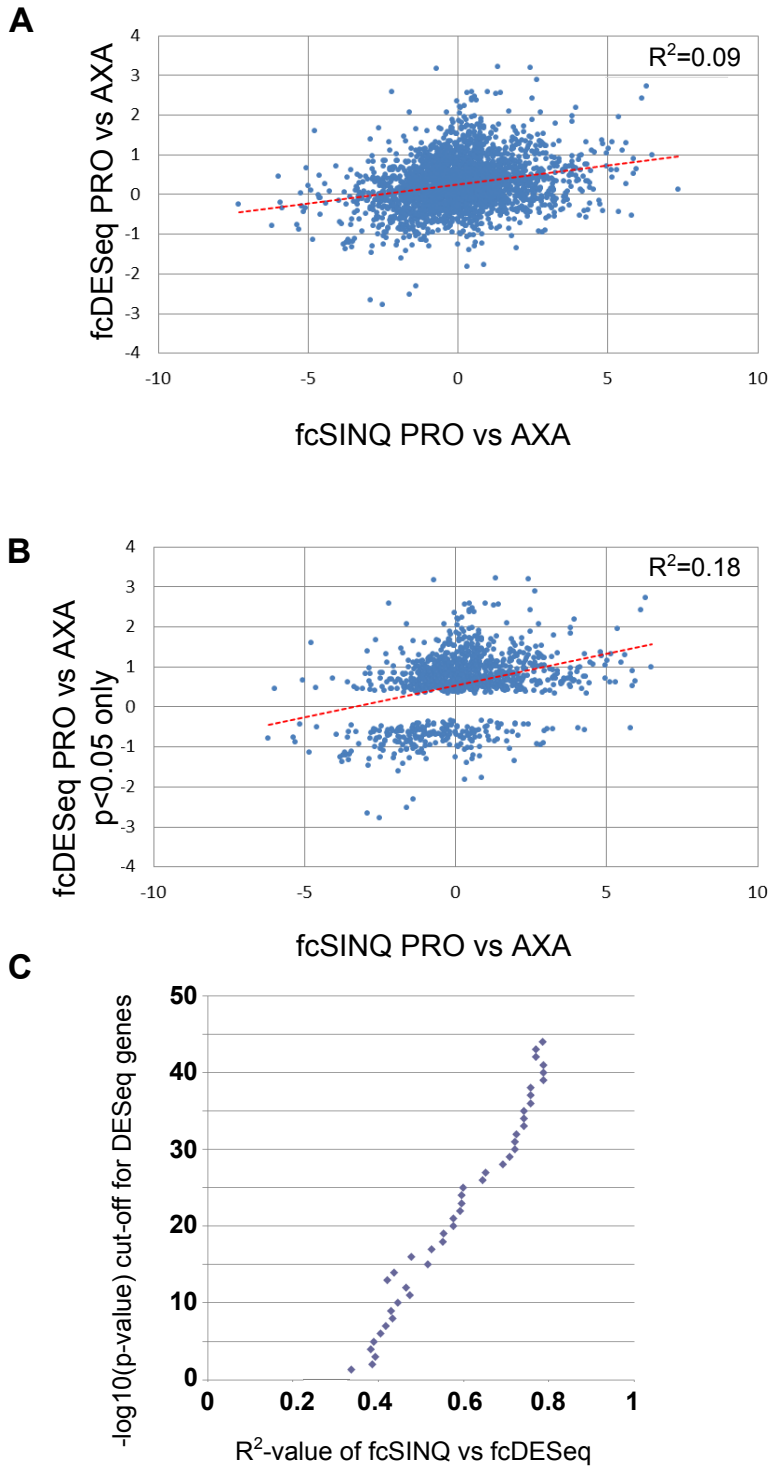


Figure 3.11 Correlation of fcSINQ and fcDESeq Plots of $\log_2$(fold change SINQ-score) (fcSINQ) against $\log_2$(fold change transcript abundance) (fcDESeq) for (A) all genes these information were available for and (B) a subset of these genes which have a p-value < 0.05 for differential transcript expression. (C) shows the effect of more stringent p-value cut-offs on linear correlation between fcSINQ and fcDESeq.

clusters of genes with similar expression patterns, in particular clusters of DE-AMA genes that may reveal a "pathogenicity island" within the genome.

Figure 3.12 shows all genes of *L. mexicana* plotted over a chromosome scaffold with a colour code indicating constitutive or differential expression between AMA and PRO. Genes located on the "top" or "bottom" strand are displayed as such to show unidirectional gene clusters, giving a partial indication of the starts and ends polycistronic transcription units (PTUs) situated at strand switch regions. This is by no means an exhaustive map of PTUs as transcription termination and start sites have also found within stretches of genes located on the same in *L. tarentolae* and *L. major* strand (van Luenen et al. 2012; Reynolds et al. 2014), however no such information is available for *L. mexicana.* It may have been possible to transfer PTU-boundaries from the data for *L. major,* these were not published at the time this study was performed (April 2014). With the exception of an amastin-rich cluster of DE-AMA genes located close to the 3rd strand-switch region from the right on chromosome LmxM.08_29, it is difficult to discern clusters of DE-AMA or DE-PRO genes in this analysis.

With no information about the boundaries of PTUs available for *L. mexicana*, it is not possible to calculate enrichment of DE-genes within a single PTU. I therefore determined the proportion of DE-PRO and DE-AMA genes on each chromosome. Table 3.14 shows the number of DE-AMA , DE-PRO and constitutively expressed genes on each chromosome as well as the ratios of these to each other on that chromosome as well as statistical testing results for significant difference to the genome average.

Calculations for DE-AMA were performed under inclusion and exclusions of amastins on that chromosome to exclude heavy skewing of results by this expanded gene family (Jackson 2010) with a predominantly amastigote-enriched expression reported in the literature (A. Rochette et al. 2005) and in my data-set (54/74 amastins are DE-AMAvPRO) (see Materials and Methods for identification of Amastins).

Chromosomes LmxM.05 and LmxM.14 show significant enrichment (Chi-squared test

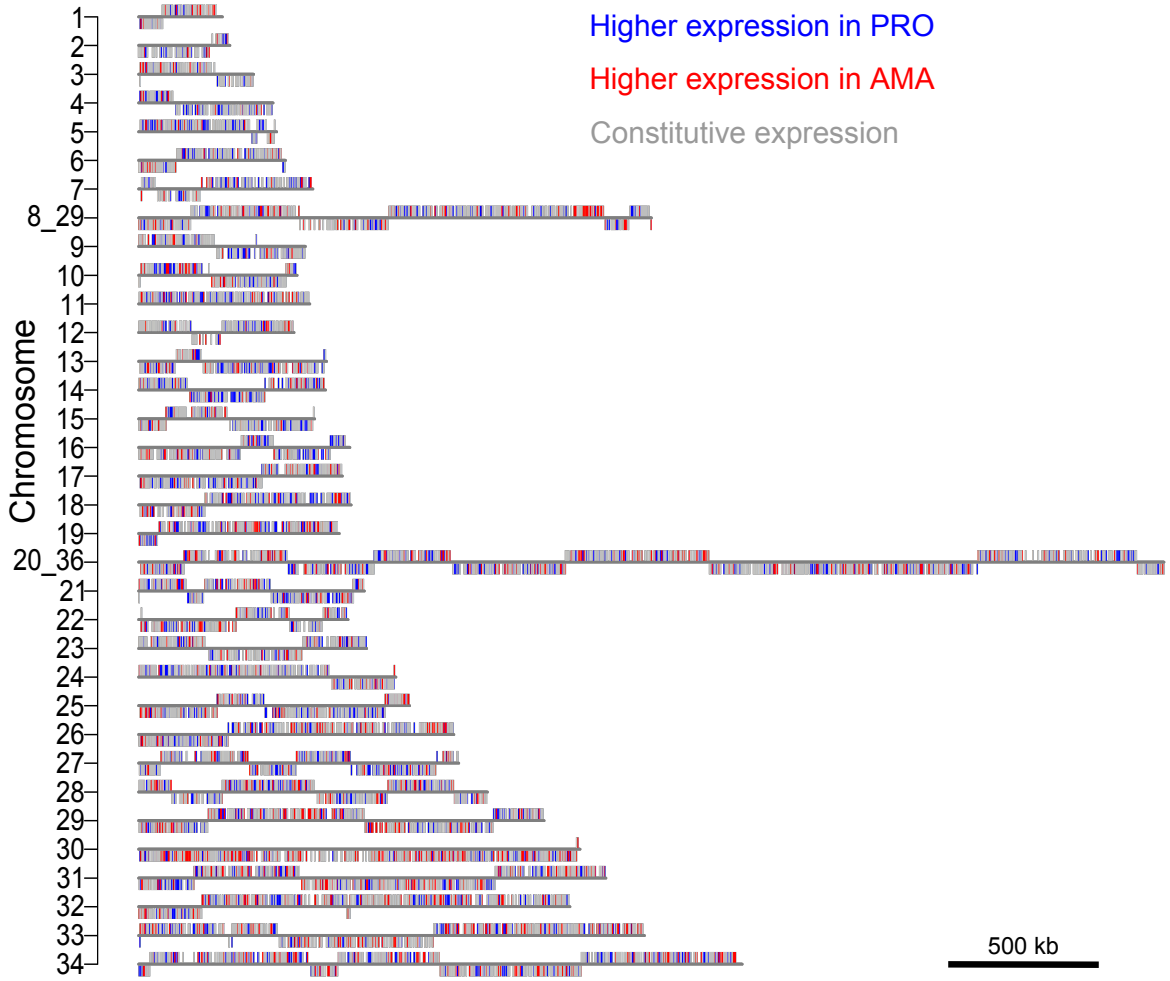


Figure 3.12 Distribution of differentially and constitutively expressed genes across chromosomes of *L. mexicana* Figure showing the location of all genes across all chromosomes of *L. mexicana*. Genes preferentially expressed in AMA or PRO are shown in red or blue respectively. Constitutively expressed genes are shown in grey. Genes are displayed as being on the top (+) or bottom (-) strand. As position of eachthe gene, the position of the stop codon is used.

Table 3.14 Distribution of differentially expressed genes across *L. mexicana* chromosomes Table showing the number of DE-PRO, DE-AMA and constitutively expressed genes across *L. mexicana* chromosomes, including and excluding (in brackets) amastins. Moreover, the ratio of DE-AMA and DE-PRO relative to the gene number on each chromosome is given, with a blue highlighting indicating a promastigote signature, i.e. either enriched for DE-PRO or depleted of DE-AMA relative to genome average (light blue), or both (dark blue). A red colour highlighting an amastigote signature, i.e. enriched for either DE-AMA or depleted of DE-PRO relative to genome average (light red), or both (dark red). Green highlights statistically significant (p<=0.05) differences to the global distribution DE-PRO or DE-AMA by Chi-squared test. "TotalChr" stands for all genes present on the chromosome.

Chromosome	Amastins	Constitutive	PRO-enriched	AMA-enriched	AMAvPRO	AMAv TotalChr with amastins.	AMAvTotalChr without amastins	PROvTotalChr with amastins	Chi-square p-value for AMA with amastins	Chi-square p-value for AMA without amastins	Chi-square p-value for PRO with amastins
LmxM.01		60	15	15	1.00	0.17	0.17	0.17	4.43E-01	5.10E-01	2.41E-01
LmxM.02		45	17	10	0.59	0.14	0.14	0.24	2.02E-01	2.36E-01	7.09E-01
LmxM.03		72	15	14	0.93	0.14	0.14	0.15	1.30E-01	1.59E-01	9.27E-02
LmxM.04		81	37	12	0.32	0.09	0.09	0.28	2.40E-03	3.44E-03	6.76E-02
LmxM.05		78	39	14	0.36	0.11	0.11	0.30	8.52E-03	1.19E-02	2.83E-02
LmxM.06		103	27	22	0.81	0.14	0.14	0.18	9.55E-02	1.25E-01	2.33E-01
LmxM.07		83	34	19	0.56	0.14	0.14	0.25	8.47E-02	1.10E-01	3.68E-01
LmxM.08_29	11	291	81	104 (93)	1.28 (1.15)	0.22	0.20	0.17	3.03E-01	7.60E-01	1.36E-02
LmxM.09		109	36	24	0.67	0.14	0.14	0.21	6.51E-02	8.84E-02	8.80E-01
LmxM.10	2	84	42 (41)	31 (30)	0.74 (0.73)	0.20	0.19	0.27	9.91E-01	9.82E-01	1.36E-01
LmxM.11		91	35	23	0.66	0.15	0.15	0.23	1.74E-01	2.22E-01	6.18E-01
LmxM.12		85	20	18	0.90	0.15	0.15	0.16	1.45E-01	1.82E-01	1.40E-01
LmxM.13		106	44	30	0.68	0.17	0.17	0.24	2.80E-01	3.54E-01	3.93E-01
LmxM.14	2	91	59 (58)	23 (22)	0.39 (0.38)	0.13	0.13	0.34	3.30E-02	3.13E-02	1.08E-04
LmxM.15		113	38	21	0.55	0.12	0.12	0.22	1.20E-02	1.75E-02	9.23E-01
LmxM.16		119	46	32	0.70	0.16	0.16	0.23	2.02E-01	2.63E-01	5.99E-01
LmxM.17		107	42	32	0.76	0.18	0.18	0.23	4.57E-01	5.56E-01	6.48E-01
LmxM.18		101	49	38	0.78	0.20	0.20	0.26	9.18E-01	7.87E-01	1.60E-01
LmxM.19		99	46	38	0.83	0.21	0.21	0.25	7.74E-01	6.51E-01	2.78E-01
LmxM.20_36	3	583 (581)	221	204 (203)	0.92 (0.92)	0.20	0.20	0.22	7.81E-01	5.58E-01	8.95E-01
LmxM.21		124	65	42	0.65	0.18	0.18	0.28	5.16E-01	6.37E-01	2.13E-02
LmxM.22		106	41	36	0.88	0.20	0.20	0.22	9.37E-01	9.34E-01	8.41E-01
LmxM.23		140	40	37	0.93	0.17	0.17	0.18	2.97E-01	3.82E-01	2.36E-01
LmxM.24	4	170 (168)	54 (53)	39 (38)	0.72 (0.72)	0.15	0.15	0.21	5.01E-02	5.59E-02	6.72E-01
LmxM.25		158	65	54	0.83	0.19	0.19	0.23	8.65E-01	9.77E-01	5.05E-01
LmxM.26		171	58	68	1.17	0.23	0.23	0.20	2.06E-01	1.38E-01	3.53E-01
LmxM.27	1	170 (169)	77	57	0.74	0.19	0.19	0.25	6.38E-01	7.90E-01	1.33E-01
LmxM.28	3	206 (205)	76	69 (67)	0.91 (0.88)	0.20	0.19	0.22	9.29E-01	9.36E-01	9.74E-01
LmxM.29	3	224	90 (89)	100 (98)	1.10 (1.10)	0.24	0.24	0.22	3.28E-02	2.74E-02	9.82E-01
LmxM.30	15	223	57	131 (116)	2.23 (2.04)	0.32	0.29	0.14	4.03E-09	1.45E-06	1.31E-04
LmxM.31		265	95	97	1.02	0.21	0.21	0.21	4.92E-01	3.44E-01	6.13E-01
LmxM.32		230	73	88	1.21	0.23	0.23	0.19	2.09E-01	1.33E-01	1.43E-01
LmxM.33	30	262 (252)	99	116 (96)	1.17 (1.07)	0.24	0.21	0.21	9.60E-03	2.86E-01	7.64E-01
LmxM.34		320	130	135	1.04	0.23	0.23	0.22	6.39E-02	3.13E-02	8.05E-01
						Genome mean= 0.20	Genome mean= 0.19	Genome mean= 0.20			

p<0.05) for DE-PRO, around 1.5 fold increased ratio of DE-PRO genes against total number of genes on chromosome, and significant depletion (Chi-squared test p<0.05) for DE-AMA genes, around 1.5 fold decrease of DE-AMA genes compared to total number of genes on chromosome. Chromosomes LmxM.04 and LmxM.15 are similarly depleted of DE-AMA, but show no statistically significant (Chi-squared test p<0.05) enrichment of DE-PRO genes. LmxM.21 is enriched for DE-PRO, but not significantly depleted for DE-AMA, inversely LmxM.07 is depleted of DE-PRO, but not enriched for DE-AMA. Significant enrichment for DE-AMA genes, but no significant depletion for DE-PRO is found on LmxM.29. LmxM.33 and LmxM.34 show a similar pattern as LmxM.29, only upon inclusion or exclusion of amastins from calculations respectively. Chromosome LmxM.30 in turn shows significant enrichment of DE-AMA and depletion of DE-PRO (including and excluding amastins).

Recent advances in nucleotide sequencing technologies have advanced research in the long-standing field (Bastien, Blaineau, and Pages 1992; Cruz, Titus, and Beverley 1993) of investigation into aneuploidy in *Leishmania spp.* and its role in pathogenicity (M. B. Rogers et al. 2011; Sterkers et al. 2012; Mannaert et al. 2012; Lachaud et al. 2014). Curiously, in recent studies, *L. mexicana* chromosome 30, and its orthologous chromosomes in other *Leishmania* species, e.g. LmjF.31, LinJ.31, LbrM.31, have emerged as being supernumerary in all examined strains and at least tetrasomic in all but one out of 17 isolates from four *Leishmania Leishmania spp.* (tetrasomic in *L. mexicana*) (M. B. Rogers et al. 2011; Mannaert et al. 2012), which suggests a requirement of increased transcript-dosage from these chromosomes.

Taking these data together, a key role of chromosome LmxM.30 (and its orthologues) in the amastigote stage may be emerging based on differential expression data and pattern of aneuploidy.

3.4 Conclusions

Quantification of mapped sequencing reads generated gene expression profiles for AMA, AXA and PRO with transcript abundances ranging over 5 orders of magnitude. Expression levels detected for AMA samples were consistently lower than those of AXA and PRO, owing to the mixed-species nature of the AMA sample with a proportion of sequencing reads mapping to the murine part of the hybrid-genome used in quantification. The very highest expressed genes were shared between the three cell types analysed and comprise e.g. histones, ribosomal proteins and heat-shock proteins. Amongst the highest expressed transcripts I also found novel genes discovered in Chapter 2. Whilst amongst the 13 novel genes in this category four were ribosomal proteins and one was a histone, the remaining novel genes had no predictable function and represent interesting candidates for biochemical characterisation.

Amongst the genes for which no FPKM values were determined are also novel genes annotated in Chapter 2. The gene models were predicted using the position of mapped SLAS and all feature PAS, but the total number of reads mapping to the transcript were too low to determine. However, there exists the possibility that these are technical artefacts from the generation of the annotation. As discussed in Chapter 2, we employed as CDS-centric approach to transcript definition. A transcript-centric approach, would have also captured these novel genes, as they feature both SLAS and PAS (a requirement for all the novel genes). A third method of defining transcripts based on sequencing read-coverage to delineate transcript could have flagged these "non-expressed transcripts" up as potential false positives, but such a method comes with its own caveats as illustrated in Section 2.3.3. Rastrojo *et al.* (Rastrojo et al. 2013) employed a combination of coverage- and transcript-centric approaches, which, the light of these "non-expressed transcripts" may have its merits. But the precise hierarchy in which these methods are implemented is a contentious topic as will be discussed in Chapter 5.

DE testing showed that at 24 h post-infection around 41.8 % of all genes were differentially expressed based on a statistical (p-value) cut-off between AMA and PRO. Moreover, 23.7 % and 13.5 % of transcripts are differentially expressed in AXA versus PRO and AMA versus AXA respectively. The latter comparison clearly shows that AXA and AMA are not equivalent, supporting previous findings (Holzer, McMaster, and Forney 2006; Rochette et al. 2009), and warranting re-consideration of whether or not AXA really are a good model *in lieu* of AMA (see Chapter 5). However, these reported differences are at a single 24 h time-point of amastigogenesis, and may not necessarily represent terminal differences, but may be transient owing to differ kinetics of intracellular and axenic differentiation.

There was a significant enrichment of novel genes amongst the genes preferentially expressed in AMA versus PRO, compared to the global proportion of differentially expressed genes raising the tantalising notion that these genes may be involved in host-parasite interaction.

Comparison of the DE-data to previously published reports of differentially expressed genes, as well as the results of the enrichment testing amongst DE genes, reflected the transition from the promastigote stage with a motile flagellum, to an immotile amastigote, with increased expression of transmembrane- (and potentially surface-) proteins, are consistent with previous reports. The role of transmembrane proteins for the amastigote stage, especially their role in establishing infection, will receive further attention in Chapter 4.

The vast catalogue of yet uncharacterised differentially expressed genes will be an important resource in future studies into promastigote and amastigote biology. However, the definition of differential expression was based solely on a p-value cut-off, with no fold-change threshold. Arguing for a fold-change cut-off might intuitively make sense, but, the degree of translational control in *Leishmania spp.* (Lahav et al. 2011), is a highly

complicating factor that precludes intuitive interpretation of fold-changes. Moreover, I show that there is very little correlation between fold-changes at the transcript versus at the protein level, albeit this particular analysis would have benefitted from replicate proteomic data and therefore has to be viewed with care.

The distribution of DE genes across chromosomes shows that *L. mexicana* chromosome 30 is enriched for amastigote enriched genes, which, taken together with the persistent aneuploidy of LmxM.30 (and its orthologous chromosomes in *Leishmania Leishmania spp.*), may point to a particular role of this chromosome in the amastigote stage and possibly in pathogenicity. We were not able to discern any obvious "pathogenicity islands" within chromosomes, however integration of the precise boundaries of polycistronic transcription units into enrichment analyses may in the future reveal these, which may have important implications for our understanding of how intracellular parasitism in *Leishmania spp.* evolved.

Chapter 4 – Identification, bioinformatic characterisation and sub-cellular localisation of amastigote upregulated proteins

4.1 Introduction

Pathogens utilise a diverse array of mechanisms to influence host responses. The search for factors utilised by pathogens to achieve this goal is interesting not only from a purely intellectual point of view, but also bears promises for the design of medical interventions. A variety of *Leishmania spp.* proteins have in the past been shown to be involved in or indeed crucial to establishing infection. (For a recent review, see (Kima 2014)).

Proteases form a well-characterised group of virulence factors in *Leishmania*. Well-established examples are members of a small multigene family of cell surface zinc metalloproteases called Major Surface Protease, Leishmanolysin or GP63. These membrane-anchored surface endopeptidases (Chaudhuri and Chang 1988; Ip et al. 1990) are found on promastigotes and are able to interact with a variety of host-cell surface components (Miller, Reed, and Parsons 1990). For example, they directly adhere to the complement receptor type 3 on macrophages (Russell and Wright 1988). This adhesion has in turn been postulated to promote uptake by macrophages. GP63 is also found, in amastigotes, albeit with different glycosylation patterns and with addition of the phosphatidylinositol membrane anchor restricted to only a sub-population of GP63 molecules (Medina-Acosta et al. 1989; Frommel et al. 1990). One function of GP63 is a role in preventing complemented-mediated lysis. The complement system is branch of the innate immune system that can recognise foreign organisms and initiate their destruction directly and indirectly by promoting inflammation, leading to killing and clearance of the foreign cell through other immune cells such as neutrophils and macrophages. In brief, the complement protein C3 may be cleaved by a convertase complex recruited to a microbial surface by lectins or antibody mediated antigen recognition. The cleavage product C3a

acts as a proinflammatory chemoattractant for other immune cells, whist C3b attaches covalently to the microbial surface. Additionally, C3 may spontaneously hydrolyse and C3b can, in the absence of inhibitory molecules present on host cells, attach to any cell surface. C3b recruits additional factors to form the C5-convertase, which in turn generates the proinflammatory chemoattractant C5a and the C5b protein, responsible for initiating assembly of the membrane attack complex, a multi-protein pore in the cell-surface of the microbial cell, and lysis of the microbe (for recent reviews of the interactions of trypanosomatids with the complement system refer to (Evans-Osses, de Messias-Reason, and Ramirez 2013; Cestari et al. 2013)). GP63 converts C3b into iC3b , which is unable to form the C5-convertase and additionally is the ligand of the complement receptor 3, promoting phagocytosis of the parasite by macrophages (Brittingham et al. 1995; P. B. Joshi et al. 1998). GP63 knock-out mutants of *Leishmania major* show delayed and reduced lesion formation in a mouse model (Phalgun B. Joshi et al. 2002).

Moreover GP63 was shown to cleave intracellular host-cells proteins, like the transcription factors NF-κB, CREB and AP-1 (Gregory et al. 2008; Gomez et al. 2009), inhibiting production of inflammatory cytokine (e.g. TNF-α, IL-12) and nitric oxide as well as reducing sensitivity to INF-γ (Contreras et al. 2010), promoting parasite survival.

Another important group of pathogen factors involved in infection are the amastigote-upregulated Cysteine Peptidases B (CPB) (Robertson and Coombs 1993). Displaying diverse substrate specificities (Mottram et al. 1997), they have been established as virulence factors (Mottram et al. 1996; Bart et al. 1997) essential for intracellular survival (Frame, Mottram, and Coombs 2000). Indeed deletion of of CBPs in *L. mexicana* leads to a protective host Th1 immune- response compared to the Th2 responses typical of high parasite burdens during infection (Denise et al. 2003; Buxbaum et al. 2003). This was put down to CPB-mediated degradation of IκBα, IκBβ, and NF-κB, leading do a suppression of IL-12 production (Cameron et al. 2004) required for Th1 T-cell differentiation (Hsieh et al. 1993).

Lipophosphoglycan (LPG) is a member of a very different group of virulence factors. *Leishmania spp.* synthesise a variety of glycoconjugates associated with virulence (Descoteaux et al. 1995), most notably the membrane bound lipophosphoglycans (Descoteaux et al. 1995; Descoteaux and Turco 1999). Different LPG species are found on procyclic promastigotes and complement resistant, infective metacyclic promastigotes (Howard, Sayers, and Miles 1987; Turco and Descoteaux 1992), where a thicker LPG coat is formed. Upon internalisation of promastigotes by phagocytes, LPG is crucial to initially delay maturation of the phagosome and permit differentiation of promastigotes to amastigotes. This is thought to happen by a number of effects LPG has in the early stages of infection. LPG promotes F-actin accumulation around the phagosome, interfering with vesicular fusions and host-cell component recruitment. This is achieved by LPG-mediated inhibition of PKC-alpha, which participates in F-actin breakdown and regulation of phagosome maturation (Holm et al. 2001). Moreover, LPG may be transferred to the phagosome membrane (Dermine et al. 2005; Winberg et al. 2009) where, possibly by disruption of lipid microdomains, it prevents the assembly of NADPH-oxidase (Pham, Mouriz, and Kima 2005; Lodge, Diallo, and Descoteaux 2006) and recruitment of v-ATPases (Vinet et al. 2009), responsible for anti-microbial respiratory burst and vacuolar acidification, respectively (Flannagan, Cosío, and Grinstein 2009).

Additionally, LPG interacts with toll-like receptor-2 (TLR-2) on macrophages, reducing the expression of TLR-9 in a process mediated by transforming growth factor β (TGF- β) and interleukin 10 (IL-10) (Srivastava et al. 2013). As a result, host-protective responses triggered by activation of TLR-9 are supressed.

Another example of proteins enabling amastigotes to thrive intracellularly focus on the uptake of nutrients by the parasite. *Leishmania spp.* are deficient for the heme-biosynthetic pathway (Kořený, Oborník, and Lukeš 2013) and are reliant on the *Leishmania* Heme Response 1 protein (LHR1). LHR1 is an essential protein that mediates heme uptake (Chau Huynh et al. 2012). *L. amazonensis* amastigotes deficient for one copy

of LHR1 failed to replicate intracellularly and were severely impaired in their ability to develop cutaneous lesions in a mouse model (Miguel et al. 2013). Equally, the uptake of iron is crucial to amastigote differentiation and survival. The LIT1 ferrous iron (Fe^{2+}) transporter is induced following iron starvation (C. Huynh, Sacks, and Andrews 2006; Miguel et al. 2013). LIT1 deficient *L. amazonensis* were able to persist intracellularly, but failed to proliferate and form lesions in mouse models (C. Huynh, Sacks, and Andrews 2006). Crucially, LIT1 functions to accumulate intracellular iron, which promotes growth arrest of promastigotes and differentiation to amastigotes, in a process dependent on production of superoxide radicals by the parasite's own iron superoxide dismutase (Miguel et al. 2013). This illustrates the intricate relationship between the host-cell milieu inhabited by the parasite and parasite factors that promote cellular adaptions enabling the parasite to survive.

The above examples illustrate the understanding we have of some mechanisms and pathways enabling *Leishmania spp.* to thrive intracellularly, but large parts of *Leishmania* genomes are still functionally uncharacterised. So it may not surprise that we know of some parasite factors that are known to be important for virulence, yet their functions remains largely elusive. The four A600-genes are a group of genes conserved in *Leishmania spp.* but absent from *T. brucei* (Murray, Lynn, and McMaster 2010). They are preferentially expressed in amastigotes and encode membrane bound or secreted proteins (Bellatin et al. 2002). Deletion-studies in *L. mexicana* showed that they are essential for amastigote proliferation, but the exact mechanism of this action is unknown (Murray, Lynn, and McMaster 2010).

The A2 genes, have been implicated in influencing disease-tropism between *Leishmania spp.* as they enable a visceral disease manifestation. Transfer of the A2 gene from *L. donovani*, which causes visceral disease, to *L.* major, which causes a cutaneous form of leishmaniasis, led to a visceral-like disease manifestation in a mouse-model caused by the transgenic *L. major* species (Zhang et al. 2003). A2 localise to the endoplasmic reticulum

(McCall and Matlashewski 2010) where they protect against heat-shock and oxidative stress. Again however, the precise mechanism of action is unclear.

And even seemingly well-known proteins are often ill-characterised: Little is known about the functions of the abundant amastin surface proteins (Teixeira et al. 1994; Wu et al. 2000; Rochette et al. 2005; Jackson 2010). Amastin proteins have four transmembrane domains, and two extracellular serine-threonine-rich domains containing glycosylation sites (Teixeira et al. 1994; Rochette et al. 2005). Rochette (Rochette et al. 2005) found the sequences of the transmembrane and extracellular domains to vary considerably between amastin-family members, as do the sequences of the C-terminal domains. Moreover, a 24-amino acid signal peptide was found to be a common feature of amastins (Rochette et al. 2005). Amastins are arranged in loci composed of tandem gene arrays (Wu et al. 2000; Ivens et al. 2005). Four amastin subtypes are known, α-, β-, γ-amastins, which each are located at a single locus in *Leishmania spp.* genomes, and δ-amastins, which are highly expanded in *Leishmania* and present on several chromosomes (Rochette et al. 2005; Jackson 2010). δ -amastin genes are often adjacent to tuzin genes (Jackson 2010), a group of transmembrane proteins of unknown function on their own and in relation to amastins (Teixeira, Kirchhoff, and Donelson 1995; Teixeira, Kirchhoff, and Donelson 1999).

Amastins are preferentially expressed in amastigotes (Wu et al. 2000; Rochette et al. 2005), which is corroborated by the transcriptomic analyses in Chapter 3. An ectopic expression study found them to localise to the cell surface in PRO and AXA and also to the flagellum in PRO. Amastins, particularly δ-amastins, are highly expanded in the *Leishmania* lineage (Jackson 2010), with the exception in the lizard-infective *L. tarentolae* (Raymond et al. 2011), suggesting an important role during infection of mammalian hosts. Roles in signalling have been postulated (Jackson 2010), but to date no functional data has been published.

Without a shadow of doubt will the vast cohort of yet uncharacterised genes encode further proteins that enable establishment of infection and replication of intracellular amastigotes and in the following study I sought to identify such proteins.

4.2 Aims

The aim of the following study was to use RNA-sequencing data to identify genes important to *Leishmania* amastigote survival, particularly to shed light on possible functions of putative surface proteins found to be upregulated in amastigotes. A cohort of candidates was tagged with green-fluorescent protein (GFP) to study sub-cellular localisation in PRO and AXA. Moreover, for a subset of promising candidates, a comprehensive bioinformatic analysis was undertaken.

4.3 Results

4.3.1 Identification of candidate genes

To identify proteins potentially involved in the interaction of amastigotes with the host cell, a candidate pool of proteins was established starting with all proteins for which the corresponding transcript was found to significantly ($p <= 0.05$) higher expressed in AMA compared to PRO and subsequently a set of criteria were used to narrow down further these candidates (Figure 4.1).

At the time these genes were chosen, only preliminary differential expression data was available, i.e. sequencing reads from all 3 replicate of AMA and PRO mapped to the reference CDS models and quantified using methods that have now been superseded by more recently developed versions. These data were generated by mapping and quantification of sequencing reads to a hybrid mouse-*L. mexicana* genome using RSEM (Li and Dewey 2011) (default parameters) using annotated coding-sequence models Differential expression testing was performed using DEseq (version 1 as opposed to DESeq2 used in Chapter 3) (Anders and Huber 2010) (see Materials and Methods). As a

Selection Criteria	Candidate Genes
• Upregulated in AMA vs PRO with p=< 0.05	732
• Absent from *T. brucei* (by annotation and synteny)	395
• Predicted transmembrane domains	154
• Not a known protein/ protein with predictable function based on annotations, homologies and and Pfam domains e.g. – No amastins – No proteases – No metabolic enzymes – No transporters (with exceptions)	64
• Predicted intracellular C-terminus	
• Clonable 3' IGS <6 kb	26

Figure 4.1 Selection criteria for tagging screen The selection criteria used to identify candidate genes for the tagging screen from the preliminary RNA-seq data set as described in section 4.3.1. The number of genes remaining after application of selection criteria is given in the right-hand column.

result, the number of transcripts preferentially expressed in AMA vs PRO was smaller (732) than the (1853) genes subsequently identified as differentially expressed (Chapter 3) and therefore, too, the pool of potential tagging candidates (Figure 4.1).

Intracellular parasitism is found amongst different kinetoplastid parasites, namely *Leishmania spp.* and *T.* cruzi. However, this lifestyle is thought to have arisen after the split of the *Leishmania* and *Trypanosoma* lineage (Fernandes, Nelson, and Beverley 1993). Therefore, factors specifically involved in host-parasite interaction in *Leishmania* either arose after the split of lineages and therefore are absent from e.g. *T. brucei*, or these proteins have been subject to expansion in the *Leishmania* lineage, leading to extensive diversification (e.g. delta-amastins (Jackson 2010)) and have an increased likelihood of no returned reciprocal best blast hits in sequence comparisons (Wall, Fraser, and Hirsh 2003). To identify proteins that play a specific role in *Leishmania*-biology, as opposed to kinetoplastid-wide biology, I therefore decided to exclude all proteins with reciprocal best-blast hits in *T. brucei* from the pool of tagging candidates, resulting in 395 remaining candidates (Figure 4.1).

The interface between parasite and host-cell can be envisaged in three ways: First, transmembrane or membrane associated proteins of both species may interact directly as seen for GP63 and complement receptor type 3 (Russell and Wright 1988). Second, proteins secreted by the parasite, via vesicles (J Maxwell Silverman et al. 2008; Judith Maxwell Silverman et al. 2010) may interact with host components of the parasitophorous vacuole or cytosol. Third, parasite factors secreted directly into the lumen of the parasitophorous vacuole (e.g. secreted acid phosphatase of *L. mexicana* (Ilg et al. 1991; Stierhof et al. 1994)) either as single molecules or other non-vesicular aggregates may interact with host-cell machinery (Martin et al. 1998; Holm et al. 2001). Parasite transmembrane proteins may be part of all three of these interactions, therefore I decided

to focus the screen on the 154 predicted transmembrane proteins in the pool of tagging candidates (Figure 4.1).

Amongst these were many genes with predicted or, based on presence of PFAM-domains predictable, functions. In an attempt to enrich for proteins that may be involved in yet uncharacterised *Leishmania* biology, I decided to exclude all proteases, glycosyl-transferases, metabolic enzymes and transporters, allowing for the exceptions of LmxM.04.0380 and LmxM.04.0400, which featured a PFAM domain (Grp1_Fun34_YaaH) implicating them in sensing of low pH, a key differentiation signal in *Leishmania spp.* Moreover, all proteins with a significant Amastin PFAM domain (by gathering threshold) were excluded, even though the precise function of Amastins is yet unknown. This decision was taken as the wide expansion of amastins in *Leishmania* genomes (> 70 amastins in *L. mexicana*, see Section 3.3.4.2) indicates a very important role of these which would need to be addressed in a study solely focussing on their biology and not in this exploratory screen. Following this selection, 64 candidates remained (Figure 4.1).

To prevent interference of the protein-tag with N-terminal signal peptides, all tagging was performed C-terminally, i.e. the fluorescent protein added to the C-terminus of the target protein. Due to the extensive expression of proteases on the surface of amastigotes (see Section 4.1), the low pH amastigotes are exposed to in the parasitophorous vacuole (Antoine et al. 1990) and the effects this may have on GFP (TN campbel 2001), it was decided to only tag intracellular moieties of proteins. As a result the transmembrane proteins were required to have an intracellular C-terminus as predicted TMHMM 2.0.

Finally, all genes with a 3' intergenic sequence larger than 6 kb were excluded (Figure 4.1). This was to make ultimately possible the cloning of the complete 3' UTR into the tagging vector to preserve elements that may influence RNA-levels and stage-regulation, albeit that stage of investigations was not reached in this study. The whole intergenic sequence

needed to be considered as this selection was performed prior to the availability of SLAS and PAS information.

The resulting list of 26 candidate genes and their differential expression established by RNA-seq (from both preliminary and final differential expression analyses) is summarised in Table 4.1. Where applicable, PFAM-domain predictions and annotated or predicted GO-terms are shown.

4.3.2 Tagging of candidate genes

Before embarking on the tagging screen, it was important to show whether or not the employed tagging vector can be used to fluorescently tag proteins in *Leishmania spp.* The employed vector is an adaptation of the pNUS vectors (Tetaud et al. 2002) called pLENT (Gluenz, unpublished) allowing endogenous tagging in *Leishmania spp.* (Figure 4.2) Importantly, C-terminal tagging in the endogenous locus using pLENT introduces a new 3'UTR for the tagged gene. The introduced 3'UTR is that of *Crithidia fasciculata* phopshoglycerate kinase A (PGKA). PGKA is absent from *Leishmania spp.* and *C. fasciculata* does not form amastigotes, therefore the *C. fasciculata* PGKA (CfPGKA) 3'UTR is unlikely to contain sequences conferring stage specific expression. Moreover, a UTR from *C. fasciculata* was chosen as it should be divergent enough minimise the possibility of recombination mediated by this sequence, yet sufficiently related to be biologically active in *Leishmania spp*. To enhance fluorescent signal from the fusion protein, enhanced GFP (eGFP) (Cormack, Valdivia, and Falkow 1996) was used as the fluorescent protein.

As a methodological positive control, I tagged the *L. mexicana* glucose transporter 2 (LmGT2, LmxM.36.6290), which was previously reported to localise solely to the pellicular membrane of *L. mexicana* and be excluded from the flagellar and flagellar pocket membrane (Tran et al. 2012).

Transfection of *L. mexicana* promastigotes with a pLENT vector targeting eGFP to the C-terminus of LmGT2 resulted in cells with a fluorescent signal highlighting the cell-body

Table 4.1 Candidate genes for tagging-screen, their differential expression and localisation Summary of the 26 candidate genes used in the tagging screen. For each gene physical characteristics and annotations from TriTrypDB are given, as well as results of differential expression analyses (both final (blue) and preliminary results (red) used in the intial selection of tagging candidates). Where applicable, localisations as determined in Section 4.3.1 are given.

Gene ID	Product Description	PFAM domains	Possible Localisation	PROvAMA Log_2 (Fold Change)	PROvAMA padj	PROvAXA Log_2 (Fold Change)	PROvAXA padj	Mean AMA FPKM	Mean AXA FPKM	Mean PRO FPKM	PROvAMA Log_2 (Fold Change) (prelim.)	PROvAMA padj (prelim.)	PROvAXA Log_2 (Fold Change) (prelim.)	PROvAXA padj (prelim.)	TMDs	Predicted Signal Peptide	Protein Length	Annotated GO Function	Annotated GO Process
LmxM.03.0380	unspecified product	Grp1_Fun34_YaaH (yeast acetate transporter, acid sensitivity?)	Endoplasmic Reticulum	-1.85	5.67E-10	-0.33	3.58E-01	54.4	34.7	22.8	-2.58	2.71E-11	-0.85	2.21E-02	6	No	372	null	null
LmxM.03.0400	unspecified product	Grp1_Fun34_YaaH (yeast acetate transporter, acid sensitivity?)	Endoplasmic Reticulum	-1.08	3.13E-03	-0.48	2.39E-01	126.6	164.6	86.8	-1.45	5.80E-03	-0.94	3.00E-01	6	No	294	null	null
LmxM.08_29.1200	hypothetical protein, conserved		Mitochondrium	-0.84	4.47E-05	-0.73	6.05E-06	14.1	27.6	14.4	-0.88	2.57E-02	-0.83	2.39E-02	4	Yes	598	null	null
LmxM.09.1330	hypothetical protein, conserved		Flagellar pocket	-1.42	5.02E-17	-0.84	6.17E-08	31.5	45.1	21.5	-1.53	1.48E-06	-0.83	1.53E-02	2	Yes	601	null	null
LmxM.15.0600	hypothetical protein, conserved		Diffuse / No Signal	-0.75	2.95E-04	-0.05	8.66E-01	14.4	18.8	16.1	-0.90	6.82E-03	-0.14	8.52E-01	1	No	1251	null	null
LmxM.16.0400	hypothetical protein, conserved	Pfam-B_18331	N/A	-0.79	1.28E-05	-0.44	1.50E-02	36.3	62.0	39.9	-1.49	5.48E-03	-0.73	4.91E-01	2	Yes	363	null	null
LmxM.16.0500	hypothetical protein, unknown function		Surface	-1.51	5.37E-10	-0.12	7.33E-01	699.5	524.5	418.6	-2.29	1.23E-09	-0.80	4.14E-02	4	Yes	216	null	null
LmxM.18.0240	hypothetical protein, conserved		Mitochondrium	-1.09	7.50E-05	-0.48	1.02E-01	12.8	17.3	10.3	-1.70	3.41E-02	-0.99	4.71E-01	1	Yes	404	null	null
LmxM.21.0070	hypothetical protein, conserved		N/A	-0.86	1.25E-04	-0.30	2.12E-01	49.1	70.5	49.4	-1.12	1.68E-03	-0.44	4.81E-01	1	null	285	null	null
LmxM.22.0410	hypothetical protein, conserved		N/A	-2.01	8.20E-19	-1.42	3.03E-10	16.6	26.2	7.5	-2.16	1.10E-02	-1.82	1.54E-08	2	Yes	1192	protein binding	null
LmxM.23.1267	hypothetical protein, unknown function		Golgi	-1.08	6.14E-05	-0.68	4.54E-03	21.4	33.2	17.2	-1.24	1.27E-03	-0.93	2.12E-02	1	Yes	389	null	null
LmxM.24.0552	hypothetical protein, conserved		Endoplasmic Reticulum	-1.73	2.68E-05	-0.79	4.03E-02	9.7	10.1	3.5	-3.81	2.46E-02	-1.87	7.31E-01	1	No	130	null	null
LmxM.25.0780	hypothetical protein, unknown function		Surface	-1.14	8.94E-08	-0.47	3.01E-02	15.8	21.2	13.2	-1.24	4.88E-03	-0.69	2.40E-01	1	No	306	null	null
LmxM.25.2240	hypothetical protein, conserved		N/A	-1.54	1.92E-06	-0.56	1.09E-01	4.6	4.7	2.6	-1.89	2.27E-04	-0.79	3.43E-01	1	No	559	null	null
LmxM.26.1460	hypothetical protein, unknown function		Golgi	-3.72	1.13E-78	-2.27	2.75E-23	79.0	71.9	10.0	-3.98	4.17E-37	-2.69	2.02E-12	2	No	333	null	null
LmxM.27.1930	hypothetical protein, unknown function		Golgi	-1.15	1.60E-06	-0.49	8.38E-02	11.3	16.1	9.7	-1.29	2.85E-03	-0.80	1.42E-01	4	No	354	null	null
LmxM.27.2310	hypothetical protein, conserved		Mitochondrium	-1.02	1.76E-05	-0.42	1.17E-01	14.3	20.6	13.2	-1.02	3.23E-02	-0.37	6.56E-01	4	No	247	null	null
LmxM.29.0290	hypothetical protein, unknown function		Diffuse / No Signal	-1.14	2.24E-09	-0.79	5.98E-05	6.0	9.9	4.9	-0.94	1.69E-02	-0.85	6.90E-02	1	Yes	1521	null	null
LmxM.29.1750	hypothetical protein, unknown function		Golgi	-1.56	5.40E-07	-0.44	2.51E-01	12.3	12.0	7.0	-1.91	1.84E-03	-0.71	5.53E-01	1	No	165	null	null
LmxM.30.2010	hypothetical protein, unknown function		Multi-vesicular tubule	-1.11	3.32E-06	-0.42	8.30E-02	24.4	30.7	19.7	-1.29	2.79E-03	-0.47	3.15E-01	1	No	606	null	null
LmxM.31.3400	hypothetical protein, conserved		Mitochondrium	-2.40	1.65E-66	-1.26	1.31E-11	57.6	59.5	20.2	-2.58	1.25E-17	-1.50	2.08E-07	2	Yes	602	null	null
LmxM.31.3420	hypothetical protein, conserved		Endoplasmic Reticulum	-0.79	4.59E-05	-0.09	6.37E-01	14.6	18.3	15.3	-1.76	9.97E-04	-0.19	9.18E-01	10	Yes	401	null	null
LmxM.31.3590	hypothetical protein, conserved		Undefined vesicular compartment	-1.40	1.95E-09	-0.55	3.23E-02	125.1	144.1	82.0	-1.89	7.28E-04	-0.92	4.14E-01	2	No	112	null	null
LmxM.33.3620	hypothetical protein, conserved	DUF3336, Patatin	Undefined vesicular compartment	-1.39	9.69E-08	-0.50	5.51E-02	11.2	12.9	7.8	-1.48	9.96E-05	-0.54	3.41E-01	2	No	619	GTP binding	lipid metabolic process
LmxM.34.2590	hypothetical protein, conserved	zf-DHHC	Diffuse / No Signal	-1.06	1.17E-06	-0.56	9.40E-03	16.7	24.5	14.2	-1.19	6.29E-03	-0.62	3.03E-01	4	No	280	zinc ion binding	null
LmxM.36.3860	similar to *Leishmania major* . l411.4-like protein	HAP2-GCS1	Diffuse / No Signal	-1.57	4.94E-10	-0.98	3.13E-06	8.8	11.7	4.9	-2.04	1.91E-04	-1.24	1.97E-02	1	No	293	null	null

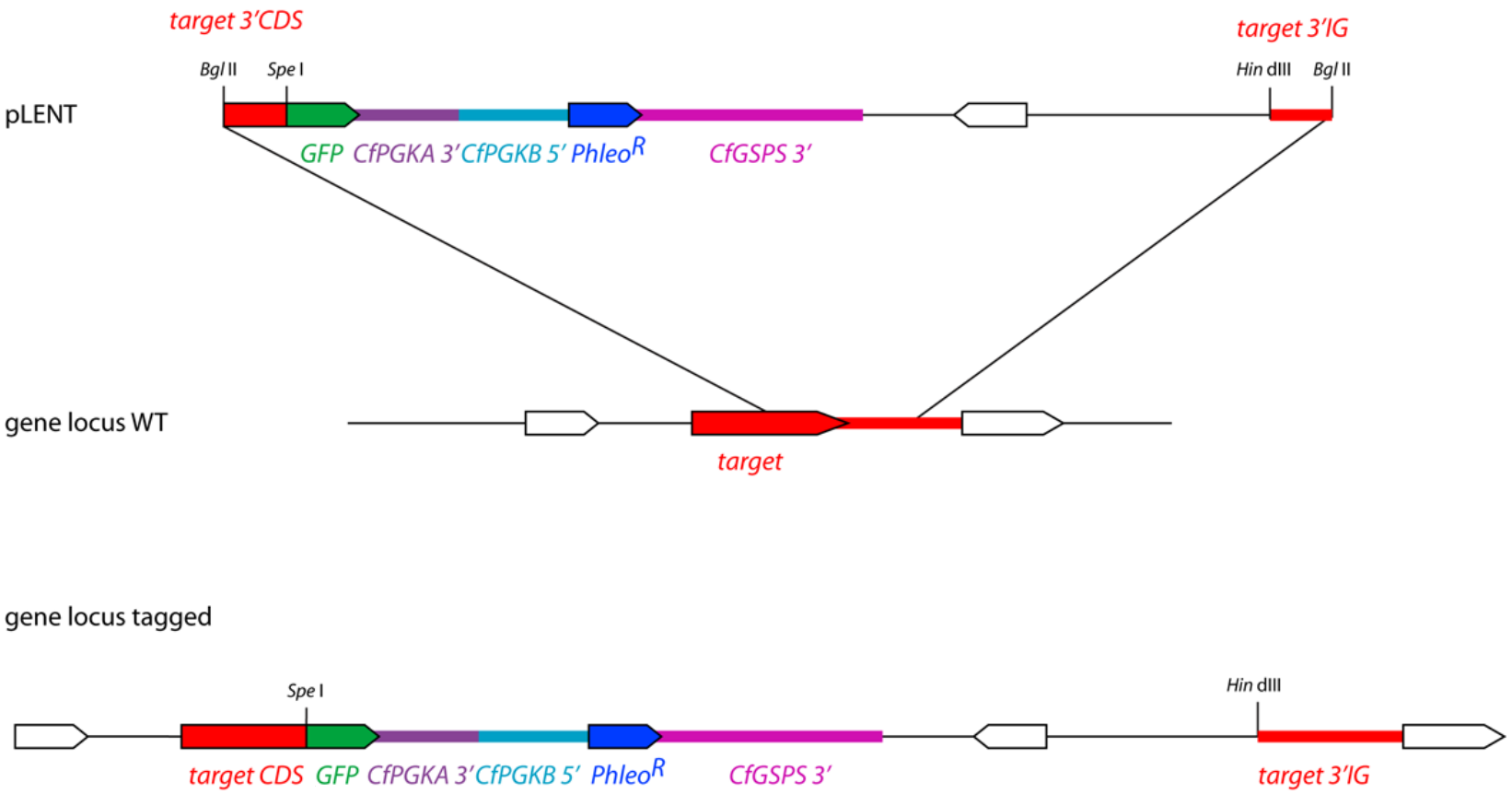


Figure 4.2 pLENT tagging strategy Schematic showing the basic structure of the pLENT tagging vector, its relationship to the target locus and the resulting recombinant genomic locus. Abbreviations: CDS= Coding Sequence, GFP= Green Fluorescent Protein, CfPGKA = *Crithidia fasciculata* phosphoglycerate kinase A, CfPGKB = *Crithidia fasciculata* phosphoglycerate kinase B, PhleoR= Phleomycin resistance gene, CfGSPS = *Crithidia fasciculata* glutathionylspermidine synthetase, 3' and 5' indicate 3' and 5' untranslated regions, respectively.

with a clear "edge-effect", suggesting that a protein on the pellicular membrane has been tagged (Figure 4.3). No fluorescent signal emanating from either the flagellum or highlighting the flagellar pocket was detectable. These observations are consistent with the successful tagging of LmGT2. Moreover, fluorescent signal could also be seen in AXA (Figure 4.4), providing proof that the pLENT-vector can be used to tag membrane proteins in *L. mexicana* promastigotes that are subsequently expressed in both PRO and AXA.

With this proof of principle in hand, I sought to tag the chosen 26 candidate genes using pLENT vectors in promastigote-form *L. mexicana* and analyse the localisation of the fusion protein in PRO and AXA. For four cell lines, no drug-resistant population were obtained and due to time-constraints, only analyses of 22 cell lines are presented. The observed localisations were grouped and representative images of PRO and AXA cells shown in Figures 4.5-10 and a summary is given in Table 4.1

Four fusion proteins, appeared to localise to a reticulated structure forming a ring around the nucleus in AXA (Figure 4.5 A & B), PRO (Figure 4.5 C), or both AXA and PRO (Figure 4.5 C). This is suggestive of a localisation to the endoplasmic reticulum, albeit further experiments would be required to prove this unequivocally (see Discussion). Four proteins localised to a tubular structure in PRO (Figure 4.6 A & B) or showed a heterogeneous signal throughout the cytoplasm in AXA without forming a distinct ring around the nucleus (Figure 4.6 C & D). It is suspected these proteins localise to the mitochondrion, but again, further experiments would be required to fully establish this finding (see Discussion).

Four cell lines expressed a fusion protein that may localise to the Golgi apparatus, with one cell line showing a punctate localisation anterior of the nucleus in AXA (Figure 4.7 A) and one showing a punctate localisation close to the nucleus in AXA (Figure 4.7 B). Moreover, one cell line showed a punctate localisation between nucleus and kinetoplast in PRO (Figure 4.7 C), and finally one cell line showing punctate localisations posterior to and

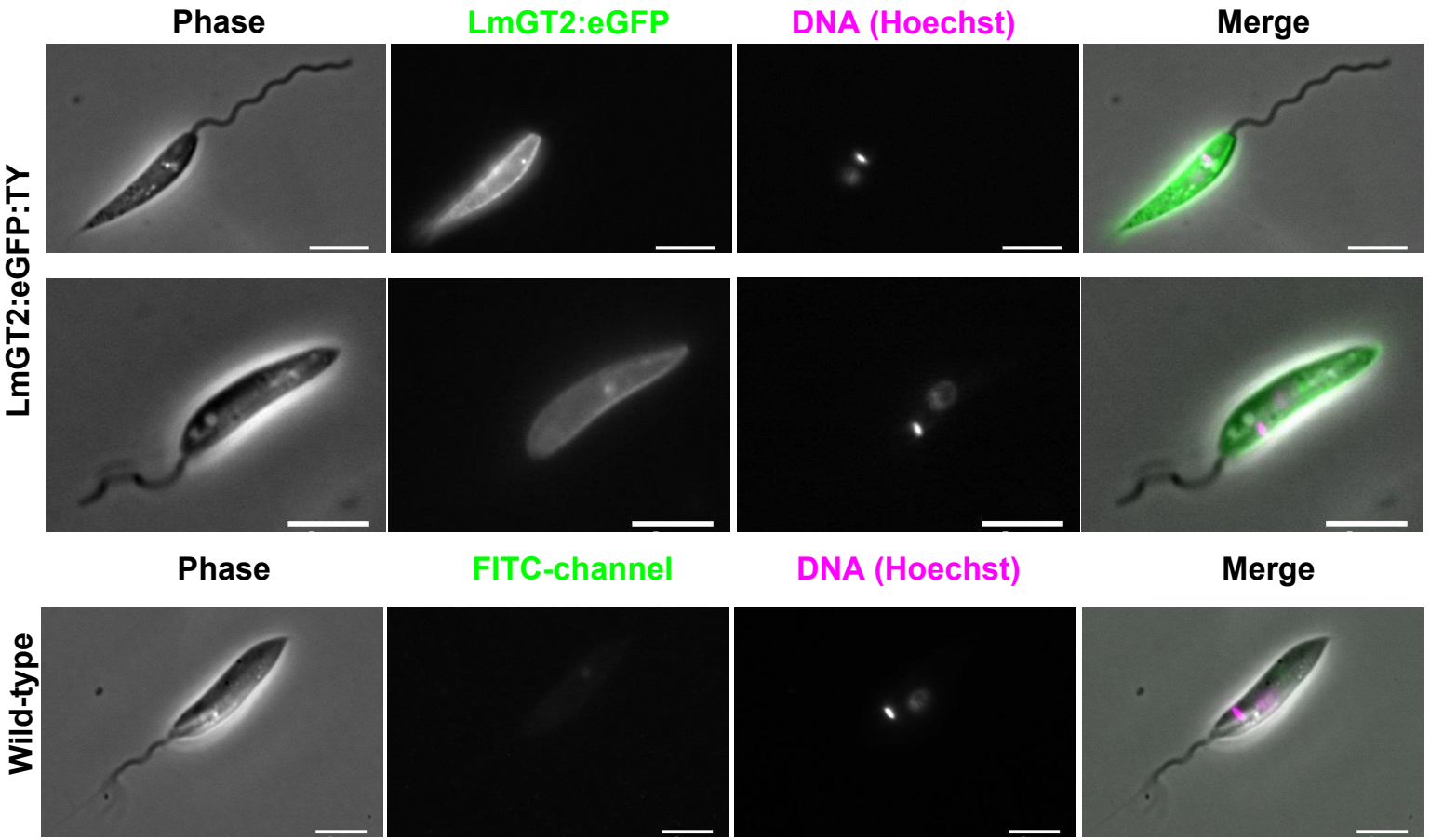


Figure 4.3 eGFP tagging of LmGT2 Live-cell light microscopic analysis of promastigotes expressing the LmGT2:eGFP:TY fusion protein at 100x magnification. For comparison, images of wild-type cells acquired and processed using same parameters to match tagged cells are shown. Scale bar is 5 µm.

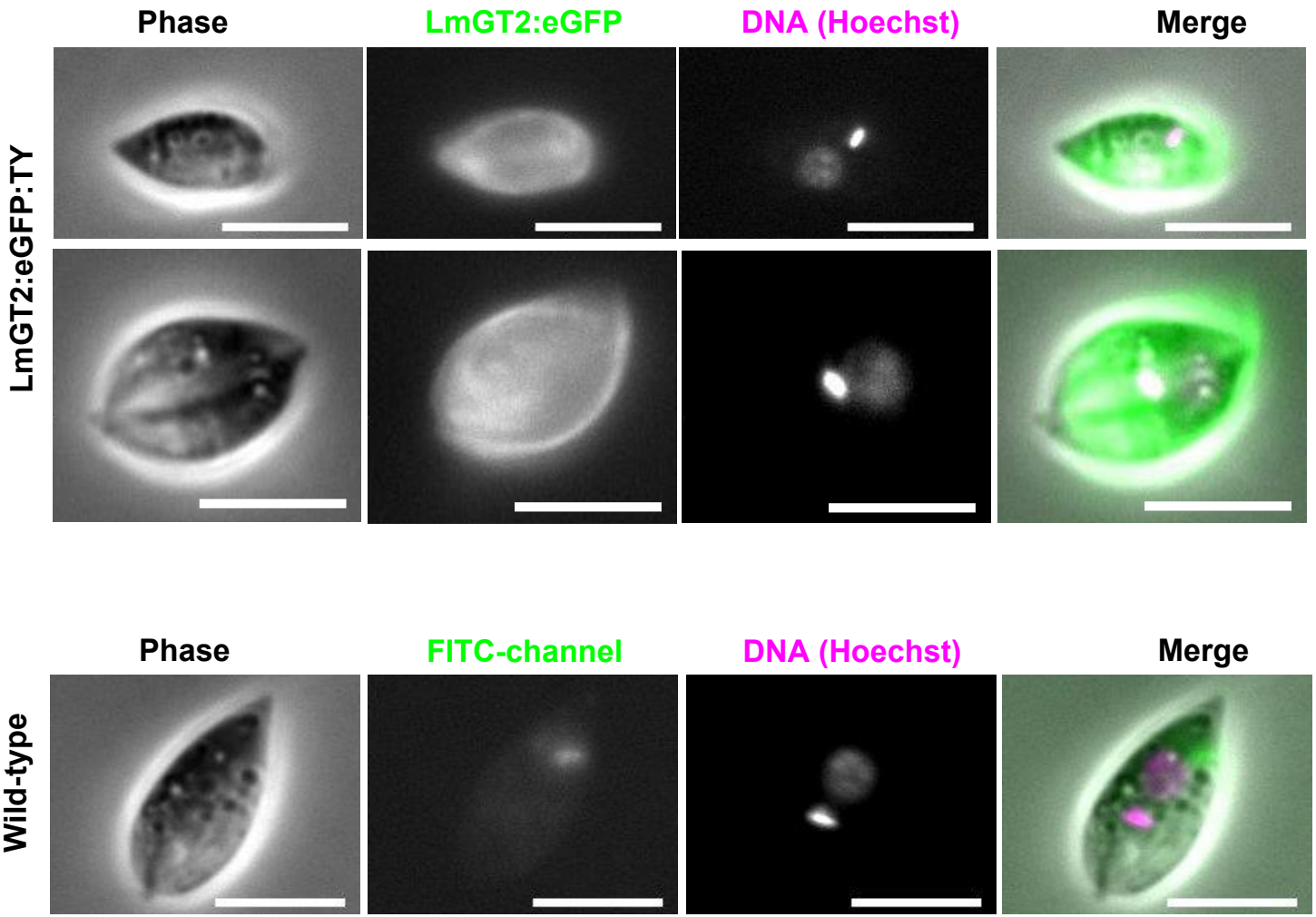


Figure 4.4 eGFP tagging of LmGT2 Live-cell light microscopic analysis of 24 h axenic amastigotes expressing the LmGT2:eGFP:TY fusion protein at 100x magnification. For comparison, images of wild-type cells acquired and processed using same parameters to match tagged cell lines are shown. Scale bar is 5 µm.

on the same height as the kinetoplast in PRO and AXA (Figure 4.7 D). However one has to entertain the possibility that the localisations presented in Figure 4.7 could also correspond to either the endosomal system or other multi-vesicular structures. In Figure 4.8 A an extended localisation posterior to the nucleus is seen in PRO, whilst a signal anterior to the nucleus is seen in AXA. This could correspond to a lysosomal structure. The signal observed in Figure 4.8 B shows a distributed punctate localisation, the nature of which is unclear, whilst in Figure 4.8 C a localisation right next to the kinetoplast is detected, which could correspond to a localisation to the endosomal system, Golgi or another vesicular body. Again, one has to bear in mind that without distinct markers of these various compartments allowing co-localisation studies, the proposed localisations are fairly speculative.

Three fusion proteins localised to the cell surface, i.e. the pellicular membrane in PRO and AXA (Figure 4.9 A), the pellicular membrane and what appear to be structures released from the cell in PRO (Figure 4.9 B) and the flagellar pocket, strictly speaking a cell-surface membrane, in AXA (Figure 4.9 C).

Four cell lines showed no or too diffuse fluorescence signal to allow determination of localisation of the fusion protein (Figure 4.10), raising the possibility that the fusion protein was not expressed.

In this screen, two proteins localising to the cell surface particularly stood out and were more closely examined and characterised: LmxM.16.0500, which showed a surface localisation in PRO, with considerable material appearing to localise to structures outside of the cell body, and LmxM.09.1330, which appeared to show a flagellar-pocket localisation in AXA.

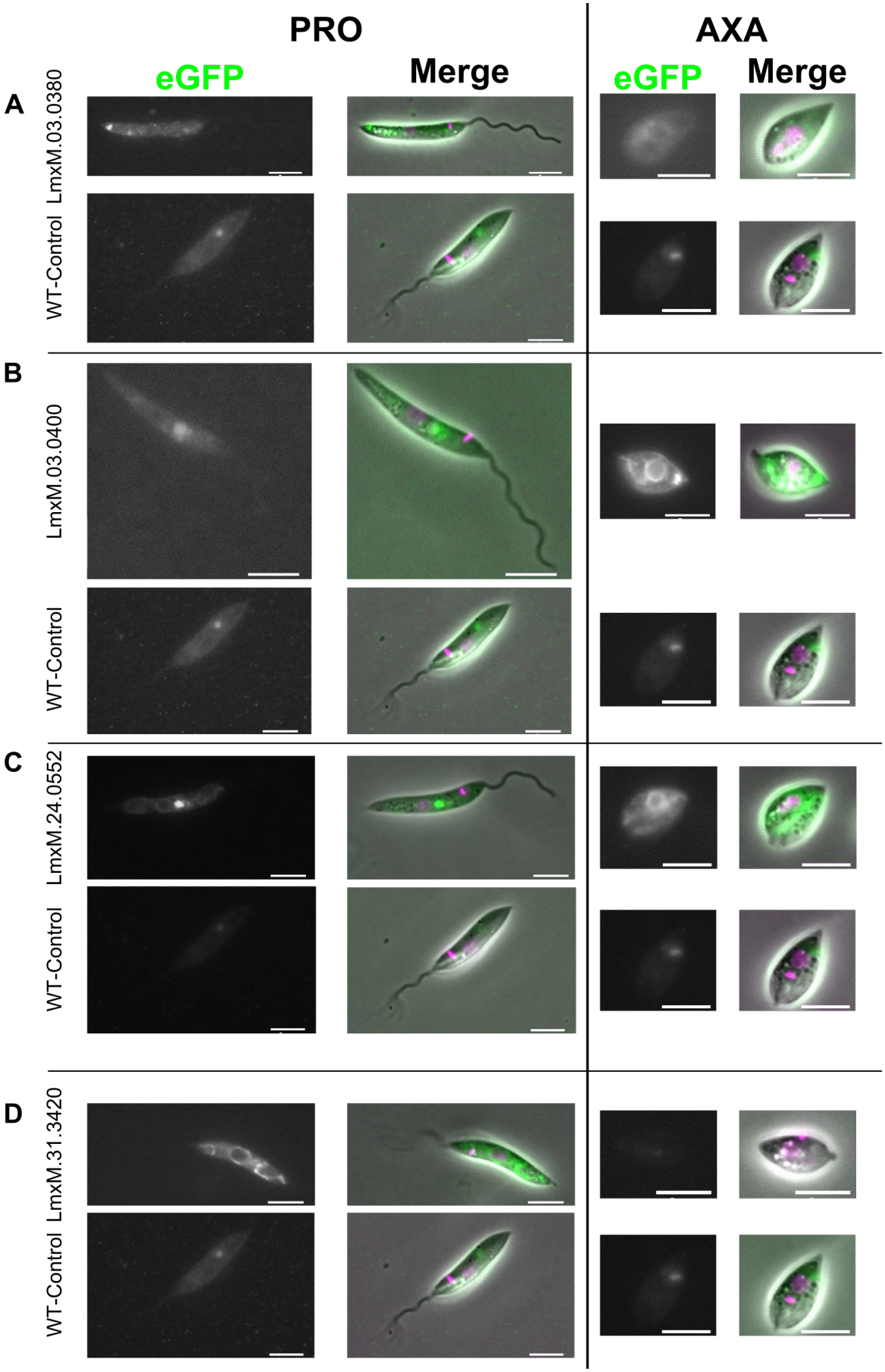


Figure 4.5 Cell lines expressing fusion protein possibly localising to the endoplasmic reticulum Summary of cell lines expressing a pLENT tagged fusion protein most likely localising to the endoplasmic reticulum in AXA (**A**, **B**,**C**) or in PRO (**D**). Representative images are shown for each cell line along with images of untagged wild-type (WT) control acquired and processed using the same parameters to match tagged cell line. Scale bar is 5 μm.

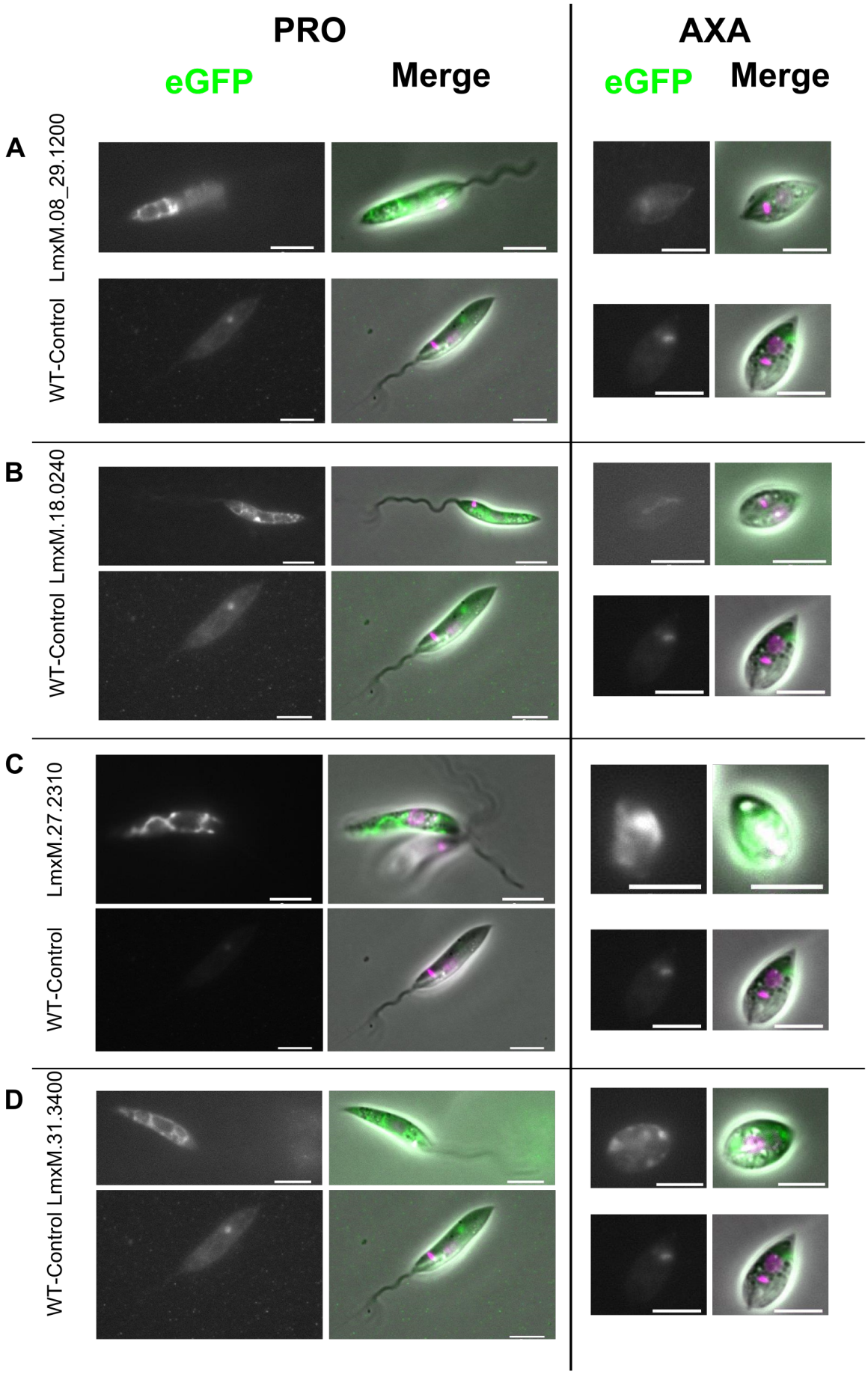


Figure 4.6 Cell lines expressing fusion protein possibly localising the mitochdrion Summary of cell lines expressing a pLENT tagged fusion protein most likely possibly localising to the mitochondrion in PRO (A & B) or PRO and AXA (C &D). Representative images are shown for each cell line along with images of untagged wild-type (WT) control acquired and processed using the same parameters to match tagged cell line. Scale bar is 5 μm.

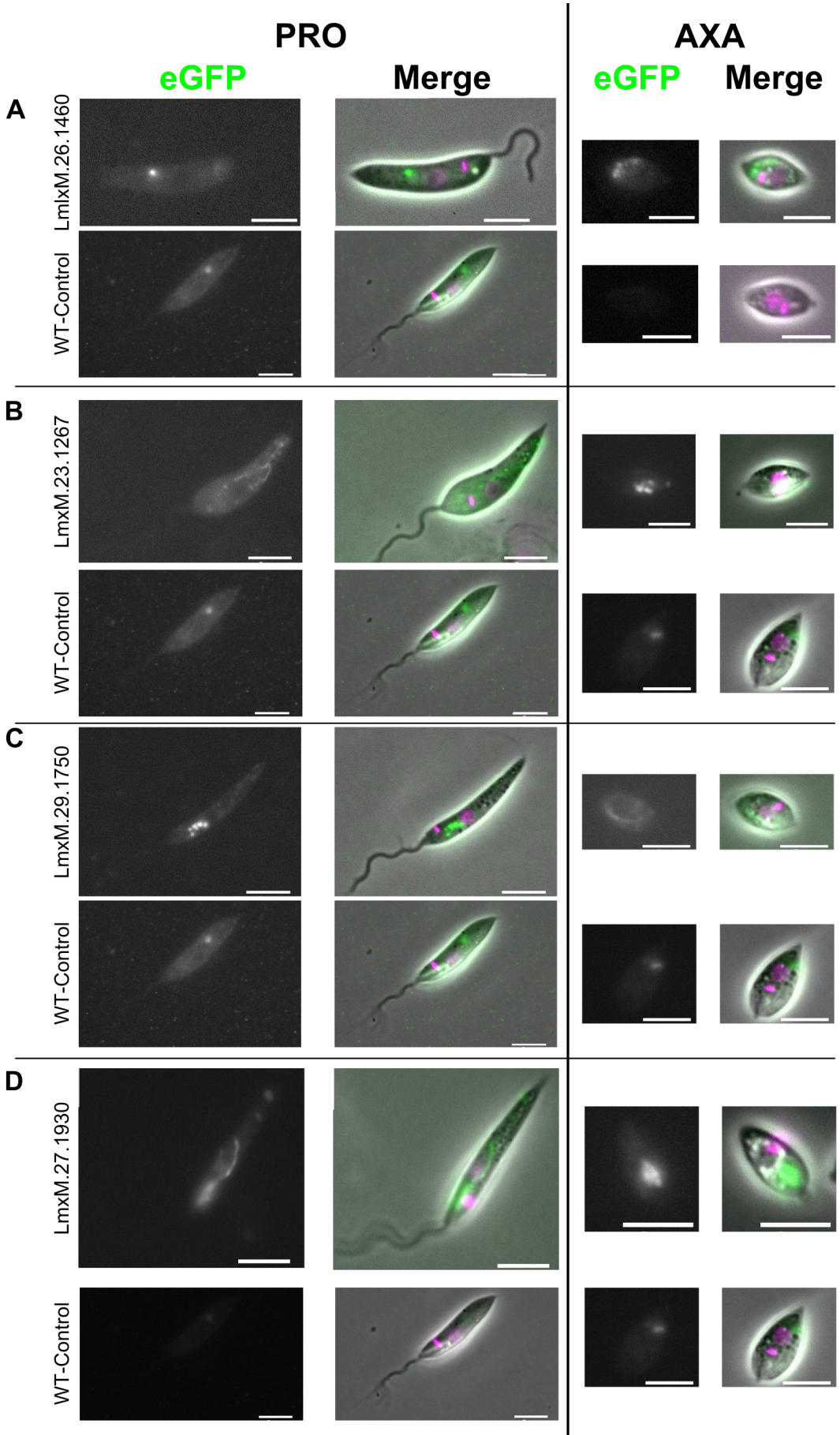


Figure 4.7 Cell lines expressing fusion protein possibly localising to the Golgi apparatus
Summary of cell lines expressing a pLENT tagged fusion protein possibly localising to the Golgi in AXA (A & B), in PRO (C) or in both PRO and AXA (D). Representative images are shown for each cell line along with images of untagged wild-type (WT) control acquired and processed using the same parameters to match tagged cell line. Scale bar is 5 μm.

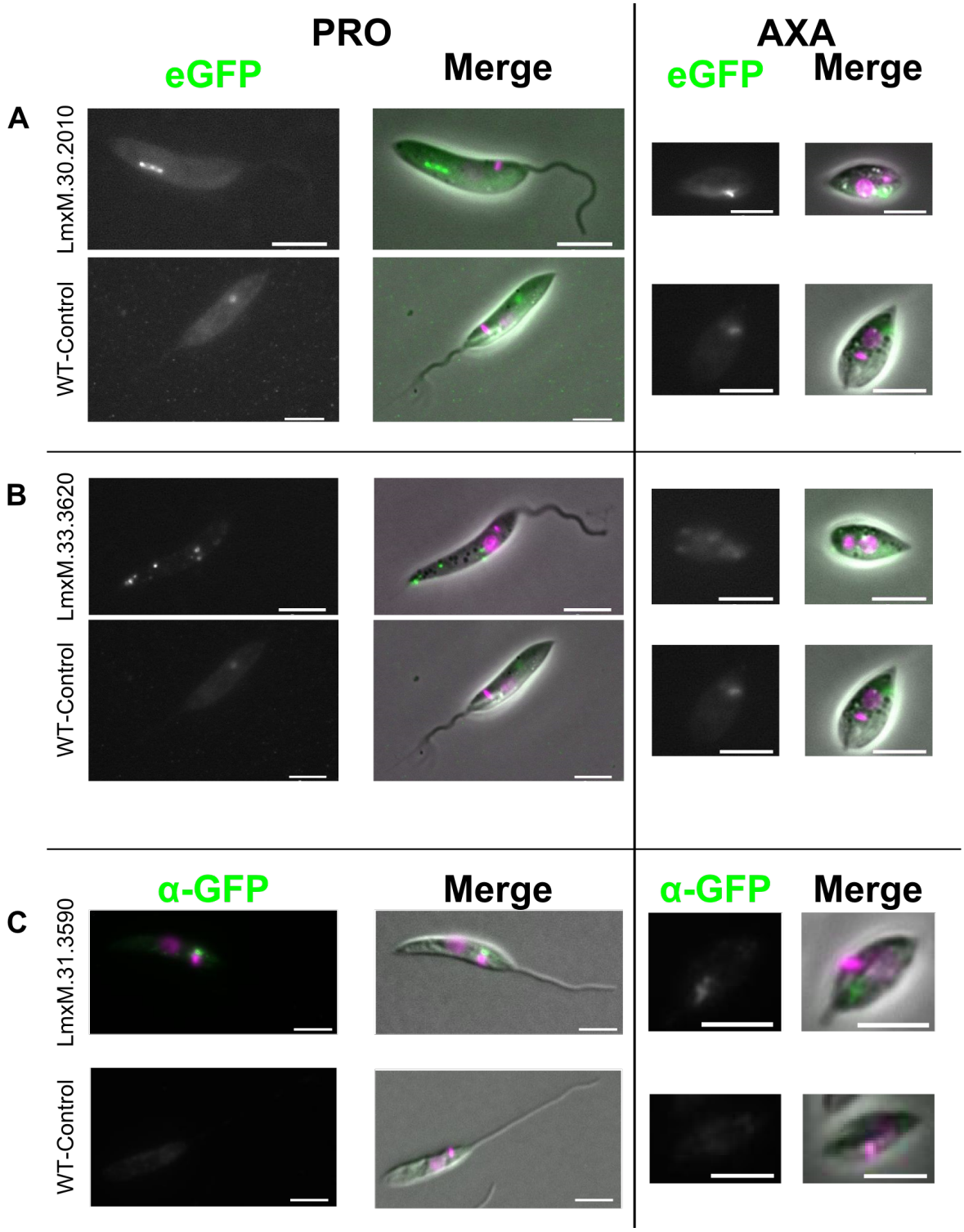


Figure 4.8 Cell lines expressing fusion protein possibly localising to the multi-vesicular tubule or another vesicular compartment Summary of cell lines from pLENT-tagging screen expressing a fusion protein possibly localising to the multi-vesicular tubule PRO and AXA (A) and localising to another, less defined vesicular compartment (B & C). Representative 100x live-cell microscopic images are shown for each cell line along with images of untagged wild-type (WT) control acquired and processed using the same parameters to match tagged cell line, with the exception of LmxM.31.3590 where immunofluorescence microscopic images are shown. Scale bar is 5 μm.

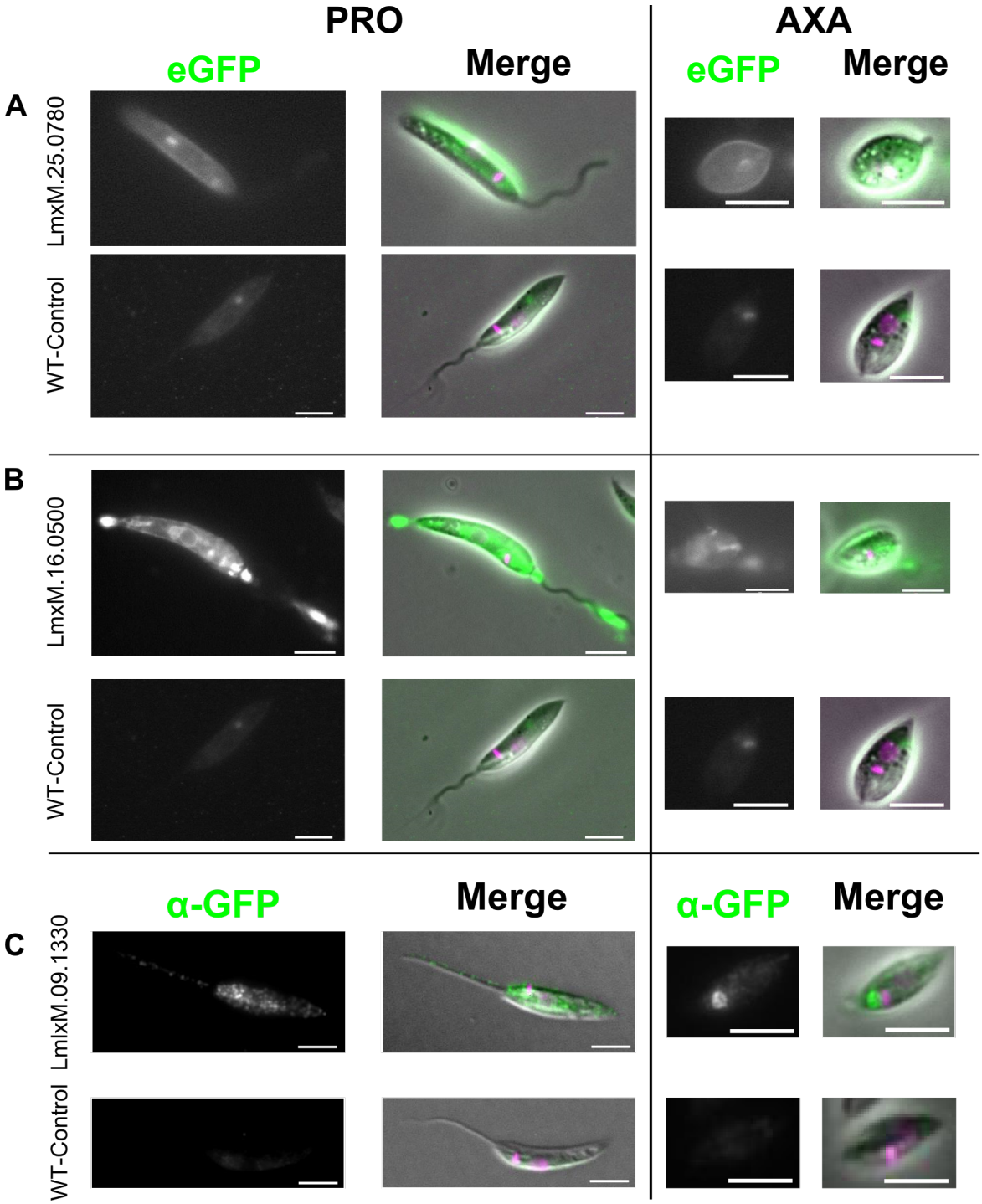


Figure 4.9 Cell lines expressing fusion protein with cell surface or flagellar pocket localisation
Summary of cell lines from pLENT-tagging screen expressing a fusion protein with putative cell surface localisation in PRO and AXA (A), surface localisation with what appears to be shedding of material containing the fusion protein from the cell (B) or localisation to the flagellar in AXA (C). Representative 100x live-cell microscopic images are shown for each cell line along with images of untagged wild-type (WT) control acquired and processed using the same parameters, to match tagged cell lines, with the exception of LmxM.09.1330 where immunofluorescence microscopic images are shown. Scale bar is 5 μm.

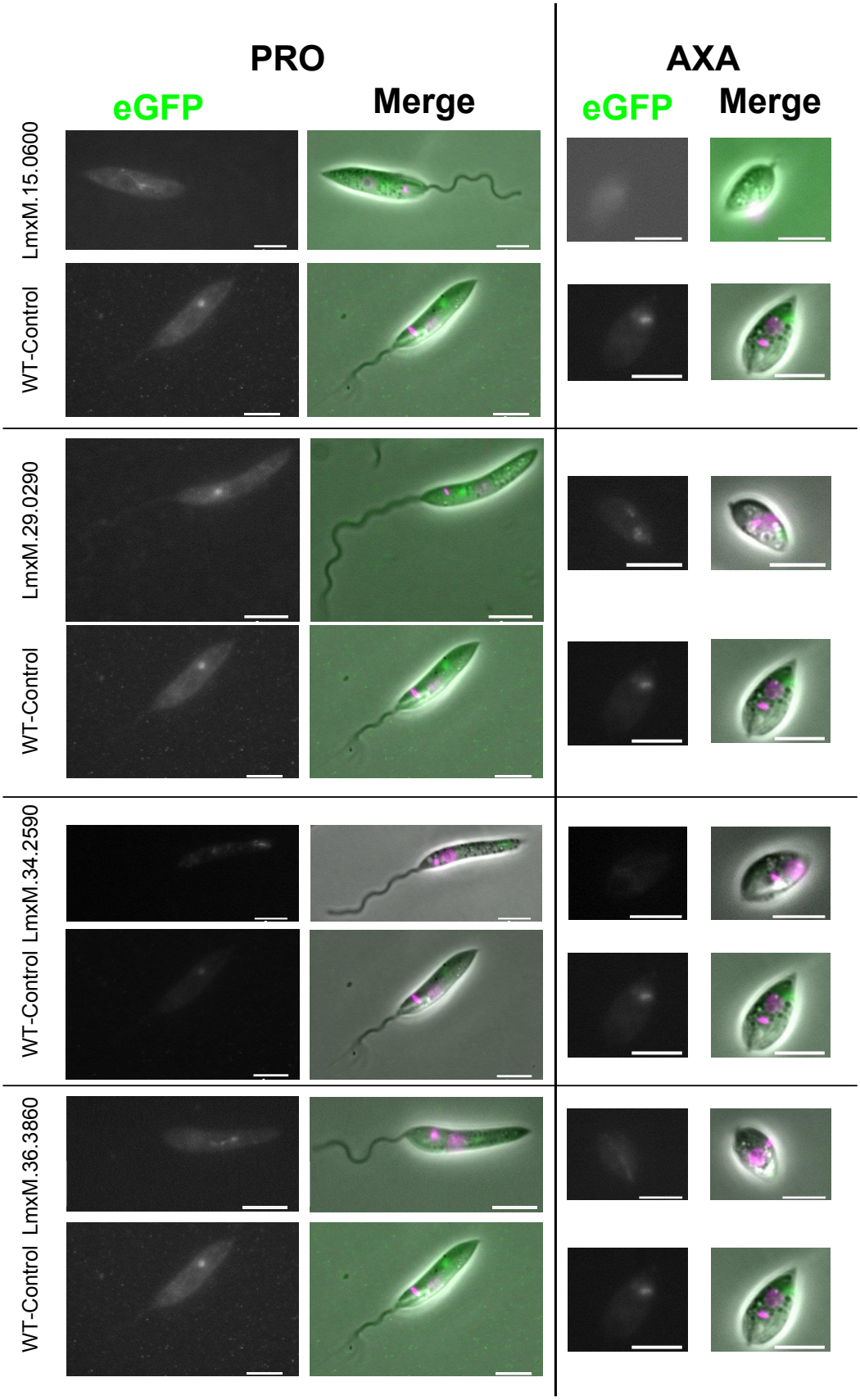


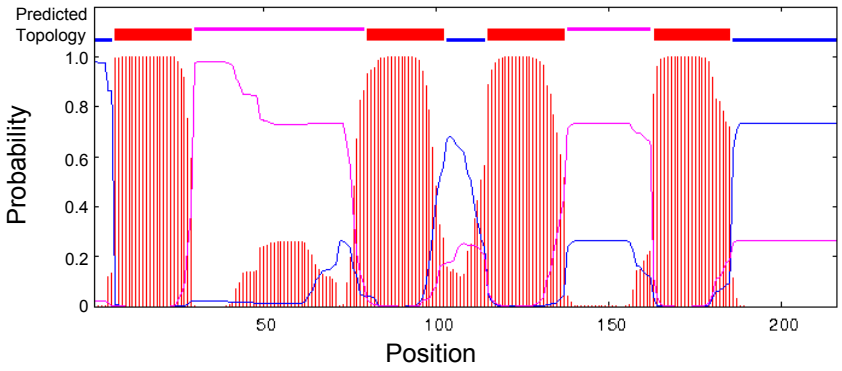
Figure 4.10 Cell lines expressing fusion protein with diffuse or no clear localisation Summary of cell lines from pLENT-tagging screen expressing a fusion protein with punctate localisation in PRO or AXA. Representative 100x live-cell microscopic images are shown for each cell line along with images of untagged wild-type (WT) controls acquired and processed using the same parameters to match tagged cell lines. No fluorescent signal corresponding to the fusion protein is apparent, either indicating absence of the fusion protein or a localisation too diffuse to be able detectable using this method.

4.3.3 LmxM.16.0500 is a highly expressed cell surface protein that is extensively shed from the cell

LmxM.16.0500 is a protein with 4 predicted TMDs and a signal peptide. Its predicted topology is shown in Figure 4.11. No significant PFAM domains (by gathering threshold) are detected for the protein. LmxM.16.0500 is syntenically conserved between *Leishmania spp.* except for in *L. tarentolae* (Figure 4.12), where a similar gene is found, which however is an orthologue of LmxM.16.0490, which will be discussed later. An alignment of the protein sequences of the *Leishmania* homologues is shown in Figure 4.13. Whilst being well conserved between species, the C-terminal region of the protein shows divergence between *Leishmania* species. No similar proteins are found outside kinetoplastids. A search for paralogous sequences of LmxM.16.0500 in *L. mexicana* returned two proteins, the most similar, LmxM.16.0490 (BlastP e-value: $5x10^{-36}$), at an adjacent locus, and the next, LmxM.16.0470 (BlastP e-value: $9x10^{-17}$), separated from the first two by a fucose kinase gene (LmxM.16.0480) (Figure 4.4). An alignment of the protein sequences of LmxM.16.0500, 490 and 470 is shown in Figure 4.14. The proteins are well conserved throughout and most divergence is found at the C-terminus. The most similar sequences retrieved by BlastP bear considerably less sequence similarity, with e-values of 4.4 and 9.3 for LmxM.10.1320 (fatty acid desaturase) and LmxM.29.0870 (amastin-like surface protein-like), respectively, indicating that the sequences of LmxM.16.0500, 490 and 470 stand apart from other protein sequences in the genome.

The fact that between orthologues and paralogues of LmxM.16.0500 most of the protein sequence divergence is located at the C-terminus may play an important role in the function and localisation of the protein.

LmxM.16.0500, 490 and 470 all have a predicted “Prokaryotic membrane lipoprotein lipid attachment site profile”(based on PROSITE data-base (Bairoch 1991)), a pattern involved in diglyceride attachment to proteins in Gram-positive bacteria (Sutcliffe and Harrington 2002). The e-values for these predictions however are fairly poor (e=5, e=5 & e=6,



Colour code: Transmembrane, Intracelllular, Extracellular

Figure 4.11 The topology of LmxM.16.0500 Diagram showing the predicted topology of LmxM.16.0500. A colour-code indicates the transmembrane domains as well as intracellular and extracellular loops. Predictions performed using TMHMM Server 2.0.

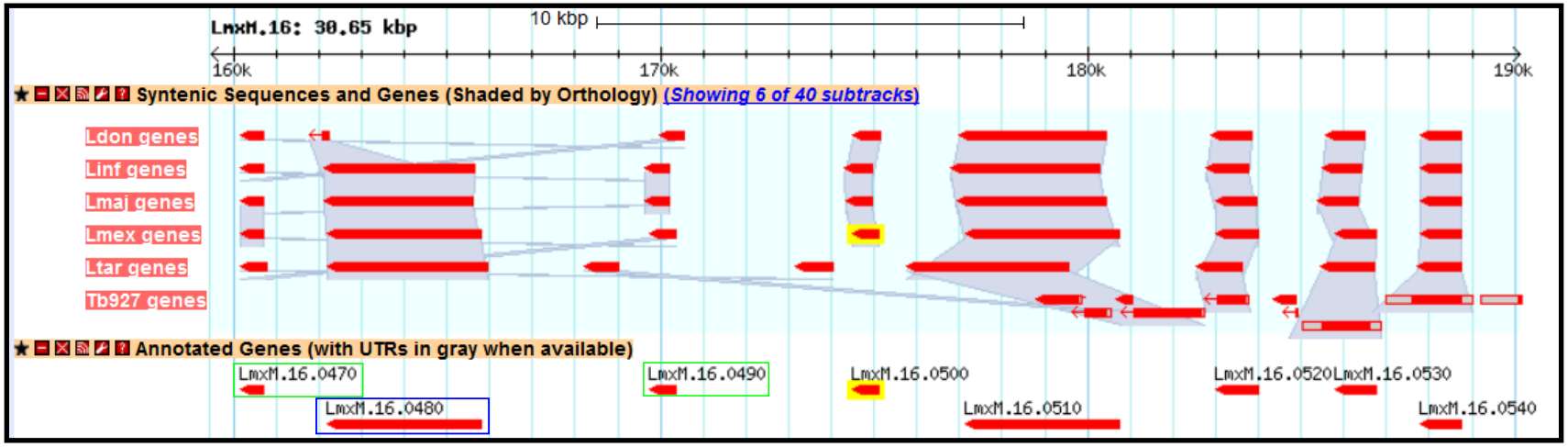


Figure 4.12 The composition and synteny of the locus surrounding LmxM.16.0500 Diagram of the genomic locus surrounding LmxM.16.0500 (highlighted in yellow). The two proteins with highest ranking BlastP hits to LmxM.16.0500 are marked in green, fucose kinase in blue. Synteny between species is highlighted in gray. Diagram adapted from Genome Browser view on TriTrypDB. Ldon= *L. donovani*, Linf= *L. infantum*, Lmaj =*L. major*, Lmex= *L. mexicana*, Ltar = *L. tarentolae*,Tb927 = *T. brucei 927*

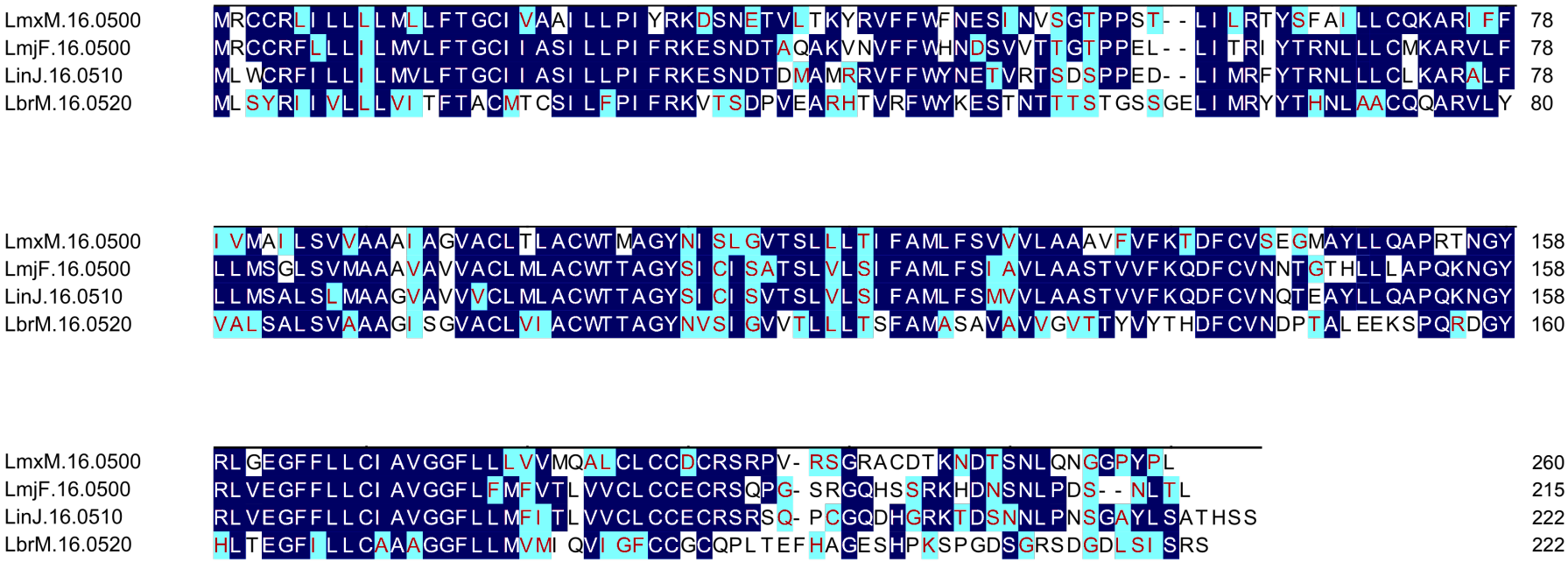


Figure 4.13 Alignment of protein sequences of LmxM.16.0500 orthologues from other *Leishmania spp.* Sequence alignments of LmxM.16.0500 homologues from *L. braziliensis* (LbrM.), *L. infantum* (LinJ.), *L. major* (LmjF.) and *L. mexicana* (LmxM.). Alignments created using Clustal W. Dark blue indicates conserved amino acid, light blue indicates conservation of a functional group.

Figure 4.14 Alignment of putatively expanded gene family comprising LmxM.16.0500 Protein sequence alignments of LmxM.16.0500, LmxM.16.0490 and LmxM.16.0470 which may form an amastin-like gene family expanded in *Leishmania spp.* compared to *T. brucei*. Alignment generated using Clustal W. Dark blue indicates conserved amino acid, light blue indicates conservation of a functional group.

respectively), suggesting that these are non-significant predictions. Like LmxM.16.0500, LmxM.16.0490 has no PFAM-domain, however LmxM.16.0470 has an amastin-domain.

LmxM.16.0500 has no *T. brucei* orthologue (by annotation and R.B.B.). LmxM.16.0470 and LmxM.16.0490 are both annotated to have the *T. brucei* orthologue Tb927.4.3520, which is an R.B.B. of LmxM.16.0490. Tb927.4.3520 has a predicted amastin-domain and RNA interference mediated ablation of the expression of this protein is lethal for *T. brucei* (Alsford et al. 2011). Indeed, even amongst the orthologues of LmxM.16.0500, some members have an amastin-domain (in *L. infantum, L. braziliensis*), whilst others do not (*L. major, L. donovani*), indicating that LmxM.16.0500 does bear a relationship to amastins.

Analysis of the synteny-relationship between *Leishmania spp.* and *T. brucei* at this locus shows that the "LmxM.16.0500-family" is exquisitely situated at a breakpoint of synteny (Figure 4.12). The fucose kinase gene (LmxM.16.0480), which is situated between LmxM.16.0490 and 0470 is absent from *T. brucei* (by R.B.B. and annotation). This may suggest loss of chromosomal segment in *T. brucei.*

The distribution of paralogues of LmxM.16.0500 in *Leishmania* and complicated evolutionary relationship of these paralogues with an essential *T. brucei* gene, with one being the R.B.B. and the other sharing an amastin PFAM domain, may suggest an expansion event in the *Leishmania* lineage, through duplication and re-purposing of proteins facilitated by resulting relaxed evolutionary constraints (Jackson 2007). Members of the "LmxM.16.0500-family" have an evolutionary link to amastins. Jackson (Jackson 2010) mapped three delta-amastins to chromosome 16, and whilst no accession numbers are given in that study, I could not identify any other amastins on chromosome 16 (by PFAM domain search), leading me to believe that the "LmxM.16.0500-family" had been identified in Jackson's study, thereby rooting this family amongst the δ-amastins. Unfortunately, these three amastins were not explored in close detail in that study (Jackson 2010), leaving questions regarding their (dis-)similarity to other amastins open.

Rochette (Rochette et al. 2005) did not identify any amastins on Chromosome 16 of *L. major* and *L. infantum.* Bearing all of the analyses in mind, the "LmxM.16.0500-family" constitutes an amastin-derived, but divergent gene family.

As shown in Table 4.1, LmxM.16.0500 is differentially expressed, with 2.8 fold higher expression in AMA compared to PRO ($p=5.37x10^{-10}$). Its absolute expression is high with 699.5 FPKM in AMA, placing it in the top percentile of highest expressed genes in AMA (c.f. Section 3.3.4.2). The stage-regulation observed in *L. mexicana* is consistent with microarray data for *L. major, L. infantum* and *L. donovani*: The *L. major* orthologue (LmjF.16.0500) showed higher expression in AMA compared to PRO as well as higher expression in metacyclics compared to PRO, with higher fold change in the comparison of metacyclics to PRO than AMA to PRO (Akopyants et al. 2010) (Figure 4.15 A). The *L. infantum* orthologue (LinF.16.0510) was shown to have higher expression in AMA and AXA compared to PRO (Rochette et al. 2009) with a higher fold-change in AXA vs. PRO than in AMA vs. PRO (Figure 4.15 B). Expression of the transcript was shown to increase during the timecourse of axenic differentiation of *L. donovani* (albeit, mapped to *L. infantum* genome, therefore the corresponding gene-accession is LinJ.16.0510) (Lahav et al. 2011) (Figure 4. 15 C).

Western blot analysis of whole-cell lysates of PRO and AXA expressing the LmxM.16.0500:eGFP fusion protein (see Materials and Methods) showed a doublet band between 45 and 50 kDa (Figure 4.16 A), consistent with the expected size of the fusion protein of 50.5 kDa (LmxM.16.0500 = 23.6 kDa, eGFP= 26.9 kDa). The doublet may indicate the presence of post-translational modifications. Expression levels of the fusion-protein between PRO and AXA were not markedly different and may either be a function of slightly different loading as assessed by Ponceau-staining of the blot membrane (Figure 4. 16 B), or suggesting that the exogenous CfPGKA 3'UTR has abolished stage-specific regulation. Alternatively, translational regulation may preclude observation of changes in

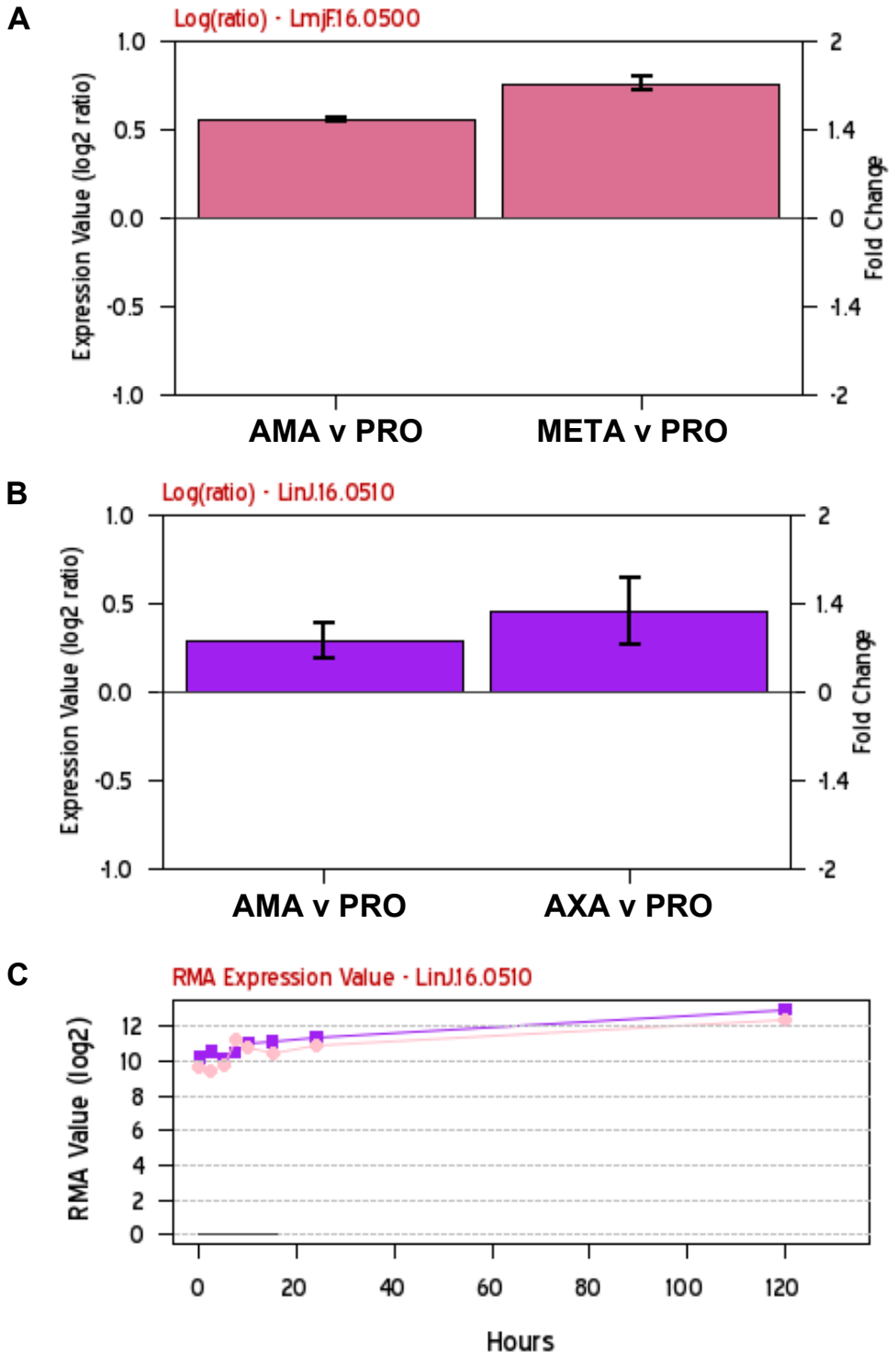


Figure 4.15 Expression summaries for orthologues of LmxM.16.0500 Graphs summarising differential expression analyses of *L. major* (A), *L. infantum* (B) and *L. donovani* (mapped to *L. infantum* genome) (C) orthologues of LmxM.16.0500. (C) is a time-course of axenic amastigote differentiation. The diagrams were taken from TriTrypDB. RMA = Robust Multi-Array Average.

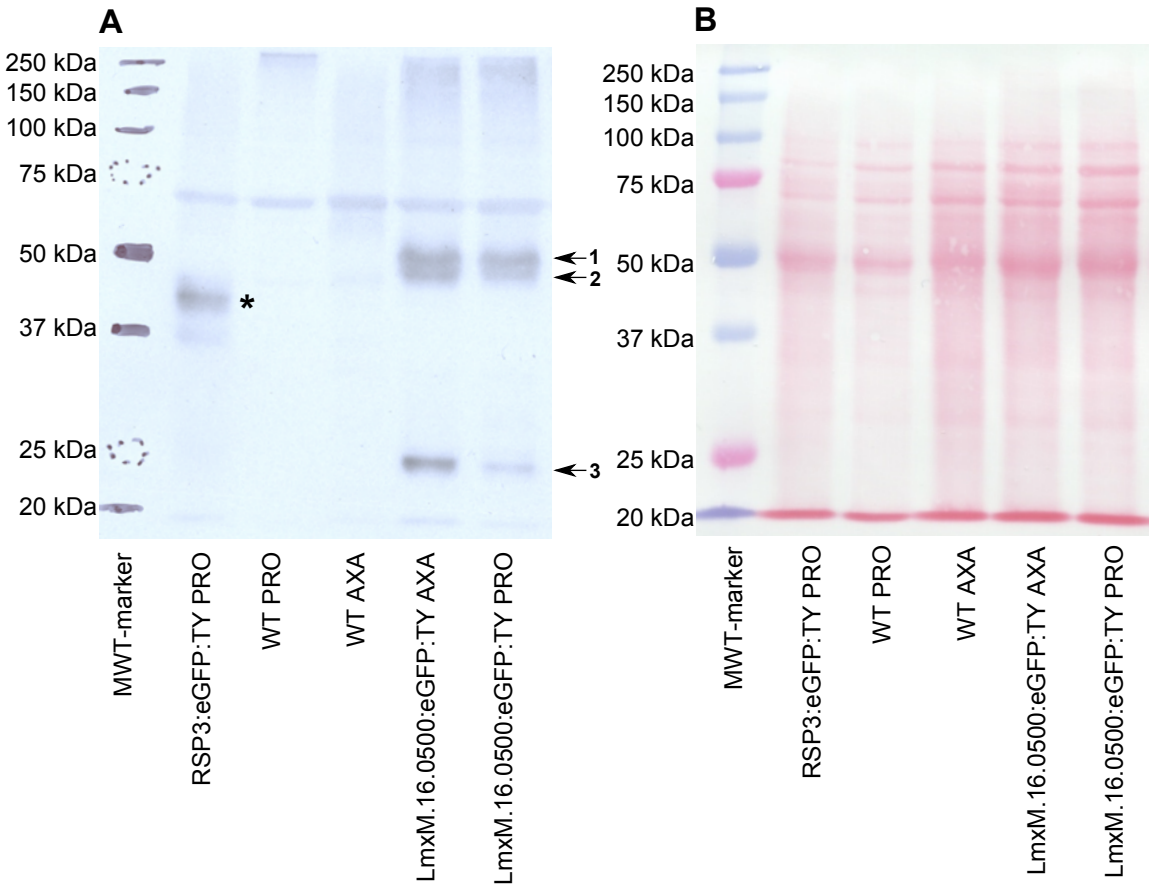


Figure 4.16 Western blot analysis of LmxM.16.0500:eGFP:TY expression Scan of (A) a membrane probed with anti-GFP antibody. A pLent-tagged cell line expressing RSP3:eGFP (Radial Spoke Protein 3, LmxM.27.0520) was used as positive control (*, expected: 69.0 kDa, reason for smaller observed band-size unclear, but sufficient for use as positive control for blotting and staining procedure) and untagged wild type (WT) PRO and AXA used as negative controls. Per lane, whole-cell lysates of ~5x10^6 used for all samples and separated on 10%-SDS-PAGE gel. Expected size of LmxM.16.0500:eGFP is 50.5 kDa, to which bands 1 & 2 may correspond. Band 3 may constitute a degradation product. Ponceau stain of the same membrane is shown in (B).

protein expression correlating with differential mRNA abundance. An additional strong band at 25 kDa was observed in PRO and AXA expressing LmxM.16.0500:eGFP. The significance of this band is not clear, but may represent a degradation product of the fusion protein.

The LmxM.16.0500:eGFP fusion protein localisation in PRO, as determined by live-cell light microscopy, is shown in Figures 4.17 and 4.18. At 40x magnification (Fig. 4.17) fluorescent signal can be seen in all cells, albeit an irregular signal. Inspecting individual cells at 100x magnification (Fig. 4.18) revealed fluorescent signal from the pellicular membrane and the flagellum. Moreover, strong fluorescent signal is seen from what appear to be membranous filaments ("streamers"), a type of appendage frequently seen on trypanosomatid cell extremities and flagellar membranes (Schelipewsky 1912; Ellis, Ormerod, and Lumsden 1976). In addition, punctate signals were observed within cells, possibly highlighting part of the endosomal system trafficking protein to the cell surface. Further punctate signals lying outside of cells may constitute shed streamer material or secreted vesicles. Whether or not the presence of the observed streamers is an artefact of the expression, or even over-expression, of the fusion protein is unclear and warrants further investigation.

The localisation in AXA is difficult to discern (Figure 4.19). Whilst the cell body is still highlighted, the edge effect one would expect from a surface membrane protein is only faintly evident. Signals localised to the anterior of the cell can be seen, which may be localising to the endosomal system and/or flagellum.

These localisations are distinct from reports of amastin-localisation in *Leishmania spp.* (Wu et al. 2000) where no released material was reported and instead a cell surface localisation in both PRO and AXA shown as well as flagellar localisation in PRO. Rochette in turn localised amastins to the cell-surface, flagellar pocket and flagellar membrane, which does not correlate with my findings for the LmxM.16.0500:eGFP fusion protein.

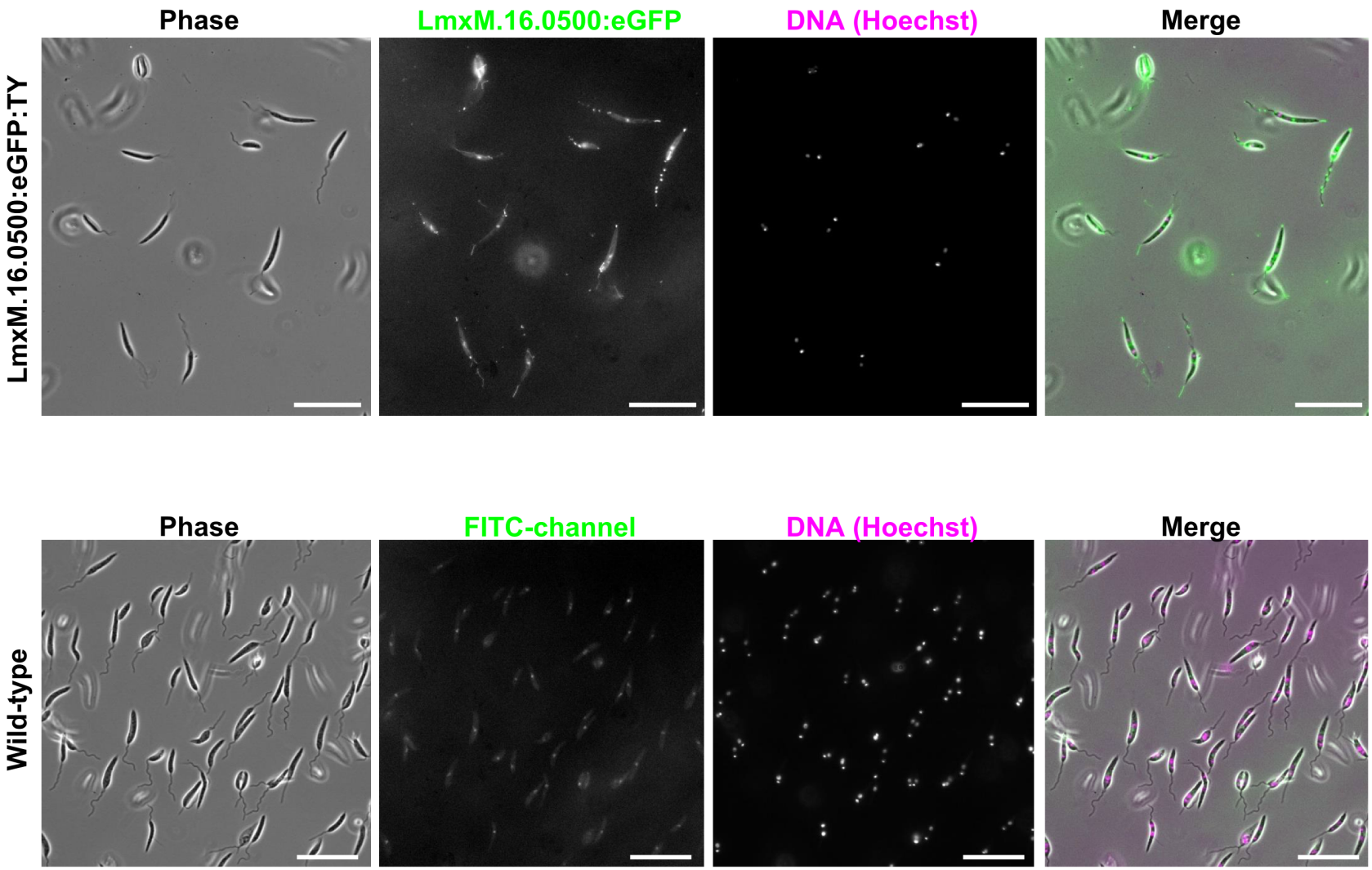


Figure 4.17 Distribution of LmxM.16.0500:eGFP:TY in a population of promastigotes Fluorescence-microscopic analysis of the distribution of the LmxM.16.0500:eGFP:TY fusion protein at 40x magnification. For comaprison, images of wild-type cells acquired and processed using the same parameters to match tagged cell line are shown. Scale bar is 20 µm.

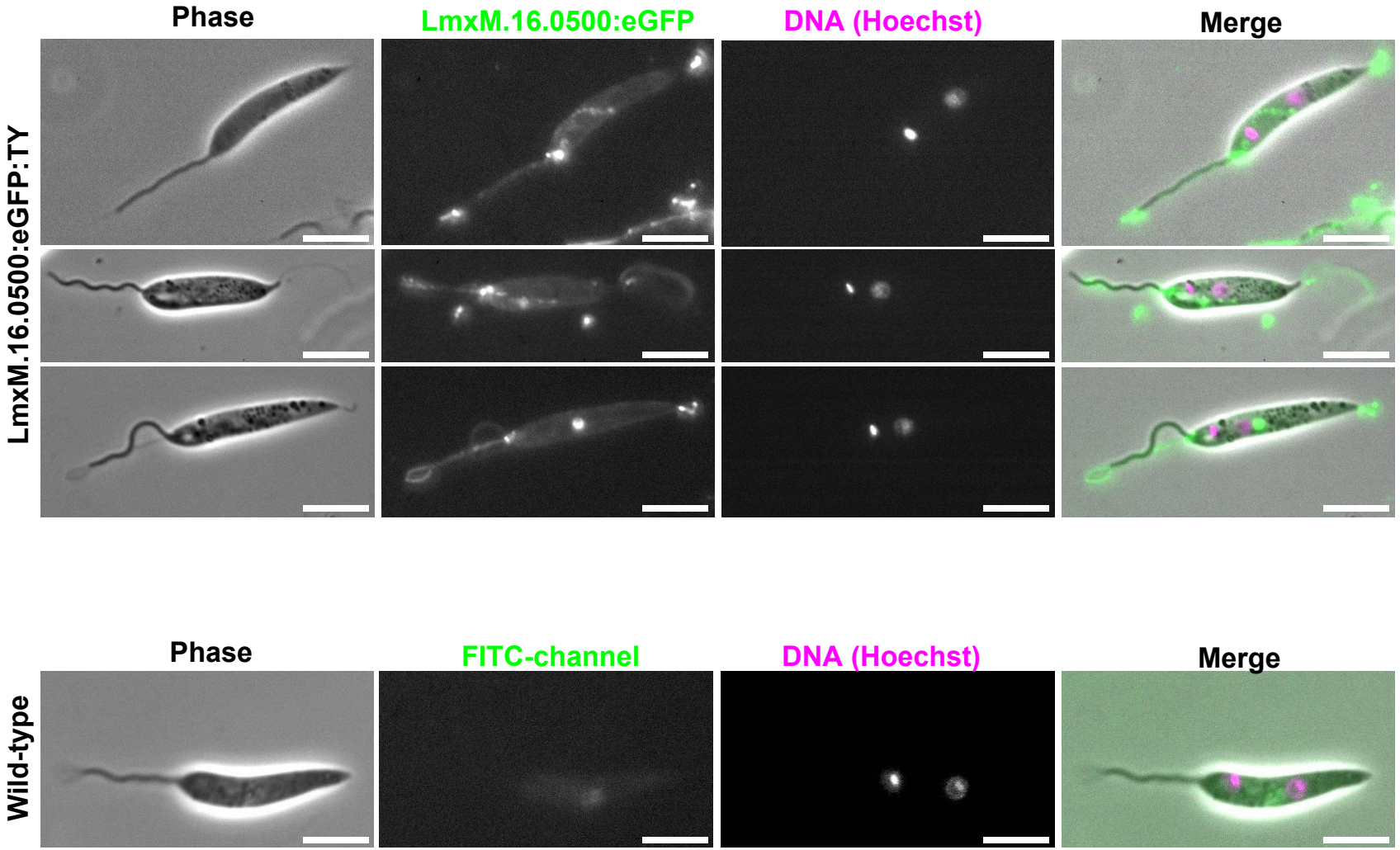


Figure 4.18 Sub-cellular localisation of LmxM.16.0500:eGFP:TY fusion protein in promastigotes Live-cell microscopic analysis of cells promastigotes expressing the LmxM.16.0500:eGFP:TY fusion protein at 100x magification. For comparion, images of wild-type cells acquired and processed using the same parameters to match tagged cell line are shown. Scale bar is 5 µm.

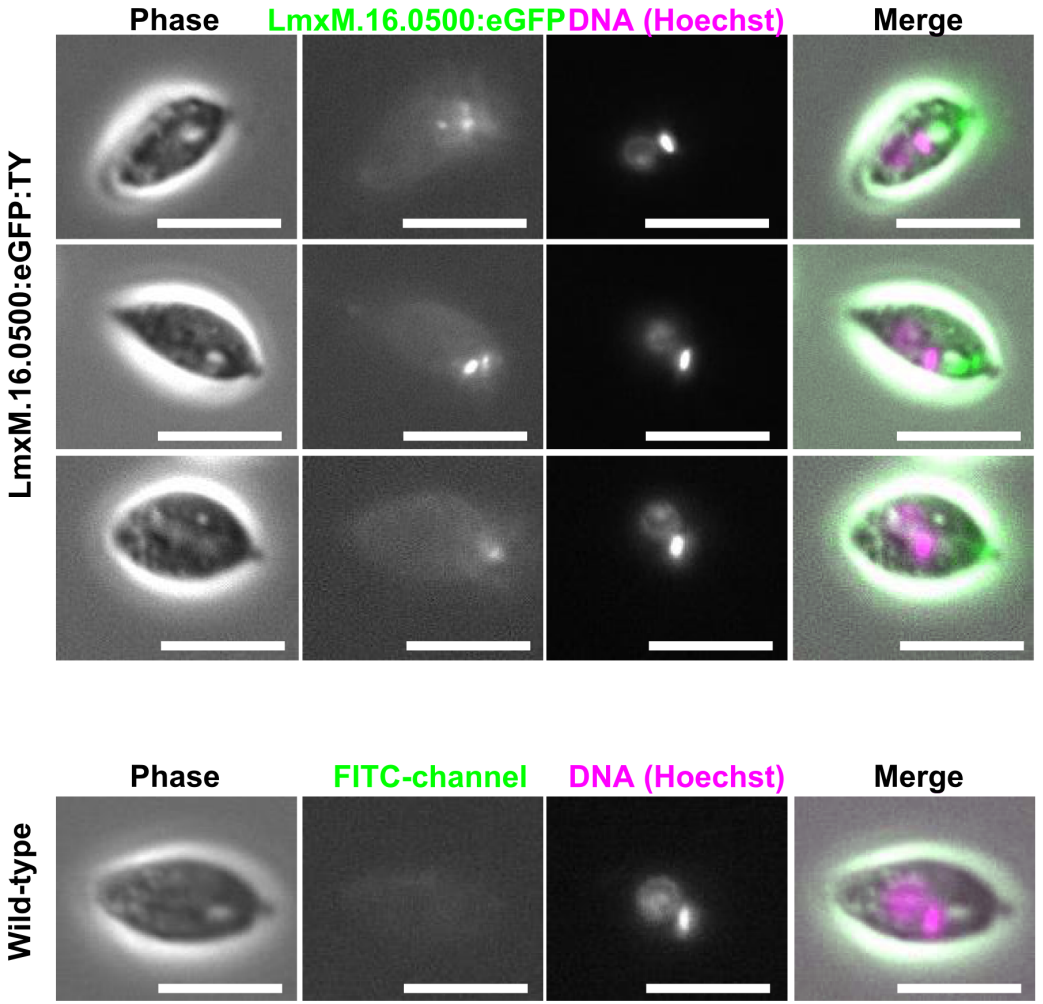


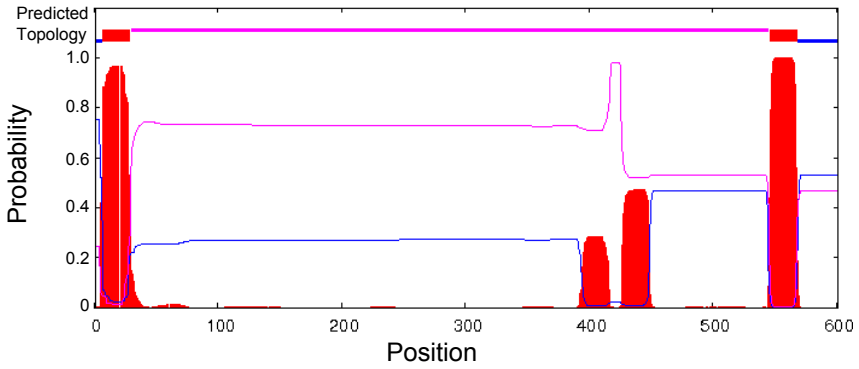
Figure 4.19 Sub-cellular localisation of LmxM.16.0500:eGFP:TY in 24 h axenic amastigotes Fluorescence-microscopic analysis of the sub-cellular localisation of the LmxM.16.0500:eGFP:TY fusion protein. For comaprison images of wild-type cells, acquired and processed using the same parameters are shown. Scale bar is 5 µm.

4.3.4 LmxM.09.1330 is a marker of the amastigote flagellar pocket

LmxM.09.1330 is a protein with 2 predicted TMDs, separated by a large extracellular loop (Figure 4.20), and has a predicted signal peptide (SignalP). It is conserved amongst *Leishmania spp.*, *Crithidia fasciculata*, *T. cruzi, T. grayi* and *T. vivax*, but absent from *T. brucei* and species outside of kinetoplastids. No similar proteins are found in *L. mexicana* by BlastP.

Alignment of the orthologous sequences of LmxM.09.1330 (Figure 4.21) revealed that the N-terminus of the protein is divergent between species, whilst the C-terminal region is highly conserved. The transcript is 2.7 fold higher expressed in AMA compared to PRO ($p= 5.02x10^{-17}$) (Table 4.1), and absolute expression levels (21.5 – 45.1 FPKM) are in the range of mean expression levels observed globally (25.8 – 53.6 FPKM). In microarray studies the *L. major* homologue (LmjF.09.1330) was previously found to have higher expression in AMA compared to PRO as well as higher expression in metacyclics compared to PRO, with higher fold change in the comparison of metacyclics to PRO than AMA to PRO (Akopyants et al. 2010) (Figure 4.22 A). The *L. infantum* homologue (LinJ.09.1400) was shown to have higher expression in AMA and AXA compared to PRO (Rochette et al. 2009) with a higher fold-change in AXA vs. PRO than in AMA vs. PRO (Figure 4. 22 B). During the timecourse experiment analysing axenically differentiating *L. donovani* cells (albeit, mapped to *L. infantum* genome, therefore the corresponding gene-accession is LinJ.09.1400) (Lahav et al. 2011), the expression level of the transcript was high from the start, and only increased slightly (Figure 4. 22 C).

Western blot analysis (see Materials and Methods) showed a distinct band at around 120 kDa and a more diffuse signal between 85 – 100 kDa (Figure 4.23), suggesting a degree of post-translational modifications as the expected fusion protein size was around 92.2 kDa (LmxM.09.1330 = 65.3 kDa, eGFP = 26.9 kDa). LmxM.09.1330 is predicted to be O- and N-glycosylated (based on NetNGlyc 1.0 Server and NetOGlyc 4.0 Server), both types of



Colour code: Transmembrane, Intracelllular, Extracellular

Figure 4.20 The topology of LmxM.09.1330 Diagram showing the predicted topology of LmxM.09.1330. A colour-code indicated the transmembrane domains as well as intracellular and extracellular loops. Predictions performed using TMHMM Server 2.0.

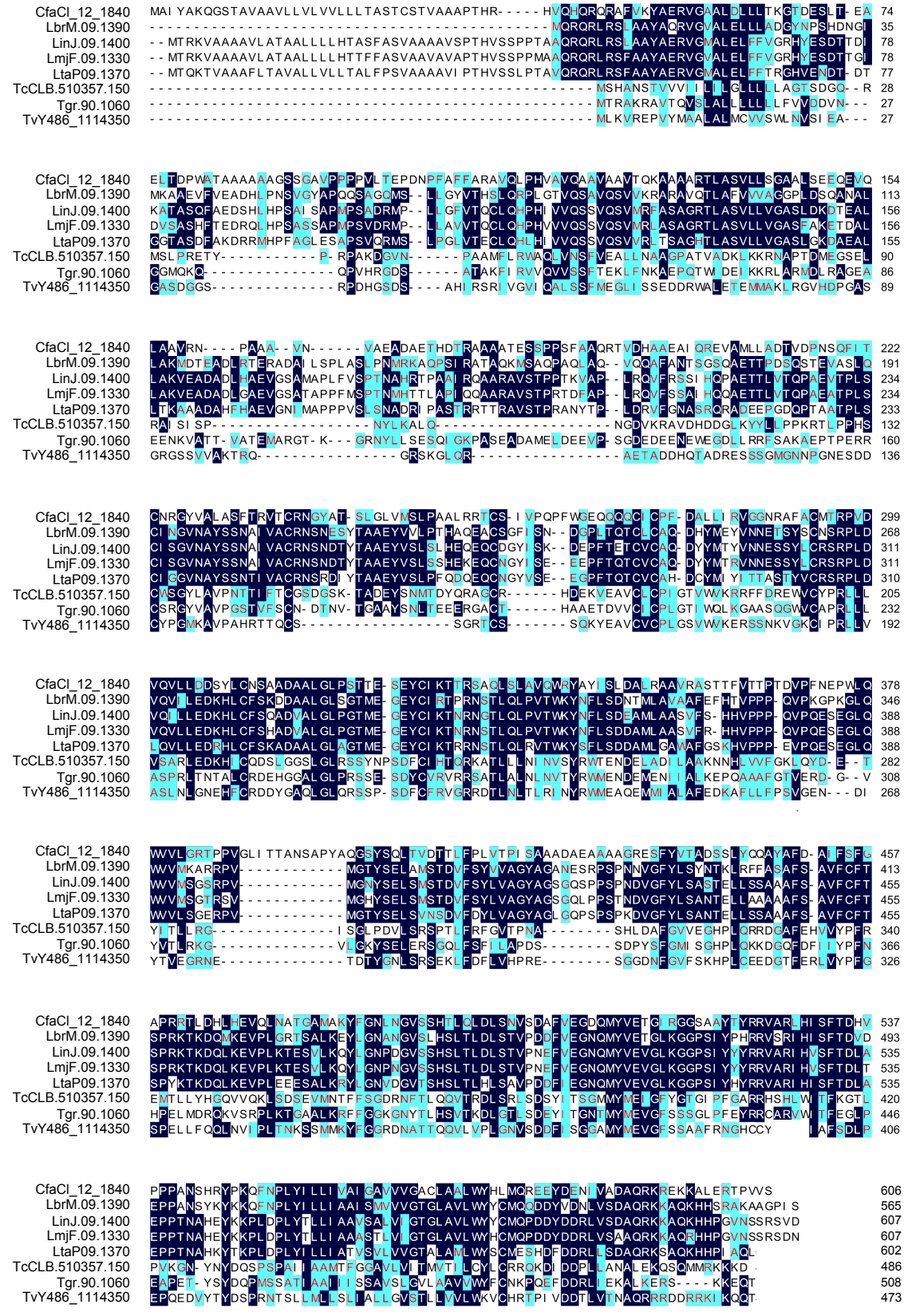


Figure 4.21 Alignment of LmxM.09.1330 orthologues Protein sequence alignment of homologues of LmxM.09.1330 from *C. fasciculata (CfaCl)*, *L. braziliensis(LbfM.)*, *L. infantum (LinJ.)*, *L. major (LmjF.)*, *L. mexicana (LmxM.)*, *L. tarentolae (LtaP)*, *T. cruzi (TcCLB.)*, *T. grayi (Tgr.)* and *T. vivax (TvY).* Aligments generated using Clustal W. Dark blue colouring of sequence indicates conserved amino acid, light blue colouring of sequence indicates conservation of a functional group in at least 4 species.

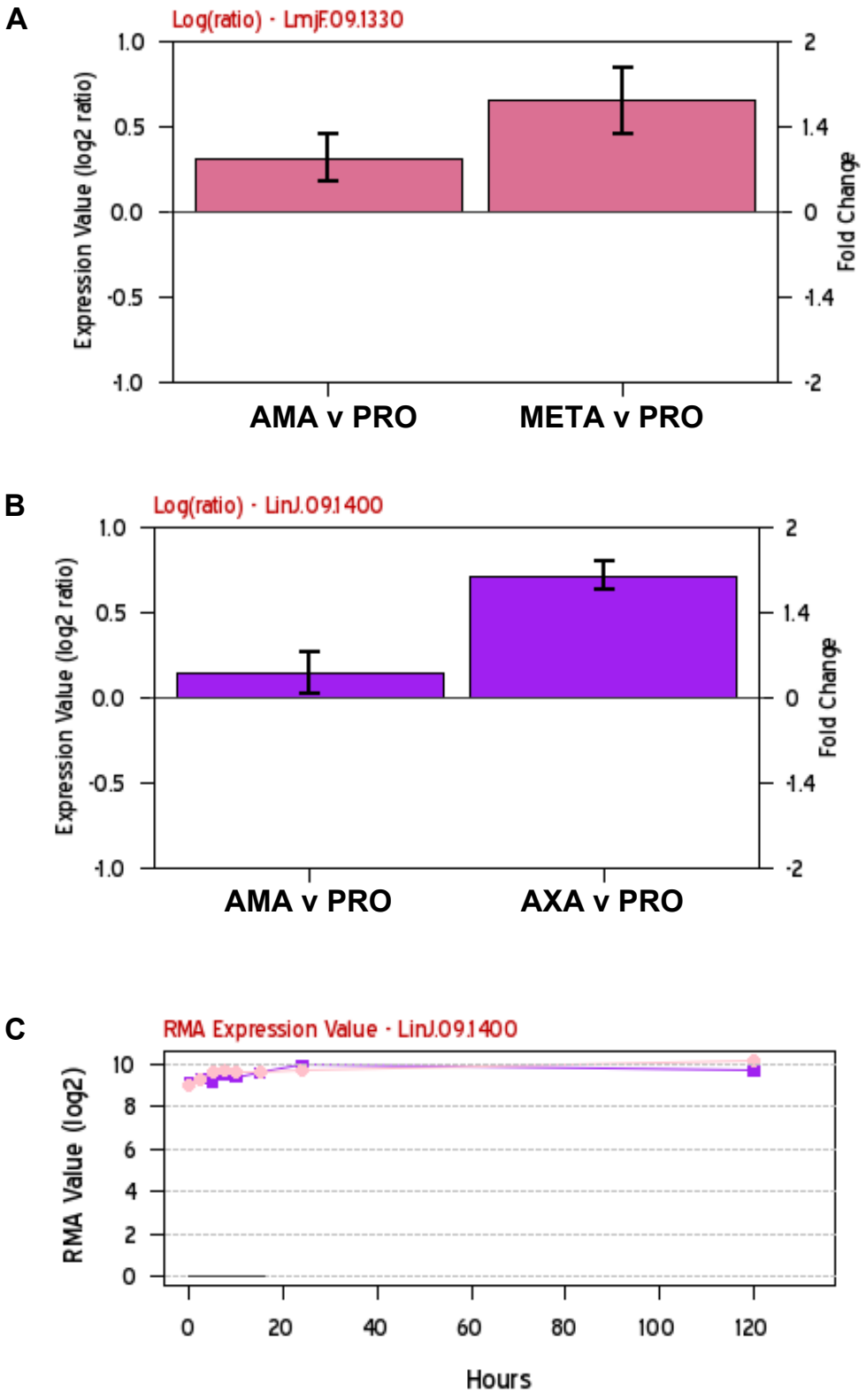


Figure 4.22 Expression summaries for orthologues of LmxM.09.1330 Graphs summarising differential expression analyses of *L. major* (A), *L. infantum* (B) and *L. donovani* (mapped to *L. infantum* genome) (C) orthologues of LmxM.09.1330. (C) is a time-course of axenic amastigote differentiation. The diagrams were taken from TriTrypDB. RMA = Robust Multi-Array Average.

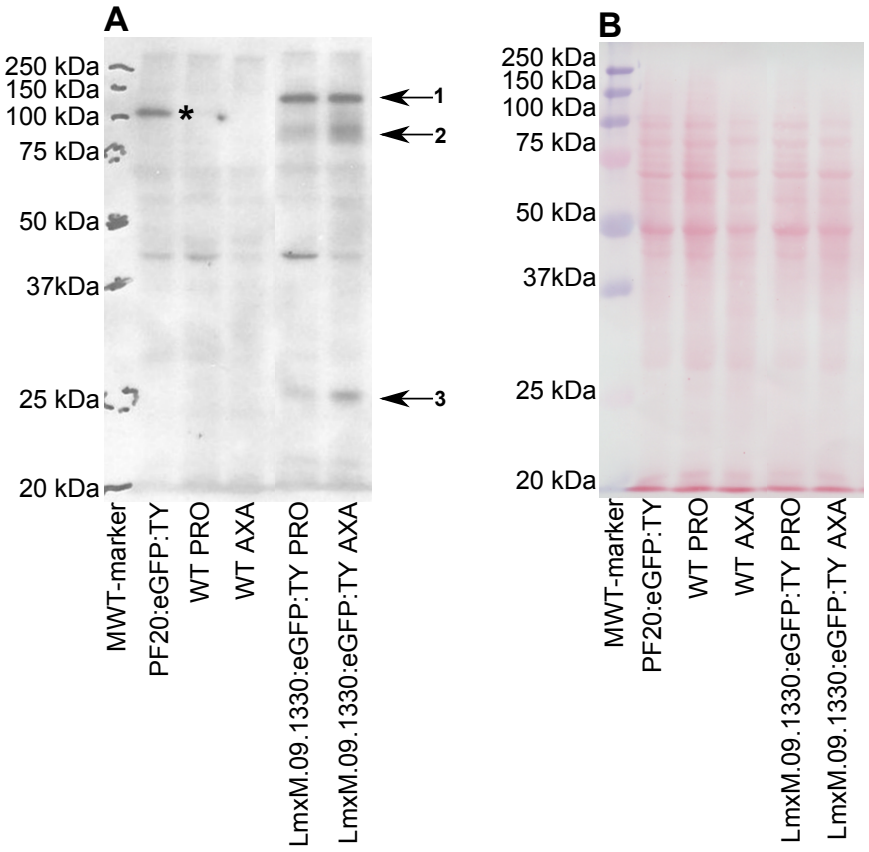


Figure 4.23 Western blot analysis of LmxM.09.1330:eGFP:TY expression Scan of (A) an membrane probed with anti-GFP antibody. A pLent-tagged cell line expressing PF20:eGFP (LmxM.18.0470) was used as positive control (*, expected: 97.3 kDa) and untagged wild type (WT) PRO and AXA used as negative controls. Per lane, whole-cell lysates for ~5x10^6 cells used for all samples and separated on 10%-SDS-PAGE gel. Expected size of LmxM.09.1330:eGFP is 92.2 kDa, to which band 2 may correspond, whilst band 1 may represent a post-translationally modified form of the fusion protein. Band 3 may represent a degradation product. Ponceau stain of the same membrane is shown in (B).

modifications found in *L. mexicana* (Ilg et al. 1994). An additional band at 25 kDa was observed in PRO and AXA expressing LmxM.09.1330:eGFP. The significance of this band is not clear, but may represent a degradation product of the fusion protein. The absolute signal intensity was low and the diffuse signal between 85-100 kDa as well the band at 25 kDa were more intense in the AXA sample compared to the PRO sample.

Using immuno-fluorescence microscopy, only a very weak fluorescent signal was seen on the cell body and flagellum of PRO (Figure 4.24). Analysis of AXA by this method showed a consistent focus of fluorescent signal anterior of the kinetoplast (Figure 4.25). Closer inspection showed that the signal forms a ring-like structure, consistent with the fusion-protein localising to the flagellar pocket (Figure 4.26). Therefore, LmxM.09.1330 appears to be selectively localised to the flagellar pocket membrane in AXA, which may point to a mechanism of restriction of localisation activated during differentiation, as all membrane proteins are trafficked via the flagellar pocket. Whether this is due to differential membrane protein-trafficking machinery present in amastigotes, or a result of differential processing of LmxM.09.1330, which may correspond to the increase in relative intensity of the 85-100 kDa bands observed in the Western blot of AXA compared to PRO, will be an interesting aspect of biology to investigate.

This fusion protein is one of the very few reported proteins localising exclusively to the amastigote (in this case axenic amastigote) flagellar pocket and the first such report (to my knowledge) in *L. mexicana.*

4. 4 Discussion and Conclusions

An eGFP-tagging screen of 26 putative transmembrane proteins with higher mRNA expression in amastigotes compared to promastigotes was performed. 22 of these cell lines were analysed microscopically, and for 18 a fluorescent signal possibly corresponding to the fusion-protein was detected. Observed localisations such as the reticulate and tubular localisation possibly corresponding to the endoplasmic reticulum or mitochondrion, as well as those on the cell surface were consistent with the fusion

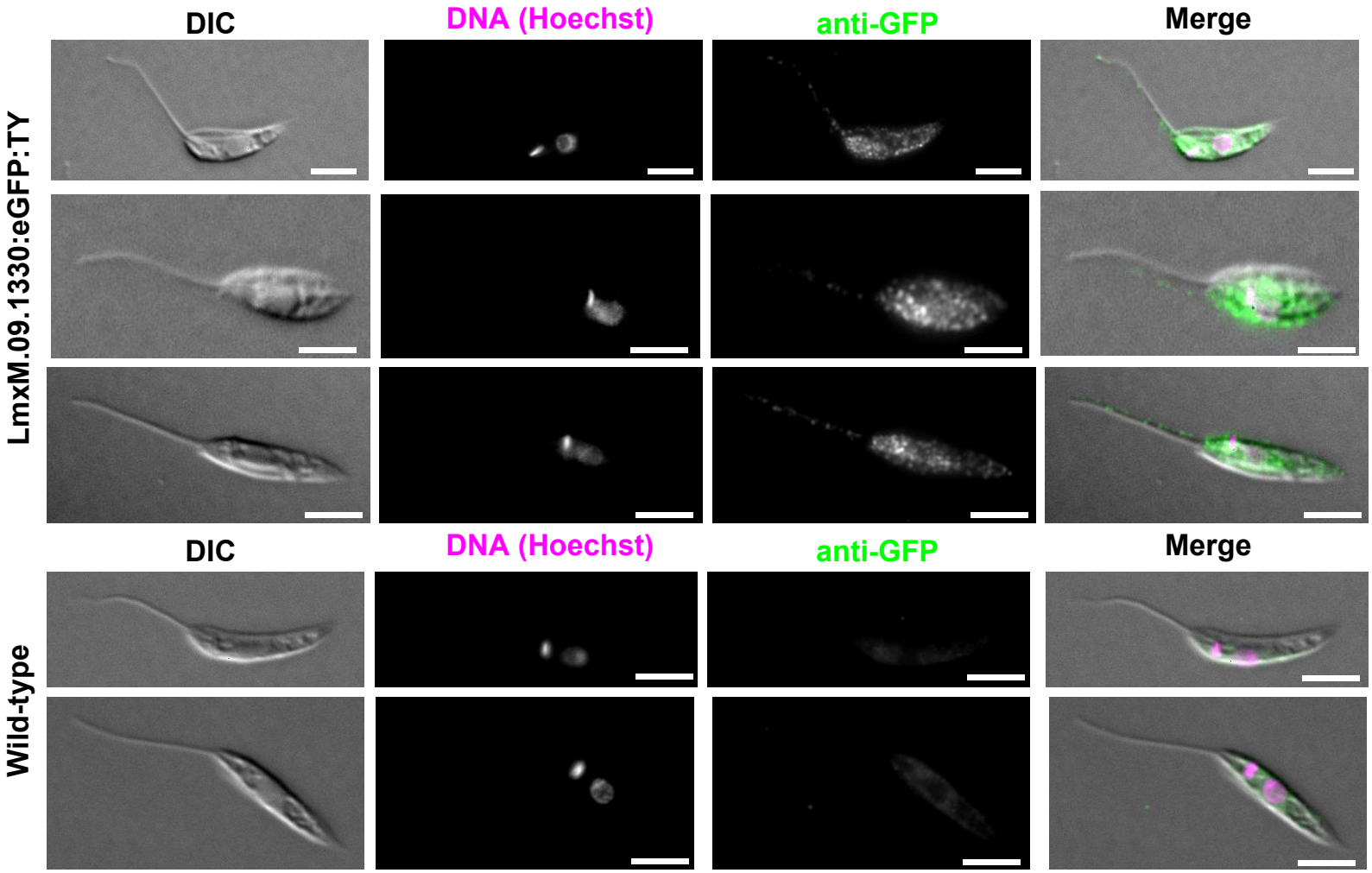


Figure 4.14 Sub-cellular localisation of LmxM.09.1330:eGFP:TY in promastigotes Immuno-fluorescence microscopic analysis of the sub-cellular localisation of the LmxM.09.1330:eGFP:TY fusion protein at 100x magnification. Cells stained with anti-GFP antibody. For comaprison, images of wild-type cells, stained like mutant cells, acquired and processed using the same parameters are shown to match tagged cell line. Scale bar is 5 µm.

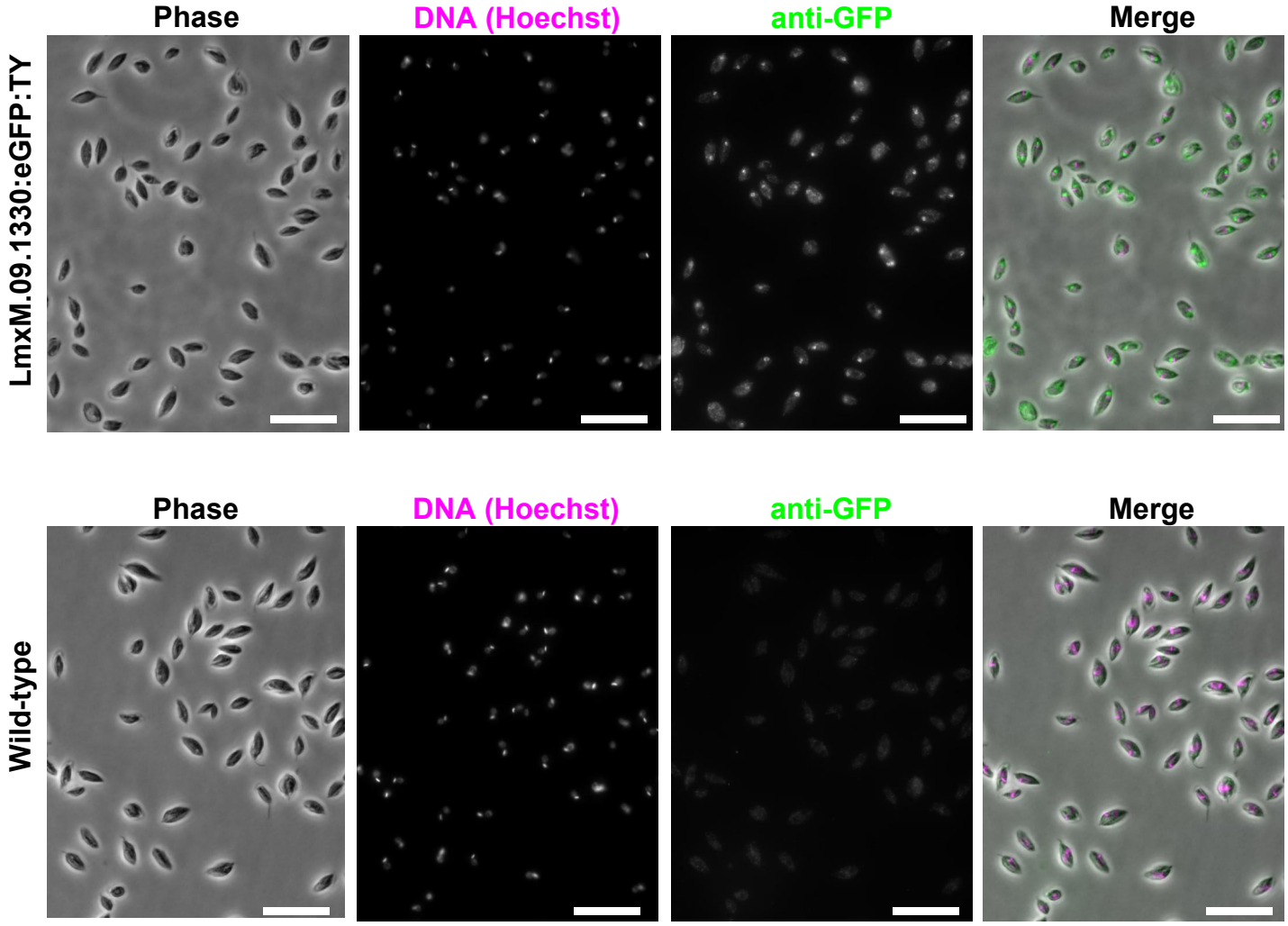


Figure 4.25 Distribution of LmxM.09.1330:eGFP:TY in a population of 24 h axenic amastigotes Immuno-fluorescence microscopic analysis of the distribution of the LmxM.09.1330:eGFP:TY fusion protein at 40x magnification. Cells stained with anti-GFP antibody. For comaprison, images of wild-type cells, stained like mutant cells, acquired and processed using the same parameters are shown.

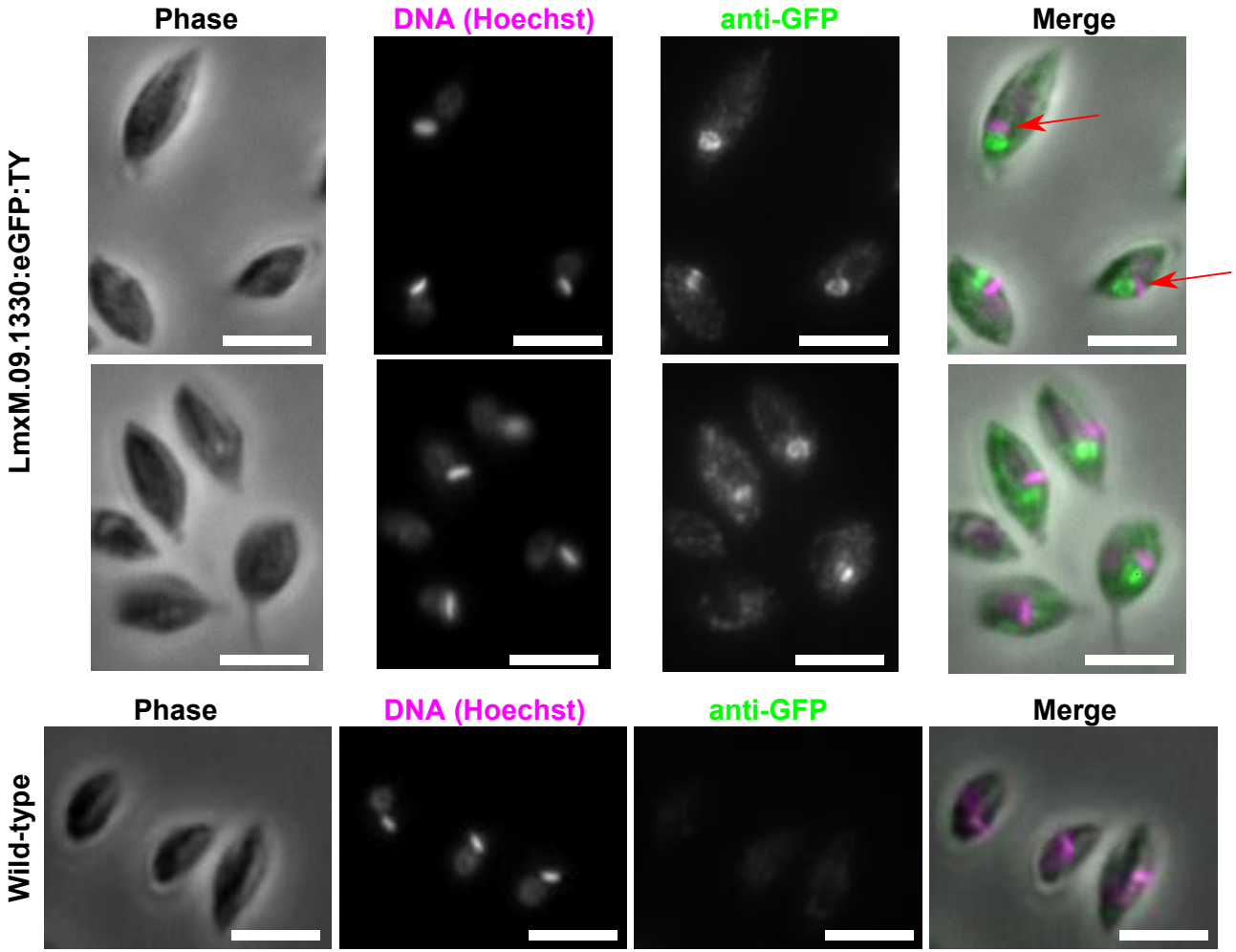


Figure 4.26 Sub-cellular localisation of LmxM.09.1330:eGFP:TY in 24 h axenic amastigotes Immuno-fluorescence microscopic analysis of the sub-cellular localisation of the LmxM.09.1330:eGFP:TY fusion protein at 100x magnification. Cells stained with anti-GFP antibody. Red arrow indicates the kinetoplast which is stained more intensely by Hoechst DNA stain than the nucleus (Wheeler, Gull, and Gluenz, 2012) For comaprison, images of wild-type cells, stained like mutant cells, acquired and processed using the same parameters are shown to match tagged cell-lines. Scale bar is 5 μm.

proteins being transmembrane proteins, albeit further biochemical analyses would be required to conclusively say that all observed fusion proteins are indeed transmembrane proteins.

Two cell lines were chosen for more in-depth analyses, because of the intriguing localisation of the fusion proteins expressed by them.

LmxM.16.0500 a member of an expanded gene family in *Leishmania spp.* compared to *T. brucei*. Other members of this family, with a likely evolutionary root within δ-amastins (Jackson 2010), are the two very similar proteins, LmxM.16.0490 and LmxM.16.0470. The fact that LmxM.16.0500 is amongst the top percentile of highest expressed genes in AMA, suggests that it plays an important role in the amastigote stage. LmxM.16.0500 has a surface membrane localisation and is found localised to extracellular structures resembling streamers or vesicles. It is highly expressed and its release from the cell body may play a role in interaction with its host, providing that it is not an artefact of dysregulation of its expression or localisation due to the creation of the eGFP-fusion protein. To test this, tissue culture supernatants from wild-type and LmxM.16.0500:eGFP *L. mexicana* could be compared by proteomic methods to discern whether native LmxM.16.0500 is found released from cells and whether relative abundances are similar between wild-type and mutant cell lines.

The putative membrane protein LmxM.09.1330 was observed faintly distributed across the cell body and flagellum in promastigotes, but appeared to form a clear ring around the flagellar pocket in AXA. Its restriction to the flagellar pocket may be a result of differential expression of trafficking machinery or of differential processing of the protein itself. Investigation into whether or not LmxM.09.1330 truly localises to the flagellar pocket could be aided by co-staining of cells expressing LmxM.09.1330:eGFP with e.g. lectins that selectively stain the flagellar pocket as has been shown in *T. brucei* (Brickman and Balber 1990). Providing that the creation of the fusion protein did not interfere with its native

localisation, LmxM.09.1330 is, to my knowledge, the first (axenic-) amastigote-specific flagellar pocket marker in *Leishmania mexicana* and may as such be a useful tool in investigating the flagellar pocket in amastigotes. Generation of antibodies against LmxM.09.1330 followed by immunofluorescent staining of fixed *L. mexicana* cells could permit microscopic investigation of the localisation of native LmxM.09.1330. The identity and origins of the multiple high-molecular weight bands observed in Western-Blot analysis could be further investigated, in a first instance by de-glycosylating protein samples prior to gel-electrophoretic separation: If the multiple bands are a result of glycosylations, de-glycosylation should result in only a single high molecular weight band corresponding to the calculated molecular weight being observed. Alternatively, other post-translational modifications such as phosphorylation may be considered as the origin of the multiple banding patter. This failing, mass-spectrometric approaches could reveal the differences between the species constituting the different bands.

Ultimately, for both LmxM.16.0500 and LmxM.09.1330, null-mutants and infection studies will inform whether these proteins play a role in infection and amastigote survival.

The eGFP-tagging screen yielded fewer fusion proteins localising to the cell surface of the parasite than hoped, whilst a considerable number of proteins are localising to cell-internal compartments. Transmembrane proteins have notoriety for being difficult to work with, in particular for accepting tags. The GFP protein can misfold upon being fused to another protein and may be preventing it from passing to and through the Golgi to the surface membrane (Vashist and Ng 2004). This resulted in the development of "superfolder" GFP-variants (Pédelacq et al. 2006). The possibility that use of a "superfolder" GFP-variant could have increased the number of fusion-proteins localising to the cell surface needs to be entertained. Alternatively, a different type of tag could circumvent the problem of protein-misfolding, such as a small epitope tag (e.g. c-myc tag). Finally, the pLENT vector was designed to provide expression of the fusion protein in both

promastigotes and amastigotes. The expression of an amastigote-specific protein in the promastigote stage could be deleterious for the parasite, resulting in loss of the protein, a process well-documented in acquisition of drug-resistance (Mukherjee et al. 2013). Therefore, the results of the tagging screen could have potentially been improved by using either the native 3'UTR of the tagged gene, or use of a 3'UTR of a known amastigote-specific transcript.

Finally, I must concede that the use of axenic amastigotes instead of intracellular amastigotes in the screening of the cell lines created may have precluded us from identifying the correct localisation of some proteins. Whilst it is convenient to work with axenic amastigotes, we have seen in chapter 3 that they do differ from intracellular amastigotes. Moreover, should a protein indeed be secreted and accumulate in the parasitophorous vacuole, at the vacuolar membrane or any other compartment in the host cell, this exciting finding would be missed using an axenic system. I would have particularly wanted to investigate the localisation of LmxM.16.0500:eGFP in intracellular amastigotes, but time constraints prevented me from doing so and I hope this experiment will be performed in the near future.

Chapter 5 - Discussion

5.1 Aims

I set out to generate a full RNA-seq based description of the transcriptome of *L. mexicana* promastigotes (PRO), 24 h axenic (AXA) and intracellular (AMA) amastigotes at single-nucleotide resolution. I sought to generate a description of the dimension of transcripts and coding sequences as well as a description of the transcript abundances within each cell type and a description of the differential abundances between cell types. Subsequently, I aimed to use these data to identify transmembrane proteins preferentially expressed in the amastigote stage that may play a role in parasite survival or infection.

5.2 Chapter 2 - Prediction of gene models using RNA-sequencing guided definition of transcript boundaries

In this work I present the first description of the transcriptome of *L. mexicana* PRO, AXA and AMA at single-nucleotide resolution. For PRO and AXA we employed what I would term "conventional" RNA-seq, i.e. only material from a single species was present in the sample. For the AMA sample we employed "dual" RNA-seq, i.e. material from two species was present in the sample, namely RNA from *L. mexicana* and mouse. The dual RNA-seq approach relies on *in silico* separation of the transcriptomes during analysis. As I was able to show, by comparing the relative abundances of ribosomal RNA from either species in the AMA samples to the proportion of sequencing reads mapping to the mouse or *L. mexicana* part of a hybrid-genome, our *in silico* separation method was able to recapitulate the estimated composition of the samples. This has important consequences for the transcriptomic study of intracellular pathogens: Not only do methods for the purification of the pathogen that may affect its transcriptome and often have very low yields (using protocols in our laboratory <10% of intracellular amastigotes can be recovered) become obsolete, but also do we obtain transcriptomic material from the infected host-cells. In the study presented here, I did not address the mouse-transcriptome, but I am certain that the

data gathered here will form part of a wider study into the response of the host cell to infection.

One of the many advantages of RNA-seq based transcriptomic approaches compared to microarray-based studies is that one does not have to define a genome model in the form of a hybridisation platform prior to data-acquisition. Therefore, it was possible to use the RNA-seq data first to refine gene models and only in a second step perform quantitation of transcript abundances.

The presence of a 5' spliced-leader in every mRNA (Lenardo, Dorfman, and Donelson 1985; Sutton and Boothroyd 1986), along with polyadenylation sites at the 3' end of mRNAs, makes genome-wide definition of transcript boundaries not only conceptually simple, but with recent advances in software tools (Fiebig et al. 2014) also practically simple.

Mapping of SLAS and PAS revealed considerable heterogeneity, with PAS sites being more diverse (median= 9 sites per gene) than SLAS sites (median= 2 sites per gene). This correlates well with findings in *T. brucei* (Kolev et al. 2010).

Using these 5' and 3' transcript-end features, I predicted the dimensions of all protein coding transcripts in *L. mexicana*, constituting the first genome-wide description of these. These predictions correlate well with Northern-blot derived size estimates of transcripts, validating this approach of defining transcript dimensions. One previously used alternative method to defining transcript boundaries could have been an approach as employed by Rastrojo *et al.* (Rastrojo et al. 2013) who first defined transcripts in *L. major* based on sequencing-read coverage and subsequently used SLAS and PAS positions to divide up the resulting transcript models. Whilst there may be merits to this when it is suspected that the number of SLAS and PAS mapped is insufficient, there are dangers to this approach as I illustrate on the example of the glucose transporter in Figure 2.16. A sharp dip in read coverage was observed within a coding sequence, which would likely have resulted in two different transcripts being defined in that locus. However, if the

assumption truly is that insufficient SLAS and PAS are mapped and therefore there may be merit to utilising transcript coverage to define transcript boundaries, I find it bold of the authors (Rastrojo et al. 2013) to propose the presence of polycistrons based on the lack of SLAS within a coverage-defined transcript.

Using the transcript dimensions I was able to define the sizes of the 5' and 3' UTRs. Their sizes are larger than previously reported for *T. brucei* (Siegel et al. 2010; Kolev et al. 2010; Nilsson et al. 2010), but correlate well with data from *L. major* (Rastrojo et al. 2013). That differences in the dimensions of UTRs may exist was to be expected considering the larger dimensions of intergenic sequences reported after the sequencing of the TriTryp genomes in 2005 (El-Sayed et al. 2005). Whilst the reasons for this remain unclear, the analysis of sequence biases surrounding SLAS presented in Section 2.3.5 showed that the dimensions of sequences with a lower mean base-pairing probability around the SLAS is also larger in *L. mexicana* than in *T. brucei*, which could lead one to envisage slightly different splicing machinery being present in *T. brucei* and *L. mexicana*. As a result, the question begs what would happen if a large chromosomal fragment of *T. brucei* were to be transfected into *L. mexicana* and vice-versa. If the precise dimensions of the sequences with lower propensity to secondary structure do play a crucial role, inefficient or indeed incomplete splicing should be observed, which could be demonstrated by reverse-transcriptase PCR approaches.

The SLAS and PAS positions enabled me to identify 936 novel genes, bringing to the total number of protein coding genes to 9169. Moreover, I proposed 1253 of extensions and 184 truncations of protein N-termini based on SLAS-positions.

All of the novel genes described in my data set have both SLAS and PAS, setting them apart from predictions made by Rastrojo *et al.* (Rastrojo et al. 2013), where this was not a requirement. In turn however, not all of the transcript identified by my method were found to be expressed, possibly indicating an artefact that would not have occurred by the

method employed by Rastrojo *et al.* (Rastrojo et al. 2013). To date, no other studies have sought to use RNA-seq data to improve *Leishmania* genome annotation. This may primarily be hampered by their inability to map both SLAS and PAS due to the RNA-seq approach employed (Mittra et al. 2013; Martin et al. 2014), using a cDNA library preparation preferentially amplifying fragments containing parts of the spliced-leader sequence and sequencing only short (36 nt) single-ended reads. Due to this library preparation, the resulting sequencing data is unlikely to contain any reads covering polyadenylation sites. Moreover, the single-ended nature of this data does not permit assembly of overlapping reads to increase mapping-accuracy as we have employed in the work presented herein and implemented in the SLaP Mapper tool (Fiebig et al. 2014).

As the novel genes were first defined as transcripts, the actual coding sequences (CDS) needed to be defined within them. We decided to define the first open-reading frame (ORF) that could encode for a protein of 25 AA or larger without stretching of the next downstream SLAS as the coding sequence of the transcript. This opens up two important points of discussion. First, is the first ORF the coding sequence? Second, what is the minimum size of protein that is realistic and needs to be considered in an annotation?

The assumption that the first ORF in a transcript is the CDS is clearly naïve, considering that upstream ORFs (uORFs) are found in 12.2 % or transcripts and have been shown to be important affectors of translational efficiency in kinetoplastids (Vasquez et al. 2014). Therefore however, without any other information available, calling the first ORF the CDS only has a 12.2% error-rate, outperforming even the simplest random-choice scenario with at least a 50% error rate.

Illustrative of this, using mass-spectrometry I was able to obtain peptide-evidence for the translation of 42 of the novel transcripts (47 when incl. the data from (Paape et al. 2010)). The way the data-base used in the assignment of mass-spectra was designed also permitted the identification of other ORFs other than the initially proposed CDS. Five out

of 42 proteins detected did not correspond the first ORF in the transcript, giving my (naïve) assumption an error rate of 11.9 %, reflecting the genome wide uORF occurrence in 12.2 % of genes. The presence of orthologues in other species also opens up avenues of predicting the CDS in transcripts based on conservation of sequence elements in other species. Using the best consensus reciprocal blast (BCRB) method I devised, I was able to correctly predict 40 out of 42 of the proteins I detected by mass-spectrometry. The BCRB method predicted all CDS which were preceded by a uORF correctly. Therefore, using multiple sequence comparisons, we can improve the prediction of CDS within novel transcripts, even when the resulting CDS may be untraditionally small. The difference of this approach to homology-based annotation methods used in the past lies in the guidance of the BCRB method by experimental data, namely the SLAS and PAS defined transcript dimensions.

The smallest protein detected by mass-spectrometry, LmxM.09_94875, was 39 amino acids in length. Minimum protein sizes in genome annotations have traditionally been considered to lie at 100 AA ((Storz, Wolf, and Ramamurthi 2014) and references therein). Recently, more and more attention has been paid to proteins below that size cut off. In prokaryotes, proteins smaller than 100 AA have been implicated in a variety of processed, e.g. cell morphogenesis (CmpA, 37 AA (Ebmeier et al. 2012)) and cell division (MciZ, 40 AA, (Handler, Lim, and Losick 2008)). In eukaryotes, RNA-seq studies of common model organisms ranging from yeast, over plants and nematodes and to vertebrates have started have started revealing a great diversity of small, translated ORFs (Ruiz-Orera et al. 2014). It is therefore not surprising that novel transcripts identified initially by Kolev *et al.* (Kolev et al. 2010) also encoded for some small proteins, a subset of which was functionally characterised by Ericson *et al.* (Ericson et al. 2014). In line with a minimal size cut-off set

by the latter two studies, we also used 25 AA as a minimal protein size, fully recognising that even smaller proteins may being encoded and functional.

The small size of the proteins is likely to be the main reason why they had been missed in previous genome annotation. The novel genes are highly conserved amongst *Leishmania spp.* in particular *Leishmania Leishmania spp.*, albeit conservation amongst other, more distantly related kinetoplastids is also seen. This suggests that a subset of these proteins were present in ancestral kinetoplastids, whilst others were gained during the divergence of *Leishmania spp.* (or selectively lost in other kinetoplastids). If these sequences were gained during the divergence of *Leishmania spp.* from other kinetoplastids, one may propose that they play a role in the adaptation to intracellular parasitism and form an exciting new group of proteins to investigate further experimentally.

The N-terminal extensions of proteins I proposed based on start-codons in the 5' UTR in frame with annotated start codons also received some confirmation from mass-spectrometric evidence. Peptides mapping to 116 of proposed extensions were detected. However, *in silico* digestion of extended sequences suggest that, in theory, peptides ranging between 7 and 29 amino acids can be obtained for 411 of the 433 proteins with predicted extensions amongst the mass-spectrometrically detected proteins following tryptic digest. How much of this discrepancy is due to experimental limitations and how much is due to the first ATG-codon not being the actual start codon is unclear. Mass-spectrometric approaches seeking to enrich for N-termini of proteins may shed light onto this question and certainly will be a valuable asset in the continuation of the refinement of the genome annotation for *L mexicana* (and genomes in general). One such approach involves using an engineered enzyme called Subtiligase, derived from a bacterial serine protease, which, with high specificity, can add a biotin labelled tag to N-terminal α-amines (Wildes and Wells 2010). Using avidin-coated beads, the N-termini of proteins may be

captured and following tryptic digestion, N-terminal peptides released from the beads and analysed mass-spectrometrically.

5.3 Chapter 3 – Transcriptomic Characterisation of Promastigotes, Axenic Amastigotes and Intracellular Amastigotes

Before proceeding to the quantification of transcript levels in AMA, AXA and PRO I sought to assess how "clean" the transcriptomic data we had was. This mainly focussed on the levels of non-polyadenylated transcripts in the data set, mainly tRNAs and rRNAs. As I show in section 3.3.4.3 we do not observe tRNAs, but do observe rRNAs, albeit only one species. With the abundance of rRNAs and tRNAs in the cell (estimates lie at around 80 % rRNA and 15 % tRNA (Lodish et al. 2000)) it is not surprising that some will have "leaked" through poly-A-selection. How much "leaks" through selection in other studies is difficult find out, as it is not reported. I suspect, since this is a metric giving, perhaps uncomfortable, insight into the efficiency of poly-A selection and thereby quality of sample preparation, this is "just not talked about".

We observe quite different levels of rRNA contamination between samples, which had the potential to impact quantification. However, the gene models subsequently used in quantification (namely the transcript models created in the previous chapter) did not contain rRNA transcripts and therefore selectively mask these from the quantification, giving me reason to believe that the rRNA contamination should not impact transcript level quantification.

To obtain a measure of relative transcript abundances within cell types, we determined the fragments per kilobase per million mapped reads (FPKM) values for each transcript in a hybrid *L. mexicana –M. musculus*-genome. Subsequently we separated the obtained quantifications for each species to consider the *L. mexicana* transcriptome on its own. The FPKM values for the AMA, i.e. mixed species samples, were consistently lower than those of PRO and AXA, likely due to the presence of ~40-80% mouse mRNA in the sample,

skewing the values due to the total library-size normalisation in performed for FPKM. Therefore, FPKM values without any further normalisation are unsuitable for differential expression testing. In all samples however, transcript abundances were detected over 4-5 orders of magnitude.

There is considerable overlap between the highest expressed transcript in each cell type, and the types of transcripts found in amongst those (e.g. Histones, ribosomal proteins, heat-shock proteins) agree well with findings in *L. major* (Rastrojo et al. 2013). Amongst the highest expressed transcripts we find novel transcript discovered by SLAS and PAS mapping, of which some are e.g. ribosomal proteins, but others have no obvious function and warrant further investigation. Overall, we find between 5-7 % of sequencing reads mapping to novel transcripts, showing that they do form a considerable part of the global transcriptome.

Comparison in PRO and AXA of the highest expressed transcripts with the most abundant proteins as determined by mass-spectrometry showed very little correlation between the transcriptomic and proteomic data sets. This is however in no way surprising considering the well-established and substantial degree of translational regulation in *Leishmania spp.* and other trypanosomatids (e.g. (Lahav et al. 2011; Vasquez et al. 2014)). Equally, these comparisons are technically difficult as the sampling of peptides from a *gemisch* of tryptic digests with a plethora of different chemical characteristics and the potential of post-translational modifications will harbour different biases than the sequencing of nucleotides.

The only exceptions to these observations are the very highest expressed transcripts, which are also some of the most abundant proteins, however this a comparatively small subset.

Turning to the analyses of differential expression of genes, we employed a method for calling differentially expressed genes that did not rely on FPKM values, but instead used

median normalised counts per gene (Dillies et al. 2013). Using this method we were able to compare count data from different library sizes, which was necessary as counts in AMA were consistently lower than in AXA and PRO, without having the drawbacks of total library size normalisations: The presence of highly and at the same time differentially expressed genes can lead to skewing of the global distribution of expression values in one sample relative to the other, impacting accurate differential expression testing (Dillies et al. 2013).

In the study presented herein, differentially expressed transcripts were defined solely based on statistical cut-offs, i.e. $p<=0.05$. As a result, transcripts with sometimes as little as a 1.5-fold change in transcript abundance may be considered differentially expressed. In some previous studies (e.g. (Siegel et al. 2010)), fold-change-cut-offs were also implemented as one may ask the question what biological significance a 1.3-fold differential abundance of a transcript can have?

There are three reasons why I would argue that a statistical cut-off is sufficient and indeed the only appropriate threshold in the present study. First, without absolute knowledge of all the reaction equilibria in a cell, it is impossible know whether or not a small increase in amount could push a transcript level above a given threshold and as a result lead to activation of a particular pathway or process. Second, we are looking at a dynamic process, differentiation, where transcript may be transiently, expressed, or only change expression after a given time post-induction of differentiation (Saxena et al. 2007). As a result, a small, but consistent fold change may be an indicator of a much bigger change that has already occurred or will occur. The third argument is regarding the degree of translational control observed in *Leishmania* (Lahav et al. 2011; Tsigankov et al. 2012). One has to entertain the possibility of differential translational efficiencies of single transcripts between promastigotes and amastigotes. Whilst a transcript may only be 1.5-fold differentially abundant, if this is combined with 2-fold increased translational efficiency, the net

outcome may indeed be biologically significant. Which brings us to the metaphorical elephant in the room: We have to concede that a down-regulated transcript could have increased translational efficiency, nullifying any observed differential regulation, or indeed promoting increased protein production despite down-regulation of the transcripts or vice-versa. Worryingly, such cases are not hypothetical, but have been shown using a combination of microarray and mass-spectrometry (Lahav et al. 2011). Only an approach such as ribosome profiling could possibly identify these cases without facing the drawbacks of combined transcriptomic and proteomic studies as I described above.

Differential transcript expression analysis comparing PRO and AMA reveals that 41.8% of genes are differentially expressed. This is by far more than the 2.9-12.5% (Leifso et al. 2007b; Rochette et al. 2009) of transcripts detected to be differentially expressed in numerous microarray studies. Only the study of Almeida *et al.* (Almeida et al. 2004) indicated around 35% of transcripts to be differentially expressed, albeit this study may suffer from artefacts derived from the cDNA-based generation of hybridisation platform and the lack of multiple-testing correction. However, I do not think the results presented here are over-estimating the number of differentially expressed transcripts given the ramifications of what constitutes a differentially expressed transcript of biological significance as described above. Moreover, the number of differentially expressed genes correlates well with RNA-seq based findings in *T. brucei* by Nilsson *et al.* (Nilsson et al. 2010) who found around 40% of genes to be differentially expressed in at least one of either long slender bloodstream forms, stumpy forms or procyclic forms. Crucially, only a p-value cut off was used in this study, making their findings, in that respect, comparable to ours.

Differential expression testing also revealed that 23.7% of transcripts are differentially expressed between PRO and AXA, whist 13.5% of transcripts are differentially expressed

between AMA and AXA. Whilst this shows us that AMA and AXA are quite different, it also shows us that, in terms of the number of differentially expressed genes, AXA lie between AMA and PRO, albeit closer to AMA than to PRO. The latter aspect lies in contrast to findings by Rochette *et al.* (Rochette et al. 2009) who proposed that AXA are more similar to PRO than to AMA based on the percentage of genes differentially expressed between PRO and AMA versus PRO and AXA.

Transcripts found to be differentially expressed ($p<0.05$) between promastigotes and the amastigote cell-types analysed are largely consistent with previous reports of differential expression. Studies providing disagreeing or inconclusive comparisons with our study were generally performed in a *Leishmania* species other than *L. mexicana*. As I showed in Section 3.3.5.2, the differential expression of well-established developmentally regulated transcripts are reflected in our study. The overriding agreement of findings however speaks for the validity of the approaches taken in this study.

GO-term and Pathway analyses as well as PFAM-domain enrichment analyses largely recapitulate previous findings with regard to lack of flagellar motility in amastigotes as well as metabolic shift away from glycolysis and reduced translational activity during promastigote-to-amastigote differentiation (Lahav et al. 2011; Tsigankov et al. 2012).

Comparison of AXA against AMA showed that AXA express higher levels of transcripts required for assembly of nucleosomes and DNA replication. Importantly, various histones are enriched in AXA and PRO compared to AMA, indicating that this is a promastigote-like signature, suggesting a reduced proliferation of AMA, compared to AXA, at the stage of intracellular differentiation analysed. We find amastins enriched in AMA and AXA compared to PRO, albeit also in AMA compared to AXA, indicating that both amastigote cell types are inducing expression of these, but AMA with more intensity than AXA. The expression of amastins in the amastigote stage, as well as generally increased expression of predicted transmembrane proteins in amastigote cell types compared to promastigotes

have been reported in the past (Rochette et al. 2008). So, *en gros*, these analyses are not contributing surprising new insights into the biology of *Leishmania spp,* but provide important validation of this data set. This may also be a function of the quality of e.g. GO-term annotation, which is not as far developed for *Leishmania spp.* as maybe for other model organisms. This will particularly concern amastigote biology, which is not found in other common model organisms and therefore will not have been part of functional annotations. That notwithstanding, improved functional annotation of the *L. mexicana* genome could in the future be combined with these data-sets to provide insight into the biology of *Leishmania spp.* we are currently still prevented from gaining.

However, new insights gained through enrichment analyses are perhaps not the most important contributions this study could have made to our understanding of *Leishmania spp.* biology. Instead, the unprecedentedly large catalogue of transcript found to be differentially expressed forms a reliable resource for other analyses, be they computational or experimental. This is particularly exemplified by two important findings.

First, we find a significant overrepresentation of the novel transcripts identified in Chapter 2 amongst the transcripts preferentially expressed in AMA compared to PRO. Taken together with their strong conservation amongst *Leishmania spp.*, with a possible origin upon divergence of *Leishmania spp.* from other trypanosomatids, I propose that some of these novel proteins may function to enable intracellular parasitism and possibly contain crucial virulence factors. Of course, biochemical approaches will be required to establish the existence and individual roles of these proteins, but considering that these predominantly small proteins have been categorically ignored in previous genome annotations and therefore in many studies into virulence, we are looking at a great opportunity to discover exciting and novel biology.

Second, looking at the distribution of differentially expressed genes across chromosomes, we find chromosome LmxM.30 enriched for transcripts preferentially expressed in AMA

and depleted of transcripts preferentially expressed in PRO. Combining this observation with reports of constitutive supernumerousness, mainly tetrasomy (Mannaert et al. 2012; Sterkers et al. 2012), of orthologous chromosomes in other *Leishmania spp.* gives tantalising indication of an important role of these chromosomes in *Leishmania*-amastigote biology. Comparison of the gene content of syntenic chromosomes in kinetoplastids that are non-pathogenic to vertebrates, e.g. *Crithidia spp.* or *Phytomonas spp.*, may reveal particular genes present on LmxM.30 of importance in establishing infection. The peculiar biology of LmxM.30, and its orthologous chromosomes in other *Leishmania spp.* certainly warrant further investigation.

Comparison between PRO and AXA of differential transcript expression and differential protein abundances determined by mass-spectrometry showed poor correlation. Similar to the comparison of transcript and protein levels within a particular cell type, correlation improves only amongst the very extreme observations, in this case the transcript with the lowest p-values for their differential transcript abundance. As the experimental design was not initially aimed at addressing this type of comparison, and the proteomic study was not performed in replicates, I encourage care when trying to interpret these results. In the light of the extensive translational control found in *Leishmania spp.*, especially after the first approximately 15 h of differentiation (Lahav et al. 2011), a clear-cut linear correlation between transcript and protein levels is highly unlikely. But an experimental design where both protein and RNA were extracted from the same cultures in addition to replicates of the proteomic design could have resulted in more comparable findings.

5.4 Are axenic amastigotes a good model *in lieu* of intracellular amastigotes?

Whether AXA and AMA are the same cell type or whether AXA display only some of the characteristics of AMA has been a long-standing question. Clearly, morphological similarities like the small, rounded cell shape and lack of motile flagellum, exist, (Bates 1994), however gene expression studies have questioned whether AXA are more similar

to PRO or AMA in other aspects (Holzer, McMaster, and Forney 2006; Rochette et al. 2009). Metabolically, *Leishmania spp.* appear to undergo similar transitions from glycolysis to amino-acid catabolism and beta-oxidation in axenic as in intracellular differentiation models (Rochette et al. 2009; Lahav et al. 2011; Tsigankov et al. 2012). However, active synthesis of long-chain fatty acyl-groups has been put forward as crucial metabolic difference, between AXA and AMA (Rochette et al. 2009) , due to the rich nutrient availability in axenic differentiation media as opposed to the sugar-poor environment of the parasitophorous vacuole.

The gene expression profiles obtained in this study correlate with the morphological and metabolic adaptations previously reported for amastigotes. Cell surface adaptations such as the expression of amastins and generally increased expression of transmembrane proteins previously reported are also reflected in my data. However, these also reveal a very important pattern: Axenic amastigotes may show the same trend of differential regulation as intracellular amastigotes, but the magnitude of the difference observed is bigger in AMA than in AXA. Notably, amastins are actually significantly differentially expressed between AMA and AXA, with higher expression in AMA.

Furthermore, the differential expression of nucleosome components in AXA and PRO compared to AMA, places AXA away from an AMA-typicatranscriptomic signature.

A final piece of evidence to integrate into this are the plain percentages of differentially expressed genes, where AXA and PRO (23.7% DE) are more similar to each other than AMA and PRO (40.8% DE).

We may therefore conclude that the 24 h axenic amastigote is an intermediate form between the promastigote and the 24 h intracellular amastigote. Whether it is closer to PRO or to AMA depends on the aspect of biology that is being considered. In my data, in terms of morphology, metabolism and expression of surface proteins it is situated closer to AMA, whilst in terms of cell-replicative biology it is closer to PRO.

The question now begs, whether this is just a transient difference due to different kinetics of differentiation, or whether this is a terminal difference between the axenic system and intracellular amastigote model. Rochette (Rochette et al. 2009) described axenic amastigotes as "stuck" in an early stage of differentiation but to fully address this question, further transcriptomic timepoints comparing AMA and AXA would be required.

I think both cases apply. The current axenic differentiation media are rich media, providing a wealth of nutrients along, of course, with the two main differentiation signals, higher temperature and low pH. However, iron deprivation, reactive oxygen species and purine starvation are also emerging as potent differentiation signals (Mittra et al. 2013; Martin et al. 2014). The cumulative effect of all differentiation signals could be key to inducing the same kinetics of differentiation *in vitro* and *in vivo*. It is however, tempting to speculate that direct interaction of parasite receptors with host-cell ligands relay signals to the parasite that fine-tune particular aspects of parasite biology, enabling the parasites to respond to changing responses of the host-cell. As a result, not every dynamic aspect of amastigote biology may be replicable in axenic systems.

Therefore, depending on the question that is being addressed, AXA can be a good model, when one considers the appropriate caveats, but AXA are unlikely to ever fully replace intracellular amastigotes in experiments.

5.4 Chapter 4 – Identification, bioinformatic characterisation and sub-cellular localisation of amastigote-upregulated proteins

The identification of proteins upregulated in the amastigote stage, with a particular focus on putative cell surface proteins started off with the pool of transcripts preferentially expressed in AMA compared to PRO. Subsequently a set of selection criteria were applied before the remaining candidate genes were tagged with GFP in promastigote form *L. mexicana* and their subcellular localisation analysed in PRO and AXA.

At the time of the choosing the tagging candidates, only a preliminary data set, generated using CDS-models only and using now superseded software tools (see Section 4.3.1), was available and therefore the pool of transcripts preferentially expressed in AMA vs. PRO was smaller than in the final data set. Therefore, using the final data-set, more transcripts could have qualified for selection, however comparison of the differential expression data set revealed that no transcript chosen using the preliminary data would have been excluded in the final data-set.

The selection criteria were aimed at identifying novel biology, so I sought to avoid proteins with existing annotation. One could argue that even with an existing annotation (e.g. a transporter) novel biology can be discovered, however with the wealth of hypothetical genes of unknown function in the genome of *Leishmania mexicana*, pursuing a screen of the wholly unknown is just as warranted.

Of course, not all of the proteins chosen were completely enigmatic, e.g. LmxM.03.0380 and LmxM.03.0400, which feature a PFAM domain found in transporters involved in yeast acid-sensitivity. Considering that low pH is an important trigger for promastigote-to-amastigote differentiation I opted to retain these proteins in the screen.

All of the proteins were tagged on their C-terminus with eGFP using pLENT vectors, which were shown to permit tagging of surface transmembrane proteins using the example of the *Leishmania mexicana* glucose transporter 2 (LmGT2), which was previously reported to localise solely to the pellicular membrane of *L. mexicana* and be excluded from the flagellar and flagellar pocket membrane (Tran et al. 2012). All of the 26 proteins that satisfied the selection criteria were tagged, however for technical reasons only 22 were analysed in PRO and AXA. Even though the selection was initially based on differential expression data between PRO and AMA, the microscopic analysis was performed in AXA for practical reasons and because, as discussed above, they appear to represent a good model for amastin regulation and show increased expression of transmembrane proteins

like AMA, suggesting that the expression of cell surface proteins, which the screen was aimed at identifying, may be similar in AXA and AMA.

Of the 22 fusion proteins only three showed the initially desired localisation, i.e. a surface localisation. Most (15) appeared to localise to intracellular compartments of the parasite, albeit without co-staining with markers of these compartments like mitotracker for the mitochondrion (c.f. (Chazotte 2011)) or anti-BIP antibody staining to label the ER (Bangs et al. 1993), it is not possible to say with certainty to which precise compartment the fusion proteins localise. Based on the data presented herein, eight fusion proteins possibly localise to the mitochondrion or ER and seven possibly to the Golgi or other another less defined vesicular compartment. Five showed no, or only very diffuse signal preventing precise localisation of the fusion protein, albeit it is not sure if fusion protein is actually being made. Indeed most of the fluorescent signals were very faint compared to the native fluorescence of untagged wild-type controls. Maybe the tagging screen could have been improved by incorporating mass-spectrometric evidence during the selection of target proteins: Whilst the correlation between transcriptomic and mass-spectrometric data was poor, mass-spectrometric detection could have provided evidence for actual expression of the protein. Searching the mass-spectrometric data set obtained in Section 2.3.7 for peptides corresponding to proteins that were part of the tagging screen revealed that only three proteins were detected, LmxM.22.0410 and LmxM26.1460 (in AXA only) and LmxM.27.1930 (with a 2.4-fold enrichment in AXA compared to PRO). This is a surprisingly low number considering that around a third of the total proteome was detected mass-spectrometrically. This may be explained by the protocol used to prepare the protein samples (1% Sodiumdeoxycholate in 8 M Urea), which was not specifically optimised to solubilise membrane proteins, but instead was aimed at simply disrupting cells. As a result these proteins may have precipitated preferentially during sample preparation. Moreover, no de-glycosylation steps were preformed which could have

increased detection of peptides derived from extracellular moieties of cell-surface proteins.

The two fusion proteins that were chosen for more in-depth analyses due to their localisation were LmxM.16.0500:eGFP and LmxM.09.1330:eGFP.

LmxM.16.0500 is a member of a small, divergent, δ-amastin-derived gene family featuring four transmembrane domains. Whilst its relationship to amastins superficially violates one of the selection criteria of the candidate genes (no amastins), this relationship only became apparent after in-depth analysis of the entire gene family in several *Leishmania spp.* suggesting that the LmxM.16.0500 family are not simply more amastins, but a divergent group. Moreover, amastins are usually found in tandem arrays (Wu et al. 2000; Rochette et al. 2005), often interspersed by tuzins (Jackson 2010). We do not see this arrangement in the LmxM.16.0500 family.

The LmxM.16.0500 transcript has 2.8- fold higher expression in AMA compared to PRO but equal (1.1-fold higher) expression in AXA compared to PRO. It ranks amongst the top percentile of highest expressed genes in AMA being the 20th highest expressed gene with 700 FPKM, 65 positions higher than the next amastin (LmxM.33.1720c, 354 FPKM), again indicating a special position for this gene in the biology of *L. mexicana* amastigotes. In PRO, the LmxM.16.0500:eGFP fusion protein highlights the surface membrane but very notably appears to be released from the cell surface either via streamers and possibly vesicles from the anterior end of the cell, resulting in foci of fluorescence along the flagellum. It is not clear whether this is an artefact of the possible over-expression of the protein in the promastigote stage or whether LmxM.16.0500 is usually found in streamers. Equally, it is not clear whether the intensity of streamer release seen highlighted by the fluorescence of LmxM.16.0500:eGFP is normal and just not as readily observed in wild-type cells. In AXA, the fluorescence signal of the fusion is only faintly discernible, but does not outline the cell body as one would expect based on previous characterisations of

amastins (Rochette et al. 2005). It is possible that the transcript stability or translational efficiency conferred by the 3' UTR introduced in by the tagging method (*C. fasciculata* PGKA), is not sufficiently high to see expression of the protein under elevated temperature and acidic pH as AXA are exposed to.

Making speculations about function based on localisation would be a bold move and considering the yet enigmatic biological function, of amastins, one cannot gain insight into the function of LmxM.16.0500 based on its shared origins with amastins. A role in signalling has been proposed for amastins (Jackson 2010) whereby amastins are able to bind ligands at the host-parasite interface and relay signals to intracellular transducers. The apparent release of LmxM.16.0500, should this not be a technical artefact, would be difficult to reconcile with such a function. However, one could speculate that if LmxM.16.0500 were released from phagocytosed promastigotes, LmxM.16.0500 would find itself in the parasitophorous vacuole, where, upon fusion of the lipid derived vesicle it is contained in (streamers, vesicles) with the membrane of the parasitophorous vacuole, LmxM.16.0500 could effectuate its function in that localisation. This could e.g. be to provide a ligand for other parasite-surface molecules, allowing *Leishmania spp.* amastigotes to adhere to the membrane of the parasitophorous vacuole or by acting as a one- or bi-directional signal transducer.

The release of material from *Leishmania spp.* cells in the form of filamentous structures has in the past been reported for polymerised secreted acid phosphatase in a non-covalent complex with high-molecular weight phosphoglycans (Ilg et al. 1991; Stierhof et al. 1994). The appearance of these filamentous secretions by immune-fluorescence microscopy released from the flagellar pocket (Stierhof et al. 1994), superficially resemble the fluorescent material released from promastigotes expressing LmxM.16.0500:eGFP, however they are not thought to contain membranous material and did not appear to detach from the posterior end of the cell as observed for the structures highlighted by LmxM.16.0500:eGFP. The four transmembrane domains present in LmxM.16.0500 would

suggest that it ought to be present in a membranous phase, however this is just an assumption and one has to entertain the possibility that LmxM.16.0500:eGFP could be released in a structure not composed of membranous material, but instead of another polymer, e.g. a phosphoglycan aggregate. Structural analysis of the structures containing LmxM.16.0500:eGFP by electron microscopy will shed further insight into the nature of the structure bearing LmxM.16.0500:eGFP and I understand such investigations have been initiated.

Infection of macrophages with *L.mexicana* promastigotes expressing LmxM.16.0500:eGFP could provide crucial insight into the fate of the fusion protein during infection which should, considering the intensity of the fluorescent signals observed during live-cell microscopy, be easily tractable.

The only way to truly investigate the native expression of LmxM.16.0500 would be to raise an antibody against it, circumventing the requirement of mutating the LmxM.16.0500-locus. A middle ground between latter approach and the approach chosen in this study here, would be tag LmxM.16.0500 with a small epitope tag, such as a c-MYC tag, whilst preserving the endogenous 3'UTR. This may prevent artefacts derived from the presence of a large protein tag as well as dysregulated expression.

LmM.09.1330 is a two-transmembrane protein with a large extracellular domain. Its transcript has 2.7-fold higher expression in AMA than PRO and 1.8-fold higher expression in AXA than PRO. LmxM.09.1330 is conserved in *Leishmania spp.* and some *Trypanosoma* species such as *T. cruzi* and *T. grayi* but absent from *T. brucei* and organisms outside of kinetoplastids. No extra copies are found in the genome of *L. mexicana*. The presence in *T. cruzi* made this an interesting tagging candidate due to the fact that *T. cruzi*, too, has an intracellular amastigote stage. Discovery of the sequence in the genome of the extracellular, crocodile-infecting *T. grayi*, published after the tagging of this protein, cast a small shadow of doubt on whether the pattern of conservation would truly correspond to

a protein involved in amastigote biology. However, the close evolutionary position of *T. cruzi* and *T. grayi* (see Figure 1.3) may indicate that this finding is not straightforward to interpret.

No PFAM-domains are found in the proteins, however the C-terminal region of the protein is highly conserved between species, whilst some heterogeneity is found at the N-terminus of the protein sequence, which may have functional implications that may warrant further investigations.

Upon C-terminal tagging with eGFP using the pLENT vector, only a very faint signal can be detected in promastigotes by immuno-fluorescence. The signal highlights the cell body and traces along the flagellum, which, along with the presence of its two predicted transmembrane domains suggests that it is present on the cell surface membrane, albeit the microscopic evidence for this is inconclusive, I suspect primarily due it the faintness of the signal observed. Moreover, Western-blot analysis indicates that the protein may be post-translationally modified, raising the possibility of it being glycosylated, which would furthermore indicate surface localisation. Upon axenic differentiation a very different signal is seen, which localises anterior to the kinetoplast in what appears to be a ring structure. This is consistent with LmxM.09.1330:eGFP localising to the flagellar pocket membrane. Our understanding of amastigote biology is still rather limited, and it is safe to say that our understanding of the amastigote flagellar pocket suffers the same limitations, if not more. Indeed, to my knowledge, only the Trp-Asp protein LAWD (Campbell, Popov, and Soong 2004) has in the past in amastigotes been shown to specifically localise to the flagellar pocket. Whether or not this is specific to amastigotes or may also be the case for promastigotes is no known. Therefore LmxM.09.1330 could be the first report of a protein specific to the amastigote flagellar pocket, albeit only assessed in an axenic system. By reciprocal best blast the LAWD protein corresponds to LmxM.06.0030, which is not

predicted to have transmembrane domains (prediction using TMHMM 2.0), making these two proteins localising to the amastigote flagellar pocket structurally distinct.

To investigate the composition and biology of the amastigote flagellar pocket further, LmxM.09.1330:eGFP may be used as a handle using which flagellar-pocket derived membrane fractions from amastigotes may be purified, enabling characterisation of their composition by mass-spectrometry. Moreover, the differential localisation of LmxM.09.1330:eGFP between PRO and AXA suggests that a mechanism restricting LmxM.09.1330 localisation is activated during the first 24 h of axenic differentiation that may be identifiable by its interaction with LmxM.09.1330. Identification of components of such a complex could provide insight into mechanisms differentially regulating surface membrane population in promastigotes and amastigotes.

Again, it is difficult to make predictions with regard to the possible function of LmxM.09.1330 based on localisation of the LmxM.09.1330:eGFP fusion protein. The flagellar pocket has traditionally been viewed as the sole site of endo- and exocytosis in trypanosomatids, albeit, as we have seen with LmxM.16.0500:eGFP, other mechanisms of release of membrane material may be seen under certain circumstances. LmxM.09.1330 could conceivably act as a receptor in the flagellar pocket, mediating uptake of a ligand, either directly or by stimulating endocytosis. Latter function could be tested by deletion of LmxM.09.1330 and quantification of fluorescently labelled lectin uptake from the flagellar pocket into the endosomal system. In the past, a 46 kDa protein binding to haemoglobin and mediating uptake via endocytosis was localised to the flagellar pocket of *L. donovani* promastigotes, showing proteins acting as receptors and influencing endocytic activity exist, of course bearing in mind this was a protein expressed in promastigotes (Sengupta et al. 1999). A direct role in solute uptake, without knowledge of what this solute may prove more difficult. However, injection of an mRNA encoding LmxM.09.1330 into *Xenopus laevis* oocytes, followed by electrophysiological comparison to control oocytes

could reveal perturbation of electrochemical potentials, thus indicating expression of a surface solute transporter. Subsequently, mass-spectrometric comparison of injected and control oocytes could reveal enrichment (or depletion!) of a particular solute.

Alternatively, LmxM.09.1330 could act to relay the signal of binding a ligand to local effectors, triggering release of vesicles into the flagellar-pocket and beyond. It is of course conceivable that LmxM.09.1330 has a more global signalling role. An argument against this would be that examples of globally signalling surface proteins have been found on the entire cell surface of trypanosomatids, like the PAD-proteins in *T. brucei* mediating stumpy-form formation triggered by citrate/*cis*-aconitate (Dean et al. 2009). However the ligands for these are small molecules able to permeate the protective surface coating of the parasite to interact with receptors hidden underneath. The flagellar pocket provides a protected environment where surface molecules otherwise hidden from the surroundings may be exposed. Therefore, if a receptor for a larger-molecule, like a soluble protein were required, the flagellar pocket would constitute a safe location for it, speaking for the possibility that LmxM.09.1330 could have a role in global signalling, whilst having to be restricted to the flagellar pocket.

To test whether LmxM.09.1330 binds a particular protein-ligand, either extracellular or intracellular, a carefully designed immuno-precipitation experiment followed by proteomic analysis of the interacting partners could be devised.

Finally, one has to entertain the possibility that LmxM.09.1330 has a purely structural role. The elevated temperature amastigotes experience compared to promastigotes, will also alter the chemical properties of membranes. The flagellar pocket membrane has been shown to have a distinct composition to the flagellar and pellicular membrane and may require specific accessory factors supporting its morphology and function in the amastigote stage. Such a function may be investigated by deleting LmxM.09.1330 in PRO and analysing flagellar pocket morphology in AXA or AMA by transmission electron microscopy.

In summary, a substantial amount of experimental work will be required to fully characterise LmxM.16.0500 and LmxM.09.1330, but the evidence presented herein give tantalising indications that that further investigation of these proteins will reveal novel, interesting and potentially important biology.

5.6 Synthesis

In this work, I have presented the first analysis of the transcriptome of three *L. mexicana* cell types, namely promastigotes, 24 h axenic amastigotes and 24 h intracellular amastigotes using both conventional and dual-RNA-seq methods. In this, we mapped the positions of spliced-leader acceptor sites as well as poly-adenylation sites, permitting the first genome wide description of transcript dimensions in *L. mexicana.* Additionally, these data were used to refine the existing genome annotation, resulting in the identification of 936 novel genes bringing the total gene number in *L. mexicana* to 9169 genes, as well proposing alterations to existing gene models. Using the transcript dimensions we have been able to establish genome wide expression profiles for promastigotes as well as 24 h axenic and intracellular amastigotes. From these, differentially expressed transcripts were identified in pairwise comparisons, resulting in an unprecedentedly large catalogue of differentially expressed transcripts, which correlated well to previous reports. These data will form a valuable resource for future investigations into *Leishmania spp.* and in particular into amastigote biology.

Moreover, these data were used to search for putative surface membrane proteins preferentially expressed in amastigotes resulting in the identification of two remarkable proteins.

The transmembrane protein LmxM.16.0500 is a highly expressed member of a small divergent amastin-derived gene family that, under the here presented experimental conditions, appeared to be released from promastigote cells. Pending further confirmation of this phenomenon, characterisation of this protein may provide novel insight into the interaction of *Leishmania spp.* with its hosts.

The transmembrane protein LmxM.09.1330 appears to be selectively restricted to the flagellar pocket in 24 h axenic amastigotes, constituting the first report of an amastigote specific flagellar pocket marker. A wealth of functions are conceivable for this protein, which will undoubtedly provide fascinating insight into the biology of the parasite, both through elucidation of its function, but also as a tool for the investigation of the properties of the amastigote flagellar pocket.

Finally, the identification of the two aforementioned proteins and the promises they bear, only serves to support the potential of the here generated transcriptomic data sets and it will be exciting to see where else this resource will take us.

6 – Materials and Methods

6.1 Generation of BMDMs

Fumuri, humeri and tibiae were dissected from C57BL/6 mice, ethanol sterilised and the bone marrow flushed out with DMEM (high glucose GlutaMAX DMEM (Dulbecco's modified eagle medium) (4.5 g/l glucose) (Life technologies 10566-016) supplemented with 10 % FCS (Life Technologies 10500-064)). Bone marrow was homogenised by passage through a 70 µm cell strainer, centrifuged (800 g, 10 min) and resuspended in 100 µl/bone erythrocyte lysis buffer; 155 mM NH4Cl, 10 mM KHCO3, 125 µM EDTA, and incubated at room temperature for 2 min. Unlysed cells were washed twice by centrifugation (800 g, 10 min) and resuspension in 30 ml DMEM. Bone marrow cells were resuspended in MCSF-DMEM; 70 % DMEM and 20 % L929 conditioned DMEM and 10 % FCS. L929 is a murine fibrosarcoma cell line producing macrophage colony stimulating factor (MCSF) which induces differentiation of monocytes into bone marrow derived macrophages.

Bone marrow cells were differentiated into macrophages over 5 days, then cultured for 2 days in DMEM prior to further procedures.

6.2 Cell Culture, Infection Protocol and RNA-extraction

For macrophage infection *Leishmania mexicana* promastigotes (WHO strain MNYC/BZ/62/M379), grown in M199 complete medium (Invitrogen) with 10% FCS (Life Technologies 10500-064) were seeded at $1x10^5$ cells/ml and grown into stationary phase over 5 days. *Leishmania* cells were harvested by centrifugation (800g, 5 min) and added to BMDMs adherent to a petri-dish at a multiplicity of infection of 20 (MOI:20). After 2h at 34°C the medium was discarded and adherent cells washed three times in medium.

For axenic differentiation stationary phase *L. mexicana* were transferred into differentiation medium (Schneider's Drosophila Medium (Invitrogen) with 20% FCS , 25mM MES, pH 5.5).

After 22h cells were washed in phosphate buffered saline (PBS, 137mM NaCl, 2.7mM KCl, 10mM Na_2HPO_4, 2mM KH_2PO_4, pH 7.4) and RNA extracted using QIAgen RNeasy Mini Kit as per manufacturer's instructions. Samples AMA, AXA and PRO 2-3 were subsequently treated using Ambion Turbo DNAase kit (cat. num. AM1907). Purity was determined using a NanoDrop 1000 spectrophotometer and RNA integrity was assessed by evaluating ribosomal RNA integrity using a Bioanalyzer2100 (Agilent Technologies), both as per manufacturer's instructions.

6.3 Flow Cytometry

BMDMs were harvested, washed twice in PBS and stained with anti-mouse F4/80 (Alexa Fluor488, clone BM8), anti-mouse MAC-1 (Alexa Fluor488, clone M1/70), anti-mouse GR-1 (Alexa Fluor647, clone RB6-8C5) and isotype controls (Alexa Fluor488 & 647 IgG2b, κ) in PBS with 2% FCS, 0.1% BSA and 0.1% NaN_3 for 20min. Cells were washed in PBS three times prior to flow-cytometric analysis.

Leishmania cells were fixed in 4% formaldehyde for 10 min prior to flow-cytometry. All flow-cytometry was performed on a BD FACSCalibur flow-cytometer (Beckton Dickinson).

6.4 Light Microscopy

Macrophages were cultured on microscopy coverslips, fixed (4% formaldehyde, 10min) and stored in methanol (-20°C). Cells were rehydrated in PBS and treated with 400µg/ml Rnase A in PBS for 30min. Cells were washed in PBS, stained in 40µg/ml propidium iodide for 5 min, washed in PBS and mounted on slides in

mounting medium (1% 1,4-diazabicyclo[2.2.2]octane, 90% Glycerol, 10% 50mM Na_2HPO_4).

For live-cell microscopy *L. mexicana* cells were washed 3x in PBS and settled on glass slides in PBS with 1 µg/ml Hoechst 33342 DNA stain (Invitrogen H3570).

For immunofluorescence microscopy, cells were settled on glass slides in PBS and fixed in 4 % formaldehyde for 10 min. Slides were blocked in 1 mg/m Bovine serum albumin (Sigma A9418) in PBS for 15 min and stained with Invitrogen rabbit anti-GFP antibody (Invitrogen A 11122), 1:200 dilution from manufacturer's stock in 1 mg/ml BSA in PBS) for 30 min followed by three PBS washes and staining with goat anti-rabbit IgG Alexa Fluor 488 conjugated antibody (Invitrogen A11008, 1:200 dilution from manufacturer's stock in 1 mg/ml BSA in PBS) for 30 min. After three PBS washed slides were mounted in mounting medium containing 1 µg/ml Hoechst 33342 DNA stain (Invitrogen H3570).

Microscopy was performed on a Zeiss Axioplan 2 microscope with a Photometrics CoolSNAP HQ camera. Images were processed using ImageJ software.

6.5 cDNA library preparation & sequencing

cDNA libraries were prepared and sequenced at the Beijing Genomics Institute (Shenzhen, China). In brief, RNA was positively selected for polyadenylated species and fragmented. Fragments between 200 – 400 nt were gel-purified and cDNA synthesis performed using Illumina TruSeq V2 Kit. First strand cDNA synthesis performed using random-hexamer primers for Library 1 and 5'-T15VN-3' primers (with V = any based but T and N= any base) for Library 2. Second strand synthesis was performed using random hexamer primers. Sequencing was performed on an Illumina Hiseq 2000 (Illumina, CA).

6.6 Quality filtering and mapping for SLAS and PAS mapping

SLAS and PAS mapping was performed as described in (Fiebig et al. 2014) using the SLaP Mapper program. In brief, sequencing reads from raw .fastq files were quality filtered using Trimmomatic (Bolger, Lohse, and Usadel 2014) using the following settings: Seed Mismatches =2, palindrome clip threshold = 10, leading strand = 10, sliding window = 5:10, minimum length =30. In all samples >99% of paired reads passed trimming, i.e. both read mates were still longer than 30 nt. Read mate-pairs were examined for overlap and where possible joined using fastiq-join utility (Aronsky 2011). To map SLAS, reads containing the last 12 nucleotides of the spliced-leader sequence in *L. mexicana* (i.e. TGT ACT TTA TTG) at their 5' end were identified, the spliced-leader sequence removed using Perl scripts and the remainder mapped to the genome using bowtie2, the 5' end of the mapped read constituting the SLAS. To map PAS, all reads featuring 5 or more A or T (i.e. the complement of the poly-A sequence) residues. Reads are split after the run of As or Ts and the non-poly(A/T) part mapped to the reference genome using bowtie2. The 5' end of the mapped read constitutes the poly-A site. All reads mapping to positions where the spliced-leader sequence or a poly-A run is present in the genome were filtered out to obtain *bona fide* SLAS and PAS.

6.7 SLAS-based gene prediction

Each SLAS detected was considered a possible primary splice-site for a transcript. The sequence between this SLAS and the next strand-appropriately downstream SLAS was scanned for all ATG-trinucleotides that could constitute possible translation start sites, and all TAA- , TAG- and TGA-trinucleotides that could constitute translation termination sites. For each possible start, it was queried

whether a potential in-frame stop codon exists at a distance of >75 nt, therefore constituting an ORF encoding a protein ≥ 25 AA. When the first such an ORF was detected, it was recorded in GFF format and the next SLAS analysed. When no such ORF could be detected for a SLAS, no data was recorded and the next SLAS considered.

All predicted ORFs (pORFs) that lie in the same genomic inter-SLAS space as an entire reference CDS were considered uORFs. All pORFs that lie within reference CDS which have been segmented by SLAS positions were manually curated to generate the most likely CDS model, considering functional annotations as well as conservation in other kinetoplastid species using information on TriTrypDB. When these analyses proved inconclusive, the reference annotation was favoured. pORFs with the same stop codon as a reference CDS but differing start codon were divided into two groups based on whether they propose N-terminal extension or truncation. All proposed extensions were accepted and subject to proteomic confirmation. Truncations were manually curated based on the splicing-pattern and frequency as well as sequencing read-coverage and referral to existing annotations and conservation between kinetoplastid species using information on TriTrypDB. Importantly, truncations were only accepted when a sole splice site was present or when a dominant cluster of splice sites was apparent. In ambiguous cases, preference was always given to the reference annotation. All resulting CDS were compiled into GFF format and compared to the reference annotation, to fill in CDS not captured by the SLAS-based CDS-annotation to generate a first complete CDS-library.

6.8 PAS based filtering of novel genes

Using the generated CDS-library, the SLaP Mapper program was used to assign SLAS and PAS to individual CDS. The resulting PAS positions were used to filter out possible false-positives from the detection of novel CDS. All novel CDS lacking PAS were removed and SLAS and PAS re-assigned using SLaP Mapper. Subsequently, all novel CDS with more than 10 % of PAS being CDS-internal were excluded and all CDS with up to 10% of PAS located CDS-internal were manually inspected and unlikely candidates removed, based on the distribution of PAS within the transcript: Tight clustering of PAS towards the 3' CDS end was deemed biologically plausible for a novel gene, whilst even distribution of PAS over the CDS was deemed implausible and removed. All SLAS and PAS were again re-assigned to remaining CDS. Finally, all novel CDS lying downstream of a gene featuring in the reference annotation without an assigned PAS were manually inspected whether these novel CDS lie within the likely 3'UTR of the gene from the reference annotation based on sequencing read-coverage. All novel CDS found to be in the likely 3'UTR of a gene from the reference annotation were excluded. SLAS and PAS re-assigned to the remaining CDS and transcript models generated from the most distal SLAS to the most distal PAS assigned to a CDS.

For visualisation, the Integrative Genome Browser (IGV) was used (Robinson et al. 2011; Thorvaldsdóttir, Robinson, and Mesirov 2013).

6.9 UTR processing

Extraction of UTRs from the *L. mexicana* genome annotation was performed by extracting the genomic segments from the 5' end of the transcript to the start of the CDS to generate the 5' UTR, and from the CDS end to the 3' end of the transcript to generate the 3' UTR. For the T. brucei data, SLAS and PAS data were obtained

from the supplementary material from Kolev *et al.* (Kolev et al. 2010) and 3' UTRs defined as the genomic segment from the most distal SLAS reported for a CDS to the start of the CDS, whilst 5' UTRs were defined as the genomic segment between the end of a CDS to the most distal PAS reported for that CDS. All analyses and plots were generated in R.

6.10 TM, SP, PFAM prediction

Transmembrane domains were predicted using TMHMM Server v. 2.0. Signal peptides were predicted using SignalP 3.0 based, with significant predictions based either or both Hidden-Markov-Model or Neural Network. PFAM domains were predicted online with significance-cut offs set at the Gathering Threshold for PFAM-A domains.

6.11 Nucleotide composition and secondary structure

Nucleotide compositions around splice sites were determined using custom R-scripts available in supplementary material. *T. brucei* SLAS data was obtained courtesy of Dr. Steven Kelly, Department of Plant Sciences, University of Oxford (unpublished data). Mean base-pairing probabilities were determined by Dr. Steven Kelly using RNAfold, part of the Vienna RNA folding package (Lorenz et al. 2011) and data visualised in GraphPad Prism.

6.12 Reciprocal Best Blast method

Reciprocal Best Blast (R.B.B.) (Hirsh and Fraser 2001) analyses were performed using BlastAll software run within an wrapping R-script. For queries against protein sequences, the a Blastx method was used for the out-going blast, the protein sequence corresponding to the highest ranking Blastx-hit queried against all transcript sequences in *L. mexicana* using a tBlastn method. When the same

transcript used in the initial query was the highest ranking return blast, a reciprocal best blast was detected.

For queries against genome sequences a tBlastx method was used in the initial query using the novel transcript sequence against the target genome. The sequence of the highest-ranking outgoing blast hit, was extracted and used in a tBlastx method against the all transcript sequences of *L. mexicana*. When the same transcript used in the initial query was the highest ranking return blast, a reciprocal best blast was detected.

For all R.B.B.-hits, the negative logarithm (base 10) of the e-value of the return Blast-result were recorded. Using a custom R-script a matrix of all R.B.B. results was generated and plotted using the heatmap() function from the "stats"-library in R, permitting for hierarchical clustering of transcripts based on conservation pattern using hclust() defaults, but retaining a manually determined order of organisms, based on the evolutionary relationship of kinetoplastids (see Figure 1.4). All genomes and proteomes used in these analyses were obtained from TriTrypDB v.7.0, for *L. braziliensis* "LbraziliensisMHOMBR75M2904" was used, for *T.* brucei "TbruceiTREU927" and *T. cruzi* "TcruziCLBrener-Esmeraldo-like" were used. Phytomonas EM1 and HART1 genomes are from (Porcel et al. 2014).

6.13 Three-frame PFAM prediction

Three frame translations of all novel transcripts were generated using the SeqinR R-package. All three translations were batch-submitted to the PFAM server and all PFAM domains recorded, with the Gathering Threshold used for PFAM-A domains.

6.14 Mass spectrometric analysis

PRO and AXA cultures were prepared as described above. Cells were washed 3x by centrifugation at 800 g for 7 min and resuspension in PBS for PRO and PBS plus custom protease inhibitor cocktail for AXA. Final protease inhibitor concentrations: Leupeptin hydrochloride (LS783, Sigma-Aldrich) 50 µM, E-64 (E3132, Sigma-Aldrich) 5µM).

Cells were lysed in ice-cold lysis buffer composed of 8M Urea (Sigma-Aldrich) in 125mM Tris (pH 6.8) with 1% Sodiumdeoxycholate and protease inhibitors at the final concentrations shown above. Lysis was performed on ice with 5 s vortexing every 30 s for 5 min. Protein concentrations of samples were determined using Thermo Scientific Pierce 660nm Protein Assay (Product # 22662) as per manufacturer's instructions.

AXA and PRO protein samples were submitted to the Central Proteomics Facility at the Sir William Dunn School of Pathology for mass-spectrometric analyses. In brief, detergent was removed precipitation using trifluoroacetic acid (0.5 % v/v final concentration) and subsequent centrifugation at 13,000 g for 10 mins. The supernatant was removed and the proteins contained in it were denatured in 8M Urea, 10 mM dithiothreitol and 10mM iodoacetamide. In-solution tryptic digests were performed at 10-fold excess of trypsin. Samples were desalted on a C18-column and injected into an HPLC-coupled QExactive mass-spectrometer. The obtained .mgf files were combined with .mgf files from sub-cellular fractionation experiments performed by Tom Beneke and François Demay in the Gluenz laboratory (Beneke, Demay & Gluenz, unpublished). Spectra were searched against custom data-bases in the Central Proteomics Facility Pipeline allowing for two missed tryptic cleavages with a precursor tolerance of 20 ppm, fragment tolerance

of 0.1 Da with fixed Carbamidomethyl and variable N-terminal acetylation and Oxidation (M). Quantitation tolerance was set to 0.02.The data from Paape *et al.* (Paape et al. 2010) was obtained as raw spectra analysed using the Central Proteomic Facility Pipeline. Search parameters were as described above only with the Fragment Tolerance increased to 0.5 Da. All peptide-data was exported from MASCOT (Koenig et al. 2008), all label free quantitation performed using (Trudgian et al. 2011).

6.15 N-terminal extension prediction and rendering to GFF

From the mass-spectrometric data for AXA and PRO samples, only peptides with a p-value ≥0.95 for correct identification and assignment protein sequence were considered for the identification of peptides corresponding to predicted N-terminal extensions. Peptides were converted into genomic coordinates by generating an alignment of the peptides with the protein sequence it was assigned to by Mascot (Koenig et al. 2008). This was performed using custom R-scripts. The positional off-set, counted as numbers of amino-acids, of the alignment site relative to the protein start was recorded and used to generate the genomic coordinate of the peptide by addition of the positional offset, converted to numbers of nucleotides, to the GFF coordinate of the CDS-start in a strand-appropriate manner. The end-coordinate of the peptide was generated by addition of the peptide length, converted to nucleotide numbers, to the peptide start coordinate in a strand-appropriate manner. These results were recorded in GFF format. Overlap of genomic coordinates for peptides with the genomic coordinates corresponding to the proposed extensions of CDS were assessed and quantified by using a custom

R-script. All custom scripts used in this analysis are provided in the Supplementary Material to this thesis.

6.16 in silico prediction of extension

In silico predictions of peptides corresponding to predicted N-terminal extensions of proteins were generated first by *in silico* digestion of all protein sequences proposed to be N-terminally extended using the "CleaveR" R-package (Enzyme = Trypsin, allowing for up to 2 missed cleavages), recording for each peptide, the gene accession number they were derived from. Peptides smaller than 7 AA, and peptides larger than 29 AA (95^{th} size percentile of peptides observed in real mass-spectrometric data used in this work) were discarded. All remaining peptides were converted into genomic coordinates and analysed for co-localisation with predicted N-terminal protein extension as described above.

6.17 Best-Consensus Reverse Blast method

A reciprocal best tBlastx analysis (R.B.tBx.) was performed using the transcript sequences of the novel transcripts. When an R.B.tBx. was returned, the genomic stretch in *L. mexicana* this corresponded to was recorded. For each novel transcript, the genomic stretch covered by at least 80 % of returned R.B.tBx.-hits was used in a Blastx analysis against a library containing translations of every ORF ≥ 25 AA in the novel transcript. The highest-ranking sequence in this analysis was proposed as the ORF constituting the most likely CDS.

6.18 Quality filtering, mapping and quantification

Low-quality reads were filtered from the sequencing data using Trimmomatic software (Bolger, Lohse, and Usadel 2014) using the following settings: Seed Mismatches =2, palindrome clip threshold = 10, leading strand = 10, sliding window = 5:10, minimum length =30. In all samples, at least 99.77 % of reads

passed quality filtering. A hybrid genome consisting of the protein coding transcripts of *L. mexicana* generated in Chapter 2 and transcripts of *Mus musculus* (Mus_musculus.GRCm38.75.cdna.all.fa) was generated. Sequencing were aligned to the hybrid-genome using RSEM v1.2.11 (Li and Dewey 2011) with the –Bowtie2 (Langmead and Salzberg 2012) aligner option enabled. Reads mapping to multiple sites are assigned according to the statistical model devised by (Li et al. 2010), whereby assignment of multi-mapping reads is influenced but surrounding sequencing read coverage. After aligning, the transcripts corresponding to *L. mexicana* and *Mus musculus* were separated *in silico.* FPKM values and read counts per gene were obtained from RSEM. Median normalisation of read counts was performed using a custom Perl script (courtesy of Dr. Steven Kelly), and differential expression testing performed using DESeq2 (Anders and Huber 2010) using default settings. False discovery rate was controlled according to Benjamini-Hochberg (Benjamini et al. 2001) with a p-value cut-off for significance set <0.05. All plots generated in R using "base"-library elements, whilst volcano-plots were generated using the ggplot2-package in R.

For the FPKM values for rRNAs and tRNAs, the raw data was quantified against the *L. mexicana* genome (v7) using RSEM (-bowtie2 enabled).

6.19 FPKM Saturation

FPKM-saturation was determined using a command pipeline written in R provided in the Supplementary Material. From the raw sequencing data files (.fq format) fractions of paired-end reads were extracted (5, 10, 20, 30, 40 ,50, 60, 70, 80, 90, 95, 100 %) and quantified against the *L. mexicana* genome (v7) using RSEM (-Bowtie2 enabled) to obtain FPKM values for each gene. From these data transcripts were grouped based on the size of their final FPKM, and the fraction of

transcripts at ±10 % of the FPKM at 100 % of the total sequencing read library determined. Results were plotted in GraphPad Prism.

6.20 Identification of amastins

Amastins were identified by batch-submitting all protein sequences of the *L. mexicana* genome to the PFAM server and obtaining a whole-genome PFAM-domain prediction. All proteins which, by Gathering Threshold, had a predicted "Amastin" PFAM domain, were considered amastins.

6.21 Enrichment analyses

Enrichment analyses were performed using the GOseq R-package (Young et al. 2010). For the length bias (Oshlack and Wakefield 2009) correction we used the transcript lengths and controlled the false-discovery rate according to Benjamini-Hochberg (Benjamini et al. 2001) setting a p-value cut-off for significance at $p < 0.05$. GO-term and Pathway maps were transferred from *L. major* from the LeishCyc data-base (Doyle et al. 2009). Signal-peptide maps were obtained using SignalP 3.0 server (using Hidden-Markov model and/or Neural Network detection), transmembrane domain maps were generated using TMHMM 2.0 server. PFAM domain map was generated by batch-submitting all protein sequences of the *L. mexicana* genome to the PFAM server and obtaining a whole-genome PFAM-domain prediction. For proteins with multiple detections of the same PFAM domain, only a single instance of the PFAM domain detection was considered.

6.22 Preliminary transcriptomic data analysis used in Chapter 4

The preliminary data-set used in Chapter 4 obtained as follows. Paired-end sequencing reads were quality-trimmed using FASTX toolkit (Goecks et al. 2010),

with a PHRED quality threshold of 20, discarding all reads <21 nt in length. Transcripts were aligned to a hybrid *L. mexicana* (v4.1) – *Mus musculus* (NCBI37.65) genome composed of annotated coding sequences only and quantified using RSEM (Li and Dewey 2011) (using default parameters). Read-count values for each gene were median normalised (see above) and differential expression testing was performed using DESeq1 (Anders and Huber 2010). Differentially expressed genes were those with a Benjamini-Hochberg corrected p-value <0.05 (Benjamini et al. 2001).

6.23 Generation of mutant cell lines

All mutant cell lines were generated using pLENT vectors (Figure 4.2). Primers used for each construct are given in Supplementary Table 6.1. For each primer the appropriate restriction enzyme is shown. Individual targeting fragments were amplified by polymerase chain reaction (PCR) from genomic DNA (gDNA) of *L. mexicana* (used at 120 ng/µl). For all reactions Taq-polymerase (Invitrogen 18038-042) was used in the following reaction mix: 1 µl dNTPs (10 mM), 1 µl gDNA, 0.5 µl Taq polymerase, 1 µl DMSO, 1 µl Primer 1 (100 µM), 1 µl Primer 2 (100 µM), 5 µl 10x standard reaction buffer provided by polymerase supplier, 40 µl ddH_20). Thermal cycling parameters: 5 min denaturation at 94 C, followed by 25 cycles of 30s (58 C), 45 s extension (72 C), denaturation 30s (92 C). Amplified fragments were gel purified, digested with the appropriate enzyme (see Supplementary Table 6.1) (all enzymes procured from New England Bioloabs) and ligated into pLENT vector using T4-DNA ligase (Roche 10716359001) as per manufacturer's instructions and reactions transformed into *E. coli* by heat shock (45 s at 42 C) before selection of LB-Agar plates with 100 µg/ml Ampicillin. Following overnight incubation at 37 C, colonies were picked and grown in liquid

LB medium. Plasmids were extracted using QIAprep Spin Miniprep Kit (Cat. No. 27104) as per manufacturer's instructions, sequenced by the Source BioScience sequencing service to verify correct plasmid composition. Larger volumes of correct plasmid were grown up in liquid LB-culture and extracted using HiSpeed Plasmid Midi Kit (Cat. No. 12643) as per manufacturer's instructions. Plasmids were linearised by the appropriate enzymes (see Supplementary Table 6.1). 10 µg linearised plasmid were washed by acetate-ethanol precipitation and transfected into $2x10^7$ logarithmic growth phase ($3x10^6$ – $1x10^7$ cells/ml) wild-type *L. mexicana* promastigotes suspended in 500 µl Zimmerman's post-fusion medium (135 mM NaCl, 8mM Na_2HPO_4, 1.5 mM KH_2PO_4, 0.5 mM Mg acetate, 0.09 mM Ca acetate, pH 7.0). Transfection was performed using two 1.7 kV pulses with 30 s intervals. Cultures were selected in 10 ml M199 with 25 µg/ml phleomycin (Sigma-P9564).

RSP3 and PF16 cell lines use as positive controls were obtained from Dr. Richard Wheeler (Wheeler *et al.*, unpublished).

6.24 Western Blot analysis

For Western Blot analyses, whole-cell protein lysates were performed as follows: $2.5x10^7$ PRO or AXA cells were washed 3 times in PBS at 800g for 5 min, re-suspended in 1 ml PBS into a 1.5 ml centrifuge tube and pelleted at 1000g for 3 min. All PBS was removed and pellets suspended in 50 µl boiling hot Laemmli buffer (Laemmli 1970) and boiled for 5 min. 10 µl of each protein sample were separated on 10 % SDS-PAGE gels and proteins transferred to nitrocellulose membranes by wet electrophoretic transfer. Membranes were blocked for 1 h in Tris-buffered saline (TBS, 200mM Tris, 3 M NaCl, pH 7.4) with 5 % w/v dry milk

(Marvel). Membranes were stained using anti-GFP monoclonal antibody (Roche 11814460001) at 1:1000 dilution from manufacturer's stock (0.4 μg / ml) in TBS+5% milk. For 1 h. Membranes were then washed 3 times for 5 min in TBS + 0.05% v/v Tween-20 (Sigma, P5927), before being stained for 1 h using rabbit anti-mouse IgG polyclonal peroxidase-conjugated antibody (Sigma, A9044) for 1 h. Membranes were washed a further three times using TBS+0.05% Tween-20 and bands resolved using Western Lightning Plus-ECL system (Perkin Elmer, NEL 103001EA) on Kodak X-OMAT LS photographic film (F1274-50EA).

References

Abanades, Daniel R, Laura Ramírez, Salvador Iborra, Ketty Soteriadou, Victor M González, Pedro Bonay, Carlos Alonso, and Manuel Soto. 2009. "Key Role of the 3' Untranslated Region in the Cell Cycle Regulated Expression of the Leishmania Infantum Histone H2A Genes: Minor Synergistic Effect of the 5' Untranslated Region." *BMC Molecular Biology* 10: 48. doi:10.1186/1471-2199-10-48.

Agami, R, and M Shapira. 1992. "Nucleotide Sequence of the Spliced Leader RNA Gene from Leishmania Mexicana Amazonensis." *Nucleic Acids Research* 20 (7): 1804.

Akiyoshi, Bungo, and Keith Gull. 2013. "Evolutionary Cell Biology of Chromosome Segregation: Insights from Trypanosomes." *Open Biology* 3 (5): 130023. doi:10.1098/rsob.130023.

Akopyants, Natalia S., Sandra W. Clifton, John Martin, Deana Pape, Todd Wylie, Li Li, Jessica C. Kissinger, David S. Roos, and Stephen M. Beverley. 2001. "A Survey of the Leishmania Major Friedlin Strain V1 Genome by Shotgun Sequencing: A Resource for DNA Microarrays and Expression Profiling." *Molecular and Biochemical Parasitology* 113 (2): 337–40. doi:10.1016/S0166-6851(01)00227-4.

Akopyants, Natalia S., Elizabeth Kruvand, Iris wong, and Stephen M Beverley. 2010. "(Manuscript in Preparation)."

Akopyants, Natalia S, Robin S Matlib, Elena N Bukanova, Matthew R Smeds, Bernard H Brownstein, Gary D Stormo, and Stephen M Beverley. 2004. "Expression Profiling Using Random Genomic DNA Microarrays Identifies Differentially Expressed Genes Associated with Three Major Developmental Stages of the Protozoan Parasite Leishmania Major." *Molecular and Biochemical Parasitology* 136 (1): 71–86. doi:10.1016/j.molbiopara.2004.03.002.

Alcolea, Pedro, Ana Alonso, Manuel Gomez, Alicia Sanchez-Gorostiaga, Mercedes Moreno-Paz, Eduardo Gonzalez-Pastor, Alfredo Torano, Victor Parro, and Vicente Larraga. 2010. "Temperature Increase Prevails over Acidification in Gene Expression Modulation of Amastigote Differentiation in Leishmania Infantum." *BMC Genomics* 11: 31.

Allen, James W. A., Andrew P. Jackson, Daniel J. Rigden, Antony C. Willis, Stuart J. Ferguson, and Michael L. Ginger. 2008. "Order within a Mosaic Distribution of Mitochondrial c-Type Cytochrome Biogenesis Systems?" *The FEBS Journal* 275 (10): 2385–2402. doi:10.1111/j.1742-4658.2008.06380.x.

Almeida, Renata, Brian J Gilmartin, Sharon H McCann, Alan Norrish, Alasdair C Ivens, Danial Lawson, Mark P Levick, et al. 2004. "Expression Profiling of the Leishmania Life Cycle: cDNA Arrays Identify Developmentally Regulated Genes Present but Not Annotated in the Genome." *Molecular and Biochemical Parasitology* 136 (1): 87–100. doi:10.1016/j.molbiopara.2004.03.004.

Alsford, Sam, Daniel J. Turner, Samson O. Obado, Alejandro Sanchez-Flores, Lucy Glover, Matthew Berriman, Christiane Hertz-Fowler, and David Horn. 2011. "High-Throughput Phenotyping Using Parallel Sequencing of RNA Interference Targets in the African Trypanosome." *Genome Research* 21 (6): 915–24. doi:10.1101/gr.115089.110.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10. doi:10.1016/S0022-2836(05)80360-2.

Alvar, Jorge, Sergio Yactayo, and Caryn Bern. 2006. "Leishmaniasis and Poverty." *Trends in Parasitology* 22 (12): 552–57. doi:10.1016/j.pt.2006.09.004.

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11 (10): R106. doi:10.1186/gb-2010-11-10-r106.

Antoine, J. C., E. Prina, C. Jouanne, and P. Bongrand. 1990. “Parasitophorous Vacuoles of Leishmania Amazonensis-Infected Macrophages Maintain an Acidic pH.” *Infection and Immunity* 58 (3): 779–87.

Archer, Stuart K, Diana Inchaustegui, Rafael Queiroz, and Christine Clayton. 2011. “The Cell Cycle Regulated Transcriptome of Trypanosoma Brucei.” *PloS One* 6 (3): e18425. doi:10.1371/journal.pone.0018425.

Aronsky, E. 2011. “Command-Line Tools for Processing Biological Sequencing Data.” http://www.sciencedirect.com/science/article/pii/0166685195025006.

Ashok, Devika, and Hans Acha-Orbea. 2014. “Timing Is Everything: Dendritic Cell Subsets in Murine Leishmania Infection.” *Trends in Parasitology*, September. doi:10.1016/j.pt.2014.08.001.

Aslett, Martin, Cristina Aurrecoechea, Matthew Berriman, John Brestelli, Brian P. Brunk, Mark Carrington, Daniel P. Depledge, et al. 2010. “TriTrypDB: A Functional Genomic Resource for the Trypanosomatidae.” *Nucleic Acids Research* 38 (suppl 1): D457–62. doi:10.1093/nar/gkp851.

Austyn, Jonathan M., and Siamon Gordon. 1981. “F4/80, a Monoclonal Antibody Directed Specifically against the Mouse Macrophage.” *European Journal of Immunology* 11 (10): 805–15. doi:10.1002/eji.1830111013.

Bairoch, A. 1991. “PROSITE: A Dictionary of Sites and Patterns in Proteins.” *Nucleic Acids Research* 19 Suppl (April): 2241–45.

Balanco, J. M., E. M. Pral, S. da Silva, A. T. Bijovsky, R. A. Mortara, and S. C. Alfieri. 1998. “Axenic Cultivation and Partial Characterization of Leishmania Braziliensis Amastigote-like Stages.” *Parasitology* 116 (Pt 2) (February): 103–13.

Balogun, R. A. 1974. “Amino Acids in the Excreta of the Tsetse Fly, Glossina Palpalis.” *Experientia* 30 (3): 239–40.

Bangs, J. D., L. Uyetake, M. J. Brickman, A. E. Balber, and J. C. Boothroyd. 1993. “Molecular Cloning and Cellular Localization of a BiP Homologue in Trypanosoma Brucei. Divergent ER Retention Signals in a Lower Eukaryote.” *Journal of Cell Science* 105 (Pt 4) (August): 1101–13.

Barak, Efrat, Sigal Amin-Spector, Elena Gerliak, Sophie Goyard, Neta Holland, and Dan Zilberstein. 2005. “Differentiation of Leishmania Donovani in Host-Free System: Analysis of Signal Perception and Response.” *Molecular and Biochemical Parasitology* 141 (1): 99–108. doi:10.1016/j.molbiopara.2005.02.004.

Bart, G, M J Frame, R Carter, G H Coombs, and J C Mottram. 1997. “Cathepsin B-like Cysteine Proteinase-Deficient Mutants of Leishmania Mexicana.” *Molecular and Biochemical Parasitology* 88 (1-2): 53–61.

Bastien, P., C. Blaineau, and M. Pages. 1992. “Leishmania: Sex, Lies and Karyotype.” *Parasitology Today (Personal Ed.)* 8 (5): 174–77.

Bastin, P., K. R. Matthews, and K. Gull. 1996. “The Paraflagellar Rod of Kinetoplastida: Solved and Unsolved Questions.” *Parasitology Today (Personal Ed.)* 12 (8): 302–7.

Bates, P. A. 1994. “Complete Developmental Cycle of Leishmania Mexicana in Axenic Culture.” *Parasitology* 108 (01): 1–9. doi:10.1017/S0031182000078458.

Bates, P. A., C. D. Robertson, L. Tetley, and G. H. Coombs. 1992. “Axenic Cultivation and Characterization of Leishmania Mexicana Amastigote-like Forms.” *Parasitology* 105 (02): 193–202. doi:10.1017/S0031182000074102.

Bates, Paul A. 2007. “Transmission of Leishmania Metacyclic Promastigotes by Phlebotomine Sand Flies.” *International Journal for Parasitology* 37 (10): 1097–1106. doi:10.1016/j.ijpara.2007.04.003.

Beer. 1852. “Bestimmung Der Absorption Des Rothen Lichts in Farbigen Flüssigkeiten.”

Bellatin, J.A., A.S. Murray, M. Zhao, and W.R. McMaster. 2002. “Leishmania Mexicana: Identification of Genes That Are Preferentially Expressed in Amastigotes.” *Experimental Parasitology* 100 (1): 44–53. doi:10.1006/expr.2001.4677.

Benjamini, Yoav, Dan Drai, Greg Elmer, Neri Kafkafi, and Ilan Golani. 2001. “Controlling the False Discovery Rate in Behavior Genetics Research.” *Behavioural Brain Research* 125 (1–2): 279–84. doi:10.1016/S0166-4328(01)00297-2.

Bente, Meike, Simone Harder, Martina Wiesgigl, Jochen Heukeshoven, Christoph Gelhaus, Eberhard Krause, Joachim Clos, and Iris Bruchhaus. 2003. “Developmentally Induced Changes of the Proteome in the Protozoan Parasite Leishmania Donovani.” *PROTEOMICS* 3 (9): 1811–29. doi:10.1002/pmic.200300462.

Berberich, C., G. Machado, G. Morales, G. Carrillo, A. Jiménez-Ruiz, and C. Alonso. 1998. “The Expression of the Leishmania Infantum KMP-11 Protein Is Developmentally Regulated and Stage Specific.” *Biochimica Et Biophysica Acta* 1442 (2-3): 230–37.

Berriman, Matthew, Elodie Ghedin, Christiane Hertz-Fowler, Gaëlle Blandin, Hubert Renauld, Daniella C. Bartholomeu, Nicola J. Lennard, et al. 2005. “The Genome of the African Trypanosome Trypanosoma Brucei.” *Science* 309 (5733): 416–22.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics*, April, btu170. doi:10.1093/bioinformatics/btu170.

Boucher, Nathalie, Ying Wu, Carole Dumas, Marthe Dubé, Denis Sereno, Marie Breton, and Barbara Papadopoulou. 2002. “A Common Mechanism of Stage-Regulated Gene Expression inLeishmania Mediated by a Conserved 3’-Untranslated Region Element.” *Journal of Biological Chemistry* 277 (22): 19511–20. doi:10.1074/jbc.M200500200.

Breton, Marie, Michel J. Tremblay, Marc Ouellette, and Barbara Papadopoulou. 2005. “Live Nonpathogenic Parasitic Vector as a Candidate Vaccine against Visceral Leishmaniasis.” *Infection and Immunity* 73 (10): 6372–82. doi:10.1128/IAI.73.10.6372-6382.2005.

Brickman, M. J., and A. E. Balber. 1990. “Trypanosoma Brucei Rhodesiense Bloodstream Forms: Surface Ricin-Binding Glycoproteins Are Localized Exclusively in the Flagellar Pocket and the Flagellar Adhesion Zone.” *The Journal of Protozoology* 37 (3): 219–24.

Bringaud, Frédéric, Michaela Müller, Gustavo Coutinho Cerqueira, Martin Smith, Annie Rochette, Najib M. A. El-Sayed, Barbara Papadopoulou, and Elodie Ghedin. 2007. “Members of a Large Retroposon Family Are Determinants of Post-Transcriptional Gene Expression in Leishmania.” *PLoS Pathogens* 3 (9): 1291–1307. doi:10.1371/journal.ppat.0030136.

Brittingham, A, C J Morrison, W R McMaster, B S McGwire, K P Chang, and D M Mosser. 1995. “Role of the Leishmania Surface Protease gp63 in Complement Fixation, Cell Adhesion, and Resistance to Complement-Mediated Lysis.” *The Journal of Immunology* 155 (6): 3102–11.

Brooks, D. R., H. Denise, G. D. Westrop, G. H. Coombs, and J. C. Mottram. 2001. “The Stage-Regulated Expression of Leishmania Mexicana CPB Cysteine Proteases Is Mediated by an Intercistronic Sequence Element.” *The Journal of Biological Chemistry* 276 (50): 47061–69. doi:10.1074/jbc.M108498200.

Brotherton, Marie-Christine, Sylvie Bourassa, Philippe Leprohon, Danielle Légaré, Guy G. Poirier, Arnaud Droit, and Marc Ouellette. 2013. “Proteomic and Genomic Analyses of Antimony Resistant Leishmania Infantum Mutant.” *PLoS ONE* 8 (11): e81899. doi:10.1371/journal.pone.0081899.

Bruce, David. 1895. “Preliminary Report on the Tsetse Fly Disease or Nagana, in Zululand.”

Burchmore, Richard J. S., and Scott M. Landfear. 1998. “Differential Regulation of Multiple Glucose Transporter Genes in Leishmania Mexicana.” *Journal of Biological Chemistry* 273 (44): 29118–26. doi:10.1074/jbc.273.44.29118.

Burki, Fabien. 2014. “The Eukaryotic Tree of Life from a Global Phylogenomic Perspective.” *Cold Spring Harbor Perspectives in Biology* 6 (5): a016147. doi:10.1101/cshperspect.a016147.

Bursell, E., K. J. Billing, J. W. Hargrove, C. T. McCabe, and E. Slack. 1973. “The Supply of Substrates to the Flight Muscle of Tsetse Flies.” *Transactions of the Royal Society of Tropical Medicine and Hygiene* 67 (2): 296.

Buxbaum, Laurence U., Hubert Denise, Graham H. Coombs, James Alexander, Jeremy C. Mottram, and Phillip Scott. 2003. “Cysteine Protease B of Leishmania Mexicana Inhibits Host Th1 Responses and Protective Immunity.” *Journal of Immunology (Baltimore, Md.: 1950)* 171 (7): 3711–17.

Cameron, Pamela, Adrienne McGachy, Mary Anderson, Andrew Paul, Graham H. Coombs, Jeremy C. Mottram, James Alexander, and Robin Plevin. 2004. “Inhibition of Lipopolysaccharide-Induced Macrophage IL-12 Production by Leishmania Mexicana Amastigotes: The Role of Cysteine Peptidases and the NF-kappaB Signaling Pathway.” *Journal of Immunology (Baltimore, Md.: 1950)* 173 (5): 3297–3304.

Campbell, Kimberly, Vsevolod Popov, and Lynn Soong. 2004. “Identification and Molecular Characterization of a Gene Encoding a Protective Leishmania Amazonensis Trp-Asp (WD) Protein.” *Infection and Immunity* 72 (4): 2194–2202.

Cargill, Michele, David Altshuler, James Ireland, Pamela Sklar, Kristin Ardlie, Nila Patil, Charles R. Lane, et al. 1999. “Characterization of Single-Nucleotide Polymorphisms in Coding Regions of Human Genes.” *Nature Genetics* 22 (3): 231–38. doi:10.1038/10290.

Castro, Helena, and Ana M. Tomás. 2008. “Peroxidases of Trypanosomatids.” *Antioxidants & Redox Signaling* 10 (9): 1593–1606. doi:10.1089/ars.2008.2050.

Cavalier-Smith, Thomas. 2010. “Kingdoms Protozoa and Chromista and the Eozoan Root of the Eukaryotic Tree.” *Biology Letters* 6 (3): 342–45. doi:10.1098/rsbl.2009.0948.

Cestari, I., I. Evans-Osses, L.J. Schlapbach, I. de Messias-Reason, and M.I. Ramirez. 2013. “Mechanisms of Complement Lectin Pathway Activation and Resistance by Trypanosomatid Parasites.” *Molecular Immunology* 53 (4): 328–34. doi:10.1016/j.molimm.2012.08.015.

Chang, K. P., and D. Fong. 1982. “Antigenic Changes during Intracellular Differentiation of Leishmania Mexicana in Cultured Macrophages.” *Infection and Immunity* 36 (1): 430–31.

Charest, H, and G Matlashewski. 1994. “Developmental Gene Expression in Leishmania Donovani: Differential Cloning and Analysis of an Amastigote-Stage-Specific Gene.” *Molecular and Cellular Biology* 14 (5): 2975–84.

Charest, H., W. W. Zhang, and G. Matlashewski. 1996. “The Developmental Expression of Leishmania Donovani A2 Amastigote-Specific Genes Is Post-Transcriptionally Mediated and Involves Elements Located in the 3’-Untranslated Region.” *The Journal of Biological Chemistry* 271 (29): 17081–90.

Chaudhuri, G., and K. P. Chang. 1988. “Acid Protease Activity of a Major Surface Membrane Glycoprotein (gp63) from Leishmania Mexicana Promastigotes.” *Molecular and Biochemical Parasitology* 27 (1): 43–52.

Chazotte, Brad. 2011. “Labeling Mitochondria with MitoTracker Dyes.” *Cold Spring Harbor Protocols* 2011 (8): 990–92. doi:10.1101/pdb.prot5648.

Chow, Conan, Serge Cloutier, Carole Dumas, Marie-Noelle Chou, and Barbara Papadopoulou. 2011. “Promastigote to Amastigote Differentiation of Leishmania Is Markedly Delayed in the Absence of PERK eIF2alpha Kinase-Dependent eIF2alpha Phosphorylation.” *Cellular Microbiology* 13 (7): 1059–77. doi:10.1111/j.1462-5822.2011.01602.x.

Clayton, C. E. 2014. “Networks of Gene Expression Regulation in Trypanosoma Brucei.” *Molecular and Biochemical Parasitology*, Special Issue on the 35th Anniversary of

Molecular and Biochemical Parasitology, 195 (2): 96–106. doi:10.1016/j.molbiopara.2014.06.005.
Clayton, Christine, and Michal Shapira. 2007. “Post-Transcriptional Regulation of Gene Expression in Trypanosomes and Leishmanias.” *Molecular and Biochemical Parasitology* 156 (2): 93–101. doi:10.1016/j.molbiopara.2007.07.007.
Cloutier, Serge, Maxime Laverdière, Marie-Noelle Chou, Nathalie Boilard, Conan Chow, and Barbara Papadopoulou. 2012. “Translational Control through eIF2alpha Phosphorylation during the Leishmania Differentiation Process.” *PloS One* 7 (5): e35085. doi:10.1371/journal.pone.0035085.
Contreras, Irazú, María Adelaida Gómez, Oliver Nguyen, Marina T. Shio, Robert W. McMaster, and Martin Olivier. 2010. “Leishmania-Induced Inactivation of the Macrophage Transcription Factor AP-1 Is Mediated by the Parasite Metalloprotease GP63.” *PLoS Pathog* 6 (10): e1001148. doi:10.1371/journal.ppat.1001148.
Coombs, G H, J A Craft, and D T Hart. 1982. “A Comparative Study of Leishmania Mexicana Amastigotes and Promastigotes. Enzyme Activities and Subcellular Locations.” *Molecular and Biochemical Parasitology* 5 (3): 199–211.
Cormack, Brendan P., Raphael H. Valdivia, and Stanley Falkow. 1996. “FACS-Optimized Mutants of the Green Fluorescent Protein (GFP).” *Gene*, Flourescent Proteins and Applications, 173 (1): 33–38. doi:10.1016/0378-1119(95)00685-0.
Cruz, A. K., R. Titus, and S. M. Beverley. 1993. “Plasticity in Chromosome Number and Testing of Essential Genes in Leishmania by Targeting.” *Proceedings of the National Academy of Sciences of the United States of America* 90 (4): 1599–1603.
Cuervo, Patricia, Jose Batista de Jesus, Magno Junqueira, Leila Mendonça-Lima, Luis Javier González, Lázaro Betancourt, Gabriel Grimaldi, Gilberto Barbosa Domont, Octavio Fernandes, and Elisa Cupolillo. 2007. “Proteome Analysis of Leishmania (Viannia) Braziliensis by Two-Dimensional Gel Electrophoresis and Mass Spectrometry.” *Molecular and Biochemical Parasitology* 154 (1): 6–21. doi:10.1016/j.molbiopara.2007.03.013.
Cupolillo, E., G. Grimaldi, and H. Momen. 1994. “A General Classification of New World Leishmania Using Numerical Zymotaxonomy.” *The American Journal of Tropical Medicine and Hygiene* 50 (3): 296–311.
Cuvillier, A., F. Redon, J. C. Antoine, P. Chardin, T. DeVos, and G. Merlin. 2000. “LdARL-3A, a Leishmania Promastigote-Specific ADP-Ribosylation Factor-like Protein, Is Essential for Flagellum Integrity.” *Journal of Cell Science* 113 (Pt 11) (June): 2065–74.
David, Maya, Idan Gabdank, Miriam Ben-David, Alon Zilka, Irit Orr, Danny Barash, and Michal Shapira. 2010. “Preferential Translation of Hsp83 in Leishmania Requires a Thermosensitive Polypyrimidine-Rich Element in the 3’ UTR and Involves Scanning of the 5’ UTR.” *RNA (New York, N.Y.)* 16 (2): 364–74. doi:10.1261/rna.1874710.
Dean, Samuel, Rosa Marchetti, Kiaran Kirk, and Keith R. Matthews. 2009. “A Surface Transporter Family Conveys the Trypanosome Differentiation Signal.” *Nature* 459: 213–17.
Debrabant, Alain, Manju B Joshi, Paulo F. P Pimenta, and Dennis M Dwyer. 2004. “Generation of Leishmania Donovani Axenic Amastigotes: Their Growth and Biological Characteristics.” *International Journal for Parasitology*, Annual Scientific Meeting of the Australian Society for Parasitology, Darwin, Carlton Hotel, The Esplande, 2003. Highlights., 34 (2): 205–17. doi:10.1016/j.ijpara.2003.10.011.
Demers, Eric, David M. Forrest, and Gabriele E. Weichert. 2013. “Cutaneous Leishmaniasis in a Returning Traveller.” *CMAJ: Canadian Medical Association Journal = Journal de l’Association Medicale Canadienne* 185 (8): 681–83. doi:10.1503/cmaj.120694.
Denise, Hubert, Kathryn McNeil, Darren R. Brooks, James Alexander, Graham H. Coombs, and Jeremy C. Mottram. 2003. “Expression of Multiple CPB Genes Encoding Cysteine

Proteases Is Required for Leishmania Mexicana Virulence in Vivo." *Infection and Immunity* 71 (6): 3190–95.

Depledge, Daniel P., Krystal J. Evans, Alasdair C. Ivens, Naveed Aziz, Asher Maroof, Paul M. Kaye, and Deborah F. Smith. 2009. "Comparative Expression Profiling of Leishmania: Modulation in Gene Expression between Species and in Different Host Genetic Backgrounds." *PLoS Negl Trop Dis* 3 (7): e476. doi:10.1371/journal.pntd.0000476.

Depledge, Daniel P., Lorna M. MacLean, Michael R. Hodgkinson, Barbara A. Smith, Andrew P. Jackson, Saufung Ma, Silvia R. B. Uliana, and Deborah F. Smith. 2010. "Leishmania-Specific Surface Antigens Show Sub-Genus Sequence Variation and Immune Recognition." *PLoS Negl Trop Dis* 4 (9): e829. doi:10.1371/journal.pntd.0000829.

Dermine, Jean-François, Guillaume Goyette, Mathieu Houde, Salvatore J. Turco, and Michel Desjardins. 2005. "Leishmania Donovani Lipophosphoglycan Disrupts Phagosome Microdomains in J774 Macrophages." *Cellular Microbiology* 7 (9): 1263–70. doi:10.1111/j.1462-5822.2005.00550.x.

Descoteaux, A., Y. Luo, S. J. Turco, and S. M. Beverley. 1995. "A Specialized Pathway Affecting Virulence Glycoconjugates of Leishmania." *Science (New York, N.Y.)* 269 (5232): 1869–72.

Descoteaux, A., and S. J. Turco. 1999. "Glycoconjugates in Leishmania Infectivity." *Biochimica Et Biophysica Acta* 1455 (2-3): 341–52.

Desjeux, Philippe, Raj S. Ghosh, Pritu Dhalaria, Nathalie Strub-Wourgaft, and Ed E. Zijlstra. 2013. "Report of the Post Kala-Azar Dermal Leishmaniasis (PKDL) Consortium Meeting, New Delhi, India, 27–29 June 2012." *Parasites & Vectors* 6 (1): 196. doi:10.1186/1756-3305-6-196.

Dey, Ranadhir, Claudio Meneses, Poonam Salotra, Shaden Kamhawi, Hira L. Nakhasi, and Robert Duncan. 2010. "Characterization of a Leishmania Stage-Specific Mitochondrial Membrane Protein That Enhances the Activity of Cytochrome c Oxidase and Its Role in Virulence." *Molecular Microbiology* 77 (2): 399–414. doi:10.1111/j.1365-2958.2010.07214.x.

Dillies, Marie-Agnès, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, et al. 2013. "A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis." *Briefings in Bioinformatics* 14 (6): 671–83. doi:10.1093/bib/bbs046.

Donelson, J. E., M. J. Gardner, and N. M. El-Sayed. 1999. "More Surprises from Kinetoplastida." *Proceedings of the National Academy of Sciences of the United States of America* 96 (6): 2579–81.

Donovan, Charles. 1903. *British Medical Journal*.

Doyle, Maria A, James I MacRae, David P De Souza, Eleanor C Saunders, Malcolm J McConville, and Vladimir A Likic. 2009. "LeishCyc: A Biochemical Pathways Database for Leishmania Major." *BMC Systems Biology* 3 (June): 57. doi:10.1186/1752-0509-3-57.

Dupé, Aurélien, Carole Dumas, and Barbara Papadopoulou. 2014. "An Alba-Domain Protein Contributes to the Stage-Regulated Stability of Amastin Transcripts in Leishmania." *Molecular Microbiology* 91 (3): 548–61. doi:10.1111/mmi.12478.

Ebmeier, Sarah E., Irene S. Tan, Katie Rose Clapham, and Kumaran S. Ramamurthi. 2012. "Small Proteins Link Coat and Cortex Assembly during Sporulation in Bacillus Subtilis." *Molecular Microbiology* 84 (4): 682–96. doi:10.1111/j.1365-2958.2012.08052.x.

El Fakhry, Youssef, Marc Ouellette, and Barbara Papadopoulou. 2002. "A Proteomic Approach to Identify Developmentally Regulated Proteins in Leishmania Infantum." *Proteomics* 2 (8): 1007–17. doi:10.1002/1615-9861(200208)2:8<1007::AID-PROT1007>3.0.CO;2-G.

Ellis, D. S., W. E. Ormerod, and W. H. Lumsden. 1976. "Filaments of Trypanosoma Brucei: Some Notes on Differences in Origin and Structure in Two Strains of Trypanosoma (Trypanozoon) Brucei Rhodesiense." *Acta Tropica* 33 (2): 151–68.

El-Sayed, Najib M., Peter J. Myler, Daniella C. Bartholomeu, Daniel Nilsson, Gautam Aggarwal, Anh-Nhi Tran, Elodie Ghedin, et al. 2005. "The Genome Sequence of Trypanosoma Cruzi, Etiologic Agent of Chagas Disease." *Science (New York, N.Y.)* 309 (5733): 409–15. doi:10.1126/science.1112631.

El-Sayed, Najib M., Peter J. Myler, Gaëlle Blandin, Matthew Berriman, Jonathan Crabtree, Gautam Aggarwal, Elisabet Caler, et al. 2005. "Comparative Genomics of Trypanosomatid Parasitic Protozoa." *Science (New York, N.Y.)* 309 (5733): 404–9. doi:10.1126/science.1112181.

Elwasila, Mohamed. 1988. "*Leishmania Tarentolae* Wenyon, 1921 from the Gecko *Tarentola Annularis* in the Sudan." *Parasitology Research* 74 (6): 591–92. doi:10.1007/BF00531640.

Eperon, S., and D. McMahon-Pratt. 1989. "Extracellular Amastigote-like Forms of Leishmania Panamensis and L. Braziliensis. II. Stage- and Species-Specific Monoclonal Antibodies." *The Journal of Protozoology* 36 (5): 510–18.

Ericson, Megan, Michael A. Janes, Falk Butter, Matthias Mann, Elisabetta Ullu, and Christian Tschudi. 2014. "On the Extent and Role of the Small Proteome in the Parasitic Eukaryote Trypanosoma Brucei." *BMC Biology* 12 (1): 14. doi:10.1186/1741-7007-12-14.

Evans-Osses, Ingrid, Iara de Messias-Reason, and Marcel I. Ramirez. 2013. "The Emerging Role of Complement Lectin Pathway in Trypanosomatids: Molecular Bases in Activation, Genetic Deficiencies, Susceptibility to Infection, and Complement System-Based Therapeutics." *The Scientific World Journal* 2013 (February): e675898. doi:10.1155/2013/675898.

Ewing, Brent, and Phil Green. 1998. "Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities." *Genome Research* 8 (3): 186–94.

Farajnia, S., M. H. Alimohammadian, N. E. Reiner, M. Karimi, S. Ajdari, and F. Mahboudi. 2004. "Molecular Characterization of a Novel Amastigote Stage Specific Class I Nuclease from Leishmania Major." *International Journal for Parasitology* 34 (8): 899–908. doi:10.1016/j.ijpara.2004.03.005.

Fernandes, A. P., K. Nelson, and S. M. Beverley. 1993. "Evolution of Nuclear Ribosomal RNAs in Kinetoplastid Protozoa: Perspectives on the Age and Origins of Parasitism." *Proceedings of the National Academy of Sciences* 90 (24): 11608–12.

Fickett, J. W. 1982. "Recognition of Protein Coding Regions in DNA Sequences." *Nucleic Acids Research* 10 (17): 5303–18.

Fiebig, Michael, Eva Gluenz, Mark Carrington, and Steven Kelly. 2014. "SLaP Mapper: A Webserver for Identifying and Quantifying Spliced-Leader Addition and Polyadenylation Site Usage in Kinetoplastid Genomes." *Molecular and Biochemical Parasitology*, August. doi:10.1016/j.molbiopara.2014.07.012.

Flannagan, Ronald S., Gabriela Cosío, and Sergio Grinstein. 2009. "Antimicrobial Mechanisms of Phagocytes and Bacterial Evasion Strategies." *Nature Reviews. Microbiology* 7 (5): 355–66. doi:10.1038/nrmicro2128.

Fong, D., M. Wallach, J. Keithly, P. W. Melera, and K. P. Chang. 1984. "Differential Expression of mRNAs for Alpha- and Beta-Tubulin during Differentiation of the Parasitic Protozoan Leishmania Mexicana." *Proceedings of the National Academy of Sciences of the United States of America* 81 (18): 5782–86.

Forget, Geneviève, David J. Gregory, and Martin Olivier. 2005. "Proteasome-Mediated Degradation of STAT1alpha Following Infection of Macrophages with Leishmania Donovani." *The Journal of Biological Chemistry* 280 (34): 30542–49. doi:10.1074/jbc.M414126200.

Frame, M. J., J. C. Mottram, and G. H. Coombs. 2000. “Analysis of the Roles of Cysteine Proteinases of Leishmania Mexicana in the Host-Parasite Interaction.” *Parasitology* 121 (Pt 4) (October): 367–77.

Frommel, T. O., L. L. Button, Y. Fujikura, and W. R. McMaster. 1990. “The Major Surface Glycoprotein (GP63) Is Present in Both Life Stages of Leishmania.” *Molecular and Biochemical Parasitology* 38 (1): 25–32.

Gagnon, Dominic, Aude Foucher, Isabelle Girard, and Marc Ouellette. 2006. “Stage Specific Gene Expression and Cellular Localization of Two Isoforms of the Serine Hydroxymethyltransferase in the Protozoan Parasite Leishmania.” *Molecular and Biochemical Parasitology* 150 (1): 63–71. doi:10.1016/j.molbiopara.2006.06.009.

Ghedin, Elodie, Frederic Bringaud, Jeremy Peterson, Peter Myler, Matthew Berriman, Alasdair Ivens, Björn Andersson, et al. 2004. “Gene Synteny and Evolution of Genome Architecture in Trypanosomatids.” *Molecular and Biochemical Parasitology* 134 (2): 183–91. doi:10.1016/j.molbiopara.2003.11.012.

Gluenz, Eva, Johanna L. Höög, Amy E. Smith, Helen R. Dawe, Michael K. Shaw, and Keith Gull. 2010. “Beyond 9+0: Noncanonical Axoneme Structures Characterize Sensory Cilia from Protists to Humans.” *The FASEB Journal* 24 (9): 3117–21. doi:10.1096/fj.09-151381.

Goecks, Jeremy, Anton Nekrutenko, James Taylor, and Galaxy Team. 2010. “Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences.” *Genome Biology* 11 (8): R86. doi:10.1186/gb-2010-11-8-r86.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, et al. 1996. “Life with 6000 Genes.” *Science (New York, N.Y.)* 274 (5287): 546, 563–67.

Gomez, Maria Adelaida, Irazu Contreras, Maxime Halle, Michel L. Tremblay, Robert W. McMaster, and Martin Olivier. 2009. “Leishmania GP63 Alters Host Signaling Through Cleavage-Activated Protein Tyrosine Phosphatases.” *Science Signaling* 2 (90): ra58. doi:10.1126/scisignal.2000213.

Gossage, Sharon M., Matthew E. Rogers, and Paul A. Bates. 2003. “Two Separate Growth Phases during the Development of Leishmania in Sand Flies: Implications for Understanding the Life Cycle.” *International Journal for Parasitology* 33 (10): 1027–34.

Goto, S, H Bono, H Ogata, W Fujibuchi, T Nishioka, K Sato, and M Kanehisa. 1997. “Organizing and Computing Metabolic Pathway Data in Terms of Binary Relations.” *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 175–86.

Gregory, David J., Marianne Godbout, Irazú Contreras, Geneviève Forget, and Martin Olivier. 2008. “A Novel Form of NF-kappaB Is Induced by Leishmania Infection: Involvement in Macrophage Gene Expression.” *European Journal of Immunology* 38 (4): 1071–81. doi:10.1002/eji.200737586.

Gutiérrez, J. A., F. Puentes, A. Moreno, M. E. Patarroyo, and L. A. Murillo. 2001. “Identification of a Differentially Expressed mRNA in Axenic Leishmania Panamensis Amastigotes.” *Memórias Do Instituto Oswaldo Cruz* 96 (6): 835–38.

Gygi, S. P., B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold. 1999. “Quantitative Analysis of Complex Protein Mixtures Using Isotope-Coded Affinity Tags.” *Nature Biotechnology* 17 (10): 994–99. doi:10.1038/13690.

Handler, Aaron A., Joo Eun Lim, and Richard Losick. 2008. “Peptide Inhibitor of Cytokinesis during Sporulation in Bacillus Subtilis.” *Molecular Microbiology* 68 (3): 588–99. doi:10.1111/j.1365-2958.2008.06173.x.

Handman, E., H. M. Jarvis, and G. F. Mitchell. 1984. “Leishmania Major: Identification of Stage-Specific Antigens and Antigens Shared by Promastigotes and Amastigotes.” *Parasite Immunology* 6 (3): 223–33.

Hansen, K. D., S. E. Brenner, and S. Dudoit. 2010. “Biases in Illumina Transcriptome Sequencing Caused by Random Hexamer Priming.” *Nucleic Acids Research* 38 (12): e131–e131. doi:10.1093/nar/gkq224.

Harder, Simone, Meike Thiel, Joachim Clos, and Iris Bruchhaus. 2010. “Characterization of a Subunit of the Outer Dynein Arm Docking Complex Necessary for Correct Flagellar Assembly in Leishmania Donovani.” *PLoS Neglected Tropical Diseases* 4 (1): e586. doi:10.1371/journal.pntd.0000586.

Harsman, Anke, Moritz Niemann, Mascha Pusnik, Oliver Schmidt, Björn M. Burmann, Sebastian Hiller, Chris Meisinger, André Schneider, and Richard Wagner. 2012. “Bacterial Origin of a Mitochondrial Outer Membrane Protein Translocase: New Perspectives from Comparative Single Channel Electrophysiology.” *The Journal of Biological Chemistry* 287 (37): 31437–45. doi:10.1074/jbc.M112.392118.

Hem, Sonia, Pier Federico Gherardini, José Osorio y Fortéa, Veronique Hourdel, Miguel A. Morales, Reiko Watanabe, Pascale Pescher, et al. 2010. “Identification of Leishmania-Specific Protein Phosphorylation Sites by LC-ESI-MS/MS and Comparative Genomics Analyses.” *Proteomics* 10 (21): 3868–83. doi:10.1002/pmic.201000305.

Hewitt, S., H. Reyburn, R. Ashford, and M. Rowland. 1998. “Anthroponotic Cutaneous Leishmaniasis in Kabul, Afghanistan: Vertical Distribution of Cases in Apartment Blocks.” *Transactions of the Royal Society of Tropical Medicine and Hygiene* 92 (3): 273–74. doi:10.1016/S0035-9203(98)91007-0.

Hirsh, A. E., and H. B. Fraser. 2001. “Protein Dispensability and Rate of Evolution.” *Nature* 411 (6841): 1046–49. doi:10.1038/35082561.

Hodges, Matthew E., Nicole Scheumann, Bill Wickstead, Jane A. Langdale, and Keith Gull. 2010. “Reconstructing the Evolutionary History of the Centriole from Protein Components.” *Journal of Cell Science* 123 (Pt 9): 1407–13. doi:10.1242/jcs.064873.

Hodgkinson, V. H., L. Soong, S. M. Duboise, and D. McMahon-Pratt. 1996. “Leishmania Amazonensis: Cultivation and Characterization of Axenic Amastigote-like Organisms.” *Experimental Parasitology* 83 (1): 94–105. doi:10.1006/expr.1996.0053.

Holm, Å., K. Tejle, K.-E. Magnusson, A. Descoteaux, and B. Rasmusson. 2001. “Leishmania Donovani Lipophosphoglycan Causes Periphagosomal Actin Accumulation: Correlation with Impaired Translocation of PKCα and Defective Phagosome Maturation.” *Cellular Microbiology* 3 (7): 439–47. doi:10.1046/j.1462-5822.2001.00127.x.

Holzer, Timothy R., W.R. McMaster, and James D. Forney. 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.” *Molecular and Biochemical Parasitology* 146 (2): 198–218. doi:10.1016/j.molbiopara.2005.12.009.

Holzer, Timothy R, Krishna K Mishra, Jonathan H LeBowitz, and James D Forney. 2008. “Coordinate Regulation of a Family of Promastigote-Enriched mRNAs by the 3’UTR PRE Element in Leishmania Mexicana.” *Molecular and Biochemical Parasitology* 157 (1): 54–64. doi:10.1016/j.molbiopara.2007.10.001.

Howard, M. K., G. Sayers, and M. A. Miles. 1987. “Leishmania Donovani Metacyclic Promastigotes: Transformation in Vitro, Lectin Agglutination, Complement Resistance, and Infectivity.” *Experimental Parasitology* 64 (2): 147–56.

Hsieh, C. S., S. E. Macatonia, C. S. Tripp, S. F. Wolf, A. O’Garra, and K. M. Murphy. 1993. “Development of TH1 CD4+ T Cells through IL-12 Produced by Listeria-Induced Macrophages.” *Science (New York, N.Y.)* 260 (5107): 547–49.

Huang, J., and L. H. van der Ploeg. 1991. “Maturation of Polycistronic Pre-mRNA in Trypanosoma Brucei: Analysis of Trans Splicing and poly(A) Addition at Nascent RNA Transcripts from the hsp70 Locus.” *Molecular and Cellular Biology* 11 (6): 3180–90. doi:10.1128/MCB.11.6.3180.

Huang, J, and L H Van der Ploeg. 1991. “Requirement of a Polypyrimidine Tract for Trans-Splicing in Trypanosomes: Discriminating the PARP Promoter from the Immediately Adjacent 3’ Splice Acceptor Site.” *The EMBO Journal* 10 (12): 3877–85.

Huynh, Chau, Xiaojing Yuan, Danilo C. Miguel, Rebecca L. Renberg, Olga Protchenko, Caroline C. Philpott, Iqbal Hamza, and Norma W. Andrews. 2012. “Heme Uptake by Leishmania Amazonensis Is Mediated by the Transmembrane Protein LHR1.” *PLoS Pathogens* 8 (7): e1002795. doi:10.1371/journal.ppat.1002795.

Huynh, C., D. L. Sacks, and N. W. Andrews. 2006. “A Leishmania Amazonensis ZIP Family Iron Transporter Is Essential for Parasite Replication within Macrophage Phagolysosomes.” *J Exp Med* 203 (October): 2363–75. doi:10.1084/jem.20060559.

Hyams, J S. 1982. “The Euglena Paraflagellar Rod: Structure, Relationship to Other Flagellar Components and Preliminary Biochemical Characterization.” *Journal of Cell Science* 55 (June): 199–210.

Ilg, T., P. Overath, M. A. Ferguson, T. Rutherford, D. G. Campbell, and M. J. McConville. 1994. “O- and N-Glycosylation of the Leishmania Mexicana-Secreted Acid Phosphatase. Characterization of a New Class of Phosphoserine-Linked Glycans.” *The Journal of Biological Chemistry* 269 (39): 24073–81.

Ilg, T, Y D Stierhof, R Etges, M Adrian, D Harbecke, and P Overath. 1991. “Secreted Acid Phosphatase of Leishmania Mexicana: A Filamentous Phosphoglycoprotein Polymer.” *Proceedings of the National Academy of Sciences of the United States of America* 88 (19): 8774–78.

Ip, H. S., A. Orn, D. G. Russell, and G. A. Cross. 1990. “Leishmania Mexicana Mexicana gp63 Is a Site-Specific Neutral Endopeptidase.” *Molecular and Biochemical Parasitology* 40 (2): 163–72.

Ivens, Alasdair C., Christopher S. Peacock, Elizabeth A. Worthey, Lee Murphy, Gautam Aggarwal, Matthew Berriman, Ellen Sisk, et al. 2005. “The Genome of the Kinetoplastid Parasite, Leishmania Major.” *Science* 309 (5733): 436–42. doi:10.1126/science.1112680.

Ives, Annette, Catherine Ronet, Florence Prevel, Giulia Ruzzante, Silvia Fuertes-Marraco, Frederic Schutz, Haroun Zangger, et al. 2011. “Leishmania RNA Virus Controls the Severity of Mucocutaneous Leishmaniasis.” *Science (New York, N.Y.)* 331 (6018): 775–78. doi:10.1126/science.1199326.

Jackson, Andrew P. 2007. “Evolutionary Consequences of a Large Duplication Event in Trypanosoma Brucei: Chromosomes 4 and 8 Are Partial Duplicons.” *BMC Genomics* 8: 432. doi:10.1186/1471-2164-8-432.

Jackson, Andrew P. 2010. “The Evolution of Amastin Surface Glycoproteins in Trypanosomatid Parasites.” *Molecular Biology and Evolution* 27 (1): 33–45. doi:10.1093/molbev/msp214.

Jaffe, C. L., and N. Rachamim. 1989. “Amastigote Stage-Specific Monoclonal Antibodies against Leishmania Major.” *Infection and Immunity* 57 (12): 3770–77.

Johnson, P J, J M Kooter, and P Borst. 1987. “Inactivation of Transcription by UV Irradiation of T. Brucei Provides Evidence for a Multicistronic Transcription Unit Including a VSG Gene.” *Cell* 51 (2): 273–81.

Jordan, I. King, Igor B. Rogozin, Yuri I. Wolf, and Eugene V. Koonin. 2002. “Essential Genes Are More Evolutionarily Conserved than Are Nonessential Genes in Bacteria.” *Genome Research* 12 (6): 962–68. doi:10.1101/gr.87702. Article published online before print in May 2002.

Joshi, P. B., D. L. Sacks, G. Modi, and W. R. McMaster. 1998. “Targeted Gene Deletion of Leishmania Major Genes Encoding Developmental Stage-Specific Leishmanolysin (GP63).” *Molecular Microbiology* 27 (3): 519–30.

Joshi, Phalgun B., Ben L. Kelly, Shaden Kamhawi, David L. Sacks, and W. Robert McMaster. 2002. “Targeted Gene Deletion in Leishmania Major Identifies Leishmanolysin (GP63) as a Virulence Factor.” *Molecular and Biochemical Parasitology* 120 (1): 33–40. doi:10.1016/S0166-6851(01)00432-7.

Kautz-Neu, Kordula, Kirsten Schwonberg, Michael R. Fischer, Anja I. Schermann, and Esther von Stebut. 2012. “Dendritic Cells in Leishmania Major Infections: Mechanisms of Parasite Uptake, Cell Activation and Evidence for Physiological Relevance.” *Medical Microbiology and Immunology* 201 (4): 581–92. doi:10.1007/s00430-012-0261-2.

Kelly, S, S Kramer, A Schwede, P K Maini, K Gull, and M Carrington. 2012. “Genome Organization Is a Major Component of Gene Expression Control in Response to Stress and during the Cell Division Cycle in Trypanosomes.” *Open Biology* 2 (4): 120033. doi:10.1098/rsob.120033.

Kelly, Bill Wickstead, Philip K. Maini, and Keith Gull. 2011. “Ab Initio Identification of Novel Regulatory Elements in the Genome of Trypanosoma Brucei by Bayesian Inference on Sequence Segmentation.” *PLoS ONE* 6 (10): e25666. doi:10.1371/journal.pone.0025666.

Killick-Kendrick, R. 1999. “The Biology and Control of Phlebotomine Sand Flies.” *Clinics in Dermatology* 17 (3): 279–89. doi:10.1016/S0738-081X(99)00046-2.

Killick Kendrick, R., D.H. Molyneux, and R.W. Ashford. 1974. “Leishmania in Phlebotomid Sandflies. I. Modifications of the Flagellum Associated with Attachment to the Mid Gut and Oesophageal Valve of the Sandfly.” *Proceedings of the Royal Society of London - Biological Sciences* 187 (1089): 409–19.

Kima, Peter E. 2014. “Leishmania Molecules That Mediate Intracellular Pathogenesis.” *Microbes and Infection*. Accessed September 13. doi:10.1016/j.micinf.2014.07.012.

Koenig, Thomas, Bjoern H. Menze, Marc Kirchner, Flavio Monigatti, Kenneth C. Parker, Thomas Patterson, Judith Jebanathirajah Steen, Fred A. Hamprecht, and Hanno Steen. 2008. “Robust Prediction of the MASCOT Score for an Improved Quality Assessment in Mass Spectrometric Proteomics.” *Journal of Proteome Research* 7 (9): 3708–17. doi:10.1021/pr700859x.

Kolev, Nikolay G, Joseph B Franklin, Shai Carmi, Huafang Shi, Shulamit Michaeli, and Christian Tschudi. 2010. “The Transcriptome of the Human Pathogen Trypanosoma Brucei at Single-Nucleotide Resolution.” *PLoS Pathogens* 6 (9): e1001090. doi:10.1371/journal.ppat.1001090.

Kolev, Nikolay G., Kiantra Ramey-Butler, George A. M. Cross, Elisabetta Ullu, and Christian Tschudi. 2012. “Developmental Progression to Infectivity in Trypanosoma Brucei Triggered by an RNA-Binding Protein.” *Science* 338 (6112): 1352–53. doi:10.1126/science.1229641.

Kořený, Luděk, Miroslav Oborník, and Julius Lukeš. 2013. “Make It, Take It, or Leave It: Heme Metabolism of Parasites.” *PLoS Pathogens* 9 (1): e1003088. doi:10.1371/journal.ppat.1003088.

Lachaud, Laurence, Nathalie Bourgeois, Nada Kuk, Christelle Morelle, Lucien Crobu, Gilles Merlin, Patrick Bastien, Michel Pagès, and Yvon Sterkers. 2014. “Constitutive Mosaic Aneuploidy Is a Unique Genetic Feature Widespread in the Leishmania Genus.” *Microbes and Infection / Institut Pasteur* 16 (1): 61–66. doi:10.1016/j.micinf.2013.09.005.

Laemmli, U. K. 1970. “Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4.” *Nature* 227 (5259): 680–85.

Lahav, T., D. Sivam, H. Volpin, M. Ronen, P. Tsigankov, A. Green, N. Holland, et al. 2011. “Multiple Levels of Gene Regulation Mediate Differentiation of the Intracellular Pathogen Leishmania.” *The FASEB Journal* 25 (2): 515–25. doi:10.1096/fj.10-157529.

Lambert, J.H. 1760. “Photometria Sive de Mensura et Gradibus Luminis, Colorum et Umbrae.”

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. doi:10.1038/nmeth.1923.

LeBowitz, J H, H Q Smith, L Rusche, and S M Beverley. 1993. “Coupling of poly(A) Site Selection and Trans-Splicing in Leishmania.” *Genes & Development* 7 (6): 996–1007.

Leifso, Kirk, Gabriela Cohen-Freue, Nisha Dogra, Angus Murray, and W. Robert McMaster. 2007. “Genomic and Proteomic Expression Analysis of Leishmania Promastigote and Amastigote Life Stages: The Leishmania Genome Is Constitutively Expressed.” *Molecular and Biochemical Parasitology* 152: 35–46. doi:10.1016/j.molbiopara.2006.11.009.

Leishman, W. B. 1903. “ON THE POSSIBILITY OF THE OCCURRENCE OF TRYPANOSOMIASIS IN INDIA.” *British Medical Journal* 1 (2213): 1252–54.

Lenardo, M J, D M Dorfman, and J E Donelson. 1985. “The Spliced Leader Sequence of Trypanosoma Brucei Has a Potential Role as a Cap Donor Structure.” *Molecular and Cellular Biology* 5 (9): 2487–90.

Li, Bo, and Colin N. Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics* 12: 323. doi:10.1186/1471-2105-12-323.

Li, Bo, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. 2010. “RNA-Seq Gene Expression Estimation with Read Mapping Uncertainty.” *Bioinformatics* 26 (4): 493–500. doi:10.1093/bioinformatics/btp692.

Liu, K., S. Zinker, C. Argüello, and L. M. Salgado. 2000. “Isolation and Analysis of a New Developmentally Regulated Gene from Amastigotes of Leishmania Mexicana Mexicana.” *Parasitology Research* 86 (2): 140–50.

Lodge, Robert, Tamsir O. Diallo, and Albert Descoteaux. 2006. “Leishmania Donovani Lipophosphoglycan Blocks NADPH Oxidase Assembly at the Phagosome Membrane.” *Cellular Microbiology* 8 (12): 1922–31. doi:10.1111/j.1462-5822.2006.00758.x.

Lodish, H., A. Berk, L. Zipursky, P. Matsudaira, D. Baltimore, and James Darnell. 2000. “Processing of rRNA and tRNA - Molecular Cell Biology - NCBI Bookshelf.” http://www.ncbi.nlm.nih.gov/books/NBK21729/.

Logan-Klumpler, Flora J., Nishadi De Silva, Ulrike Boehme, Matthew B. Rogers, Giles Velarde, Jacqueline A. McQuillan, Tim Carver, et al. 2012. “GeneDB--an Annotation Database for Pathogens.” *Nucleic Acids Research* 40 (D1): D98–108. doi:10.1093/nar/gkr1032.

Lorenz, Ronny, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. 2011. “ViennaRNA Package 2.0.” *Algorithms for Molecular Biology* 6 (1): 26. doi:10.1186/1748-7188-6-26.

Lukes, J., M. Jirkû, D. Dolezel, I. Kral’ová, L. Hollar, and D. A. Maslov. 1997. “Analysis of Ribosomal RNA Genes Suggests That Trypanosomes Are Monophyletic.” *Journal of Molecular Evolution* 44 (5): 521–27.

Lunter, Gerton, and Martin Goodson. 2011. “Stampy: A Statistical Algorithm for Sensitive and Fast Mapping of Illumina Sequence Reads.” *Genome Research* 21 (6): 936–39. doi:10.1101/gr.111120.110.

Lye, Lon-Fye, Katherine Owens, Huafang Shi, Silvane M. F. Murta, Ana Carolina Vieira, Salvatore J. Turco, Christian Tschudi, Elisabetta Ullu, and Stephen M. Beverley. 2010. “Retention and Loss of RNA Interference Pathways in Trypanosomatid Protozoans.” *PLoS Pathog* 6 (10): e1001161. doi:10.1371/journal.ppat.1001161.

Maguire, G. P., I. Bastian, S. Arianayagam, A. Bryceson, and B. J. Currie. 1998. “New World Cutaneous Leishmaniasis Imported into Australia.” *Pathology* 30 (1): 73–76.

Manfredi, R., M. A. di Bari, L. Calza, and F. Chiodo. 2001. “American Cutaneous Leishmaniasis as a Rare Imported Disease in Europe: A Case Report Favourably Treated with Antimonial Derivatives.” *European Journal of Epidemiology* 17 (8): 793–95.

Mannaert, An, Tim Downing, Hideo Imamura, and Jean-Claude Dujardin. 2012. “Adaptive Mechanisms in Pathogens: Universal Aneuploidy in Leishmania.” *Trends in Parasitology* 28 (9): 370–76. doi:10.1016/j.pt.2012.06.003.

Martínez-Calvillo, Santiago, Kenneth Stuart, and Peter J. Myler. 2005. “Ploidy Changes Associated with Disruption of Two Adjacent Genes on Leishmania Major Chromosome 1.” *International Journal for Parasitology* 35 (4): 419–29. doi:10.1016/j.ijpara.2004.12.014.

Martin, Isabelle, Salvatore J. Turco, Richard M. Epand, and Jean-Marie Ruysschaert. 1998. “Lipophosphoglycan of Leishmania Donovani Inhibits Lipid Vesicle Fusion Induced by the N-Terminal Extremity of Viral Fusogenic Simian Immunodeficiency Virus Protein.” *European Journal of Biochemistry* 258 (1): 150–56. doi:10.1046/j.1432-1327.1998.2580150.x.

Martin, Jessica L., Phillip A. Yates, Radika Soysa, Joshua F. Alfaro, Feng Yang, Kristin E. Burnum-Johnson, Vladislav A. Petyuk, et al. 2014. “Metabolic Reprogramming during Purine Stress in the Protozoan Pathogen Leishmania Donovani.” *PLoS Pathog* 10 (2): e1003938. doi:10.1371/journal.ppat.1003938.

Martínez-Calvillo, Santiago, Shaofeng Yan, Dan Nguyen, Mark Fox, Kenneth Stuart, and Peter J Myler. 2003. “Transcription of Leishmania Major Friedlin Chromosome 1 Initiates in Both Directions within a Single Region.” *Molecular Cell* 11 (5): 1291–99. doi:10.1016/S1097-2765(03)00143-6.

Matthews, K R, C Tschudi, and E Ullu. 1994. “A Common Pyrimidine-Rich Motif Governs Trans-Splicing and Polyadenylation of Tubulin Polycistronic Pre-mRNA in Trypanosomes.” *Genes & Development* 8 (4): 491–501.

McCall, Laura-Isobel, and Greg Matlashewski. 2010. “Localization and Induction of the A2 Virulence Factor in Leishmania: Evidence That A2 Is a Stress Response Protein.” *Molecular Microbiology* 77 (2): 518–30. doi:10.1111/j.1365-2958.2010.07229.x.

McNicoll, François, Jolyne Drummelsmith, Michaela Müller, Eric Madore, Nathalie Boilard, Marc Ouellette, and Barbara Papadopoulou. 2006. “A Combined Proteomic and Transcriptomic Approach to the Study of Stage Differentiation in Leishmania Infantum.” *Proteomics* 6 (12): 3567–81. doi:10.1002/pmic.200500853.

McNicoll, François, Michaela Müller, Serge Cloutier, Nathalie Boilard, Annie Rochette, Marthe Dubé, and Barbara Papadopoulou. 2005. “Distinct 3’-Untranslated Region Elements Regulate Stage-Specific mRNA Accumulation and Translation in Leishmania.” *Journal of Biological Chemistry* 280 (42): 35238–46.

Meade, J. C., K. M. Hudson, S. L. Stringer, and J. R. Stringer. 1989. “A Tandem Pair of Leishmania Donovani Cation Transporting ATPase Genes Encode Isoforms That Are Differentially Expressed.” *Molecular and Biochemical Parasitology* 33 (1): 81–91.

Meade, J. C., J. Shaw, S. Lemaster, G. Gallagher, and J. R. Stringer. 1987. “Structure and Expression of a Tandem Gene Pair in Leishmania Donovani That Encodes a Protein Structurally Homologous to Eucaryotic Cation-Transporting ATPases.” *Molecular and Cellular Biology* 7 (11): 3937–46. doi:10.1128/MCB.7.11.3937.

Medina-Acosta, Enrique, Roger E. Karess, Heinz Schwartz, and David G. Russell. 1989. “The Promastigote Surface Protease (gp63) of Leishmania Is Expressed but Differentially Processed and Localized in the Amastigote Stage.” *Molecular and Biochemical Parasitology* 37 (2): 263–73. doi:10.1016/0166-6851(89)90158-8.

Miguel, Danilo C, Andrew R Flannery, Bidyottam Mittra, and Norma W Andrews. 2013. “Heme Uptake Mediated by LHR1 Is Essential for Leishmania Amazonensis Virulence.” *Infection and Immunity*, July. doi:10.1128/IAI.00687-13.

Miller, R. A., S. G. Reed, and M. Parsons. 1990. “Leishmania gp63 Molecule Implicated in Cellular Adhesion Lacks an Arg-Gly-Asp Sequence.” *Molecular and Biochemical Parasitology* 39 (2): 267–74.

Mishra, Krishna K., Timothy R. Holzer, Landon L. Moore, and Jonathan H. LeBowitz. 2003. “A Negative Regulatory Element Controls mRNA Abundance of the Leishmania Mexicana Paraflagellar Rod Gene PFR2.” *Eukaryotic Cell* 2 (5): 1009–17.

Mittra, Bidyottam, Mauro Cortez, Andrew Haydock, Gowthaman Ramasamy, Peter J. Myler, and Norma W. Andrews. 2013. “Iron Uptake Controls the Generation of Leishmania Infective Forms through Regulation of ROS Levels.” *The Journal of Experimental Medicine* 210 (2): 401–16. doi:10.1084/jem.20121368.

Monnerat, Séverine, Santiago Martinez-Calvillo, Elizabeth Worthey, Peter J. Myler, Kenneth D. Stuart, and Nicolas Fasel. 2004. “Genomic Organization and Gene Expression in a Chromosomal Region of Leishmania Major.” *Molecular and Biochemical Parasitology* 134 (2): 233–43. doi:10.1016/j.molbiopara.2003.12.004.

Monod, J., D. Perrin, C. Sanchez, and J. Monod. 1960. “L’Operon: Groupe de Genes a Expression Coordonee Par Un Operateur - Google Search.” https://www.google.co.uk/search?q=L%27Operon%3A+Groupe+de+genes+a+expression+coordonee+par+un+operateur&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a&channel=sb&gfe_rd=cr&ei=qrn0U5S9HvOq8we5sIGABA.

Moore, Landon L., Cecilia Santrich, and Jonathan H. LeBowitz. 1996. “Stage-Specific Expression of the Leishmania Mexicana Paraflagellar Rod Protein PFR-2.” *Molecular and Biochemical Parasitology* 80 (2): 125–35. doi:10.1016/0166-6851(96)02688-6.

Morales, Miguel A., Reiko Watanabe, Mariko Dacher, Philippe Chafey, José Osorio y Fortéa, David A. Scott, Stephen M. Beverley, et al. 2010. “Phosphoproteome Dynamics Reveal Heat-Shock Protein Complexes Specific to the Leishmania Donovani Infectious Stage.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (18): 8381–86. doi:10.1073/pnas.0914768107.

Morales, Miguel A., Reiko Watanabe, Christine Laurent, Pascal Lenormand, Jean-Claude Rousselle, Abdelkader Namane, and Gerald F. Späth. 2008. “Phosphoproteomic Analysis of Leishmania Donovani pro- and Amastigote Stages.” *Proteomics* 8 (2): 350–63. doi:10.1002/pmic.200700697.

Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28. doi:10.1038/nmeth.1226.

Mottram, J. C., M. J. Frame, D. R. Brooks, L. Tetley, J. E. Hutchison, A. E. Souza, and G. H. Coombs. 1997. “The Multiple Cpb Cysteine Proteinase Genes of Leishmania Mexicana Encode Isoenzymes That Differ in Their Stage Regulation and Substrate Preferences.” *The Journal of Biological Chemistry* 272 (22): 14285–93.

Mottram, J. C., A. E. Souza, J. E. Hutchison, R. Carter, M. J. Frame, and G. H. Coombs. 1996. “Evidence from Disruption of the Lmcpb Gene Array of Leishmania Mexicana That Cysteine Proteinases Are Virulence Factors.” *Proceedings of the National Academy of Sciences of the United States of America* 93 (12): 6008–13.

Mottram, Jeremy C., and Graham H. Coombs. 1985. “Leishmania Mexicana: Enzyme Activities of Amastigotes and Promastigotes and Their Inhibition by Antimonials and Arsenicals.” *Experimental Parasitology* 59 (2): 151–60. doi:10.1016/0014-4894(85)90067-0.

Muhich, M. L., and J. C. Boothroyd. 1988. “Polycistronic Transcripts in Trypanosomes and Their Accumulation during Heat Shock: Evidence for a Precursor Role in mRNA Synthesis.” *Molecular and Cellular Biology* 8 (9): 3837–46.

Mukherjee, Angana, Sébastien Boisvert, Rubens Lima do Monte-Neto, Adriano C. Coelho, Frederic Raymond, Rita Mukhopadhyay, Jacques Corbeil, and Marc Ouellette. 2013. “Telomeric Gene Deletion and Intrachromosomal Amplification in Antimony-Resistant Leishmania.” *Molecular Microbiology* 88 (1): 189–202. doi:10.1111/mmi.12178.

Müller, Michaela, Prasad K Padmanabhan, Annie Rochette, Debdutta Mukherjee, Martin Smith, Carole Dumas, and Barbara Papadopoulou. 2010. “Rapid Decay of Unstable

Leishmania mRNAs Bearing a Conserved Retroposon Signature 3'-UTR Motif Is Initiated by a Site-Specific Endonucleolytic Cleavage without Prior Deadenylation." *Nucleic Acids Research* 38 (17): 5867–83. doi:10.1093/nar/gkq349.

Müller, Michaela, and Barbara Papadopoulou. 2010. "Stage-Specific Expression of the Glycine Cleavage Complex Subunits in Leishmania Infantum." *Molecular and Biochemical Parasitology* 170 (1): 17–27. doi:10.1016/j.molbiopara.2009.11.009.

Murray, Angus, Christine Fu, Golareh Habibi, and W Robert McMaster. 2007. "Regions in the 3' Untranslated Region Confer Stage-Specific Expression to the Leishmania Mexicana a600-4 Gene." *Molecular and Biochemical Parasitology* 153 (2): 125–32. doi:10.1016/j.molbiopara.2007.02.010.

Murray, Angus S., Miriam A. Lynn, and W. Robert McMaster. 2010. "The Leishmania Mexicana A600 Genes Are Functionally Required for Amastigote Replication." *Molecular and Biochemical Parasitology* 172 (2): 80–89. doi:10.1016/j.molbiopara.2010.03.008.

Murray, Henry W, Jonathan D Berman, Clive R Davies, and Nancy G Saravia. 2005. "Advances in Leishmaniasis." *The Lancet* 366 (9496): 1561–77. doi:10.1016/S0140-6736(05)67629-5.

Myler, P. J., L. Audleman, T. deVos, G. Hixson, P. Kiser, C. Lemley, C. Magness, et al. 1999. "Leishmania Major Friedlin Chromosome 1 Has an Unusual Distribution of Protein-Coding Genes." *Proceedings of the National Academy of Sciences of the United States of America* 96 (6): 2902–6.

Myler, P. J., S. M. Beverley, A. K. Cruz, D. E. Dobson, A. C. Ivens, P. D. McDonagh, R. Madhubala, et al. 2001. "The Leishmania Genome Project: New Insights into Gene Organization and Function." *Medical Microbiology and Immunology* 190 (1-2): 9–12.

Myler, P. J., E. Sisk, P. D. McDonagh, S. Martinez-Calvillo, A. Schnaufer, S. M. Sunkin, S. Yan, R. Madhubala, A. Ivens, and K. Stuart. 2000. "Genomic Organization and Gene Function in Leishmania." *Biochemical Society Transactions* 28 (5): 527–31.

Nagalakshmi, Ugrappa, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. 2008. "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing." *Science* 320 (5881): 1344–49. doi:10.1126/science.1158441.

Nature. 2002. "Initial Sequencing and Comparative Analysis of the Mouse Genome" 420 (6915): 520–62. doi:10.1038/nature01262.

Nilsson, Daniel, Kapila Gunasekera, Jan Mani, Magne Osteras, Laurent Farinelli, Loic Baerlocher, Isabel Roditi, and Torsten Ochsenreiter. 2010. "Spliced Leader Trapping Reveals Widespread Alternative Splicing Patterns in the Highly Dynamic Transcriptome of Trypanosoma Brucei." *PLoS Pathogens* 6 (8): e1001037. doi:10.1371/journal.ppat.1001037.

Nirujogi, Raja Sekhar, Harsh Pawar, Santosh Renuse, Praveen Kumar, Sandip Chavan, Gajanan Sathe, Jyoti Sharma, et al. 2014. "Moving from Unsequenced to Sequenced Genome: Reanalysis of the Proteome of Leishmania Donovani." *Journal of Proteomics*, Special Issue: Trends in Microbial Proteomics, 97 (January): 48–61. doi:10.1016/j.jprot.2013.04.021.

Nugent, Philip G., Saiful A. Karsani, Robin Wait, Jane Tempero, and Deborah F. Smith. 2004. "Proteomic Analysis of Leishmania Mexicana Differentiation." *Molecular and Biochemical Parasitology* 136 (1): 51–62. doi:10.1016/j.molbiopara.2004.02.009.

Ochsenreiter, Torsten, Sedrick Anderson, Zachary A. Wood, and Stephen L. Hajduk. 2008. "Alternative RNA Editing Produces a Novel Protein Involved in Mitochondrial DNA Maintenance in Trypanosomes." *Molecular and Cellular Biology* 28 (18): 5595–5604. doi:10.1128/MCB.00637-08.

Ogbadoyi, Emmanuel O., Derrick R. Robinson, and Keith Gull. 2003. "A High-Order Trans-Membrane Structural Linkage Is Responsible for Mitochondrial Genome Positioning

and Segregation by Flagellar Basal Bodies in Trypanosomes." *Molecular Biology of the Cell* 14 (5): 1769–79. doi:10.1091/mbc.E02-08-0525.

Oshlack, Alicia, and Matthew J Wakefield. 2009. "Transcript Length Bias in RNA-Seq Data Confounds Systems Biology." *Biology Direct* 4: 14. doi:10.1186/1745-6150-4-14.

Paape, Daniel, Martin E. Barrios-Llerena, Thierry Le Bihan, Logan Mackay, and Toni Aebischer. 2010. "Gel Free Analysis of the Proteome of Intracellular Leishmania Mexicana." *Molecular and Biochemical Parasitology* 169 (2): 108–14. doi:10.1016/j.molbiopara.2009.10.009.

Paape, Daniel, Christoph Lippuner, Monika Schmid, Renate Ackermann, Martin E. Barrios-Llerena, Ursula Zimny-Arndt, Volker Brinkmann, et al. 2008. "Transgenic, Fluorescent Leishmania Mexicana Allow Direct Analysis of the Proteome of Intracellular Amastigotes." *Molecular & Cellular Proteomics* 7 (9): 1688–1701. doi:10.1074/mcp.M700343-MCP200.

Pace, David. 2014. "Leishmaniasis." *Journal of Infection*. Accessed September 28. doi:10.1016/j.jinf.2014.07.016.

Padmanabhan, Prasad K, Carole Dumas, Mukesh Samant, Annie Rochette, Martin J Simard, and Barbara Papadopoulou. 2012. "Novel Features of a PIWI-like Protein Homolog in the Parasitic Protozoan Leishmania." *PloS One* 7 (12): e52612. doi:10.1371/journal.pone.0052612.

Pan, A. A. 1986. "Leishmania Mexicana Pifanoi: Analysis of the Antigenic Relationships between Promastigotes and Amastigotes by Gel Diffusion, Immunoelectrophoresis, and Immunoprecipitation." *The Journal of Protozoology* 33 (2): 192–97.

Pan, A. A., and D. McMahon-Pratt. 1988. "Monoclonal Antibodies Specific for the Amastigote Stage of Leishmania Pifanoi. I. Characterization of Antigens Associated with Stage- and Species-Specific Determinants." *Journal of Immunology (Baltimore, Md.: 1950)* 140 (7): 2406–14.

Paniz Mondolfi, Alberto E., Gabrielle Baker Duffey, Lucy E. Horton, Mariantonieta Tirado, Oscar Reyes Jaimes, Alexandra Perez-Alvarez, and Olga Zerpa. 2013. "Intermediate/borderline Disseminated Cutaneous Leishmaniasis." *International Journal of Dermatology* 52 (4): 446–55. doi:10.1111/j.1365-4632.2012.05709.x.

Peacock, Christopher S., Kathy Seeger, David Harris, Lee Murphy, Jeronimo C. Ruiz, Michael A. Quail, Nick Peters, et al. 2007. "Comparative Genomic Analysis of Three Leishmania Species That Cause Diverse Human Disease." *Nature Genetics* 39 (7): 839–47. doi:10.1038/ng2053.

Pédelacq, Jean-Denis, Stéphanie Cabantous, Timothy Tran, Thomas C. Terwilliger, and Geoffrey S. Waldo. 2006. "Engineering and Characterization of a Superfolder Green Fluorescent Protein." *Nature Biotechnology* 24 (1): 79–88. doi:10.1038/nbt1172.

Pham, Nam-Kha, Jennifer Mouriz, and Peter E. Kima. 2005. "Leishmania Pifanoi Amastigotes Avoid Macrophage Production of Superoxide by Inducing Heme Degradation." *Infection and Immunity* 73 (12): 8322–33. doi:10.1128/IAI.73.12.8322-8333.2005.

Polando, Rachel, Upasna Gaur Dixit, Cristina R. Carter, Blake Jones, James P. Whitcomb, Wibke Ballhorn, Melissa Harintho, Christopher L. Jerde, Mary E. Wilson, and Mary Ann McDowell. 2013. "The Roles of Complement Receptor 3 and Fcγ Receptors during Leishmania Phagosome Maturation." *Journal of Leukocyte Biology* 93 (6): 921–32. doi:10.1189/jlb.0212086.

Porcel, Betina M, France Denoeud, Fred Opperdoes, Benjamin Noel, Mohammed-Amine Madoui, Tansy C Hammarton, Mark C Field, et al. 2014. "The Streamlined Genome of Phytomonas Spp. Relative to Human Pathogenic Kinetoplastids Reveals a Parasite Tailored for Plants." *PLoS Genetics* 10 (2): e1004007. doi:10.1371/journal.pgen.1004007.

Porter-Kelley, Johanna M., Noel J. Gerald, Juan C. Engel, Elodie Ghedin, and Dennis M. Dwyer. 2004. “LdARF1 in Trafficking and Structural Maintenance of the Trans-Golgi Cisternal Network in the Protozoan Pathogen Leishmania Donovani.” *Traffic (Copenhagen, Denmark)* 5 (11): 868–83. doi:10.1111/j.1600-0854.2004.00229.x.

Ramiro, M. J., T. Hanke, S. Taladriz, and V. Larraga. 2002. “DNA Polymerase Beta mRNA Determination by Relative Quantitative RT-PCR from Leishmania Infantum Intracellular Amastigotes.” *Parasitology Research* 88 (8): 760–67. doi:10.1007/s00436-002-0653-0.

Rastrojo, Alberto, Fernando Carrasco-Ramiro, Diana Martín, Antonio Crespillo, Rosa M Reguera, Begoña Aguado, and Jose M Requena. 2013. “The Transcriptome of Leishmania Major in the Axenic Promastigote Stage: Transcript Annotation and Relative Expression Levels by RNA-Seq.” *BMC Genomics* 14 (1): 223. doi:10.1186/1471-2164-14-223.

Raymond, Frédéric, Sébastien Boisvert, Gaétan Roy, Jean-François Ritt, Danielle Légaré, Amandine Isnard, Mario Stanke, et al. 2011. “Genome Sequencing of the Lizard Parasite Leishmania Tarentolae Reveals Loss of Genes Associated to the Intracellular Stage of Human Pathogenic Species.” *Nucleic Acids Research*, October. doi:10.1093/nar/gkr834.

Ready, Paul D. 2013. “Biology of Phlebotomine Sand Flies as Vectors of Disease Agents.” *Annual Review of Entomology* 58 (1): 227–50. doi:10.1146/annurev-ento-120811-153557.

Ready, P. D. 2010. “Leishmaniasis Emergence in Europe.” *Euro Surveillance: Bulletin Européen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 15 (10): 19505.

Rettig, Jochen, Yimu Wang, André Schneider, and Torsten Ochsenreiter. 2012. “Dual Targeting of Isoleucyl-tRNA Synthetase in Trypanosoma Brucei Is Mediated through Alternative Trans-Splicing.” *Nucleic Acids Research* 40 (3): 1299–1306. doi:10.1093/nar/gkr794.

Reynolds, David, Laura Cliffe, Konrad U. Förstner, Chung-Chau Hon, T. Nicolai Siegel, and Robert Sabatini. 2014. “Regulation of Transcription Termination by Glucosylated Hydroxymethyluracil, Base J, in Leishmania Major and Trypanosoma Brucei.” *Nucleic Acids Research*, August. doi:10.1093/nar/gku714.

Ritter, Uwe, Freddy Frischknecht, and Ger van Zandbergen. 2009. “Are Neutrophils Important Host Cells for Leishmania Parasites?” *Trends in Parasitology* 25 (11): 505–10. doi:10.1016/j.pt.2009.08.003.

Robertson, C. D., and G. H. Coombs. 1993. “Cathepsin B-like Cysteine Proteases of Leishmania Mexicana.” *Molecular and Biochemical Parasitology* 62 (2): 271–79.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26. doi:10.1038/nbt.1754.

Rochette, A., François McNicoll, Joanne Girard, Marie Breton, Éric Leblanc, Michel G. Bergeron, and Barbara Papadopoulou. 2005. “Characterization and Developmental Gene Regulation of a Large Gene Family Encoding Amastin Surface Proteins in Leishmania Spp.” *Molecular and Biochemical Parasitology* 140 (2): 205–20. doi:10.1016/j.molbiopara.2005.01.006.

Rochette, Annie, Frédéric Raymond, Jacques Corbeil, Marc Ouellette, and Barbara Papadopoulou. 2009. “Whole-Genome Comparative RNA Expression Profiling of Axenic and Intracellular Amastigote Forms of Leishmania Infantum.” *Molecular and Biochemical Parasitology* 165 (1): 32–47. doi:10.1016/j.molbiopara.2008.12.012.

Rochette, Annie, Frédéric Raymond, Jean-Michel Ubeda, Martin Smith, Nadine Messier, Sébastien Boisvert, Philippe Rigault, Jacques Corbeil, Marc Ouellette, and Barbara Papadopoulou. 2008. “Genome-Wide Gene Expression Profiling Analysis of Leishmania

Major and Leishmania Infantum Developmental Stages Reveals Substantial Differences between the Two Species." *BMC Genomics* 9 (1): 255. doi:10.1186/1471-2164-9-255.
Rogers, L. 1904. *Quart. J. Micr. Sci.*, no. 48: 367.
Rogers, Matthew B., James D. Hilley, Nicholas J. Dickens, Jon Wilkes, Paul A. Bates, Daniel P. Depledge, David Harris, et al. 2011. "Chromosome and Gene Copy Number Variation Allow Major Structural Change between Species and Strains of Leishmania." *Genome Research* 21 (12): 2129–42. doi:10.1101/gr.122945.111.
Rogers, Matthew E., Martina Hajmová, Manju B. Joshi, Jovana Sadlova, Dennis M. Dwyer, Petr Volf, and Paul A. Bates. 2008. "Leishmania Chitinase Facilitates Colonization of Sand Fly Vectors and Enhances Transmission to Mice." *Cellular Microbiology* 10 (6): 1363–72. doi:10.1111/j.1462-5822.2008.01132.x.
Rogers, Matthew E., Thomas Ilg, Andrei V. Nikolaev, Michael A. J. Ferguson, and Paul A. Bates. 2004. "Transmission of Cutaneous Leishmaniasis by Sand Flies Is Enhanced by Regurgitation of fPPG." *Nature* 430 (6998): 463–67. doi:10.1038/nature02675.
Rogers, M. E., M. L. Chance, and P. A. Bates. 2002. "The Role of Promastigote Secretory Gel in the Origin and Transmission of the Infective Stage of Leishmania Mexicana by the Sandfly Lutzomyia Longipalpis." *Parasitology* 124 (Pt 5): 495–507.
Rosenzweig, Doron, Derek Smith, Peter J. Myler, Robert W. Olafson, and Dan Zilberstein. 2008. "Post-Translational Modification of Cellular Proteins during Leishmania Donovani Differentiation." *Proteomics* 8 (9): 1843–50. doi:10.1002/pmic.200701043.
Rosenzweig, Doron, Derek Smith, Fred Opperdoes, Shay Stern, Robert W. Olafson, and Dan Zilberstein. 2008. "Retooling Leishmania Metabolism: From Sand Fly Gut to Human Macrophage." *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* 22 (2): 590–602. doi:10.1096/fj.07-9254com.
Ross, Philip L., Yulin N. Huang, Jason N. Marchese, Brian Williamson, Kenneth Parker, Stephen Hattan, Nikita Khainovski, et al. 2004. "Multiplexed Protein Quantitation in Saccharomyces Cerevisiae Using Amine-Reactive Isobaric Tagging Reagents." *Molecular & Cellular Proteomics: MCP* 3 (12): 1154–69. doi:10.1074/mcp.M400129-MCP200.
Ruiz-Orera, Jorge, Xavier Messeguer, Juan Antonio Subirana, and M. Mar Alba. 2014. "Long Non-Coding RNAs as a Source of New Peptides." *eLife* 3 (September): e03523. doi:10.7554/eLife.03523.
Russell, D G, R J Newsam, G C Palmer, and K Gull. 1983. "Structural and Biochemical Characterisation of the Paraflagellar Rod of Crithidia Fasciculata." *European Journal of Cell Biology* 30 (1): 137–43.
Russell, D. G., and S. D. Wright. 1988. "Complement Receptor Type 3 (CR3) Binds to an Arg-Gly-Asp-Containing Region of the Major Surface Glycoprotein, gp63, of Leishmania Promastigotes." *The Journal of Experimental Medicine* 168 (1): 279–92.
Saar, Yehoshua, Asamoa Ransford, Ella Waldman, Salam Mazareb, Sigal Amin-Spector, Jodie Plumblee, Salvatore J Turco, and Dan Zilberstein. 1998. "Characterization of Developmentally-Regulated Activities in Axenic Amastigotes of Leishmania Donovani." *Molecular and Biochemical Parasitology* 95 (1): 9–20. doi:10.1016/S0166-6851(98)00062-0.
Salzberg, S. L., A. L. Delcher, S. Kasif, and O. White. 1998. "Microbial Gene Identification Using Interpolated Markov Models." *Nucleic Acids Research* 26 (2): 544–48.
Saxena, A., T. Lahav, N. Holland, G. Aggarwal, A. Anupama, Y. Huang, H. Volpin, P.J. Myler, and D. Zilberstein. 2007. "Analysis of the Leishmania Donovani Transcriptome Reveals an Ordered Progression of Transient and Permanent Changes in Gene Expression during Differentiation." *Molecular and Biochemical Parasitology* 152 (1): 53–65. doi:10.1016/j.molbiopara.2006.11.011.

Saxena, Elizabeth A Worthey, Shaofeng Yan, Aaron Leland, Kenneth D Stuart, and Peter J Myler. 2003. “Evaluation of Differential Gene Expression in Leishmania Major Friedlin Procyclics and Metacyclics Using DNA Microarray Analysis.” *Molecular and Biochemical Parasitology* 129 (1): 103–14.

Schelipewsky, E. 1912. “Fadenförmige Anhängsel Bei Den Trypanosomen.” *Zbl. Bakt.*

Schnarwiler, Felix, Moritz Niemann, Nicholas Doiron, Anke Harsman, Sandro Käser, Jan Mani, Astrid Chanfon, et al. 2014. “Trypanosomal TAC40 Constitutes a Novel Subclass of Mitochondrial B-Barrel Proteins Specialized in Mitochondrial Genome Inheritance.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (21): 7624–29. doi:10.1073/pnas.1404854111.

Schroeder, Andreas, Odilo Mueller, Susanne Stocker, Ruediger Salowsky, Michael Leiber, Marcus Gassmann, Samar Lightfoot, Wolfram Menzel, Martin Granzow, and Thomas Ragg. 2006. “The RIN: An RNA Integrity Number for Assigning Integrity Values to RNA Measurements.” *BMC Molecular Biology* 7 (1): 3.

Schürch, N, A Hehl, E Vassella, R Braun, and I Roditi. 1994. “Accurate Polyadenylation of Procyclin mRNAs in Trypanosoma Brucei Is Determined by Pyrimidine-Rich Elements in the Intergenic Regions.” *Molecular and Cellular Biology* 14 (6): 3668–75.

Sendler, Edward, Graham D. Johnson, and Stephen A. Krawetz. 2011. “Local and Global Factors Affecting RNA Sequencing Analysis.” *Analytical Biochemistry* 419 (2): 317–22. doi:10.1016/j.ab.2011.08.013.

Sengupta, Shantanu, Jalaj Tripathi, Ruchi Tandon, Manoj Raje, Rajendra P. Roy, Sandip K. Basu, and Amitabha Mukhopadhyay. 1999. “Hemoglobin Endocytosis in Leishmania Is Mediated through a 46-kDa Protein Located in the Flagellar Pocket.” *Journal of Biological Chemistry* 274 (5): 2758–65. doi:10.1074/jbc.274.5.2758.

Sereno, D., M. Cavaleyra, K. Zemzoumi, S. Maquaire, A. Ouaissi, and J. L. Lemesre. 1998. “Axenically Grown Amastigotes of Leishmania Infantum Used as an in Vitro Model to Investigate the Pentavalent Antimony Mode of Action.” *Antimicrobial Agents and Chemotherapy* 42 (12): 3097–3102.

Shapira, M., J. G. McEwen, and C. L. Jaffe. 1988. “Temperature Effects on Molecular Processes Which Lead to Stage Differentiation in Leishmania.” *The EMBO Journal* 7 (9): 2895–2901.

Siegel, Kapila Gunasekera, George A. M. Cross, and Torsten Ochsenreiter. 2011. “Gene Expression in Trypanosoma Brucei: Lessons from High-Throughput RNA Sequencing.” *Trends in Parasitology* 27 (10): 434–41. doi:10.1016/j.pt.2011.05.006.

Siegel, Doeke R Hekstra, Louise E Kemp, Luisa M Figueiredo, Joanna E Lowell, David Fenyo, Xuning Wang, Scott Dewell, and George A M Cross. 2009. “Four Histone Variants Mark the Boundaries of Polycistronic Transcription Units in Trypanosoma Brucei.” *Genes & Development* 23 (9): 1063–76. doi:10.1101/gad.1790409.

Siegel, Doeke R Hekstra, Xuning Wang, Scott Dewell, and George A M Cross. 2010. “Genome-Wide Analysis of mRNA Abundance in Two Life-Cycle Stages of Trypanosoma Brucei and Identification of Splicing and Polyadenylation Sites.” *Nucleic Acids Research* 38 (15): 4946–57. doi:10.1093/nar/gkq237.

Silverman, J Maxwell, Simon K Chan, Dale P Robinson, Dennis M Dwyer, Devki Nandan, Leonard J Foster, and Neil E Reiner. 2008. “Proteomic Analysis of the Secretome of Leishmania Donovani.” *Genome Biology* 9 (2): R35. doi:10.1186/gb-2008-9-2-r35.

Silverman, Judith Maxwell, Joachim Clos, Carolina Camargo de’ Oliveira, Omid Shirvani, Yuan Fang, Christine Wang, Leonard J. Foster, and Neil E. Reiner. 2010. “An Exosome-Based Secretion Pathway Is Responsible for Protein Export from Leishmania and Communication with Macrophages.” *Journal of Cell Science* 123 (6): 842–52. doi:10.1242/jcs.056465.

Siman-Tov, M. M., R. Aly, M. Shapira, and C. L. Jaffe. 1996. “Cloning from Leishmania Major of a Developmentally Regulated Gene, c-lpk2, for the Catalytic Subunit of the cAMP-Dependent Protein Kinase.” *Molecular and Biochemical Parasitology* 77 (2): 201–15.

Singh, Nisha, Manish Kumar, and Rakesh Kumar Singh. 2012. “Leishmaniasis: Current Status of Available Drugs and New Potential Drug Targets.” *Asian Pacific Journal of Tropical Medicine* 5 (6): 485–97. doi:10.1016/S1995-7645(12)60084-4.

Smircich, Pablo, Diego Forteza, Najib M. El-Sayed, and Beatriz Garat. 2013. “Genomic Analysis of Sequence-Dependent DNA Curvature in Leishmania.” *PloS One* 8 (4): e63068. doi:10.1371/journal.pone.0063068.

Somanna, Ashwini, Vasanthakrishna Mundodi, and Lashitew Gedamu. 2002. “In Vitro Cultivation and Characterization of Leishmania Chagasi Amastigote-like Forms.” *Acta Tropica* 83 (1): 37–42. doi:10.1016/S0001-706X(02)00054-2.

Sopwith, William F., Alain Debrabant, Mat Yamage, Dennis M. Dwyer, and Paul A. Bates. 2002. “Developmentally Regulated Expression of a Cell Surface Class I Nuclease in Leishmania Mexicana.” *International Journal for Parasitology* 32 (4): 449–59.

Souza, A. E., S. Waugh, G. H. Coombs, and J. C. Mottram. 1992. “Characterization of a Multi-Copy Gene for a Major Stage-Specific Cysteine Proteinase of Leishmania Mexicana.” *FEBS Letters* 311 (2): 124–27.

Springer, Timothy, Giovanni Galfré, David S. Secher, and Cesar Milstein. 1979. “Mac-1: A Macrophage Differentiation Antigen Identified by Monoclonal Antibody.” *European Journal of Immunology* 9 (4): 301–6. doi:10.1002/eji.1830090410.

Srivastava, S, S P Pandey, M K Jha, H S Chandel, and B Saha. 2013. “Leishmania Expressed Lipophosphoglycan Interacts with Toll-like Receptor (TLR)-2 to Decrease TLR-9 Expression and Reduce Anti-Leishmanial Responses.” *Clinical and Experimental Immunology* 172 (3): 403–9. doi:10.1111/cei.12074.

Stanke, Mario, Rasmus Steinkamp, Stephan Waack, and Burkhard Morgenstern. 2004. “AUGUSTUS: A Web Server for Gene Finding in Eukaryotes.” *Nucleic Acids Research* 32 (suppl 2): W309–12. doi:10.1093/nar/gkh379.

Stanke, Mario, Ana Tzvetkova, and Burkhard Morgenstern. 2006. “AUGUSTUS at EGASP: Using EST, Protein and Genomic Alignments for Improved Gene Prediction in the Human Genome.” *Genome Biology* 7 (Suppl 1): S11. doi:10.1186/gb-2006-7-s1-s11.

Sterkers, Yvon, Laurence Lachaud, Nathalie Bourgeois, Lucien Crobu, Patrick Bastien, and Michel Pagès. 2012. “Novel Insights into Genome Plasticity in Eukaryotes: Mosaic Aneuploidy in Leishmania.” *Molecular Microbiology* 86 (1): 15–23. doi:10.1111/j.1365-2958.2012.08185.x.

Sterkers, Yvon, Laurence Lachaud, Lucien Crobu, Patrick Bastien, and Michel Pagès. 2011. “FISH Analysis Reveals Aneuploidy and Continual Generation of Chromosomal Mosaicism in Leishmania Major.” *Cellular Microbiology* 13 (2): 274–83. doi:10.1111/j.1462-5822.2010.01534.x.

Stierhof, Y. D., T. Ilg, D. G. Russell, H. Hohenberg, and P. Overath. 1994. “Characterization of Polymer Release from the Flagellar Pocket of Leishmania Mexicana Promastigotes.” *The Journal of Cell Biology* 125 (2): 321–31.

Storz, Gisela, Yuri I. Wolf, and Kumaran S. Ramamurthi. 2014. “Small Proteins Can No Longer Be Ignored.” *Annual Review of Biochemistry* 83: 753–77. doi:10.1146/annurev-biochem-070611-102400.

Sunkin, S. M., P. Kiser, P. J. Myler, and K. Stuart. 2000. “The Size Difference between Leishmania Major Friedlin Chromosome One Homologues Is Localized to Sub-Telomeric Repeats at One Chromosomal End.” *Molecular and Biochemical Parasitology* 109 (1): 1–15.

Sutcliffe, Iain C., and Dean J. Harrington. 2002. “Pattern Searches for the Identification of Putative Lipoprotein Genes in Gram-Positive Bacterial Genomes.” *Microbiology (Reading, England)* 148 (Pt 7): 2065–77.

Sutton, Richard E., and John C. Boothroyd. 1986. “Evidence for Trans Splicing in Trypanosomes.” *Cell* 47 (4): 527–35. doi:10.1016/0092-8674(86)90617-3.

Teixeira, S. M., L. V. Kirchhoff, and J. E. Donelson. 1995. “Post-Transcriptional Elements Regulating Expression of mRNAs from the Amastin/tuzin Gene Cluster of Trypanosoma Cruzi.” *The Journal of Biological Chemistry* 270 (38): 22586–94.

Teixeira, S M, L V Kirchhoff, and J E Donelson. 1999. “Trypanosoma Cruzi: Suppression of Tuzin Gene Expression by Its 5’-UTR and Spliced Leader Addition Site.” *Experimental Parasitology* 93 (3): 143–51. doi:10.1006/expr.1999.4446.

Teixeira, S. M., D. G. Russell, L. V. Kirchhoff, and J. E. Donelson. 1994. “A Differentially Expressed Gene Family Encoding ‘Amastin,’ a Surface Protein of Trypanosoma Cruzi Amastigotes.” *Journal of Biological Chemistry* 269 (32): 20509–16.

Tetaud, Emmanuel, Isabelle Lecuix, Tara Sheldrake, Théo Baltz, and Alan H. Fairlamb. 2002. “A New Expression Vector for Crithidia Fasciculata and Leishmania.” *Molecular and Biochemical Parasitology* 120 (2): 195–204.

Thomas, Sean, Amanda Green, Nancy R. Sturm, David A. Campbell, and Peter J. Myler. 2009. “Histone Acetylations Mark Origins of Polycistronic Transcription in Leishmania Major.” *BMC Genomics* 10: 152. doi:10.1186/1471-2164-10-152.

Thorvaldsdóttir, Helga, James T. Robinson, and Jill P. Mesirov. 2013. “Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration.” *Briefings in Bioinformatics* 14 (2): 178–92. doi:10.1093/bib/bbs017.

Tran, Khoa D., Dayana Rodriguez-Contreras, Ujwal Shinde, and Scott M. Landfear. 2012. “Both Sequence and Context Are Important for Flagellar Targeting of a Glucose Transporter.” *Journal of Cell Science* 125 (14): 3293–98. doi:10.1242/jcs.103028.

Trudgian, David C, Gabriela Ridlova, Roman Fischer, Mukram M Mackeen, Nicola Ternette, Oreste Acuto, Benedikt M Kessler, and Benjamin Thomas. 2011. “Comparative Evaluation of Label-Free SINQ Normalized Spectral Index Quantitation in the Central Proteomics Facilities Pipeline.” *Proteomics* 11 (14): 2790–97. doi:10.1002/pmic.201000800.

Tsigankov, Polina, Pier Federico Gherardini, Manuela Helmer-Citterich, Gerald F. Spaeth, Peter J. Myler, and Dan Zilberstein. 2014. “Regulation Dynamics of Leishmania Differentiation: Deconvoluting Signals and Identifying Phosphorylation Trends.” *Molecular & Cellular Proteomics: MCP*, April. doi:10.1074/mcp.M114.037705.

Tsigankov, Polina, Pier Federico Gherardini, Manuela Helmer-Citterich, and Dan Zilberstein. 2012. “What Has Proteomics Taught Us about Leishmania Development?” *Parasitology* 139 (Special Issue 09): 1146–57. doi:10.1017/S0031182012000157.

Tull, Dedreia, James E. Vince, Judy M. Callaghan, Thomas Naderer, Tim Spurck, Geoffrey I. McFadden, Graeme Currie, Kris Ferguson, Antony Bacic, and Malcolm J. McConville. 2004. “SMP-1, a Member of a New Family of Small Myristoylated Proteins in Kinetoplastid Parasites, Is Targeted to the Flagellum Membrane in Leishmania.” *Molecular Biology of the Cell* 15 (11): 4775–86. doi:10.1091/mbc.E04-06-0457.

Turco, Salvatore J., and Albert Descoteaux. 1992. “The Lipophosphoglycan of Leishmania Parasites.” *Annual Review of Microbiology* 46 (1): 65–92. doi:10.1146/annurev.mi.46.100192.000433.

Turetz, Meredith L., Paulo R. Machado, Albert I. Ko, Fábio Alves, Achiléa Bittencourt, Roque P. Almeida, Niloufar Mobashery, Warren D. Johnson, and Edgar M. Carvalho. 2002. “Disseminated Leishmaniasis: A New and Emerging Form of Leishmaniasis Observed in Northeastern Brazil.” *Journal of Infectious Diseases* 186 (12): 1829–34. doi:10.1086/345772.

Ueno, Norikiyo, and Mary E. Wilson. 2012. “Receptor-Mediated Phagocytosis of Leishmania: Implications for Intracellular Survival.” *Trends in Parasitology* 28 (8): 335–44. doi:10.1016/j.pt.2012.05.002.

Ullu, E, K R Matthews, and C Tschudi. 1993. “Temporal Order of RNA-Processing Reactions in Trypanosomes: Rapid Trans Splicing Precedes Polyadenylation of Newly Synthesized Tubulin Transcripts.” *Molecular and Cellular Biology* 13 (1): 720–25.

Van der Ploeg, L. H. 1986. “Discontinuous Transcription and Splicing in Trypanosomes.” *Cell* 47 (4): 479–80.

Van Luenen, Henri G.A.M., Carol Farris, Sabrina Jan, Paul-Andre Genest, Pankaj Tripathi, Arno Velds, Ron M. Kerkhoven, et al. 2012. “Glucosylated Hydroxymethyluracil, DNA Base J, Prevents Transcriptional Readthrough in Leishmania.” *Cell* 150 (5): 909–21. doi:10.1016/j.cell.2012.07.030.

Vashist, Shilpa, and Davis T. W. Ng. 2004. “Misfolded Proteins Are Sorted by a Sequential Checkpoint Mechanism of ER Quality Control.” *The Journal of Cell Biology* 165 (1): 41–52. doi:10.1083/jcb.200309132.

Vasquez, Juan-José, Chung-Chau Hon, Jens T. Vanselow, Andreas Schlosser, and T. Nicolai Siegel. 2014. “Comparative Ribosome Profiling Reveals Extensive Translational Complexity in Different Trypanosoma Brucei Life Cycle Stages.” *Nucleic Acids Research* 42 (6): 3623–37. doi:10.1093/nar/gkt1386.

Villalba, Elena, Bias Dorta, and José Lues Ramírez. 1985. “Comparative Study of the Ribosomal RNA from Leishmania and Trypanosoma1.” *Journal of Eukaryotic Microbiology* 32 (1): 49–53. doi:10.1111/j.1550-7408.1985.tb03012.x.

Vinet, Adrien F., Mitsunori Fukuda, Salvatore J. Turco, and Albert Descoteaux. 2009. “The Leishmania Donovani Lipophosphoglycan Excludes the Vesicular Proton-ATPase from Phagosomes by Impairing the Recruitment of Synaptotagmin V.” *PLoS Pathogens* 5 (10): e1000628. doi:10.1371/journal.ppat.1000628.

Wall, D. P., H. B. Fraser, and A. E. Hirsh. 2003. “Detecting Putative Orthologs.” *Bioinformatics* 19 (13): 1710–11. doi:10.1093/bioinformatics/btg213.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nat Rev Genet* 10 (1): 57–63. doi:10.1038/nrg2484.

Webb, J. R., A. Campos-Neto, Y. A. Skeiky, and S. G. Reed. 1997. “Molecular Characterization of the Heat-Inducible LmSTI1 Protein of Leishmania Major.” *Molecular and Biochemical Parasitology* 89 (2): 179–93.

Wheeler, Richard J., Keith Gull, and Eva Gluenz. 2012. “Detailed Interrogation of Trypanosome Cell Biology via Differential Organelle Staining and Automated Image Analysis.” *BMC Biology* 10: 1. doi:10.1186/1741-7007-10-1.

WHO Report. 2013. “Second WHO Report on NTDs” Chapter 3.9*: 67–71.

Wickens, M. 1990. “How the Messenger Got Its Tail: Addition of poly(A) in the Nucleus.” *Trends in Biochemical Sciences* 15 (7): 277–81.

Wickstead, Bill, and Keith Gull. 2007. “Dyneins across Eukaryotes: A Comparative Genomic Analysis.” *Traffic (Copenhagen, Denmark)* 8 (12): 1708–21. doi:10.1111/j.1600-0854.2007.00646.x.

Wiese, M. 1998. “A Mitogen-Activated Protein (MAP) Kinase Homologue of Leishmania Mexicana Is Essential for Parasite Survival in the Infected Host.” *The EMBO Journal* 17 (9): 2619–28. doi:10.1093/emboj/17.9.2619.

Wildes, David, and James A. Wells. 2010. “Sampling the N-Terminal Proteome of Human Blood.” *Proceedings of the National Academy of Sciences* 107 (10): 4561–66. doi:10.1073/pnas.0914495107.

Winberg, Martin E., Asa Holm, Eva Särndahl, Adrien F. Vinet, Albert Descoteaux, Karl-Eric Magnusson, Birgitta Rasmusson, and Maria Lerm. 2009. “Leishmania Donovani Lipophosphoglycan Inhibits Phagosomal Maturation via Action on Membrane Rafts.”

Microbes and Infection / Institut Pasteur 11 (2): 215–22. doi:10.1016/j.micinf.2008.11.007.

Wincker, Patrick, Christophe Ravel, Christine Blaineau, Michel Pages, Yann Jauffret, Jean-Pierre Dedet, and Patrick Bastien. 1996. "The Leishmania Genome Comprises 36 Chromosomes Conserved Across Widely Divergent Human Pathogenic Species." *Nucleic Acids Research* 24 (9): 1688–94. doi:10.1093/nar/24.9.1688.

Winter, G., M. Fuchs, M. J. McConville, Y. D. Stierhof, and P. Overath. 1994. "Surface Antigens of Leishmania Mexicana Amastigotes: Characterization of Glycoinositol Phospholipids and a Macrophage-Derived Glycosphingolipid." *Journal of Cell Science* 107 (Pt 9) (September): 2471–82.

World Health Organization. 2010. "Control of the Leishmaniases." *World Health Organ Tech Rep Ser*, no. 949.

Wright, Jessica R., T. Nicolai Siegel, and George A. M. Cross. 2010. "Histone H3 Trimethylated at Lysine 4 Is Enriched at Probable Transcription Start Sites in Trypanosoma Brucei." *Molecular and Biochemical Parasitology* 172 (2): 141–44. doi:10.1016/j.molbiopara.2010.03.013.

Wurst, Martin, Beate Seliger, Bhaskar Anand Jha, Cornelia Klein, Rafael Queiroz, and Christine Clayton. 2012. "Expression of the RNA Recognition Motif Protein RBP10 Promotes a Bloodstream-Form Transcript Pattern in Trypanosoma Brucei." *Molecular Microbiology* 83 (5): 1048–63. doi:10.1111/j.1365-2958.2012.07988.x.

Wu, Ying, Youssef El Fakhry, Denis Sereno, Samira Tamar, and Barbara Papadopoulou. 2000. "A New Developmentally Regulated Gene Family in Leishmania Amastigotes Encoding a Homolog of Amastin Surface Proteins." *Molecular and Biochemical Parasitology* 110 (2): 345–57. doi:10.1016/S0166-6851(00)00290-5.

Young, Matthew D, Matthew J Wakefield, Gordon K Smyth, and Alicia Oshlack. 2010. "Gene Ontology Analysis for RNA-Seq: Accounting for Selection Bias." *Genome Biology* 11 (2): R14. doi:10.1186/gb-2010-11-2-r14.

Zhang, Wen-Wei, Susana Mendez, Anirban Ghosh, Peter Myler, Al Ivens, Joachim Clos, David L. Sacks, and Greg Matlashewski. 2003. "Comparison of the A2 Gene Locus in Leishmania Donovani and Leishmania Major and Its Control over Cutaneous Infection." *The Journal of Biological Chemistry* 278 (37): 35508–15. doi:10.1074/jbc.M305030200.

Zhao, Shanrong, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. 2014. "Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells." *PLoS ONE* 9 (1): e78644. doi:10.1371/journal.pone.0078644.

Zhao, Zhixing, Megan E. Lindsay, Arnab Roy Chowdhury, Derrick R. Robinson, and Paul T. Englund. 2008. "p166, a Link between the Trypanosome Mitochondrial DNA and Flagellum, Mediates Genome Segregation." *The EMBO Journal* 27 (1): 143–54. doi:10.1038/sj.emboj.7601956.

Zilberstein, D., N. Blumenfeld, V. Liveanu, A. Gepstein, and C. L. Jaffe. 1991. "Growth at Acidic pH Induces an Amastigote Stage-Specific Protein in Leishmania Promastigotes." *Molecular and Biochemical Parasitology* 45 (1): 175–78.