# The *Phytomonas* cell surface proteome: Insights into the evolution of the parasite-host interface in Trypanosomatids




Thesis submitted for the degree of

Doctor of Philosophy

Eleanor Ceindeg Jaskowska

Keble College

University of Oxford

## *Abstract*

The trypanosomatids are a diverse group of obligate parasitic eukaryotes that have adapted to colonise multicellular and unicellular hosts that span the eukaryotic tree of life. The genetic resources that have enabled trypanosomatids to colonise such a diverse range of hosts are largely unknown. However, there is precedent that adaptation is in part facilitated through the evolution of families of cell surface proteins. The aim of this thesis was to broaden our knowledge of how trypanosomatids have adapted their cell surface to colonise different hosts through a combination of proteomic and bioinformatic approaches. The principal body of work is focussed on *Phytomonas*, a globally distributed genus of plant infective trypanosomatids that are transmitted by insects. As a sister group to *Leishmania* parasites of animals, this study of *Phytomonas* informs our understanding of how trypanosomatids have evolved from a monoxenous parasite of insects to adapt to both mammalian and plant hosts. To provide further resolution, a second species, *Trypanosoma carassii* (an extracellular parasite of cyprinid fish) was also analysed. *T. carassii* occupies an informative position in the phylogenetic tree of trypanosomatids as an outgroup to the salivarian and stercorarian *Trypanosoma* species. The results of the bioinformatic analysis presented in this thesis reveal that there is a conserved core of trypanosomatid cell surface proteins, the majority of which have no known function. Bioinformatic and proteomic results show that the cell surface of *T. carassii* is covered in one or more novel and unique families of GPI anchored proteins. The study of *P. serpens* indicates that while evolution to parasitise plants was facilitated though adaptation of the cell surface, there was no evidence for the innovation of large families of cell surface proteins as is observed in *Trypanosoma* and *Leishmania* species. Placed in an evolutionary context, the results presented in this thesis provide some insight into how these enigmatic parasites evolved from a single common ancestor to parasitise a broad host range from coconuts to crocodiles.

## *Acknowledgements*

I would like to thank Steve Kelly for his supervision and for not giving up on me. Thanks also to Jane Langdale, Gem Toes-Crichton, Roni McGowan, Nick Kruger, John Baker and everyone in the Department of Plant Sciences for their kind words and acts of encouragement. I would also like to recognise the support from past and present group members of the Kelly and Langdale groups - Emily for being my cheerleader, David for sharing his love of bikes, Olga for partnering me in nocturnal report writing, Tom for being surprisingly interesting, Jessie for her humour, Dana for taking me off the streets, Julie for running a tight ship, and everyone else for just putting up with me.

I would like to thank Keble College and the MCR for being much more than just a hall of residence. Thank you for the opportunities, biathlons and formal dinners. My time as MCR President gave me a sense of purpose when other things weren't going so well and the experiences were invaluable in preparing for life on the outside.

My time as a graduate student would not have been the same without all the brilliant and lasting friendships, usually kindled while wearing some sort of lycra. Thanks are not enough for OUWRFC, KCBC, Cowley Road Condors (KTFU) and the rest of the peloton. I probably would have finished sooner if I hadn't discovered cycling, or maybe I would have gone insane and never finished at all. For rebuilding my confidence after leaving Oxford I'd like to thank past colleagues from EPSRC. I am forever grateful for the offers of gin, chocolate and understanding from Pria, Steph, Ollie, Tash.

Finally, and most importantly, I would like to thank my family. My Mum for her unwavering support and for being such a brilliant role model, and Mamgu for teaching me that giving up is rarely an option.

This work was funded by a BBSRC postgraduate studentship.

# *Contents*

Abstract....ii

Acknowledgements....iii

Appendix legend .... 188

References .... 221

Figure Legend....ix

Table legend....xi

List of Abbreviations....xii

Chapter 1: Introduction.... 1

1.1 Preface .... 1

1.2 *Phytomonas*: endophytes and pathogens .... 2

1.3 Host range and phylogeny .... 4

1.4 Transmission and life cycle .... 9

1.5 *Phytomonas* interaction with the insect host.... 12

1.6 *Phytomonas* interaction with the plant host .... 13

1.7 *Phytomonas* cell morphology .... 15

1.8 *Phytomonas* metabolism.... 19

1.9 The *Phytomonas* cell surface.... 21

1.10 Discussion .... 24

1.11 Thesis aims.... 27

Chapter 2: Materials and Methods .... 29

2.1 Preface .... 29

2.2 General Materials ..... 29
2.3 Cells and growth conditions ..... 29
2.3.1 Phytomonas serpens ..... 29
2.3.2 Escherichia coli ..... 29
2.3.3 Trypanosoma carassii ..... 30
2.3.4 Phytomonas françai ..... 30
2.4 Bioinformatics ..... 30
2.4.1 Phytomonas serpens genome reassembly ..... 31
2.4.2 Prediction of protein coding genes ..... 32
2.4.3 Trypanosoma carassii transcriptome ..... 33
2.4.4 Phytomonas françai genome sequencing and assembly ..... 34
2.4.5 Inference of orthogroups ..... 35
2.4.6 Phylogenetic inference ..... 36
2.4.7 Prediction of putative membrane orthogroups ..... 37
2.4.8 Prediction of putative cell surface orthogroups for Trypanosoma carassii ..... 37
2.4.9 Prediction of cell surface orthogroups for Phytomonas ..... 38
2.5 Proteomic methods ..... 38
2.5.1 Live cell biotinylation of Phytomonas serpens ..... 38
2.5.2 Quenching the biotinylation of Phytomonas serpens ..... 40
2.5.3 Live cell biotinylation of Trypanosoma carassii ..... 40
2.5.4 Enrichment of biotinylated proteins ..... 41
2.5.5 Release of biotinylated proteins from streptavidin resin ..... 41
2.5.6 SDS-polyacrylamide gel electrophoresis of proteins ..... 43

2.5.7 Coomassie staining of polyacrylamide protein gels ..... 43
2.5.8 SYPRO Ruby staining of polyacrylamide gels ..... 43
2.5.9 Transfer of proteins from gels to nitrocellulose membrane (blotting) ..... 43
2.5.10 Enhanced Chemiluminescence (ECL) detection ..... 44
2.5.11 Tandem mass spectrometry ..... 44
2.5.12 Data searches ..... 46
2.5.13 Detection of biotin tag by epifluorescence microscopy ..... 47
2.6 Molecular cloning ..... 47
2.6.1 Polymerase Chain Reaction (PCR) ..... 47
2.6.2 Nested PCR tagging ..... 48
2.6.3 Synthesis of an endogenous signal peptide ..... 49
2.6.4 Long Primer PCR Tagging ..... 52
2.6.5 DNA Sequencing ..... 52
2.6.6 Bacterial Transformation ..... 52
2.6.7 Transfection of Phytomonas for nested PCR product ..... 53
2.6.8 Transfection of Phytomonas serpens for long primer PCR product ..... 53
2.6.9 Cloning of cell lines ..... 54
2.6.10 Validation of YFP expression and localisation by microscopy ..... 54
2.6.11 Validation of YFP expression by western blot ..... 55
2.6.12 DNA extraction from Phytomonas serpens ..... 55
2.7 Infection of tomato fruit ..... 55
2.8 Microscopy ..... 56

Chapter 3: Proteomic and bioinformatic investigation into the Phytomonas serpens cell surface ..... 57

3.1 Preface ..... 57

3.2 Introduction: methods for studying cell surface proteins ..... 58

3.3 Results ..... 63

3.4 Discussion ..... 105

3.5 Conclusions ..... 109

Chapter 4: Gene tagging in Phytomonas serpens, validating the localisation of putative cell surface proteins ..... 110

4.1 Preface ..... 110

4.2 Introduction ..... 110

4.3 Prioritisation of proteins for localisation ..... 115

4.4 Results ..... 123

4.3.1 The use of eYFP expressing P. serpens cells for monitoring infection in Solanum lycopersicun ..... 130

4.5 Discussion ..... 133

4.6 Conclusions ..... 139

Chapter 5: The Trypanosoma carassii transcriptome and predicted cell surface proteome 140

5.1 Preface ..... 140

5.2 Introduction and background ..... 141

5.3 Results ..... 149

5.3.1 Sequencing of the T. carassii transcriptome and clustering of trypanosomatid orthogroups. ..... 149

5.3.2 Prediction of orthogroups containing cell surface localised genes. ..... 151

5.3.3 The set of surface orthogroups found in all species ..... 158

5.3.5 T. carassii transcriptome contains two large families of putative surface proteins not found in any other trypanosomatid ..... 160

5.3.6 Cell surface biotinylation and isolation of the T. carassii surface proteome. . 162

5.3.7 Characterisation of the T. carassii cell surface proteome. ..... 166

5.4 Discussion ..... 170

5.5 Conclusions ..... 173

Chapter 6: Conclusions and General Discussion ..... 175

6.1 Overview ..... 175

6.2 *Phytomonas serpens* as a model for plant-trypanosomatid interactions ..... 177

6.3 The adaptation of an ancestral trypanosomatid to a suite of multicellular hosts .. 179

6.4 Future directions ..... 186

# *Figure Legend*

Figure 1.1. Summary of all known *Phytomonas* isolates ..... 7

Figure 1.2. Maximum likelihood phylogenetic tree of trypanosomatids ..... 8

Figure 1.3. The life cycle of *Phytomonas* ..... 11

Figure 1.4. *Phytomonas serpens* morphology ..... 18

Figure 2.1. A schematic diagram of the nested PCR tagging approach ..... 51

Figure 3.1. The workflow of the cell surface biotinylation protocol ..... 62

Figure 3.2. The order in which stages of the biotinylation protocol were optimised .... 64

Figure 3.3. Optimisation of cell surface labelling ..... 66

Figure 3.4. Cell surface biotinylation of *Phytomonas serpens* ..... 67

Figure 3.5. Optimisation of streptavidin resin ..... 70

Figure 3.6. Results of preliminary tandem mass spectrometry ..... 76

Figure 3.7. Optimisation of cell lysis and bead wash buffers ..... 79

Figure 3.8. Post translational modification with SS-biotin ..... 81

Figure 3.9. Results of tandem mass spectrometry ..... 86

Figure 3.10. A summary of the results of orthogroup clustering ..... 97

Figure 3.11. The distribution of the surface orthogroups of *Phytomonas, Trypanosoma* and *Leishmania* ..... 103

Figure 4.1. Topogenic determinants of single-spanning membrane proteins ..... 113

Figure 4.2. Proteins shortlisted for eYFP tagging using nested PCR ..... 119

Figure 4.3. Plasmid restriction maps and schematic of nested PCR ..... 120

Figure 4.4. Surface localisation of eYFP fusion of Pse.222.1080 ..... 127

Figure 4.5. Validation of clonal cell lines ..... 128

Figure 4.6. Immunofluorescence of *P. serpens* ..... 129

Figure 4.7. The infection of tomato fruit with *P.serpens* cells expressing eYFP ..... 132

Figure 4.8. A phylogeny of the amastin surface glycoproteins in trypanosomatids ... 138

Figure 5.1. The life cycle of *T. carassii* ..... 145

Figure 5.2. A phylogenetic tree of the trypanosomatids..... 146

Figure 5.3. The results of orthogroup clustering in *Trypanosoma*..... 150

Figure 5.4. The results of orthogroup clustering following prediction of surface orthogroups ..... 154

Figure 5.5. The species distributions of surface orthogroups in *Trypanosoma* ..... 155

Figure 5.6. *Trypanosoma carassii* specific surface orthogroups..... 161

Figure 5.7. Cell surface biotinylation of *Trypanosoma carassii* ..... 165

Figure 5.8. Results from cell surface biotinylation in *Trypanosoma carassii* ..... 169

Figure 6.1. A schematic summarising the main points about what we now know about the cell surface proteins of *P. serpens* and *T. carassii*. ..... 185

# *Table legend*

Table 2.1. *P. serpens* genome assembly ..... 32

Table 2.2. *P. serpens* predicted protein sequences ..... 33

Table 3.1. Results of preliminary tandem mass spectrometry ..... 77

Table 3.2. The composition of cell lysis/bead wash buffers ..... 78

Table 3.3. Results of LC-biotin tandem mass spectrometry ..... 87

Table 3.4. Results of SS-biotin tandem mass spectrometry ..... 89

Table 3.5. Proteins common to both LC and SS biotin datasets ..... 90

Table 3.6. The predicted cell surface orthogroups found in *Phytomonas* species that are absent from *Trypanosoma* and *Leishmania* species ..... 103

Table 4.1. Proteins shortlisted for tagging using eYFP fusion proteins using nested PCR ..... 118

Table 4.2. Proteins shortlisted for eYFP tagging using long primer PCR ..... 121

Table 4.3. Transfection attempts for nested PCR ..... 125

Table 4.4. Validation primers for nested PCR ..... 128

Table 5.1. A comparison of the present study with two other investigations into the trypanosome cell surface ..... 157

Table 5.2. A summary of LC MS/MS on T. carassii ..... 163

Table 5.3. Results of LC MS/MS of cell surface biotinylated T. carassii ..... 163

Table 5.4. Putative cell surface proteins in Table 5.3 with orthologs in other species annotated as hypothetical proteins ..... 164

# *List of Abbreviations*

ATP – adenosine triphosphate

BLAST – basic local alignment search tool

DMSO – dimethyl sulfoxide

DNA – deoxyribonucleic acid

ECL – Enhanced Chemiluminescence

eYFP – enhanced yellow fluorescent protein

GFP – green fluorescent protein

GIPL – glycoinositol phospholipids

GPI – glycophosphatidylinositol

HMM – hidden Markov model

LB – Luria-Bertani media

LC MS/MS – liquid chromatography and tandem mass spectrometry

LPG – lipophosphoglycan

mRNA – messenger RNA

MS – mass spectrometry

MW – molecular weight

NADH – nicotinamide adenine dinucleotide

NCBI – National Centre for Biotechnology Information

NHS – *N*-hydroxysuccinimide

OG – orthogroup

PAGE – polyacrylamide gel electrophoresis

PCR – polymerase chain reaction

PFA – paraformaldehyde

PTM – post-translational modification

RNA – ribonucleic acid

SDS – sodium dodecyl sulphate

SP – signal peptide

TMD – transmembrane domain

TRITC – Tetramethylrhodamine

Tris – Tris (hydroxymethyl) aminomethane

UV – ultraviolet

WT – wild type

# *Chapter 1: Introduction*

## 1.1 Preface

The trypanosomatids are a monophyletic group of single-celled eukaryotic parasites that are spread between multicellular hosts, predominantly by insects. Globally these parasites cause a considerable burden on human health and welfare, with an estimated 20 million people infected with trypanosomatid pathogens (Stuart *et al.*, 2008), as well as several devastating diseases of livestock. Thus, trypanosomatids have attracted significant attention from disparate academic communities for their impact on mammalian biology. In more recent years, trypanosomatids have also drawn interest for use as model organisms because of their streamlined and often extreme biology. In this context the study of trypanosomatid biology has greatly contributed to our understanding of several biological phenomena including mechanisms of immune evasion (MacGregor *et al.*, 2012), glycosylphosphatidylinositol (GPI) anchors (Ferguson, 1999), RNA editing (Benne *et al.*, 1986), polycistronic transcription (Siegel and Gunasekera, 2011), trans-splicing (Sutton and Boothroyd, 1986) and chromosome segregation (Akiyoshi and Gull, 2014). Though trypanosomatids have been the focus of many studies, one large and diverse sub-group of plant-infecting trypanosomatids known as *Phytomonas* are relatively poorly understood. Little is known of their biology, life cycle or how they have adapted to life inside plants.

This introduction chapter provides a broad overview of *Phytomonas* in terms of disease biology, morphology and metabolism in the light of emerging genome resources for this globally distributed group. The purpose of this introduction is to provide context for the results chapters that follow and to give a rationale for the experimental procedures used. This chapter (in a more concise form) was published as a review in PLoS pathogens (Jaskowska *et al.*, 2015).

## 1.2 *Phytomonas*: endophytes and pathogens

Over 100 years have passed since the first plant trypanosomatids were isolated from the latex of *Euphorbia pilulifera* on the island of Mauritius (Lafont, 1909). The discovery of these trypanosomatids, initially named *Leptomonas davidi,* was confirmed in *E. pilulifera* in Madras (4,400 kilometres away) later the same year (Donovan, 1909). It was recognised from the outset that the plant trypanosomatids were morphologically distinct from mammalian infective species and it was on this basis that Donovan suggested a new genus termed *Phytomonas* for their classification (Donovan, 1909). The first reports of *Phytomonas* infections of plants initially described the host plants as exhibiting poor growth and wilt (Lafont, 1909). Subsequent observations in *Euphorbia segetalis* and *Euphorbia peplus* agreed that infection with *Phytomonas* was harmful to the plant (Franca, 1920), and reported that infected plants exhibited a depletion of starch granules from the latex and surrounding parenchyma (cellular tissue bulk of soft parts), and a reduction in the viscosity of the latex (Bensaude, 1925). However, in contrast to this, other authors stated that infections in different Euphorbiaceae produced no detectable effects on plant growth or yield. Moreover, they observed that parasitised and non-parasitised plants were indistinguishable without examination of the latex by microscopy (Aragão, 1931; Bensaude, 1925). Thus the early descriptions of the genus disagreed over whether these parasites were pathogenic or endophytic (non-detrimental) in their plant hosts.

Since these initial investigations two species have been definitively shown to cause plant disease. The first species, *Phytomonas staheli*, is the aetiologic agent of '*hartrot*' (fatal wilt) of coconut palm (*Cocos nucifera)*, '*marchitez sorpresiva*' (sudden wilt) and slow wilt of oil palm (*Elaeis guineensis)* (Di Lucca *et al.*, 2013; Parthasarathy *et al.*, 1976). Both diseases are acute lethal wilts which begin in the leaves and progress to the rotting of the spear and root (Camargo, 1999), although this rotting may be the result of secondary infection by bacteria and other organisms (Parthasarathy *et al.*,

1976). The second pathogenic species, *Phytomonas leptovasorum* causes coffee phloem necrosis in both Liberica and Arabica coffee (Stahel, 1931). A common feature of both of these pathogenic *Phytomonas* species is that they exclusively inhabit the phloem during the plant stage of their life cycle (Parthasarathy *et al.*, 1976; Stahel, 1931). A third species, *Phytomonas françai*, inhabits the latex ducts of cassava (*Manihot esculenta*) and has been linked with yield loss diseases known as '*chochamento de raizes'* or 'empty root syndrome'. This disease is characterised by poor root development and chlorosis of the leaves (Kitajima, 1986; Vainstein and Roitman, 1986). Several attempts have been made to show experimentally that *P. françai* is the aetiological agent of empty root syndrome, however in experimental infections infected plants appeared identical to uninfected individuals (Aragão, 1927, 1931; Kitajima, 1986). Moreover, the Unha cultivar of cassava that may have exhibited pathology when infected with *Phytomonas* is no-longer widely farmed in Brazil and therefore there have been no reports of 'empty root syndrome' in cassava since 1980 (Camargo, 1999). Thus it appears that *P. françai* poses little or no risk to food security in this crop.

Though there are only two proven pathogenic species, those species pose a degree of economic risk in South America. *P. leptovasorum* (the aetiological agent of coffee phloem necrosis) has been isolated in Suriname (Teixeira *et al.*, 1994) and Brazil (Serrano *et al.*, 1999). With the rapid expansion of coffee plantations across South America and the resulting change in the distribution of crop pathogens and pests (Bebber *et al.*, 2013), coffee phloem necrosis poses a potential risk to these expanding coffee economies. For example, Brazil is the world's largest exporter of green coffee beans, trading more than 1,700,000 tonnes in 2011 at a value of $8bn (http://faostat.fao.org/). *P. staheli* (the aetiological agent of wilt diseases of oil and coconut palm) has previously been isolated from Ecuador, Suriname, Venezuela, Brazil, Costa Rica, Honduras and Colombia (Berrio and Ocho, 1991). In these regions

oil palm is a major commercial crop with 397,000 tonnes of palm oil exported by Colombia and Ecuador alone in 2011 at a value of $493m (http://faostat.fao.org/). Currently, there are no effective treatments for these *Phytomonas* plant diseases, and strategies for infection control comprise the felling of diseased plants or the removal of the infected plant material. Attempts to develop effective chemical controls have suffered from a lack of genome resources and have focused on exploiting targets that are conserved in related trypanosomatid species (Magán *et al.*, 2004). These include perturbation of shared metabolic pathways such as aspects of nucleic acid biosynthesis (Magán *et al.*, 2004).

## 1.3 Host range and phylogeny

*Phytomonas* as a genus are globally distributed (Figure 1.1A) and have been repeatedly isolated from disparate plant hosts that span the angiosperm tree-of-life (Figure 1.1B). Appendix 1.1 is a list of all published *Phytomonas* isolations known at the time of writing and is summarised in Figure 1.1. Within their plant hosts *Phytomonas* species have been isolated from a variety of host tissues including phloem, latex ducts, fruit, flowers and seeds (Figure 1.1C & Appendix 1.1) and thus have evolved to inhabit both extracellular and intracellular plant environments (Camargo, 1999).

The first attempt to rationalise the phylogeny and sub-groupings of *Phytomonas* used isoenzyme analysis (Vainstein *et al.*, 1987). This was extended using molecular markers including the internal transcribed spacer of the ribosomal RNA locus, kinetoplast DNA sequence and the spliced leader RNA gene array (Dollet *et al.*, 2001; Dollet *et al.*, 2012; Sturm *et al.*, 2007; Teixeira *et al.*, 1997). Collectively, these studies revealed that *Phytomonas* are monophyletic and that the pathogenic phloem-limited *Phytomonas* species formed a discrete group termed the Phloemicola that are distinct from non-pathogenic trypanosomatid parasites of fruit and latex (Dollet *et al.*, 2001). Outside of this association, molecular data revealed that neither habitat nor host

species was predictive of phylogenetic sub-groupings within *Phytomonas* (Dollet *et al.*, 2012). Moreover, it was found that the same plant species could harbour distantly related *Phytomonas* species (Dollet *et al.*, 2012). For example the genome strain, *Phytomonas serpens* 9T isolated from tomato fruit (*Solanum esculentum*) in Brazil belongs to a different group to *Phytomonas* isolated from tomatoes from Southern Spain (Dollet *et al.*, 2012). Similarly, the *Phytomonas* species isolated from plant latex in Europe, Africa and India were serologically and molecularly distinct from parasites isolated from latex from South America (Dollet *et al.*, 2012). However, the low quantities of data available precluded the complete resolution of the evolutionary relationship between species.

The first *Phytomonas* genomes to be published were those of *P. serpens* 9T (Koreny *et al.*, 2012), *Phytomonas* sp. *HART1* (Porcel *et al.*, 2014) and *Phytomonas sp. EM1* (Porcel *et al.*, 2014). *P. serpens* 9T was isolated from a tomato fruit in South America (Santos *et al.*, 2006), *Phytomonas sp. HART1* was isolated from the phloem of a diseased coconut palm in Guiana (Dollet *et al.*, 2012; Porcel *et al.*, 2014), and *Phytomonas sp. EM1* was isolated from the latex of an asymptomatic *Euphorbia* in the south of France (Dollet *et al.*, 2012; Porcel *et al.*, 2014). Figure 1.2 is a phylogenetic tree constructed from 959 single copy nuclear genes, adapted from Kelly *et al.*, (2014), showing the evolutionary relationship of the sequenced *Phytomonas* species to several pathogenic and non-pathogenic taxa from *Leishmania* and *Trypanosoma*. Interestingly, these *Phytomonas* genomes are much smaller than those of their *Leishmania* or *Trypanosoma* relatives, for example, the genome of EM1 is 17.8 mega bases (Mb) (Porcel *et al.*, 2014) in comparison to 32.9 Mb of *Leishmania major* (Ivens *et al.*, 2005), 62.3 Mb of *Trypanosoma brucei* (Berriman *et al.*, 2005) and 32.5 Mb of *Trypanosoma cruzi* (El-Sayed *et al.*, 2005).

A unifying feature of phylogenetic analyses of *Phytomonas* is that there appears to be large diversity within the group. For example, the kDNA sequence data shows that the divergence between the phloem limited *Phytomonas sp. HART1* and the fruit parasite

*P. serpens* is greater than that spanning all sampled *Leishmania* species and similar to the divergence between *T. cruzi* and *T. brucei* (Porcel *et al.*, 2014). The mitochondrial DNA of *Phytomonas* is contained within the kinetoplast, the subcellular oraganelle after which the order is named (the kinetoplastids). The kinetoplast is a unique organelle which exhibits radically different modes of DNA organisation (mini and maxi circles), and expression (RNA editing), and gene fragmentation (RNA editing) compared to other eukaryotes (Benne *et al.*, 1986; Simpson, 1987). While RNA editing is relatively rare in eukaryotes it is not unique to the trypanosomatid mitochondria. The presence of a kinetoplast was a diagnostic character used by early taxonomists to group the kinetoplastids together. Subsequently, analyses of the DNA sequence of the kinetoplast (first by restriction digest and later by gene sequencing) have provided insights into the interrelationships between species in the group (Riou *et al.*, 1987; Sturm *et al.*, 2007).

This finding from kDNA that genetic diversity between *Phytomonas* species is greater than that observed for *Leishmania* species is also supported by studies of other molecular markers such as small sub-unit rRNA and glycosomal GAPDH (Maslov *et al.*, 2013; Votýpka *et al.*, 2012), as well as the concatenated protein sequence phylogeny presented in Figure 1.2.

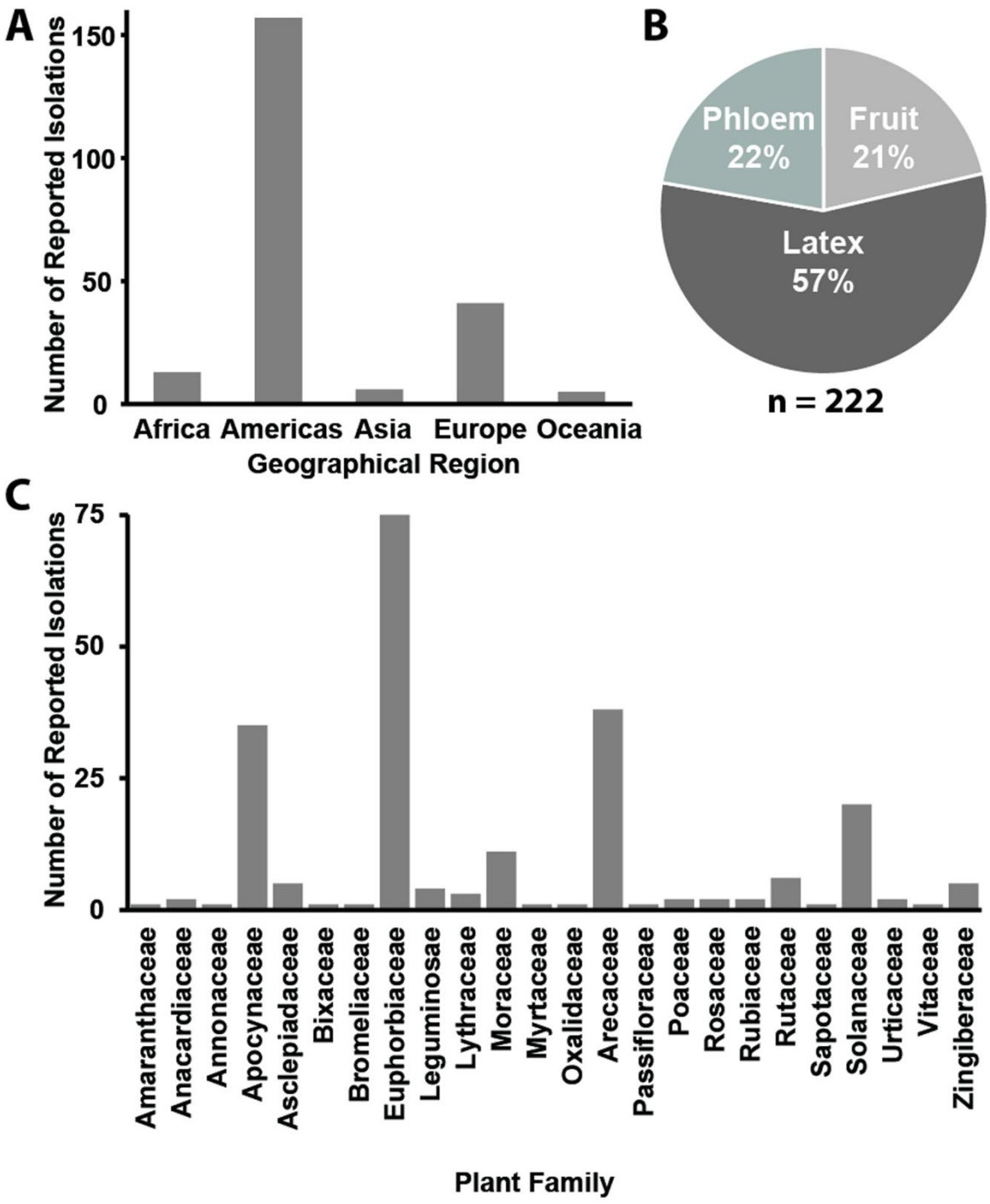



***Figure 1.1*** **Summary of all known *Phytomonas* isolates. A)** A bar chart depicting the number of reported isolations by continent. **B)** A pie chart depicting the plant host environment from which these isolations were made. For further details see Appendix 1.1. Data from Appendix 1.1 taken which was adapted and updated from Camargo, 1999. **C)** A bar chart depicting the plant families from which these isolations were made

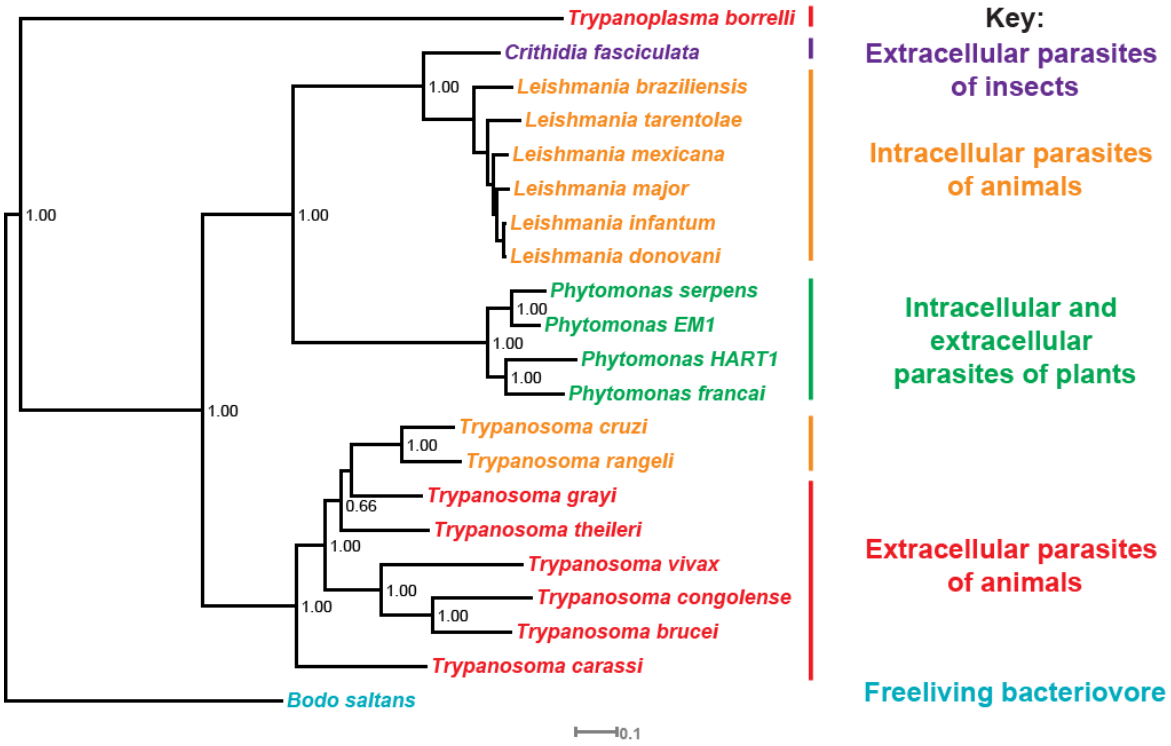



***Figure 1.2*** **A maximum likelihood phylogenetic tree of trypanosomatids.**

The phylogeny was constructed by Dr Kelly using orthologous gene groups containing only single copy genes aligned with MergeAlign (Collingridge and Kelly, 2012) and phylogenetic tree inferred using RAxML (Stamatakis, 2006) as described in Methods section 2.4.6. and the tree was rooted on *Bodo saltans*.

## 1.4 Transmission and life cycle

Though trypanosomatids that parasitise fish (Overath *et al.*, 1999), frogs (Siddall and Desser, 1992) turtles (Siddall and Desser, 1992) and possibly several Australian vertebrates (Hamilton *et al.*, 2005; Hamilton *et al.*, 2007) are transmitted by leeches, the overwhelming majority of described trypanosomatid species are spread between hosts by insects (Hamilton *et al.*, 2005; Hamilton *et al.*, 2007). Very shortly after the initial discovery of *Phytomonas*, it was shown in caged insect experiments that the phytophagous insect *Nysius euphorbiae,* a hemipteran which feeds on multiple species of Euphorbia, was able to transmit the parasites from infected plants to uninfected hosts (Lafont, 1911). Further work by Bouet and Rouband (1911) reviewed in Bensaude, (1925) demonstrated that the latex feeding insect *Dieuches humilis* could also act as the insect host, however the natural host was shown to be the nocturnal coreid spurgebug *Dicranocephalus agilis* (Franca, 1920). Taken together these early investigations suggested that individual *Phytomonas* species may be spread between plant hosts by a broad range of different insects. Consistent with these early reports, all subsequent evidence has further confirmed that *Phytomonas* are transmitted by phytophagous insects (Alves e Silva *et al.*, 2013; Camargo *et al.*, 1990; Gibbs, 1957). Though the relationship between plant host, parasite and insect host appears simple (Appendix 1.1), in reality it is unknown to what extent different *Phytomonas* species can be spread by different insects or can colonise different plants. This relationship between plant host, parasite and insect host is further complicated by the nomenclature of the parasites. For example, the name *P. serpens,* has been applied to *Phytomonas* parasites that have been isolated from the fruit of multiple different tomato cultivars (Appendix 1.1) in the Americas, Africa and Europe (Dollet *et al.*, 2012; Gibbs, 1957). Moreover, it has also been applied to isolates from multiple different insects, for example from the salivary glands of *Nezara viridula* in South Africa (Gibbs, 1957) and from *Phthia picta* in Brazil (Brazil *et al.*, 1990). These multiple disparate isolations may represent different species and thus there may be some inflation of the true host and insect host range for an individual species.

In insect-spread mammalian infections of *Leishmania* and *Trypanosoma,* parasite cells undergo differentiation to metacyclic forms that are capable of infecting the mammalian host. In *Trypanosoma brucei* these infective metacyclic cells express a specialised variant surface glycoprotein on their cell surface (Barry, 1998) while in *Leishmania major* the molecular marker for mammalian infective metacyclic cells is a specialised lipophosphoglycan (LPG) (Späth and Beverley, 2001; Späth *et al.*, 2000). There is no evidence for metacyclic variant surface glycoproteins or metacyclic LPG in *Phytomonas.* Moreover, the metacyclic specific LPG found in *L. major* is not found in other *Leishmania* species and there are no other known molecular markers of metacyclogenesis. Though there are descriptions of metacyclic *Phytomonas* cells in the literature (Gibbs, 1957), these descriptions are based solely on cell morphology and it is unknown whether they represent a true specialised plant infective stage of the life cycle. In this context it is noteworthy that there have been successful infections of tomato fruit with insect isolated *P. serpens* cells grown in culture and from allowing infected insects to feed on fruit (Gibbs, 1957; Jankevicius *et al.*, 1989). This is in contrast to *Leishmania* and *Trypanosoma* where cultured insect adapted cells are not infective to mammalian hosts and metacyclogenesis must be induced prior to infection (Zakai *et al.*, 1998). Though a pre-adapted metacyclic stage may not be present in the *Phytomonas* life cycle*, P. serpens* cells have been observed to attach to insect salivary glands (Alves e Silva *et al.*, 2013) indicative of a true developmental stage inside the insect host. However, morphologically distinguishable differentiated cell forms have yet to be defined. In addition, as Jankevicius and colleages (1989) reported, parasites are readily transmitted from tomatoes to insects and *vice versa*. These are key differences between the different host stages observed for *Phytomonas* compared with those observed from *Leishmania* and *Trypanosoma* species. The main events of the *P. serpens* life cycle are summarised in Figure 1.3.

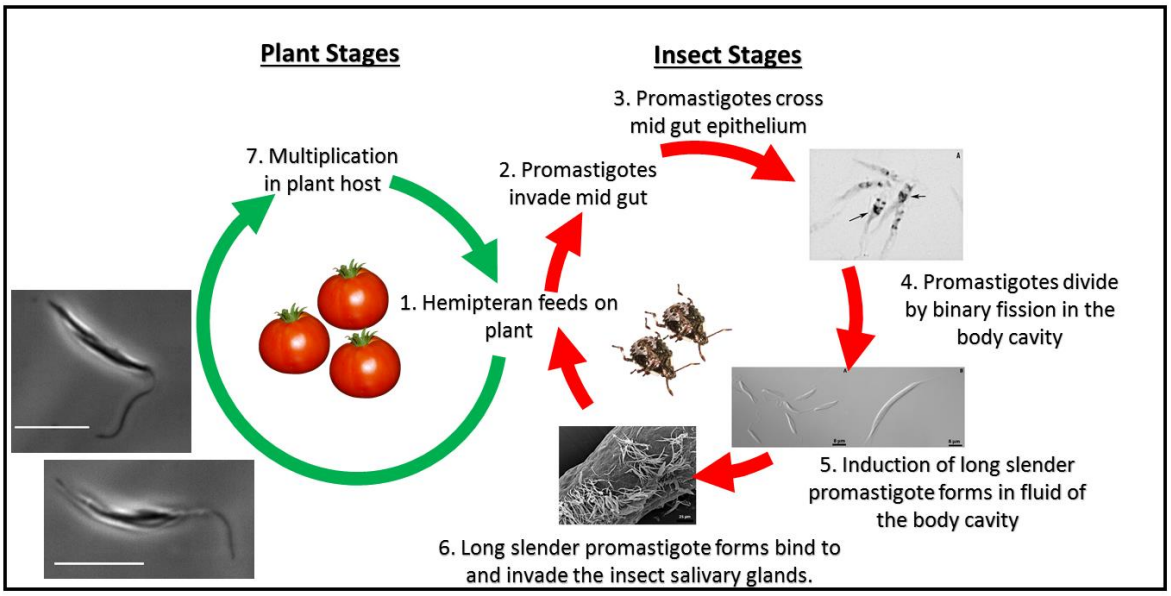



***Figure 1.3*** **The life cycle of *Phytomonas***

The main events of the *Phytomonas* life cycle are summarised here based on what is known of the life cycle of *P. serpens.* **(1)** *Phytomonas* are taken up in the insect meal. **(2)** Promastigotes invade the mid gut of *Oncopeltus fasciatus*. **(3)** Once the infection of the mid gut is established the *P. serpens* cells cross the epithelium and colonise the body cavity or hemocoel. **(4)** In the hemocoel the promastigotes have been observed to divide by binary fission. **(5)** Long slender promastigotes have been observed in the hemocoel (Alves e Silva *et al.*, 2013). **(6)** Parasites migrate to the salivary glands where they attach and invade the cells of the salivary glands, the parasites have not been observed to divide here**.** From the salivary glands the parasites are transmitted to the plant host when the insect feeds. **(7)** Parasites replicate in the plant host**.** Image of insects and tomatoes are not my own and are distributed under the creative commons license. The insect image is by Thomas Quine and the tomato image is by Fastily (Talk). Images of *P. serpens* insect forms are from Alves e Silva *et al.*, 2013. Images of *P. serpens* grown in culture are produced for this figure as described in 2.3.1, fixed and imaged as described in 2.8. Scale bar = 10µm.

## 1.5 *Phytomonas* interaction with the insect host

As described above, the extent of the true insect host range of a single *Phytomonas* species is unknown. It has been shown that multiple species of insect host can harbour the same species of *Phytomonas.* While *Phthia picta* is thought to be the most common host of *P. serpens* it has also been isolated from *Nezara viridula.* Furthermore, in experimental studies *P. serpens* has been shown to infect the insect host *Oncopeltus fasciatus*, which is the natural insect host for *P. elmassiani* (Alves e Silva *et al.*, 2013; Freymuller *et al.*, 1990; Gibbs, 1957). Though natural infections of *O. fasciatus* with *P. serpens* have not been observed, our understanding of the insect immune response to *P. serpens* is based on experimental infections of this species. The insect immune response includes humoral and cellular components. The former includes the production of antimicrobial peptides and reactive oxygen species. The cellular immune response involves specialised immune cells called hemocytes (Lemaitre and Hoffmann, 2007). The insect hemocyte response involves phagocytosis and nodulation (the binding of many hemocytes to aggregate pathogens). Phagocytosis and nodulation was observed in the hemocoel of *O. fasciatus* infected with *P. serpens* and has also been observed in other insects in response to trypanosomatid infection e.g. in *Rhodnius prolixus* infected with *Trypanosoma rangeli* (Alves e Silva *et al.*, 2013; M.A de Oliveira and de Souza, 2003; Takle, 1988).

However, the insect immune response in *O. fasciatus* does not prevent *P. serpens* from completing this stage of the life cycle (Alves e Silva *et al.*, 2013). Once taken up in the insect meal *P. serpens* promastigotes invade the *O. fasciatus* mid gut and from here were observed to cross the epithelium and colonise the body cavity or hemocoel. It is here that the parasites are exposed to the cellular host immune response. In the hemocoel the promastigotes were observed to divide by binary fission and produce long slender forms (Alves e Silva *et al.*, 2013). These *P. serpens* bear similarity with long slender forms of *T. rangeli* that have been observed in the hemolymph of *Rhodnius prolixus* and these forms were unable to establish an infection in mice following experimental infection (Hecker *et al.*,

1990, Alves e Silva *et al.*, 2013). It is unknown whether the analogous Phytomonas cells can establish an infection in tomato fruit. After colonising the hemocoel the Phytomonas parasites migrate to the salivary glands where they attached and invade cells of the salivary gland epithelium, no cell division of the parasites has been observed at this stage (Alves e Silva *et al.*, 2013). From the salivary glands the parasites can be transmitted to the plant host when the insect feeds.

Though distinct differentiated cell types have yet to be defined and may not exist, there are other notable differences *Phytomonas* and mammalian infective *Trypanosoma* and *Leishmania* in terms of transmission between hosts. These include the lack of a change in temperature between the different host stages and the absence of a change in parasite metabolism (see section 1.8 *Phytomonas* metabolism).

## 1.6 *Phytomonas* interaction with the plant host

The connection between *Phytomonas* infection and plant pathology is poorly understood. In the cases where infection of a plant has been associated with either a natural or experimental infection with a *Phytomonas* species the mechanisms responsible for causing the disease phenotype are unknown. It is unknown whether disease symptoms in host plants arise as a result of *Phytomonas* mediated degradation of plant tissues, the plant immune response, or as a by-product of depletion of plant assimilates. Those pathologies that have been observed are often coincident with secondary infection by bacteria and other organisms (Parthasarathy *et al.*, 1976). This lack of overt pathology is similar to that observed for mammalian infective trypanosomes like *Trypanosoma theileri,* that cause pathology only when secondary infection with other microorganisms has occurred (Villa *et al.*, 2008). Finally, although *Phytomonas* have not been observed to attach to the plant host the difficulties in conducting exhaustive experiments to test this mean that attachment cannot be ruled out.

It would be valuable to test the hypothesis that host range is linked to pathology. For example, by gathering improved data, either by sampling natural infections or through

experimental investigation, of the host ranges of those *Phytomonas* species associated with pathologies of plants, such as *P. HART1,* and to compare the host range observed to those species understood to be endophytic in plant hosts e.g. *P. serpens.*

The extent to which *Phytomonas* interact with and elicit the plant immune response is unknown. The plant immune response is triggered when the host recognises molecules that are common to a broad range of microbes. This recognition is achieved by extracellular leucine rich repeat (LRR) proteins in the plant. These proteins detect pathogen associated molecular patterns or microbe associated molecular patterns such as bacterial flagellin. The ability for plants and animals to detect this conserved microbial pattern evolved independently in each lineage, FLS2 in plants and TLR5 in mammals recognise different bacterial flagellin domains (Gómez-Gómez and Boller, 2000; Hayashi *et al.*, 2001). Recognition of these molecular patterns triggers a signalling cascade to modulate the transcription of downstream defence response genes (Asai *et al.*, 2002; Chang and Karin, 2001). In the 'zig-zag' model of plant-pathogen co-evolution virulent pathogens then secrete effectors to interfere with this immunity (Breitenbach *et al.*, 2014; Jones and Dangl, 2006). It is unknown whether some plants contain LRR proteins that are capable of recognising molecules secreted from or on the surface of *Phytomonas* parasites.

One particularly poorly understood aspect of *Phytomonas* biology is whether they secrete protein effectors to modulate the host plant immune response in this way. A genome-wide analysis of the predicted *Phytomonas EM1* and *P. HART1* secretomes was undertaken in the hope that it would lead to the identification of potential pathogenicity effectors in *Phytomonas* parasites. This resulted in the identification of several aspartyl proteases in both *Phytomonas* strains that were absent from the genomes of *Leishmania* and *Trypanosoma* (Porcel *et al.*, 2014). These proteases may play a role in host-parasite interactions as phloem sap often contains aspartyl protease inhibitors and aspartyl proteases are known to be secreted by plant pathogenic fungi such as *Botrytis cinerea* (ten Have *et al.*, 2010) and have been implicated in the suppression of systemic acquired resistance (Breitenbach *et al.*, 2014). However, the role of these inhibitors has so far mainly been

investigated in the context of phloem-feeding herbivores (Kehr, 2006) and there is some doubt whether these proteases are involved in plant-fungal interactions as protease deficient strains of *B. cinerea* are not attenuated in virulence (ten Have *et al.*, 2010).

Furthermore, the different plant microenvironments colonised present different challenges to the *Phytomonas* parasite. Plant latex has no known primary metabolic function and is thought to provide a defence function in the protection against herbivorous insects. Secondary metabolites found in latex include rubber, alkaloids, phenolics, proteins such as proteases and lectins, starch and other compounds such as hydrogen cyanide (Agrawal and Konno, 2009). As a result, latex is a hostile environment for a parasite such as *Phytomonas*. For example, the latex of *Euphorbia characias* contains a galactose-specific lectin (Barbieri *et al.*, 1983) that may agglutinate any parasite expressing galactose on its cell surface, as is found in some species of *Phytomonas* (Redman *et al.*, 1995). This is further discussed in section 1.9 in relation to the *Phytomonas* cell surface. The primary function of the phloem is in the transport of organic nutrients including sucrose and amino acids such as serine, glutamine and threonine. The phloem can also carry signalling molecules to activate systemic resistance in the plant (Carella *et al.*, 2016). The phloem is therefore a rich source of plant assimilates and pathogens such as bacterial phytoplasmas specifically colonise this environment (Hogenhout *et al.*, 2008). In some plants such as the cucurbits, the phloem can also contain defence compounds, including lectins (Gaupels and Ghirardo, 2013). Furthermore, tomato fruit also contain some defence compounds and it has been demonstrated that two of these (tomatine and tomatidine) can inhibit the growth of *P. serpens* in culture (Medina *et al.*, 2015).

## 1.7 *Phytomonas* cell morphology

In *Trypanosoma* and *Leishmania* morphological distinct, differentiated cell types are thought to be driven by adaptation to different host environments, for example; epimastigote forms for attachment to insect epithelia and amastigote forms inside macrophages (Wheeler *et al.*, 2013). *Phytomonas* does not appear to exhibit the same association between host, life cycle

stage and morphology as Trypanosoma and Leishmania species. They appear to have a simpler morphology than their mammalian infective sister species and exhibit promastigote morphology throughout the life cycle. The lack of requirement for morphologically discrete life cycle stages may be a contributory factor to the reduced gene content of the genome, however this has not been tested.

Consistent with their sister grouping to *Leishmania*, the majority of *Phytomonas* species isolated to date exhibit promastigote morphology (Camargo, 1999; Hoare and Wallace, 1966). Promastigote cells are characterised by a long cell body with a detached flagellum that protrudes from a flagellar pocket at the cell anterior (Figure 1.4). This morphology is similar to *Leishmania* parasites in the insect host and many monoxenous trypanosomatid insect parasites. In *Leishmania* and *Trypanosoma,* cell morphologies are replicated with high fidelity through the cell-division cycle, and altered with high fidelity when transitioning between different developmental stages depending on the host and organ or type of tissue colonised (Hoare and Wallace, 1966; Wheeler *et al.*, 2013). From the earliest descriptions of *Phytomonas* it was noted these parasites exhibited extreme morphological polymorphism (Franca, 1920; Gibbs, 1957). Polymorphism in cell body width, cell body length and flagellum length is typical of trypanosomatids isolated from their hosts, common to morphologies with a free (liberform) and attached (juxtaform) flagellum (Wheeler *et al.*, 2013). Whether the observed variation is an adaptation to changes in physiological conditions or uncharacterised life cycle stages is unknown. However, it has been hypothesised (Wheeler *et al.*, 2013), based on division of the *Leishmania mexicana* promastigote (Wheeler *et al.*, 2011), that one contributing factor to polymorphism of trypanosomatids with a detached flagellum is the pattern of morphogenesis through the cell cycle. In *L. mexicana*, the cell body increases in length during G1 and decreases post-S phase. Flagellar length is highly variable, in particular at cell division where the old flagellum is significantly longer than the new flagellum (Wheeler *et al.*, 2011). Thus it is possible that this morphological diversity during the cell division cycle may account for all of the variation in cell form that has been observed in isolates of *Phytomonas*.

The earliest drawings of parasites isolated from *Euphorbia* latex clearly show cells with a small round cell body and no external flagellum (Aragão, 1927; Franchini, 1922e, Franchini, 1922f; Lafont, 1909, Lafont, 1910) reminiscent of amastigote stages of *Leishmania* parasites (Lafont, 1911). Amastigote-like cells have also been repeatedly extracted from fruit (Camargo, 1999). However, the biological similarity of these *Phytomonas* amastigote forms to amastigote forms of *Leishmania* parasites and the mechanism(s) regulating differentiation remain completely unknown. Moreover, given the morphological plasticity of promastigote forms these may just be newly divided cells which are smaller, rounder and have short or no external flagellum.

A consistent observation from the study of *Phytomonas* species when attached to the insect salivary glands is that they have a long slender morphology with cells greater than 60 µm in length. This has been observed both for *P. serpens* in the model infective system *Oncopeltus fasciatus* and in hosts *Phthia picta* and *Nezara viridula* (Alves e Silva *et al.*, 2013; Freymuller *et al.*, 1990; Gibbs, 1957) and for *P. elmassiani* in insect host *O. fasciatus* (McGhee and Hanson, 1964). It was observed that cells of *P. elmassiani* in the insect salivary glands underwent asymmetric cell division (McGhee and Hanson, 1964). There is no published investigation into the possible function of these long slender *Phytomonas* cell forms however similar long slender forms of *T. rangeli* have been observed in the hemolymph of *Rhodnius prolixus* and these forms were unable to establish an infection in mice following experimental infection (Hecker *et al.*, 1990). Thus the long slender forms in Phytomonas may not be competent to establish an infection in plants,

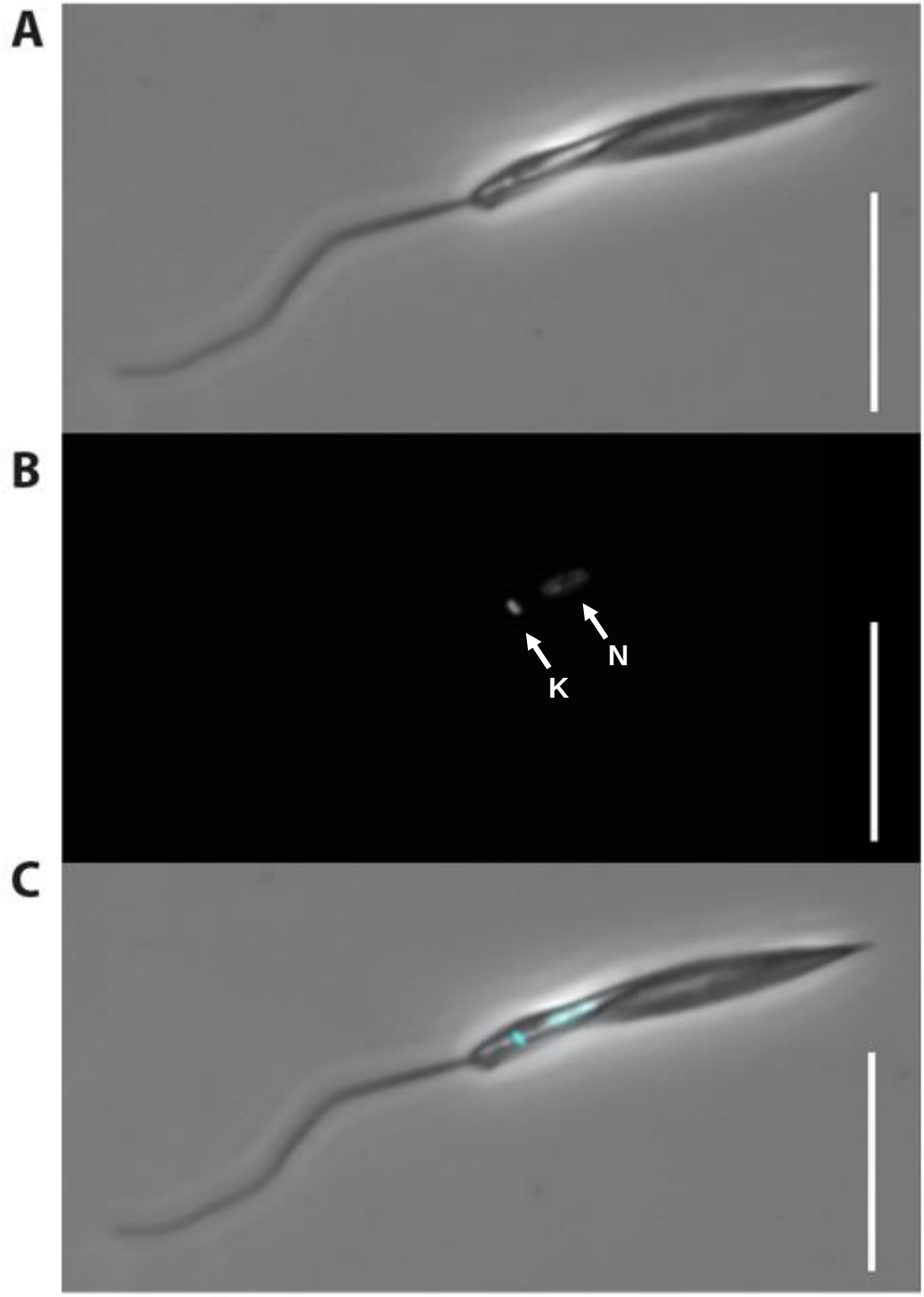



***Figure 1.4 Phytomonas serpens* morphology.** Epifluorescence and phase microscopy to show cell morphology of *P. serpens* cells grown in culture. DNA was stained using Hoechst (blue). **A)** Phase contrast image of a *P. serpens* cell. **B)** Mitochondrial DNA in the Kinetoplast (K) and nuclear DNA (N) **C)** Merge of all channels. Scale bar 10 µm.

## 1.8 *Phytomonas* metabolism

Analysis of the genes encoding metabolic proteins in the completed genome sequences revealed a cohort of enzymes consistent with life in a plant environment. Both *P. HART1* and *P. EM1* genomes encode glucoamylase, alpha-glucosidase and alpha,alpha-trehalose phosphorylase genes (acquired through horizontal gene transfer) allowing them to utilise plant carbohydrates (Porcel *et al.*, 2014). Only the phloem restricted pathogen, *P. HART1*, encodes invertase genes for degradation of sucrose (Porcel *et al.*, 2014). This finding is consistent with the diverse metabolic profiles of fruit (Carrari *et al.*, 2006) and latex (Andreia P Oliveira *et al.*, 2010) compared to the overwhelming abundance of sucrose in the phloem (Hammond and White, 2008). In addition, the *Phytomonas* genomes contain few tandemly duplicated genes in comparison to *Trypanosoma* and *Leishmania* species where the genes encoding metabolic enzymes are frequently found as tandem duplicates (Porcel *et al.*, 2014).

Another remarkable feature of *Phytomonas* parasites is the loss of genes encoding the cytochrome *c* oxidase subunits I-III (COI, COII, COIII), and cytochrome b (CyB) of the bc1 complex (Maslov *et al.*, 1999; Nawathean and Maslov, 2000; Porcel *et al.*, 2014). This loss appears to have occurred at the base of the *Phytomonas* clade (Koreny *et al.*, 2012). One aspect of *Phytomonas* biology which may have facilitated the loss of the cytochromes is the feeding behaviour of the insect host. Phytophagous insects feed exclusively on carbohydrate rich plant sap and therefore there may be no requirement for a metabolic shift from a carbohydrate to an amino acid energy substrate (Porcel *et al.*, 2014). In the case of the former, ATP production can occur at the level of substrate phosphorylation (Bringaud *et al.*, 2006) but in the case of the latter the mitochondrial respiratory chain is required for the complete oxidation of the substrate (Berriman *et al.*, 2005). The absence of a requirement to use amino acids as the main source of energy for *Phytomonas* cells in the insect host may have contributed to the loss of the mitochondrial cytochromes and the sole production of ATP via glycolysis in glycosomes (Chaumont *et al.*, 1994; Maslov *et al.*, 1999; Nawathean

and Maslov, 2000; Porcel *et al.*, 2014; Sanchez-Moreno *et al.*, 1992). *Phytomonas serpens* has been observed to produce ethanol as an end product of glucose catabolism providing an alternate pathway for the reoxidation of NADH produced in glycolysis (Ienne *et al.*, 2012). The key enzyme for fermentative metabolism- pyruvate/indolepyruvate decarboxylase was acquired by horizontal gene transfer in the common ancestor of *Phytomonas* and *Leishmania* (Ienne *et al.*, 2012).

An essential co-factor for the mitochondrial cytochromes is heme, an iron molecule contained by a porphyrin ring which functions in redox reactions and electron transport (Frankenberg *et al.*, 2003). Interrogation of the available *Phytomonas* genomes and the culture of *P. serpens* in a heme-free medium suggests that they are the only aerobic eukaryotes with the ability to survive in the absence of this molecule (Koreny *et al.*, 2012). However, the relationship between these plant trypanosomatids and heme is not straightforward as *P. serpens* cells grown in the presence of heme did contain small amounts of the compound, indicating that they are able to scavenge heme from the environment (Koreny *et al.*, 2012). To further complicate this curious aspect of *Phytomoans* biology the coding sequence for the enzyme ferrocheltase (involved in heme synthesis) is present in *P. serpens* and sequences encoding two heme-containing proteins are found in all three available *Phytomonas* genomes (Koreny *et al.*, 2012).

In addition to the alpha,alpha-trehalose phosphorylase mentioned above, there is also at least one other instance of horizontal gene transfer that conferred novel metabolic capabilities to *Phytomonas*. The transferred gene encodes a zinc-dependant alcohol dehydrogenase (Molinas *et al.*, 2003; Uttaro and Opperdoes, 1997b), which in combination with malate dehydrogenase enables *Phytomonas* species to produce lactate (Porcel *et al.*, 2014; Uttaro and Opperdoes, 1997a). It is unknown whether the production of lactate may produce a selective advantage for *Phytomonas* within the plant or insect host. Together, these changes in the complement of metabolic genes strongly indicate a cell adapted to life in a carbohydrate-rich environment.

## 1.9 The *Phytomonas* cell surface

The cell surface of *Phytomonas*, like all trypanosomatids, consists of at least three discrete domains. The domain with the largest surface area is the cell body surface; there is also the flagellar surface and next to the flagellum a distinct portion of the cell body membrane that is invaginated to form the flagellar pocket. In all trypanosomatids, these cellular surfaces lack an outer cell wall and instead are coated with membrane anchored proteins and glycoinositol phospholipids (Redman *et al.*, 1995). This exterior cell surface acts to protect the parasite cell from host responses and from changing environmental conditions.

To date, few molecular components of the *Phytomonas* cell surface have been characterised. Those that have been so far identified are those that are conserved in other trypanosomatids. One major component is the zinc-metalloprotease Major Surface Protease (MSP), also known as gp63 (D'Avila-Levy *et al.*, 2006). This protein is likely to be pan-eukaryotic (Manna *et al.*, 2013) and is found ubiquitously among the trypanosomatids in both a secreted form and a form anchored to the plasma membrane by glycosylphosphatidylinositol (GPI) (Jaffe and Dwyer, 2003; Santos *et al.*, 2007). The role of gp63 in the plant host has yet to be interrogated, however there is some evidence to suggest that gp63 is important in the insect stage of the *Phytomonas* life cycle (D'Avila-Levy *et al.*, 2006). Treatment of *P. serpens* with antibodies raised to gp63 perturbs parasite morphology and results in loss of adhesion to the salivary glands of the insect host, a step necessary for the establishment of infection in the insect host (Elias *et al.*, 2008; Santos *et al.*, 2006). Along with gp63, cysteine-proteases related to cruzipain of *T. cruz*i have also been found in *P. serpens* (Santos *et al.*, 2006). Both of these proteins have been shown to be secreted by *P. serpens* as well as existing in a form anchored to the cell surface by GPI (Elias *et al.*, 2008). Consistent with this observation, interrogation of the *Phytomonas* genomes revealed the presence of the pathways for GPI anchor biosynthesis (Porcel *et al.*, 2014).

Analysis of the genome sequences also revealed that *P. HART1* contains an expansion in the gp63 gene family with over 20 members in contrast to 2 genes in *P. EM1* (Porcel *et al.*,

2014). This expansion suggests that *P. HART1* may have a more complex cell surface that might facilitate survival in the hostile environment of the plant phloem (Beneteau *et al.*, 2010; Gaupels and Ghirardo, 2013). This increased diversity may also be further enhanced through post-translational modification as the genes required for post-translational addition of glucose, mannose, galactose, N-acetylglucosamine, glucuronic acid, xylose and fucose are also present in one or both *Phytomonas* genomes (Porcel *et al.*, 2014). While the function of these surface proteins in either the plant or insect host have yet to be established, it has been shown that the expression of these proteins in *Phytomonas* is responsive to culture conditions (Elias *et al.*, 2012). For example, incubation in media rich in complex proteins (such as serum) causes expression of gp63 to decrease (Elias *et al.*, 2012). Expression also correlates with growth rate and therefore it may play a role in allowing the *Phytomonas* parasites to adapt to different environments (Elias *et al.*, 2008).

In addition to proteins, trypanosomatids express a rich and varied array of glycolipids on their cell surface. *Leishmania* cells, for example, are coated in a layer of lipophosphoglycan (LPG) that helps to protect the parasite from host pressures such as complement-mediated lysis in mammals (Späth *et al.*, 2003). While LPG has yet to be isolated from *Phytomonas,* the complete enzyme complement for the synthesis of LPG is present in the *Phytomonas* genomes (Porcel *et al.*, 2014) and thus LPG is likely to be present at the cell surface. Moreover, these genes may also function in the biosynthesis of other classes of structurally related glycolipids known as glycoinositol phospholipids (GIPLs). Interestingly, it has been shown that the major glycolipids of the *Phytomonas* cell surface are GIPLs (Redman *et al.*, 1995). Four different GIPLs have been characterised from a *Phytomonas sp.* isolated from *Euphorbia characias* and collectively these GIPLs are present at ≥ 5 $x10^6$ copies per cell (Redman *et al.*, 1995). The GIPL abundance in *Phytomonas* promastigotes is less than observed for the *Leishmania* promastigote cell surface where there are $1x10^7$ GIPL molecules per cell (Forestier *et al.*, 2014). Given that the cell sizes are very similar, this suggests that the GIPLs coat is less dense in *Phytomonas* than in *Leishmania*. The function

of these highly abundant surface molecules is as yet unknown, however, one hypothesis is that the dense negative charge they give to the cell surface may provide some protection to the *Phytomonas* parasites in the plant host (Redman *et al.*, 1995). Specifically, we propose that in *Phytomonas* the dense negative charge may help repel negatively charged superoxide produced by the plant host as part of the oxidative burst; a rapid, transient production of reactive oxygen species that is one of the first responses in a plant's defence strategy (Jones and Dangl, 2006). It may be that the GIPLs function together with other cell surface glycolipids and glycoproteins in this role, for example in *Leishmania* there is evidence to show that LPG provides some protection against reactive oxygen species (El-On *et al.*, 1980; McNeely and Turco, 1990). The specific composition of GIPLs in *Phytomonas* may also be important for adaptation to different plant hosts. For example, in one *Phytomonas* strain it was found that the GIPLs lack galactose residues that are common in Leishmania GIPLs (Redman *et al.*, 1995). As plants express a diverse array of lectins with affinities for different hexose moieties, it may be that lectins from some plant species will effectively agglutinate individual *Phytomonas* species while others will not (Redman *et al.*, 1995). For example the latex of the host plant (*Euphorbia characias*) from which the *Phytomonas* strain above was isolated, contains a galactose-specific lectin (Barbieri *et al.*, 1983) that may agglutinate any parasite expressing galactose on its cell surface (Redman *et al.*, 1995). Interestingly, though this hypothesis was not tested in the plant host, a galactose-specific lectin from a different lactiferous plant (*Ricinus communis* RCA1) resulted in effective agglutination of the same species of *Phytomonas* (Sanchez-Moreno *et al.*, 1995). Further investigation will reveal whether the composition of GIPLs play a role in plant host specificity.

Intriguingly, immunogenic similarities between *P. serpens* and *T. cruzi* are enough that oral exposure to *P. serpens* has been observed to attenuate the symptoms of Chagas disease in C57BL/6 mice (da Silva *et al.*, 2013). It is also interesting to note that while naturally occurring *Phytomonas* infections in mammals have not been described, inoculation of mice

with infected latex has been reported to produce sustained infections (Franchini, 1922a). Following inoculation, it was noted that the spleen was increased in volume and that blood smears from this organ contained amastigote-like forms of the parasites. It may be that some species of *Phytomonas* are capable of infecting mammals but that their host range is governed by the feeding habits of their invertebrate host. However, caution should be exercised here as the parasites used were identified solely on the basis of their presence in plant latex, and this may have occurred through infection from other opportunistic trypanosomatids. Furthermore, there have yet to be reports of animal infections with *Phytomonas* in wild populations. Taken together, this suggests that if animal infections do exist either they do not persist or re-transmission does not occur.

Comparative surface proteomics between *Phytomonas*, *Leishmania* and *Trypanosoma* may yield further insight into the molecular reasons for these immunogenic similarities. Moreover, it may help to elucidate the evolution of these disparate parasite cell surfaces and identify factors that distinguish plant-adapted from mammalian-adapted trypanosomatids and facilitate colonisation of different plant environments. In Chapter 3 a proteomic integration of the *P. serpens* surface will be discussed.

## 1.10 Discussion

*Phytomonas* are a widespread and diverse group of plant parasites exhibiting both pathogenic and endophytic lifestyles. From their initial discovery in 1909 to the publication of the first *Phytomonas* genomes over 100 years later much progress has been made in their classification, taxonomy and understanding of their cell biology.

In contrast to related *Trypanosoma* and *Leishmania* species the bulk of *Phytomonas* studied to date are not known to cause pathology to the host. The apparent abundance of endophytes in the genus *Phytomonas* may be representative of the extant diversity, or may be the result of a sampling bias. Plant trypanosomatids were first discovered in the latex of a Mediterranean spurge, and as a result, the bulk of subsequent sampling was conducted in the plant family *Euphorbiaceae* (Camargo, 1999; Lafont, 1909). The apparent lack of

disease causing *Phytomonas* may simply be due to the fact that *Euphorbia* infecting *Phytomonas* are generally asymptomatic. Therefore, increased sampling needs to occur before broad conclusions about the diversity and pathogenicity in this genus are made. In this context, it is important to bear in mind that the breadth of the host range (both plant and insect) for any individual *Phytomonas* species is unknown and many aspects of the life cycle are poorly understood.

In *Trypanosoma* and *Leishmania* species, morphologically distinguishable differentiated cell forms are associated with different developmental and life cycle stages. In contrast to this, although individual species show great morphological diversity (Gibbs, 1957) true life cycle stages have yet to be defined and molecular markers that distinguish discrete morphological sub-types have yet to be identified. Thus it remains unknown whether *Phytomonas* cells differentiate to adapt to plant and insect hosts and to what extent both host stages are necessary to complete the life cycle.

Morphological diversity is another unique aspect of *Phytomonas* biology in comparison to sister taxa *Leishmania* and *Trypanosoma.* From their discovery, the extreme polymorphism of *Phytomonas* was noted as a feature which distinguished them from related trypanosomatids (Franca, 1920; Gibbs, 1957). While the functional significance of this morphological diversity is poorly understood, in other trypanosomatids changes in cell morphology are associated with adaptation to different hosts. Furthermore, it is thought that different cell shapes arise as a result of the different selection pressures associated with the contrasting host environments (Wheeler *et al.*, 2013). While not studied in this thesis, further investigation concerning *Phytomonas* cell morphology during the cell cycle could address the question of whether the diverse morphology seen in plant and insect infections is simply a product of the cell division cycle or is adaptive and bears implications for virulence.

The information regarding the host ranges of different *Phytomonas* species is not complete*.* However, the host range (the range of both the plant and insect hosts) of *P. serpens* is understood to be much broader than that observed for a single species of *Trypanosoma* or *Leishmania* (compared to the range of insect and mammalian hosts). This broad host range

may be linked to metabolism. As phytophagous insects feed exclusively on carbohydrate rich plant sap there is no need for *Phytomonas* parasites to switch from carbohydrate to amino acid metabolic substrates for different life cycle stages (Porcel *et al.,* 2014). There is also a limited temperature change when moving from plant to insect host compared to moving from mammalian to insect host. Together, the reduced necessity for metabolic and thermal adaptation may facilitate greater plasticity in the insect and plant host range of *Phytomonas* parasites and allow individual species to infect a broad range of both host types.

One further area where the extent of differences between *Phytomonas* and mammalian infective relatives is not fully understood is at the level of the cell surface proteins. These are likely to be profoundly different due to the requirements of survival in a mammalian host with an adaptive immune response as compared to survival in plant host with plant immune responses. Although it might be expected that the cell surfaces will be different, it is interesting to note that the only protein components of the *Phytomonas* cell surface characterised to date are also conserved in other trypanosomatids and these include the Major Surface Protease and GPI anchored cysteine-proteases. The *Phytomonas* surface also shares some similarity with *Leishmania* in that it contains glycoinositol phospholipids and lipophosphoglycan. This thesis aims to further investigate the conserved and unique proteins that decorate the *Phytomonas* cell surface.

The release of the first *Phytomonas* genomes has already begun to reveal clues into their biology and pathogenicity. For example, the expansion of a gene family of putative secretable aspartyl proteases in the pathogenic isolate is reminiscent of similar phenomena in plant pathogenic fungi (Porcel *et al.*, 2014). It will be interesting to see whether the publication of these novel genome resources will help elucidate mechanisms of pathogenicity in *Phytomonas*, and resolve whether disease symptoms in host plants arise as a result of a parasite mediated degradation of plant tissues, or as an indirect by-product of depletion of plant assimilates. In either case, it is likely that surface proteins play a role in the adaptation of *Phytomonas* to their plant hosts because of the involvement of such proteins in

environmental sensing and parasite-host interactions. Moreover, there is abundant precedent in *Leishmania* and *Trypanosoma* that parasite surface proteins play a crucial role in mediating parasite host-interactions and in evading host immune responses.

## 1.11 Thesis aims

Our knowledge of how an ancestral trypanosomatid evolved and adapted to life inside plant hosts is improving with the availability of experimental and genomic data. We are beginning to understand how parasite metabolism, for example, is adapted and streamlined for a life cycle involving photosynthetic hosts and phytophagous insects. In sister species *Leishmania* and *Trypanosoma* there is precedent for adapting to the host environment and evading the host immune system through a specialised repertoire of cell surface proteins. The aim of the work presented in this thesis was to better understand to what extent *Phytomonas* has adapted the ancestral trypanosomatid cell surface proteome to facilitate parasitism of plant hosts. Proteomic and biochemical characterisation of the *P. serpens* cell surface is presented in Chapter 3 and my attempts to validate the putative cell surface proteins identified in *P. serpens* are presented in Chapter 4. Over the course of the research a further opportunity to gain additional insight into the evolution of the trypanosomatid cell surface was presented. In Chapter 4, *Trypanosoma carassii* is introduced. *T. carassii* is a poorly understood kinetoplastid parasite of fish. The biology of this parasite and fish hosts are presented separately in the introduction of this chapter and the methods optimised for use on *Phytomonas* were applied to investigate the cell surface of this parasite. By using biochemical and genomic evidence from two poorly understood species I hoped to gain insights into both the biology and evolution of the cell surface of trypanosomes as a group. Furthermore, by applying the biochemical and bioinformatic framework developed in *P. serpens* I have developed a platform that enables the study of the cell surface proteins in the trypanosomatids.

Taken together, the work presented in this thesis expands our understanding of the trypanosomatid surface, further extends the dogma that surface proteins largely evolved in a

clade-specific manner, and provides new insight how these intriguing parasites have evolved to inhabit a host range that spans from carp to coconuts.

# *Chapter 2: Materials and Methods*

## 2.1 Preface

Some of the work in this thesis was conducted as part of a collaboration with Professor Geert Wiegertjes (Wageningen University) with the advice of Professor Mark Carrington (University of Cambridge). These procedures comprise those listed in 2.4.3 and 2.5.3. I conducted all other listed experimental procedures.

## 2.2 General Materials

Solutions were made using de-ionised water that was purified using MilliQ (Millipore, www.milipore.com) cartridges. Where necessary, solutions were sterilised by autoclaving at 120˚C for 15 minutes at 100 kPa or by microfiltration through a 0.45 µm filter. Chemicals were purchased from Sigma-Aldrich (www.sigmaaldrich.com), Fisher Scientific (www.fisher.com) or Melford (www.melford.com), unless otherwise stated. Synthetic oligonucleotides were purchased from Eurogentec (www.eurogentec.com) unless otherwise stated. For solutions, % is specified here as weight/volume.

## 2.3 Cells and growth conditions

### *2.3.1* Phytomonas serpens

*Phytomonas serpens* strain 9T was kindly provided by Professor Julius Lukeš, Institute for Parasitology České Budějovice. Unless otherwise stated *P. serpens* parasites were grown in Brain Heart Infusion broth (37 g per 1 L) supplemented with 5% Foetal Calf Serum (both Sigma). Parasites were grown at 27˚C. The osmolality of this completed growth medium was determined to be 410 milliosmol/kg. Where appropriate blasticidin resistant cells were selected by addition of blasticidin (Sigma) to the growth medium at a final concentration of 4 µg $ml^{-1}$.

### *2.3.2* Escherichia coli

*Escherichia coli* XL1-Blue (Agilent Technologies) was used for plasmid manipulation. XL1-Blue Genotype: recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F´ proAB

lacIqZΔM15 Tn10 (Tetr)]. (Genes listed signify mutant alleles. Genes on the F´ episome, however, are wild-type unless indicated otherwise). Cultures were grown on Luria Bertani (LB) medium (1% Tryptone, 0.5% yeast extract, 0.5% NaCl (plus 1.5% agar for solid media)) (Sigma). *E. coli* was cultured at 37 ˚C overnight, either in liquid culture shaking at 200 rpm or solid media in a static incubator. Where appropriate ampicillin resistant bacterial cultures were selected by addition of Ampicillin (Sigma) to the growth medium at a final concentration of 100 μg $ml^{-1}$.

### *2.3.3* Trypanosoma carassii

*Trypanosoma carassii* cultures were kindly provided by the group of Professor G. Wiegertjes and Dr M. Forlenza (Wageningen University). Parasites were grown in carp and isolated in collaboration with Eva Doro (Wageningen University). After 4 weeks of infection blood was collected and left to settle overnight at 4˚C. Live *T. carassii* cells were isolated from the parasite fraction of infected blood by DE52 column purification (Whatman) and grown for 14 days in culture (22.5% Minimum Essential Medium (Sigma), 22.5% L-15 Medium (Leibovitz), 10% sterile water, 45% Hank's Balanced Salt Solution (Fisher), 10% pooled carp serum, 2% L-Glutamine (Sigma) and 1% Penicillin Streptomycin (Fisher)) at 27˚C without $CO_2$. Parasites were purified by DE52 column to remove dead cells and cells originating from the host prior to mRNA extraction and cell surface biotinylation.

### *2.3.4* Phytomonas françai

*Phytomonas françai* was kindly provided by Professor E. Camargo, University of São Paulo, São Paulo, Brazil. Parasites were grown in Brain Heart Infusion broth (37 g per 1 L) supplemented with 10% Foetal Calf Serum. Parasites were grown at 27˚C.

## 2.4 Bioinformatics

Bioinformatics in sections 2.4.1 – 2.4.6 was carried out with the assistance of Steven Kelly and David Emms (Oxford University).

### *2.4.1* **Phytomonas serpens *genome reassembly***

The raw reads for the genome assembly were obtained from Koreny *et al.*, 2012. The available assembly was fragmented and Steven Kelly reassembled a draft genome from the sequence reads to obtain more complete genes. A comparison is presented in Table 2.1. Raw reads were subject to quality filtering using Trimmomatic (Bolger *et al.*, 2014). This was done to remove low quality bases and read-pairs as well as contaminating adaptor sequences prior to assembly. Sequences were searched for all common Illumina adaptors (the default option) and the settings used for read processing by trimmomatic were LEADING:10 TRAILING:10 SLIDINGWINDOW:5:15 MINLEN:50. The trimmed and quality filtered paired-end reads were then subject to read error correction using BayesHammer (Nikolenko *et al.*, 2013) using the default settings. Reads were then digitally normalised using khmer (Brown *et al.*, 2014, Brown *et al.*, 2012) with settings –C 20 -k 21 -M 8e9. Following normalisation, overlapping paired-end reads were joined together using ALLPATHS-LG (Gnerre *et al.*, 2011) and all reads (joined and those that were unable to be joined) were subject to initial *de novo* assembly using SGA (Simpson and Durbin, 2010) with a minimum overlap size of 80 nucleotides (with no mismatches). The assembled contigs were subject to 4 additional rounds of assembly using SGA, each round reducing the overlap size by 10 nucleotides such that the last round of assembly was performed with an overlap size of 40 (with no mismatches). The filtered, normalised joined reads were then mapped to this assembly using bowtie2 (Langmead and Salzberg, 2012). Reads that were absent from the assembly were identified and combined to the assembly file. The combined un-mapped reads and assembled contigs were subject to assembly using SGA with an overlap size of 70. This process of identifying unmapped reads and reassembling with SGA was repeated, each time decreasing the overlap size by 10 nucleotides until the final round of assembly was performed with an overlap size of 40 nucleotides. Contigs were then subject to scaffolding using SSPACE (Boetzer and Pirovano, 2014) and the full set of corrected, normalised paired-end reads using the settings -k 10 -a 0.7 -n 50 -o 20. Scaffolds were

subject to gap filling using the SGA gap filling function. Finally contigs are subject to assembly base-error correction using Pilon (Walker *et al.*, 2014). The final assembly contained 1841 sequences with a scaffold N50 of 28054 nucleotides. A comparison of the original assembly (Koreny *et al.*, 2012) and the reassembly by Steve Kelly is shown in (Table 2.1).

## Table 2.1 *Phytomonas serpens* genome reassembly statistics

A summary to compare the statistics for the *P. serpens* genome from the original assembly AIHY01.1fsa_n (Koreny, 2012) and the reassembly (this thesis). N50 is the weighted median statistic such that 50% of the entire assembly is contained in contigs or scaffolds equal to or longer than this length. Units are DNA base pairs.

| **FEATURE** | Original Assembly AIHY01.1.fsa_n | Reassembly |
|---|---|---|
| **NUMBER OF SEQUENCES:** | 23748 | 1841 |
| **MEDIAN SEQUENCE LENGTH:** | 401 | 3649 |
| **MEAN SEQUENCE LENGTH:** | 1081 | 9974 |
| **MAX SEQUENCE LENGTH:** | 53524 | 194433 |
| **MIN SEQUENCE LENGTH:** | 200 | 91 |
| **NO. SEQUENCES > 1K BP:** | 3973 | 1442 |
| **NO. SEQUENCES > 10K BP:** | 405 | 481 |
| **NO. GAPS:** | 1 | 3818 |
| **N50:** | 3687 | 28054 |
| **COMBINED SEQUENCE LENGTH:** | 25693183 | 18363416 |

### *2.4.2 Prediction of protein coding genes*

The assembled draft genome of *P. serpens* was subject to gene model prediction using Augustus (Keller *et al.*, 2011). An initial set of gene models were predicted using gene prediction parameters inferred by training using the *Leishmania mexicana* genome. These gene model parameters were used to predict a training set of genes in the draft assembly of *P. serpens*. Multiple iterations of prediction and training were performed until

no further genes could be detected. The final set of gene predictions contained 8320 protein coding genes (Table 2.2).

## Table 2.2 *Phytomonas serpens* predicted protein sequences

A table summarizing the predicted protein sequences from the reassembled *P. serpens* genome compared to *L. mexicana* and *T. brucei.* Units are amino acids.

| FEATURE | *P. serpens* | *L. mexicana* | *T. brucei* |
|---|---|---|---|
| **NUMBER OF SEQUENCES:** | 8320 | 8250 | 11567 |
| **MEDIAN SEQUENCE LENGTH:** | 316 | 465 | 398 |
| **MEAN SEQUENCE LENGTH:** | 426 | 627 | 484 |
| **MAX SEQUENCE LENGTH:** | 5480 | 6738 | 8214 |
| **MIN SEQUENCE LENGTH:** | 66 | 4 | 9 |
| **NO. SEQUENCES > 1K AA** | 567 | 1259 | 902 |
| **NO. SEQUENCES > 10K AA** | 0 | 0 | 0 |

### *2.4.3* Trypanosoma carassii *transcriptome*

Bloodstream form *T. carassii* parasites were grown in carp (2.3.3). Total RNA was extracted from these parasites by Prof. Mark Carrington (University of Cambridge) and the mRNA fraction sequenced using Illumina paired-end sequencing technology by the Beijing Genomics Institute (BGI,http://bgi-international.com/). Raw reads were subject to quality filtering using Trimmomatic (Bolger et al., 2014) and searched for all common Illumina adaptors (the default option) and the settings used for read processing by Trimmomatic were LEADING:10 TRAILING:10 SLIDINGWINDOW:5:15 MINLEN:50. The trimmed and quality filtered paired-end reads were then filtered to remove reads coming from ribosomal RNA using SortMeRNA using the default settings, before being subject to read error correction using BayesHammer (Nikolenko et al., 2013) with the default settings. Reads were then digitally normalised using khmer (Brown et al., 2014) with settings –C 20 -k 21 -M 8e9. Following normalisation, overlapping paired-end reads

were joined together using ALLPATHS-LG (Gnerre et al., 2011) and all reads (joined and those that were unable to be joined) were subject to initial *de novo* assembly using SGA (Simpson, 2014) using minimum overlap size of 80 nucleotides (with no mismatches). The assembled contigs were subject to four additional rounds of assembly using SGA, each round reducing the overlap size by 10 nucleotides such that the last round of assembly was performed with an overlap size of 40 (with no mismatches). The filtered, normalised joined reads were then mapped to this assembly using bowtie2 (Langmead and Salzberg, 2012). Reads that were absent from the assembly were identified and combined to the assembly file. The combined un-mapped reads and assembled contigs were subject to assembly using SGA with an overlap size of 70. This process of identifying unmapped reads and reassembling with SGA was repeated each time decreasing the overlap size by 10 nucleotides until the final round of assembly was performed with an overlap size of 40 nucleotides. Contigs were then subject to scaffolding using SSPACE (Boetzer and Pirovano, 2014) and the full set of non-ribosomal, corrected, normalised paired-end reads using the settings -k 10 -a 0.7 -n 50 -o 20. Scaffolds were subject to gap filling using the SGA gap filling function. Finally the assembled contigs were subject to base-error correction using Pilon (Walker et al., 2014) with the default settings prior to error correction with BayesHammer (Nikolenko *et al.*, 2013).

### *2.4.4* Phytomonas françai *genome sequencing and assembly*

To provide a draft genome assembly, total genomic DNA was isolated from *Phytomonas françai* cells grown in culture (section 2.3.4) using Qiagen Blood/Tissue DNA Extraction Kit (Qiagen, UK). Isolated DNA was sent to BGI and subjected to Illumina sequencing using a 5 library approach comprising three libraries of 100bp long paired-end reads with differing insert sizes (170bp, 500bp and 800bp) and two libraries of 100bp long mate-pair reads (2000bp and 5000bp insert size respectively). The genome assembly was performed by Steven Kelly. The raw reads were subject to quality trimming and filtering

using Trimmomatic (Bolger *et al.*, 2014) using the following settings LEADING:10 TRAILING:10 SLIDINGWINDOW:5:15 MINLEN:50. The quality filtered reads were then subject to *de novo* assembly using the ALLPATHS-LG algorithm. The resultant assembly was then subject to twelve rounds of assembly error correction and gap filling with Pilon using the –fix all option and setting the expected ploidy to diploid (Walker *et al.*, 2014). The final draft assembly contained 96 scaffolds greater than 1000bp in length with an N50 of 784kb and a total assembly length of 17.7Mb. The assembled draft genome of *P. françai* was subject to gene model prediction using Augustus (Keller *et al.*, 2011). To develop specific gene model prediction parameters for *P. françai*, a preliminary set of gene models were first predicted using gene prediction parameters inferred by training using gene predictions in existing *Phytomonas* genomes (Porcel *et al.*, 2014). The genome wide distribution of these initial predictions were interrogated. Gene models that were predicted to occur in polycistronic transcription units on the opposite strand to the direction of transcription were discarded and the remaining gene predictions were used to train *P. françai* specific gene model parameters. Multiple iterations of prediction, filtering and training were performed until no further genes could be detected. The final prediction comprised 7108 genes. The spliced leader acceptor and polyadenylation sites for these genes were identified and quantified using SLaP mapper (Fiebig *et al.*, 2014). This assembly is deposited in GenBank under the accession number MJCC00000000. Raw genome reads are deposited in NCBI SRA under the accession numbers SRR4244756-SRR4244760. RNAseq reads are available through EBI under the accession number E-MTAB-5110.

### *2.4.5 Inference of orthogroups*

An orthogroup is defined here as the set of genes that are descended from a single gene in the last common ancestor of the species used to generate the gene families (Emms and Kelly, 2015). The annotated proteins for *Trypannosoma brucei* TREU927, *Trypanosoma congolense* IL3000, *Trypanosoma cruzi* CL brenner Esmereldo-like,

*Trypanosoma grayi* ANR4, *Trypanosoma vivax* Y486, *Leishmania tarentolae* ParrotTarII, *Leishmania major* Friedlin, *Leishmania infantum* JPCM5, *Leishmania mexicana* MHOMGT2001U1103, *Leishmania donovani* BPK282A1, *Leishmania braziliensis* MHOMBR71973M2269 and *Crithidia fasiculata* CfCl were downloaded from TriTrypDb V8.0 (Aslett et al. 2010). The predicted protein sequences from *Trypanosoma rangeli* were downloaded from NCBI (Stoco *et al.*, 2014). The predicted proteins from *Phytomonas EM1* and *Phytomonas Hart1* were obtained from Porcel *et al.* (2014). The predicted proteins from *Phytomonas serpens* were obtained as described in section 2.3.2 and *Phytomonas françai* as in 2.3.4. Finally, two *de novo* assembled transcriptomes for *Trypanosoma carassii* and *Trypanoplasma borelli* were included (M. Carrington, Personal Communication). Orthogroups were inferred with OrthoFinder (Emms and Kelly, 2015) using default parameters.

### *2.4.6 Phylogenetic inference*

Orthogroups containing only single copy genes in all species were selected for inclusion in the phylogenetic analysis. The individual orthologous groups were aligned using MergeAlign (Collingridge and Kelly 2012) and edited to remove all gap-containing columns and columns containing fewer than two character states. MergeAlign was selected as it has previously been shown to increase topological congruity between individual gene trees in multi-gene phylogenomic analyses (Collingridge and Kelly 2012). This resulted in a set of 496 alignments. Individual phylogenetic trees were inferred from each alignment using RAxML (Stamatakis 2006) using the PROTGAMMAAUTO model of sequence evolution. A subset of the fastest evolving gene families, those phylogenetic trees with substitutions site$^{-1}$ branch$^{-1}$ > 0.3, were discarded to minimise long-branch attraction artefacts. 100 bootstrapped concatenated multi-gene multiple sequence alignments were then constructed from this subset using an equi-sampling strategy. In this strategy each bootstrap replicate is sampled (at random with replacement) 70 un-gapped phylogenetically informative columns from each multiple sequence alignment

containing >70 columns. This equi-sampling strategy was performed so that each gene contributed equally to the final phylogenetic tree, thus preventing longer genes from biasing the result. Each bootstrap replicate contained 30,030 phylogenetically informative, un-gapped aligned positions spanning all species. Trees were inferred from each resampled multi-gene alignment using RAxML utilising the PROTGAMMAGTR model of sequence evolution and using minimum evolution principle (with log corrected scores) implemented in FastME (Desper and Gascuel 2002).

### *2.4.7 Prediction of putative membrane orthogroups*

Signal peptide, transmembrane domain, and GPI anchor) were be predicted from the amino acid sequence using computational methods. SignalP (http://www.cbs.dtu.dk/services/SignalP/) was used to predict signal peptides, FragAnchor (https://omictools.com/fraganchor-tool) was used to predict GPI anchor attachment and TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) was used to predict transmembrane domains. These computational methods are not perfect and can produce both false positives (i.e. predict features that are absent) and false negatives (i.e. fail to detect *bona fide* features). Given that predictions for an individual gene may be unreliable I sought to increase the classification robustness of genes encoding putative surface proteins by analysing the presence or absence of predicted surface-associated features for all genes in the same orthogroup. In this way, predictions for homologous genes in other species were used as evidence to support the presence/absence of membrane associated domains. This approach assumes that the features are evolutionarily conserved and therefore the localisation of the protein is also evolutionarily conserved.

### 2.4.8 *Prediction of putative cell surface orthogroups for* Trypanosoma carassii

Predicted protein sequences used in the clustering described in section 2.4.5 from the species *T. carassii*, *T. brucei, T. vivax, T. congolese, T. cruzi, T. grayi* and *T. rangeli*

were subject to computational analysis to identify protein features that are characteristic of cell surface proteins. Signal peptides were predicted using SignalP 4.0 (Petersen *et al.*, 2011), GPI anchors were predicted using FragAnchor (Poisson *et al.*, 2007) and transmembrane helices predicted using TMHMM v2.0 (Krogh *et al.*, 2001). Pseudogenes were discarded and not considered in this analysis. *T. carassii* proteins that were truncated at either the N′ or C′ terminus were deemed ineligible for signal peptide or GPI anchor predictions respectively. Orthogroups were designated as surface orthogroups if ≥40% of the genes in the orthogroup contained a signal peptide and at least one transmembrane helix, or a signal peptide and a GPI anchor. This cut-off of ≥40% was employed on the basis of observed values for a gp63, a ubiquitously conserved cell surface protein (of all *gp63* genes in trypanosomatids 48.4% have a computationally predicted GPI anchor and 46.4% have a computationally predicted signal peptide. All gp63 proteins are thought to be cell surface localised).

### *2.4.9 Prediction of cell surface orthogroups for Phytomonas*

The methods were the same as described in section 2.4.7 with the inclusion of *L. major, L. mexicana, L. infantum, L. tarentolae, L. braziliensis, P. EM1, P. HART1, P. serpens* and *P. françai.* The inclusion of these species gave the observed frequencies of GPI anchors and signal peptide in gp63 at 40.1% GPI and 35.9% signal peptide. Orthogroups were designated as surface orthogroups if ≥35% of the genes in the orthogroup contained a signal peptide and at least one transmembrane helix, or a signal peptide and a GPI anchor.

## 2.5 Proteomic methods

### *2.5.1 Live cell biotinylation of* **Phytomonas serpens**

A total of $1x10^9$ *P. serpens* promastigotes in mid-log phase growth were used for cell surface biotinylation. Protease and phosphatase inhibitor cocktails (Sigma) were added to solutions to prevent degradation of proteins in the cell extracts. A number of different combinations of beads, buffers, timings and reagents were tested in the optimisation of

the surface biotinylation method presented in this thesis. These disparate combinations are presented here as separate protocols. All biotinylation was performed after washing.

**Protocol 1** Cells were washed 3 times in 1x phosphate-buffered saline (PBS: 0.01 M $PO_4^{3-}$, 0.0027 M KCl and 0.137 M NaCl, pH 7.4) (Sigma) and were biotinylated as per 2.5.1.1, this was first done at room temperature as performed by Yao *et al.* (2012) and optimised to 4ºC on ice, as described in Chapter 3 (3.3.2, Figure 3.3). Labelling was quenched with 2-amino-2-(hydroxymethyl)-1,3-propanediol (Tris) (Sigma) as per 2.5.2.1 (Yao et al., 2012). Cells were washed 3 times with 1x Tris buffered saline (TBS: 50 mM TrisCl, 150 mM NaCl pH 7.5). Cells were lysed as per 2.5.4 and 2.5.4.1 (Yao *et al.*, 2012). The whole cell lysate was centrifuged at 10,000 *g* to pellet detergent-insoluble material. The detergent-soluble fraction was incubated on streptavidin agarose beads (Thermo Scientific) as Yao *et al.* (2012), but this was subsequently optimised to streptavidin Dyna beads (Invitrogen). Beads were washed 5 times in lysis buffer (2.5.4) and prepared for mass spectrometry (2.5.5.1), optimisation and results are described in Chapter 3 (3.3.3, Figure 3.5).

**Protocol 2** Given that cells were grown in a medium which had an osmolality of 410 milliosmol/kg, all wash buffers used for live cells were constituted to be osmotically balanced to prevent unwanted osmotic lysis of cells during wash and incubation steps. Cells were washed 3 times in 10ml ice cold osmotically balanced PBS and biotinylated as per 2.5.1.2. Excess biotin was quenched as per 2.5.2.2. Cell lysis was carried out with buffer L4 (2.5.4 and 2.5.4.2). Optimisation of lysis buffer is presented in Chapter 3 (3.3.4, Figure 3.7). Whole cell lysate was incubated on Dynabeads (2.5.4). The elution methods for NHS-LC and NHS-SS biotin were different, NHS-LC biotin as described in 2.5.5.2 and NHS-SS biotin in 2.5.5.3. For both NHS-LC and NHS-LC, unbound material was prepared for LC MS/MS as described in 2.5.6. These are the results described in Chapter 3 (3.3.6).

#### *2.5.1.1 Biotinylation with PBS and 1mM EZlink Sulfo-NHS-LC-Biotin*

Cells were washed 3 times in 1x PBS and were biotinylated by the addition of 1mM EZlink Sulfo-NHS-LC-Biotin to the wash buffer.

*2.5.1.2 Osmotically balanced biotinylation with PBS*

Cells were washed with osmotically balanced PBS (OPBS: 262 mM NaCl, 2.7 mM KCl, 10 mM $Na_2HPO_4$, 1.8 mM $KH_2PO_4$, pH 7.5) and biotinylated by addition of either 1mM EZ-link Sulfo-NHS-SS-Biotin or 1mM EZ-link Sulfo-NHS-LC-Biotin (both Thermo Scientific) on ice for 10 minutes.

### *2.5.2 Quenching the biotinylation of* Phytomonas serpens

Labelling reactions were quenched with Tris.

*2.5.2.1 Quenching with Tris-buffered saline*

Cells were biotinylated as in 2.5.1.1, then washed 3 times with 1x Tris-buffered saline (TBS: 50 mM TrisCl, 150 mM NaCl pH 7.5).

*2.5.2.2 Quenching with osmotically balanced Tris-buffered saline*

Cells were biotinylated as in 2.5.1.2, then washed 3 times with 1x osmotically balanced Tris-buffered saline (OTBS: 50 mM TrisCl, 180 mM NaCl pH 7.5).

### *2.5.3 Live cell biotinylation of* Trypanosoma carassii

Following isolation of cells from carp as described in section 2.3.3, *T. carassii* cells were washed once in 10 ml ice cold trypanosome dilution buffer (TBD: 118 mM NaCl, 1.2 mM $KH_2PO_4$, 30 mM TES, 16 mM $NaHPO_4$, 5 mM $NaHCO_3$, 5 mM KCl, pH 8) and biotinylated by addition of EZ-link Sulfo-NHS-SS-Biotin (Thermo Scientific) to cells in trypanosome dilution buffer and incubated on ice for 10 minutes. The labelling reaction was quenched with ethanolamine (Sigma) and cells were washed 3 times with 1x TBS.

### *2.5.4 Enrichment of biotinylated proteins*

Surface biotinylated cells were lysed with lysis buffer. Cell lysate was then incubated on streptavidin Dynabeads (Invitrogen) for 1h at 4˚C with rotation. Two different cell lysate buffers were used for *P. serpens.*

#### *2.5.4.1 Nonidet P-40 lysis*

After quenching, cells were lysed by addition of 1ml 2% Nonidet P-40 (Sigma) in 1x TBS.

#### *2.5.4.2 L4 buffer lysis*

After quenching, cells were lysed by addition of 1ml L4 lysis buffer (8 M urea, 1% sodium deoxycholate, 0.2% SDS (Sodium dodecyl sulfate), 1 M NaCl in 1x TBS, pH 7.5).

Unbound material from the cell lysate i.e. the supernatant after bead incubation, is referred to as flow through or unbound fraction. Material bound to the beads and later released is called the eluate or elution fraction.

### *2.5.5 Release of biotinylated proteins from streptavidin resin*

To release the biotinylated material from the resin the following methods were used:

#### *2.5.5.1 Elution with lysis buffer*

Following incubation of the detergent-soluble fraction on streptavidin resin, the bound material was collected by either centrifugation for streptavidin agarose beads or by using a magnet for Dynabeads, the beads washed 5 times in 2% Nonidet P-40 in 1x TBS and resuspended in 50 µl 2x Laemmli buffer (100 mM Tris pH 6.8, 4% SDS, 0.2% Bromophenol blue, 20% glycerol, 200 mM β Mercaptoethanol) and boiled for 5 minutes. The unbound material was mixed 1:1 with 2x Laemmli buffer (100 mM Tris pH 6.8, 4% SDS, 0.2% Bromophenol blue, 20% glycerol, 200 mM β-Mercaptoethanol) and boiled for 5 minutes. To prepare samples for mass spectrometry, the eluate, unbound material and negative control eluate were loaded into an SDS-PAGE gel 7-14% NuPage gel (Invitrogen) and run at 25V/cm for 5

minutes. Gel slices were reduced and alkylated with 20 mM iodoacetamide (Sigma) for 30 minutes in the dark. *N*-linked glycans were removed with Peptide-N-Glycosidase F (PNGase F) (New England Biosystems) according to the manufacturer's instructions and proteins digested in-gel with 1 mg trypsin (Sigma) at 37 ˚C with agitation.

*2.5.5.2 Elution of NHS-LC biotin*

Following incubation of whole cell lysate on Dynabeads, the beads were retained using a magnet and the unbound material removed. The beads were washed 3 times with buffer L4 and 3 times with bead wash buffer (50 mM Tris-Cl, pH 7.5 and 1% sodium deoxycholate) to remove any residual SDS from the sample. For analysis by enhanced chemiluminescence (ECL) as in 2.5.10, the beads were retained by magnetisation and the solution removed. The beads were resuspended in 50 µl 2x Laemmli buffer. The unbound material mixed 1:1 with 2x Laemmli buffer. Both samples boiled for 5 minutes. To prepare samples for mass spectrometry, the beads were retained and separated from the final wash using a magnet. The sample was reduced with 10 mM tris(2-carboxyethyl)phosphine (TCEP, Sigma) for 15 minutes at 37 ˚C, alkylated with 20 mM iodoacetamide for 30 minutes in the dark, *N*-linked glycans were removed with PNGase F and tryptic digestion performed on bead, with 1 mg trypsin at 37 ˚C with agitation. Finally, the solution was removed and the beads retained with a magnet.

*2.5.5.3 Elution of NHS-SS-biotin*

After labelling with NHS-SS biotin and cell lysis with buffer L4. the beads were washed 3 times with buffer L4 and 3 times with bead wash buffer. Bound material was removed by reduction with 10 mM TCEP for 15 minutes at 37 ˚C. Using a magnet. the beads were retained and the solution removed. The bead eluate (solution) was alkylated with 20 mM iodoacetamide for 30 minutes in the dark, *N*-

linked glycans removed with PNGase F and tryptic digestion in solution, using 1 mg trypsin at 37 ˚C with agitation.

*2.5.5.4 NHS-LC and NHS-SS biotin sample preparation for mass spectrometry*

For both 2.5.5.2 and 2.5.5.3 the unbound material was loaded into an SDS-PAGE gel 7-14% NuPage gel (Invitrogen) and run at 25 V/cm for 5 minutes. Gel slices were reduced and alkylated with 20 mM iodoacetamide for 30 minutes in the dark and proteins digested in-gel with 1 mg trypsin at 37 ˚C with agitation.

### *2.5.6 SDS-polyacrylamide gel electrophoresis of proteins*

10% acrylamide resolving gels (10% bis-acrylamide (Bio-Rad), 1.5 M Tris pH 8.8, 10% SDS, 10% APS, 0.001% TEMED) were cast and overlaid with 5% stacking gels (5% bis-acrylamide (Bio-Rad), 1 M Tris pH 6.8, 10% SDS, 10% APS, 0.002% TEMED). Protein samples were loaded in Laemmli buffer and electrophoresed at 20A per gel with running buffer (25 mM Tris base, 250 mM glycine, 0.1% SDS).

### *2.5.7 Coomassie staining of polyacrylamide protein gels*

To visualise protein bands within SDS-PAGE gels, gels were stained with Coomassie stain (0.5% Coomassie Blue R, 10% acetic acid, 50% methanol) for 30 minutes and destained for 4 hours or longer using destain solution (10% acetic acid, 20% methanol). Destained gels were imaged using a ChemiDoc MP (BioRad).

### *2.5.8 SYPRO Ruby staining of polyacrylamide gels*

For a more sensitive protein stain SYPRO Ruby (Sigma) was used. Gels were stained overnight in the dark at room temperature with SYPRO Ruby stain. Gels were washed (10% methanol, 7% acetic acid) for 30 minutes at room temperature. Gels were imaged under UV light on a Biospectrum Image System with Vision Works software (UVP).

### *2.5.9 Transfer of proteins from gels to nitrocellulose membrane (blotting)*

The blotting sandwiches were assembled according to the manufacturer's instructions in semi-dry blotting buffer (28 mM Tris pH9.2, 39 mM Glycine 20% methanol) with Protran

nitrocellulose membrane (Whatman) and extra thick blotting paper. Transfer was performed at 12 V 1 A for 30 minutes on a Trans-Blot Turbo (BioRad).

### *2.5.10 Enhanced Chemiluminescence (ECL) detection*

Following transfer of proteins to nitrocellulose, membranes were incubated in 20 ml blocking buffer (5% dried milk (Tesco) in 1X TBS, 0.05% Tween-20 (Sigma)) for 1 hour with shaking. Membranes were washed 3 times in TBST (1X TBS, 0.05% Tween-20 ) before incubation with 10 ml TBST containing streptavidin HRP (Cell Signalling Technology) at a dilution of 1:2000 at room temperature with shaking for 1 hour. Membranes were washed 5 times with 400ml TBST, 5 minutes per wash with shaking. 1 ml of enhanced luminol reagent and 1 ml oxidising reagent (both PerkinElmer) were added to the membrane and incubated with gentle agitation for 1 minute. Membranes were put in a plastic sleeve and fixed inside a light-proof cassette and exposed to Kodak X-OMART AR film (Kodak). Film was developed and fixed using Kodak AR chemicals.

### *2.5.11 Tandem mass spectrometry*

Tandem mass spectrometry was performed by the University of Oxford Central Proteomics facility (CPFP, http://www.proteomics.ox.ac.uk/). Samples from 2.5.5.1, 2.5.5.2 and 2.5.5.3 were acidified with 0.5% Trifluoroacetic acid to remove sodium deoxycholate. The supernatants were desalted on C18 and analysed by LC MS/MS using a QExactive Orbitrap mass spectrometer (Thermo Electron) coupled to an Ultimate 3000 RSLCnano HPLC (Dionex) system run in direct injection mode. Samples were resolved on a 25 cm by 75 micron inner diameter home-packed column with ProntoSIL 120-3 C18 Ace-EPS phase, 3 µm bead (Bischoff Chromatography,). The system was operated at a flow-rate of 300nL $min^{-1}$. A 100 minute gradient was used to separate the peptides with a linear gradient ranging from 5-50% buffer B. The mass spectrometer was operated in a “Top *20*” data dependent acquisition mode with a scan range of 350-1650 *m/z*. Precursor scans were performed in the orbitrap at a resolving power of 70,000 from which the twenty most intense precursor ions were selected by the quadrupole and

fragmented by higher energy collision dissociation at a normalised collision energy of 30%. The quadrupole isolation window was set at *3 m/z.* Charge state +1 ions and undetermined charge state ions were rejected from selection for fragmentation. Dynamic exclusion was enabled for 40s.

SINQ Quantitation was carried out as described by Trudgian *et al.*, (2011) and is summarised here. Raw Spectral Index values are obtained for a protein using a weighted sum of the matched-ion-current intensity (the intensity of matched fragment ions) for all spectra assigned to peptides which belong to that protein. Then ProteinProphet (Nesvizhskii *et al.*, 2003) was used to infer protein identifications. This groups protein identifications that share common peptides. Common peptides make a weighted contribution to the probability score of each protein they occur in. The weights are used to assign quantitation to individual proteins within groups. This introduces a weighted term to the SI calculation where *P* is the number of peptides assigned to protein *x*, *wy* is the weight given by ProteinProphet for the assignment of peptide *y* to protein *x*, *S* is the number of spectrum matches for peptide *y* and $i_{y,z}$ is the summed fragment ion-intensity for the assignment of spectrum *x* to peptide *y*.

$$SI_x = \sum_{y=1}^{P} W_y \left( \sum_{x=1}^{S} i_{y,z} \right)$$

The $SI_x$ value for each protein is normalised by the sum of SI values across all N proteins in the dataset. The resulting value is divided by the protein length to correct for longer proteins producing a greater number of peptides (and consequently higher SI values than shorter proteins).

$$SIN_x = \frac{SI_x}{\sum_{j=1}^{N} SIj} / l$$

### *2.5.12 Data searches*

Data were converted from ".raw" file format to ".mgf" file format using ProteoWizard MSconvert (Chambers et al., 2012). ".mgf" files were uploaded onto the Central Proteomics Facility Pipeline server (Trudgian et al., 2010).

**Data searches for *T. carassii*:** data were searched against an in-house constructed database containing the predicted proteins encoded in the *T. carassii* transcriptome using the following search parameters: full tryptic cleavage; 2 missed cleavages; carbamidomethyl c fixed modification; oxidised (M) and addition of the reacted and reduced NHS-SS-biotin ($C_5H_7O_2NS$) (K, N-terminal protein) variable modifications; precursor peptide tolerance 20ppm, fragment tolerance 0.1 Da and charge states =+1,+2,+3.

**Data searches for *P. serpens:*** data were searched against an in-house constructed database of the *P. serpens* predicted protein sequences using the following parameters: full tryptic cleavage; 2 missed cleavages; carbamidomethyl c fixed modification; oxidised (M), addition of the reacted and reduced NHS-SS biotin (for SS biotin labelled samples) and deamidated (N) and (Q) variable modifications; precursor peptide tolerance 20ppm, fragment tolerance 0.1 Da and charge states =+1,+2,+3.

The CPFP uses a combination of two search engines; Mascot (Matrixscience) and OMSSA (Geer et al., 2004), and combines the search results for increased coverage using the Trans Proteomic Pipeline (TPP) InterProphet tool (Deutsch et al., 2010). Label free quantitation was performed using the Spectral Index Normalised Quantitation (SINQ) method (Trudgian et al., 2011) based on the Normalised Spectral Index method (Griffin et al., 2010). In all cases the SINQ values for enriched proteins eluted from beads were compared to the SINQ values for the same proteins in the flow-through. Matched Ion Current intensity was used for quantitation and protein identifications were filtered for a protein false discovery rate of 1% with at least 2 peptide sequences. Where spectral

indices for a given protein coding gene were missing in one or other sample the value obtained for the lowest abundant protein that was detected was used.

### *2.5.13 Detection of biotin tag by epifluorescence microscopy*

Biotinylated and non-biotinylated control cells were fixed in 4% ice cold PFA (paraformaldehyde, Sigma), washed in 1x PBS and settled onto poly-L-lysine slides (Sigma) for 20 minutes. Cells were permeabilised in -20˚C methanol for 5 minutes and washed with 1x PBS. Slides were stained with Streptavidin AlexaFluor 488 (Life Technologies) and Hoechst 33342 (10mg/ml stock diluted 1:1000 working concentration) (ThermoFisher) for 60 minutes. Slides were then washed 3 times in 1x PBS, mounted with FluroMount (Sigma) and sealed with nail varnish. Cells were imaged on a Leica DM5500B microscope described in section 2.8.

## 2.6 Molecular cloning

### *2.6.1 Polymerase Chain Reaction (PCR)*

Constructs were designed *in silico* using A Plasmid Editor (ApE) and primers that were 18-25 nucleotides in length were designed using the 'Find Primer' tool in ApE (http://biologylabs.utah.edu/jorgensen/wayned/ape/). For standard reactions GoTaq polymerase (Promega) was used to amplify DNA sequences. Reactions were assembled on ice and consisted of 1x GoTaq green buffer, 1-4 mM $MgCl_2$, 0.2 mM each dNTP, 0.1-1 µM of each primer, 1.25 unit enzyme and <0.5 µg/50 µl template DNA, where necessary DMSO was used as a PCR additive at 2%. Amplification was carried out in thermocyclers (Applied Biosystems or BioRad) as below:

| | | | |
|---|---|---|---|
| Denaturation | 94˚C | 4 minutes | |
| Denaturation | 94 ˚C | 15 seconds | x 35 cycles |
| Annealing | 50-60 ˚C | 30 seconds | |
| Extension | 72 ˚C | 1 minute | |
| Extension | 72 ˚C | 4 minutes | |

### *2.6.2 Nested PCR tagging*

For the nested PCR approach for tagging candidate genes with yellow fluorescent protein (Dean *et al.*, 2015), Phusion polymerase (Thermo Scientific) was used. Reactions were assembled on ice and consisted of 1x GC or HF buffer, 200 µM each dNTP, 0.5 µM of each primer, 1 unit enzyme and <250 ng template DNA. Where necessary DMSO was used as a PCR additive at 2%.

500 bp target DNA sequences were amplified using the cycling conditions:

| | | | |
|---|---|---|---|
| Denaturation | 98˚C | 30 seconds | |
| Denaturation | 98 ˚C | 10 seconds | x 35 cycles |
| Annealing | 50 ˚C | 30 seconds | |
| Extension | 72 ˚C | 1 minute | |
| Extension | 72 ˚C | 5 minutes | |

The pPOTv2 vector, kindly provided by the laboratory of Professor Keith Gull, was used for tagging proteins at the N-terminus and is shown in Figure 2.1. This vector contains the TY epitope tag is a 10 amino acid (EVHTNQDPLD) fragment of a protein from *Saccharomyces cerevisiae* transposable element Ty1 (Bastin et al., 1996). The TY tag is recognised by the BB2 monoclonal antibody (Bastin et al., 1996), this is helpful when it is advantageous to double label e.g. in immunofluorescence or for alternatives to anti-GFP antibodies in Western Blotting. The pJ1258 vector, given by Dr J Sunter, is a modification of the pLENT tagging plasmid with a shorter intergenic region between the Blasticidin resistance gene and eYFP (Dean *et al.*, 2015).

The pJ1258 vector was used in the C-terminal tagging of proteins. pPOTv2 was used in the N-terminal tagging of candidate proteins. The tagging vectors were used in this way because of the orientation of the coding sequences for eYFP and selectable marker, i.e. to create a fusion protein at the N′ terminus (pPOTv2) or C′ terminus (pJ1258) of the target protein.

The overlap PCR was performed using the following thermal cycle scheme with addition of the primers after the completion of the first 6 cycles:

| | | | |
|---|---|---|---|
| Denaturation | 98˚C | 30 seconds | |
| Denaturation | 98 ˚C | 10 seconds | x 6 cycles |
| Annealing | 50 ˚C | 30 seconds | |
| Extension | 72 ˚C | 1 minute 30s | |
| Denaturation | 98 ˚C | 10 seconds | x 30 cycles |
| Annealing | 55 ˚C | 30 seconds | |
| Extension | 72 ˚C | 5 minutes | |

PCR was validated by gel electrophoresis on 1% agarose gels (tanks and power packs from BioRad), visualised using 0.1% ethidium bromide (Sigma) and imaged at 365nm using a Biospectrum Image System with Vision Works software (UVP). Where multiple gel bands were produced, the band of the correct length was cut out and gel purified using the QIAquick Gel Extraction Kit (Qiagen). Nested primers and validation primers are listed in Appendix 1.

### *2.6.3 Synthesis of an endogenous signal peptide*

The endogenous signal peptide cleavage site for the protein Pse.270.1010 was predicted using signalP 4.0 (Petersen *et al.*, 2011) and the coding sequence synthesised *de novo* by Eurofins. The original and adapted sequences are shown below. The sequence was adapted by Eurofins proprietary software GENEius.The signal peptide was inserted into the pPOTv2 vector using the HindIII and SpeI restriction sites and transformed into heat shock competent XL1 Blue *E. coli (Agilent Technologies)*. Correct integration was validated by colony PCR using the following primer sequences: 5′-CTGCTGGTGAGCCTCTTACA-3′ and 3′-TTGAAGAAGATGGTGCGCTCCT-5′. Colonies were then grown in 5 ml LB medium and plasmids extracted by miniprep using

a Qiagen QIAprep spin miniprep kit (Qiagen). Correct integration of the signal peptide was validated by PCR.

Original Sequence:

ATG CGC ACC AAA ACA GAA TTG ATG GCG CTC TTT TGC GCT GCG CTC CTA CTG GTG AGC TTG TTG CAG GCC GAC

Adapted Sequence:

ATG CGC AC**G** AAA AC**C** GAA **C**TG ATG GCG CT**G** TTT TGC GCT GC**A** **T**T**G** CT**G** CTG GTG AGC **C**T**C** TT**A** CAG GCC GA**T**

**Eurofins proprietary software GENEius optimises and adapts amino acid or DNA sequences.**

The adaptation of the codon usage of cloned DNA fragments is of capital importance for optimum gene expression.

Generally, codon optimisation can improve gene expression in heterologous expression systems. The main functionalities are:

- Avoidance of rare codons of the expressing organism
- Usage of codons with high frequencies based on the selected organism
- Removal of unwanted motifs or secondary structures like repeat or hairpin structures
- Even out the distribution of G and C bases

https://www.eurofinsgenomics.eu/en/gene-synthesis-molecular-biology/geneius/

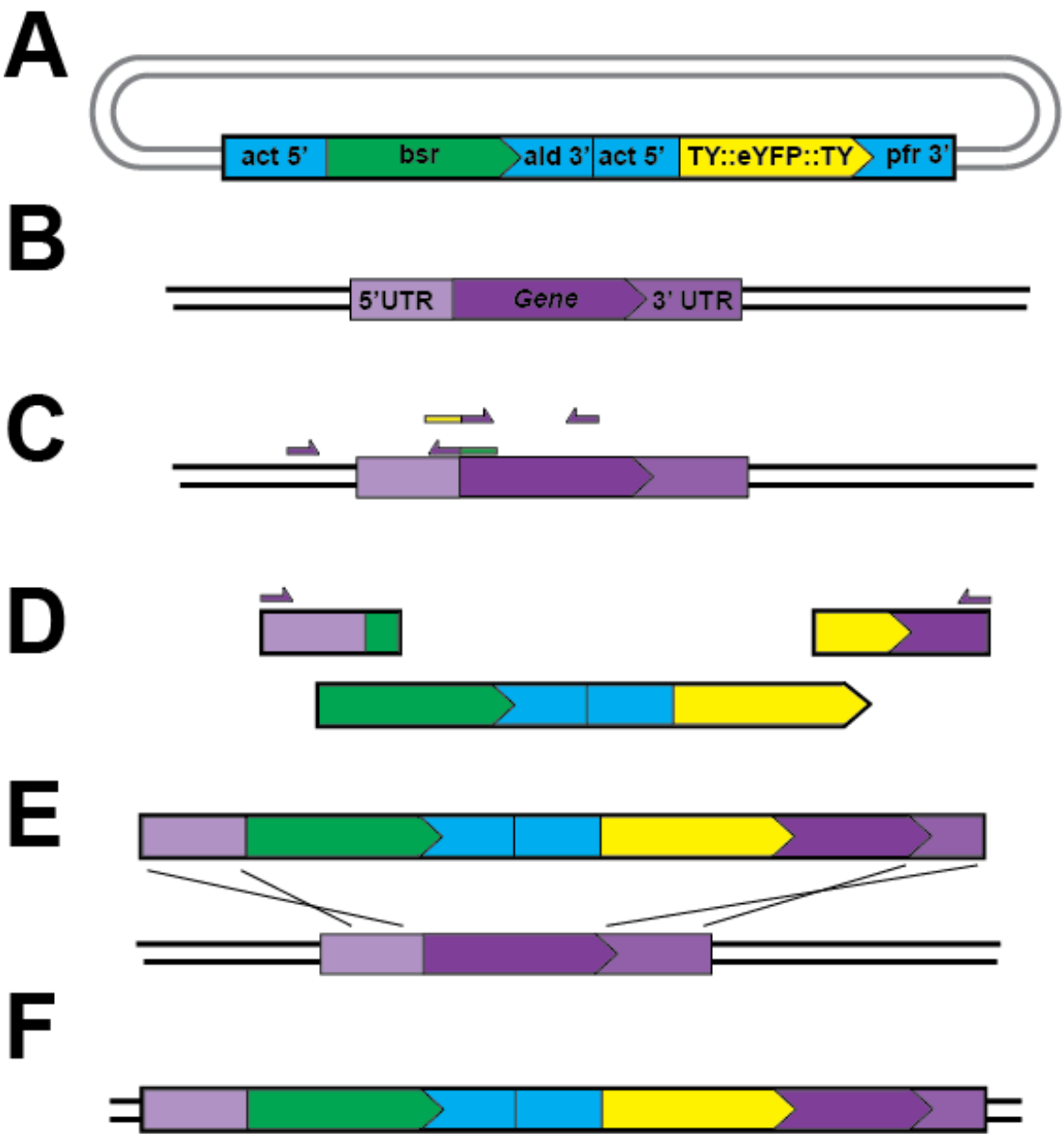



***Figure 2.1 A schematic diagram of the nested PCR tagging approach***

**A)** A map of the pPOTv2 vector used for the N-terminal tagging of target proteins with yellow fluorescent protein. **B)** The gene of interest. **C)** The first round of PCR to amplify 500 bp of the 5′ UTR with a 20 bp overlap to the blasticidin resistance gene and 500 bp of the 5′ end of the gene of interest with a 20 bp overlap with the eYFP. **D)** The second PCR with nested primers to join 500 bp fragments with the tagging cassette. **E)** Integration into the endogenous locus. **F)** The resulting DNA encoding an eYFP fusion protein.

### *2.6.4 Long Primer PCR Tagging*

Long primer PCR tagging using oligos ≥100 bases was performed according to the method developed by Dean *et al.*, 2015. 140 nucleotide primers were designed in ApE and synthesised by Sigma-Aldrich and are listed in Appendix 2. PCR was carried out as below with Phusion polymerase (Thermo Scientific). Reactions were assembled on ice and frozen at -80˚C before addition of polymerase total reaction consisted of 1x GC buffer, 200 µM each dNTP, 0.5 µM of each primer, 1 unit enzyme and <250 ng circular pJ1258 plasmid, DMSO was used as a PCR additive at 2%.

| | | | |
|---|---|---|---|
| Denaturation | 95˚C | 3 minutes | |
| Denaturation | 94 ˚C | 15 seconds | ┐ |
| Annealing | 60 ˚C | 30 seconds | ├ x 35 cycles |
| Extension | 72 ˚C | 3 minutes | ┘ |
| Extension | 72 ˚C | 5 minutes | |

PCR was validated by gel electrophoresis on 1% agarose gels (tanks and power packs from BioRad) and visualised using 0.1% ethidium bromide and imaged at 365nm using a Biospectrum Image System with Vision Works software (UVP).

### *2.6.5 DNA Sequencing*

Sequencing was carried out using the BigDye terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) as described in the manufacturer's guidelines. Following the sequencing reaction the DNA was precipitated using 2.5x volumes of ethanol and 0.08M NaOAc before being analysed using an ABI Prism 3100 capillary sequencer (Applied Biosystems).

### *2.6.6 Bacterial Transformation*

Competent cells were produced using the Inoue Method (Sambrook and Russell, 2006). Restriction digests were carried out according to New England Biolabs instruction. Ligations were carried using 3:1 molar ratio insert:backbone according to enzyme manufacturer's instruction and left at 4ºC overnight.

DNA constructs containing selectable resistance markers were transformed into *E.coli* XL1-Blue (Agilent Technologies) using heat shock. Constructs were added to recently thawed heat-shock competent cells on ice. The cells were then incubated at 42˚C for 45 seconds and returned to ice for 2 minutes to recover. Cells were then added to plates of LB medium containing 100µg $ml^{-1}$ ampicillin (Sigma) and incubated overnight at 37˚C. Individual colonies were isolated and grown overnight in 1.5 ml LB containing 100 µg $ml^{-1}$ ampicillin. Plasmid purification was carried out using a QIAprep spin miniprep kit (Qiagen) according to the manufacturer's instructions.

### *2.6.7 Transfection of Phytomonas for nested PCR product*

Transfection was carried out according to protocols published by the laboratory of Professor G. Cross and may be obtained from www.tryps.rockerfeller.edu. Prior to transfection, 10 µg of linearised DNA was precipitated in ethanol and re-suspended in cytomix (2 mM EGTA pH 7.6, 120 mM KCl, 0.15 mM $CaCl_2$, 10 mM $K_2HPO_4/KH_2PO_4$ pH 7.6, 2.5 mM HEPES pH7.6, 5 mM $MgCl_2.6H_2O$, 0.5% glucose, 100ug/ml BSA, 1 mM Hypoxanthine, filter sterilised). $1x10^8$ cells were spun at 800 *g* for 5 minutes, the supernatant was discarded and the cell pellet was washed with cytomix. Cells were re-suspended in cytomix and mixed with DNA to a total volume of no more than 500 µl. Cells were transferred to a 0.4 mm electroporation cuvette and transfected using a BTX electroporator (1600 V for 100 µs, 2 pulses with a 200 ms interval). Cells were allowed to recover for 12 hours before addition of blasticidin (Sigma). Correct plasmid integration was confirmed by standard GoTaq PCR (Promega) using a combination of primers flanking the integration site and primers located within the linearised plasmid that was integrated.

### *2.6.8 Transfection of* **Phytomonas serpens** *for long primer PCR product*

50 µl PCR product was precipitated in ethanol and resuspended in cytomix. $1x10^7$ cells were spun at 800 *g* for 5 minutes, the supernatant was discarded and the cell pellet was washed with cytomix. Cells were re-suspended in cytomix and then mixed with DNA to

a total volume of no more than 500 µl. Cells were transferred to a 0.4 mm electroporation cuvette and transfected using BioRad Gene Pulser at 1.5 kV. Cells were allowed to recover for 12 hours before addition of blasticidin (Sigma).

### *2.6.9 Cloning of cell lines*

Cells were counted using a CASY counter (Roche) and diluted to <0.5 cells $ml^{-1}$ in Brain Heart Infusion medium (Sigma) and plated out in 96 well plates. Clones were obtained after ~2 weeks.

### *2.6.10 Validation of YFP expression and localisation by microscopy*

#### *2.6.10.1 Native YFP fluorescence*

Cells containing YFP-tagged proteins and control wild-type cells were fixed in 4% ice cold PFA, washed in 1x PBS and stained with Hoechst 33342 (Thermo Fisher) (10mg/ml stock diluted to 1:1000 working concentration) for 5 minutes, washed in PBS and settled onto poly-L-lysine slides for 20 minutes. Slides were mounted with FluroMount (Sigma) and sealed with nail varnish. Native YFP fluorescence was imaged using a Leica DM5500B microscope described in section 2.8.

#### *2.6.10.2 Immunofluorescence*

Cells were fixed in 4% ice cold PFA and washed in 1x PBS. Cells were permeabilised in -20˚C methanol for 5 minutes and washed with 1x PBS. The slides were blocked for 5 minutes at room temperature in TB buffer (1xPBS with 0.1% Triton X100 and 0.1% bovine serum albumen (Sigma)). BB2 primary antibody (Thermo Fisher) (diluted 1:10) was added to the slide and incubated for 1 hour at room temperature. Slides were washed 3 times with TB and Tetramethylrhodamine (TRITC) (Sigma), conjugated secondary antibody (diluted 1:200) was added to the slides and incubated for 1 hour in the dark at room temperature. Slides were stained with Hoechst 33342, mounted and imaged as described in 2.6.10.1.

### *2.6.11 Validation of YFP expression by western blot*

Cells were grown to 1x$10^7$ cells/ml, spun, washed, resuspended in laemmli buffer and boiled for 5 min. Whole cell lysates were run on SDS PAGE and transferred onto nitrocellulose as in Section 2.5.9. Membranes were blocked in 5% fish gelatine (Sigma) in TBST overnight at 4˚C with continuous shaking. For immunodetection of YFP tagged proteins primary antibody BB2 serum was kindly provided by the laboratory of Professor Keith Gull (Bastin *et al.*, 1996). Membranes were incubated with primary antibodies at room temperature for 1 hour. Primary antibody serum used was BB2 at a 1:20 dilution in TBST. The BB2 antibody is a monoclonal mouse IgG raised against the TY epitope. Membranes were washed with 1 L of TBST and probed with secondary antibody: anti-mouse IgG HRP (Abcam) at 1:20000 dilution for 1 hour at room temperature. 1 ml of enhanced luminol reagent and 1 ml oxidising reagent (both PerkinElmer) were added to the membrane and incubated with gentle agitation for 1 minute. The membrane was imaged on ChemiDoc MP (BioRad).

### *2.6.12 DNA extraction from* **Phytomonas serpens**

Template DNA for PCR was extracted from *P. serpens* using a DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer's instructions. In each case DNA was extracted from 1x$10^8$ cells.

## 2.7 Infection of tomato fruit

Four tomato (*Solanum lycopersicon)* types were used: Cherry (Morocco), Beefsteak (Morocco), Plum (Netherlands) and Vine (Spain). The surface of each tomato was sterilised using 70% ethanol. Inoculation was at the base of the fruit at the scar left by the tomato flower. *P. serpens* cell cultures grown in Brain Heart Infusion Broth were inoculated into the tomato flesh at a depth of around 1cm. Tomatoes were inoculated by 10 ml syringe using a 1.1 x 50 mm needle. *P. serpens* cells expressing eYFP were injected at a density of 1x$10^7$ cells $ml^{-1}$. The following volumes were used: 0.5 ml Cherry, 1 ml Vine and Plum, 1.5 ml Beefsteak. Tomatoes were then placed in sealed plastic

boxes at 28˚C for 7 days. To re-isolate *P. serpens* skin of the fruit was pierced using a scalpel and the sap was collected on a glass slide. The sap was stained with Hoechst 33342 and examined by microscopy as in 2.8.

## 2.8 Microscopy

Unless stated otherwise cells were fixed in 4% ice cold PFA, washed in 1x PBS and settled onto poly-L-lysine slides for 20 minutes. Cells were permeabilised in -20˚C methanol for 5 minutes and washed with 1x PBS. To detect DNA, slides were stained with Hoechst 33342 (10mg/ml stock diluted 1:1000 for working concentration) for 1 minute. Slides were then washed 3 times in 1x PBS, mounted with FluroMount and sealed with nail varnish.

Images were captured with Andor Neo sCMOS Camera Model DC-152Q-C00-FI using MicroManager version 1.4.15 (Edelstein *et al.*, 2014). Raw images were subject to manual processing in ImageJ (Collins, 2007). Using the tools Image> Adjust > Brightness/Contrast the minimum and maximum pixel intensities were adjusted so that the background of the slide was black and that the brightest pixels were not saturated. This was to provide an image that was qualitatively representative of what was observed down the microscope.

Objectives lenses used were: 100x 1.4 oil immersion, 63x 1.4 oil immersion, 40x 1.2 oil immersion, 40x long working distance (not oil immersion)

Filter cubes used were:

| Filter cube | Excitation range | Excitation Filter | Dichromatic Mirror | Suppression Filter. BP = bandpass filter |
|---|---|---|---|---|
| A4 (blue) | UV | BP 360/40 | 400 | BP 470/40 |
| L5 (green) | Blue | BP 480/40 | 505 | BP 527/30 |
| YFP | Blue | BP 500/20 | 515 | BP 535/30 |
| Y5 | Red | BP 620/60 | 660 | BP 700/75 |
| TX2 | Green | BP 560/40 | 595 | 595 |

# *Chapter 3: Proteomic and bioinformatic investigation into the* **Phytomonas serpens** *cell surface*

## 3.1 Preface

In *Trypanosoma* and *Leishmania* the parasite cell surface provides a diverse array of functions including endocytosis and protection from host immune responses. Despite these shared functions, it is thought that the protein composition of the cell surface of different species is largely unique with only a small number of conserved surface proteins found in multiple species (Jackson *et al.*, 2013). However, these observations were made following the analysis of a small number of species from *Trypanosoma* and *Leishmania* and it is not known whether these observations hold true across the wider trypanosomatid group. *Phytomonas* are a sister group to *Leishmania* and thus information concerning the cell surface of *Phytomonas* has the potential to improve our understanding of the evolution of the parasite-host interface in *Leishmania* and reveal how *Phytomonas* have adapted their common ancestral cell surface to enable colonisation of plants. To provide new insight into these areas this chapter presents a proteomic analysis of the cell surface of the plant parasite *Phytomonas serpens* and a comparative bioinformatic analysis of cell surface proteins in multiple *Phytomonas*, *Trypanosoma* and *Leishmania* species.

Following the prediction of the cell surface proteome for four *Phytomonas* species the comparative bioinformatic analysis of these proteins across trypanosomatids will indicate which are conserved across the species studied and which are clade specific. It is predicted that proteins required for cell form and function are likely to be conserved across all genera e.g. structural proteins and transporters, and these will form an invariant surface proteome. Proteins that are not conserved may be adaptive for specific hosts and environments such as Variant Surface Glycogen (VSG) and mucins. These latter proteins form part of the cell surface that is not conserved between species and

will form a variant surface proteome. By looking at the surface proteins conserved across the trypanosomatids it is possible to hypothesise the composition of the cell surface of the ancestral trypanosomatid. However, this does not take into account that different invariant proteins may have evolved different localisation patterns following speciation and would exclude for genes inherited from the last common ancestor but lost from one genus.

## 3.2 Introduction: methods for studying cell surface proteins

The study of cell surface proteins presents many challenges. The first is stoichiometric: cell surface proteins can represent a small percentage of total cell protein, which can impede their detection in whole cell lysates by tandem mass spectrometry (MS/MS). Secondly, some cell surface proteins have properties that make them difficult to manipulate, for example transporter proteins with multiple transmembrane helices are hydrophobic and can aggregate and precipitate out of solution following the solubilisation of whole cells (Wu and Yates, 2003).

As cell surface proteins are important for many cellular processes there have been various methods developed to attempt to specifically isolate them from whole cell material. Some methods attempt to directly release proteins from the surface of the cell. This can be done in a non-specific manner through proteolytic treatment e.g. trypsinisation or in a more specific manner e.g. using GPI-PLC. The former releases peptides from exposed proteins which can then be identified by mass spectrometry (Solis *et al.*, 2010). The latter involves cleaving the GPI anchor to release GPI anchored proteins which can later be treated with a protease prior to mass spectrometry (Elortza et al., 2003). Like *Leishmania*, *Phytomonas* lack a GPI-PLC. Treatment with GPI-PLC may only recognise a subset of GPI anchors (e.g. some modifications such as acylation at certain positions of the GPI anchor can prevent cleavage (Ferguson, 1992) and will not release proteins present on the plasma membrane that are not GPI anchored such as transmembrane proteins. Detergent extraction is another method to isolate cell

surface proteins (Filho *et al.*, 2001) via solubilisation of lipid membranes. A number of different detergents with subtly different effects have been employed in studies of cell membrane proteins. Membrane proteins are readily soluble in detergents because the detergent micelles recreate the environment of the lipid bilayer (Seddon *et al.*, 2004). Detergent extraction is often combined with another method to improve the enrichment of the cell surface membrane above the cell background, for example; freeze thaw prior to detergent extraction (Yao *et al.*, 2010), purification of different membranes by a sucrose density gradient or application of a molecular label to the surface of cells prior to solubilisation. Each method suffers from varying degrees of contamination from non-cell surface proteins. Given the success and precedent of biochemical techniques, in particular the combining of protein labelling followed by detergent extraction, we decided to pursue this approach for the isolation of cell surface proteins in *Phytomonas*. Example labels include fluorescein (Gadelha *et al.*, 2015) and biotin (Yao *et al.*, 2010). Purification of labelled proteins is achieved using antibodies or avidin/streptavidin immobilised on a resin. Biotin was chosen for the labelling of the *P. serpens* cell surface because it has been previously used for the identification of plasma membrane proteins in *Leishmania* (Yao *et al.*, 2010) and is easily purified by utilising the strong interaction with avidin and streptavidin.

The interaction between avidin and biotin is extremely stable with a low dissociation constant ($K_d$). The $K_d$ between two molecules, A and B, is the concentration of free A at which half of the total molecules of B are associated with A. The smaller the $K_d$ the tighter the two molecules are bound together. The $K_d$ for biotin and streptavidin is ≈0.000001nM (Green, 1975) which is around 10,000 times stronger than the majority of commercial pharmaceuticals ($K_d$ of 0.1-10nM). The sub picomolar binding affinity observed between biotin and avidin is rare, and has only been observed for a few interactions such as for some ribonuclease inhibitors (Johnson *et al.*, 2007). This strong interaction between biotin and avidin has been exploited in the study of proteins as biotin is easily modified

for attachment to proteins without disrupting its high affinity for avidin. As a consequence of this, a number of reagents have been developed to covalently bind biotin to proteins in different ways (Elia, 2008).

A biotinylation reagent used for labelling proteins is made up of three building blocks: the biotin moiety itself; a spacer arm; and the reactive moiety that facilitates the covalent attachment of biotin to target proteins. The biotin molecule can be modified at its terminal carboxylic acid group without disrupting the binding to avidin (Elia, 2008). This allows for the incorporation of reactive groups at this point to attach the biotin to other molecules. A carbon skeleton spacer arm connects the biotin to a reactive moiety. This spacer can affect the solubility properties of the label and can have a significant role in reducing steric hindrance increasing the ability to recover labeled material (Elia, 2008). For example, biotinyl-ε-aminocaproic-N-hydroxysuccinimide (NHS long chain (LC) biotin) has a long carbon chain that helps to overcome steric hindrance and improves the availability of the biotin residue to avidin/streptavidin. Similarly, *N*-hydroxysulfosuccinimide (Sulfo-NHS-LC-biotin) is a highly water-soluble analog which also has a long spacer arm (~22.4 Å) that can be further extended by addition of aliphatic aminocaproid acid chains. The reactive moiety positioned at the end of the spacer arm covalently attaches the entire molecule to a protein.

Reactive moieties are positioned at the terminal end of these space arms. The most widely used reactive groups are reactive esters, such as *N*-hydroxysuccinimide which undergoes nucleophilic substitution in the presence of primary amines (e.g. the ε-amino group in lysine residues) and to a lesser extent the secondary amines (e.g. glucosamine in the GPI core). The biotinyl-NHS ester (NHS-biotin) is widely used in the biotinylation of proteins, in part because lysine residues are numerous in most proteins and often occupy accessible surface positions in the protein structure (Elia, 2008). Other secondary amines e.g. in the GPI anchor will react but are not identified by the mass spectrometry methods used in this study. An additional drawback of these reactive moieties is that when biotinylation is followed by tryptic digests (such as in preparation

for mass spectrometry), the addition of the NHS-biotin will inhibit cleavage at that biotin modified lysine residue. If labelling results in missed cleavages, then the peptide may become too large for identification by mass spectrometry or increase the error rate of peptide identification. If reaction with glucosamine in GPI or too many missed cleavages had limited the detection of peptides by tandem mass spectrometry then alternative labels could be pursued. The presence of LPG is another factor that compromise a biotin labelling approach. This complication arises due to the negative charge of LPG that may repel the sulphated NHS of the biotinylation reagent. This is a potential limitation of this labelling approach.

As discussed, the interaction between biotin and avidin/streptavidin is strong and stable. While this has many benefits, it also poses a challenge to disrupt this interaction to facilitate release of the protein after capture (i.e. the elution of labelled material from immobilised avidin resin). The removal or elution of the biotinylated material from the avidin/streptavidin conjugated resin requires very harsh conditions, for example boiling in reducing reagents with detergent. These elution conditions limit the utility of the material following elution (i.e. the function of the material can be destroyed by elution) and can cause selective aggregation of proteins (Gallagher, 2006). To overcome this problem, easily cleavable spacer arms have been developed. The most commonly used of these is a spacer arm that is held together by a single disulphide bridge, which can be broken under reducing conditions. The further addition of a charged group to the spacer arm has an additional advantage for the analysis of surface proteins as it makes the entire molecule membrane impermeable. This prevents mis labelling of cytoplasmic proteins. Moreover, in the event that the label enters the cell, reducing conditions found in the cytosol would cleave the disulphide bridge, separating the biotin from the NHS-ester thus further preventing contamination from cytoplasmic proteins (Elia, 2008). A schematic of cell surface biotinylation is shown in Figure 3.1.

The full extent of existing knowledge concerning the *Phytomonas* cell surface and its similarities to that of *Leishmania* was summarised in the introduction chapter of this

thesis. This chapter presents the development of a *Phytomonas* surface protein isolation method and the characterisation of a cell surface proteome for this organism. It also includes a comparative bioinformatic investigation into the cell surface proteins of multiple *Phytomonas* species including the published genome sequences of *P. EM1* and *P. HART1,* and the genome of an additional species, *P. françai* that was sequenced specifically to provide additional data for this study.

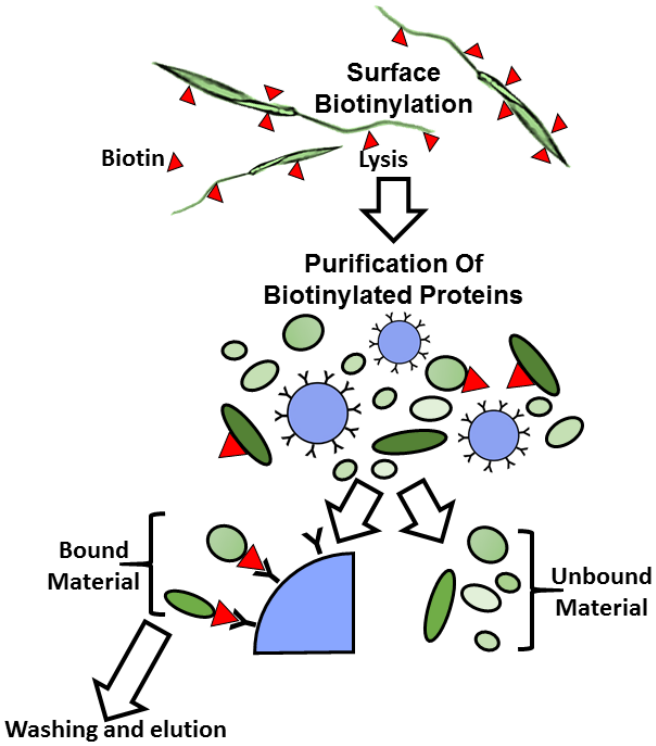



***Figure 3.1*** **The workflow of the cell surface biotinylation protocol**

Whole *P. serpens* cells are surface labelled with biotin (red triangles). Surface biotinylated cells are lysed. Fractions of the lysate from biotinylated whole cells are incubated on streptavidin resin (blue). The beads are separated from solution. What remains in solution is the unbound material. The bound material is then washed several times prior to elution. Isolated proteins are identified by tandem mass spectrometry.

## 3.3 Results

### *3.3.1 Criteria used for optimisation of cell surface labelling*

The aim of this chapter was to isolate and identify the cell surface proteins of the tomato fruit parasite *P. serpens*. To do this a biotin-based protocol for the isolation and characterisation of cell surface proteins in a related parasite (Yao *et al.*, 2010) was adapted and optimised for use with *P. serpens*. During the optimisation process several objective functions were applied. These were to 1) minimise perturbation to cells; 2) minimise contamination (defined below); 3) maximise retrieval and identification of labelled material in a non-biased manner. The success of these objective functions was assessed using multiple different techniques including 1) fluorescence microscopy to visualise the distribution of biotinylated material in labelled cells. 2) blotting and detection of biotin labelled material using enhanced chemiluminescence (ECL) to visualise the distribution of biotinylated material in different fractions i.e. detergent soluble vs insoluble. 3) liquid chromatography and tandem mass spectrometry (LC MS/MS) to elucidate the identity of the proteins in different fractions and their relative abundances. Due to the solubility of cell membranes in detergent it was expected that more biotinylated material would be detected in detergent soluble vs insoluble material and also more in the material released from the beads vs the whole cell lysate. The order in which different stages of the protocol were optimised is described in Figure 3.2.

While success was achieved in several steps, the major challenge encountered during the optimisation of the method was the reduction of contamination. Contamination is defined in this context as the presence of proteins in the final eluted material that are definitively not cell surface proteins (e.g. tubulin, ribosomal proteins, nucleolar proteins etc.). Such non-surface proteins can become contaminants in several ways. For example: 1) by direct biotinylation of material from lysed dead cells. 2) By the uptake of unreacted biotin by live cells resulting in the biotinylation of cytosolic proteins. 3) By sticking non-specifically to the streptavidin conjugated resin during wash steps but

detaching during the harsher elution conditions. 4) By sticking to biotinylated material that is attached to beads.

From this point forward I define the eluate as the set of proteins that are isolated by the biotin-streptavidin isolation protocol. The negative control is the eluate from an identical experiment performed without the addition of biotinylation reagent. The set of proteins from the whole cell lysate that did not bind to the resin is termed the unbound material. The methods used in this chapter are described in sections 2.4 - 2.5.

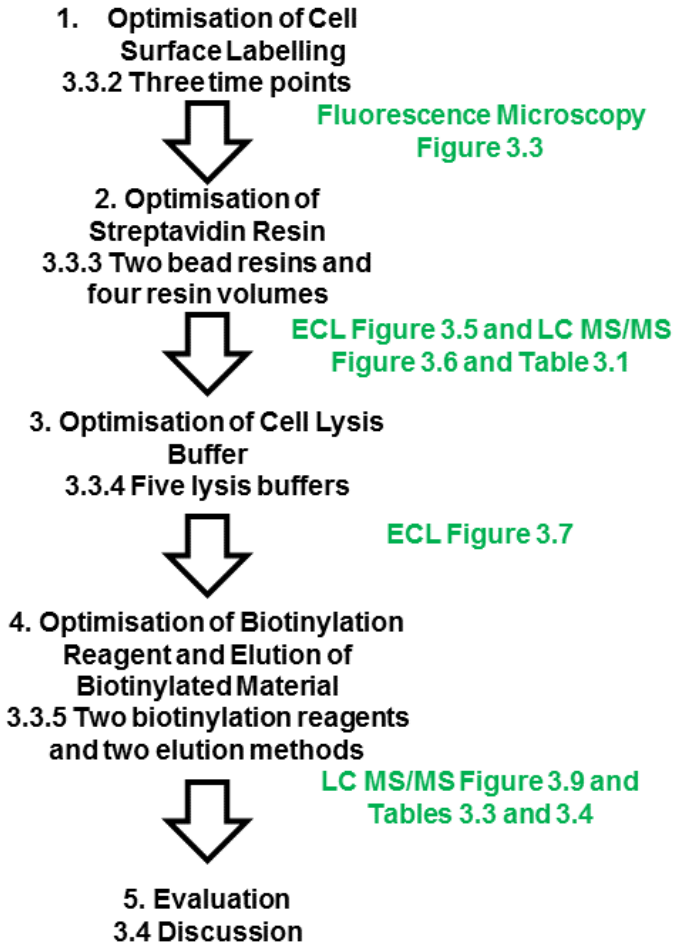



***Figure 3.2*** **The order in which stages of the biotinylation protocol were optimised**

A schematic diagram showing the order in which the steps in a protocol for the isolation of cell surface proteins in *P. serpens* was optimised. The stages of optimisation are shown in black text and the methods used to assess the success of each stage are shown in green text indicating where in this chapter each stage is presented.

### *3.3.2 Optimisation of cell surface labeling*

The protocol used by Yao *et al.,* (2010) (Protocol 1 outlined in Chapter 2 where cells were biotinylated at room temperature, lysed in 2% NP-40 and the detergent soluble fraction incubated on streptavidin agarose beads) provided a starting point for a method that was developed for application to *P. serpens*. In this published work, cells were biotinylated for 1hr at room temperature. This was optimised as contamination can arise from the biotinylation of the proteins in the cell interior e.g. as a result of endocytosis of the biotin label or the redistribution of biotinylated proteins in the cell, or from death and lysis of cells left for periods of time in buffer. This protocol was modified to reduce the perturbation to cells and also reduce the likelihood of biotinylation of non surface proteins. In the protocol developed all cells were kept on ice and biotinylation time was reduced. Figure 3.3 shows that there is some qualitative difference in biotin labelling when biotin is detected using Streptavidin Alexa Fluor 488 (which labels biotin) if cells are biotinylated for 10, 30 or 60 minutes on ice. Figure 3.4 shows a higher magnification image of a single cell from the 10 minute time point. Figure 3.4 E shows that there was significant labelling of the cell surface following the 10 minute incubation time and the linescan shows labelling that is consistent with being at the cell surface, therefore this time was selected to minimise perturbation to the cells. This decision was made as prolonged retention of cells in incubation buffer can lead to cell lysis and potentially to endocytosis of labelling reagent or labelled materials. Lysis of cells or endocytosis of material will all increase the amount of contamination from non-surface proteins.

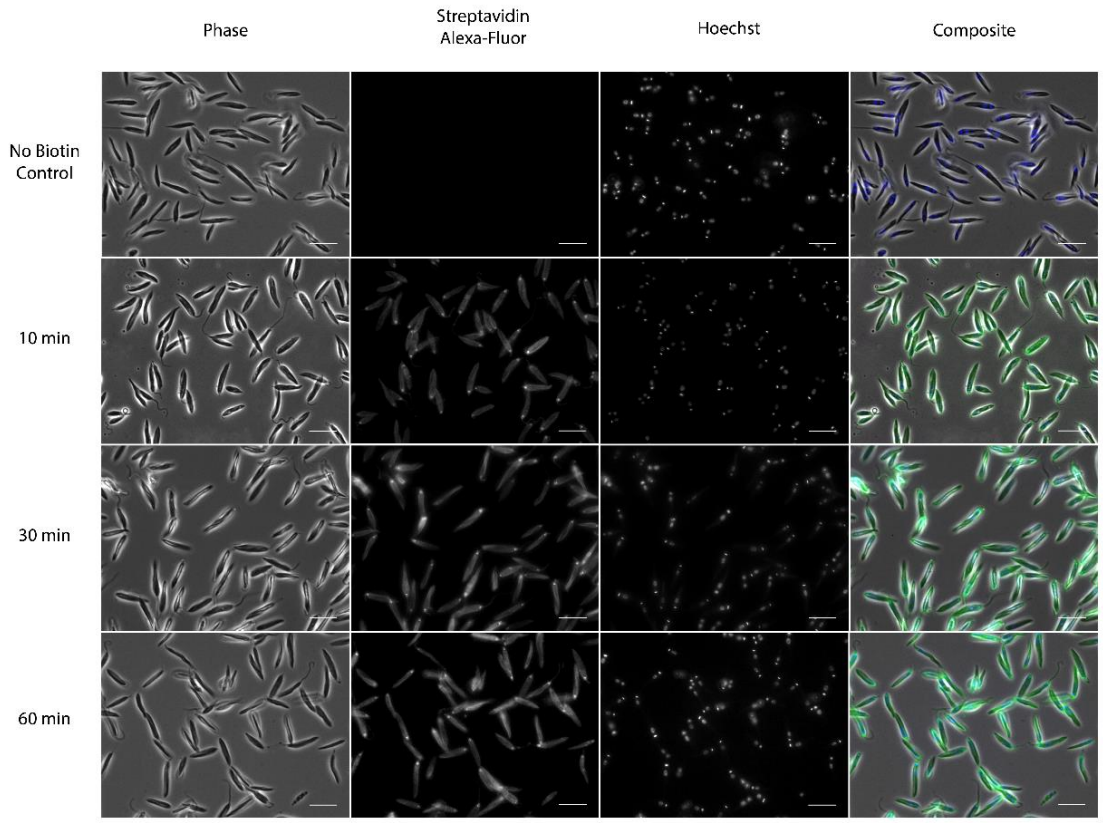



***Figure 3.3*** **Optimisation of cell surface labelling**

Images showing the difference in cell surface labelling of *P. serpens* following incubation with 5 mM Sulfo NHS LC-biotin for 10, 30 or 60 min. Scale bars represent 10 µm. All images taken using the same exposure time in each channel and processed to remove background using identical settings using ImageJ (Collins, 2007).

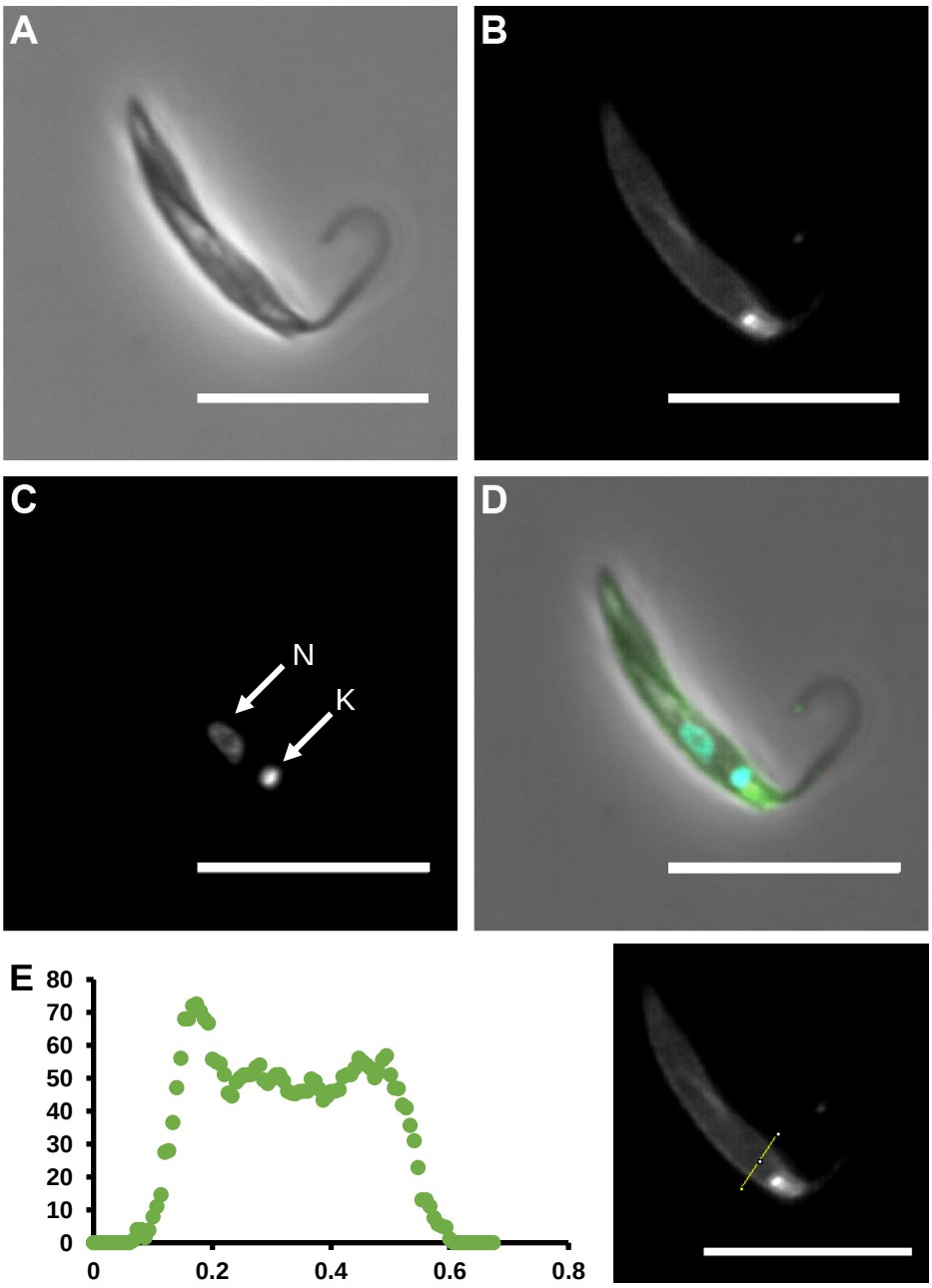



***Figure 3.4*** **Cell surface biotinylation of *P. serpens***

High magnification images of a *P. serpens* cell following incubation with 5 mM Sulfo NHS LC-biotin for 10 minutes (Protocol 1). **A)** Phase contrast **B)** streptavidin Alexa Fluor **C)** nucleus (N) and kinetoplast (K) stained with Hoechst **D)** False colour image representing the merge of all channels with streptavidin Alexa Fluor in green and Hoechst stain in blue. **E)** A pixel intensity plot of the biotin label across a *P. serpens* cell. Scale bars represent 10 µm. Cells were permeablised prior to staining.

### *3.3.3 Optimisation of the streptavidin resin*

The protocol for the isolation of cell surface proteins in *Leishmania* (Yao *et al.*, 2010) was used to perform an analogous isolation of cell surface proteins in *P. serpens*. The protocol was carried out as described but with a single change whereby cells were incubated with the biotinylation reagent on ice for 10 minutes instead of room temperature for 1 hour. The results are shown in Figure 3.5A. Biotinylated material was observed in whole cell lysate (Figure 3.5A, +W). Partitioning this whole cell lysate into detergent soluble (+S) detergent insoluble (+P) via centrifugation resulted in partitioning of biotinylated material into both fractions (Figure 3.5A, +P and +S). The soluble fraction (+S) was incubated on streptavidin-agarose beads. The beads were then washed in lysis buffer, and the bound material was eluted by boiling in Laemmli buffer as described in the published method (Yao *et al.*, 2010). There was little detectable biotinylated material in the eluate (Figure 3.5A, +B) and the biotinylated material was not present in the flow through (+F). Thus the biotinylated material was depleted by the addition of the streptavidin beads but was not eluted by boiling in Laemmli buffer.

Despite several attempts biotinylated proteins were not recovered following incubation with streptavidin agarose beads. This may have been due to the properties of the agarose beads which are both porous and anionic. There was clearly biotinylated material in the soluble fraction (+S) but none in the elution (+B) or flow through (+F). An alternative widely used resin is streptavidin conjugated magnetic beads, (Dynabeads) which are not anionic and not porous. Figure 3.5B shows exactly the same protocol as for Figure 3.5A but with a single modification of using the Dynabeads.

Use of Dynabeads successfully depleted biotinylated material from the detergent soluble fraction such that none was observed in the unbound material (Figure 3.5B, +F). Given that few bands were detected in the eluate by SYPRO-ruby staining (Figure 3.5D, +B), and that many non-biotinylated proteins were detected in the unbound material, this indicates that selective purification of the biotinylated material had occurred (Figure 3.5D,

+F). Moreover, there were fewer bands detected following elution from Dynabeads (Figure 3.5D, +B) than from the equivalent elution from agarose beads (Figure 3.5C, +B). Therefore as there was less non-specific binding to the Dynabeads and there was improved recovery of biotinylated material, Dynabeads were used for all subsequent analyses.

To determine the optimal number of cells-per-bead for use in subsequent analyses soluble fractions from different quantities of cells were incubated with a fixed number of beads (Figure 3.5E). To minimise the relative quantity of non-specific binding, it is important to use a concentration of beads that does not completely deplete the sample of biotinylated material. This is because non-specific binding is proportional to bead volume and thus use of excess beads causes more contaminants to be recovered without increasing recovery of biotinylated material. A ratio of $1x10^7$ cells $bead^{-1}$ was selected as a ratio at which the maximum amount of biotinylated material was recovered without fully depleting the biotinylated material from the detergent soluble fraction (Figure 3.5E).

It was not possible to detect biotinylated material eluted from the streptavidin agarose beads (+B lane Figure 3.5 A). The absence of biotinylated material from the flow through is also unexplained (+FT lane Figure 3.5 A). The blot shows that a large amount of biotinylated material is lost in the detergent insoluble fraction (+P lane Figure 3.5 A), however biotinylated material is present in the detergent soluble fraction (+S lane Figure 3.5 A) and therefore the beads are being incubated with biotinylated material. I can only conclude that despite boiling in Laemmli buffer the agarose beads did not release the biotinylated material, possibly because of their porous nature. I did not pursue this further as the streptavidin Dynabeads released biotinylated material, displayed low levels of nonspecific binding and were easier to handle than agarose beads (washes by retaining beads in Eppendorf with magnet and replacing supernatant instead of using centrifugation to separate the beads and supernatant).

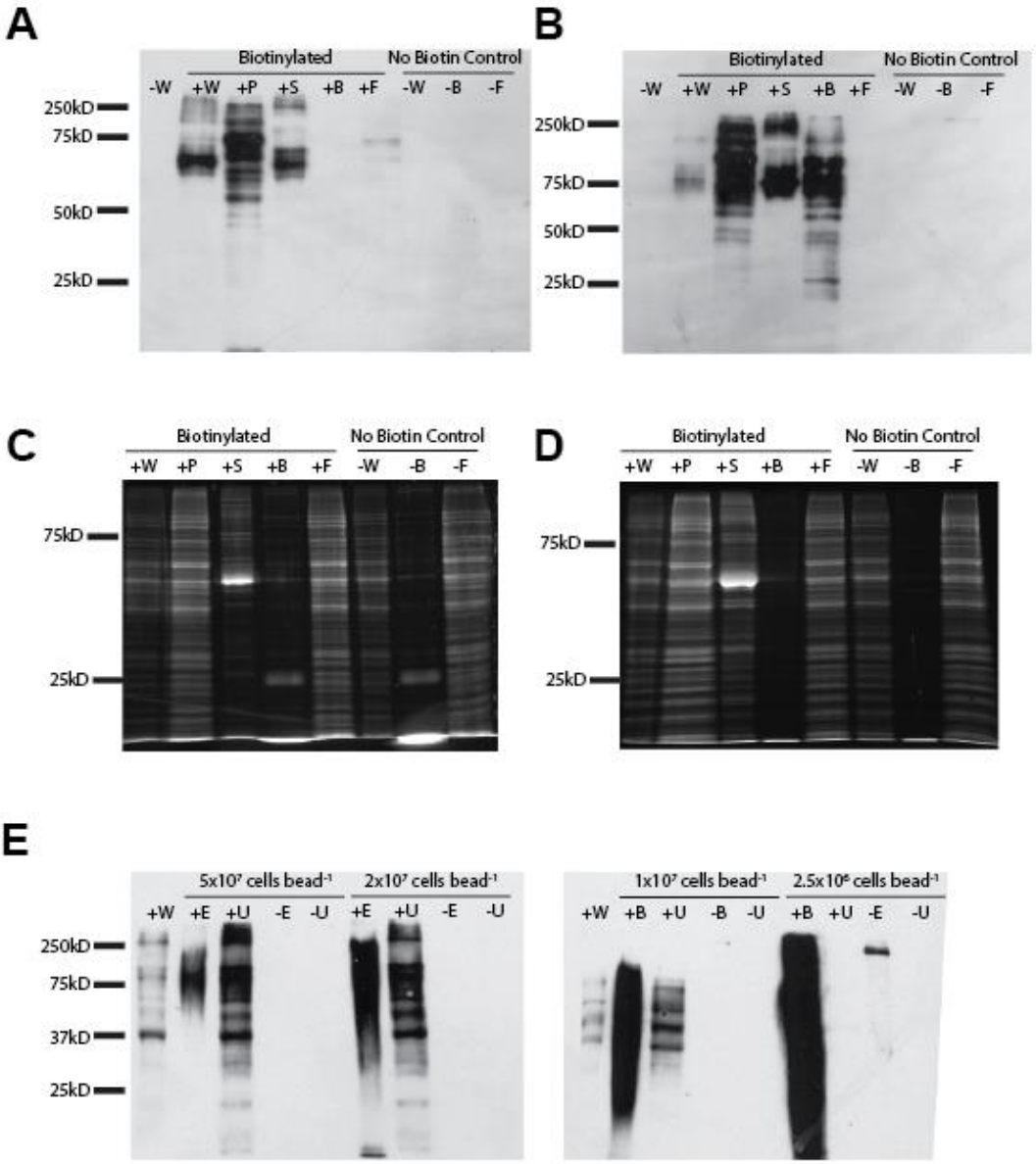

A
Biotinylated
No Biotin Control
-W +W +P +S +B +F -W -B -F
250kD
75kD
50kD
25kD
B
Biotinylated
No Biotin Control
-W +W +P +S +B +F -W -B -F
250kD
75kD
50kD
25kD
C
Biotinylated
No Biotin Control
+W +P +S +B +F -W -B -F
75kD
25kD
D
Biotinylated
No Biotin Control
+W +P +S +B +F -W -B -F
75kD
25kD
E
5x10⁷ cells bead⁻¹
2x10⁷ cells bead⁻¹
+W +E +U -E -U +E +U -E -U
250kD
75kD
37kD
25kD
1x10⁷ cells bead⁻¹
2.5x10⁶ cells bead⁻¹
+W +B +U -B -U +B +U -E -U

***Figure 3.5*** **Optimisation of streptavidin resin**

Comparison of different streptavidin-conjugated substrates for use in pull down experiments. Cell equivalents loaded are $5x10^5$ for whole cell lysates and $1x10^6$ for detergent soluble and insoluble fractions. Proteins were transferred to nitrocellulose membrane and probed with streptavidin conjugated horseradish peroxidase to detect the presence of biotinylated material in each fraction. Two different avidin conjugated resins were used **A)** Streptavidin agarose beads **B)** Streptavidin Dyna beads. Equivalent SYPRO Ruby stained SDS-PAGE gels corresponding to A and B are shown in **C)** and **D)** respectively. Key: -W = whole cell lysate from cells that were not surface biotinylated, +W = whole cell lysate from surface biotinylated cells, +P = detergent insoluble fraction from surface biotinylated cells, +S = detergent soluble fraction from surface biotinylated cells, +B = eluate following bead incubation in detergent soluble fraction from surface biotinylated cells, +F = unbound material following bead incubation in detergent soluble fraction from surface biotinylated cells, -B = bead elute (negative control) as above except from incubation in the detergent soluble fraction from cells that were not surface biotinylated, -F = unbound material as above except from incubation in detergent soluble fraction from cells that were not surface biotinylated, **E)** Titrating resin volume, calculated as number of cells per bead +W= whole cell control, +E = eluate following bead incubation in whole cell lysate from surface biotinylated cells, +U = unbound material following bead incubation in whole cell lysate from surface biotinylated cells, -E = eluate following bead incubation in whole cell lysate from cells that were not biotinylated, -U = unbound material following bead incubation in whole cell lysate from cells that were not biotinylated.

From the gel based analysis above it was evident that biotinylated material was being enriched in the eluate using this protocol. To determine the identity of the proteins contained in the eluate, and assess the success of the procedure, the samples corresponding to Figure 3.5B, +B, -B and +F were prepared as described in 2.5.5.1 and subjected to tandem mass spectrometry (LC MS/MS) as described in 2.4.11. The results from this are summarised in Figure 3.6 and Table 3.1.

In total 50797 spectra were identified across all samples. Data searches against the predicted *P. serpens* proteome matched to 60.43% of spectra, including 12620 unique peptides. Data searches identified a total of 1285 proteins, of which 261 were found in the elution fraction. Figure 3.6A shows a comparison of the estimated abundance of each protein in the elution and the unbound material. Here, a normal distribution centred on 0 would indicate that the proteins in the elution were just a random sample of the proteins in the whole cell lysate with no selective enrichment. That is proteins in the eluate are found at their expected abundance +/- some quantity of normally distributed sampling error. However, the data show a clear shift in location of the data to the right with an extended right-handed tail (Figure 3.6A). To determine whether this represented a significant deviation from expected sampling bias, a Shapiro-Wilk test for normality revealed that ratios of observed abundances in the eluate compared to the unbound material were not normally distributed. This analysis indicated that there was a significant difference in the observed ratios ($p < 0.01$) and thus that there is both a population of proteins present in the eluate at expected abundances (contamination) and a population of proteins present in the eluate with abundances that are much higher than expected (enriched proteins).

To determine whether the set of isolated proteins was enriched for putative surface proteins a statistical analysis was used to look at the frequency of occurrence of proteins containing the membrane associated domains in the set of proteins that were identified by mass spectrometry. These domains comprised the signal peptide, GPI anchor and

transmembrane helices. The presence of a transmembrane helix and a signal peptide does not guarantee that a protein will localise to the cell surface membrane. These features could also target a protein to a cell internal membrane such as the endoplasmic reticulum, secretory vesicle, or Golgi apparatus. An additional limitation is that computational methods to predict GPI anchors, signal peptides and transmembrane domains are not perfect and can result in both false positive errors (i.e. prediction of features that are absent) and false negative errors (i.e. fail to detect *bona fide* features). The hypergeometric test calculates the probability of obtaining a number of successes given a number draws without replacement, from a finite population that contains a known number of success states. In this instance, a success is a protein with a signal peptide, GPI anchor or transmembrane helix. The number of draws are the number of proteins identified, which were enriched above a given fold change in abundance. The population size is the total number of proteins identified by tandem mass spectrometry, and the number of successes that exist in the population is the total number of proteins identified by tandem mass spectrometry with a signal peptide, GPI anchor or transmembrane helix. The low p-values obtained in the hypergeometric test indicate that for high fold change in enrichment it is very unlikely that such a number of putative surface proteins would be observed by chance alone.

The distribution of the number of proteins against fold-change in abundance is shown in Figure 3.6B. If proteins were being enriched at random we would expect to see a normal distribution here. The histogram shows a multimodal distribution. The first mode is centered around 1 and typical of what would be expected from contamination. The second mode centered on 11 (and associated right hand tail) is strongly indicative of specifically enriched proteins. This second distribution would not be expected is the proteins isolated were simply a random sample of cellular proteins. To further probe the significance of the proteins in this second mode and right hand tail I have shown the *p*-values obtained for each membrane protein domain for each fold-change cut-off in Figure

3.6B. The test shows that the set of proteins with fold-changes in abundance >0 (i.e. abundance is higher in the eluate than in the unbound material) contains a higher number of proteins with predicted signal peptides than expected by chance (Figure 3.6B, green line). A similar phenomenon was observed for proteins with transmembrane domains (Figure 3.6B, purple line). Here, the set of proteins with fold-changes in abundance ≥ 8 (i.e. $2^3$ times more abundant in the eluate than in the unbound material) contains a higher number of proteins with transmembrane domains than expected by chance. Statistical significance was not obtained for the set of proteins with an abundance of ≥ 128 (i.e. $2^7$-fold higher in the eluate than in the unbound material), due to the low numbers of proteins detected above this threshold. Few proteins with GPI anchors were detected in this dataset and as a result there was no significant over-representation of proteins with a predicted GPI anchor for any fold-change in abundance. This absence of GPI anchored proteins from the eluate may be caused by the fact that GPI anchored proteins show poor solubility in non-ionic detergent (Mayor and Maxfield, 1995). Therefore they may have been partitioned into the detergent insoluble fraction prior to bead incubation.

Of the 261 proteins detected in the eluate 80 were detected with a change in abundance 2 fold or greater in the eluate than in the unbound material. Proteins were annotated according to the annotation information available for orthologous genes in *L. mexicana.* The negative control was the eluate from an identical experiment performed without the addition of biotinylation reagent to the whole cells prior to cell lysis. The negative control controlled for proteins that were bound by the beads in a manner that is not specific to the labelling of proteins with biotin. Proteins whose abundance was 2-fold higher in the eluate than in the unbound material and contained either A) a Pfam domain specific to surface proteins, or B) a predicted SP, TMD or GPI or C) are annotated as a protein that has been shown to be surface associated in *L. mexicana* are listed in table 3.1.

These 17 proteins include a total of 16 proteins with at least 1 predicted membrane associated domain (TMD, SP and GPI). Of these, 5 are orthologs of probable surface

proteins in *Leishmania* and *Trypanosoma* including a surface antigen-like protein (Pse.12.1070 and 270.1010), and receptor-type adenylate cyclase (Pse.1381.1000 and 556.1030). Another protein, Pse.1.1460 is the ortholog of KHARON1 (LmxM.36.5850), a protein localised to the base of the flagellum in *L. mexicana* and involved In the targeting of the LmxGT1 glucose transporter to the flagellar membrane (Tran *et al.*, 2013). However, the *P. serpens* protein does not contain any predicted membrane associated domains. This could either represent a false positive BLAST hit, a false negative from the domain prediction tools or a problem with the *P. serpens* gene model. Similarly, Pse.270.1010 is predicted to be an ortholog of *L. mexicana* surface antigen-like protein and was only detected in the eluate, however the sequence does not contain any predicted membrane associated domains.

Also included in this dataset are proteins that are likely not located on the *Phytomonas* cell surface. For example, Pse.29.1160 with a TRAP beta PFAM domain which is predicted to contain a signal peptide but with no anchor sequence is likely secreted. Likely contaminants from the dataset in Table 3.1 shown in red include Pse.21.1380 for which available evidence (protein size and PFAM domain) points to this being a heat shock protein and Pse.84.1120 which contains the PFAM domain for mitochondrial pyruvate carriers.

Taken together the analysis of the data shows that while the method produced significant enrichment of putative membrane associated proteins the proteome presented in Table 3.1 is not robust. The dataset also contains cytosolic proteins and no GPI anchored proteins were detected as enriched in the eluate. This method requires further optimisation to reduce contamination and improve the recovery of putative cell surface plasma membrane proteins.

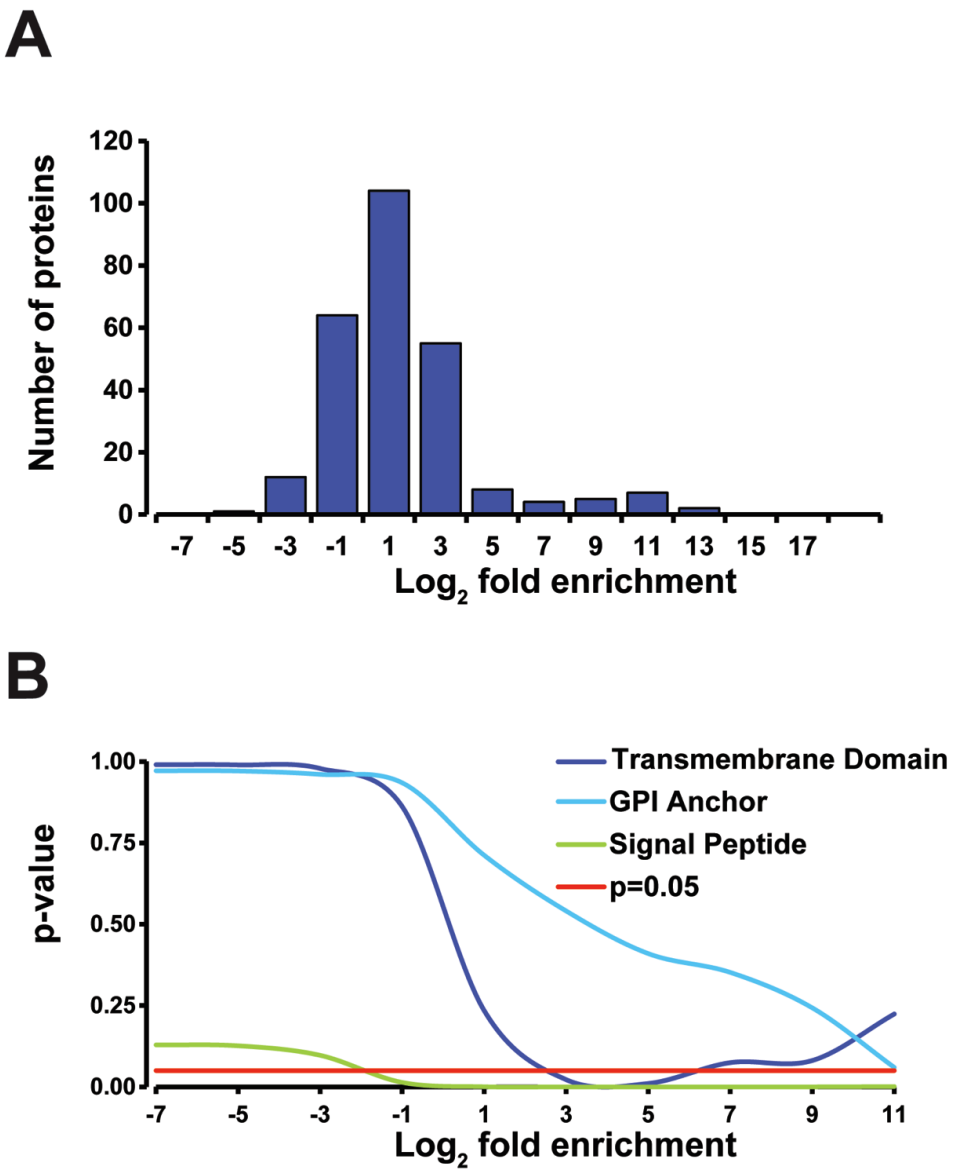



***Figure 3.6*** **Results of preliminary tandem mass spectrometry**

Comparison of the abundance estimates of proteins in the eluate and the unbound material. These figures summarize all data from mass spectrometry. **A)** The distribution of ratios of abundance in the eluate vs the unbound material i.e. $ratio = \frac{eluate}{unbound\ material}$. **B)** The results of hypergeometric tests to determine whether the set of proteins with that fold-change or higher contains more than the expected number of proteins that contain predicted membrane associated domains.

## TABLE 3.1 RESULTS OF PRELIMINARY LC-BIOTIN TANDEM MASS SPECTROMETRY: PUTATIVE CELL SURFACE PROTEINS 2-FOLD MORE ABUNDANT IN THE ELUATE COMPARED TO THE UNBOUND MATERIAL

Proteins are listed in order of decreasing $log_2$(fold enrichment). Putative membrane associated proteins are those either containing either a TM helix, or a signal peptide, or a GPI anchor, or if the annotation of the ortholog in *L. mexicana* or PFAM domain indicate possible interaction with the plasma membrane. The change in abundance in the eluate compared to the unbound material is calculated form the spectral indices for these two fractions. Likely contaminants are indicated in red.

| Protein | *Molecular Weight (Da)* | *PFAM* | *L. Mexicana* annotation | *L. mexicana* ortholog | TM helices | GPI anchor | Signal peptide | Spectral Index Bound | Spectral Index Flow Through | $log_2$ Fold CHANGE | SINQ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Pse.422.1010 | 44018.6 | | | | 0 | 0 | 1 | 1.13E-05 | 4.17E-09 | 11.30 | +B only |
| Pse.604.1040 | 62733.9 | | | | 4 | 0 | 1 | 7.79E-06 | 4.17E-09 | 10.80 | +B only |
| Pse.12.1070 | 37417.3 | | surface antigen-like protein | LmxM.04.0190 | 0 | 0 | 0 | 5.65E-06 | 4.17E-09 | 10.50 | +B only |
| Pse.154.1060 | 79766.8 | | hypothetical protein conserved | LmxM.10.0620 | 2 | 0 | 1 | 4.50E-06 | 4.17E-09 | 9.98 | +B only |
| Pse.270.1010 | 42437.2 | | surface antigen-like protein | LmxM.04.0190 | 0 | 0 | 1 | 3.26E-06 | 4.17E-09 | 9.51 | +B only |
| Pse.556.1030 | 57325.7 | Guanylate_cyc | receptor-type adenylate cyclase putative | LmxM.17.0190 | 1 | 0 | 0 | 1.08E-06 | 4.17E-09 | 7.92 | +B only |
| Pse.50.1060 | 22955.7 | | i/6 autoantigen-like protein | LmxM.22.1460 | 0 | 0 | 0 | 1.38E-05 | 1.19E-07 | 7.04 | 0.01 |
| Pse.179.1020 | 31019.1 | | hypothetical protein conserved | LmxM.27.0070 | 2 | 0 | 1 | 1.51E-06 | 9.66E-08 | 5.99 | 0.06 |
| Pse.422.1040 | 30570.9 | | hypothetical protein conserved | LmxM.26.2590 | 5 | 0 | 0 | 3.61E-06 | 8.99E-08 | 5.52 | 0.02 |
| Pse.777.1000 | 9142.84 | | hypothetical protein conserved | LmxM.30.0900 | 2 | 0 | 1 | 1.25E-05 | 7.57E-07 | 4.24 | 0.06 |
| Pse.1.1460 | 42876.7 | | flagellum targeting protein kharon1 (KH1) | LmxM.36.5850 | 0 | 0 | 0 | 8.55E-07 | 5.48E-08 | 4.16 | 0.06 |
| Pse.29.1160 | 22289.2 | TRAP beta | hypothetical protein conserved | LmxM.31.0840 | 0 | 0 | 1 | 2.22E-06 | 4.68E-07 | 2.44 | 0.21 |
| Pse.220.1100 | 39325.9 | | hypothetical protein conserved | LmxM.34.1260 | 1 | 0 | 0 | 2.26E-06 | 5.26E-07 | 2.40 | 0.23 |
| Pse.677.1000 | 33999.6 | Choline transport | hypothetical protein conserved | LmxM.14.0380 | 6 | 0 | 0 | 1.35E-06 | 2.52E-07 | 2.09 | 0.19 |
| Pse.84.1120 | 16005.5 | MPC domain | hypothetical protein conserved | LmxM.14.1460 | 2 | 0 | 0 | 4.76E-06 | 1.61E-06 | 1.76 | 0.34 |
| Pse.28.1240 | 60867.9 | Sugar transport | glucose transporter/membrane transporter D2 putative | LmxM.32.0290 | 12 | 0 | 0 | 1.84E-06 | 6.50E-07 | 1.69 | 0.35 |
| Pse.21.1380 | 72613.2 | HSP70 domain | heat shock protein 70 putative | LmxM.28.2820 | 0 | 0 | 1 | 5.95E-06 | 2.96E-06 | 1.26 | 0.5 |

### *3.3.4 Optimisation of cell lysis and bead washing conditions*

In the above analysis much of the biotinylated material is partitioned into the detergent insoluble fraction which is then discarded. As many surface proteins, such as GPI anchored proteins (Mayor and Maxfield, 1995), may be detergent insoluble I sought to develop an alternative lysis buffer that would enable these proteins to be captured by the streptavidin coated beads. In agreement with this, Figure 3.6B and Table 3.1 show that proteins with predicted GPI anchors were absent in the set of eluted proteins.

Here, 5 different lysis/wash buffers were constructed with increasing harshness. These were tested for their ability to solubilise the maximum amount of protein while maintaining the biotin streptavidin interaction. The composition of these buffers is shown in Table 3.2.

## TABLE 3.2 THE COMPOSITION OF CELL LYSIS/BEAD WASH BUFFERS

Reducing background from non cell surface proteins by optimising cell lysis and bead wash buffers. Five lysis/wash buffers were trialed with varying compositions of 8 molar (M) urea, 1% sodium deoxycholate, 0.2% and 2% sodium dodecyl sulfate (SDS), 1 molar (M) sodium chloride (NaCl).

| Buffer Name | Composition |
|---|---|
| L0 | 8M urea, 1% sodium deoxycholate |
| L1 | 8M urea, 1% sodium deoxycholate + 0.2% SDS |
| L2 | 8M urea, 1% sodium deoxycholate + 2% SDS |
| L3 | 8M urea, 1% sodium deoxycholate + 1M NaCl |
| L4 | 8M urea, 1% sodium deoxycholate, 0.2% SDS + 1 M NaCl |
| L5 | 8M urea, 1% sodium deoxycholate, 2% SDS + 1 M NaCl |

In addition, to reduce the perturbation to cells from incubation in buffered conditions, the osmolarity of the culture medium was measured and buffers for live cells were adjusted to be osmotically balanced. The results of using these different lysis/wash buffers are shown in Figure 3.7. Here, the addition of SDS and NaCl did not inhibit the biotin-streptavidin interaction as biotinylated material remained associated with the beads though all wash steps until it was eluted by boiling in Laemmli buffer. No insoluble material was observed in any of the above lysis buffers and thus all material was successfully solubilised. As discussed in Section 3.3.3, I did not want or expect the unbound material to be fully depleted of biotinylated material. There appeared to be some additional loss of material in buffer L5. Moreover, buffer L5 was unstable and it precipitated out of solution when chilled or left on the bench. For this reason, buffer L4 was chosen was chosen for all subsequent lysis and wash steps as the maximal stringency lysis and wash buffer.

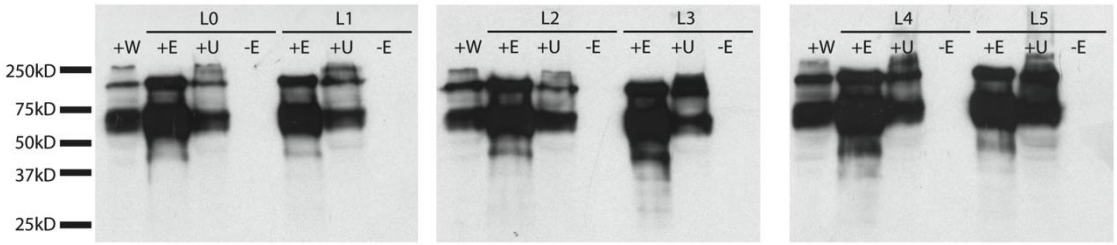



***Figure 3.7*** **ECL blots to follow optimisation of cell lysis and bead wash buffers**

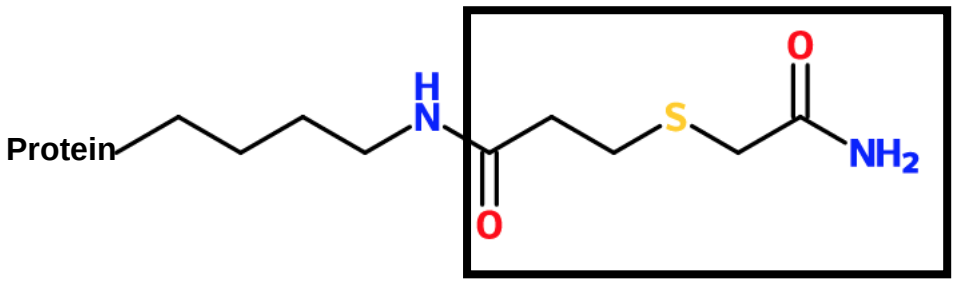
Comparison of different lysis/wash buffers on the biotin streptavidin interaction. Cell equivalents loaded are $5x10^5$ for whole cell lysates and $1x10^6$ for detergent soluble and insoluble fractions. Proteins were transferred to nitrocellulose membrane and probed with streptavidin conjugated horseradish peroxidase to detect the presence of biotinylated material in each fraction. Buffer composition is described in Table 3.2. Sample Key: +W= whole cell control, +E = eluate following incubation in whole cell lysate from surface biotinylated cells, +U = unbound material following bead incubation in whole cell lysate from surface biotinylated cells, -E = eluate following bead incubation in whole cell lysate from cells that were not biotinylated.

### *3.3.5 Interrogation of an alternative biotinylation reagent and elution strategies*

The biotin-streptavidin interaction is very strong and stable, thus the release of biotinylated proteins presents a challenge. Proteins cannot be eluted by incubation in excess biotin, and harsh elution conditions such as boiling in Laemmli buffer are required. However, this also disrupts the non-specific interactions of contaminant proteins with the streptavidin beads leading to high rates of contamination. Moreover, boiling is known to cause aggregation of hydrophobic proteins such as multimembrane pass proteins (Santoni *et al.*, 2000). Therefore two alternative strategies were tested. The first comprises an alternative biotinylation reagent (Sulfo-NHS-SS-biotin) that can be released from streptavidin without boiling. The second strategy was to digest proteins directly on streptavidin beads. While this latter strategy would abrogate effects due to boiling, it would not solve the non-specific contaminant problem and would also have an additional complication of introducing peptides derived from streptavidin.

Sulfo-NHS-SS-biotin contains a cleavable disulphide bond in the spacer arm allowing the captured proteins to be eluted from streptavidin beads by addition of a reducing agent. In this case the mass of the remaining portion of the spacer arm will be left attached to the peptide. The predicted additional mass remaining after this reduction and alkylation was added to the proteomic searches as a variable post translational modification (PTM) (Figure 3.8). The equivalent PTM cannot be determined for the on-bead digestion following labelling with LC-biotin.

**A**




**B**

| Unique | Peptide |
|---|---|
| U | K.NFLSQLSK.T |
| U | K.NFLSQLSK.T |
| U | K.NFLSQLSK.T + Deamidated (N) |
| | K.DGWGPCTQR.Y |
| | K.DGWGPCTQR.Y |
| U | K.FTDACISIWK.I |
| U | R.GPHEIEPCGPMK.N |
| | K.CYDSSKSAGIK.D + RandA_Sulfo-NHS-SS-Biotin (K) |
| U | R.GPHEIEPCGPMK.N + Oxidation (M) |
| U | K.IYAKCYDSSK.S + RandA_Sulfo-NHS-SS-Biotin (K) |
| | K.SAGIKDGWGPCTQR.Y + RandA_Sulfo-NHS-SS-Biotin (K) |
| U | R.FKVNGKFTDACISIWK.I + Deamidated (N); RandA_Sulfo-NHS-SS-Biotin (K) |

***Figure 3.8*** **Post translational modification with SS-biotin**

Post translational modification with EZ-link Sulfo NHS-SS-biotin following reduction with TCEP and alkylation with iodoacetamide **A)** The box shows the predicted additional mass of the post translational modification of a lysine residue with SS biotin after reduction and alkylation calculated as 139.98 atomic mass units. **B)** An example of the detection of the post translational modification with SS biotin on missed cleavage lysine residues in the peptide results.

### *3.3.6 Isolation and identification of putative surface proteins*

To determine if the L4 lysis buffer and use of a cleavable biotinylation reagent and specific elution step improved the selective recovery of biotinylated material from the cell lysates several samples were analysed by LC MS/MS. These samples were +E, -E and +U for both LC-biotin and SS-biotin elutions as described in 2.5.5.2 (LC) and 2.5.5.3 (SS). Unbound material was prepared as described in 2.5.5.4. The workflow for sample preparation is described as Protocol 2 in Chapter 2, All samples were subject to tandem mass spectrometry as described in 2.5.11 and the data searched as described in 2.5.12. The results are summarised in Figure 3.9. The full results of the data searches are available in Appendix 3.1 (LC-botin) and Appendix 3.2 (SS-biotin). The distribution of the $\log_2$-fold change in abundance for each protein between the eluate and the unbound material is shown in Figure 3.9A. This revealed that both biotinylation reagents produced selective recovery of specific proteins, which was not a random sample of the unbound material ($p < 0.05$ SS-biotin and $p < 0.01$ LC-biotin). As before, the distribution was shifted to the right and contained an additional right hand tail. To see if putative surface proteins were enriched in these sets of selectively recovered proteins hypergeometric tests were carried out to test for any overrepresentation of proteins containing predicted membrane associated domains. i.e. transmembrane helices, the signal peptide and the GPI anchor (Figure 3.9B and C).

For the LC-biotin, the set of proteins whose abundance was ≥ 2-fold higher in the eluate than in the flow-through was significantly enriched for proteins containing predicted membrane associated domains (Figure 3.9B). In contrast, for the SS-biotin the entire set (regardless of abundance) was enriched for proteins containing predicted membrane associated domains (Figure 3.9C). Some proteins were identified in both the LC and SS biotin datasets, those which represent likely contaminants are listed in red (Table 3.5). Many of the proteins identified belonged to the same orthogroups (gene family) and thus I will refer to orthogroup numbers as well as accession numbers in the following sections.

The inference of orthogroups is described in 2.4.5 and they are listed in Appendix 3.3. An orthogroup is defined here as the set of genes that are descended from a single gene in the last common ancestor of the species used to generate the gene families (Emms and Kelly, 2015). In this case that last common ancestor is the common ancestor of *Trypanosoma*, *Leishmania* and *Phytomonas*.

In the LC-biotin labelled samples 69,000 spectra were obtained and data searches against the *P. serpens* proteome identified 15,490 (22.45%) of these spectra, including matches to 9,893 unique peptides. These search results correspond to a total of 1,145 proteins identified across all 3 fractions, of which 290 were present in the eluate. Using the SINQ normalisation procedure, the abundance of 79 of these proteins was determined to be 2-fold higher in the eluate than in the unbound material. The SS-biotin labelled samples yielded 55,654 spectra across all samples. Of these spectra, 14,470 (26.22%) were matched to the predicted *P. serpens* proteome including 9,623 unique peptides. Data searches identified 1,220 proteins of which 86 were found in the eluate. Using the SINQ normalisation procedure, the abundance of 36 of these proteins was determined to be 2-fold higher in the eluate than in the unbound material. Proteins identified in the SS-biotinylation sample (Table 3.4) with a SS-biotin PTM detected are denoted with an asterisk. Among the proteins found in the eluate fraction determined to be less than 2-fold higher in abundance in the eluate compared to the unbound material there were 11 proteins that were identified by peptides containing a lysine residue labelled with a detectable biotin PTM.

Full proteomic datasets are available in Appendix 3.2 and Appendix 3.5 contains the predicted secondary structures for proteins identified in 3.3.6.1 and 3.3.6.2.

#### *3.3.6.1 Proteins identified using LC-biotin*

The 79 enriched proteins include a total of 23 proteins with at least 1 predicted surface associated domain (TM, GPI, SP). Adopting the criteria that proteins must be enriched 2-fold or higher in the bound material compared to the flow through and have either a

predicted cell surface associated domains or a *Leishmania* ortholog annotated as a putative cell surface protein there are a total of 20 highly probable cell surface proteins listed in Table 3.3.

Pse.422.1070 and 422.1030 were also included because of the enrichment value obtained despite only containing one membrane associated domain. Again, this highlights possible false negatives either as a result of the gene models or inability for the predictive tools to identify domains in *Phytomonas* protein sequences. Of these, a number are orthologs of probable surface proteins in *Leishmania* and *Trypanosoma* including amastin (Pse.222.1080 and Pse.264.1080), glucose transporter (Pse.28.1240) and receptor-type adenylate cyclase (Pse.556.1020 and Pse.556.1030). This dataset also includes 12 hypothetical proteins that have surface associated domains, of which 6 were not identified in *L. mexicana* and 6 are found in other trypanosomes.

These contain two proteins with a predicted signal peptide and at least one transmembrane domain: Pse.777.1000 and Pse.604.1040. Both of these proteins were also detected in the negative control elution (not evaluated to be a good way of identifying non-specific binding, discussed above). There were no proteins detected with a signal peptide and GPI anchor. A further 3 proteins were predicted to contain a GPI anchor and transmembrane domain: Pse.926.1000, Pse.222.1080 and Pse.264.1080. These unlikely topologies probably represent false positive transmembrane predictions. Two of these, Pse.222.1080 and Pse.264.1080 belong to the same gene family and are orthologs of known surface proteins in *L. mexicana* (amastin) also conserved in *Trypanosoma*. Other orthologs of likely cell surface proteins include Pse.270.1010, which is orthologous to a surface antigen-like protein with a predicted signal peptide, and Pse.556.1030 orthologous to receptor-type adenylate cyclase with a predicted transmembrane helix.

### *3.3.6.2 Proteins identified using SS-biotin*

Adopting the same criteria for evaluating surface proteins as in 3.5.6.2 there are 19 putative cell surface proteins identified with this labelling and reduction elution method listed in Table 3.4. Of these, 5 are orthologs of probable surface proteins in *Leishmania* and *Trypanosoma* including amastin (Pse.222.1080, Pse1124.1000 and Pse.264.1080), glucose transporter (Pse.28.1240) and surface antigen-like protein (Pse.270.1010). This dataset also includes 7 hypothetical proteins that have predicted membrane associated domains, of which 3 are unique to *Phytomonas* and 4 are found in other trypanosomes. Also included in this dataset are at least 4 proteins that are likely contaminants (those containing PFAM domains for pro isomerase, PEX11, AOX and PTPLA). However, at this stage the application of a stricter filter for candidate proteins may result in the exclusion of putative cell surface proteins so at this stage the aim was to cast a wide net for validation in Chapter 4.

Neither LC nor SS biotin selectively recovered any proteins that are predicted to localise to the cell surface *via* a signal peptide and GPI anchor. However, 3 proteins (Pse.960.1000, Pse.222.1080 and Pse.264.1080) were each predicted to encode a GPI anchor and 4 transmembrane helices. Pse.946.1080 is a member of a gene family only found in *P. serpens* and *P. EM1*. The other 2 proteins are members of a family of amastin proteins conserved in *Trypanosoma* and *Leishmania* species. However, the predicted protein topology of a GPI anchor and 4 transmembrane passes is unlikely and it is more probable that this represents a false positive prediction of TMDs possibly as a result of a hydrophobic protein core. Comparison of the proportion of known contaminants observed in the LC and SS biotin labelling experiments revealed that the SS biotin, presumably through the use of a cleavable spacer arm, produced fewer contaminants.

Despite the recovery of many likely surface proteins via both methods, there are also absences of known surface proteins. The most notable absence is gp63, the major surface protease. There are several possible explanations for this absence that are

reviewed in the discussion section. Furthermore, both proteomes suffer a degree of contamination of known non-cell surface proteomes and possible candidate surface proteins are predicted to exhibit non-canonical topologies in the plasma membrane giving reasons to question the computational predictions TM, GPI and SP sequences.

In order to expand the investigation into the cell surface of *Phytomonas* species I applied a bioinformatic filter to the set of predicted orthogroups. These results are presented in the remainder of this chapter.

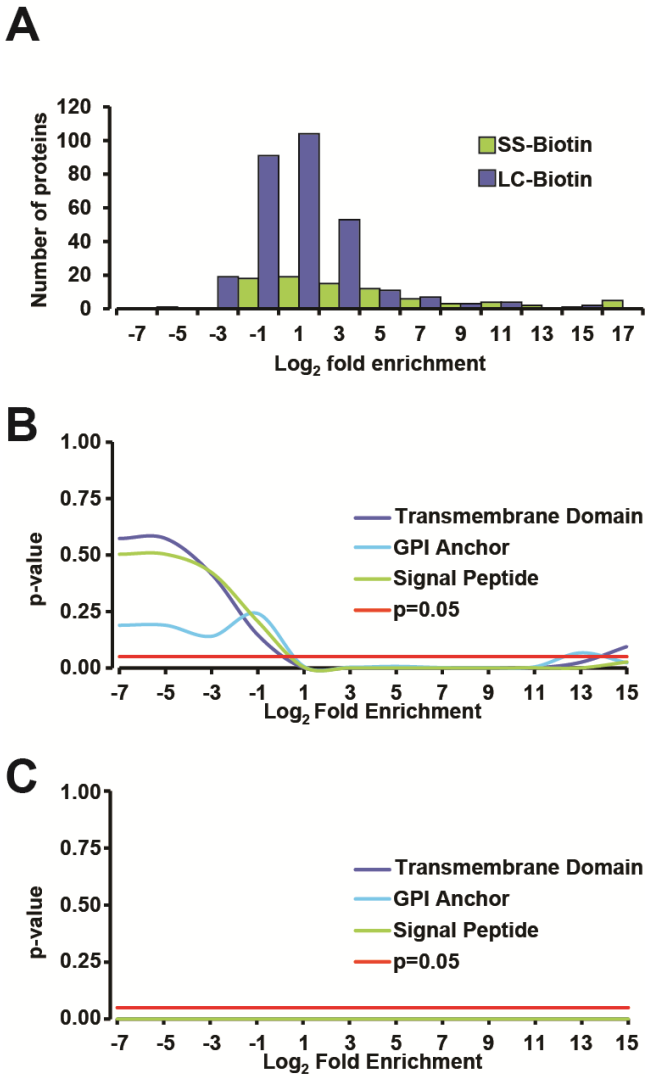



***Figure 3.9*** **Results of tandem mass spectrometry**

A) Comparison of the abundance estimates of proteins in the eluate vs unbound material for two biotinylation reagents. The results of hypergeometric tests to determine whether the set of proteins at that fold-change or higher contain more than the expected number of proteins that contain predicted domains that are associated with membrane proteins for B) LC-biotin and C) SS-biotin.

## TABLE 3.3 RESULTS OF LC-BIOTIN TANDEM MASS SPECTROMETRY: PUTATIVE CELL SURFACE PROTEINS LOG$_2$-FOLD MORE ABUNDANT IN THE ELUATE COMPARED TO THE UNBOUND MATERIAL

Putative cell surface proteins are those either containing either a TM helix, or a signal peptide, or a GPI anchor, or if the annotation of the ortholog in *L. mexicana* or PFAM domain indicate possible interaction with the plasma membrane. The change in abundance in the eluate compared to the unbound material is calculated form the spectral indices for these two fractions. Likely contaminants are indicated in red.

| Protein | Molecular weight (Da) | PFAM | L. mexicana annotation | L. Mexicana Ortholog | TM Domains | GPI | SP | Spectral Index Bound | Spectral Index Flow Through | Log$_2$-Fold Change In Abundance | SINQ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Pse.604.1040 | 62733.9 | | | | 4 | 0 | 1 | 6.51E-05 | 2.82E-09 | 14.50 | + B only |
| Pse.422.1070 | 25087.8 | | | | 0 | 0 | 1 | 4.56E-05 | 2.82E-09 | 13.98 | + B only |
| Pse.422.1030 | 50098.2 | | | | 2 | 0 | 0 | 8.51E-06 | 2.82E-09 | 11.56 | + B only |
| Pse.264.1080 | 20774.9 | Amastin domain | amastin, putative | LmxM.30.0450 | 4 | 1 | 0 | 6.40E-06 | 2.82E-09 | 11.15 | + B only |
| Pse.556.1020 | 24265.4 | | receptor-type adenylate cyclase, putative | LmxM.17.0190 | 0 | 0 | 0 | 3.22E-06 | 2.82E-09 | 10.16 | + B only |
| Pse.270.1010 | 42437.2 | | surface antigen-like protein | LmxM.04.0190 | 0 | 0 | 1 | 2.98E-06 | 2.82E-09 | 10.05 | + B only |
| Pse.926.1000 | 21849.3 | Amastin domain | | | 4 | 1 | 0 | 2.67E-06 | 2.82E-09 | 9.89 | + B only |
| Pse.556.1030 | 57325.7 | Guanyla te cyc | receptor-type adenylate cyclase, putative | LmxM.17.0190 | 1 | 0 | 0 | 4.55E-07 | 2.82E-09 | 7.34 | + B only |
| Pse.155.1000 | 424517 | | hypothetical protein, conserved | LmxM.10.0500 | 1 | 0 | 0 | 1.94E-07 | 2.82E-09 | 6.11 | + B only |
| Pse.422.1010 | 44018.6 | | | | 0 | 0 | 1 | 1.49E-05 | 2.30E-07 | 6.02 | 0.02 |
| Pse.777.1000 | 9142.84 | | hypothetical protein, conserved | LmxM.30.0900 | 2 | 0 | 1 | 2.56E-05 | 6.93E-07 | 5.21 | 0.03 |
| Pse.53.1220 | 35457 | GDPD domain | hypothetical protein, conserved | LmxM.24.2160 | 1 | 0 | 0 | 2.79E-06 | 9.44E-08 | 4.88 | 0.03 |
| Pse.677.1000 | 34165.5 | Amastin domain | hypothetical protein, conserved | LmxM.14.0380 | 6 | 0 | 0 | 4.39E-06 | 1.68E-07 | 4.71 | 0.04 |
| Pse.222.1080 | 20371.8 | Amastin domain | amastin, putative | LmxM.30.0450 | 4 | 1 | 0 | 2.64E-05 | 1.39E-06 | 4.25 | 0.05 |
| Pse.45.1050 | 27903.7 | PEX 11 domain | hypothetical protein, conserved | LmxM.34.3690 | 2 | 0 | 0 | 4.24E-06 | 4.70E-07 | 3.17 | 0.11 |
| Pse.104.1140 | 58667.1 | UbiA domain | prenyltransfer ase, putative | LmxM.28.1320 | 8 | 0 | 0 | 6.93E-07 | 1.40E-07 | 2.30 | 0.2 |
| Pse.682.1000 | 15509.7 | | | | 1 | 0 | 0 | 1.54E-05 | 5.35E-06 | 1.52 | 0.35 |
| Pse.61.1150 | 22069.4 | Arf domain | small GTP-binding protein, putative | LmxM.05.0030 | 0 | 0 | 1 | 3.31E-06 | 1.16E-06 | 1.52 | 0.35 |
| Pse.28.1240 | 60867.9 | Sugar transport | glucose transporter/me mbrane | LmxM.32.0290 | 12 | 0 | 0 | 1.08E-05 | 3.98E-06 | 1.45 | 0.37 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | transporter D2, putative | | | | | | | | |
| Pse.37.1110 | 30350.7 | DUF1295 domain | hypothetical protein, conserved | LmxM.25.1445 | 4 | 0 | 0 | 2.82E-06 | 1.34E-06 | 1.07 | 0.48 |

## TABLE 3.4 RESULTS OF SS-BIOTIN TANDEM MASS SPECTROMETRY: PUTATIVE CELL SURFACE PROTEINS $LOG_2$-FOLD MORE ABUNDANT IN THE ELUATE COMPARED TO THE UNBOUND MATERIAL

Putative cell surface proteins are those either containing either a TM helix, or a signal peptide, or a GPI anchor, or if the annotation of the ortholog in *L. mexicana* or PFAM domain indicate possible interaction with the plasma membrane. The change in abundance in the eluate compared to the unbound material is calculated form the spectral indices for these two fractions. Proteins identified with a SS-biotin PTM detected are denoted with an asterisk. Likely contaminants are indicated in red.

| Protein | Molecular weight (Da) | PFAM | *L. mexicana* annotation | *L. Mexicana* Ortholog | TM Domains | GPI | SP | Spectral Index Bound | Spectral Index Flow Through | $Log_2$ Fold Change | SINQ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Pse.264.1080 | 20774.9 | Amastin domain | amastin, putative | LmxM.30.0450 | 4 | 1 | 0 | 1.37E-04 | 2.15E-09 | 15.83 | + B only |
| Pse.604.1040 | 62733.9 | | | | 4 | 0 | 1 | 9.75E-05 | 2.15E-09 | 15.34 | + B only |
| Pse.926.1000 | 21849.3 | Amastin domain | | | 4 | 1 | 0 | 8.56E-05 | 2.15E-09 | 15.16 | + B only |
| Pse.270.1010 | 42437.2 | | surface antigen-like protein | LmxM.04.0190 | 0 | 0 | 1 | 4.83E-06 | 2.15E-09 | 11.01 | + B only |
| Pse.777.1000 | 9142.84 | | hypothetical protein, conserved | LmxM.30.0900 | 2 | 0 | 1 | 2.03E-04 | 3.11E-07 | 9.96 | 0 |
| Pse.677.1000 | 34165.5 | Choline transport domain | hypothetical protein, conserved | LmxM.14.0380 | 6 | 0 | 0 | 8.20E-05 | 6.48E-08 | 9.87 | 0 |
| Pse.1047.1010 | 66811.9 | Amino acid transport | amino acid permease | LmxM.30.1800 | 3 | 0 | 0 | 7.52E-06 | 4.96E-08 | 7.77 | 0.01 |
| Pse.422.1010 | 44018.6 | | | | 0 | 0 | 1 | 4.79E-05 | 3.92E-07 | 7.46 | 0.01 |
| Pse.73.1140 | 60206.6 | Amino acid transport | amino acid permease | LmxM.30.1800 | 11 | 0 | 0 | 6.47E-06 | 1.22E-07 | 6.05 | 0.02 |
| Pse.216.1030 | 29876.7 | | transmembrane protein, putative | LmxM.20.0240 | 0 | 0 | 1 | 7.34E-06 | 1.82E-07 | 5.86 | 0.02 |
| Pse.45.1050 | 27903.7 | PEX11 domain | hypothetical protein, conserved | LmxM.34.3690 | 2 | 0 | 0 | 8.85E-06 | 2.40E-07 | 5.54 | 0.03 |
| Pse.412.1030 | 25976.8 | PTPLA domain | protein tyrosine phosphatase, putative | LmxM.05.0280 | 5 | 0 | 0 | 6.96E-06 | 4.34E-07 | 4.53 | 0.06 |
| Pse.403.1040 | 72417.7 | ABC transporter | ATP-binding cassette protein subfamily G, member 2, putative (ABCG2) | LmxM.06.0090 | 5 | 0 | 0 | 7.01E-06 | 3.80E-07 | 4.39 | 0.05 |
| Pse.33.1240 | 21162.9 | Pro isomaerase | cyclophilin | LmxM.06.0120 | 1 | 0 | 1 | 5.38E-06 | 6.82E-07 | 3.50 | 0.13 |
| Pse.28.1240 | 60867.9 | Sugar transport | glucose transporter/ membrane transporter D2, putative | LmxM.32.0290 | 12 | 0 | 0 | 2.48E-05 | 4.45E-06 | 2.96 | 0.18 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Pse.205.1130 | 36601.9 | AOX domain | hypothetical protein, conserved | LmxM.36.4380 | 1 | 0 | 0 | 8.09E-06 | 2.50E-06 | 2.02 | 0.31 |
| Pse.1124.1000 * | 24206.5 | Amastin domain | amastin, putative | LmxM.30.0450 | 4 | 0 | 0 | 4.56E-05 | 2.15E-09 | 14.25 | + Elu only |
| Pse.22.1190 * | 249555 | ABC transporter | ABC transporter-like protein | LmxM.33.0670 | 9 | 0 | 0 | 1.48E-05 | 2.15E-09 | 10.26 | 0 |
| Pse.222.1080 * | 20371.8 | Amastin domain | amastin, putative | LmxM.30.0450 | 4 | 1 | 0 | 3.76E-04 | 7.61E-07 | 9.54 | 0 |

## Table 3.5 Proteins common to both LC and SS biotin datasets

Proteins common to both LC and SS biotin datasets (all proteins, not just putative surface listed in Tables 3.3 and 3.4). Proteins common to both datasets include putative surface proteins such as those annotated as amastin in *L. mexicana.* Proteins annotated as known contaminants e.g. ribosomal proteins in *L. mexicana* are shown in red.

| **$Log_2$-fold chance in abundance** | | **Protein** | **L. mexicana annotation** | **TMD** | **GPI** | **SP** |
|---|---|---|---|---|---|---|
| SS = 11.14859 | LC = 15.00445 | Pse.163.1060 | protein kinase, putative | 0 | 0 | 0 |
| SS = 15.3426 | LC = 14.49561 | Pse.604.1040 | | 1 | 0 | 1 |
| SS = 15.83021 | LC = 11.14998 | Pse.264.1080 | amastin, putative | 1 | 1 | 0 |
| SS = 11.00836 | LC = 10.04551 | Pse.270.1010 | surface antigen-like protein | 0 | 0 | 1 |
| SS = 15.15588 | LC = 9.891136 | Pse.926.1000 | | 1 | 1 | 0 |
| SS = 7.455338 | LC = 6.022948 | Pse.422.1010 | | 0 | 0 | 1 |
| SS = 9.956861 | LC = 5.209876 | Pse.777.1000 | hypothetical protein, conserved | 1 | 0 | 1 |
| SS = 9.866249 | LC = 4.708547 | Pse.677.1000 | hypothetical protein, conserved | 1 | 0 | 0 |
| SS* = 9.536701 | LC = 4.250058 | Pse.222.1080 | amastin, putative | 1 | 1 | 0 |
| SS = 5.54095 | LC = 3.174253 | Pse.45.1050 | hypothetical protein, conserved | 1 | 0 | 0 |
| SS = 4.451705 | LC = 2.216318 | Pse.259.1020 | unspecified product | 0 | 0 | 0 |
| SS = 1.429201 | LC = 1.728955 | Pse.21.1220 | ribosomal protein S20, putative | 0 | 0 | 0 |
| SS = 2.959507 | LC = 1.445162 | Pse.28.1240 | glucose transporter/membrane transporter D2, putative | 1 | 0 | 0 |
| SS = 4.322902 | LC = 1.099919 | Pse.1326.1010 | 60S acidic ribosomal protein, putative | 0 | 0 | 0 |

### *3.3.7 Bioinformatic interrogation of the predicted* P. serpens *cell surface proteome*

The proteomic dataset described in this chapter presented an insight into the protein components of the *P. serpens* cell surface. However, all datasets suffered from a degree of contamination and the absence of the major surface protease gp63 from any of the mass spectrometry data searches raises the potential that other proteins may also have been missed because of some proteins not being expressed in the chosen culture conditions. To complement and extend the proteomic investigation and to gain additional insight from comparison with other species the full set of predicted gene models for *P. serpens* was clustered with the full set of putative proteins from the following species: *T. carassii, T. brucei, T. vivax, T. congolese, T. cruzi, T. grayi and T. rangeli L. major, L. mexicana, L. infantum, L. tarentolae, L. braziliensis, P. EM1, P. HART1,* and *P. françai.* Clustering resulted in 21,425 total orthogroups (Appendix 3.3) of which 7047 (33%) were found in *P. serpens* and 8226 (38%) found in at least one species of *Phytomonas*. Figure 3.10 shows the relationship between shared and unique orthogroups across 3 genera of digenetic trypanosomatids. In total, 5,478 (26%) of the orthogroups were common to *Trypanosoma, Leishmania* and *Phytomonas.*

OrthoFinder did not cluster all genes into orthogroups. A number these remain as single, unclustered genes (21425 orthogroups generated by OrthoFinder containing 141764 genes (excluding pseudogenes) of which 11817 are unclustered genes). The two most likely sources of unclustered genes are 1) novel genes with no similarity to any other gene sequences in the species set being clustered 2) genes that have undergone rapid sequence evolution such that similarity to related genes is too weak to facilitate clustering in an orthogroup. Orthogroups and unclustered genes that are unique to a species are of potential interest from a biological point of view because they might contribute to the unique biology of the species. For example, the SRA gene is a unique gene that is only found in sub-populations of *T. brucei* where it confers resistance to normal human serum

(De Greef and Hamers, 1994). Therefore these unclustered genes have not been removed these from the analysis or from the summary presented in Figure 3.10.

There were 963 orthogroups present in *Leishmania* and *Trypanosoma* that are absent from any of the sequenced *Phytomonas* genomes. As *Phytomonas* and *Leishmania* evolved from a common ancestor following the divergence of *Trypanosoma* these genes appear to have been lost from the lineage that gave rise to *Phytomonas.* The majority of the 963 orthogroups lost from *Phytomonas* (n = 652, 68%) comprise hypothetical proteins with no known functions and may represent gene families that are required for survival in animal hosts. Additionally, consistent with the absence of a requirement for heme in *Phytomonas* (Koreny *et al.*, 2012), the orthogroups lost from *Phytomonas* include cytochrome c oxidase assembly factors and subunits and two ferric reductase transmembrane protein families.

One common feature of each genus is the presence of a large number of orthogroups that are specific to that genus. This phenomenon is most pronounced in *Trypanosoma,* (Figure 3.10A).

Further investigation was carried out to elucidate the functions of the proteins found in the *Phytomonas* lineage that are not present in *Trypanosoma* and *Leishmania.* The *P. serpens* sequences from these *Phytomonas* specific orthogroups were searched against the Pfam database of protein families represented by multiple sequence alignments and hidden Markov models (Finn *et al.*, 2014). Of these 1328 proteins, 1265 had no significant similarity (e-value<$1x10^{-5}$) to sequences present in the Pfam database (Finn *et al.*, 2014). This low hit rate may be at least partially explained by the fact that PFAM is a collection of domains built by identifying regions of similarity between related sequences. The lack of Phytomonas genome data means that PFAM domains for Phytomonas specific proteins have not yet been generated.

A total of 63 proteins produced significant hits to protein domains in the database. The annotations of these hits included 1 amino acid transport domain, 6 protease domains

and an amastin domain. The recent publication of the *P. HART1* and *P. EM1* reported the expansion of a gene family of putative secretable aspartyl proteases in the pathogenic isolate. Coupled with the 6 protease domains available here an expansion in proteases is reminiscent of similar phenomena in plant pathogenic fungi (Porcel *et al.*, 2014).

#### *3.3.7.1 Inference of putative cell surface orthogroups*

In trypanosomatids, as in other eukaryotes, proteins that localise to the exterior surface of the plasma membrane do so either through transmembrane helices or GPI anchors. To achieve this cell surface localisation the majority of these proteins are targeted to the cell surface by the secretion system and thus the majority contain N-terminal signal peptides. These three features (signal peptides, transmembrane domains, and GPI anchors) can be predicted from the amino acid sequence of a protein using computational methods. In this chapter I sought to build upon a computational analysis that was carried out in *Trypanosoma* species (Jackson *et al.*, 2013) in which proteins with a potential cell surface localisation were identified by a genome-wide screen for proteins that contained either a signal peptide and one or more transmembrane domains, or a signal peptide and a GPI anchor domain. While this type of computational screen for putative cell surface localised proteins has merit, there are three major limitations to using this approach. First, the presence of a transmembrane helix and a signal peptide does not guarantee that a protein will localise to the cell surface membrane. These features could also target a protein to a cell internal membrane such as the endoplasmic reticulum, secretory vesicle, or Golgi apparatus. Second, this screen will exclude surface localised proteins that are Type II transmembrane proteins or other transmembrane proteins where the first transmembrane domain is more than a few residues from the amino terminus (and therefore not detected by these computational predictive tools). Third, computational methods to predict GPI anchors, signal peptides and transmembrane domains are not perfect and can result in both false positive errors (i.e.

prediction of features that are absent) and false negative errors (i.e. fail to detect *bona fide* features).

I acknowledge that the first two limitations described above are not addressed in this thesis or in previous computational analyses of putative cell surface proteins in trypanosomatids (Jackson *et al.* 2013). Thus in this chapter some cell surface proteins will be missed and some proteins that localise to cell internal membranes will be mistakenly identified as potential surface proteins. While these two limitations are not addressed, I have attempted to address the issue of computational prediction error. Specifically, to minimise the false positive and false negative error associated with computational prediction of signal peptides, transmembrane domains and GPI-anchors I analysed the presence or absence of these features for all genes in each orthogroup. Each orthogroup was then analysed and the consensus for the presence and absence of surface associated features was computed. In this way, evidence to support the presence/absence of surface associated features are drawn from all members of an orthogroup. This approach protects against prediction error for an individual gene and assumes that if the features are evolutionarily conserved then the localisation of the protein is also likely evolutionarily conserved.

The *P. serpens* transcriptome contains 13960 transcripts encoded an ORF≥100 amino acids. Of these there were 3872 transcrips where the ORF began at the first, second or third nucleotide of the transcript and thus were likely incomplete transcripts resulting in N-terminally truncated ORFs. Similarly, 2148 transcripts terminated before an in frame stop codon and thus encoded C-terminally truncated ORFs. Finally, 3980 transcripts encoded ORFs that spanned the entire transcript and thus represented ORFs that were truncated at both the N and C termini. As a result of this evidence for truncation I have not rejected proteins that are not predicted to encode a signal peptide if there is other evidence indicating that it could be a surface protein, specifically if it is predicted to encode other surface associated domains or is annotated as a putative cell surface protein in *Leishmania*

Where available I have included the annotation from *L.mexicana* orthologs in the results of the bioinformatic screens and in the Appendices to this thesis. This provides some additional information and in some cases can confirm the presence of known surface proteins e.g. gp63. Potential next steps to parse this dataset would be to remove proteins which are predicted to contain a mitochondrial targeting sequence as identified by MitoFates or similar programme. Following this remaining sequences could be searched for Pfam domains, where present the type of Pfam domain would provide additional information to support or discard the hypothesis for cell surface localisation.

To identify the set of orthogroups that may contain surface targeted genes, a bioinformatic screen was applied to the full orthogroup dataset described above. Orthogroups were designated as putative surface orthogroups if ≥35% of the genes in the orthogroup fell into one of two categories: 1) they contained a signal peptide and at least one transmembrane helix or 2) a signal peptide and a GPI anchor. The cut off that was used was informed by the observed values for the ubiquitously conserved surface protein gp63. Here, the percentages of proteins predicted to encode GPI anchors and signal peptides was 40.1% and 35.9% respectively (Appendix 3.4). Application of this filter to the complete set of orthogroups resulted in a set of 1542 putative cell surface orthogroups (containing 9796 genes of which 1032 are unclustered genes Appendix 3.4). Figure 3.10B shows a Venn diagram detailing the relationships between the orthogroups after the application of the 'putative cell surface' orthogroups filter. There are limitations to this method, namely that type II transmembrane proteins will be excluded. However, these are not excluded in the biochemical isolation reported earlier and the two different methods of identifying putative cell surface proteins complement each other in this respect. Beyond GPI, SP and TM domain prediction, no further targeting sequences were searched for. For example, mitochondrial targeting sequences were not predicted for this dataset. This was in order to conduct an analysis directly comparable to the *Trypanosoma* cell surface phylome (Jackson *et al.*, 2013). This also prevented the exclusion of sequences that could encode cell surface localised proteins but could be

mispredicted to be targeted to the mitochondria. While this is a limitation of the study, the intention was to compensate for this by validating surface localisation with YFP tagging (Chapter 4).

In total 29% (2358 unique orthogroups (UO) containing 3604 genes of which 1853 are unclustered genes. A total of 8226 orthogroups (TO) were identified in *Phytomonas,* UO/TO=0.287 Figure 3.10A) of all orthogroups (including unclustered genes) found in *Phytomonas* were not found in *Trypanosoma* or *Leishmania*. Analogous values for *Leishmania* and *Trypanosoma* were 16% (UO=1249 containing 4388 genes of which 476 are unclustered genes and TO=7880, UO/TO= 0.159 Figure 3.10A) and 62% (UO= 10987 containing 21422 genes of which 9488 are unclustered genes and TO=17628, UO/TO=0.623 Figure 3.10A) respectively. Proportionally fewer (22%) surface orthogroups were found to be unique to *Phytomonas* (54 unique surface orthogroups (USO) containing 87 genes of which 40 are unclustered genes. In total, 250 surface orthogroups (TSO) were identified in *Phytomonas*. USO/TSO=0.216 Figure 3.10B). In contrast, both *Leishmania* and *Trypanosoma* exhibited a higher proportion of unique surface orthogroups than would be expected given the distribution of all orthogroups and unclustered genes. Specifically, 24% (USO=78 containing 282 genes of which 28 are unclustered genes and TSO=324, USO/TSO=0.241, Figure 3.10B) of *Leishmania* surface orthogroups and unclustered genes were unique to the genus while 81 % (USO=1128 containing 3895 genes of which 964 are unclustered genes and TSO=1384, USO/TSO=0.814 Figure 3.10B) of *Trypanosoma* surface orthogroups and unclustered genes were unique to the genus. This data is consistent with a scenario where adaptation to hosts with adaptive immune systems may have resulted in a selective increase in the number of surface localised proteins. In contrast adaptation to plant environments does not seem to have produced the same effect.

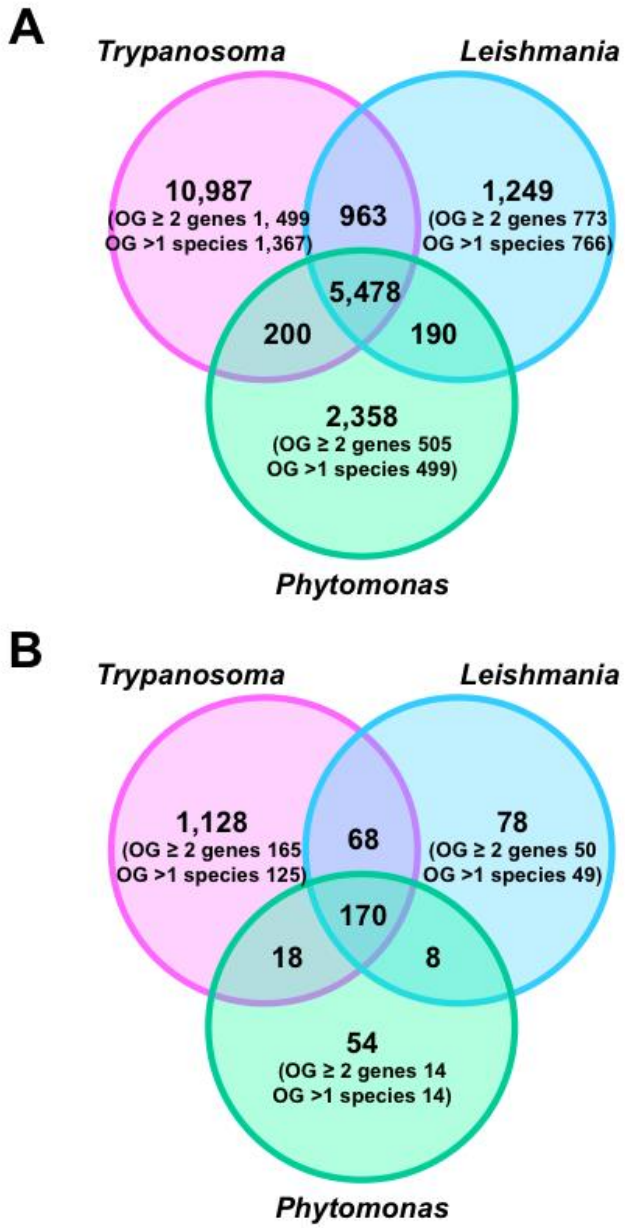



***Figure 3.10*** **A summary of the results of orthogroup clustering**

Venn diagrams summarising the results of orthogroup clustering with OrthoFinder (Emms and Kelly, 2015). The number of gene families (orthogroups) that are common or unique to *Phytomonas*, *Leishmania* and *Trypanosoma*. Also shown are the number of orthogroups (OG) containing a total of greater or equal to two genes (OG ≥2 genes) and the number of orthogroups in more than one species (OG > 1 species). **A)** Shows information for all orthogroups (Appendix 3.3) and **B)** for orthogroups of predicted surface genes (Appendix 3.4).

This comparative method identified a core set of 170 putative cell surface orthogroups that were present in the last common ancestor of the trypanosomatids used in this analysis, and that were conserved in at least one species of *Phytomonas*, *Leishmania* and *Trypanosoma* (Appendix 3.5). Of these, 116 (68%) are annotated as hypothetical proteins in the *L. mexicana* and *T. brucei* genomes. The largest family of genes in this core set is the major surface protease gp63, which has undergone a lineage specific expansion in multiple species. The largest expansion is found in *P. HART1* in agreement with the findings presented in the publication of the genome sequence (Porcel et al., 2014). Also conserved are receptor type adenylate cyclase (GRESAG4); membrane bound acid phosphatase 1 precursor; flagellum adhesion glycoprotein fla1; genes involved in GPI anchor synthesis such as GPI-anchor transamidase subunit 8 and a family of endosomal integral membrane proteins. There is likely to be much fundamental biology as yet undiscovered in this set of genes.

In total, 68 putative cell surface orthogroups were found to be conserved in *Leishmania* and *Trypanosoma* but absent from *Phytomonas*. Given the phylogenetic relationship of these species this indicates that these genes have been lost from the *Phytomonas* lineage. Of these 68, 53 families (78%) are annotated as hypothetical proteins indicating that a large proportion of *Phytomonas* surface biology may be uncharacterised. Analysis of the 15 orthogroups missing from *Phytomonas* that had functional annotation terms in one of more trypanosomatid revealed no clear biological trend.

Similarly, the bioinformatic investigation also identified 54 putative surface orthogroups that were unique to the *Phytomonas* lineage (Table 3.6). Table 3.6 shows that of the ten predicted surface orthogroups that are unique to the Phytomonas lineage and found in three out of the four species studied were predicted to be localised to the cell surface via a transmembrane domain and signal peptide. None of the predicted cell surface orthogroups found in three or more species of Phytomonas studied were predicted to localise to the surface by a GPI anchor. Transmembrane proteins are integral to the

plasma membrane and are often involved with roles such as ion transport. As a result, they tend to be functionally constrained and therefore conserved. Only one predicted cell surface protein retrieved any BLAST result, protein disulfide isomerase. As this is an enzyme found in the endoplasmic reticulum, it represents either a false positive BLAST hit or false positive surface prediction and demonstrates the limits of the dataset produced by this method. However, the absence of any significant BLAST hits for the remaining nine proteins found in three or more *Phytomonas* species demonstrates that there may be some novel biology to be uncovered here. The proportion of orthogroups that are predicted to be surface specific that are unique to *Phytomonas* is comparatively lower than that which is observed for *Leishmania* or *Trypanosoma.* In *Trypanosoma* species, parasites have evolved clade or species specific gene families encoding surface proteins involved in evading the adaptive immune response in the animal host. Similarly, in *Leishmania* these intracellular parasites of animals have a number of clade specific surface orthogroups such as families of amastins that are only found in this genus. *Phytomonas* are adapted to parasitism in a host where there is no adaptive immune system, therefore there may have been reduced selective pressure to evolve new cell surface gene families. As a result, a greater proportion of the *Phytomonas* predicted cell surface orthogroups are shared with other genera of trypanosomatids and are likely to be involved with core cellular processes and morphology.

Only one of these *Phytomonas* specific putative surface orthogroups OG009377, was identified by biotinylation and streptavidin pull down described above. This likely cell surface orthogroup exhibits some similarity to amastin-like proteins in *Leishmania*. Amastins are transmembrane glycoproteins localised to the cell surface of trypanosomatids and are highly immunogenic but their precise function is unknown (Jackson, 2010). Of the 54 *Phytomonas* specific orthogroups, only 7 were found in all *Phytomonas* species analysed. BLAST searching the NCBI non-redundant protein database with these 7 orthogroups did not reveal any clues as to biological function.

Interestingly, all 7 putative surface orthogroups are present in a single copy in each *Phytomonas* genome except for OG008665, which has undergone a duplication event in *P. EM1*. All 7 of these putative surface orthogroups contain genes encoding proteins predicted to localise to the cell surface by transmembrane domains and signal peptides consistent with the majority of surface proteins in this genus. Although they are not found in *Leishmania* and *Trypanosoma,* the family OG008538 contains an ortholog in the insect parasite *Crithidia fasculata*. It is also possible that these proteins do not localise to the cell surface and that their predicted signal peptides and transmembrane helices target them to the endomembrane system.

Some known cell surface components did not pass the surface orthogroups filter. These orthogroups include OG000024 and OG007162. The former is an orthogroup of 199 amastin genes with 100% of genes predicted to contain a transmembrane domain but only 30% predicted to encode a signal peptide. The latter contains 7 amastin-like genes that also have 100% of genes with a predicted transmembrane domain but 14% of sequences containing a predicted signal peptide sequence. These low percentages observed for known cell surface components may be because of incomplete gene models in *Leishmania* and *Phytomonas* species. Furthermore, likely contaminants passed the filter including OG019962 orthologous to *Leishmania* peroxisome assembly proteins and OG021130 containing putative short chain dehydrogenase. These false positives and false negatives indicate that further validation of this list is required.

To gain a finer resolution of the distribution of genes in the putative surface orthogroups the content of each orthogroup was analysed at the individual gene and species level (Figure 3.11A). For genes in all orthogroups, the majority exhibited conservation across multiple species (Figure 3.11A). In contrast, for putative cell surface orthogroups while there are several orthogroups that have ubiquitous distribution, a larger proportion of genes were found in just one or two species. Figure 3.11B shows the same data but comparing the proportion of orthogroups rather than individual genes that constitute the

orthogroups. Here, the majority of putative surface gene families are found in just a single species and very few are conserved across multiple species of trypanosomatids. The species distribution of the “single species” surface genes is shown in Figure 3.11C. This shows that *Trypanosoma* species contain the majority of these species specific predicted surface genes with relatively few observed for the *Leishmania* and *Phytomonas* species studied. *T. vivax* was shown to have the most unique surface gene families followed by *T. brucei* and *T. congolese. L. tarentolae* and *P. serpens* have the most species specific putative surface genes in the *Leishmania* and *Phytomonas* genera respectively. One possible explanation for the differences for the number of unique orthogroups between genera observed in Figure 3.11 part C is that it is an artefact resulting from non-equal species sampling. As the *Trypanosoma* species sampled are more closely related to each other than the *Leishmania* or *Phytomonas* species. More equal species sampling is required here to test the observed distribution. If indeed there are more unique orthogroups in the closely related *Trypanosomatid* species then these orthogroups become even more evolutionarily significant.

In addition, Figure 3.11 shows that for genes in orthogroups predicted to localise to membranes there are a higher proportion of genes found only in one species than there are of the genes found in more than one species. This trend may be a result of limited species sampling across the kinetoplastida. This is not observed for all orthogroups or all genes. This suggests that these putative membrane proteins tend to be found in only one species and therefore probably evolved after speciation or are too divergent for OrthoFinder to identify orthologs. Part C shows that most unique orthogroups tend to be low or single copy, with the exception of the *Trypanosoma* species where there has been large expansion of genes surface orthogroups such as VSG in adaptation to escaping the adaptive host immune response.

Taken together, the fact that predicted membrane genes have such limited species distribution indicates that there has been higher rates of birth and death of putative

surface genes than expected from an analysis of all genes. An alternative explanation could be that the rates of sequence evolution are much higher for surface genes, reducing similarity between sequences and so leading to errors in detecting homologous sequences across multiple species. In either case, this enhanced rate of change in surface gene content likely reflects the fact that surface proteins are central to survival in different host environments.

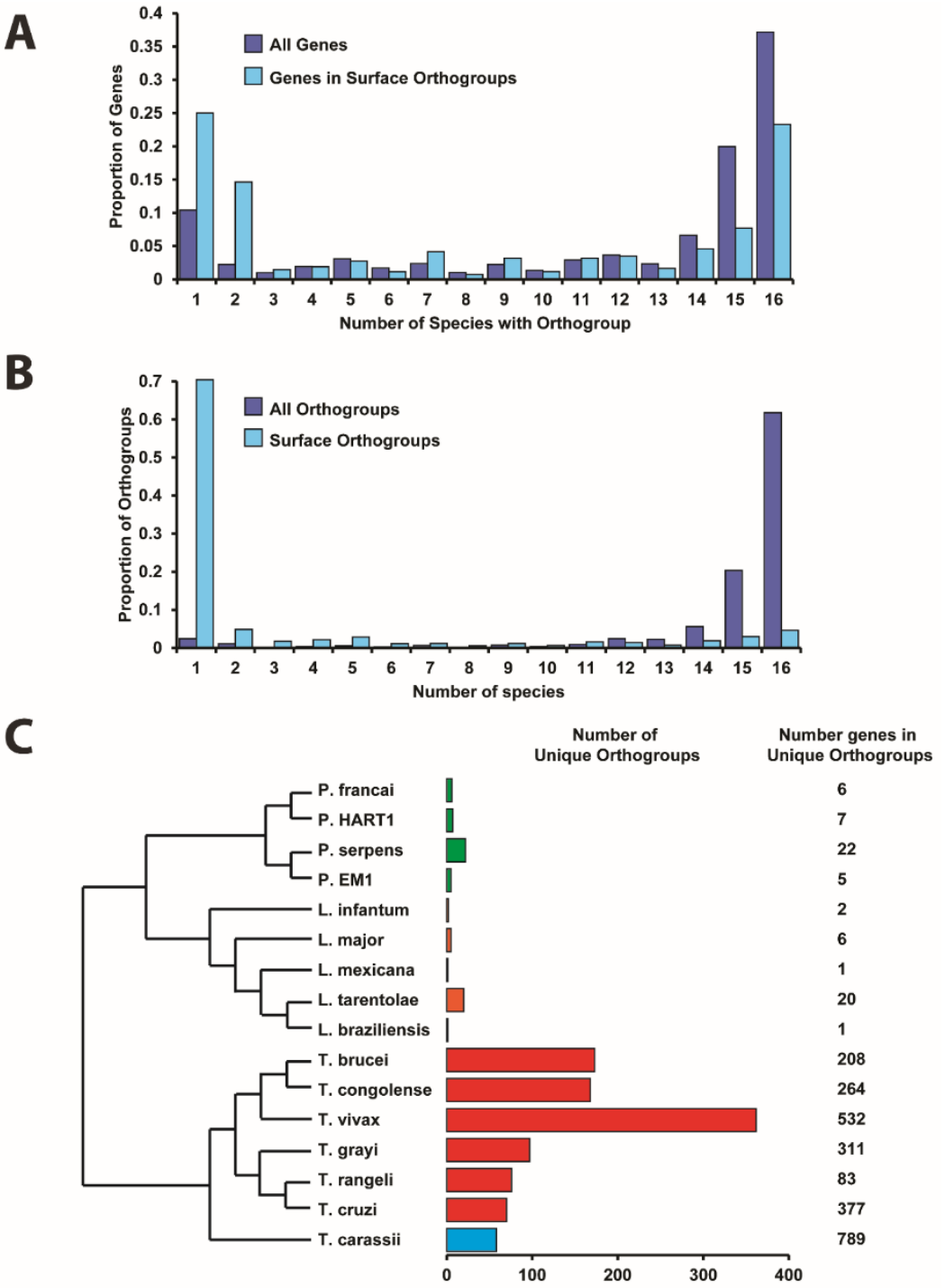



***Figure 3.11*** **The distribution of the surface orthogroups of *Phytomonas, Trypanosoma* and *Leishmania***

**A)** The proportion of genes in all orthogroups and surface orthogroups plotted against the number of species with that orthogroup. **B)** The proportion of orthogroups plotted against the number of species containing that orthogroups for all orthogroups and surface orthogroups. **C)** A distribution of the species specific (unique) orthogroups with the total number of genes contained in those unique orthogroups indicated on the right.

## TABLE 3.6 THE PREDICTED CELL SURFACE ORTHOGROUPS FOUND IN *PHYTOMONAS* SPECIES THAT ARE ABSENT FROM *TRYPANOSOMA* AND *LEISHMANIA* SPECIES

Numbers in column under species name indicate the number of genes from that species in the listed orthogroup. %GPI is the percentage of all genes in the orthogroup that have predicted GPI anchors. %TMD is the percentage of all genes in the orthogroup that have predicted transmembrane domains. %SignalP is the percentage of all genes in the orthogroup that have predicted signal peptides. BLAST result column shows the annotation information for the best BLAST hit in the NCBI nr database. Likely contaminants are indicated in red.

| Orthogroup | BLAST result | P. EM1 | P.HART1 | P. serpens | P. francai | % GPI | % TMD | % SignalP |
|---|---|---|---|---|---|---|---|---|
| **OG008665** | no significant hits outside phyto | 2 | 1 | 1 | 1 | 0 | 100 | 80 |
| **OG008968** | no significant hits outside phyto | 1 | 1 | 1 | 1 | 0 | 100 | 50 |
| **OG009064** | | 1 | 1 | 1 | 1 | 0 | 100 | 50 |
| **OG008538** | | 1 | 1 | 1 | 1 | 0 | 100 | 75 |
| **OG009057** | hypothetical proetin, conserved | 1 | 1 | 1 | 1 | 0 | 100 | 50 |
| **OG009041** | hypothetical conserved protein | 1 | 1 | 1 | 1 | 0 | 100 | 50 |
| **OG009059** | | 1 | 1 | 1 | 1 | 0 | 50 | 50 |
| **OG009447** | | 1 | 1 | 1 | 0 | 33 | 100 | 67 |
| **OG009434** | Putative cell surface-expressed gene family, protein disulfide isomerase, putative | 1 | 1 | 1 | 0 | 0 | 100 | 67 |
| **OG007880** | | 0 | 1 | 1 | 1 | 0 | 67 | 67 |
| **OG009377** | amastin-like protein | 1 | 0 | 2 | 0 | 100 | 100 | 100 |
| **OG010158** | | 1 | 0 | 1 | 0 | 100 | 100 | 50 |
| **OG010109** | | 1 | 0 | 1 | 0 | 0 | 100 | 50 |
| **OG010186** | | 0 | 1 | 1 | 0 | 0 | 100 | 50 |
| **OG019962** | peroxisome assembly protein, putative | 1 | 0 | 0 | 0 | 100 | 100 | 100 |
| **OG020160** | | 1 | 0 | 0 | 0 | 0 | 100 | 100 |
| **OG020041** | | 1 | 0 | 0 | 0 | 0 | 100 | 100 |
| **OG020137** | | 1 | 0 | 0 | 0 | 0 | 100 | 100 |
| **OG020150** | | 1 | 0 | 0 | 0 | 0 | 100 | 100 |
| **OG020187** | | 0 | 1 | 0 | 0 | 0 | 100 | 100 |
| **OG020532** | | 0 | 1 | 0 | 0 | 100 | 100 | 100 |
| **OG020394** | | 0 | 1 | 0 | 0 | 0 | 100 | 100 |
| **OG020507** | | 0 | 1 | 0 | 0 | 0 | 100 | 100 |
| **OG020543** | | 0 | 1 | 0 | 0 | 0 | 100 | 100 |
| **OG020476** | | 0 | 1 | 0 | 0 | 0 | 100 | 100 |
| **OG020288** | hypothetical protein, conserved | 0 | 1 | 0 | 0 | 0 | 100 | 100 |
| **OG021247** | | 0 | 0 | 1 | 0 | 0 | 100 | 100 |
| **OG021084** | | 0 | 0 | 1 | 0 | 0 | 100 | 100 |
| **OG021276** | | 0 | 0 | 1 | 0 | 0 | 100 | 100 |
| **OG020934** | | 0 | 0 | 1 | 0 | 0 | 100 | 100 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **OG021179** | | 0 | 0 | 1 | 0 | 0 | 100 | 100 |
| **OG021373** | | 0 | 0 | 1 | 0 | 100 | 100 | 100 |
| **OG021399** | | 0 | 0 | 1 | 0 | 0 | 100 | 100 |
| **OG020998** | | 0 | 0 | 1 | 0 | 100 | 100 | 100 |
| **OG021669** | | 0 | 0 | 1 | 0 | 100 | 0 | 100 |
| **OG021059** | | 0 | 0 | 1 | 0 | 100 | 0 | 100 |
| **OG021071** | | 0 | 0 | 1 | 0 | 100 | 0 | 100 |
| **OG021744** | | 0 | 0 | 1 | 0 | 0 | 100 | 100 |
| **OG021130** | short chain dehydrogenase, putative | 0 | 0 | 1 | 0 | 0 | 100 | 100 |
| **OG021229** | | 0 | 0 | 1 | 0 | 100 | 0 | 100 |
| **OG021335** | | 0 | 0 | 1 | 0 | 0 | 100 | 100 |
| **OG021034** | | 0 | 0 | 1 | 0 | 100 | 0 | 100 |
| **OG021318** | | 0 | 0 | 1 | 0 | 100 | 0 | 100 |
| **OG021559** | | 0 | 0 | 1 | 0 | 0 | 100 | 100 |
| **OG021011** | nucleobase transporter | 0 | 0 | 1 | 0 | 100 | 100 | 100 |
| **OG021178** | | 0 | 0 | 1 | 0 | 100 | 0 | 100 |
| **OG021159** | | 0 | 0 | 1 | 0 | 100 | 0 | 100 |
| **OG021238** | | 0 | 0 | 1 | 0 | 0 | 100 | 100 |
| **OG020898** | | 0 | 0 | 0 | 1 | 0 | 100 | 100 |
| **OG020586** | | 0 | 0 | 0 | 1 | 100 | 0 | 100 |
| **OG020617** | | 0 | 0 | 0 | 1 | 0 | 100 | 100 |
| **OG020740** | | 0 | 0 | 0 | 1 | 0 | 100 | 100 |
| **OG020790** | | 0 | 0 | 0 | 1 | 0 | 100 | 100 |
| **OG020600** | | 0 | 0 | 0 | 1 | 0 | 100 | 100 |

## 3.4 Discussion

A single origin of parasitism in an ancestral cell gave rise to trypanosomes specialised for infecting hosts from across the animal and plant kingdom (Lukeš *et al.*, 2014). Our knowledge of the deep evolution of the cell surface proteins that have facilitated colonisation of this broad range of hosts is lacking in resolution, in part because only a few mammalian infective species have been studied comprehensively. *Phytomonas* is positioned as a sister group to intracellular animal infective *Leishmania* species. *Leishmania* and *Phytomonas* are both equally closely related to *Trypanosoma*. Since the split of *Trypanosoma, Phytomonas* and *Leishmania* from their common ancestor over 200 million years ago (Lukeš *et al.*, 2014) there has been lineage-specific expansion of ancestral gene families as well as *de novo* evolution of genes. For example, the extracellular *T. brucei* evolved a large, novel family of VSG genes to produce antigenic

variation which is central to the strategy of immune evasion employed by this parasite (MacGregor *et al.*, 2012). In contrast, *Leishmania* underwent a large expansion of the ancestral amastin genes (Jackson, 2010) and this expanded repertoire is thought to function in the invasion and maintenance of the intracellular stage of the parasite. The evolutionary context of *Phytomonas* represents an ideal opportunity to investigate the deep evolution of surface proteins in these enigmatic parasites to help identify previously undiscovered, conserved biology which may help piece together the evolutionary history of the trypanosomatids and how they have adapted to different host environments.

Two approaches were taken to characterise the cell surface of *Phytomonas* parasites. The first approach was to isolate cell surface proteins and identify them using mass spectrometry. The second approach was to use bioinformatic methods to predict families of putative cell surface proteins from the available genome sequences. In the first approach, multiple different strategies, reagents and conditions were interrogated to optimise a proteomic workflow for identification of cell surface proteins. Success was evaluated by computational analysis of the enrichment of proteins predicted to contain membrane associated domains and tagging of a putative surface protein described in Chapter 4. Analysis of the proteins identified by this biotin labelling approach revealed that contamination from non-surface components was difficult to eliminate. The source of these contaminants was also difficult to ascertain. For example, some proteins were isolated non-specifically, presumably through interaction with the streptavidin resin while other obvious contaminants had detectable biotin PTMs indicating that these proteins were biotinylated possibly from lysed cells present in the sample. While levels of contamination were reduced through optimisation, it still presented a significant challenge to the analysis such that not all isolated proteins could be considered to have come from the cell surface. In this context it is noteworthy that contamination of this nature is not unusual in cell surface biotinylation and has been reported by previous studies in *Leishmania* (Yao *et al.*, 2010). Of the two biotinylation reagents used in this

analysis, the SS-biotin produced better results. Specifically, there was a statistical overrepresentation of domains associated with membrane proteins irrespective of the cut-off applied to the relative abundance of the eluted proteins. In contrast only a subset of the eluted proteins form the LC biotin showed the same enrichment. This suggests that the use of a cleavable biotin reagent coupled with a specific elution step is more effective for specific isolation of cell surface proteins. Use of an alternative labelling reagent may have produced reduced or different levels of contamination. Additionally, few GPI anchored cell surface proteins were detected - perhaps more of these would have been identified if a method such as GPI-PLC had been used.

The proteomic study suggested the presence of amastin-like genes on the *Phytomonas* cell surface. Amastin genes encode highly immunogenic transmembrane glycoproteins that are conserved across the trypanosomatids with lineage-specific gene family expansion in the *Leishmania* (Jackson, 2010). Their function is unknown but it has been speculated that they may be membrane bound signal transducers and/or function in the interaction with host macrophages (Jackson, 2010). The subcellular localisation of one of these amastin genes is described in Chapter 4. Interestingly, I also identified a *Phytomonas* specific gene-family that exhibited weak similarity to amastin genes and also encompassed a weak hit to an amastin PFAM domain. This amastin-like gene family was only found in species of *Phytomonas* that are not associated with any plant pathology (*P. serpens* and *P. EM1*). The gene family may represent an innovation specific to the non-pathogenic lineage of plant parasites or may have evolved in the common ancestor and were subsequently lost in the disease causing *Phytomonas* species. Further investigation of this gene family under the context of parasitism and disease could help to unravel the relationship between parasitism and pathology in the plant infective trypanosomatids.

The bioinformatic prediction and analysis of the cell surface proteomes of 16 species of trypanosomatid revealed 170 putative surface orthogroups that comprise a conserved

set of predicted cell surface proteins that were present in the last common ancestor of all species analysed. However much of this conserved surface biology is uncharacterised as 68% of these core surface gene families are annotated as hypothetical proteins in *T. brucei* and *L. mexicana*. It would be worthwhile to utilise high-throughput techniques to validate localisation and investigate the function of many of the orthologs in the related model organism *T. brucei* (Dean *et al.*, 2015).

This investigation confirmed the previously reported expansion of the major surface protease gene family in *P. HART1,* the species of *Phytomonas* that causes hartrot of palm trees (Porcel *et al.*, 2014). There are multiple copies of the major surface glycoprotein gp63 in the *P. serpens* genome but there was no evidence of gp63 in the eluate fractions analysed by mass spectrometry. However, this is not a cause for concern as the parasites were grown in medium containing 5% foetal calf serum (2.3.1) and serum has been shown to inhibit the expression of gp63 (Elias *et al.*, 2012). This may indicate that the surface composition of *Phytomonas* parasites can be remodelled according to environmental conditions. It will be interesting to determine to what extent this applies to other surface proteins and whether this is true for parasites in plant environments as well as those in the insect host.

Taken together, the combined outputs of the proteomics and bioinformatics revealed that *Phytomonas* did not undergo an expansion in cell surface protein repertoire. This is in contrast to what has occurred in *Leishmania* and *Trypanosoma,* where species and lineage specific expansion of surface gene families is intrinsically linked to host adaptation. The one exception to this is the expansion of the gp63 gene family in *P. HART1*. This lack of expansion of surface protein gene families likely represents differences in the selective pressures experienced by the parasites in adapting to plant as compared to animal hosts. These differences are likely attributable to differences in the immune mechanisms of the hosts and the parasite requirement to evade these immune responses in order to establish and maintain an infection.

## 3.5 Conclusions

The combined proteomic and bioinformatic investigations provide new insight into the biology of the trypanosomatid cell surface. Furthermore, it identified a core suite of protein families found in the ancestor of *Phytomonas, Trypanosoma* and *Leishmania* that have been subject to lineage specific adaptations and expansions. Remarkably, the bulk of this conserved biology is unknown due to the large number of hypothetical proteins. The primary finding is that *Phytomonas* species did not undergo an analogous evolutionary strategy as observed in *Trypanosoma* and *Leishmania* species for parasite adaptation through the expansion of the repertoire of cell surface proteins. The data presented in this thesis does not provide any evidence that *Phytomonas* species underwent an expansion in the repertoire of cell surface proteins as they adapted to their plant hosts. The expansion of cell surface proteins has been observed in related species of *Leishmania* and *Trypanosoma* in adapting to the evasion of an adaptive immune system in the animal host. The data show that *Phytomonas* species have proportionally fewer unique predicted cell surface orthgroups than their animal infective relatives. The predicted *Phytomonas* cell surface is composed of a number of proteins that are conserved across the trypanosomatids and therefore likely play a role in core cell processes such as amino acid transporters. A small number of *Phytomonas* unique orthogroups predicted to localise to the cell surface have been identified and Chapter 4 describes my attempts to validate the localisation of these proteins at the plasma membrane.

# *Chapter 4: Gene tagging in* Phytomonas serpens, *validating the localisation of putative cell surface proteins*

## 4.1 Preface

The proteomic workflow in Chapter 3 indicated that there are a number of putative cell surface proteins identifiable in *Phytomonas serpens* some of which are conserved in *Trypanosoma* and *Leishmania* species and some which are unique to *Phytomonas*. Bioinformatic investigation identified 250 families of putative membrane associated and surface protein genes in the available *Phytomonas* genomes. Of these, 54 were absent from *Trypanosoma* and *Leishmania* species, 18 shared only with *Trypanosoma*, 8 shared only with *Leishmania* and 170 families conserved across all 3 genera. To determine whether the putative surface proteins identified in Chapter 3 localise to the cell surface, a subset of these proteins were tagged with a fluorescent protein. This enables visualisation of the subcellular localisation of the tagged protein. Though the approach was challenging and yielded only one *bona fide* surface protein, the results presented in this chapter provide the first example of endogenous protein tagging in the genus *Phytomonas*.

## 4.2 Introduction

Determining the localisation of a protein within a cell is often central to understanding protein function (Stadler et al., 2013). The primary method for determining the subcellular localisation of proteins is to use electron or light microscopy. A comprehensive range of methods are available for use in determining the subcellular localisation of proteins using light microscopy of which the most popular are immuno-detection and fluorescent protein tagging. Here, a high throughput gene tagging approach was used due to limitations of funding in the project which precluded generation of antibodies.

While immuno-detection methods are widely used, they require the generation of antibodies that bind to the protein of interest. These must be raised against each protein of interest individually making this process relatively low throughput and expensive. In contrast, fluorescent protein tagging based techniques are less expensive as multiple genes can be readily tagged and the fluorescent protein can be visualised directly. However, the disadvantage of this tagging approach is that the fluorescent protein tag itself may cause the tagged protein to mislocalise. Multiple locations of protein fusion e.g. at both the N′ and C′ termini can help to minimise artefacts (Stadler *et al.*, 2013). Compared to immunofluorescence methods, fluorescent protein based methods are relatively quick, cheaper and can therefore be used in high throughput workflows (Dean *et al.*, 2015).

For the purposes of this work it was determined that native fluorescence provided sufficient resolution for the purpose of validating the putative components of the *P. serpens* cell surface. Native fluorescence has been used successfully in other studies of trypanosomatids to determine the subcellular localisation of proteins (Dean *et al.*, 2015) and can be combined with controls such as labelling the cell surface carbohydrates with concanavalin A (Gadelha *et al.*, 2015). In addition, this method could be combined with the trypsin ‘shave’ method (where whole cells are incubated with trypsin to cleave surface proteins (Tjalsma *et al.*, 2008)) or immuno-detection with an antibody to the fluorescent protein to show that the signal is present on the external side of the plasma membrane although this was not performed.

As stated above, tagging the endogenous gene carries the risk of affecting the native protein. Enhanced Yellow Fluorescent Protein (eYFP) is a 26.4 kDa protein and the addition of this molecule to the target protein could perturb protein folding and result in mislocalisation and/or degradation of the chimeric protein. This problem is particularly pertinent for surface proteins as the sequences at the N′ and C′ termini of surface localised proteins often contain critical sequences for targeting of the protein to the cell

surface. Specifically, the signal peptide is an N′ terminal peptide usually 15-30 amino acids long that targets a protein for translation at the endoplasmic reticulum, the peptide is then cleaved during translocation (Nielsen *et al.*, 1997). Additionally, signal-anchor sequences are responsible for insertion and anchoring in the membrane in Type II and Type III membrane proteins which are targeted to cell membranes without a cleavable signal peptide (Figure 4.1) (Goder and Spiess, 2001). Signal-anchor sequences are usually longer in length than the signal peptide and can be internal to the polypeptide chain. They are not cleaved, and form a transmembrane helix in the protein. Type II membrane proteins induce the translocation of the C-terminus across the bilayer and the opposite is true for Type III membrane proteins (Figure 4.1) (Goder and Spiess, 2001). The glycosylphosphatidylinositol (GPI) anchor attaches a protein to the plasma membrane at the C-terminus of the protein. The GPI anchor is the result of a hydrophobic signal at the carboxyl terminus which is retained in the endoplasmic reticulum during translation where it is cleaved and the GPI glycolipid attached (Ferguson, 1999). The addition of a fluorescent protein tag at either the N′ or C′ termini thus has the potential to disrupt the signal peptide or GPI anchor if they are present but not signal anchors.

This problem can be overcome by inserting an additional signalling sequence in the tagging construct (Gadelha *et al.*, 2015). However, when using an artificial signal peptide in place of the endogenous signal sequence other caveats for misintegration or mislocalisation apply. The signal sequence prediction may be a false positive, and replacement of this signal sequence with another may result in mislocalisation of the protein. However it has been shown in *T. brucei* that addition of a known surface signal sequence to a non-surface protein does not result in mislocalisation of the protein to the cell surface (Gadelha *et al.*, 2015). This is presumably due to the fact that these proteins lack other features (such as transmembrane domains) that would cause them to be retained at this location. It may also be that the combination of the known signal

sequence with tagged protein may lead to problems with cleavage of the signal sequence and potential mistargeting.

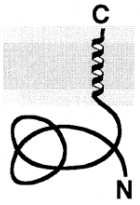
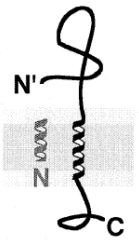
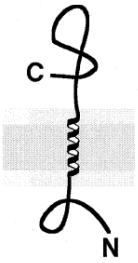
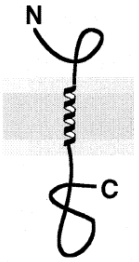
| Signal type: | C-Terminus-translocating signals | | | N-Terminus-translocating |
|---|---|---|---|---|
| Topogenic determinants: | C-terminal signal | Cleaved signal + stop transfer sequence | Signal-anchor | Reverse signal-anchor |
| Machinery: | unknown | SRP/SR/Sec61 + signal peptidase | SRP/SR/Sec61 | SRP/SR/Sec61 |
| exoplasmic<br>Final topology:<br>cytoplasmic |  |  Type I |  Type II |  Type III |
| Examples: | Synaptobrevin; cytochrome $b_5$ | Glycophorin; LDL receptor | Transferrin receptor; galactosyl transferase | Synaptotagmin; neuregulin; cytochromes P-450 |

***Figure 4.1*** **Topogenic determinants of single-spanning membrane proteins**

The different types of membrane proteins and the signalling mechanisms associated with each type. Reprinted from (Goder and Spiess, 2001) with permission from Elsevier.

*Phytomonas serpens* has great potential as a model trypanosome system for the study of protein localisation and function. It is closely related to biomedically important species but is non-pathogenic to humans; has a short doubling time relative to the other trypanosomes and can be grown to high densities in inexpensive growth media (Miranda *et al.*, 2015). However, because the plant infective trypanosomes have attracted limited scientific interest compared to their mammalian infective relatives the methods available for genetic manipulation are not as well developed.

It has been shown that *Phytomonas* cells can be transfected using electroporation (Lukeš *et al.*, 2006; Miranda *et al.*, 2015). *P. serpens* was transfected with a *Leishmania* expression vector to introduce enhanced Green Fluorescence Protein (eGFP) to the small subunit (SSU) locus (Lukeš *et al.*, 2006). In another study, a *T. cruzi* expression vector was used to transfect GFP into *Phytomonas Jma* (Miranda *et al.*, 2015). Here, expression of the heterologous protein was achieved by the vector forming a stable episome. In comparison with *T. cruzi*, drug selection and expression of GFP took longer in the plant parasite. Qualitative analysis by Western blot showed that expression levels of GFP were comparable in both *T. cruzi* and *P. Jma* (Miranda *et al.*, 2015). This demonstrates that while the genetic tools for *Phytomonas* are not as well established as they are for other trypanosomatid systems it is possible to take technologies and techniques designed for other trypanosomes and apply them for use in the plant parasite.

In order to tag the endogenous protein the tagging construct must integrate into the genome to tag one copy (allele) of the gene (Kelly *et al.*, 2007). While no examples of endogenous protein tagging in *Phytomonas* exist, successful genomic integration and expression of GFP in *P. françai* was achieved in our laboratory (Butler *et al.*, in preparation). This was achieved by using 500 base pairs of genomic sequence to target the integration of the tagging construct into the desired locus by electroporation and homologous recombination.

In Chapter 3, cell surface biotinylation and streptavidin pull down yielded a list of proteins that were shown to contain a significant overrepresentation of hallmarks of surface proteins (i.e. signal peptides, transmembrane domains and GPI anchors). A number of putative surface proteins were shortlisted from the proteomic and bioinformatic datasets for experimental validation of surface localisation using gene tagging. In order to interrogate the cellular localisation of these proteins both N′ and C′ terminal tagging were used as appropriate. Two tagging methods were attempted - the first used a nested PCR approach (described in section 2.6.2) to make C′ and N′ terminal tagging constructs with ~500bp of homology up and downstream to the target insertion site. The second technique was to use a long primer PCR approach (described in section 2.6.4) where 140bp oligonucleotides were designed for the amplification of a C-terminal eYFP and resistance marker sequence flanked on each end with a 120bp sequence homologous to the sites up and down stream of the target sequence. The former method was successful in yielding cell surface localisation of one of the target proteins identified by proteomics in Chapter 3. This was validated by sequencing and Western Blot. All molecular techniques are described in Chapter 2, section 2.6.

## 4.3 Prioritisation of proteins for localisation

It was necessary to prioritise the proteins identified by proteomic and bioinformatic methods in Chapter 3 for validation of localisation by eYFP tagging. Proteins identified by cell surface biotinylation in Chapter 3 are shown in Figure 4.2. The proteins shown are those with the greatest $\log_2$-fold change in the eluate compared to the unbound material. Proteins shown in orange were prioritised for tagging. The pPoTv2 and pJ1258 vectors were used and tagging constructs for TY:eYFP:TY were designed for either N′ terminal tags (pPoT) or C′ terminal (pJ1258) tags as described in 2.6.2 (Figure 4.3). The tagging vectors were kindly provided by the Laboratory of Professor Keith Gull (Dean *et al.*, 2015).

The TY epitope tag is a 10 amino acid (EVHTNQDPLD) fragment of a protein from *Saccharomyces cerevisiae* transposable element Ty1 (Bastin *et al.*, 1996). The TY tag is recognised by the BB2 monoclonal antibody (Bastin *et al.*, 1996). This is helpful when it is advantageous to double label e.g. in immunofluorescence or for alternatives to anti-GFP antibodies in Western Blotting.

Proteins were prioritised for tagging if they contained predicted membrane and cell surface associated domains (these are the N-terminal signal peptide, C-terminal GPI anchor and transmembrane domains). Proteins shortlisted and the rationale for shortlisting and location of eYFP fusion are described in Table 4.1. Cells were transfected using the transfection protocol described in 2.6.6. Where a GPI anchor was detected proteins were tagged on the N-terminus. However, while preventing the disruption of the GPI anchor site an N-terminal tag will disrupt a signal sequence targeting the protein to the endoplasmic reticulum and hence to the surface. Thus, GPI anchored proteins pose a challenge for tagging with fluorescent proteins. For proteins predicted to be targeted to the plasma membrane with a signal peptide the tag was designed on the C-terminus**.**

Table 4.1 contains the candidates identified by the biochemical approach and shortlisted for tagging used the nested PCR approach as per 2.6.2 (Figure 4.3) including the tagging of Pse.270.1010 with the attempted reproduction of the endogenous sequence predicted to encode a signal peptide as per 3.6.3. The proteins listed here once again demonstrate the limitations and problems with predicting membrane associated domains in the predicted *P. serpens* proteome as observed in Chapter 3. For example. Pse.246.1080, Pse.1124.1000, Pse.926.1000 and Pse.222.1080 are all predicted the unusual topologies of 4TMD and a GPI anchor, and Pse.422.1000 may be a contaminant secreted protein as it is only predicted to contain a signal peptide.

Table 4.2 contains the candidates identified by both biochemical and bioinformatic approach and shortlisted for tagging using the long primer approach as per 2.6.4. The method that was chosen to tag the candidate genes in Table 4.2 was a high throughput method designed to enable a large number of genes to be tagged. Thus criteria were relaxed so more genes could be included in the experimental validation pipeline. The targets listed in Table 4.2 were all prioritised because they were either hypothetical proteins with surface associated domains or were orthologs to known surface proteins in *Trypanosoma* and *Leishmania.* The glucose transporter GT1 was used as a positive control. All PCR was carried out with a negative control containing no DNA. Files have been included in Appendix 4 that were used to design primers for nested PCR tagging and single knockouts. They were desined in ApE (http://biologylabs.utah.edu/jorgensen/wayned/ape/) and show the sequence of the tagged locus as it was designed in accordance with the method used in *Leishmania* (Dean *et al.*, 2015).

All proteins were predicted to localise to the cell surface by internal transmembrane domains, thus all eYFP tags were designed to be C-terminal to prevent the disruption of any N-terminal signal peptide. Unfortunately, this experimental validation (methods described in 2.6.3 and 2.6.7) pipeline did not work, and possible explanations for this are explored in the discussion of this chapter.

## Table 4.1 *P. serpens* proteins shortlisted for eYFP tagging using nested PCR

Proteins are indicated by their accession number. Membrane associated domains are included when they have been predicted. Molecular weight is also shown. Where a GPI anchor was detected proteins were tagged on the N-terminus. Where a signal peptide was detected proteins were tagged on the C-terminus. A brief rationale for the tagging of each protein is included to indicate why each protein was included in this shortlist.

| Protein | Predicted surface hallmarks | Terminus for eYFP | Molecular weight (Da) | Rationale for tagging |
|---|---|---|---|---|
| Pse.19.1260 | - | C′ | 57674.5 | The protein with the greatest observed change in abundance in the eluate compared to the flow through for the SS biotin labelled sample. |
| Pse.246.1080 | 4 TMD & GPI | N′ | 20774.9 | Members of an amastin-like orthogroup conserved in *P. françai, P. EM1, Leishmania* species, *T.cruzi, T. rangeli* and *T. grayi.* N-terminal tag to prevent disrupting GPI anchor (possible false positive). |
| Pse.1124.1000 | | | 24206.5 | |
| Pse.222.1080 | | | 20371.8 | |
| Pse.604.1040 | SP & 4 TMD | C′ | 62733.9 | Unique to *P. serpens.* C-terminal eYFP tag to prevent disrupting signal peptide. |
| Pse.926.1000 | 4 TMD & GPI | N′ | 21849.3 | Pfam predicts an amastin domain. N-terminal tag to prevent disrupting GPI anchor (possible false positive). |
| Pse.677.1000 | 6 TMD | C′ | 33999.6 | Hypothetical protein with no orthologs detected outside *P. serpens.* C-terminal tag as there was no N-terminal intergenic sequence (incomplete genome). |
| Pse.422.1030 | 2 TMD, no signal peptide: alignment shows sequence is truncated at N terminus | C′ | 50098.2 | Paralogous sequences from OG007413. Hypothetical protein. C-terminal tag to prevent disrupting signal peptide. |
| Pse.422.1010 | SP | C′ | 44018.6 | Paralogous with Pse.422.1030 |

| | | | | |
|---|---|---|---|---|
| Pse.270.1010 | SP & GPI (if undetermined amino acids (not at C-terminus), removed) | N′ terminal tag signal peptide | 42437.2 | Orthologous to a surface antigen-like protein. Signal peptide was synthesised *de novo* and inserted into the tagging construct at the N-terminus of the eYFP described in 2.6.3 |
| Pse.352.1060 | 2 TMD | N′ | 92700.1 | ATP-dependant zinc metallopeptidase with 2 transmembrane helices. No signal peptide predicted. |
| Pse.777.1000 | SP & 2 TMD | C′ | 9142.84 | Hypothetical protein, conserved and with surface hallmarks. C-terminal tag to prevent disrupting signal peptide. |

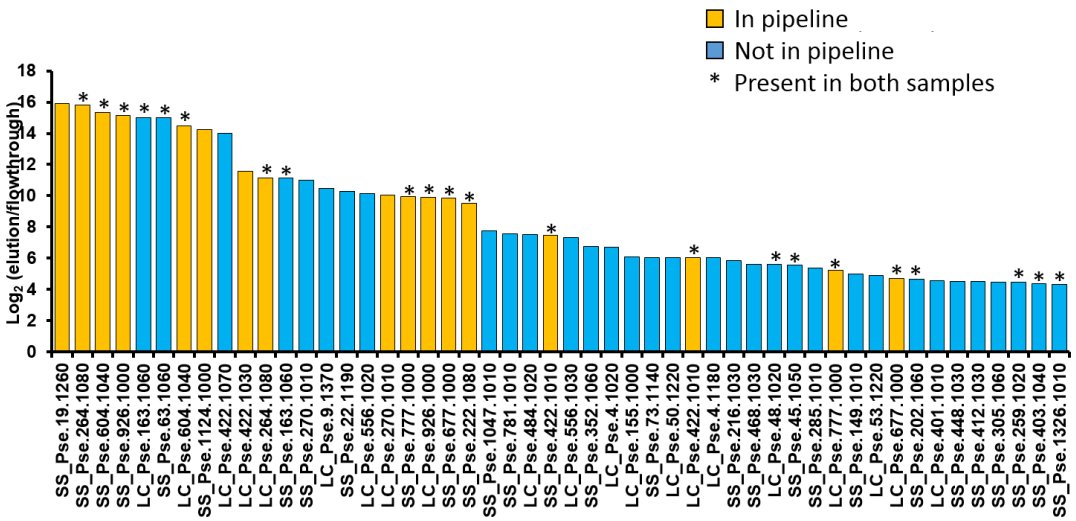



***Figure 4.2*** **Proteins shortlisted for YFP tagging using nested PCR**

Proteins with the highest enrichment scores identified by cell surface biotinylation of *P. serpens* identified in Chapter 3 ranked by $\log_2$ fold enrichment. Proteins shown in yellow were shortlisted for tagging. Proteins shown with an asterisk were identified in both the SS-biotin and LC-biotin labelling experiments.

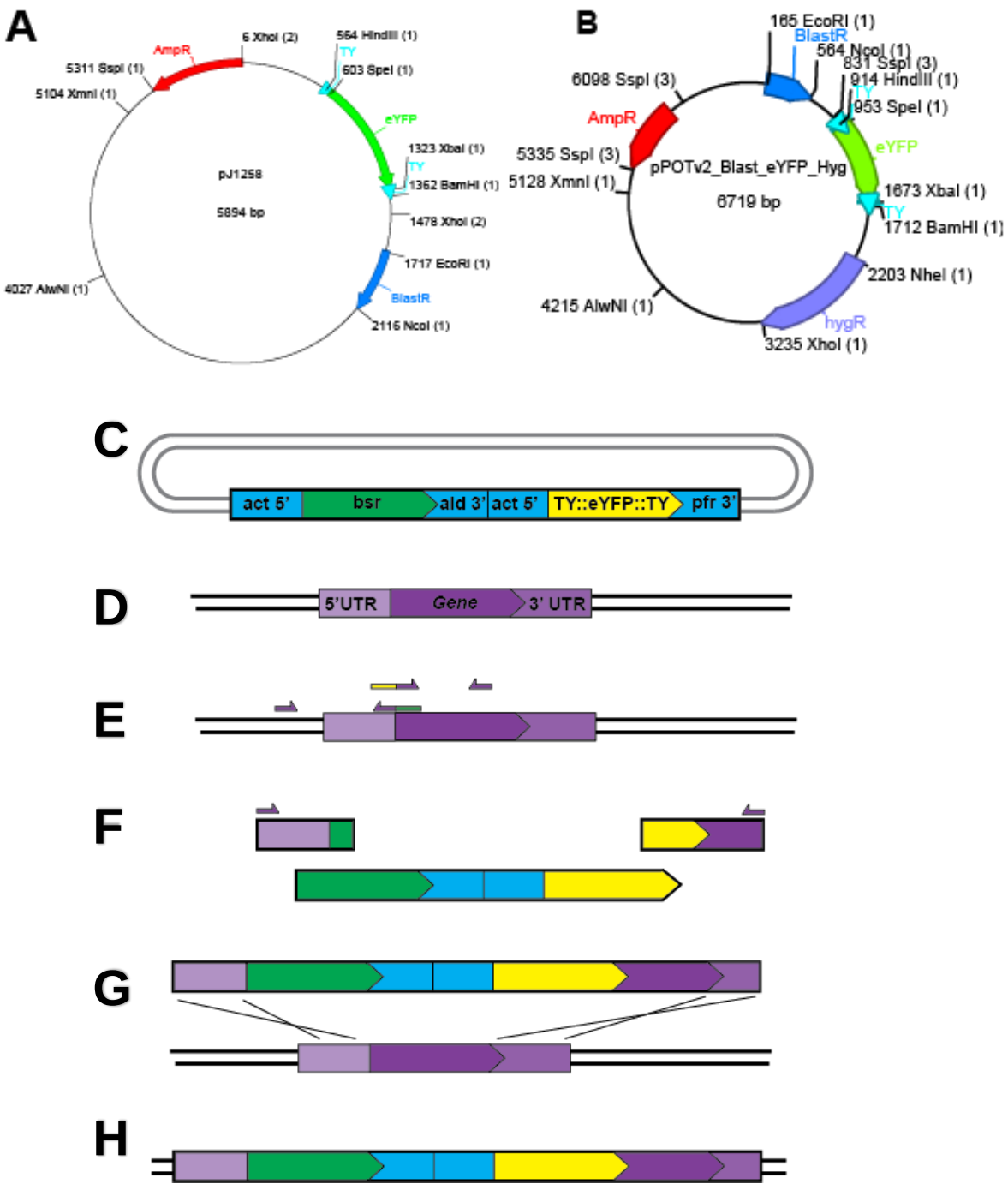



***Figure 4.3 A schematic diagram of the nested PCR tagging approach***

**A)** Restriction map of tagging vector pJ11257 **B)** Restriction map of pPOT. Both maps show eYFP, blasticidin resistance selectable marker for selection of *P. serpens* and ampicillin resistance selectable marker for cloning in *E. coli*. All vector sequences and maps are available in Appendix 4. **C-H)** Reproduced form Figure 2.1 showing a schematic of the nested PCR approach for N' terminal tagging.

## Table 4.2 Proteins shortlisted for eYFP tagging using long primer PCR

This high throughput technique enabled the attempted tagging of a greater number of candidate proteins. Proteins here were identified from both bioinformatic and experimental methods. Proteins identified in the biochemical method include the molecular weights calculated from mass spectrometry. Predicted molecular weights from ExPASy. The annotation column shows the annotation information for the best BLAST hit in the NCBI nr database. The three final columns indicate the proportion of genes in the orthogroup that have predicted GPI anchors, transmembrane domains (TMD) and signal peptides respectively.

| Accession | Orthogroup | Annotation from orthologs | PFAM | Molecular Weight (Da) | % GPI in orthogroup | % TMD in orthogroup | %SP in orthogroup |
|---|---|---|---|---|---|---|---|
| Pse.87.1000 | OG020934 | | None detected | 13287.15* | 0 | 100 | 100 |
| Pse.926.1000 | OG009377 | | Amastin | 21849.3 | 100 | 100 | 100 |
| Pse.956.1020 | OG007413 | | None detected | | 0 | 25 | 50 |
| Pse.71.1110 | OG003247 | hypothetical protein, conserved | TRAP alpha | 31757.93* | 0 | 100 | 93 |
| Pse.73.1140 | OG000324 | ABC transporter, putative | Transmembrane amino acid transporter | 60206.6 | 84 | 100 | 0 |
| Pse.794.1010 | OG021179 | | None detected | 11674.04* | 0 | 100 | 100 |
| Pse.8.1300 | OG003069 | hypothetical protein, conserved | None detected | 30459.3 | 0 | 93 | 100 |
| Pse.850.1010 | OG000166 | hypothetical protein, conserved | Choline transporter | 30263.27* | 3 | 100 | 3 |
| Pse.63.1060 | OG000708 | hypothetical protein, conserved | None detected | 32324.1 | 0 | 10 | 79 |
| Pse.632.1010 | OG005017 | hypothetical protein, conserved | None detected | 33479.81* | 7 | 79 | 93 |
| Pse.70.1160 | OG007152 | | Amastin | 22806.18* | 38 | 100 | 63 |
| Pse.39.1150 | OG005570 | | IGP family C-type lectin domain | 50456.70* | 0 | 93 | 93 |
| Pse.40.1010 | OG006457 | hypothetical protein, conserved | None detected | 63584.09* | 9 | 100 | 73 |
| Pse.40.1240 | OG004853 | hypothetical protein, conserved | Nicastrin | 69419.47* | 64 | 79 | 50 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Pse.45.1050 | OG005627 | hypothetical protein, conserved | Peroxisomal biogenesis factor 11 | 27903.7 | 73 | 53 | 0 |
| Pse.514.1040 | OG004589 | hypothetical protein, conserved | None detected | 38153.95* | 0 | 100 | 100 |
| Pse.304.1010 | OG005866 | hypothetical protein, conserved | Yos1-like | 13662.93* | 100 | 100 | 100 |
| Pse.33.1130 | OG005527 | hypothetical protein, conserved | None detected | 24860.74* | 15 | 100 | 64 |
| Pse.352.1060 | OG001789 | ATP-dependent zinc metallopeptidase, Clan MA(E) Family M41, | ATPase family associated with various cellular activities and Peptidase family M41 | 92700.1 | 6 | 81 | 0 |
| Pse.25.1140 | OG006195 | hypothetical protein, conserved | None detected | 9657.23* | 0 | 100 | 100 |
| Pse.25.1170 | OG000107 | hypothetical protein, conserved | None detected | 62189.67* | 2 | 100 | 76 |
| Pse.270.1010 | OG000063 | surface antigen-like protein | None detected | 42437.2 | 37 | 40 | 69 |
| Pse.277.1010 | OG004362 | hypothetical protein, conserved | DUF1620 | 102844 | 13 | 93 | 71 |
| Pse.28.1250 | OG000056 | glucose transporter, putative | Sugar transporter | 31364.74* | 1 | 99 | 5 |
| Pse.179.1020 | OG003814 | hypothetical protein, conserved | None detected | 30966.67* | 0 | 88 | 100 |
| Pse.201.1030 | OG008538 | | None detected | 46150.26* | 0 | 100 | 75 |
| Pse.216.1030 | OG002872 | transmembrane protein, putative | None detected | Cannot be predicted | 69 | 63 | 88 |
| Pse.1.1110 | OG005230 | hypothetical protein, conserved | emp24/gp25L/p24 family/GOLD | 26636.18* | 0 | 100 | 93 |
| Pse.103.1080 | OG001975 | hypothetical protein, conserved | Sulfite exporter TauE/SafE | 52244.36* | 0 | 100 | 100 |
| Pse.1047.1010 | OG000233 | amino acid permease | Transmembrane amino acid transporter | 66811.9 | 68 | 95 | 0 |
| Pse.149.1010 | OG000135 | P-type H -ATPase, putative | Cation transporter, E1-E2 ATPase and Hydrolase | 104028 | 3 | 92 | 0 |
| Pse.1195.1000 | OG021011 | | Nucleoside transporter | 16602.99* | 100 | 100 | 100 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Pse.550.1020 | OG000852 | hypothetical protein, conserved | None detected | 9198.67* | 0 | 100 | 83 |
| Pse.556.1030 | OG000007 | receptor-type adenylate cyclase GRESAG 4 | Adenylate and Guanylate cyclase catalytic domain | 57325.7 | 8 | 62 | 48 |
| Pse.280.1070 | OG001999 | | Major facilitator superfamily x 2 | 59767.95 | 9 | 100 | 55 |
| Pse.29.1160 | OG002314 | hypothetical protein, conserved | Translocon-associated protein beta | 22289.2 | 0 | 94 | 100 |
| Pse.155.1000 | OG000142 | hypothetical protein, conserved | None detected | Cannot be predicted | 0 | 55 | 19 |
| Pse.167.1010 | OG021335 | | None detected | Cannot be predicted | 0 | 100 | 100 |

## 4.4 Results

Of the 11 proteins shortlisted for tagging by the nested PCR method, only one clonal cell line of *P. serpens* was obtained, this line showed cell surface expression of eYFP for the protein *Pse.222.1080.* Populations of transfectants were obtained for two further genes (*Pse.926.1000* and *Pse.264.1080*). For *Pse.926.1000* there were very few cells expressing eYFP and it was not possible to obtain a clonal cell line with uniform eYFP expression, the *Pse.264.1080* transfectants did not show surface localisation of eYFP.

Pse.222.1080 was the only protein for which a clonal cell line expressing surface localised N-terminal TY:eYFP:TY was obtained (Figure 4.4). Positive transfectants were obtained after 2 weeks and clones were obtained after ~2 weeks from seeding cultures in 96 well plates at <0.5 cells per well. Correct integration was confirmed by sequencing across the integration junction and the molecular weight of the fusion protein was identified by Western Blot (Figure 4.5A). Pse.222.1080 is predicted to weigh 20 kDa. The blot (Figure 4.5A) probed with BB2 antibody (anti-TY) (Bastin *et al.*, 1996) shows a band between 48 kDa and 63 kDa, consistent with the calculated molecular weight of 50.18 kDa that is predicted for the fusion protein. Figure 4.4B shows that the eYFP is localised to all three surface membrane domains: the cell body membrane, the flagellar pocket

and the flagellar membrane. Pse.222.1080 was found to be surface localised as a fusion product however, mislocalisation as a result of the addition of the YFP cannot be discounted despite the use of the native signal peptide. Although simultaneous concanavalin A labelling was not done, the pixel intensity plot (Figure 4.4 D) shows a profile consistent with the signal being localised at the plasma membrane. From the data provided it is not possible to determine whether the GFP is located on the intracellular or extracellular face of the membrane. This feature could be independently validated by an antibody based approach in the presence or absence of cell permeabilisation, or by immuno-electron microscopy (with anti-GFP or anti-TY in this instance). An alternative method of interrogation could have been to remove the proteins on the outside of the plasma membrane by using a trypsin shave or similar. If the signal from the eYFP was reduced or removed without disruption of the structure of the plasma membrane then this would be consistent with localisation on the external face of cell surface. However, there was insufficient time to complete these experiments and thus although the localisation is consistent with a localisation at the cell surface, it remains unknown whether the GFP resides on the inside or outside face of the cell surface membrane.

The predicted transmembrane helices and membrane topology from TMHMM (Krogh *et al.*, 2001) is confirmed by the peptide results from LC MS/MS as the lysine residue with a detected SS-biotin post translational modification is contained in a portion of the peptide predicted to be outside the cell.

The tagging of Pse.264.1080 with N-terminal TY:eYFP:TY produced a population of cells expressing eYFP, but microscopy showed that this signal was not localised to the cell surface membrane (Figure 4.7A). The tagging of Pse.926.1000 with N-terminal TY:eYFP:TY produced positive transfectants after drug selection but very few cells were seen to be expressing eYFP and attempts to clone the cell line expressing eYFP failed. To elucidate why the majority of cells in the blasticidin resistant population of cells transfected with the eYFP fusion construct of Pse.926.1000 did not express eYFP, blot

validation was carried out on non-eYFP expressing clones. Validation by PCR indicated the correct insertion of the tagging construct (Figure 4.5C). For *Pse.926.1000* the validation primer combinations of 1+3 and 4+2 each anneal with one of a primer pair outside the integration site and the other of the primer pair inside the integrated DNA. In this case PCR generated fragments of the expected size of 722 and 776 respectively as predicted. However, the blot confirmed that there was no expression of TY:eYFP:TY (Figure 4.5A). This was further investigated by immunofluorescence with the BB2 antibody and anti-mouse TRITC conjugated secondary antibody (Figure 4.6). This confirmed that there is no expression either of the TY epitope tag or eYFP in the blasticidin resistant clone obtained from the transfection despite PCR showing the correct integration of the tagging construct.

**Table 4.3 Transfection attempts for nested PCR tagging and single knock out (KO) in *P. serpens***

Number of transfection attempts listed, whether any viable cells. If no transfection attempt is given nested PCR failed to produce full length construct. Y indicates whether viable cells of *P. serpens* were recovered following selection with antibiotics. Y or N also indicates whether cell lines were validated by PCR or Western blot.

| | YFP tag | | | | Single KO | | |
|---|---|---|---|---|---|---|---|
| | Transfected | Viable Cells | PCR validated | Western Validated | Transfected | Viable cells | PCR validated |
| **Pse.1124.1000** | 0 | | | | | | |
| **Pse.222.1080** | 2 | Y | 4 and 2 primer pair only | Y | 2 | y | |
| **Pse.352.1060** | 1 | | | | 2 | | |
| **Pse.422.1010** | 1 | | | | | | |
| **Pse.604.1040** | | | | | | | |
| **Pse.677.1000** | 1 | | | | 2 | | |
| **Pse.422.1030** | 1 | | | | | | |
| **Pse.270.1010** | | | | | | | |
| **Pse.926.1000** | 2 | Y | Y | N | 2 | y | |
| **Pse.264.1080** | 1 | Y | | | | | |
| **Pse.19.1260** | 2 | | | | 1 | | |
| **Pse.777.1000** | | | | | | | |

Of the 38 proteins shortlisted for tagging with the long primer approach in Table 4.1, full constructs were obtained by PCR for 25 proteins, one transfection attempt was made for each of the 25 proteisn but no viable cells were recovered following selection with antibiotics.

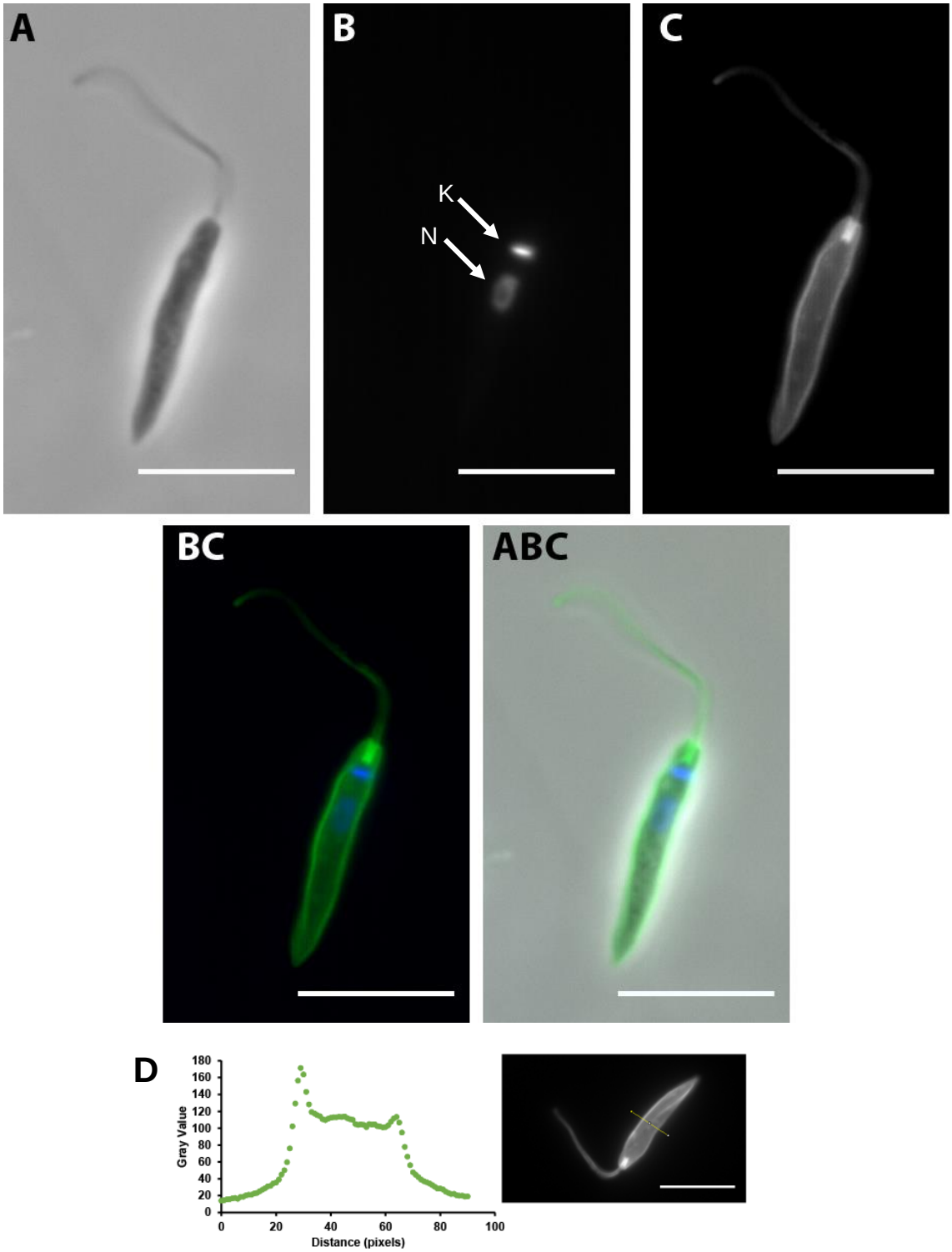



***Figure 4.4*** **Surface localisation of eYFP fusion of Pse.222.1080**

Native eYFP fluorescence for Pse.222.1080 tagged with N-terminal TY:eYFP:TY. **A)** Phase **B)** eYFP native fluorescence **C)** Nuclear DNA (N) and kinetoplast (K) stained with Hoechst and **BC)** and **ABC)** merge of channels **D)** A pixel intensity plot of the eYFP native fluorescence across of a *P. serpens* cell. Scale bar represents 10 µm.

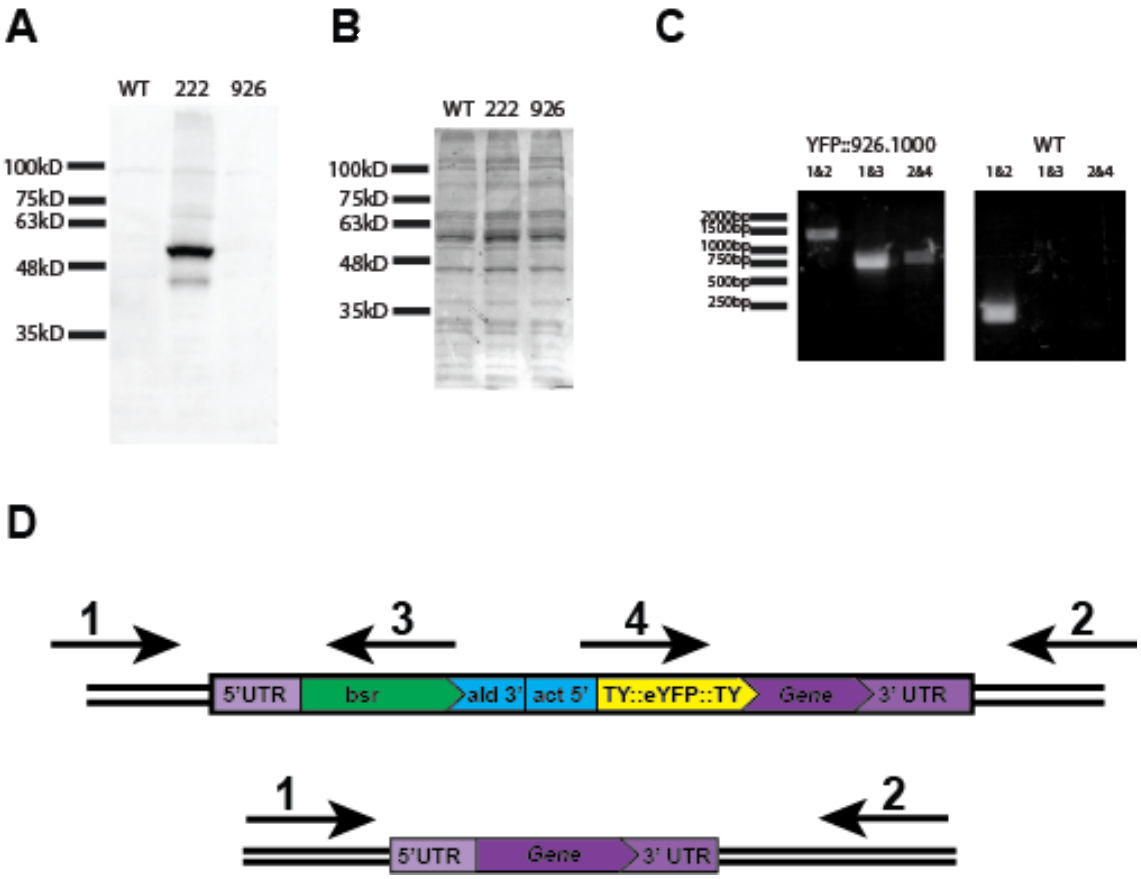



**Table 4.4 Validation primers for Pse.926.1000 tagged with YFP**
Primer sequences are 5'>3' and Position given is relative to the start codon of Pse.926.1000. Positions are given for the WT and 926.1000 YFP tagged genotype.

| | Primer sequence | Position | | Primer Sequence | Position | Size expected |
|---|---|---|---|---|---|---|
| **1** | attccagctag acgcgagtaa | -2210 in YFP 660 in WT | 2 | gcaacgaaacgagtga gaaa | 674 | 1334 in WT<br>2884 in YFP |
| **1** | attccagctag acgcgagtaa | -2210 in YFP | 3 | TTGATTGTAGCC GTTGCTC | -1507 | 722 in YFP |
| **4** | GTCCTGC TGGAGTT CGTGA | -102 | 2 | gcaacgaaacgagtga gaaa | 674 | 776 |

***Figure 4.5*** **Validation of clonal cell lines**

**A)** Western Blot of wild type (WT) and clonal cell lines of eYFP fusion of Pse.222.1080 probed with BB2 primary antibody. **B)** Equivalent SDS-PAGE to show protein loading. **C)** PCR validation of eYFP fusion of *Pse.926.1000*. **D)** Schematic diagram to show location of different primer pairs for PCR validation for N terminal tag (upper) and WT locus (lower). Primer sequences, positions and fragment sizes are detailed in Table 4.4

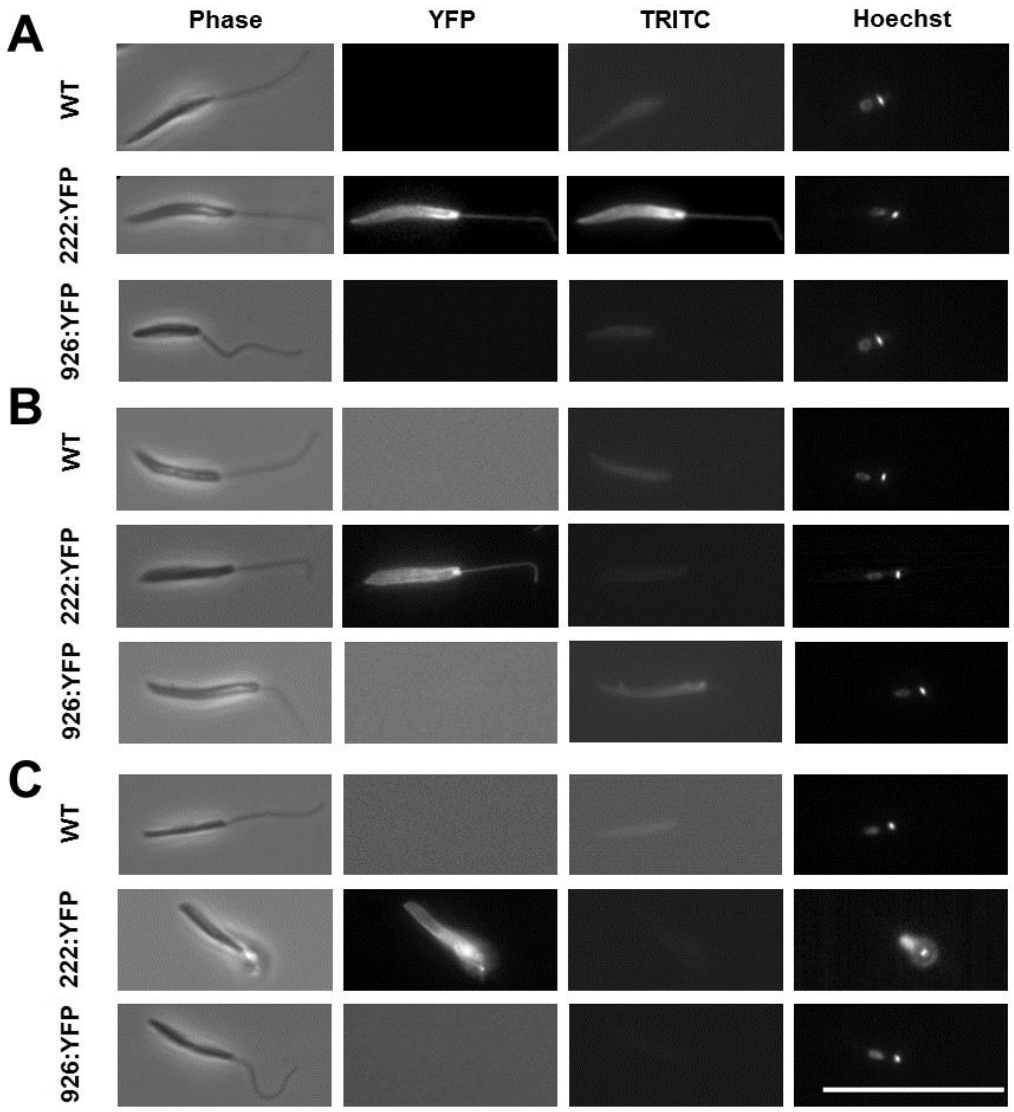



***Figure 4.6*** **Immunofluorescence of** ***P. serpens***

**A)** Immunofluorescent labelling using anti-TY antibody BB2 with TRITC labelled secondary showing signal in the eYFP and TRITC channels for eYFP fusion of Pse.222.1080 and no signal in either channel for eYFP fusion of Pse.926.1000. **B)** The same as A, with omission of primary antibody to show no signal in the TRITC channel for any cell line. **C)** The same as A, with omission of the secondary antibody to show no signal in the TRITC channel for any cell line. Scale bar represents 10 µm.

### 4.3.1 The use of eYFP expressing *P. serpens* cells for monitoring infection in *Solanum lycopersicun*

The cells obtained from tagging Pse.264.1080 with TY:eYFP did not show cell surface localisation of eYFP. However, the native fluorescence was very bright and easily visible by fluorescence microscopy. For this reason the cells were used for an experimental infection of tomatoes (*Solanum lycopersicun*) in order to determine if (genetically modified) parasites could be isolated following the inoculation of fruit. Ripe tomatoes were used because they provide better conditions for the parasites as the levels of the glycoalkoid tomatine and its aglycone tomatidine are reduced compared to non-ripe fruit. These defence compounds have been shown to inhibit the growth of *P. serpens* in axenic cultures (Medina *et al.*, 2015). Tomato infection is described in section 2.7. Two experiments were performed: a time course and a tomato variety test. The time course of infection was performed first. This was attempted with Cherry tomatoes (country of origin: Morocco) because they were small and therefore amenable to the large numbers of fruit required for the experiment. Time points were 0, 24, 48, 72, 96, 168, 192, 216, 240, 264h. 6 fruit were inoculated per time point, with 0.5ml of *P. serpens* cells expressing *YFP* or control WT cells were inoculated at 2 x $10^7$ cells $ml^{-1}$. The parasites were difficult to recover from the fruit and it was challenging to identify *P. serpens* from the fruit sap which contained large numbers of bacterial and fungal cells in addition to plant matter. It was not possible to complete the time course as despite sterilising the surface of the tomatoes with 70% ethanol all fruit developed fungal infection and rotted after incubation at 28˚C. As the surface of the fruit were sterilised with 70% ethanol it was likely that these infections were originating from bacteria and fungi inside the fruit. Gamma irradiation as a method of sterilising these infections was pursued however it was concluded that the irradiation time required would be far too long.

A separate experiment attempted to evaluate which tomato varieties were optimal for *P. serpens* growth: 5 fruit each of Cherry (Morocco), Plum (Netherlands) and Vine (Spain)

and 2 Beefsteak (Morocco) were inoculated with: 0.5 ml Cherry, 1 ml Vine and Plum, 1.5 ml of *YFP* expressing *P. serpens* cells at 2 x $10^7$ cells $ml^{-1}$. Fungal infection and fruit rotting was also observed in this experiment. It was possible to isolate live *P. serpens* cells from Beefsteak and Plum tomatoes from 5 days post inoculation. Whole cells were recovered from Cherry tomatoes but in this instance were not live. After 7 days, yellow spots had appeared on the surface of the Plum variety of tomato fruit. These discoloured lesions are likely because of the ethanol produced during the fermentative metabolism of *P. serpens* (Ienne *et al.*, 2012) disrupting the red pigmentation. Fruit were dissected at the discoloured site which was distantly located from the site of innoculation (Figure 4.7C). Fruit sap extracted from these discoloured areas was examined by live cell microscopy and is shown in Figure 4.7. The *P. serpens* cells found in the fruit sap are still expressing *eYFP* and the cells appeared to be promastigote in morphology (Figure 4.7B), though this was not evaluated morphometrically. The nucleus was present but not readily visible due to high levels of background fluorescence. No 2k2n or 2 flagella cells were imaged or observed directly so it is not possible to definitively state whether or not *P. serpens* cells were dividing in the tomato fruit in this experiment.

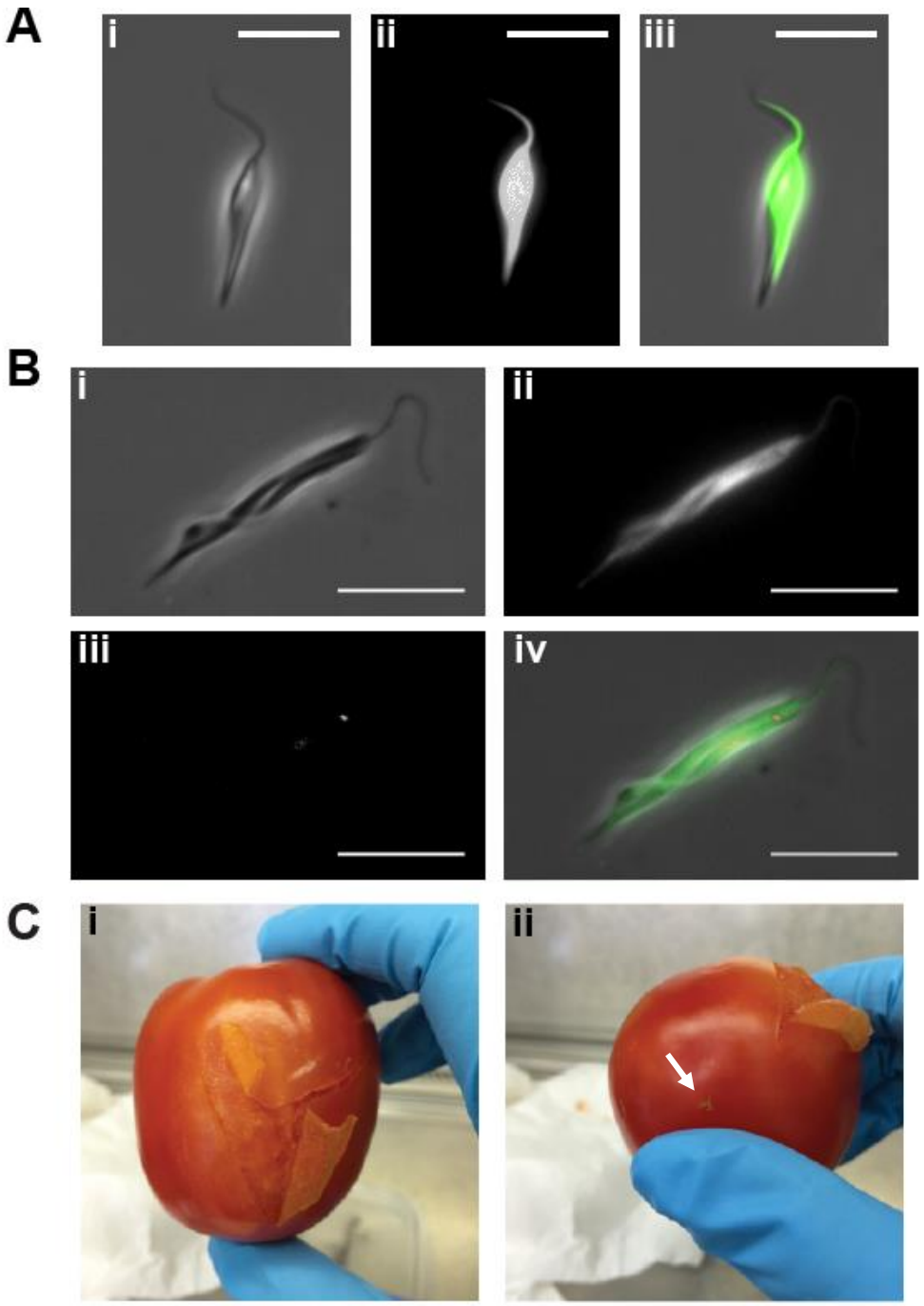



***Figure 4.7*** **The infection of tomato fruit with *P.serpens* cells expressing *eYFP***

**A)** live cell microscopy of cultured cells **i)** phase **ii)** eYFP **iii)** merge of channels. **B)** live cell microscopy of the cells recovered from the tomato fruit **i)** phase **ii)** eYFP **iii)** DNA stained with Hoechst and **iv)** merge of channels. **C)** infected tomato **i)** yellow spotting visible underneath skin at site where *Phytomonas* were recovered **ii)** The site of inoculation at the base of the tomato as indicated by the arrow.

## 4.5 Discussion

Presented here is the first example of tagging an endogenous gene with eYFP in *Phytomonas*. It is possible that this protein is trafficked to the plasma membrane via a non-cleaved signal-anchor (Figure 4.1) and thus the addition of the N-terminal tag did not interfere with cell surface localisation. However, the tagging of the paralog Pse.264.1080 using the same strategy did not result in cell surface localisation of eYFP. Both proteins show high levels of amino acid sequence conservation with other members of OG000024, including at the N-terminus. Indeed, the N-terminus is well conserved across the orthogroup. However, only 30% of the proteins in this orthogroup are predicted to encode a signal peptide. It may be that these proteins are targeted to the membrane by a hydrophobic signal-anchor sequence at the N-terminus or that the methods for predicting signal peptides were not effective in this instance. However, this assumes that the gene predictions are correct. The gene prediction software may not have predicted the correct start codon for each coding sequence and thus some gene models may be erroneously truncated. This problem of gene model truncation has previously been observed to affect ~14% of the predicted proteome in *Leishmania mexicana* (Fiebig et al 2015). Given more time and resources, further work would benefit from refining the gene models using RNAseq data.

Tagging may disrupt protein localisation by disrupting signal sequence and protein size. In this example the tag is inserted at the endogenous N-terminus of the Pse.222.1080 coding sequence therefore utilising the native sequences and did not introduce signal peptides from other genes. The eYFP fusion of Pse.222.1080 was observed to localise to all 3 surface membrane domains. BLAST searches and investigation of orthologs in *Leishmania* show Pse.222.1080 is an amastin-like gene. The phylogeny in Figure 4.8 shows that Pse.222.1080 is a member of the δ amastin sub-family and that *P. HART1, P. serpens* and *P. EM1* also contain members of the α amastin sub-family. A member of the δ amastin sub-family was found to be localised to the *T. cruzi* amastigote cell surface

by cryoimmuno-gold (Teixeira *et al.*, 1994). Subsequently, GFP fusion proteins of members of the β and δ sub-families were localised at the *T. cruzi* epimastigote (insect stage) cell surface (Kangussu-Marcolino *et al.*, 2013). Transcripts of δ amastins were upregulated in the intracellular stage of *T. cruzi,* while β amastins are more abundant in epimastigotes (Kangussu-Marcolino *et al.*, 2013). Here, a δ amastin eYFP fusion protein is shown to localise to the plasma membrane of promastigote cells of the plant parasite *P. serpens.* This is noteworthy as, with the exception of the β amastins above, amastins show stage specific expression in amastigote forms of *T. cruzi* (Kangussu-Marcolino *et al.*, 2013) and *L. donovani* (Wu *et al.*, 2000). Pse222.1080 is a member of an orthogroup of amastin-like proteins (OG000024 in Appendix 3.3) shown in darker blue in Figure 4.8. There are 7 identified sequences in *P. serpens*, 2 each in *P. francai* and *P. EM1*, there were no members of this orthogroup identified in *P. HART1*. These are large multicopy gene families in *L. major* (46 genes), *L. Mexicana* (55), *L. infantum* (33) and *L. braziliensis* (40). Expression of *L. major* ortholog *LmjF.34.0980* is downregulated in PNA - Metacyclic Promastigote vs. Early Log Procyclic Promastigote (0.388 fold change) and downregulated in Lesion Derived Amastigote vs. Early Log Procyclic Promastigote (-0.081 fold change) (trytrypdb.org). It will be interesting to see if the amastin genes in *Phytomonas* are under developmental regulation.

Without localisation data for a greater number of putative cell surface proteins it is not possible to validate the datasets presented in Chapter 3. Optimisation is needed both of the transfection protocols for *Phytomonas* and investigation into the optimal length of homology for a greater percentage of correct transfectants to increase transfection efficiency is required.

As there was a bottleneck in obtaining PCR constructs to tag genes of interest my efforts to optimise focused on improving the PCR. I began with 12 candidate genes for generating knock outs and YFP tagged lines but following high attrition with the nested PCR I focused my efforts on attempting to generate YFP cell lines to validate the

candidates obtained in Chapter 3. With hindsight it would have been prudent to have optimised the nested PCR prior to the selection of candidate proteins for tagging with this method. With foreknowledge of these challenges I would have attempted more traditional tagging methods that would not have the benefit of allowing high throughput tagging but may have yielded more robust results for a smaller number of proteins of interest.

The method described in 2.6.7 used nested PCR to generate the tagging construct and BTX electroporator to transfect cells. This method was chosen because it has been successful in *Trypanosoma* and *Leishmania* and in *P. francai* (Butler, personal communication*)* The method in 2.6.8 describes using long primer PCR to produce the tagging construct and was modified from a method successful in the high throughput tagging of *T. brucei* (Dean *et al.*, 2015)*.* This method was adapted for *P. serpens* to use a BioRad Gene Pulser for transfection as used by Miranda and colleagues (2015) instead of the Amaxer used by Dean and colleagues.

Using the BTX electroporator, two variations of the method described in 2.6.7 were used for transfected cell selection: (a) recovering populations of resistant cells in a single culture for 12h followed by drug selection in a single culture then once a population of cells was obtained these were diluted to 0.5 cells per well of a 96-well plate and cultured to yield clonal cell lines and (b) recovering for 12 hours in a single culture and performing drug resistance and clonal selection simultaneously at a density of 0.5 cells per well in a 96-well plate. Success was only achieved with variant (a). Only one attempt of 2.6.8 was attempted due to time limitations. There are a number of possible reasons why the electroporation method yielded such a low frequency of success. The methods used to transfect *P. serpens* may have been inefficient and therefore resulted in insufficient numbers of cells taking up the tagging construct. Due to limitations on time there was little opportunity to optimise electroporation conditions (voltage, pulse length, pulse length, interval time, number of pulses, number of cells, and concentration of DNA etc)

and different ‘cytomix’ buffer conditions, thus all of these factors are candidates for optimisation. Further possible causes of low transfection efficiency include factors relating to homologous recombination in *Phytomonas*. It may be that homologous recombination in *Phytomonas* requires longer homologous regions of DNA to direct effective homologous recombination, or it may be that this process in *Phytomonas* is not as precisely controlled as it is in *T. brucei* and mis-integration may be common. A possible way to troubleshoot and optimise the transfection protocol would be to use an episome expressing fluorescent protein. Maintenance of an episome has been shown in a different species of *Phytomonas* (Miranda *et al.*, 2015).The episome, if maintained by *P. serpens,* would not need to integrate into the genome by homologous recombination and cytoplasmic YFP would make it possible to follow transfection and determine whether electroporation needed optimising (i.e. DNA was not getting in to cells) or whether the PCR product needed optimising (i.e. DNA not integrating into the genome and length of DNA overlap used to direct recombination needed optimising). Optimisation of fluorescent protein used should also be explored. Furthermore, it isn’t known what effect the negative charge of cells may have on the interaction of DNA and the diffusion of DNA into a cell.

Overall, the data presented here suggests that a number of the proteins identified by the biochemical and bioinformatical approach presented in this thesis are surface proteins in *Phytomonas*, this work has also demonstrated that it is possible to create eYFP fusion proteins to attempt to validate the localisation of these proteins at the plasma membrane, however examples of this are limited. The dataset is not conclusive but represents a leap forward in our understanding of *Phytomonas*. Other possible validation strategies would have been to try to raise antibodies to recombinant proteins, and to tag the orthologous genes of the predicted surface proteins in *Trypanosoma* or *Leishmania*. The advantage of tagging orthologous genes in these sister taxa is that they are more genetically tractable and high throughput methods for tagging genes with fusion proteins are well

established. Tagging orthologous genes in different species would also allow the investigation of whether conserved genes have conserved localisation at the plasma membrane.

Also presented in this chapter is the first experimental infection of tomato fruit with transgenic *Phytomonas serpens* cells. These fluorescent cells may facilitate the development of an infection assay that allows the progress of infection to be monitored using fluorescence.

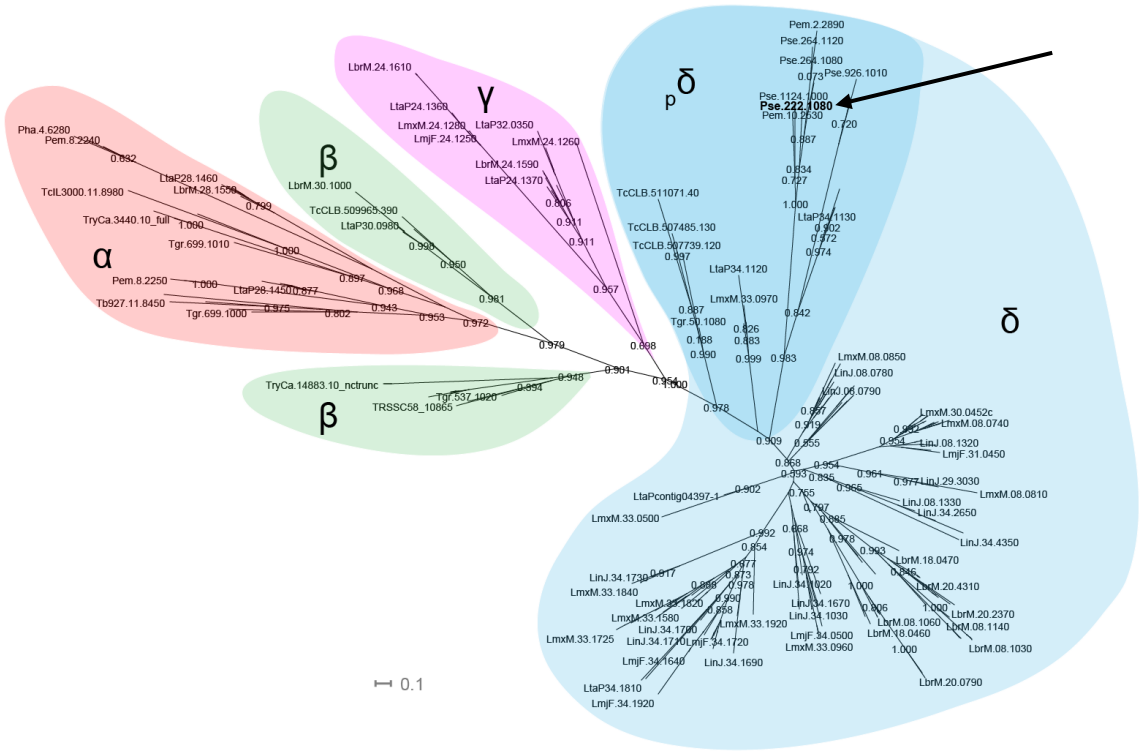



***Figure 4.8*** **A phylogeny of the amastin surface glycoproteins in trypanosomatids**

Maximum likelihood protein sequence phylogenetic tree of the amastin proteins in 16 trypanosomatid species including *Phytomonas*, *Leishmania* and *Trypansoma*. Sequences were aligned using MergeAlign (Collingridge and Kelly, 2012) and tree built using FastTree (Price et al., 2010) all under default settings. Clades are shaded by subfamily in accordance with Jackson (2010). *P. serpens* sequence that was successfully validated as Pse.222.1080 is identified by the arrow.

## 4.6 Conclusions

The validation of the putative cell surface proteins identified in Chapter 3 yielded the first example of endogenous gene tagging in the plant parasite *P. serpens.* However, attempts to validate other putative cell surface proteins of *P. serpens* were not successful. A large number of methods exist for gene tagging in *Trypanosoma* and *T. brucei* is a well-developed model system. However, the use of *Phytomonas* species as model systems for investigating protein localisation and function requires significant further development before levels of success equivalent to that of *T. brucei* are obtained.

# *Chapter 5: The* Trypanosoma carassii *transcriptome and predicted cell surface proteome*

## 5.1 Preface

The parasite cell surface plays many key roles in host-parasite interactions, including parasite protection, transmission, initiation and maintenance of infection. Trypanosomatid parasites have evolved to colonise a broad range of hosts from across the plant and animal kingdoms. Previous comparisons between *Trypanosoma cruzi* and *Trypanosoma brucei* have shown that the cell-surface proteomes of different species are largely taxon specific (Jackson *et al.*, 2013). Given this specificity and the fact that information on the composition of the cell surface is only available for a limited number of species, little is known about the evolution or composition of the cell surface in the vast majority of species in this group. This chapter presents a transcriptomic and proteomic analysis of the cell surface of the leech-transmitted extracellular parasite of carp, *Trypanosoma carassii*. This parasite was selected due to its position as an out-group to both *T. brucei* and *T. cruzi,* and thus information gained here has the potential to provide further insight into the ancestral surface proteome of these globally important parasites as well as the origins and evolution of antigenic variation in *Trypanosoma brucei.*

Following the optimisation of the biotinylation protocol in *P. serpens,* I had the opportunity to work on *T. carassii*. The analysis presented here is similar to that in Chapter 3. The process of isolating the parasites from carp is extremely difficult and due to the number of parasites isolated and the constraints of only being able to make a single trip to the Netherlands there was only enough material for one sample. Due to limited material I did not perform blots as in Chapter 3 but rather processed everything for mass spectrometry to try and maximise the information acquired from the available material.

## 5.2 Introduction and background

The trypanosomatids are a group of single-celled eukaryotes that encompass numerous important pathogens of crops, livestock and people. Though there is thought to have been a single evolution of parasitism within this large group (Simpson *et al.*, 2006), there have been repeated adaptations to intracellular and extracellular lifestyles that have resulted in a concomitant diverse array of immune evasion strategies (Gupta *et al.*, 2013; Higgins and Carrington, 2014; Manna *et al.*, 2013). In Sub-Saharan Africa, one species of trypanosomatid, *Trypanosoma brucei*, causes human African trypanosomiasis. The entire life cycle of *T. brucei* occurs extracellularly, and within the human host the parasites reside both the bloodstream and lymphatic system (http://www.cdc.gov/parasites/sleepingsickness/biology.html). This exposed lifestyle necessitates that *T. brucei* is continually challenged by both innate and adaptive immune responses and thus has evolved mechanisms to avoid elimination.

In the mammalian host the *T. brucei* cell surface is concealed beneath a densely packed coat of identical GPI-anchored variant surface glycoprotein (VSG) (Barry and McCulloch, 2001; Donelson, 2003; Field and Carrington, 2009; Pays *et al.*, 2004). Though this uniform proteinaceous coat may shield some surface antigens from antibody recognition (Borst and Cross, 1982; Mehlert *et al.*, 2002), its primary function is in immune evasion *via* antigenic variation; the process by which an individual cell changes its expressed VSG gene, and hence presents an entirely novel surface to the host immune system. It is thought that additional immune evasion is also provided by clathrin-dependent endocytosis and degradation of antibodies bound to surface proteins (Allen *et al.*, 2003; Field and Carrington, 2009; Pal *et al.*, 2003). However, this additional safeguard is thought to be beneficial to the parasite primarily for clearance of antibodies bound to invariant surface proteins which are not changed by antigenic variation.

Antigenic variation underpins the success of *T. brucei* as an extracellular parasite of the mammalian bloodstream as it facilitates escape from immune clearance and thus

regeneration of parasite populations and prolongation of infection. Though the molecular mechanisms of antigenic variation are now well characterised, and the evolution of antigenic diversity within the VSG repertoire of individual salivarian trypanosome species is understood (Jackson *et al.*, 2012), little is known of how this immune evasion strategy evolved. Moreover, in addition to VSG there are also invariant proteins that localise to the *T. brucei* cell surface. Until recently, few of these invariant surface proteins had been identified, however novel insight into these proteins was provided by a proteomic investigation of the *T. brucei* cell surface (Gadelha *et al.*, 2015). This study identified several flagellar pocket proteins in addition to many proteins whose functions and localizations were previously unknown.

To provide further insight into the evolution of the trypanosomatid cell surface we have sequenced the transcriptome and performed cell surface proteomics on the bloodstream form of *Trypanosoma carassii*. *T. carassii* is a kinetoplastid parasite of cyprinid fish such as carp (*Cyprinus carpio L.*) as well as some noncyprinids (Overath *et al.*, 1998). Similar to *Trypanosoma brucei*, *T. carassii* lives extracellularly in the blood and tissue fluids of the fish (Figure 5.1) (Haag *et al.*, 1998). This life cycle necessitates that *T. carassii* is subject to attack from both innate and adaptive immune responses of the fish host. However, there is no evidence that antigenic variation occurs in this parasite and the mechanism by which this parasite is able to achieve chronic infections is unknown.

Although *T. carassii* lives extracellularly like salivarian trypanosomes, it diverged from the trypanosome lineage before the divergence of the stercorarian and salivarian trypanosomes (Figure 5.2). The stercorarian trypanosomes, of which *T.cruzi*, *T, rangeli*, *T. grayi* and *T. theileri* are included in this chapter, are transmitted by contact with contaminated faeces of hematophagous flies. The salivarian trypanosomes, of which *T. brucei*, *T. vivax, T. congolense* are included in this chapter, are transmitted directly during feeding by hematophagous insects. As *T. carassii* diverged from the trypanosome lienage prior to the emergence of the salivarian or stercorarian clades it is ideally placed in the

phylogenetic tree of trypanosomatids to improve our knowledge of how these different groups evolved taxon-specific cell surfaces specialised for their different parasitic lifestyles. Specifically, analysis of the gene complement and surface proteome of *T. carassii* has the potential to provide insight into the cell surface proteome of the last common ancestor of the stercorarian and salivarian trypanosomes enabling comparative analysis of gene gain and loss in these lineages to be conducted.

The innate immune system of teleost fish (including carp) is highly similar to that present in mammals and the adaptive immune system is also similar with immunoglobulins, thymus gland and some immune memory (Magnadóttir, 2006; Magor and Magor, 2001; Sunyer, 2013). However, the mechanism by which *T. carassii* evades the immune system of carp is unknown (Haag *et al.*, 1998). This makes it an appealing candidate upon which to apply the biochemical and bioinformatics toolkits developed in Chapter 4. Infections occur in both wild and farmed fish populations, in aquaculture prevalence of the parasite may be high, but parasitaemia is often low (Lom and Dyková, 1992) and infection is rarely lethal to the common carp (Ruszczyk *et al.*, 2008). The parasites used here were isolated directly from carp and therefore transcripts and surface proteins that are expressed in this stage of the parasite life cycle are those that have the capacity to be detected. Those transcripts and proteins that are expressed specifically in life cycle stages that are adapted to the leech host are not expected to be present in this analysis. Similarly, given that the *T. carassii* lives extracellulary in the bloodstream of the fish host it is possible that putative cell surface proteins similar to those found in the bloodstream form of *T. brucei,* such as orthologs of VSG or cell surface receptors, will be found. However, given the rapid sequence evolution of gene families associated with immune evasion it is possible sequence divergence may be too large to facilitate bioinformatic inference of orthologs. The other possibility is that there may be similarities that have occurred as the result of convergent evolution in *T. carassii* and salivarian trypanosomes. For example, if cell surface proteins in *T. carassii* are involved in immune evasion in a similar way to VSG then it would be expected that

there would be a large family of such sequences, that they would have high primary sequence divergence, and that they may have evolved a similar mechanism of attachment to the cell surface and similar secondary structure to facilitate packing on the cell surface. e.g. GPI anchored and alpha helical, densely packed protein surface coat (Cohen *et al.*, 1984) .

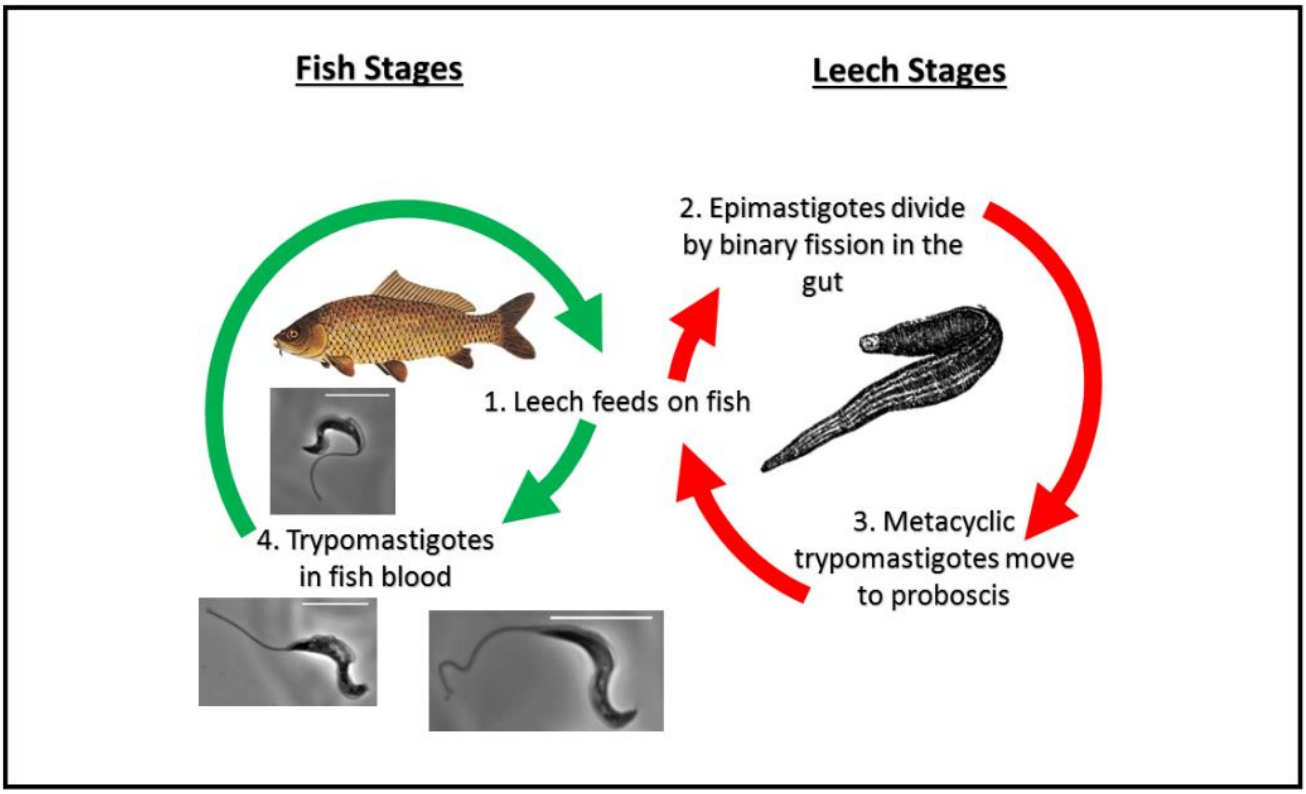



***Figure 5.1*** **The life cycle of *T. carassii***

The main events of what is known of the *T. carassii* life cycle are summarised here. When the leech feeds on a fish (1), trypomastigotes forms of *T. carassii* are ingested as part of the blood meal. Once inside the leech, the parasite cells proliferate as epimastigotes in the cop (2). The cop is a distention of the alimentary canal in leeches. From the cop, non-dividing trypomastigotes move to the proboscis of the leach (3) and can be transferred to the fish host from here (Overath *et al.*, 1998). Once in the fish bloodstream (4) the parasites maintain the trypomastigote morphology (Overath *et al.*, 1998). Parasitemia in the fish host are often low and persistent (Overath *et al.*, 1998) and mortality rates vary between host fish species (Ruszczyk *et al.*, 2008). The *T. carassii* images are my own. Scale bars 10 µm. Carp image from http://www.glerl.noaa.gov/ and leech from Brockhaus Kleines Konversationslexikon 1906, both distributed under the creative commons license.

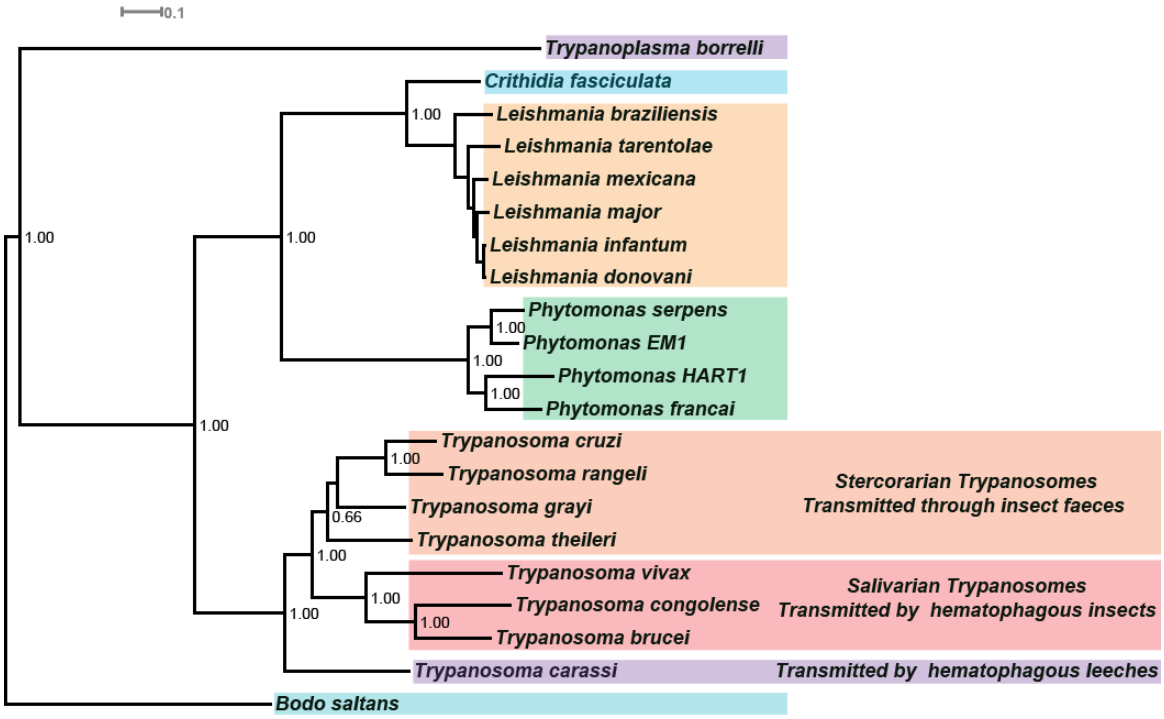



***Figure 5.2*** **A phylogenetic tree of the trypanosomatids**

A maximum likelihood protein sequence phylogeny shows that *T. carassii* diverged before the split between the salivarian and stercorarian trypanosomes. The phylogeny was constructed by Dr Kelly using orthologous gene groups containing only single copy genes aligned with MergeAlign (Collingridge and Kelly, 2012) and phylogenetic tree inferred using RAxML (Stamatakis, 2006) as described in Methods section 2.4.6. and the tree was rooted on *Bodo saltans*.

As *T. carassii* diverged before the emergence of the stercorarian and salivarian trypanosomes it is also possible that the surface proteome and immune evasion strategy has some commonality with that of the sister group containing *Leishmania*. In contrast to *Trypanosoma*, the cell surface of *Leishmania* parasites is characterised by a non-protein cell surface which is densely packed with molecules of lipophosphoglycan (LPG) and glycoinositolphospholipid (GIPL) (de Assis, Ibraim, Nogueira, Soares, and Turco, 2012). During the intracellular stage species-specific proteins are also expressed on the cell surface of *Leishmania* amastigotes to complement the LPG/GIPL coat, these are multi-copy and often species-specific (Jackson, 2010). These include the transmembrane glycoproteins encoded by the amastin genes (Jackson, 2010). While the sub-families of α and β amastin genes are conserved in *Leishmania* and *Trypanosoma* species, the repertoire expanded in *Leishmania* and the subfamily δ-Amastin is highly diverse in this lineage (Jackson, 2010). Amastins are linked to another transmembrane glycoprotein of unknown function, tuzins (Ivens *et al.*, 2005; Jackson, 2010). Both amastins and tuzins are hypothesised to be involved in the adaptation to the intracellular invertebrate stage. In addition to functioning in parasite development, pathogenicity and transmission these cell surface molecules often have immunogenic roles because of their prominent position at the vanguard of the host-parasite interface (MacGregor *et al.*, 2012).

While *Trypanosoma cruzi* and *Leishmania* species escape host immune responses by living intracellularly, *Trypanosoma carassii* is known to be extracellular in the fish host. However, to date there is no evidence that antigenic variation occurs in this parasite so the mechanism by which this species is able to achieve chronic infections is unknown. Previously identified components of the *T. carassii* cell surface include a GPI anchored mucin-like glycoprotein expressed at $6.0 \pm 1.7 \times 10^6$ molecules per cell (Campo *et al.*, 2002). This cell surface molecule was not observed to undergo antigenic variation (as observed for VSG in *T. brucei)* and was shown to be a target of the fish humoral immune response. The authors hypothesise that the high incidence of chronic infections by *T. carassii* in

nature and in aquaculture is the result of an ineffective host immune response (Campo *et al.*, 2002). *Trans*-sialidase like proteins were also identified (Campo *et al.*, 2002) and recombinant *T. carassii* GP63 was shown to attenuate anti-microbial responses of goldfish monocytes and macrophages (Oladiran and Belosevic, 2012).

Within the Trypanosomatids it is believed that the genes coding for cell surface proteins are predominantly lineage specific. However, previous studies have concentrated on a limited number of species. To further explore this part of trypanosome biology we undertook a genetic and proteomic study of the extracellular parasite of cyprinid fish, *Trypanosoma carassii.* This species is ideally suited to provide insight into the lineage specificity of the trypanosomatid surface as it diverged from the trypanosomatid lineage prior to the emergence of the stercorarian and salivarian trypanosomes. In this chapter I show that 80 surface protein families previously thought to be unique to stercorarian trypanosomes were present in the common ancestor of the stercorarian and salivarian clades. Similarly 2 surface protein families previously thought to be unique to salivarian trypanosomes were also present in the common ancestor. Taken together these findings reveal that, in addition to the evolution of lineage specific protein components, the trypanosomatid cell surface has also evolved by loss of ancestral components. In addition to these findings I identify two novel large families of GPI anchored cell surface proteins that are unique to *T. carassii.* I propose that these act in immune evasion in these parasites.

## 5.3 Results

### 5.3.1 *Sequencing of the* T. carassii *transcriptome and clustering of trypanosomatid orthogroups.*

*T. carassii* cells were isolated from the bloodstream of carp and mRNA was extracted, sequenced and assembled (see sections 2.2.3 and 2.3.2). The *T. carassii* transcriptome contains 13960 transcripts that encode open reading frames ≥100 amino acids. The predicted protein coding genes contained within these transcripts were subject to clustering using OrthoFinder (Emms and Kelly, 2015) with the bioinformatically predicted proteomes of 6 other trypanosomes including *T. brucei*, *T. vivax* and *T. cruzi*. The clustering resulted in the inference of 17608 orthogroups containing 70435 genes (excluding pseudogenes) of which 9625 are unclustered genes. A total of 7615 orthogroups were found in *T. carassii* containing 12346 genes of which 485 are unclustered (Appendix 5.2). In total *T. carassii* shared 7059 orthogroups with stercorarian trypanosomatids of which 834 (containing 3039 genes) are not found in any salivarian trypanosomes (Figure 5.3). Similarly *T. carassii* shared 41 orthogroups (containing 212 genes) with salivarian trypanosomes that are not found in any sequenced stercorarian trypanosomes (Figure 5.3). This differential presence of orthogroups indicates that there have been different patterns of gene loss in the two clades since they last shared a common ancestor. Moreover, these data imply that there has been larger number of gene families lost in the salivarian lineage than in the stercorarian lineage.

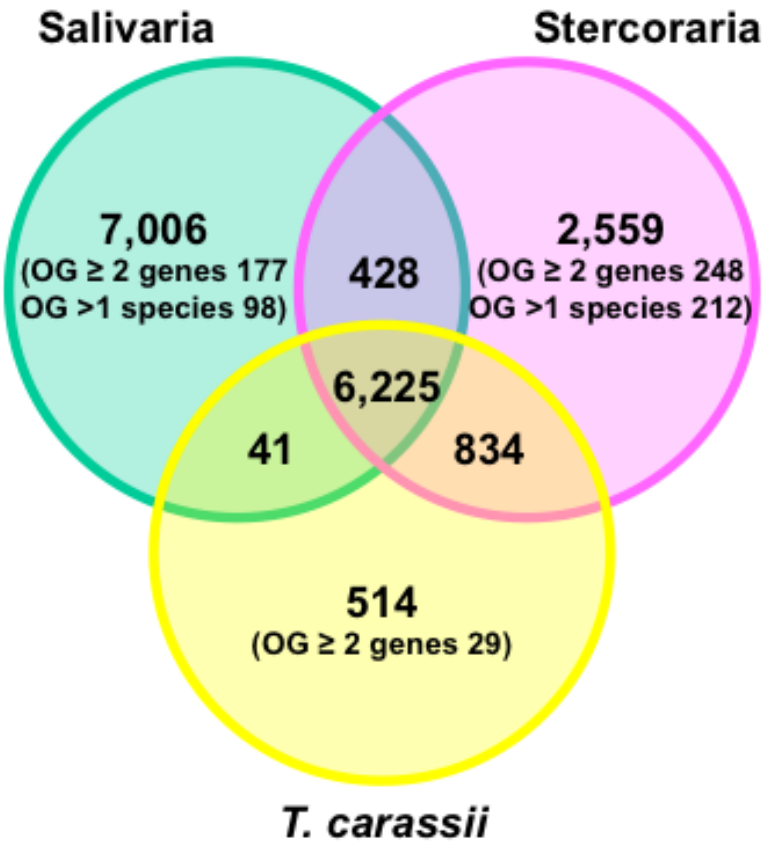



***Figure 5.3*** **The results of orthogroup clustering in *Trypanosoma***

A Venn diagram showing the relationship between the orthogroups of *T. carassii*, the stercorarian and salivarian trypanosomes. Also shown are the number of orthogroups (OG) containing a total of greater or equal to two genes (OG ≥2 genes) and the number of orthogroups in more than one species (OG > 1 species). The larger number of gene families shared between *T. carassii* and the stercorarian trypanosomes suggest that up 834 gene families present in the ancestor of these trypanosomes were lost in the salivarian lineage. The full dataset is available in Appendix 5.2.

### *5.3.2 Prediction of orthogroups containing cell surface localised genes.*

To predict the cell surface proteomes of *T. carassii* and related trypanosomatid parasites of animals, the predicted protein sequences were searched for sequences encoding domains associated with membrane proteins. These comprise the signal peptide (SP), transmembrane domains (TMD) and the glycophosphotidylinositol (GPI) anchor. This method is analogous to that presented in Chapter 3 and is therefore subject to the same limitations. First, the presence of a transmembrane helix and a signal peptide does not guarantee that a protein will localise to the cell surface membrane. These features could also target a protein to a cell internal membrane such as the endoplasmic reticulum, secretory vesicle, or Golgi apparatus. Second, this screen will exclude surface localised proteins that are Type II transmembrane proteins or other transmembrane proteins where the first transmembrane domain is more than a few residues from the amino terminus (and therefore not detected by these computational predictive tools). Third, computational methods to predict GPI anchors, signal peptides and transmembrane domains are not perfect and can result in both false positive errors (i.e. prediction of features that are absent) and false negative errors (i.e. fail to detect *bona fide* features).

I acknowledge that the first two limitations described above are not addressed in this thesis or in previous computational analyses of putative cell surface proteins in trypanosomatids (Jackson *et al.* 2013). Thus in this chapter some cell surface proteins will be missed and some proteins that localise to cell internal membranes will be mistakenly identified as potential surface proteins. While these two limitations are not addressed, I have attempted to address the issue of computational prediction error. Specifically, to minimise the false positive and false negative error associated with computational prediction of signal peptides, transmembrane domains and GPI-anchors by analyzing the presence or absence of these features for all genes in each orthogroup was analysed. Given the limitations of these predictive algorithms, a family of genes was considered to be a surface

gene family if 40% of the constituent genes in the family were predicted to have either a signal peptide and transmembrane domain (SP & TMD) or a signal peptide and GPI anchor (SP & GPI). This cut-off value of 40% of gene-family members was informed by the observed values for the conserved surface glycoprotein gp63 gene family. In this gene family 48% of proteins were predicted to encode a SP & GPI (Appendix 5.3).

Where available I have included the annotation from *T. brucei* and *T.cruzi* orthologs in the results of the bioinformatic screens in the Appendices to this thesis. This provides some additional information and in some cases can confirm the presence of known surface proteins e.g. gp63. Potential next steps to parse this dataset would be to remove proteins which are predicted to contain a mitochondrial targeting sequence as identified by MitoFates or similar programme. Following this remaining sequences could be searched for Pfam domains, where present the type of Pfam domain would provide additional information to support or discard the hypothesis for cell surface localisation.

Application of these filtration criteria to the full set of 17608 orthogroups and unclustered genes identified a set of 1384 orthogroups containing 6488 genes (excluding pseudogenes) of which 975 are unclustered genes (7.9% of all orthogroups) that comprised putative cell surface localised proteins. 261 surface orthogroups and unclustered genes (19% of all surface orthogroups) are present in both the SP & TMD and SP & GPI categories of surface orthogroups (Figure 5.4 B). These 261 surface orthogroups include the mucin associated surface protein (MASP) of *T. cruzi* and *T. rangeli.* The presence of all three features is most likely caused by errors in predicting the correct N- and C-termini of proteins and limitations in the sensitivity and specificity of the feature prediction methods.

The transcriptome of *T. carassii* was found to contain transcripts encoding proteins that could be clustered into 353 different surface orthogroups containing 1454 genes of which 54 are unclustered. In total *T. carassii* shared 292 surface orthogroups (containing 2573 genes) with stercorarian trypanosomatids of which 80 (containing 928 genes) are not

found in any salivarian trypanosomes (Figure 5.4 A). Similarly *T. carassii* shared 2 surface orthogroups (containing 4 genes) with salivarian trypanosomes that are not found in the stercorarian trypanosomes (Figure 5.4 A).The results are summarised in Figure 5.4 and a table of the 212 pan-trypanosomatid surface orthogroups are listed in Appendix 5.1.

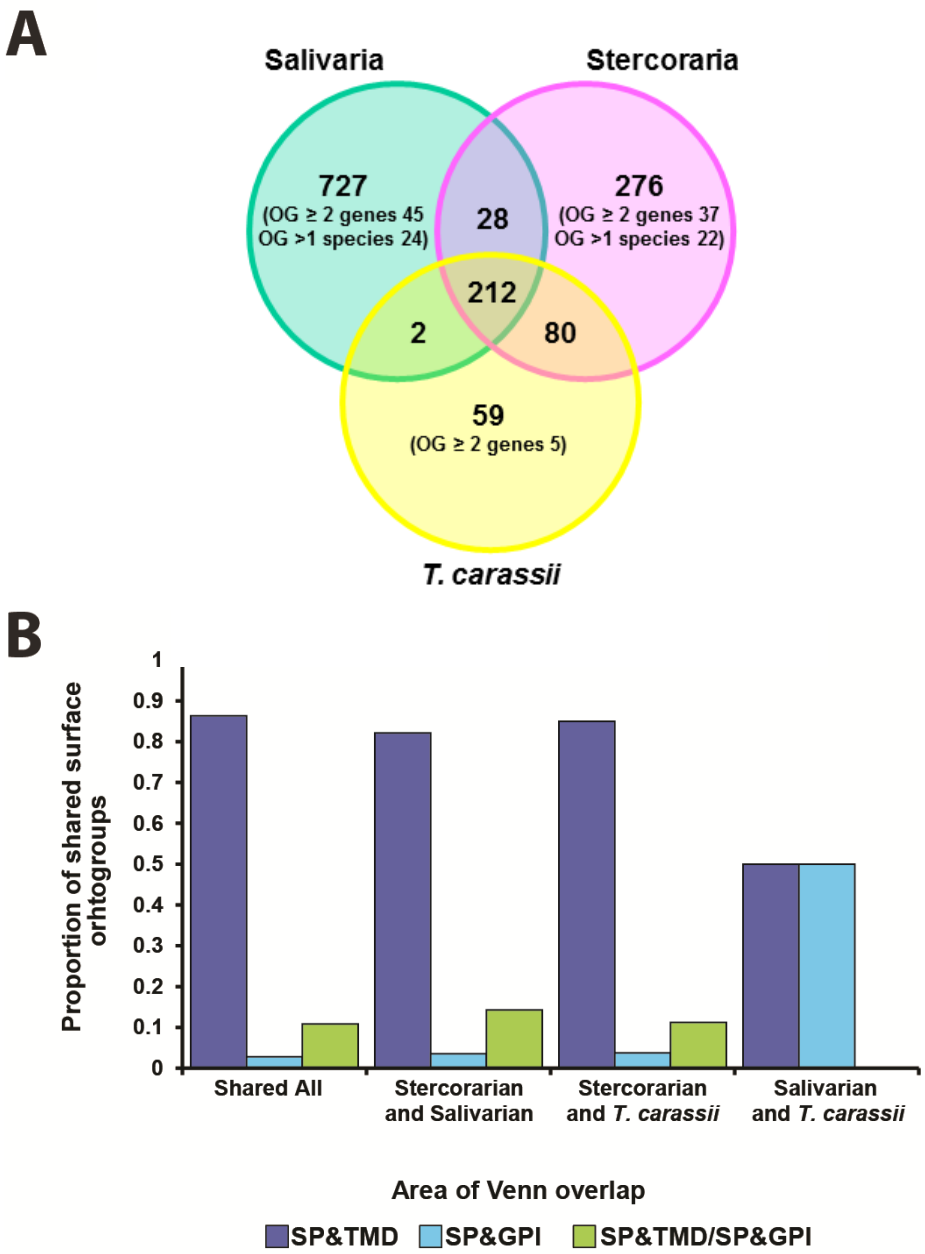



***Figure 5.4*** **The results of orthogroup clustering following prediction of surface orthogroups**

**A)** A Venn diagram showing the overlap between the surface orthogroups and the different groups of trypanosomes. Also shown are the number of orthogroups (OG) containing a total of greater or equal to two genes (OG ≥2 genes) and the number of orthogroups in more than one species (OG > 1 species). **B)** A breakdown of the shared surface orthogroups showing the proportion of surface orthogroups constituted by genes predicted to encode the surface signalling domains. The full dataset is available

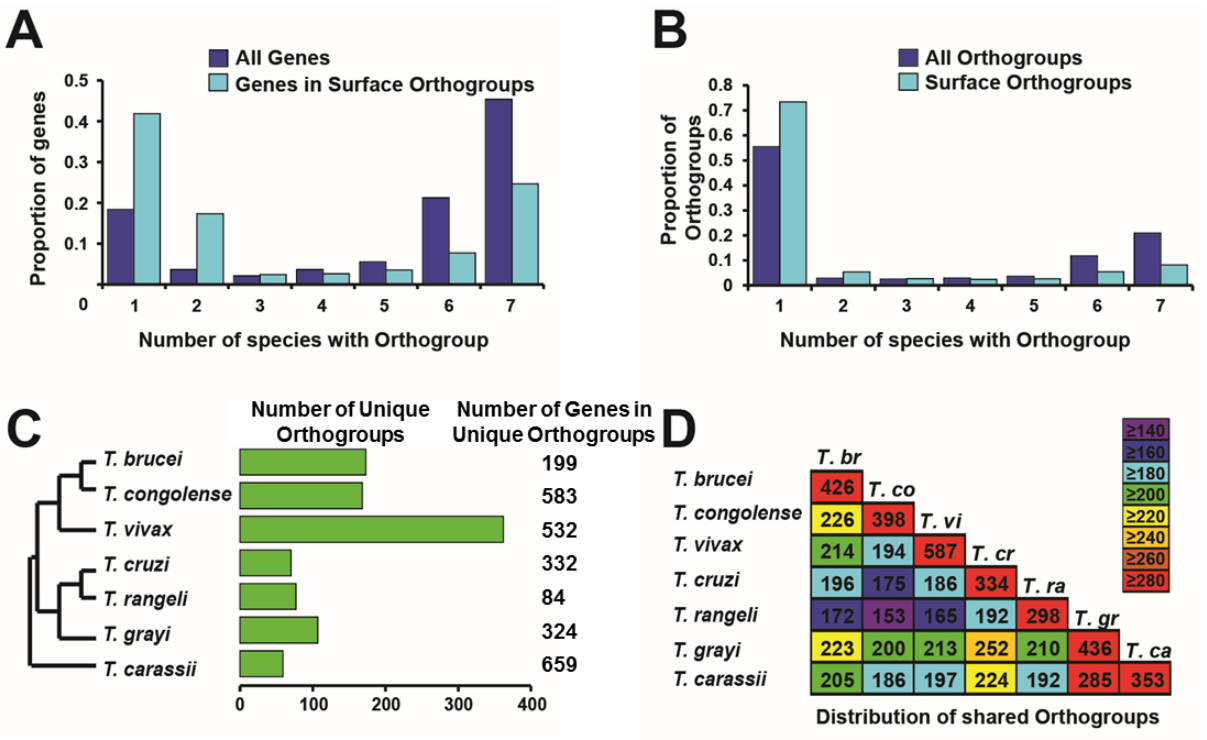



***Figure 5.5*** **The species distributions of surface orthogroups in *Trypanosoma***

**A)** The proportion of genes from all species that are contained within surface orthogroups spanning the given number of species (for example 40% of all surface protein genes are contained within surface orthogroups that are unique to a single species, whereas only 18% of all genes have the same distribution). **B)** As in A with analysis performed for surface orthogroups rather than genes. **C)** The number of species-specific surface orthogroups. **D)** A matrix showing the number of surface orthogroups that are shared between species, species order on the bottom is the same as on the side.

A similar trend is observed in this analysis as was seen in the previous analysis with *Phytomonas, Leishmania* and *Trypanosoma* (Chapter 3). By proportion, there are more genes in predicted surface orthogroups found in one species than you would expect given the species distribution of all orthogroups. At the other end of the spectrum there are proportionally more genes found in all orthogroups shared by six or seven species than there are genes from predicted surface orthogroups shared by these numbers of species. This result means that surface genes are found in fewer species than you would expect from the distribution of genes. Furthermore, orthogroups of surface genes are found more often in a single species than you would expect by chance. Moreover, within those cell surface orthogroups that are unique to a single species there are examples where individual orthogroups are very large such as OG8 and OG42 of *T. carassii*. While the rate of expansion of these orthogroups is unknown, given that members of the orthogroup could not be detected in other species it is likely that expansion occurred since the divergence of *T. carassii* from its closest relatives in the tree.

The results agree and extend previous studies that the ancestral cell surface proteome has been adapted within each lineage following the split from the common ancestor (Jackson *et al.*, 2013). The proteins identified by this bioinformatics method that are present in *T. brucei* and common to the proteomic analysis performed by Gadelha (2015) are summarised in Table 5.1. The corresponding family numbers from the Cell Surface Phylome of African trypanosomes (Jackson *et al.*, 2013) are also shown.

## Table 5.1 A comparison of the present study with two other investigations into the trypanosome cell surface

Given here are the *T. brucei* accession numbers identified by Gadelha *et al.*, 2015, the corresponding gene family numbers identified by Jackson *et al.*, 2013 and the orthogroups that these genes belong to identified in this study.

| Identified CG 2015 | Description | CSP AJ 2013 | OG EJ 2015 |
|---|---|---|---|
| Tb427VSG-1637 | Tb427VSG-1637 | | OG000534 |
| Tb927.7.4230 | hypothetical protein, conserved | Fam79 | OG000765 |
| Tb05.5K5.210 | hypothetical protein,BAC from homologous region on chr5 | Fam73 | OG000821 |
| Tb427.04.1110 | hypothetical protein, conserved | | OG001099 |
| Tb427.08.2380 | ABC transporter, putative | Fam56 | OG001259 |
| Tb427tmp.01.6860 | hypothetical protein, conserved | | OG002194 |
| Tb427tmp.01.6880 | cytosolic coat protein, putative | | OG002275 |
| Tb427.03.4630 | UDP-glucose:glycoprotein glucosyltransferase, putative | | OG002755 |
| Tb427tmp.02.0960 | endosomal integral membrane protein, putative | | OG003106 |
| Tb427.10.9150 | hypothetical protein, conserved | | OG003746 |
| Tb427.10.6870 | short-chain dehydrogenase, putative | | OG004129 |
| Tb427.10.5090 | hypothetical protein, conserved | | OG005542 |
| Tb427.04.3040 | hypothetical protein, conserved | | OG005593 |
| Tb427.07.6490 | hypothetical protein, conserved | Fam10 | OG006115 |
| Tb427.05.360 | 75 kDa invariant surface glycoprotein (ISG75) | Fam49 | OG006303 |
| Tb427.05.630 | acidic phosphatase, putative | Fam49 | OG006305 |
| Tb427tmp.v1.0650 | hypothetical protein, conserved | | OG006409 |
| Tb427.10.10840 | hypothetical protein, conserved | | OG006457 |
| Tb427tmp.02.4900 | hypothetical protein, conserved | | OG006881 |
| Tb927.6.440 | haptoglobin-hemoglobin receptor (HpHbR) | | OG008395 |
| Tb927.7.470 | hypothetical protein | | OG008716 |
| Tb427.05.610 | acidic phosphatase, putative | Fam49 | OG009114 |
| Tb427.10.990 | hypothetical protein, conserved | | OG010343 |
| Tb427.BES65.1 | transferrin-binding protein, putative | Fam15 | OG010345 |
| Tb427.07.4690 | hypothetical protein | | OG010354 |
| Tb427.08.5910 | hypothetical protein | | OG010365 |
| Tb427.10.5700 | hypothetical protein, conserved | Fam05 | OG028173 |
| Tb927.1.5110 | expression site-associated gene 11 (ESAG11) protein, putative | | OG028768 |
| Tb927.3.600 | hypothetical protein | Fam05 | OG028978 |
| Tb427.10.9450 | hypothetical protein | | OG029038 |
| Tb427.05.4010 | hypothetical protein | | OG029096 |

### *5.3.3 The set of surface orthogroups found in all species*

A core set of 212 putative surface proteins are conserved in all three lineages, this includes transporters such as the ABC transporter, UDP-galactose transporter and other putative transporters. Also found in this core of the pan-trypanosome surface proteome are gp63, lipophosphoglycan and a large number (64% of ubiquitously conserved surface orthogroups) of hypothetical proteins (Appendix 5.1). The high proportion of conserved putative surface protein genes annotated as hypothetical proteins indicates that there may be a great deal of biology associated with the trypanosome cell surface that is conserved across the trypanosomatids but that has yet to be characterised.

Consistent with previous reports, surface orthogroup clustering found that cell surface proteins such as the haptoglobin-haemoglobin receptor, *Brucei* Alanine-Rich Protein (BARP), Procyclin-Associated Genes (PAGs) and Expression Site-Associated Genes (ESAGs) are unique to the salivarian lineage and were not inherited from a common ancestor and subsequently lost from other trypanosomatids (Appendix 5.2). Similarly Mucins and Mucin-Associated Surface Proteins (MASPs) were found to be unique to the stercorarian trypanosomes (Appendix 5.2). Trans-sialidases were found to be mostly unique to the stercorarian trypanosomes but a putative trans-sialidase gene was found to be shared between the salivarian and stercorarian groups but absent from *T. carassii*. This may represent a gene that evolved after the divergence of *T. carassii* from the trypanosomatid lineage or may represent a false negative missing gene from *T. carassii* due to the limitations inherent in transcriptome sequencing.

Some clear trends emerge; of those surface orthogroups that are found in all trypanosoma clades the majority (86%) are predicted to localise to the cell surface by a signal peptide and transmembrane domain. Only 3% are predicted to localise to the surface by SP&GPI, and the remaining 11% of these core surface orthogroups are predicted to encode both SP&TMD and SP&GPI and thus, in the absence of other information, may localise to the surface by either method. Furthermore, 27% (58 of the 212 pan trypanosome surface

orthologs) are present as single copy genes in each of the species included in this study (Appendix 5.1) and so lack the gene copy number expansions associated with proteins involved in immune evasion. These observations are consistent with those made by Gadelha *et al.,* (2015) in that the GPI-anchored proteins are those with receptor-like properties and are largely parasite-specific, in contrast to transmembrane proteins with transporter-like properties that are conserved across the trypanosomatids and, in some cases, across the eukaryotes (Gadelha *et al.*, 2015). The common proteins identified in both the Gadelha dataset and the cell surface phylome of African trypanosomes (Jackson *et al.*, 2013) are shown in Table 5.1. The proteins listed are a subset of these two studies.

### 5.3.4 The set of surface orthogroups lost from either the salivarian or stercorarian lineages

Of the surface orthogroups common to both the salivarian and stercorarian trypanosomes but absent from *T. carassii,* 82% contain SP & TMD compared to 4% targeted by SP & GPI, the remaining 14% are members have all three features predicted (SP & TMD & GPI probably representing a false positive TM or GPI prediction). The majority of surface orthogroups shared by *T. carassii* and the stercorarian trypanosomes but absent from the salivarian trypanosomes contain a signal peptide and transmembrane domain: 85% compared to 4% SP & GPI anchor and 11% SP & TMD & GPI. Of the two surface orthogroups shared between *T. carassii* and the salivarian clade but absent from the stercorarian trypanosomes one is predicted to localise to the surface by a signal peptide and transmembrane domain (OG007019) while the other is predicted to have a signal peptide and GPI anchor (OG009313).

Surface orthogroups shared between *T. carassii* and the stercorarian trypanosomes but absent from the salivarian lineage include a phosphate transporter and a putative cysteine protease. Characterised cysteine proteinases include cruzipain, the major cysteine protease of *T. cruzi* with roles including host cell invasion and evasion of the host immune response (Cazzulo *et al.*, 2001). It may be that cysteine protease are ancestral virulence

factors and this function may have been conserved in *T. carassii.* A notable absence is that of any mucin or mucin-like protein families shared between *T. carassii* and the stercorarian lineage. The presence of these proteins was suggested by previous work (Campo *et al.*, 2002).

### *5.3.5 T. carassii* transcriptome contains two large families of putative surface proteins not found in any other trypanosomatid

Following inference of orthogroups and identification of putative surface localised orthogroups we identified two large orthogroups that were specific to *T. carassii*. The first of these two families is orthogroup 8 (OG8) (Figure 5.6, Appendix 5.2), which contains 500 members. Of OG8 99% of the complete predicted open reading frames (i.e. those that have a start codon and stop codon within the boundaries of the assembled transcript) are predicted to encode a GPI anchor. Similarly 63% of those that begin with a start codon that is not located at the start of the assembled transcript are predicted to encode a signal peptide. A consensus alignment of these regions shows a high degree of conservation of the N- and C-termini representing the signal peptide and GPI anchor signals respectively (Figure 5.6 B). The logos (representing both sequence alignments and profile hidden Markov models) of these orthogroups shows that the internal portion of the proteins are degenerate with a low degree of sequence conservation (Figure 5.6 B). The predicted secondary structure of a full length representative of this family TryCa.5662.10_full by Phyre2 (Buchan *et al.*, 2013) consists mainly of alpha helices, a feature shared with the surface antigen VSG genes of *T. brucei*.

The other surface orthogroup unique to *T. carassii* is OG42 (Figure 5.6, Appendices 5.1 and 5.2, Table 5.3). This surface orthogroup contains fewer genes (n = 128) of which 83% are predicted to encode a GPI anchor sequence and 65% are predicted to encode a signal peptide. Unlike surface orthogroup 8, the central part of the protein between the N and C terminal sequences shows a high degree of sequence conservation (Figure 5.6D).

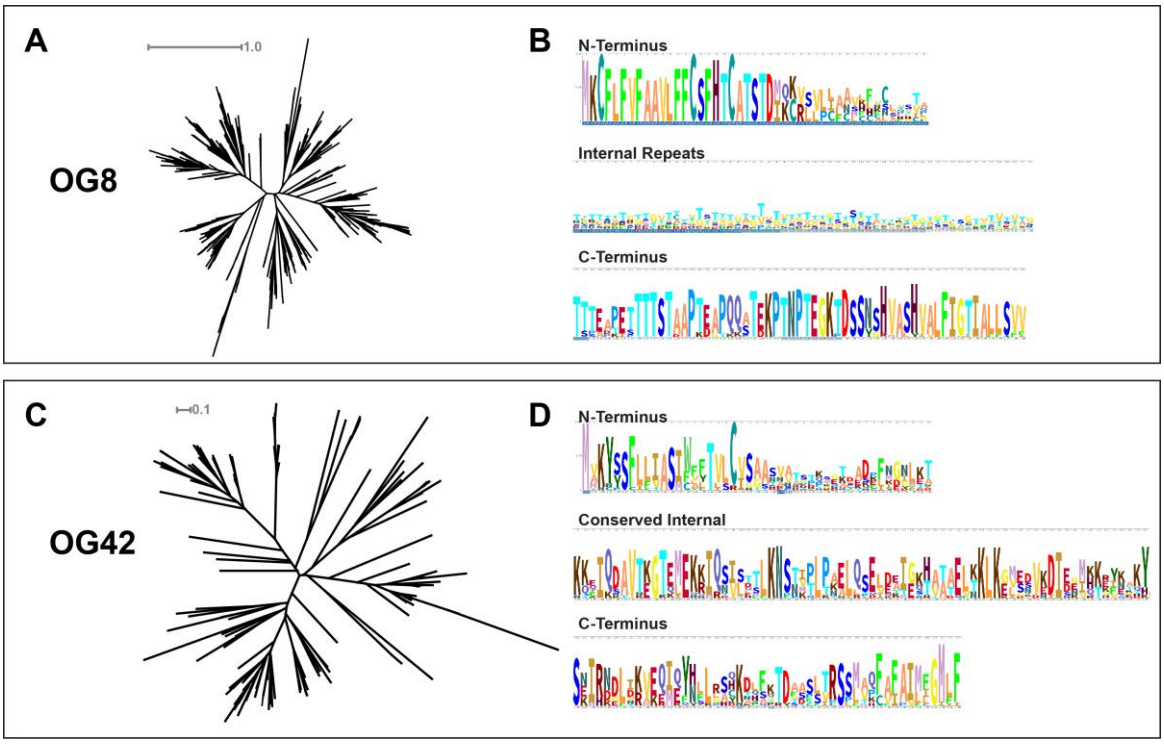



***Figure 5.6 Trypanosoma carassii* specific surface orthogroups**

A) OG8 phylogenetic tree showing the relationship of the assembled gene family members. B) LOGOs of OG8 where letter height shows the probability of a residue in that position. Conservation at the N-terminus of the sequence is predicted to be a result of functional constraint on this part of the sequence to encode a signal peptide, the second line shows internal degenerate region with many repeats and the third line the conserved C-terminal region. Sequence similarity at C terminus could be as a result of GPI attachment signal placing constraint on sequence evolution. C) OG42 phylogenetic tree showing the relationship of the assembled gene family members. D) LOGOs of OG42 where the first line shows conservation at the N-terminus also predicted to be because of the signal peptide, the second line shows the internal region with a high degree of sequence conservation. A conserved C-terminal region is shown in the third line, again conservation at the C terminus is believed to be the result of the sequence for the addition of a GPI anchor.

Logos generated with http://skylign.org/

### 5.3.6 Cell surface biotinylation and isolation of the *T. carassii* surface proteome.

To determine if any of the predicted cell surface proteins could be detected on the surface of *T. carassii* a proteomic investigation was conducted. Bloodstream form *T. carassii* parasites were grown (see section 2.2.3), cells were biotinylated (see section 2.4.1), biotinylated proteins were enriched (see section 2.4.2), tandem mass spectrometry performed on all fractions (see section 2.4.8) and data were searched against a database of *T. carassii* protein sequences inferred from the transcriptome (see section 2.4.9).

The epifluorescence microscopy of biotinylated cells (for method see section 2.4.10) labelled with streptavidin Alexa-Fluor shows that a surface localisation of the biotin tag was achieved (Figure 5.7). This localisation is analogous to that previously observed for *P. serpens* in Chapter 3, and in Leishmania parasites (Yao et al., 2010). Moreover, the fluorescence imaging supports the observation made in Chapter 3 that the flagellar pocket can be labelled with biotin. This may indicate that the plasma membrane proteins of the flagellar pocket are accessible to reaction with the biotinylation reagent or that biotinylated proteins located in the membrane of the cell body and flagellum are relocated to the flagellar pocket during the incubation period.

Table 5.2 gives a summary of the data from the mass spectrometry of the total combined *T. carassii* protein samples. Table 5.3 details the likely cell surface proteins i.e. those that contain at least one predicted membrane associated domain (TMD, SP, GPI), detected in the biotinylated pull down fraction in decreasing order of enrichment score (spectral index in eluate over unbound material). Table 5.4 lists the putative cell surface proteins identified by this method in *T. carassii* that have orthologs in other trypanosomatids annotated as hypothetical proteins.

**Table 5.2 A Summary of LC MS/MS on *T. carassii***
showing total number of spectra identified, the percentage of spectra identified and the number of unique peptides.

| | |
|---|---|
| **Spectra** | 45541 |
| **% Spectra Identified** | 25.06% |
| **Uniqiue peptides** | 7391 unique |

## Table 5.3 Results of LC MS/MS of biotinylated *T. carassii*

Proteins are listed in order of decreasing $log_2$(fold enrichment). Three proteins were identified only in the bound fraction so $log_2$(fold enrichment) was calculated using the lowest value observed for a protein in the unbound fraction. To search for putative function sequences were searched against the Pfam and NCBI nr databases and searched against other Trypanosomatid proteins. Secondary structure was also predicted in addition to predictions of GPI anchor, signal peptide and transmembrane domain. In the Secondary structure column D = disordered, A = alpha helix, B = beta sheet, TM = transmembrane helix. The final three columns indicate the presence and number of features present.

| LOG2 FOLD ENRICHMENT | PROTEIN | PFAM | MOLECULAR WEIGHT (DA) | DESCRIPTION | SECONDARY STRUCTURE | ORTHOGROUP | SI BOUND | SI UNBOUND | UNBOUND/ BOUND | TMD | GPI | SP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8.42 | TryCa.2544.10 ctrunc | Peptidase M8 | 63366.2 | surface protease GP63, putative,metallopeptidase, putative | D25%, A17%, B23%, TM3%, | 000071 | 1.05E-06 | 3.06E-09 | BOnly | 1 | 0 | 1 |
| 6.89 | TryCa.8020.10 ntrunc | | 29298.1 | no significant hits | D22%, A79%, B2%, | 000042 | 3.62E-07 | 3.06E-09 | BOnly | 0 | 1 | 0 |
| 6.87 | TryCa.5961.10 full | | 21018.2 | no significant hits | D20%, A84%, B2%, TM9%, | 000042 | 3.57E-07 | 3.06E-09 | BOnly | 0 | 1 | 0 |
| 1.48 | TryCa.5636.10 full | | 36604.9 | no significant hits | D23%, A90%, | 030905 | 6.75E-07 | 2.42E-07 | 0.36 | 0 | 1 | 0 |
| 0.30 | TryCa.5956.10 full | | 30633.1 | no significant hits | D19%, A82%, TM6%, | 000042 | 6.91E-07 | 5.6E-07 | 0.81 | 1 | 1 | 1 |
| -0.81 | TryCa.8764.10 ntrunc* | | 22322.8 | no significant hits | D17%, A88%, TM16%, | 030584 | 6.27E-07 | 1.1E-06 | 1.75 | 0 | 1 | 0 |
| -0.89 | TryCa.7013.10 ctrunc* | | 32481.1 | no significant hits | D24%, A85%, | 000042 | 2.16E-07 | 4.01E-07 | 1.86 | 0 | 0 | 1 |
| -1.28 | TryCa.5723.10 full* | | 33172.1 | no significant hits | D31%, A85%, | 030853 | 2.25E-07 | 5.47E-07 | 2.44 | 4 | 1 | 1 |
| -3.42 | TryCa.5286.10 full | | 27731.7 | no significant hits | D17%, A79%, B4%, | 029363 | 4.74E-07 | 5.08E-06 | 10.7 | 1 | 0 | 1 |

## Table 5.4 Putative cell surface proteins in Table 5.3 with orthologs in other species annotated as hypothetical proteins

Using the orthogroup data two of the proteins listed in Table 5.3 had orthologs in other *Trypanosoma* species, they are given here along with differential expression data from *T. brucei.* i.e. mRNA abundance in bloodstream form (BSF) divided by mRNA abundance in procyclic form (PCF).

| | *T. brucei* | *BSF/ PCF* | *T. congolense* | *T. vivax* | *T. cruzi* | *T. rangeli* | *T. grayi* |
|---|---|---|---|---|---|---|---|
| *TryCa.1036. 10ntrunc* | Tb927.11 .6370, | 0.91 | TcIL3000.1 1.6930, | TvY486_11 06760, | TcCLB.5101 01.250, | TRSC58_ 06105, | Tgr.116. 1010, |
| *TryCa.10527 .10 ntrunc* | Tb927.6. 4020, | 0.73 | TcIL3000_6 _3420, | TvY486_06 03490, | TcCLB.5117 37.6, TcCLB.5117 35.70, | | Tgr.27.1 050, Tgr.27.1 040, |

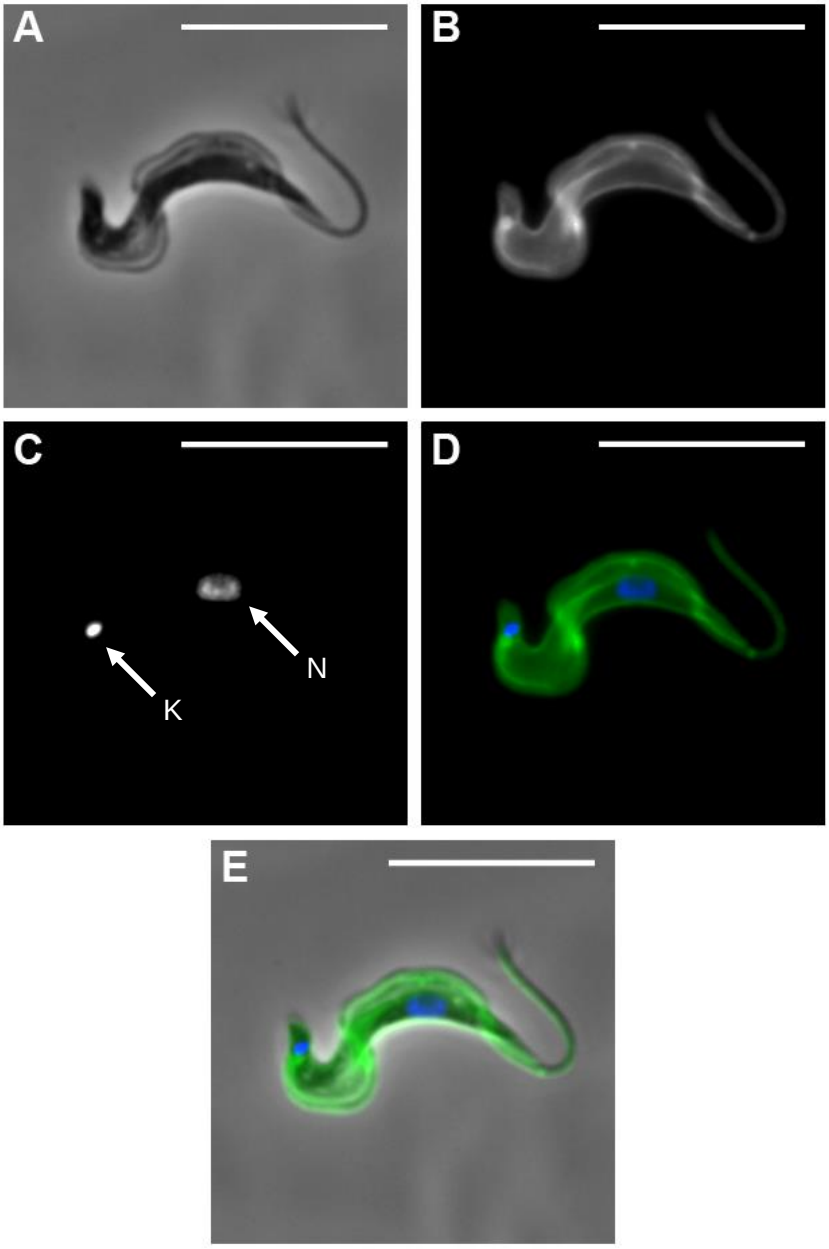



***Figure 5.7*** **Cell surface biotinylation of** ***Trypanosoma carassii***

Fluorescence micrograph images of fixed *T. carassii* cells to show cell surface localisation of biotin label. Cells were labelled with EZ-link-sulfo-NHS-SS-Biotin (Thermo Scientific) and fixed in 4% paraformaldehyde. Biotin was detected using streptavidin AlexaFluor488 (green) (Invitrogen). DNA was stained using Hoechst (blue). **A**) Phase contrast image of a *T. carassii* cell. **B**) Epifluorescence image of AlexaFluor label. **C**) Epifluorescence image of Hoescht DNA stain showing nucleus (N) and kinetoplast (K). **D**) False colour image representing the merge of B and C. **E)** merge of all channels. Scale bar 10 μm.

### 5.3.7 Characterisation of the *T. carassii* cell surface proteome.

Peptides corresponding to 22 proteins were identified in the enriched fraction (Table 5.3). Similarly, peptides corresponding to 1326 proteins were found in the unbound material and 10 proteins in the negative control. Of the proteins found both in the negative control and enriched sample we identified a glucose regulated putative luminal binding protein 1, a putative tryparedoxin peroxidase and beta tubulin. The presence of these proteins indicates that there was some contamination in the enriched fraction from abundant cytosolic and mitochondrial proteins.

The enrichment scores were calculated from the SI values. Where a protein was not detected in the unbound fraction the SI was set as the lowest value observed in the bound fraction which was 3.06E-09. An asterisk following the protein accession number indicates that a biotin PTM was detected on peptides that matched to that protein. The secondary structure was predicted using Phyre2 AND (D = disordered, A = alpha helix, B = Beta strand and TM = Transmembrane domain). From the Phyre2 predictions, the majority of proteins isolated (8/9) are predicted to be predominantly alpha helical. All secondary structure predictions from Phyre2 for identified *T. carassii* proteins are available in Appendix 5.6.

Figure 5.8A shows a histogram of the $\log_2$ fold enrichment of proteins in the enriched fraction compared to total cell soluble fraction. The right hand skew in this histogram indicates that proteins detected in the enriched fraction do not have relative abundances that you would expect from a random sample of the control fraction and is indicative of a successful enrichment of specific subsets of proteins. Table 5.3 and Figure 5.8C both show the proteins that were detected in the enriched fraction in order of $\log_2$ fold enrichment over the unbound fraction. The predicted secondary structures of these proteins are provided in Appendix 5.6. A putative gp63 metallopeptidase is the top enriched protein followed by two GPI anchored proteins unique to *T. carassii*. These two proteins ranking 2$^{nd}$ and 3$^{rd}$ most enriched along with the 7$^{th}$ and 10$^{th}$ most enriched proteins are each

members of one of the surface orthogroups unique to *T. carassii,* OG42, discussed earlier in this chapter. We also identified 4 hypothetical proteins of unknown function (i.e. no annotation of orthologous genes and no BLAST hits), two of these had a biotinylation modification detected by mass spec (Table 5.3).

TryCa.5636.10_full, TryCa.5956.10_full, TryCa.8764.10_ntrunc, TryCa.7013.10_ctrunc, TryCa.5723.10_full, TryCa.5286.10_full are members of orthogroups which were not found to be present in any other species of *Trypanosoma* included in this study. Of the hypothetical proteins with orthologs in *T. brucei,* all *T. brucei* orthologs have been observed to be downregulated in bloodstream forms (BSF) compared to procyclic forms (PCF) (Table 5.4).

We were able to detect the modification of a lysine residue within a peptide with the covalently attached portion of the EZ-link NHS-SS-Biotin (i.e. what remains after reaction with the lysine, reduction of the disulphide bond by TCEP and alkylation by Iodoacetamide). This modification was detected in three peptides TryCa.5723.10_full, TryCa.8764.10_ntrunc and TryCa.7013.10_ctrunc. The former is a member of the *T. carassii* surface orthogroup OG42 and all three proteins contain the membrane associated domains often found in surface localised proteins (Table 5.3). Proteins with a detectable biotin PTM were also detected in the unbound material- one such protein is TryCa.4555.10_full which has a BLAST hit to a flagellar calcium-binding protein in *Trypanosoma rangeli*. The full dataset is available in Appendix 5.4.

Once again there is evidence that the computational methods for predicting TMDs, GPI and SP domains in these trypanosomatid predicted proteomes is not working perfectly. For example TryCa.8020.10 ntrunc, TryCa.5961.10 full, TryCa.5636.10 full, TryCa.8764.10 ntrunc are all predicted GPI anchors with no signal peptide targeting the protein to the endomembrane system. TryCa.5956.10 full, TryCa.5723.10 full are predicted to localise by TMD, GPI and SP domains and as in Chapter 3 likely represent false positive TM predictions.

A number of carp proteins were detected in the proteomic dataset. Appendix 5.5 shows the results of searching the peptide results of both LC-MSMS runs against a database of predicted carp proteins. The detection of transferrin variants, complement factors and Alpha-1 antitrypsin is consistent with previous proteomic studies of carp blood plasma (Dietrich *et al.*, 2014). However, carp proteins were detected in the proteomic samples when searched against the *T. carassii* transcriptome. In particular, a carp transferrin variant present in the *T. carassii* transcriptome (TryCa.10460.10_ntrunc) was detected in the biotinylated pull-down sample with a $\log_2$ enrichment score of 0.074. While it is possible that the gene encoding this protein was horizontally transferred to the *T. carassii* genome, a much more parsimonious explanation is that both the protein and transcript were contaminants of the proteome and transcriptome preparation respectively as both required purification of trypanosome cells from carp blood.

With respect to the two *T. carassii* specific surface orthogroups discussed in 5.3.5 only members of OG42 were detected in the enriched sample. In total, proteins corresponding to three different transcripts TryCa.5956.10_full, TryCa.5961.10_full and TryCa.7013.10_ctrunc were identified as enriched in the eluate vs unbound material. Members of the *T. carassii* unique orthogroup *OG8* were not detected in the eluate but peptides corresponding to members of the orthogroup were detected in the unbound material. The results obtained in Figure 5.8D are a reflection of the low numbers of detected proteins (n = 22). Here, small differences in the numbers of genes that are recovered from this set result in large differences in the magnitude of the hypergeometric test p-value that is returned. This plot is provided to be consistent with the figure provided in Chapter 3 (Figure 3.6B). In Chapter 3 there is a much larger sample size and thus the p-values behave less erratically. Therefore, not much can be interpreted from the data presented in Figure 5.8D however it is provided for consistency with the previous chapter.

Similarly, the negative values (implying less abundance in the elution than the flow through) in the log2 fold enrichment of OG42 family members are a mathematical artefact

cause by the data normalization procedure again caused by the low number of proteins detected (n = 22). Again this figure was provided for consistencies sake and the important thing to note is not the normalized relative values but the fact the proteins were detected at all.

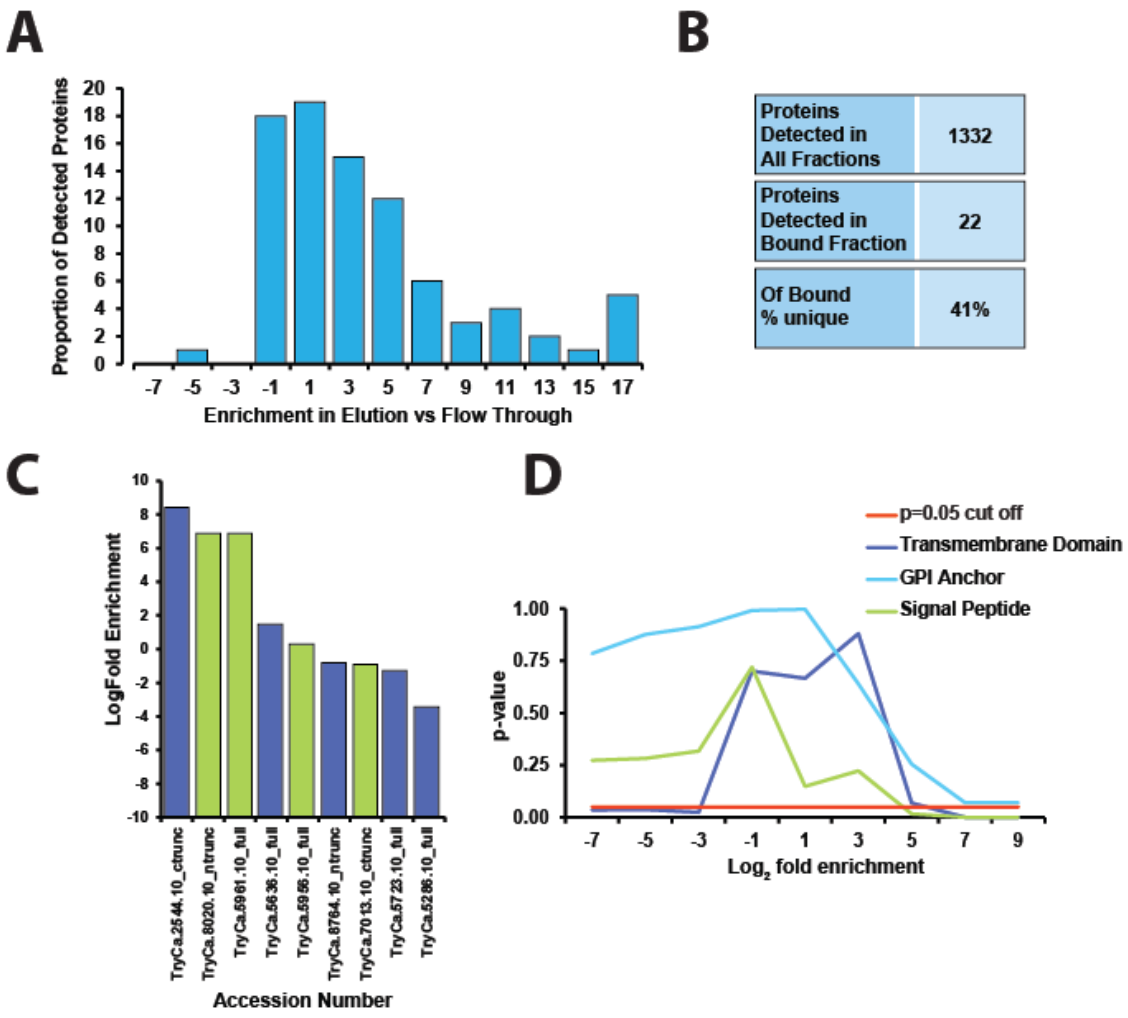



***Figure 5.8*** **Results from cell surface biotinylation in *T. carassii***

**A**) A histogram plotting the $log_2$ enrichment scores of proteins detected in the bead elution vs the unbound fractions. **B**) A table giving the breakdown of the number of proteins detected in each fraction and the proportion of enriched proteins that belong to surface orthogroups unique to *T. carassii.* **C**) Proteins detected in the eluate that are possible surface proteins listed in Table 5.3 ranked by $log_2$-fold enrichment of the elate vs unbound

material. Members of the *T. carassii* unique OG42 orthogroup are in green. **D)** The results of hypergeometric tests to determine whether the set of proteins with that fold-change or higher contains more than the expected number of proteins that bear the membrane associated domains: SP, GPI, TMD.

## 5.4 Discussion

In this chapter I present a comparative bioinformatic analysis of the predicted surface proteome of *T. carassii,* an extracellular parasite of cyprinid fishes. Comparison of the genes present in the *T. carassii* transcriptome with those present in the genomes of other trypanosomatids reveals a new dimension to the evolution of the cell surface in this group. As the *T. carassii* sequence data is from a transcriptome, not a genome, only genes expressed at the time of RNA isolation will be present. This means that the sequence data gives no indication of genome organisation, or whether there are large numbers of genes in the same gene family that were either not expressed or expressed at levels too low to be detected by *de novo* assembly of RNAseq. In Figure 5.4 the number of surface orthogroups in *T. carassii* is lower than in the salivarian or stercorarian trypanosomes. This is in part due to the fact that genome data was used for the salivarian and stercorarian trypanosomes while transcriptome data is used for *T. carassii*. As not all genes in the genome are expressed to levels sufficient to be assembled, it is likely that there are surface orthogroups in *T. carassii* genome that are not present in the assembled transcriptome.

As *T. carassii* is an outgroup to both the stercorarian and salivarian trypanosomes, analysis of this transcriptome provides insight into those proteins that were present in the last common ancestor of these organisms. This revealed that there was widespread loss of surface protein genes in the lineage that lead to salivarian trypanosomes. In total 5.8% of the predicted ancestral surface orthogroups (those orthogroups with genes present in *T. carassii* and one species of the stercorarian clade) were lost in the lineage that lead to the salivarian trypanosomes. In contrast to this, the extent of gene loss in the stercorarian trypanosomes is much less, with only 0.1% of ancestral surface protein gene families lost

in this lineage. In total, 15.3% of the predicted cell surface orthogroups are present in *T. carassii* and at least one species of both stercorarian and salivarian clades. This work also reinforces previous observations that the complement of surface proteins have largely evolved independently in each lineage, such as the evolution of the VSG genes in the salivarian clade. Absence of mucins and MASPs from *T. carassii* transcriptome provide further evidence that they are unique to the stercorarian clade. The majority of our knowledge of these two groups of proteins is from the stercorarian parasite *T. cruzi* where mucins are glycoproteins with functions in cell-cell and cell-molecule interactions and cell surface protection (Frasch, 2000) and MASPs are a polymorphic family of cell surface proteins thought to function in the evasion of the host immune system (dos Santos *et al.*, 2012).

Beyond GPI, SP and TM domain prediction, no further targeting sequences were searched for. For example, mitochondrial targeting sequences were not predicted for this dataset. This was in order to conduct an analysis directly comparable to the *Trypanosoma* cell surface phylome (Jackson *et al.*, 2013). This also prevented the exclusion of sequences that could encode cell surface localised proteins but could be mispredicted to be targeted to the mitochondria. This is a limitation of the dataset.

The general trend seen in the pan-trypanosome core bioinformatic predicted surface proteome and the various patterns of conservation is that shared surface orthogroups consist principally of transmembrane proteins predicted to be targeted to the cell surface by a signal peptide as opposed to signal peptide and GPI anchored proteins. It is tempting to speculate, given anecdotal evidence concerning a limited number of surface proteins, that the reason for this differential pattern may be due to differences in the functionality of the two types of surface genes. Specifically, that transmembrane proteins provide the bulk of essential membrane bound functions while GPI anchored proteins function mostly in immune evasion. It is interesting to note that of the set of conserved gene families predicted to be present on the cell surface, the majority of these are of unknown function.

Thus there is a large portion of the conserved biology of these enigmatic parasites that remains to be characterised. The *T. carassii* transcriptome revels that this parasite expresses the orthologs of the genes involved in LPG synthesis but the proteomic and bioinformatic methods are unable to provide the required evidence to confirm the presence or absence of GIPLs on the cell surface (Appendix 3.3 and blast searches to confirm).

Bioinformatic analysis of the predicted surface protein genes in *T. carassii* also led to the discovery of two large groups of putative surface proteins unique to *T. carassii*. Lack of detectable similarity by BLAST to known surface proteins such as trans-sialidase, mucin and VSG suggests that they are unique to this clade. The presence of predicted GPI anchors and signal peptides in members of these two groups of orthologous proteins provides strong evidence that they are located on the plasma membrane of *T. carassii* and the identification of representatives from one surface orthogroup, OG42, in the biotinylated sample is evidence that they are indeed present on the cell surface during this life cycle stage. OG42 is a large family of putative cell surface proteins, which are the 42nd largest family of orthologous genes in the trypanosomatid sequences included in this analysis. The number of individual transcripts sequenced and proportion of proteins detected in the proteomic cell surface proteome indicated that this group might form a majority of the cell surface protein coat in *T. carassii*. OG42 is predicted to have a predominantly alpha helical structure, other large families of surface proteins with this structure include the VSG of *T.brucei*. Thus it is tempting to speculate that this protein may perform a similar role in host-parasite interactions, though there is no functional data to support this hypothesis. Moreover, the full extent of this gene family is unknown. The limited number of proteins detected in the proteomic sample may be due to the differential presence of members of this highly diverse family in both the proteome and transcriptome samples, which were isolated from temporally separated populations of cells isolated from different fish. Thus the cohort of surface proteins expressed at the time of isolation of the transcriptome and the proteome may have been different.

It is possible that this study has touched on an exciting find; specifically that the OG42 orthogroup appears to be a large families of GPI anchored surface localised proteins that are predicted to have a predominantly alpha helical structure. These features of the gene family are in some ways similar to the major surface gene families in *T. cruzi* and *T. brucei* and hints that these gene families may share a common genetic origin that is not detectable due to rapid sequence divergence. Alternatively, it is possible that it represents convergent evolution of three large families of GPI anchored proteins. It would be interesting to sequence the *T. carassii* genome to see if genome organisation could reveal any further information to support this, e.g. genes organised at telomeres.

The proteomics list is much shorter compared to that for *P. serpens* in Chapter 3. There are some different ways to interpret the list of enriched proteins. Either, the list is poor and few surface proteins were isolated. Or that the OG42 proteins represent the majority of the cell surface in these parasites when they inhabit the host bloodstream, similar to VSG in BSF *T. brucei*. Also, given that 100's of members of OG42 were assembled from the transcriptome (most lacking N-termini due to assembly artefacts such as low coverage and sequence complexity), and that they are extremely diverse at the protein sequence level, it is possible that the overlap between the genes that were expressed at the time of the RNA isolation and those expressed at the time of the protein isolation were different, and thus the mass spectra could not be matched to the protein sequences predicted from the transcriptome.

## 5.5 Conclusions

Previous observations on the evolution of the trypanosomatid cell surface were made based on analysis of a small number of species representing a small fraction of the diversity contained within the group. A genetic and proteomic study to investigate the cell surface proteome of the fish parasite *T. carassii*, which evolved prior to the evolutionary split between the stercorarian and salivarian trypanosomes, provides new insight into the reductive evolution of the salivarian cell surface as well as lineage specific invention and

expansion of surface expressed genes. Furthermore, the discovery of two large orthogroups, unique to the fish parasite and predicted to encode proteins present on the cell surface, raises the possibility that there may be many such species specific gene families and many different ways to evade the adaptive immune system within this diverse clade of parasites.

# *Chapter 6: Conclusions and General Discussion*

## 6.1 Overview

A single origin of parasitism in an ancestral trypanosome gave rise to the trypanosomatida, a successful group of single-celled eukaryotic parasites with an extensive host range that includes crops, livestock and humans (Camargo, 1999; Simpson *et al.*, 2006). While the trypanosomatids evolved from a single origin of parasitism, there have been parallel adaptations to facilitate the colonisation of intracellular and extracellular host environments (Lukeš *et al.*, 2014). Primary amongst these adaptations are the lineage specific evolution of immune evasion strategies (Gupta *et al.*, 2013; Higgins and Carrington, 2014; Manna *et al.*, 2013). A number of mammalian infective *Trypanosoma* and *Leishmania* have drawn attention because of the disease burden they place on mankind. In addition, *T. brucei* has been used as a model system in the study of a number of eukaryotic cellular processes (Akiyoshi and Gull, 2014; Broadhead *et al.*, 2006). The result is a very uneven knowledge landscape across the trypanosomatid phylogeny as our understanding of how these enigmatic parasites have adapted to different hosts is limited to just a few trypanosomatid parasites of mammals. This poor resolution has some major implications. Firstly, this limited species sampling precludes us from inferring evolutionary processes and thus limits our understanding of how human infective species have evolved to parasitise us. Secondly, the limited species sampling outside of the mammalian infective species means that we have almost no knowledge of the diversity of immune evasion mechanisms that exist within the group. Therefore, improving species sampling has the potential to improve our understanding of how this fascinating group of parasites evolved. In addition, improved knowledge of the host-parasite interface at the parasite cell surface may lead to improvements in how infections with these parasites may be controlled, for example through the design of novel therapeutic to conserved protein targets. In Chapter 5 I showed that by including *T. carassii* in a comparative bioinformatic study of the predicted

cell surface proteins in the *Trypanosoma* species, it can be seen that 30% of the hypothetical ancestral surface orthogroups were lost in the lineage that gave rise to the salivarian trypanosomes. Without the inclusion of *T. carassii* it was not possible to determine whether these surface orthogroups were lost from salivarian trypanosomes or gained in the ancestor of the stercorarian trypanosomes. It is likely that increased species sampling in this part of the tree will yield further insights into the patterns of molecular evolution (i.e. selection, as well as gene gain and loss) associated with the evolution of antigenic variation.

The proteomic method optimised in *P. serpens* and then applied to *T. carassii* was successful in providing some enrichment of proteins containing the hallmarks domains of cell surface proteins: signal peptides, transmembrane domains and GPI anchors. However, the method suffered from contamination, probably from dead and lysed cells, internalisation of biotin by live cells, and non-specific binding interactions with the beads and associated bound proteins. Further optimisations could be to use a different labelling agent e.g. fluorescein, or to perform a mechanical separation of the plasma membrane e.g. freeze thaw or sucrose gradient. A combination of biochemical labelling, mechanical separation and then affinity purification of the label may also have improved the discovery of cell surface proteins. A different label may have reduced background as contamination is commonly reported in biotinylation methods in the literature (Yao *et al.*, 2010). Given the characteristics of the *T. carassii* protein coat being more complex than *P. serpens* and covered in a dense GPI anchored protein coat a GPI-PLC treatment or trypsin ‘shave’ may have been more successful than cell surface biotinylation to selectively purify these GPI anchored proteins. A whole genome sequence for *T. carassii* would have been an asset to this study as it would have allowed better resolution of gene families and enabled an investigation into the organisation of these gene families in the genome. The proteomic datasets presented in this thesis are likely to contain a number of false positives as proteins potentially targeted to cell-internal membranes (proteins

including a signal peptide and transmembrane domains) were not bioinformatically removed from the analysis as I wanted to conduct a study directly comparable to previous studies in *Trypanosoma* species (Jackson *et al.*, 2013). It is likely that problems with genome assembly have affected gene predictions e.g. if the wrong start codon was used. I acknowledge that these limitations are not addressed in this thesis. However, to improve the stringency of the analysis I only included orthogroups that contained two of the surface associated domains (SP & TMD, or SP & GPI). This way the proportion of false positives will be less than if surface proteins had been predicted on the presence of a single surface associated domain.

Where available I have included the annotation from *T. brucei* and/or *L.mexicana* orthologs in the results of the bioinformatic screens in the Appendices to this thesis. This provides some additional information and in some cases can confirm the presence of known surface proteins e.g. gp63. Potential next steps to parse this dataset would be to remove proteins which are predicted to contain a mitochondrial targeting sequence as identified by MitoFates or similar programme. Following this remaining sequences could be searched for Pfam domains, where present the type of Pfam domain would provide additional information to support or discard the hypothesis for cell surface localisation.

## 6.2 *Phytomonas serpens* as a model for plant-trypanosomatid interactions

Proteomic interrogation of the *P. serpens* cell surface and a comparative bioinformatic study of all available *Phytomonas* predicted protein sequences allowed a comprehensive comparison of surface protein conservation across the trypanosomatids. Comparative bioinformatics showed that the last common ancestor of the trypanosomatid species contained a core of 170 orthogroups (gene families) predicted to encode cell surface localised proteins. However, as 68% of these are annotated as hypothetical proteins in the *L. mexicana* and *T. brucei* genomes, the function of these conserved surface proteins is unknown. As methods to localise proteins in *Phytomonas* using fluorescence tagging

are in their infancy, it would be worthwhile to pursue localisation studies of these hypothetical proteins in *T. brucei* using high throughput methods, such as those developed by Dean *et al.,* (2015). Furthermore, improvement of the protocols for the genetic manipulation of *P. serpens* would allow the complete validation of the proteomic dataset presented in Chapter 3.

*Phytomonas* parasites have been previously shown to exhibit wide morphological diversity (Franca, 1920; Gibbs, 1957). *P. serpens* cells recovered from an experimentally infected tomato in Chapter 4 exhibited cell bodies that were longer and more twisted than observed for cells grown in culture. The morphology was not evaluated morphometrically and the *P. serpens* cells appeared to be in promastigote form. No amastigote-like parasite cell forms were observed among the *P. serpens* cells recovered from fruit. This conflicts with previous observations which found that these cell types were abundant in parasites isolated from plant hosts (Gibbs, 1957; Jankevicius *et al.*, 1989; Kastelein and Camargo, 1990). This may be due to the length of time since infection, or may be due to amastigote cells not expressing the eYFP and thus escaping detection. It would have been interesting to determine the extent of the morphological difference between culture and fruit isolated forms and to determine whether there is any adaptive significance to the observed morphological difference. It would also be interesting to determine whether the fruit adapted forms express a different subset of surface localised proteins, and therefore determine whether *Phytomonas* cells undergo any differentiation events in response to the different host stages of their life cycle. The ability to infect a tomato fruit with culture forms of the cells would suggest that if there are different developmental stages then these are present in the culture conditions or can be induced on inoculation into fruit. Studies of the *Phytomonas* cell cycle and its regulation could be inform whether the morphologies observed are adaptive or are simply a result of cytoskeletal restructuring that occurs during the normal cell division cycle (Wheeler *et al.*, 2013) and

may be helpful in the study of the differentiation events that occur in other digenetic trypanosomatids.

While *P. serpens* is an ideal organism for laboratory work and Chapter 4 shows that it is amenable to genetic manipulation, it is also likely to be representative of only one subset of *Phytomonas* biology. *P. serpens* colonises the fruit of tomato plants and a study of the existing *Phytomonas* literature and isolations showed that 21% of isolations had been from fruit. A further 22% were isolated from phloem and 57% isolated from latex (Figure 1.1B). It is interesting to note that the only example of a lineage specific expansion of a cell surface protein family in *Phytomonas* occurred in *P. HART1*, the only species definitively associated with disease for which there is a sequenced genome available (Porcel *et al.*, 2014). Without more genomic sequence data available for other phloem-dwelling, disease-causing species of *Phytomonas* it is not possible to link this single observation of a gene family expansion with the disease symptoms associated with *P. HART1*.

## 6.3 The adaptation of an ancestral trypanosomatid to a suite of multicellular hosts

The study of mammalian infective species of the genera *Trypanosoma* and *Leishmania* has revealed a recurring theme in trypanosome evolution; that surface proteins play a key role in facilitating adaptation for survival in different host environments (MacGregor *et al.*, 2012). Figure 6.1 draws together the main findings of this thesis and contextualises them within the scientific community's current knowledge on the cell surfaces of trypanosomatid parasites. This schematic by no means summarises all knowledge of the cell surfaces of trypanosomatid parasites but aims to capture the most important features. The tree is rooted on free living *Bodo saltans* for which genomic information has recently become available. This revealed the protein bodonin as a major component of the plasma membrane involved in attachment to substrates and prey (Jackson *et al.*, 2016), The *B. saltans* genome also encodes amastins that are monophyletic but

outgroup to all trypanosomatid amastins, indicating that the expansion of this family in *Leishmania* species were independent of those that occurred in the lineage that lead to *B. saltans* (Jackson *et al.*, 2016). This has prompted the speculation that these proteins may play a role in the adaptation to parasitism in the trypanosomatids (Jackson *et al.*, 2016). Comparison of mammalian parasitic *Trypanosoma* and *Leishmania* species with free living *B. saltans* revealed patterns of loss of membrane transporters (ABC transporters), cathepsin cysteine proteases and ion transport concomitant with a genome wide reduction and streamlining. The genomes of *Phytomonas* were reported to be even further reduced, in particular with respect to metabolic genes and therefore their genomes are often described as streamlined as a result of adapting to their plant hosts (Porcel *et al.*, 2014). This may have a consequence for parasitism in respect to limiting host range and metabolic plasticity although these hypotheses have never been tested. Comparisons with *B. saltans* revealed gains in the *Leishmania* lineage of amastins and promastigote surface antigens (PSA) and in the *T. cruzi* lineage expansions in the mucin, trans sialidase and gp63 gene families (Jackson *et al.*, 2016). In *T. brucei* recently published work has revealed a number of validated cell surface proteins on the plasma and flagellar membranes and in the flagellar pocket, these have been summarised in Figure 6.1 as ESPs (enriched surface-labelled protein) (Gadelha *et al.*, 2015).

Prior to the work described in this thesis the only known components that had been identified on the *P. serpens* cell surface were those conserved in mammalian infective trypanosomatids such as cruzipain and gp63 (Elias *et al.*, 2012). This study expands knowledge of the *P. serpens* cells surface by enabling amastins to be added to this list. The consequences that this has for the biology of *Phytomonas* is unknown as the function of amastin proteins is unknown (Jackson, 2010). The expression of an expanded of a family of amastins in the *Leishmania* lineage has led to speculation that they are involved in parasite adaptation to the intracellular stage inside the mammalian host. This

discovery of amastin proteins on the cell surface of the plant infective *P. serpens* adds an additional layer of complexity to the evolutionary origin and maintenance of this family of proteins. I was able to localise one *Phytomonas* amastin to validate the cell surface localisation. However, as it was not possible within the scope of this study to produce lines defective in both copies of the alleles, it is impossible to speculate as to what the functional significance of these proteins on the cell surface of *Phytomonas* may be. It is also worth discussing that evidence shows the *P. serpens* cell surface to respond to environmental conditions (Elias *et al.*, 2012), and thus the biochemical characterisation presented in this thesis provides only a snapshot of the cell surface of this species. With development of *in vivo* growth methods, a cell surface proteome of the plant infective stage of the life cycle may uncover more biology in *P. serpens*.

Broadly, the proteomic and bioinformatic predicted cell surfaces of *P. serpens* revealed that it is decorated mostly with transmembrane proteins. This is in contrast to the cell surfaces of trypanosoma, including the new species presented in this thesis *T. carassii* (discussed later).

The comparative bioinformatic analysis of the first *Phytomonas* genomes and proteomic investigation into the predicted membrane or cell surface proteins of cultured *P. serpens* cells has started to elucidate the mechanisms by which an ancestral trypanosomatid evolved to colonise plant hosts. Results presented here are evidence in support of a hypothesis that the cell surface proteins in *Phytomonas* parasites of plants were not subjected to similar lineage specific expansion of surface proteins. This is in contrast to what has been observed in animal infective trypanosomatid species, where lineage specific expansions of surface protein genes are thought to be critical to host adaptation.

Few investigations have sought to study *Phytomonas* species in the paradigm of a plant pathogen. As a result, we know very little of how *Phytomonas* avoids the plant immune system. This is in part because of the environments they colonise as the immune responses of the latex and phloem tissues are poorly understood. The host specificity of

*Phytomonas* implies that there may be an inherent recognition mechanism. In interactions with susceptible plants the *Phytomonas* must have a specific strategy for disrupting the host immune response.

Among the orthogroups identified by the bioinformatic analysis as unique to *Phytomonas,* six were predicted to encode a protease domain. This finding extends a previous observation that there was a lineage specific expansion of putative secreted proteases in *P. HART1* (Porcel *et al.*, 2014). Porcel and colleagues consider that this class of Cathepsin D-like A1 family proteases may play a role in pathogenicity, based on the role that related proteases play in plant pathogenic fungi and in *Leishmania*. However, as discussed in the introduction to this thesis, our understanding of the role of proteases in plant pathology is predominantly in the context of phloem feeding insects (Kehr, 2006) and data from a protease deficient strain of the fungus *B. cinerea* indicates that the relationship between proteases and virulence is not straightforward in other systems (ten Have *et al.*, 2010). However, that both myself and Porcell *et al* independently discovered different protease gene families that are unique to the *Phytomonas* lineage, they may be worthy of further study to identify whether they are expressed in the insect stage, plant stage or both, and whether they have a role in virulence or survival in the plant host.

Here, I propose that the *Phytomonas* proteases could be fulfilling an effector-type role to supress the plant immune response. In the context of the traditional model of plant immunity, we do not know what *Phytomonas* pathogen associated molecular patterns (PAMPs) are recognised by the host and activate PAMP-triggered immunity. However, when PAMP-triggered immunity occurs in the plant-*Phytomonas* interaction, proteases could interfere with the plant immune response to break down the defence related proteins. Recognition of the parasite effector by the host plant activates effector triggered immunity (the type of immunity that often triggers a hypersensitive response where host cell death isolates the pathogen) (Jones and Dangl, 2006). In the case of *P. serpens* these effectors may not be recognised by the host, or the effector-triggered immunity

may not cross the threshold for the hypersensitive response, as this has not been observed in infected tomato fruit.

Plants have an innate immune system capable of recognising pathogen associated molecular patterns and pathogen effectors to trigger an immune response, but they lack an adaptive immune system (Jones and Dangl, 2006). In this scenario it is not clear what strategy the ancestral trypanosomatid cell surface has implemented to evade the plant host and prevent the triggering of the hypersensitive response. The evolution of a large family of genes encoding GPI anchored alpha helical glycoproteins to facilitate immune evasion is unlikely in the absence of an adaptive immune system. However, it is not possible to rule out an adaptive role for the expansion of the gp63 family in *P. HART1.* Conversely, it may be that some trypanosomatid parasites are not recognised and/or do not trigger an innate immune response, depending on the host resistance gene complement and microenvironment colonised. Thus the selective pressure on the cell surface proteins to hide or evade the host immune system is not present in all plant hosts or host environments occupied by the parasite. Under this hypothesis we do not find a suite of highly specialised cell surface proteins. Indeed, *Phytomonas* parasites have retained core trypanosomatid cell surface proteins such as gp63 and the amastin surface glycoproteins and evolved a small number of putative cell surface transmembrane proteins. Phylogenetic analysis of the amastin and amastin-like orthogroups identified in Chapter 3 showed that the *Phytomonas* amastins belong to the α and δ subfamilies (Jackson, 2010). The δ subfamily member is localised to all three surface membrane domains in *P. serpens*, as shown by fluorescent protein tagging in Chapter 4. Fluorescent protein tagging will be needed to confirm that the other proteins identified in this study are indeed surface localised. However, gene tagging proved to be more technically challenging than expected (given the precedent set in *Leishmania* and *Trypanosoma*) and although this work provides a firm basis on which to build, there is

extensive scope for further investigations into understanding how a trypanosomatid parasites interact with a plant host.

The comparative genomic method also identified some previously unknown biology regarding the well-studied animal infective *Trypanosoma* lineage. Using this method I identified a core set of 170 orthogroups that are shared amongst the trypanosomatids and are predicted to localise to the internal membranes or cell surface of at least one species of *Trypanosoma, Leishmania* and *Phytomonas*. Furthermore, the study of *T. carassii* uncovered 80 families of putative surface protein orthogroups (although it is likely that some of these are localised to internal membrane structures) that were previously thought to be unique to stercorarian trypanosomatids, but were probably present in the common ancestor of stercorarian and salivarian trypanosomes and subsequently lost from the *T. brucei* lineage. This highlights the importance of better sampling in improving our understanding of the evolution of cell surface proteins in these enigmatic parasites and indicates that there may be a large component of the conserved cell surface biology of the trypanosomatids that remains uncharacterised.

Interrogation of the *T. carassii* transcriptome provided evidence that the cell surface of this parasite may be primarily composed of two large families of alpha helical, GPI anchored proteins with biochemical evidence for one of these proteins being present on the cell surface. It may be that these two novel families of genes encode surface proteins are unique to the fish parasite and provide further evidence that mucin and MASP gene families are an invention in the stercorarian clade. Given precedent from other trypanosomatids, the expanded orthogroups may represent a suite of GPI anchored alpha helical membrane proteins, with possible functions in immune evasion in *T. carassii*. No evidence was found to suggest the presence of mucins or MASPs in the transcriptome of *T. carassii* (Figure 6.1). It could be that the two novel *T. carassii* protein orthogroups are related to either MASPs or mucins but that high levels of sequence divergence precludes detection of significant sequence identity/similarity. The structure

of a full length representative of OG8 is reminiscent of the VSG of *T. brucei* and it is tempting to speculate that these two families of proteins provide *T. carassii* with a similar function that VSG provides to *T. brucei*. Moreover, given that both gene families are rapidly evolving it is possible that they both evolved from an ancestral pre-VSG though sequence similarity is no longer readily detectable. Genomic evidence to demonstrate if these *T. carassii* genes are telomeric and also evidence of recombination of these *T. carassii* genes would provide more evidence in support of this hypothesis.

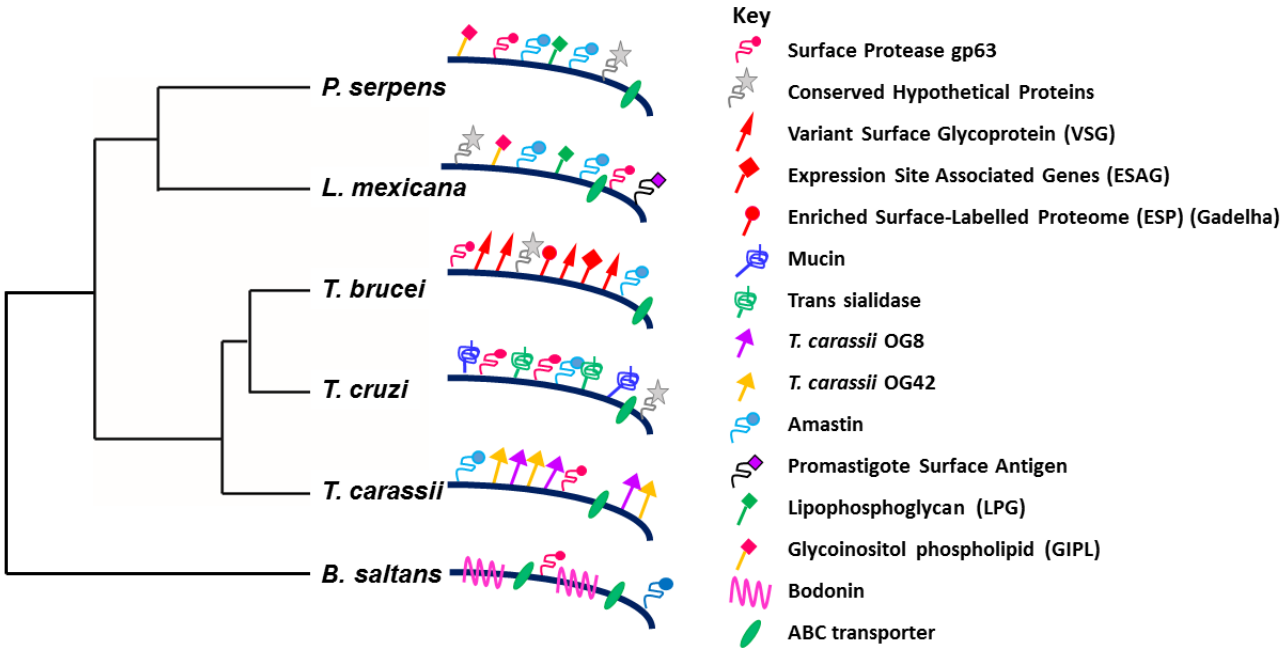



***Figure 6.1* A schematic summarising the main points about what we now know about the cell surface proteins of *P. serpens* and *T. carassii*.**

The schematic phylogeny is rooted on *Bodo saltans*. The schematic shows that the cell surface of *P. serpens* contains amastin like proteins, some of these are orthologous with amastins found in other trypanosomatids and some are unique to *Phytomonas.* The cell surface of *T. carassii* is believed to be similar to that of *T. brucei* with respect to the cell surface packed in alpha helical GPI anchored proteins from two large gene families.

## 6.4 Future directions

The core trypanosomatid cell surface proteome has been adapted for parasitism in a number of hosts and host environments, each scenario presenting a different suite of selection pressures. A general trend observed is that core surface orthogroups, conserved across all 3 genera, consist principally of transmembrane surface proteins predicted to be targeted to the cell surface by a signal peptide. Species specific cell surface proteins in *Trypanosoma* are targeted to the cell surface by a signal peptide and GPI anchor. A related observation was made in another study, that transporter-like proteins are conserved whereas receptor-like proteins are lineage specific (Gadelha *et al.*, 2015).

Transmembrane cell surface proteins are likely to be involved with cellular functions such as transport, whereas receptor-like and surface glycoproteins are more likely to be incorporated into the plasma membrane by a GPI anchor. Furthermore, the close association of a multi-pass transmembrane protein with the plasma membrane will likely place a limit on the amount of sequence variation that a protein can undergo before structural and functional conservation is lost. In contrast, structural conservation with sequence variation has been shown as a theme in the evolution of the variant surface glycoprotein (Higgins and Carrington, 2014). Investigation into the functions of these two classes of proteins in model systems such as *T. brucei* could elucidate whether this observed trend is the result of the involvement of transmembrane surface proteins in core cellular processes, thus placing functional constraint on their evolution. Furthermore, the investigation of the large number of hypothetical proteins predicted to localise to the cell surfaces of all 3 genera of digenetic parasites (Figure 6.1) has the potential to yield new insights into core functions performed by the trypanosomatid cell surface.

As a body of work, the data presented here highlight the importance of increasing our sampling within the trypanosomatids. Moreover, it reveals the limits that poor taxonomic

sampling places on our ability to elucidate the evolutionary history of gene families, even between closely related species.

A number of challenges were faced in this project. A number of these arose from working on *P. serpens,* an organism for which there were few protocols established, setting up standard protocols for working on a new organism as the first member of the lab, limited success with transfection and gene tagging, and fruit rotting when attempting to set up a fruit infection protocol. Now, with the tomato fruit infection protocol that has been established through this work, it would be interesting to develop this to repeat the biochemical isolation of the cell surface with the plant stage of the *P. serpens* life cycle. For *T. carassii,* the amount of material available limited what was possible experimentally. A greater quantity of genetic material would have enabled the isolation of genomic DNA for sequencing and thus study of genome organisation, in particular if OG8 and OG42, the two large gene families unique to *T. carassii*, are in arrays and telomeric reminiscent of VSG. It would also be important to know whether a single *T. carassii* cell expresses a single member of these families (as is the case for VSG in *T. brucei*) or whether dozens or hundreds of these genes decorate the surface of every cell.

This thesis indicates the likely involvement of amastins on the cell surface of *P. serpens* cells raising further questions concerning the role of these proteins in the adaptation of the trypanosomatids for parasitism. Optimisation and improvement of the genetic techniques for the tagging of proteins in *P. serpens* and other *Phytomonas* is necessary to test hypotheses regarding the function and localisation of putative cell surface proteins. Further sampling and genome sequencing of diverse *Phytomonas* species would provide more evidence in favour of an apparent lack of any clade-specific expansion of cell surface proteins to adapt to plant hosts. As a diverse and important class of parasites, this study has highlighted that there is a lot of core trypanosomatid biology that remains to be understood.

# *Appendix legend*

Appendix 1.1. Table of all known *Phytomonas* isolates..... 190

Appendix 2.1. Primer sequences for nested PCR tagging ..... 199

Appendix 2.2. Primer sequences for long primer PCR tagging ..... 201

Appendix 3.1. The results of LC MS/MS on *P. serpens* cells surface biotinylated with LC-biotin with purification and isolation of biotinylated material with streptavidin Dynabeads. ..... DIGITAL

Appendix 3.2. The results of LC MS/MS on *P. serpens* cells surface biotinylated with SS-biotin with purification and isolation of biotinylated material with streptavidin Dynabeads ……..DIGITAL

Appendix 3.3. Results of orthogroup clustering for *Phytomonas, Leishmania* and *Trypanosoma*..... DIGITAL

Appendix 3.4. The surface orthogroups for *Phytomonas, Leishmania* and *Trypanosoma* ..... DIGITAL

Appendix 3.5. The core set of surface orthogroups conserved in *Trypanosoma, Leishmania* and *Phytomonas* species..... 205

Appendix 4.0. Primer sequences and plasmid sequences..... DIGITAL

Appendix 5.1. The pan-trypanosome core repertoire of surface gene families. ..... 212

Appendix 5.2. The results of orthogroup clustering of the stercorarian and salivarian trypanosomes with *T. carassii* ……..……DIGITAL

Appendix 5.3. The surface orthogroups of the *Trypanosoma*…….……..……DIGITAL

Appendix 5.4. Results from tandem mass spectrometry from cell surface biotinylation and streptavidin pull down on *T. carassii*..... DIGITAL

Appendix 5.5. Carp proteins detected in the in the LC MS/MS of cell surface biotinylated *T. carassii* ……DIGITAL

## *Appendix 1.1 Table of all* Phytomonas *isolations*

| Host Family | Host Plant | Tissue | Insect Host | Place | Original name | Reference |
|---|---|---|---|---|---|---|
| **Apocynaceae** | *Allamanda cathartica* | Latex | | Americas | *Phytomonas sp.* | (Kastelein *et al.*, 1984) |
| **Zingiberaceae** | *Alpinia purpurata* | Phloem | | Americas | | Hunt in (Muller *et al.*, 1994) |
| **Zingiberaceae** | *Alpinia purpurata* | Phloem | | Americas | *Alp2* | (M. Dollet *et al.*, 2001) |
| **Zingiberaceae** | *Alpinia purpurata* | Phloem | | Americas | *Alp1* | (M. Dollet *et al.*, 2001) |
| **Amaranthaceae** | *Amaranthus sp.* | Phloem | | Europe | | Sanchez-Moreno in (Muller *et al.*, 1994) |
| **Anacardiaceae** | *Anacardium occidentale* | Fruit | | Americas | | (Conchon *et al.*, 1989) |
| **Annonaceae** | *Annona cherimola* | Fruit | | Europe | | (Sanchez-Moreno *et al.*, 1995) |
| **Asclepiadaceae** | *Araujia angustifolia* | Latex | | Americas | *Leptomonas elmassiani* | (Migone, 1916) |
| **Asclepiadaceae** | *Asclepias amplexicaulis* | Latex | | Americas | | (McGhee and McGhee, 1971) |
| **Apocynaceae** | *Sarcostemma clausum* | Latex | | Americas | | (McGhee and McGhee, 1971) |
| **Apocynaceae** | *Asclepias curassavica* | Latex | | Americas | | (Holmes, 1925) |
| **Apocynaceae** | *Asclepias curassavica* | Latex | | Americas | *Phytomonas elmassiani* | (Ruiz, 1958) |
| **Apocynaceae** | *Asclepias exaltata* | Latex | | Americas | | (McGhee and McGhee, 1971) |
| **Apocynaceae** | *Gomphocarpus fruticosus* | Latex | | Americas | | (Ayala *et al.*, 1975) |
| **Apocynaceae** | *Asclepias glaucescens* | Latex | | Americas | | (Hewitt, 1940) |
| **Apocynaceae** | *Asclepias humistrata* | Latex | | Americas | | (McGhee and McGhee, 1971) |
| **Apocynaceae** | *Asclepias incarnata* | Latex | | Americas | | (McGhee and McGhee, 1971) |
| **Apocynaceae** | *Asclepias michauxii* | Latex | | Americas | | (McGhee and McGhee, 1971) |
| **Apocynaceae** | *Asclepias nivea* | Latex | | Americas | *Herpetomonas oncopleti* | (Noguchi and Tilden, 1926) |
| **Apocynaceae** | *Asclepias obovata* | Latex | | Americas | | (McGhee and McGhee, 1971) |
| **Apocynaceae** | *Asclepias syriaca* | Latex | | Americas | *Herpetomonas elemassiani* | (Holmes, 1924) |
| **Apocynaceae** | *Asclepias syriaca* | Latex | | Americas | *Herpetomonas oncopleti* | (Noguchi and Tilden, 1926) |
| **Apocynaceae** | *Asclepias tuberosa* | Latex | | Americas | | (McGhee and McGhee, 1971) |
| **Apocynaceae** | *Asclepias viridiflora* | Latex | | Americas | | (McGhee and McGhee, 1971) |
| **Arecaceae** | *Attalea funifera* | Phloem | | Americas | | (Bezerra *et al.*, 1983) |

| | | | | | |
|---|---|---|---|---|---|
| **Oxalidaceae** | *Averrhoa carambola* | Fruit | Americas | | (Conchon *et al.*, 1989) |
| **Arecaceae** | *Bentinckia nicobarica* | Phloem | Americas | | (Kastelein, 1987) |
| **Bixaceae** | *Bixa orellana* | Fruit | Americas | *Phytomonas* | (Almeida *et al.*, 1990) |
| **Asclepiadaceae** | *Blepharodon nitidus* | Latex | Americas | *Phytomonas sp.* | (Kastelein *et al.*, 1984) |
| **Apocynaceae** | *Bonafousia tetrastachya* | Latex | Americas | | (Kastelein *et al.*, 1984) |
| **Leguminosae** | *Cajanus flavus* | Fruit | Americas | | (Cavazzana *et al.*, 1993) |
| **Solanaceae** | *Capsicum annuum* | Fruit | Americas | | (Kastelein and Camargo, 1990) |
| **Solanaceae** | *Capsicum praetermissum* | Fruit | Americas | | (Kastelein and Camargo, 1990) |
| **Urticaceae** | *Cecropia palmata* | Latex | Americas | | (Leeuw and Kastelein, 1987) |
| **Urticaceae** | *Cecropia surinamensis* | Latex | Americas | | (Kastelein *et al.*, 1984) |
| **Apocynaceae** | *Cerbera odollam* | Latex | Europe | | (Franchini, 1922a) |
| **Euphorbiaceae** | *Chamaesyce thymifolia* | Latex | Americas | *Phytomonas sp.* | (Santos *et al.*, 1994) |
| **Rutaceae** | *Citrus aurantium* | Fruit | Americas | | (Carrara *et al.*, 1992) |
| **Rutaceae** | *Citrus aurantium* | Fruit | Americas | *Phytomonas* Ora485 | (Luisa M. Catarino *et al.*, 2001) |
| **Rutaceae** | *Citrus aurantium* | Fruit | Americas | | (Fiorini *et al.*, 1990) |
| **Rutaceae** | *Citrus bergamia* | Fruit | Americas | | (Conchon *et al.*, 1989) |
| **Rutaceae** | *Citrus reticulata* | Fruit | Americas | | (Conchon *et al.*, 1989) |
| **Arecaceae** | *Cocos nucifera* | Phloem | Americas | *Phytomonas staheli* | (McGhee and McGhee, 1971) |
| **Arecaceae** | *Cocos nucifera* | Phloem | Americas | | (Menara *et al.*, 1988) |
| **Arecaceae** | *Cocos nucifera* | Phloem | Americas | | (Parthasarathy *et al.*, 1976) |
| **Rubiaceae** | *Coffea arabica* | Phloem | Americas | | (Stahel, 1931) |
| **Rubiaceae** | *Coffea liberica* | Phloem | Americas | *Phytomonas leptovasorum* | (Stahel, 1931) |
| **Apocynaceae** | *Cynanchum acutum* | Latex | Europe | | (Zotta, 1924) |
| **Apocynaceae** | *Cynanchum laeve* | Latex | Americas | | (McGhee and McGhee, 1971) |
| **Apocynaceae** | *Cynanchum sibiricum* | Latex | Asia | | (Podlipaev, 1986) |
| **Apocynaceae** | *Pergularia daemia* | Latex | Africa | | (Rodhain, 1926) |
| **Arecaceae** | *Elaeis guineensis* | Phloem | Americas | | (M. Dollet *et al.*, 1977) |
| **Arecaceae** | *Elaeis guineensis* | Phloem | Americas | | (Gargani *et al.*, 1992) |

| | | | | | |
|---|---|---|---|---|---|
| **Arecaceae** | *Elaeis guineensis* | Phloem | Americas | *Phytomonas staheli* | (McGhee and McGhee, 1971) |
| **Euphorbiaceae** | *Elaeophorbia drupifera* | Latex | Europe | *None* | (Franchini, 1922a) |
| **Myrtaceae** | *Eugenia uniflora* | Fruit | Americas | | (Cavazzana *et al.*, 1996) |
| **Euphorbiaceae** | *Euphorbia adenoptera* | Latex | Americas | | (Harvey and Lee, 1943) |
| **Euphorbiaceae** | *Euphorbia antiquorum* | Latex | Europe | *None* | (Franchini, 1922b) |
| **Euphorbiaceae** | *Euphorbia bacteriana* | Latex | Americas | | (M. Dollet *et al.*, 1979) |
| **Euphorbiaceae** | *Euphorbia braziliensis* | Latex | Americas | | (Siqueira-Jaccoud, 1955) |
| **Euphorbiaceae** | *Euphorbia buxifolia* | Latex | Americas | | (Harvey and Lee, 1943) |
| **Euphorbiaceae** | *Euphorbia callitrichoides* | Latex | Americas | *None* | (Strong, 1924) |
| **Euphorbiaceae** | *Euphorbia cereiformis* | Latex | Europe | *None* | (Franchini, 1923b) |
| **Euphorbiaceae** | *Euphorbia cereiformis* | Latex | Europe | *Herpetomonas euphorbiae* | (Nieschulz, 1924) |
| **Euphorbiaceae** | *Euphorbia characias* | Latex | Europe | *None* | (Poisson, 1930) |
| **Euphorbiaceae** | *Euphorbia characias* | Latex | Europe | *Phytomonas sp.* | (M. Dollet, 1984) |
| **Euphorbiaceae** | *Euphorbia coerulescens* | Latex | Europe | *Trypanosoma euphorbiae* | (Franchini, 1922f) |
| **Euphorbiaceae** | *Euphorbia conferta* | Latex | Americas | | (Harvey and Lee, 1943) |
| **Euphorbiaceae** | *Euphorbia cupani* | Latex | Europe | *None* | (Visentini, 1914) |
| **Euphorbiaceae** | *Euphorbia cyathophora* | Latex | Americas | | (Harvey and Lee, 1943) |
| **Euphorbiaceae** | *Euphorbia deltoidea* | Latex | Americas | | (Harvey and Lee, 1943) |
| **Euphorbiaceae** | *Euphorbia dulcis* | Latex | Europe | *None* | (Laveran and Franchini, 1920) |
| **Euphorbiaceae** | *Euphorbia esula* | Latex | Europe | *None* | (Zotta, 1921) |
| **Euphorbiaceae** | *Euphorbia falcata* | Latex | Europe | *None* | (Laveran and Franchini, 1920) |
| **Euphorbiaceae** | *Euphorbia foliolosa* | Latex | Americas | | (M. Dollet *et al.*, 1979) |
| **Euphorbiaceae** | *Euphorbia gemella* | Latex | Americas | | (Mcghee and Postell, 1976) |
| **Euphorbiaceae** | *Euphorbia gerardiana* | Latex | Europe | *None* | (Galli-Valerio, 1921) |
| **Euphorbiaceae** | *Euphorbia grandis* | Latex | Europe | *None* | (Laveran and Franchini, 1920) |
| **Euphorbiaceae** | *Euphorbia helioscopia* | Latex | Europe | *None* | (Zotta, 1921) |
| **Euphorbiaceae** | *Euphorbia heterophylla* | Latex | Americas | | (Harvey and Lee, 1943) |
| **Euphorbiaceae** | *Euphorbia heterophylla* | Latex | Americas | *Phytomonas davidi* | (Mcghee and Postell, 1976) |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Euphorbiaceae** | *Euphorbia hirta* | Latex | | Americas | | (Harvey and Lee, 1943) |
| **Euphorbiaceae** | *Euphorbia hirta* | Latex | | Asia | | Pradesh in (Marín *et al.*, 2008) |
| **Euphorbiaceae** | *Euphorbia hirta* | Latex | | Europe | *Phytomonas sp.* | (Kastelein *et al.*, 1984). |
| **Euphorbiaceae** | *Euphorbia hirtella* | Latex | | Americas | | (M. Dollet *et al.*, 1979) |
| **Euphorbiaceae** | *Euphorbia hypericifolia* | Latex | | Africa | *None* | (Lafont, 1910). |
| **Euphorbiaceae** | *Euphorbia hyssopifolia* | Latex | | Americas | | (Harvey and Lee, 1943) |
| **Euphorbiaceae** | *Euphorbia hyssopifolia* | Latex | | Americas | *Phytomonas sp.* | (Attias and De Souza, 1986) |
| **Euphorbiaceae** | *Euphorbia indica* | Latex | | Africa | *None* | (Rodhain and Bequaert, 1911) |
| **Euphorbiaceae** | *Euphorbia laro* | Latex | | Europe | *None* | (Franchini, 1922d) |
| **Euphorbiaceae** | *Euphorbia lasiocarpa* | Latex | | Americas | | (M. Dollet *et al.*, 1983) |
| **Euphorbiaceae** | *Euphorbia lasiocarpa* | Phloem | | Americas | | (M. Dollet *et al.*, 1983) |
| **Euphorbiaceae** | *Euphorbia maculata* | Latex | | Americas | | (Harvey and Lee, 1943) |
| **Euphorbiaceae** | *Euphorbia chamaesula* | Latex | | Americas | | (Harvey and Lee, 1943) |
| **Euphorbiaceae** | *Euphorbia neriifolia* | Latex | | Europe | *None* | (Laveran and Franchini, 1920) |
| **Euphorbiaceae** | *Euphorbia oblongata* | Latex | | Americas | | (Harvey and Lee, 1943) |
| **Euphorbiaceae** | *Euphorbia officinarum* | Latex | | Europe | *None* | (Franchini, 1923a) |
| **Euphorbiaceae** | *Euphorbia peploides* | Latex | | Africa | *None* | (Sergent, 1921) |
| **Euphorbiaceae** | *Euphorbia peplus* | Latex | | Europe | *None* | (Franca, 1911) |
| **Euphorbiaceae** | *Euphorbia pilulifera* | Latex | | Asia | *Phytomonas davidi* | (Donovan, 1909) |
| **Euphorbiaceae** | *Euphorbia pilulifera* | Latex | | Africa | *Leptomonas davidi* | (Lafont, 1909) |
| **Euphorbiaceae** | *Euphorbia pinea* | Latex | *Stenocephalus agilis* | Europe | *Phytomonas sp.* | (M. Dollet *et al.*, 1982) |
| **Euphorbiaceae** | *Euphorbia pinetorum* | Latex | | Americas | | (Harvey and Lee, 1943) |
| **Euphorbiaceae** | *Euphorbia prostrata* | Latex | | Americas | | (M. Dollet *et al.*, 1979) |
| **Euphorbiaceae** | *Euphorbia schimperiana* | Latex | | Europe | *None* | (Visentini, 1914) |
| **Euphorbiaceae** | *Euphorbia segetalis* | Latex | | Europe | *None* | (Franca, 1911) |
| **Euphorbiaceae** | *Euphorbia splendens* | Latex | | Europe | *None* | (Franchini, 1923a) |
| **Euphorbiaceae** | *Euphorbia striata* | Latex | | Africa | *None* | (Fantham, 1925) |
| **Euphorbiaceae** | *Euphorbia thymifolia* | Latex | | Africa | *None* | (Lafont, 1910) |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Euphorbiaceae** | *Euphorbia tirucalli* | Latex | | Africa | *Phytomonas tirucalli* | (Reichenow, 1940) |
| **Euphorbiaceae** | *Euphorbia tracyi* | Latex | | Americas | | (Harvey and Lee, 1943) |
| **Euphorbiaceae** | *Euphorbia uralensis* | Latex | | Asia | *None* | (Ioff, 1927) |
| **Euphorbiaceae** | *Euphorbia virosa* | Latex | | Europe | *None* | (Laveran and Franchini, 1920) |
| **Euphorbiaceae** | *Excoecaria emarginata* | Latex | | Europe | *None* | (Franchini, 1922a) |
| **Moraceae** | *Ficus benjamina* | Latex | | Europe | | (Franchini, 1922a) |
| **Moraceae** | *Ficus carica* | Latex | | Europe | | (Franchini, 1922d) |
| **Moraceae** | *Ficus costaricana* | Latex | | Americas | *Phytomonas tortuosa* | (Ruiz, 1958) |
| **Moraceae** | *Ficus edulis* | Latex | | Africa | *Herpetomonas ficuum* | (Fantham, 1925) |
| **Moraceae** | *Ficus hochstetteri* | Latex | | Africa | *Herpetomonas ganorae* | (Franchini, 1931) |
| **Moraceae** | *Ficus parietalis* | Fruit | | Europe | | (Franchini, 1922c) |
| **Moraceae** | *Ficus parietalis* | Latex | | Europe | | (Franchini, 1922c) |
| **Moraceae** | *Ficus scabra* | Latex | | Oceania | | (Bancroft, 1927) |
| **Moraceae** | *Ficus scabra* | Latex | | Oceania | *Herpetomonas bancrofti* | (Holmes, 1931) |
| **Asclepiadaceae** | *Funastrum clausum* | Latex | | Americas | | (Harvey and Lee, 1943) |
| **Apocynaceae** | *Funtumia elastica* | Latex | | Europe | *Herpetomonas funtumiae* | (Franchini, 1922e) |
| **Leguminosae** | *Glycine max* | Fruit | | Americas | | (Jankevicius *et al.*, 1987) |
| **Asclepiadaceae** | *Hoya australis* | Latex | | Oceania | | (Bancroft, 1927) |
| **Euphorbiaceae** | *Jatropha macrantha* | Latex | | Americas | *Phytomonas Jma* | (Burstein and Romero, 1956) |
| **Solanaceae** | *Lycopersicon esculentum* | Fruit | | Americas | | (Fiorini *et al.*, 1986) |
| **Solanaceae** | *Lycopersicon esculentum* | Fruit | *Nezara viridula* | Africa | *Phytomonas serpens* | (Gibbs, 1957) |
| **Solanaceae** | *Lycopersicon esculentum* | Fruit | | Americas | *Phytomonas serpens* | (Jankevicius *et al.*, 1989) |
| **Solanaceae** | *Lycopersicon esculentum* | Fruit | | Europe | | (Sanchez-Moreno *et al.*, 1995) |
| **Solanaceae** | *Lycopersicon esculentum* | Fruit | | Americas | *Phytomonas serpens 9T** | (Santos *et al.*, 2006) |
| **Solanaceae** | *Lycopersicum lycopersicum* | Fruit | *Phthia picta* | Americas | | (Brazil *et al.*, 1990) |
| **Rosaceae** | *Malus sp.* | Fruit | | Americas | | (Cavazzana *et al.*, 1995) |
| **Apocynaceae** | *Mandevilla scabra* | Latex | | Americas | *Phytomonas sp* | (Kastelein *et al.*, 1984) |
| **Euphorbiaceae** | *Manihot esculenta* | Latex | | Americas | *Phytomonas Francai* | (Vainstein and Roitman I, 1986) |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Euphorbiaceae** | *Manihot palmata* | Latex | | Americas | *Phytomonas Francai* | (Aragão, 1927) |
| **Arecaceae** | *Maximiliana maripa* | Phloem | | Americas | | (Slobbe *et al.*, 1978) |
| **Apocynaceae** | *Morrenia odorata* | Latex | | Americas | *Leptomonas bordasi* | Mignone in |
| **Moraceae** | *Morus sp.* | Fruit | | Americas | | (Cavazzana *et al.*, 1993) |
| **Moraceae** | *Morus sp.* | Fruit | | Americas | | (Fiorini *et al.*, 1995) |
| **Solanaceae** | *Nicandra physaloides* | Fruit | | Americas | | (Kastelein and Camargo, 1990) |
| **Passifloraceae** | *Passiflora edulis* | Fruit | | Americas | | (Conchon *et al.*, 1989) |
| **Apocynaceae** | *Pergularia extensa* | Latex | | Africa | | (Vickerman, 1962) |
| **Leguminosae** | *Phaseolus vulgaris* | Fruit | | Americas | | (Jankevicius *et al.*, 1987) |
| **Solanaceae** | *Physalis angulata* | Fruit | | Americas | | (Kastelein and Camargo, 1990) |
| **Rosaceae** | *Prunus persica* | Fruit | | Americas | | (Conchon *et al.*, 1989) |
| **Lythraceae** | *Punica granatum* | Fruit | | Americas | | (Conchon *et al.*, 1989) |
| **Lythraceae** | *Punica granatum* | Fruit | | Americas | *Phytomonas* | (Catarino *et al.*, 2001) |
| **Lythraceae** | *Punica sp.* | Fruit | | Americas | | (Catarino *et al.*, 1991) |
| **Apocynaceae** | *Rhabdadenia biflora* | Latex | | Americas | *Phytomonas sp.* | (Kastelein *et al.*, 1984) |
| **Arecaceae** | *Roystonea regia* | Phloem | | Americas | | (Attias *et al.*, 1987) |
| **Apocynaceae** | *Sarcostemma australe* | Latex | | Oceania | | (Bancroft, 1927) |
| **Apocynaceae** | *Sarcostemma clausum* | Latex | | Americas | | (McGhee and McGhee, 1971) |
| **Apocynaceae** | *Secamone elliptica* | Latex | | Oceania | | (Bancroft, 1927) |
| **Sapotaceae** | *Sideroxylon inerme* | Latex | | Europe | | (Franchini, 1922a) |
| **Solanaceae** | *Solanaceae sp.* | | *Arvelis albopunctatus* | Americas | | (Kastelein and Camargo, 1990) |
| **Solanaceae** | *Solanum americanum* | Fruit | | Americas | | (Kastelein and Camargo, 1990) |
| **Solanaceae** | *Solanum concinnum* | Fruit | | Americas | | (Kastelein and Camargo, 1990) |
| **Solanaceae** | *Solanum pseudocapsicum* | Fruit | | Americas | | (Kastelein and Camargo, 1990) |
| **Solanaceae** | *Solanum erianthum* | Fruit | | Americas | | (Kastelein and Camargo, 1990) |
| **Solanaceae** | *Solanum gilo* | Fruit | | Americas | | (Fiorini *et al.*, 1986) |
| **Solanaceae** | *Solanum robustum* | Fruit | | Americas | | (Kastelein and Camargo, 1990) |
| **Solanaceae** | *Solanum robustum* | Fruit | | Americas | | (Kastelein and Camargo, 1990) |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Solanaceae** | *Solanum viarum* | Fruit | | Americas | | (Kastelein and Camargo, 1990) |
| **Apocynaceae** | *Thevetia peruviana* | Latex | | Europe | | (Franchini, 1922e) |
| **Leguminosae** | *Trifolium sp.* | Phloem | | Europe | | Sanchez-Moreno in (Muller *et al.*, 1994) |
| **Vitaceae** | *Vitis vinifera* | Fruit | | Americas | | (Carrara *et al.*, 1992) |
| **Poaceae** | *Zea mays* | Fruit | *Veneza zonata* | Americas | *Herpetomonas mcgheei* | (Jankevicius *et al.*, 1993) |
| **Poaceae** | *Zea mays* | Fruit | | Americas | *Phytomonas mcgheei* | (Teixeira *et al.*, 1997) |
| | *Unknown (isolated from vector)* | | *Metochus* sp. | Asia | Ch12:402 | (Votýpka *et al.*, 2010) |
| **Bromeliaceae** | *Ananas ananas* | Fruit | | Americas | *Herpetomonas* Pin772 | (Catarino *et al.*, 2001) |
| **Euphorbiaceae** | *Euphorbia prostrata* | Latex | | Americas | E.pro.Vz | (Sturm *et al.*, 2007) |
| **Euphorbiaceae** | *Euphorbia heterophylla* | Latex | | Americas | E.het.Vz | (Sturm *et al.*, 2007) |
| **Rutaceae** | *Citrus bergamia* | Fruit | | Americas | Berg.Br | (Michel Dollet *et al.*, 2000) |
| **Euphorbiaceae** | *Jatropha macrantha* | Latex | | Americas | Lima.Pe | (Michel Dollet *et al.*, 2000) |
| **Euphorbiaceae** | *Euphorbia hirta* | Latex | | Americas | E.hi.Su | (Michel Dollet *et al.*, 2000) |
| **Apocynaceae** | *Mandevilla scabra* | Latex | | Americas | Mand.sc.Br | (Michel Dollet *et al.*, 2000) |
| **Anacardiaceae** | *Mangifera indica* | Fruit | | Europe | Mango.Sp | (Michel Dollet *et al.*, 2000) |
| **Solanaceae** | *Lycopersicon esculentum* | Fruit | | Europe | Tom.Sp | (Michel Dollet *et al.*, 2000) |
| **Euphorbiaceae** | *Euphorbia hirta* | Latex | | Asia | E.hi.Ind | (Michel Dollet *et al.*, 2000) |
| **Euphorbiaceae** | *Euphorbia hirta* | Latex | | Africa | E.hi.Su | (Michel Dollet *et al.*, 2000) |
| **Euphorbiaceae** | *Euphorbia pinea* | Latex | | Europe | EM1* | (Michel Dollet *et al.*, 2000) |
| **Euphorbiaceae** | *Manihot esculenta* | Latex | | Americas | Mani.Br | (Michel Dollet *et al.*, 2000) |
| **Euphorbiaceae** | *Euphorbia hirta* | Latex | | Americas | E.hi.Gu | (Michel Dollet *et al.*, 2000) |
| **Euphorbiaceae** | *Euphorbia hirta* | Latex | | Americas | E.hi.Vz | (Michel Dollet *et al.*, 2000) |
| **Euphorbiaceae** | *Euphorbia hyssopifolia* | Latex | | Americas | E.hy.Gu | (Michel Dollet *et al.*, 2000) |
| **Zingiberaceae** | *Alpinia purpurata* | Phloem | | Americas | Alp1 | (Michel Dollet *et al.*, 2000) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus croupius* | Americas | Hart1* | (Dollet *et al.*, 2001, Dollet *et al.,* 2000) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus lamelliger* | Americas | Hart2 | (Dollet *et al.*, 2001, Dollet *et al.,* 2000) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Ochlerus* sp. | Americas | Hart3 | (Dollet *et al.*, 2001, Dollet *et al.,* 2000) |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus* sp. | Americas | Hart11 | (Dollet *et al.*, 2001, Dollet *et al.,* 2000) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Ochlerus* sp. | Americas | Hart13 | (Dollet *et al.*, 2001, Dollet *et al.,* 2000) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus croupius* | Americas | Hart22 | (Dollet *et al.*, 2001, Dollet *et al.,* 2000) |
| **Arecaceae** | *Elaeis guineensis* | Phloem | *Lincus tumidifrons* | Americas | Mar.1 | (Dollet *et al.*, 2001, Dollet *et al.,* 2000) |
| **Arecaceae** | *Elaeis guineensis* | Phloem | *Lincus tumidifrons* | Americas | Mar.3 | (Dollet *et al.*, 2001, Dollet *et al.,* 2000) |
| **Arecaceae** | *Elaeis guineensis* | Phloem | | Americas | Mar.7 | (Michel Dollet *et al.*, 2000) |
| **Apocynaceae** | *Asclepias curasavica* | Phloem | | Americas | Ascl.cu.Vz | (Michel Dollet *et al.*, 2000) |
| **Apocynaceae** | *Allamanda cathartica* | Latex | | Americas | *Aca* | (Michel Dollet *et al.*, 2000) |
| **Euphorbiaceae** | *Euphorbia biflora* | Latex | | Americas | Rhab.Sur | (Michel Dollet *et al.*, 2000) |
| **Apocynaceae** | *Blepharodon nitidus* | Latex | | Americas | Bleph.Sur | (Michel Dollet *et al.*, 2000) |
| **Zingiberaceae** | *Alpinia purpurata* | Phloem | | Americas | Alp3 | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus croupius* | Americas | *Hart4* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus croupius* | Americas | *Hart5* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus* sp. | Americas | *Hart6* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus croupius* | Americas | *Hart9* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus* sp. | Americas | *Hart14* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus croupius* | Americas | *Hart15* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus croupius* | Americas | *Hart16* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus croupius* | Americas | *Hart17* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus croupius* | Americas | *Hart18* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus croupius* | Americas | *Hart19* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus croupius* | Americas | *Hart20* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Cocos nucifera* | Phloem | *Lincus croupius* | Americas | *Hart21* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Elaeis guineensis* | Phloem | *Lincus lethifer* | Americas | *Mar.2* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Elaeis guineensis* | Phloem | *Lincus tumidifrons* | Americas | *Mar.4* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Elaeis guineensis* | Phloem | *Lincus* sp. | Americas | *Mar.5* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Elaeis guineensis* | Phloem | *Lincus tumidifrons* | Americas | *Mar.6* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Elaeis guineensis* | Phloem | *Lincus* sp. | Americas | *Mar.8* | (M. Dollet *et al.*, 2001) |

| | | | | | |
|---|---|---|---|---|---|
| **Arecaceae** | *Elaeis guineensis* | Phloem | Americas | *Mar.9* | (M. Dollet *et al.*, 2001) |
| **Arecaceae** | *Elaeis guineensis* | Phloem | Americas | *Mar.10* | (M. Dollet *et al.*, 2001) |

## *Appendix 2.1: Primer sequences for nested PCR tagging*

| OLIGO NAME | SEQUENCE 5' > 3' |
|---|---|
| 19.1260_5'TF | GAACGACGGGTCCTGGAAAGAA |
| 19.1260_NF | TACGACTGAAGCCCTTGAAA |
| 19.1260_3'TF | CCCTCTGGTTATGTGTGGGAGGGCTAACCATGGTTAATGGGAGCTCTTTTTTAACTCAG |
| 19.1260_5'TR | GGATCTTGGTTAGTATGGACCTCCATAAGCTTACAAGCTGTGCTTTTAGCGCCTCCGC |
| 19.1260_3'TR | GTTACCGTCTCCTTTGCTACTT |
| 19.1260_NR | GAGTGATGGGAAACCAAGGGA |
| 264.1080_5'TF | CCATGGGTTGCGTCTTGTGGATA |
| 264.1080_5'TR | TCTTCTTGAGACAAAGGCATGAATTCTTTTGCGCAAACGAGTAG |
| 264.1080_3'TF | AACAGCCACGGACAGCCATGGATCCGTCAAGTGGATCT |
| 264.1080_3'TR | CCATGGAAATCAGTCATCTCAACACA |
| 264.1080_NF | CCTCAATGCTGTTGGGATTCT |
| 264.1080_NR | AGGAATCCAGAACAACCGCGAATA |
| 604.1040_5TF | AGTAGAAGCGTGATGTGACCT |
| 604.1040_5TR | CCCGGTGAACAGCTCCTCGCCCTTGCTCACTGGCGGTATGAGTGTTGTCGGACGT |
| 604.1040_3TF | CTGCCCTCTGGTTATGTGTGGGAGGGCTAACCATGGAGGTATGCCCCGATATCGACAAGACCGCAA |
| 604.1040_3TR | ATACTCTACGAGGACCTGGA |
| 604.1040_NF | GTGGTTAATGTGCGGCTCTTCT |
| 604.1040_NR | TTATTGTACGGGTCCTCCAGA |
| 926.1000_5'TF | ATATGCGGAGATAGTAGGGATGGT |
| 926.1000_5'TR | TCTTCTTGAGACAAAGGCATGAATTCTGCGATCCTTACGAAAAAGA |
| 926.1000_3'TF | TTACATATGCAACCCATGGATCCCTTGTACAGCTCGTCCAT |
| 926.1000_3'TR | GCGCAACCCGCAATGAAAA |
| 926.1000_NF | ACCTACGTGTTGTCTCCATCAAG |
| 926.1000_NR | CCTATATCCATGCCTTCCGCG |
| 1124.1000_3'TF | CTAACCAAGATCCACTTGACGGATCCATGGCCTTTCCTAAGCTCTA |
| 1124.1000_3'TR | CTCGTGCTCACCTTACTTAACA |
| 1124.1000_5'TF | CCCGCAGATTAGACGCTCAT |
| 1124.1000_5'TR | TCTTCTTGAGACAAAGGCATGAATTCTTGTCTGTGAACAGGTAGAT |
| 1124.1000_NR | GATGCCGTTTAGAAGCCCA |
| 1124.1000_NF | GGCTGTAGCCTTTCCAATACTT |
| 422.1010_3'TF | GTTATGTGTGGGAGGGCTAACCATGGTGATGGTGCTCTACGCCGTT |
| 422.1010_3'TR | GAGCAAGCACCCAAGCAT |
| 422.1010_5'TF | CGTTTGGCATTGATCTTCTGGA |
| 422.1010_5'TR | TGGTTAGTATGGACCTCCATAAGCTTCAAGGTAGTGGGGGGACGAT |
| 422.1010_NF | CGTCTAAGCGTTACCAACATGA |
| 422.1010_NR | CATCGCTCCTTCATTGCGGT |
| 422.1030_3'TF | GAGGTCTTTCGGCGTGTTCT |
| 422.1030_3'TR | TGGATTCTTCTTGAGACAAAGGCATGAATTCTTCACCCACCGATATATTATAGCAA |
| 422.1030_5'TF | CCATACTAACCAAGATCCACTTGACGGATCCATGCTTATAACATTTTCCAATACGA |
| 422.1030_5'TR | CTCCGGTCGATCATGTTGGT |
| 422.1030_NF | ATGCCTTGTCTTTCCACACT |
| 422.1030_NR | TTATTGTACGGGTCCTCCAGA |
| 270.1010_3'TF | GAGGTCCATACTAACCAAGATCCACTTGACGGATCCAATTTTGCGAAGGCTGTTTC |
| 270.1010_3'TR | CTGTTGTTCACTCCAAGCCCAT |
| 270.1010_5'TF | CCGCAGAGAGAACATCGCAT |
| 270.1010_5'TR | CAGCGCCATCAGTTCGGTTTTCGTGCGCATAAGCTTCGTTGCAGTTGAGTTTCTGT |
| 270.1010_NF | CTCCGACCCACACGCACTTA |

| | |
|---|---|
| 270.1010_NR | GACGTTCGTCGCGCATTTATT |
| 777.1000_3'TF | CCCTCTGGTTATGTGTGGGAGGGCTAACCATGGAATCGGTACCCC ACTCCTGCATGCAG |
| 777.1000_3'TR | TTAGAGTTAAAAAATTAAAAATC |
| 777.1000_5'TF | CACAGGTTCCTCTTGATAGACA |
| 777.1000_5'TR | GGATCTTGGTTAGTATGGACCTCCATAAGCTTTACCGCCTCATTTTC CCTCTGCTGTGG |
| 777.1000_NF | GTAGCTCACTAAAGCGGACGGATG |
| 777.1000_NR | ACACCAGAACAGACAAATTG |
| 677.1000_3'TF | GTTATGTGTGGGAGGGCTAACCATGGTAGCTTGCTGTGCTTTGCAC |
| 677.1000_3'TR | CCCTCAAGGTCTCCGATGCT |
| 677.1000_5'TF | GCAATATGCCGTCGAACCCAA |
| 677.1000_5'TR | TGGTTAGTATGGACCTCCATAAGCTTCGAATCCTGATTGCGCTCCG |
| 677.1000_NF | CCACGAGTTTCGGCTCCATCT |
| 677.1000_NR | GAGCATAAGGCACTGTTAGACAGA |
| 222.1080_5'TF | TTCAAGGCATTTCGGAGTGT |
| 222.1080_5'TR | TCTTGAGACAAAGGCATGAATTCTTGTCTGTGAACAGATA |
| 222.1080_3'TF | CCAAGATCCACTTGACGGATCCATGGCCTTTCCTAAGCTC |
| 222.1080_3'TR | CGTTTAGAAGCCCAAGTGCA |
| 222.1080_NF | GCTGTGAGGGACGTGTTT |
| 222.1080_NR | GATCACAGTGAGGACGAATCCTT |
| 352.1060_3'TF | CTAACCAAGATCCACTTGACGGATCCATGCGTTTCACGAAGGTTGC |
| 352.1060_3'TR | AATCACGCAAGGGAAGCTCT |
| 352.1060_5'TF | CCCTCTGCCTCTCACGTATA |
| 352.1060_5'TR | TCTTCTTGAGACAAAGGCATGAATTCACGCCCGATATGGTGCCCCC |
| 352.1060_NF | GCTTCGTATGAGTCCACGTAAA |
| 352.1060_NR | CGTGTCTGGGTCCTTCGGAT |
| 926_validation_1 | ATTCCAGCTAGACGCGAGTAA |
| 926_validation_2 | GCAACGAAACGAGTGAGAAA |
| 926_validation_3 | TTGATTGTAGCCGTTGCTCT |
| 926_validation_4 | GTCCTGCTGGAGTTCGTGA |

## *Appendix 2.2: Primer sequences for long primer PCR tagging*

| OLIGO NAME | SEQUENCE |
|---|---|
| 1.1110_F | CACTCCATTATGAGGGTAACAATCGGAACATTGCTGCTTAATGTGATA ATTGCAGTGGTTCTATTTTTATGGTCAGAGCGTTACATGGAAAGAGTAT TAATGAAAAAAAGGAAAAATGAAACTAGTGTGAGCAAGGGCGA |
| 1.1110_R | TGCGAAAATGGACATAATAGGGAGATGATACAGAGATGGAGCTCAGA CATGTTCATCACATACAGAGGGAATGGCCTCTATCAATTTTAGGTGTC CAATGGATAAAAAGTCTGCCATTTAGCCCTCCCACACATAACCAG |
| 103.1080_F | TGTTTAGCTTTGATCGTGGTCGGCTCCTTGGTGGCGACTGGTTCATAC GGCATTTATGATATATTAGACAAGGATGGTAGCTCAAGCATCTTTATGT TTGGAAAGTTGTGCATTCCACAAACTAGTGTGAGCAAGGGCGA |
| 103.1080_R | ATAAAACAATCAGATATAGACAAATGTATTTTATAGAAGGAATCGTTCA TCGCTATATCAAATAATCTCAATATTCCTTTGTTGTTGTTGTAAAAATAC TCACCCATAAATTGCATTTTAGCCCTCCCACACATAACCAG |
| 1047.1010_F | TACGTGATTTGTGTTGGAAATCTTTTTAGCCCGATTTTCAAGGAGGCG TACAAGCGCAATCCNGATCTCTGCAATGGTCCACACCGTCGTGCCGC CCGGTGGCATCCTCTCTGGTGCTTTACTAGTGTGAGCAAGGGCGA |
| 1047.1010_R | GAGAAGACAGTAAGGAATGTCATGACAACGAGAAGGATTATGGTGGT AACAAGGCCGGCATTGTTAGTGGCACCGGGAAGGCCAAGAATGCCG GCCCCTATAGAGCCACTTGCCATGTTAGCCCTCCCACACATAACCAG |
| 1195.1000_F | GTTTCGCTGCCTGAACTAGAAACAGATGGAGAGCGCTCGATGGCTGG GCAAATCTGTGATATGGTTCTCCTGTTTAGTGTCACAGCCGGAGAATT CCTGCAAACCTTTTGTTTGGATCCCACTAGTGTGAGCAAGGGCGA |
| 1195.1000_R | GGGCCGAGGCCATTTACGATCCCGTGATTCACAAGATACACGATGGA ATGGTCTATCCATGTTACAGAAGGTTGGAGATAGTAATCGATCTCATTA CCTCTTGTTTTTCTTGGTATTTTAGCCCTCCCACACATAACCAG |
| 149.1010_F | CATGAGTCTGATTCCGCATCTCCCAGCAAGAAAGCCAAAGAGAACGA AAGTAATGCCAGTTCTTCCGAACCCACCCACCAGACCGAAGCAGAGC AAAATAGGGATCCCAAGCATTCCGTTACTAGTGTGAGCAAGGGCGA |
| 149.1010_R | ACTCGATAAATATGAGTTGTTATTCTACTGATATTTTTTTCTTGCATACT AATAATCGAAAAACATCAAGATATTAATATACTTCGCATTGTACTGCTTT GACGCTGCAATGTGGTTTTAGCCCTCCCACACATAACCAG |
| 155.1000_F | GGAGATGATACAGTAGAGACAATGCTAGGACCTTTTGTCGTATCAGGA CACACATCACGCCCCATGGTGAATTCACAAGTATACCAAGAGAAAATA GAAGGAGATAAAAATGAATATAATACTAGTGTGAGCAAGGGCGA |
| 155.1000_R | CTTTGTTGCCTTTCATTTGTTTCTTTTTTTATTTTAACTCAACCTTTGGAT TGAGTGTGTAATTATCATCTGTTGGGATATTTGTTTTTAAAATCGTATTT TTATCCATTTATTTTTTTAGCCCTCCCACACATAACCAG |
| 167.1010_F | ATAGACAAAAAGCAACGGTATGGTGGGGGGGATATCAGAGGGCAATG GGTGTCCTTCGTTGAGCGCGCTTTCGCCATTGGGAGAGCTCGGAGCA GGAACTTCGACCCACTCACACAACGGACTAGTGTGAGCAAGGGCGA |
| 167.1010_R | CCCAAAAGCCGTAGATTATTGAAAAAGAGTAATCAGATGCCGCTTCCA GATTCAGAAAGCAGCAATGGAAAGAAAGAAATGTTAGAGGAAGAAAAA AGGCAGTACATCTCAAAGCACTTAGCCCTCCCACACATAACCAG |
| 179.1020_F | AAAAAAACACCTCCCCAACGTACCTTTTTTGAGAAGTATCGAAACTCAT TTTTATTTATCGTAGTGTTCTTAGGGTTTAGGTTAGTGAATATGATTGT CAAGAATAGGGTTTCCGGCAACACTAGTGTGAGCAAGGGCGA |
| 179.1020_R | AGGAAGAATAATATTATGCTAAAAAATCACTCATTATTTGATATAGCAAT GTTAAAATTATATTCGATAACAGAGAAATGGACGAGAAACCAACCTAG CCGGTAAAGCCAACCCCAATTAGCCCTCCCACACATAACCAG |
| 201.1030_F | GATGACGGGACGGGAAAGACACAGGTCTACATTGTTCTTTTGTTCTGC GGTTGGTTCTCCATCATCAGTTTTTTTGGTTTGCGGTGGTTGGCAAGC ATGTTCAATCGGGTGGAGCTCGTCACTAGTGTGAGCAAGGGCGA |
| 201.1030_R | ACATGGCCAATAAGTATAAATTTCTGTTCGTGCTGAGATAAAAATATGC ATGCATGAATCTATCTGCGATTTCGGCTTCTCTAGTTCCTTTACCTTCA TTTGAAAGGTCGCTGTCATTTAGCCCTCCCACACATAACCAG |

| | |
|---|---|
| 216.1030_F | AATGAAAACGGTGAGTGGGAGGAGCGGGAGGAGGTTGTTCAAGAGC<br>AGACGTTCTTGCAAAAGTATGGGCTATACATTGCTATCTTTGTGGGGG<br>TCACCATCATGAAGAACTTGTCGTCAACTAGTGTGAGCAAGGGCGA |
| 216.1030_R | GCGAAATGTAGTTAGTACCTACTTTAAAGCCTAATCAACCCCTGCTCA<br>CTATTTAGTTCATTTTATAATATGACGGTATGCATCCCTGCTCTTTCCC<br>CGAAGTCTGCAACTCTGATCTTAGCCCTCCCACACATAACCAG |
| 25.1140_F | TGTTCAATGGCATTAGTTGTGATTGCCTGTTACATGGCTGTAGTCTCAT<br>TGGTAAGCATTGACTACAGTAATGATGCATTGCTAATGGTTGAGGTCC<br>CAGATGACATCTGTCAGAATGAAACTAGTGTGAGCAAGGGCGA |
| 25.1140_R | TTTTATACGATGAGCAATCGTCAAAGAGTATGTTGCGTATCAAAATAGA<br>ATATATTAATAAATAGAAGTTCCTCGTCCCCACTTTTATCATCGGTTAAT<br>ATGAGAACCCCATCTGAGTTAGCCCTCCCACACATAACCAG |
| 25.117_F | TCAAGAATGGGTTCGTTCGCAGCAGAACCAAGGCCCCAGCCATCTGG<br>CAGGGGTATCCTTGCCGGCTCCACAAGGACAATGAGTTTTAATCAGAT<br>GCCACCACCTGCCTATGGTCAGTCGACTAGTGTGAGCAAGGGCGA |
| 25.117_R | GGGCCACTTACTTTTTTTACATTCATGTAAGCATCCCAATACATAGTCT<br>GTAAAACAATTACAGAGTTAGTTAGATATCAGCATTTAGTATGAGTAAA<br>AAGATATAGAAATATCTGCTTAGCCCTCCCACACATAACCAG |
| 270.1010_F | TGCAAGTCTGGGTATCATCTGGACAGGCAAAACGTTTGTCAGAAGAAC<br>GGCGTGGCCACTCGTGTTTCCGGTGCACTGTTTATCGGAATCATATCG<br>GCATTGGCGACAACATTGCTGCTGACTAGTGTGAGCAAGGGCGA |
| 270.1010_R | ACTGATGGACCTATAAAAAATATTTATTCCTCTAATATGAATAGTAGCG<br>CCACTTCTGTGATGGGGGGCTTGGATGAAGGTAGATCCCGGTCGTGT<br>CTGCTTGTGGCTTTAGATAATTTAGCCCTCCCACACATAACCAG |
| 277.1010_F | GACCTGCTGAATGGGGATTATAACAAACCGTTGTTGGTTGTCCTCGTC<br>GTTGGGCTTCTCCTCGCCGCCCTCGTCGCGCGCTACTTCGTCGTACG<br>GAAGAGCCTGGCGGTGGTGTGGGCGACTAGTGTGAGCAAGGGCGA |
| 277.1010_R | TATATATATATGTTTATATGGATATTTGTATATGTATGTGTATGTGTGAG<br>AGTGTACGTTAATGGAAACCTAGATCGCGAAATGCGATCCCCTCCTCC<br>ACCCCCCCCCCATCCTCGATTAGCCCTCCCACACATAACCAG |
| 28.125_F | CAGGACAAGGGACAGGCCGTGGCTTTCATATTCTTTGGCTGCATCGG<br>CCTTGTGTGCTTTGTTGTGGAAATCTTCTTCCTGCACCCATGGGAGGA<br>GAAAGTAGACGAAAACGAAGAAGCTACTAGTGTGAGCAAGGGCGA |
| 28.125_R | ATCTTTCCTTTATTTGTTATATTGCTTAATCAATTTAGATTTATCTATTTC<br>CTGAATCCTGGCGCTCTTCGTTTCACCCGACCAGTTTACAGGCAAGAC<br>GACCGTGTTAGCGGGCAGTTAGCCCTCCCACACATAACCAG |
| 280.1070_F | CTGAACAGGCGTGATGCCGAGTTGTTAGCGAAAAGACAGTTGATTTTG<br>CCTGATCATAAAGATCGTCCGACGGGTGCGGAACCCGACTCTCCTCA<br>TCCTTCGGCCGGAAAAGAAGAGTCTACTAGTGTGAGCAAGGGCGA |
| 280.1070_R | GTTTACAAAACATTCATATATGGATGTACAATAATAAAAACAAAGTATC<br>ACCGATCAAATGCCATTTAATGATCGGTCCAAAAAAATCACTCTCGCT<br>CATAAAACGTATGGCCTCCCTTAGCCCTCCCACACATAACCAG |
| 29.1160_F | CTAAAGGTTCTTACACGGGAACGGTACGATCGTCTCCACGCACGCTTT<br>ATTAAAGAGACAACGGCGTACATCTTTCTTGGATGCATTCCCACNCTA<br>TCATATAAAGCAGTCTATGATTGTACTAGTGTGAGCAAGGGCGA |
| 29.1160_R | TTACGACGATTTTTTGCTGTTAAATTTCATCAGATTTGCCTGTGTAGGC<br>GCAAATACATAATACCCTGCGATCCTTTCAGACATGTTCAGCGGTTTC<br>GCTTCGAACGGCGGAGTATTTTAGCCCTCCCACACATAACCAG |
| 304.1010_F | CAATCCCAAGGTGGAACCAAGTTGCAAATAGCTGCTCTCTTGAGTTCT<br>GTGCGAACTTTATTGCGCTGGCCTCTGATCTTTGCTAATACTTTGCTCA<br>TTGTGTTCGCTATAATATTAGGAACTAGTGTGAGCAAGGGCGA |
| 304.1010_R | CAAATACATGGCACAACATCAAGTCATGAGTAAGTGCATATTCTCAATT<br>TAACGCATGTAGAATAGACACATGTAACATCACGCTTCAGCACAGTAT<br>AAAACGGAATGAAAAGTGGGTTAGCCCTCCCACACATAACCAG |
| 33.1130_F | GTCTGTTTGCTAACACCATCTTGGTTGTTTGTAGGAGATGTTGTGCGC<br>ACCGCGCTACTGCTAACAGGTGCTGCACTATTCCCAATACGCGATCTT<br>GAGCACCAGAAAAACAAGAGAAAAACTAGTGTGAGCAAGGGCGA |
| 33.1130_R | AGTTGGAAGAAAATGATTACCTCGTAAGTCCATGTCTTCAATATCGTAT<br>ATATATATATATATTAGAAACCGGATGTAAAATCTACAACTAATGGTTAT<br>CAGGCTTATTAGTATGAATTAGCCCTCCCACACATAACCAG |

| | |
|---|---|
| 352.1060_F | AAGGCCGACATGGAGAAGCTCGCACAACACTTGCTATCCAAGGAGCT<br>ATTAACCCATCAGGATGTGCTGCACTACCTCGGCGCGCGCCCAGTGC<br>GCTCAAGCGATAAGATAAAAGGAGTTACTAGTGTGAGCAAGGGCGA |
| 352.1060_R | CAAACGCAGGAAAAATGGCAATCGAAACTATCCCTCAACACATCCTGG<br>CCTATAACACCAAAATTAACACGATCAGCAGCAAAAAAAAAAACAATGA<br>TAGTCCTAGAAAGAATGCATTTAGCCCTCCCACACATAACCAG |
| 39.1150_F | CAGAGACGCCTGGCGAGTCAAACTAGTCGACATGCTCGCAGCCATGG<br>CTCCAATCCATGCGTCTTCAGTGAGGTCGATATGCCTAATGCGGGAG<br>AGGTTCATGAAACTAATGTCGAGCTTACTAGTGTGAGCAAGGGCGA |
| 39.1150_R | GAGAAAATTGATGAAGACGACGGACCTCGCAATCGGTCATGCGTCAG<br>TGCATGTTGTCAAATGTATTTGAGGTCGTAAAGCAGAAAGGTCCCGTC<br>AGTCTTCTAGTTAGGGTAGAGCTTAGCCCTCCCACACATAACCAG |
| 40.1010_F | AGCAACCCGACCGATTCCATCATTGAGGAAGCCATTGCCAAGTGTTTT<br>AGTAGAAACAAATGCTCTGCATGTGGGCAGAAGATATCAGAATATTTA<br>TTAACATCCAAGGTTCGATTGATAACTAGTGTGAGCAAGGGCGA |
| 40.1010_R | CACTAAAGGCATATATGCACATATTTATAGAAGCGAGAACTAAGCACG<br>ACGAATAAAAAAGAAACCTTGAAGTTATATCGGGTAAATACAGAACAAT<br>ATTGCTAAGGGGCTATATCTTTAGCCCTCCCACACATAACCAG |
| 40.1240_F | TTATTTAACAAGGGCAATGGTGCACGAACCGTAATGCTAACGTCACCT<br>TCCATGGGGGCTTACGTATTCTTCAGTGGAACCTTTATGTGCGGTCTC<br>CTATTATTTGCAGCATACCATTTGACTAGTGTGAGCAAGGGCGA |
| 40.1240_R | GGTTGATGGCTCATGAAAGTTATTTATTCAGACAGTGACGTTTATATGA<br>GCATTTACCAATTCTTCAGAGAATAGTACGACCTTTTCCGCAATATTAA<br>GGGGCATAATATTAAAAAATTAGCCCTCCCACACATAACCAG |
| 45.1050_F | TCAGGAGACCTCGTCGGCATTATCACAAGCTTTGCCCCTCCAGTGGA<br>GCTTCATAACACACCGCATGGTATTCTGGGTGTTGTTGCATCAGTACC<br>TGGATTTATGTTGAGTTTTATGCATACTAGTGTGAGCAAGGGCGA |
| 45.1050_R | AATTTCTCATCATTTTTATATCTCCACCATAAATCCTCACAAAACAACTT<br>TTATTATACTCTGCTAATCCAGCATAAATATTGCCGAAAATAATGATTTT<br>AAAAATAAACAACTGATTTAGCCCTCCCACACATAACCAG |
| 514.1040_F | CTGATGATCGTCCTGACGCTTGCTCTCCAGATCATCACAGGCATCCAC<br>GACGGGAGGGAAAAAGCCGCTCAGGCAAGGAATCCGCAACCGGCTG<br>TTACCGCAGCCGCGAAGAAGACGAATACTAGTGTGAGCAAGGGCGA |
| 514.1040_R | GGTAAAAAAAAACATAACCTAACACTGATTTCGATATATTACTAATATTT<br>ATTTAATTTTAAACCACGTCCTTAAAAATATTCTCACCATTACTACTTAC<br>TTTGCCCCACACGTGGATTAGCCCTCCCACACATAACCAG |
| 550.1020_F | TACGTCGCGTACCGCCGCCCTGATGAGCTACACCTGCCCGGCATGAC<br>AGCCTTGTCGGAAGATGAGCAGAAACCCTTTGCTGGGAACCTCGAGC<br>CAAACCGCAGGGATAATGAGGGGGTCACTAGTGTGAGCAAGGGCGA |
| 550.1020_R | CATGTATATATATAAAAATAATAAATACATATATATATATGCGATTTAAA<br>GCATTTATTCAAAGAGAAATCGTCAGCTCCTAGCGGTTATGACACTGA<br>TATGCAGAGGAGTACGGTTTTAGCCCTCCCACACATAACCAG |
| 556.1030_F | TCTCGATGTGGCTCTATCAATTCACAAGGAGATGCCTCGAGGGTTCGC<br>CGCATGGCCAGCGACAGTGTCGTCAGTATTCACATCCAGATGCCGTT<br>GTCTTCTCTCGAGAAGACCGGAGGGACTAGTGTGAGCAAGGGCGA |
| 556.1030_R | GATGCCTTCAGCGACAAGCTGCTGCCAACCAAAGGTAACGCAGCGCA<br>TCCATCCCCGATACGCATGCAGCCGCACCGAAGCACACAAACAGGGA<br>GACGAATGACACAGACGTACGCGTTAGCCCTCCCACACATAACCAG |
| 63.1060_F | TCCTATTACGGGACAGACTCGAAGGGCAGAACCCTTTCTAGTGGGAC<br>CAGCAACCCTTTAAACTTCCGTGCATATGGTGGGAACCATATTTTTAAA<br>GACATTGAGCGAGCATTAAAAACTACTAGTGTGAGCAAGGGCGA |
| 63.1060_R | ACCAAATTAATATGTTGTATAAACAGGTGTGCAAAAATACAGCAAGAGT<br>ATTTCTCATCATAAAGTACTATTAAGAAAAAAACGCAACATTAGCAATG<br>ACATCATCCCGCACAACTCTTAGCCCTCCCACACATAACCAG |
| 632.1010_F | CTTTGCAGTGTTGTGGTCCTGTTGATAGCATTAATAATGGCAAATAATA<br>TGTTGTCGCGCCGAGGGAGGCTCTCAGCACTCTTTAAACAAAAAGTA<br>GTTGTGGTACGTGCTCATGAGGACACTAGTGTGAGCAAGGGCGA |
| 632.1010_R | ACAGTTATTGCTGACCATTCATTCCTCGTTTTTACTTTCTTCTTGGTGC<br>TTTCAACGAGGAAATAAATCGGCGCTGCATCAGAAATCAGAACACTTT<br>CCCGTTTCCCCTTTCCCTCCTTAGCCCTCCCACACATAACCAG |

| | |
|---|---|
| 70.1160_F | TTTAAGGATGCTGATTACAGTTTTGGAACATCCTTTATAGTGCAAATTT<br>TCTGCTGTTTGACATACTTACTATGCTTGCTGTTAACTCCTTATACACA<br>GGATTTATGGTGTGGGAAACGGACTAGTGTGAGCAAGGGCGA |
| 70.1160_R | TAAATACTTAGCTATAAGGGATTTACTTTCTTGTCTGCTCATTCAATAAT<br>CAGCGAAGCAATTTAAGATGTACCAAAATGAATGATGGCAATGTTGTT<br>TTCCGGACCTTTCTCCCCGTTAGCCCTCCCACACATAACCAG |
| 71.1110_F | ACGGAGACCGGAACTGTAGATAAGAACTATGACCCGGAATACATCTCA<br>GCAGAGCATCTTCGTTATCGTGACGCGTTGCTGAAGCGTCACAACTCT<br>CCTGCAAGTGCACGCAAGAAGAATACTAGTGTGAGCAAGGGCGA |
| 71.1110_R | AGGAAACACTGCTCCACCTAACTCGAAAGGATCTTAGATATTAATACA<br>CACATTTATCTTGCACAAAAATATACTCCCCTTCCAAACATTACCGCAA<br>CCTTTAGCCTCTTTCATCCTTTAGCCCTCCCACACATAACCAG |
| 73.1140_F | GGTTTCACGGTGGAGCAGGAGGGTCGGTTCAACTACTACACCACCTA<br>CCTGCTGCTGATTTGTGGAGTGATTGGCATCGTCTTTGGCTTGACCGG<br>CACCATCTACAACACCGTGAGCGGTACTAGTGTGAGCAAGGGCGA |
| 73.1140_R | CTTAATGTCTTTTTAATCCATAATTGAAAATATATTCCTTTTGTTAATGAT<br>ATAGATCCAATTATTATACAAAGTAATCTCTCGACACATCTATCTCACA<br>GGCCCCTATTAGCCTGATTAGCCCTCCCACACATAACCAG |
| 794.1010_F | CTGAGGTTTCTTGCCTCCCGTCGGCCGCTTCCTATGGTTTCTACCTTT<br>ATACTGCTAGAGCCACGGGCGGAATTTTTACTCTTGGGAGGTGTGCT<br>GGCATCTTTAAGAAAGCCTGTCGCTACTAGTGTGAGCAAGGGCGA |
| 794.1010_R | CAAGCTTCGAGGTAACATCTCTCATGAAACTCTCTCAGCCACTCCTTC<br>GGTAGGATCTCTCTCACATCGCTCCCTTCTGCAAGGCTCTCTTGGTAA<br>TTCCTCGTCACGGCGTTCACATTAGCCCTCCCACACATAACCAG |
| 8.1300_F | ATCATGTTTTCTCCACTTTTCATATTGATCGGTGGGGTTTATCTGATAT<br>ACAATAAGCATTTTCGAGGGATACCTAATTTGTCGAGGCGAGCTAGGT<br>ATCGTTATTCATCCCATGAAGACACTAGTGTGAGCAAGGGCGA |
| 8.1300_R | CAAACCACTCTAAGAATAATTTTGCTGTCAATGAGTGGAAATAACCGTC<br>CACTTGAACATTTAGCATACCTCCTTGTACACCCCTAGACTCTTTCTAC<br>TTTTACTTGTTTGCCTCTTTTAGCCCTCCCACACATAACCAG |
| 850.1010_F | TCACCCATTCGCAGTGCTGTCGTGACTATATTTGTATGCTTTGCTGAT<br>GCACCTGAGGCTTTGGAGAGATCCGATCCATCACTATACCAAAAGATA<br>CGCGATGCCGATGCGGAGCTCAACACTAGTGTGAGCAAGGGCGA |
| 850.1010_R | TCTTACCGAATAAATAACTAAAAGAGTAATAAAATGCAACCCTTTCTAA<br>AGTTACATACATTTCCGACAGCTATATTAATAAGAATTCCCTTTGTAAA<br>GCACAGCAAGCTACGAATCTTAGCCCTCCCACACATAACCAG |
| 87.1000_F | ATAATTTGTGGGTGGGTGTGGGTGGTGGTGGTGTTGTTGTGCGTGGT<br>GGGTAAGTGGGGTTGGGTTTGGGGTTGGGGTTGGTGTGTTGGTATGA<br>TGGGTTTGTGGTGTGGTAATGGTGGGACTAGTGTGAGCAAGGGCGA |
| 87.1000_R | CCTCAATCACACCCATAATATGGGTACCTACAATATCCACCACACAAC<br>CACCTCACTCCACCTACCCAACACTCGCTACACCACATACCACACCCA<br>CCCTCAACCCACACCCAAACCTTAGCCCTCCCACACATAACCAG |
| 926.1000_F | TCCTATACGGGCCCATTTGGTTACTTCGCGGAAGGCATGGATATAGGA<br>CTTGGGTTAAAATTTTTCATTGCGGGTTGCGCGTTGAACACATTAAATA<br>TTTTAATTCTTTTAATTCCCTTGACTAGTGTGAGCAAGGGCGA |
| 926.1000_R | CTCACAATCACCAAATTTAGAGCTGCTTGGCGTCTAACGCATACAAGC<br>AACGAAACGAGTGAGAAAGAAGTCAACAAAATACATAAATCATGCTCA<br>TTTTTCTCTTAAATATGTTACTTAGCCCTCCCACACATAACCAG |
| 956.1020_F | TTTACATCCCCATGCAATCGAAGGCAACGGTCACATCGTCCCCCCACT<br>ACCTTGTGGTGGTGCTCTACGCCGTTGNGACGCTGGGCATACTGTGC<br>ACGCCAGCGTGCTCGACAGTGCAGCACTAGTGTGAGCAAGGGCGA |
| 956.1020_R | GTCACGACCGCACTGCCAACGGCGTAGAGCACCACCACAAGATAGTG<br>GGGGTATAATGTAACCGCGATTTTTGACCAAATTGGAATGTAAAAGCG<br>CTTCTCATCAGCACTGCCCGCGTTAGCCCTCCCACACATAACCAG |

## *Appendix 3.5 The core set of surface orthogroups conserved in* Trypanosoma*,* Leishmania *and* Phytomonas *species*

| ORTHOGROUP | Annotation | T. carassii | T. brucei | T. congolense | T. vivax | T. cruzi | T. rangeli | T. grayi | L. major | L. mexicana | L. infantum | L_tarentolae | L. braziliensis | P. EM1 | P.HART1 | P. serpens | P. francai | %GPI | %TMD | %SignalP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG000003** | gp63 | 80 | 17 | 8 | 18 | 41 | 30 | 52 | 5 | 7 | 7 | 56 | 32 | 3 | 57 | 17 | 4 | 0.40 | 0.43 | 0.36 |
| **OG000007** | receptor-type adenylate cyclase, putative | 91 | 73 | 86 | 33 | 11 | 10 | 41 | 7 | 8 | 6 | 8 | 6 | 5 | 5 | 32 | 5 | 0.08 | 0.62 | 0.48 |
| **OG000041** | Cathepsin L-like (CP), cysteine peptidase precursor | 3 | 11 | 23 | 14 | 7 | 8 | 2 | 9 | 13 | 3 | 2 | 4 | 1 | 1 | 1 | 1 | 0.01 | 0.79 | 0.78 |
| **OG000092** | membrane-bound acid phosphatase 1 precursor (MBAP1), membrane-bound acid phosphatase 2 (MBAP2), | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 6 | 6 | 6 | 5 | 6 | 3 | 4 | 5 | 3 | 0.02 | 0.76 | 0.68 |
| **OG000099** | endosomal integral membrane protein, putative | 4 | 3 | 3 | 3 | 3 | 4 | 5 | 3 | 3 | 3 | 3 | 3 | 1 | 1 | 1 | 1 | 0.29 | 1.00 | 0.69 |
| **OG000107** | hypothetical protein, conserved | 4 | 5 | 11 | 4 | 5 | 1 | 4 | 2 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 0.02 | 1.00 | 0.76 |
| **OG000148** | phosphate-Repressible Phosphate Permease-like protein | 2 | 2 | 2 | 4 | 2 | 6 | 1 | 3 | 4 | 4 | 3 | 4 | 1 | 1 | 2 | 1 | 0.07 | 1.00 | 0.68 |
| **OG000160** | hypothetical protein, conserved | 4 | 1 | 3 | 1 | 1 | 1 | 4 | 1 | 1 | 1 | 4 | 1 | 1 | 3 | 12 | 1 | 0.08 | 0.90 | 0.42 |
| **OG000176** | vacuolar ATP synthase, putative | 2 | 2 | 3 | 2 | 3 | 2 | 2 | 3 | 3 | 3 | 2 | 3 | 2 | 2 | 2 | 2 | 0.00 | 1.00 | 0.95 |
| **OG000207** | protein disulfide isomerase | 2 | 2 | 2 | 3 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 1 | 1 | 8 | 2 | 0.00 | 0.63 | 0.94 |
| **OG000217** | HSP70-like protein | 5 | 2 | 3 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 3 | 1 | 0.09 | 0.50 | 0.52 |
| **OG000218** | beta-fructofuranosidase-like protein | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 3 | 3 | 2 | 2 | 0 | 7 | 0 | 0 | 0.05 | 0.43 | 0.71 |
| **OG000219** | 3'-nucleotidase/nuclease, putative | 3 | 3 | 3 | 1 | 1 | 1 | 2 | 5 | 4 | 4 | 3 | 3 | 1 | 1 | 1 | 1 | 0.06 | 0.62 | 0.82 |
| **OG000235** | glycosyl hydrolase-like protein | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 0.03 | 0.52 | 0.47 |
| **OG000276** | hypothetical protein, conserved | 3 | 2 | 2 | 2 | 1 | 2 | 1 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0.00 | 0.69 | 0.55 |
| **OG000293** | hypothetical protein, conserved | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0.03 | 1.00 | 0.62 |
| **OG000303** | chaperone protein DNAj, putative | 2 | 2 | 2 | 2 | 3 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0.00 | 0.73 | 0.87 |
| **OG000335** | hypothetical protein, conserved | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 0.00 | 0.60 | 0.66 |
| **OG000371** | hypothetical protein, conserved | 1 | 2 | 3 | 2 | 1 | 0 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 1 | 0.28 | 0.96 | 0.77 |

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG000422** | kinetoplastid-specific phospho-protein phosphatase, putative, | 2 | 1 | 0 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 0 | 0.00 | 0.54 | 0.61 |
| **OG000444** | protein kinase, putative,eukaryotic translation initiation factor 2-alpha kinase precursor, putative | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 0.00 | 0.50 | 0.44 |
| **OG000531** | nucleobase transporter (NT4) | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 3 | 8 | 3 | 0.85 | 1.00 | 0.58 |
| **OG000551** | hypothetical protein, conserved | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 0 | 1 | 1 | 0.00 | 0.91 | 0.77 |
| **OG000603** | hypothetical protein, conserved | 1 | 2 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 0.00 | 0.96 | 0.38 |
| **OG000621** | peptidyl-prolyl cis-trans isomerase, macrophage infectivity potentiator precursor, putative | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 0.00 | 0.50 | 0.57 |
| **OG000630** | hypothetical protein, conserved | 1 | 2 | 3 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 0.00 | 1.00 | 0.79 |
| **OG000652** | Selenoprotein T, putative (SelT) | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 0.00 | 0.50 | 0.55 |
| **OG000657** | subtilisin-like serine peptidase,serine peptidase, clan SB, family S8-like protein | 4 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 0.10 | 0.79 | 0.65 |
| **OG000686** | Golgi/lysosome glycoprotein, putative | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 0.00 | 0.94 | 0.61 |
| **OG000710** | hypothetical protein, conserved | 2 | 2 | 2 | 3 | 1 | 0 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0.00 | 0.95 | 0.38 |
| **OG000719** | vacuolar-type proton translocating pyrophosphatase 1, putative | 1 | 2 | 5 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 0.09 | 1.00 | 0.96 |
| **OG000787** | hypothetical protein, conserved | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.68 | 0.37 |
| **OG000821** | hypothetical protein, conserved | 1 | 6 | 4 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.65 |
| **OG000852** | hypothetical protein, conserved | 0 | 2 | 2 | 1 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 0 | 3 | 2 | 0.00 | 1.00 | 0.83 |
| **OG000854** | hypothetical protein, unknown function | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 0.00 | 0.94 | 0.82 |
| **OG001038** | hypothetical protein, conserved | 1 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 0.00 | 0.79 | 0.89 |
| **OG001099** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 2 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 0 | 0.29 | 0.82 | 0.81 |
| **OG001140** | hypothetical protein, conserved | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.94 | 0.53 |
| **OG001166** | hypothetical predicted transmembrane protein | 2 | 1 | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 5 | 1 | 0.23 | 0.77 | 0.59 |
| **OG001167** | flagellar glycoprotein-like protein | 3 | 3 | 3 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 0.68 | 0.82 | 0.80 |
| **OG001206** | lectin, putative | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0.00 | 0.56 | 0.44 |
| **OG001230** | dolichyl-lipid chaperone, putative (rft1) | 3 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.38 |
| **OG001249** | hypothetical protein, conserved | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 0.00 | 0.89 | 0.78 |
| **OG001259** | ABC transporter, putative, | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0.94 | 0.94 | 0.88 |
| **OG001296** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 0.70 | 0.35 | 0.55 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG001308** | endonuclease G, putative | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.50 | 0.37 |
| **OG001376** | hypothetical protein, conserved | 3 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 0 | 2 | 1 | 0.00 | 0.82 | 0.50 |
| **OG001381** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 0.11 | 1.00 | 0.84 |
| **OG001463** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 0.00 | 0.94 | 0.44 |
| **OG001568** | FG-GAP repeat protein, putative,intergrin alpha chain protein, putative | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.72 | 0.71 |
| **OG001608** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.07 | 1.00 | 0.64 |
| **OG001676** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.50 | 0.81 |
| **OG001701** | phosphatidylglycerolphosphate synthase, mitochondrial, putative (pgps) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0.12 | 0.94 | 0.82 |
| **OG001740** | endoplasmic reticulum oxidoreductin, putative | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.47 | 0.69 |
| **OG001935** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 3 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.74 | 0.39 |
| **OG001940** | hypothetical protein, conserved | 4 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.40 | 0.81 |
| **OG001975** | hypothetical protein, conserved | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 1.00 |
| **OG001999** | hypothetical protein, conserved | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.09 | 1.00 | 0.55 |
| **OG002030** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 1.00 |
| **OG002031** | heat shock protein 90, putative,lipophosphoglycan biosynthetic protein, putative,glucose regulated protein 94, putative (LPG3) | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.47 | 0.82 |
| **OG002096** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 3 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.94 | 0.88 |
| **OG002105** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.00 | 1.00 | 0.47 |
| **OG002163** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.94 | 0.67 |
| **OG002194** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.94 | 0.88 |
| **OG002275** | cop-coated vesicle membrane protein p24 precursor, putative,ER--golgi transport protein p24, putative | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.31 | 1.00 | 0.75 |
| **OG002277** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0.00 | 1.00 | 0.43 |
| **OG002314** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.94 | 1.00 |
| **OG002557** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.07 | 0.81 | 0.81 |
| **OG002575** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 0.00 | 0.76 | 0.59 |
| **OG002612** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0.00 | 0.94 | 0.41 |
| **OG002624** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 0 | 0.00 | 1.00 | 0.88 |

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG002632** | hypothetical protein, conserved | 2 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 0 | 0.00 | 0.76 | 0.56 |
| **OG002642** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.94 | 0.65 |
| **OG002697** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 0 | 0.06 | 0.67 | 0.89 |
| **OG002725** | hypothetical protein, conserved | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.53 | 0.89 |
| **OG002747** | COP-coated vesicle membrane protein gp25L precursor, putative,ER--golgi transport protein gp25L, putative | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 1.00 |
| **OG002872** | transmembrane protein, putative | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.69 | 0.63 | 0.88 |
| **OG002946** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.00 | 0.93 | 0.47 |
| **OG002950** | cyclophilin | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 2 | 0.00 | 0.38 | 0.88 |
| **OG003041** | hypothetical protein, conserved | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.00 | 0.81 | 0.87 |
| **OG003062** | gpi transamidase component, putative | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.88 | 0.88 |
| **OG003069** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.93 | 1.00 |
| **OG003083** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.67 | 0.40 |
| **OG003115** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.63 | 0.88 |
| **OG003203** | short chain dehydrogenase, putative | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.81 | 0.53 |
| **OG003247** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.93 |
| **OG003262** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.88 | 0.88 |
| **OG003356** | GPI anchor biosynthesis protein, putative | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.00 | 0.94 | 0.47 |
| **OG003366** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.81 | 0.50 |
| **OG003399** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.60 |
| **OG003457** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.88 | 0.44 |
| **OG003496** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 0.00 | 0.76 | 0.44 |
| **OG003500** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.00 | 0.88 | 0.50 |
| **OG003520** | glycosyltransferase family 28 protein, putative | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.80 |
| **OG003558** | GPI-anchor transamidase subunit 8 (GPI8), putative,cysteine peptidase, Clan CD, family C13, putative | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.29 | 0.64 | 1.00 |
| **OG003746** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.50 | 0.75 |
| **OG003774** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.81 | 0.69 |
| **OG003804** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.53 | 0.82 |

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG003811** | GPI transamidase component GAA1, putative | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 0.06 | 1.00 | 0.41 |
| **OG003814** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.88 | 1.00 |
| **OG003922** | hypothetical protein, conserved | 2 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.29 | 0.60 | 0.46 |
| **OG004004** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.60 | 0.80 |
| **OG004026** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.00 | 0.87 | 0.87 |
| **OG004033** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.60 | 0.60 |
| **OG004037** | lysyl-tRNA synthetase, putative | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.40 | 0.00 | 0.47 |
| **OG004055** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.38 | 0.87 |
| **OG004057** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.81 | 0.38 |
| **OG004129** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.07 | 0.64 | 0.46 |
| **OG004164** | hypothetical protein, conserved | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.60 |
| **OG004362** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.13 | 0.93 | 0.71 |
| **OG004422** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.00 | 1.00 | 0.73 |
| **OG004477** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.07 | 1.00 | 0.80 |
| **OG004492** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.73 |
| **OG004509** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.53 | 0.00 | 0.43 |
| **OG004511** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.13 | 0.94 | 0.40 |
| **OG004514** | carbonic anhydrase-like protein | 1 | 1 | 2 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.94 |
| **OG004552** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.00 | 0.67 | 0.36 |
| **OG004556** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.07 | 1.00 | 0.40 |
| **OG004564** | DNAJ domain protein, putative | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.13 | 0.81 | 0.88 |
| **OG004589** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 1.00 |
| **OG004621** | COP-coated vesicle membrane protein erv25 precursor, putative,ER--golgi transport protein erv25 precursor, putative | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.14 | 1.00 | 1.00 |
| **OG004628** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.00 | 0.57 | 0.50 |
| **OG004631** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0.00 | 0.93 | 0.53 |
| **OG004662** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.94 | 0.81 |
| **OG004663** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0.44 | 0.63 | 0.47 |

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG004738** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.63 | 0.80 |
| **OG004772** | NADH-cytochrome b5 reductase, putative | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.44 |
| **OG004827** | vesicular-fusion ATPase-like protein, putative | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.40 | 0.73 |
| **OG004832** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.79 | 0.69 |
| **OG004844** | N-acetylglucosaminyl transferase component, putative | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.07 | 1.00 | 0.38 |
| **OG004856** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.27 | 0.87 | 0.73 |
| **OG004876** | extracellular receptor, putative | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 0.00 | 1.00 | 0.69 |
| **OG004946** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.73 | 0.36 |
| **OG004975** | N-acetylglucosaminyl-phosphatidylinositol deacetylase | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.00 | 0.80 | 0.93 |
| **OG004990** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.13 | 0.40 | 0.57 |
| **OG005003** | rhomboid-like protein,serine peptidase, Clan S-, family S54, putative | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.07 | 0.87 | 0.80 |
| **OG005016** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.36 |
| **OG005017** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.07 | 0.79 | 0.93 |
| **OG005107** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0.00 | 0.36 | 0.38 |
| **OG005117** | squalene monooxygenase-like protein | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.60 | 0.67 |
| **OG005227** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.80 | 1.00 | 0.79 |
| **OG005230** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.93 |
| **OG005368** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.00 | 1.00 | 0.64 |
| **OG005372** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0.00 | 0.79 | 0.79 |
| **OG005446** | phosphate transporter, putative | 2 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.00 | 1.00 | 0.62 |
| **OG005468** | hypothetical protein, unknown function | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.00 | 1.00 | 0.42 |
| **OG005499** | hypothetical protein, conserved | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.00 | 1.00 | 0.54 |
| **OG005527** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0.15 | 1.00 | 0.64 |
| **OG005542** | hypothetical protein, conserved | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0.00 | 0.64 | 0.86 |
| **OG005545** | hypothetical protein, conserved | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.77 | 0.46 |
| **OG005557** | hypothetical protein, conserved | 0 | 1 | 1 | 1 | 2 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 0.00 | 0.56 | 0.50 |
| **OG005567** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.00 | 0.57 | 0.86 |
| **OG005577** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.50 |

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG005645** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.44 | 0.69 |
| **OG005670** | hypothetical protein, conserved | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0.38 | 1.00 | 0.85 |
| **OG005690** | hypothetical protein, conserved | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.91 |
| **OG005848** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.00 | 1.00 | 0.62 |
| **OG005862** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.07 | 1.00 | 0.86 |
| **OG005866** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1.00 | 1.00 | 1.00 |
| **OG005888** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 2 | 1 | 0.38 | 0.00 | 0.54 |
| **OG005897** | C-8 sterol isomerase-like protein | 0 | 1 | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.00 | 0.71 | 0.93 |
| **OG005948** | 3-oxo-5-alpha-steroid 4-dehydrogenase-like protein | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.00 | 0.92 | 0.42 |
| **OG006114** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0.00 | 0.36 | 0.92 |
| **OG006149** | hypothetical protein, conserved | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0.36 | 0.91 | 0.36 |
| **OG006195** | hypothetical protein, conserved | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0.00 | 1.00 | 1.00 |
| **OG006251** | hypothetical protein, unknown function | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.10 | 0.91 | 0.60 |
| **OG006413** | A2 protein | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0.25 | 0.40 | 0.75 |
| **OG006594** | hypothetical protein, unknown function | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0.70 | 1.00 | 0.80 |
| **OG006795** | hypothetical protein, conserved | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0.00 | 0.78 | 0.44 |
| **OG007150** | hypothetical protein, unknown function | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0.00 | 1.00 | 0.75 |

## *Appendix 5.1 The Pan-Trypanosome core repertoire of surface gene families*

| SURFACE ORTHOGROUP | Annotation | T. carassii | T. brucei | T. congolense | T. vivax | T. cruzi | T. ranglei | T. grayi | %GPI | %TMD | %SP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG000007** | receptor-type adenylate cyclase GRESAG 4 | 91 | 73 | 86 | 33 | 11 | 10 | 41 | 0.08 | 0.63 | 0.58 |
| **OG000003** | Gp63-1 surface protease homolog | 80 | 17 | 8 | 18 | 41 | 30 | 52 | 0.48 | 0.40 | 0.47 |
| **OG000048** | hypothetical protein,chrIX additional | 25 | 20 | 20 | 13 | 9 | 5 | 2 | 0.10 | 0.77 | 0.80 |
| **OG000041** | cysteine peptidase precursor,cysteine peptidase, Clan CA, family C1, Cathepsin L-like (CP), | 3 | 11 | 23 | 14 | 7 | 8 | 2 | 0.00 | 0.74 | 0.73 |
| **OG000107** | hypothetical protein, conserved | 4 | 5 | 11 | 4 | 5 | 1 | 4 | 0.03 | 1.00 | 0.81 |
| **OG000163** | hypothetical protein, conserved | 1 | 3 | 1 | 2 | 8 | 3 | 6 | 0.83 | 1.00 | 0.75 |
| **OG000099** | endosomal integral membrane protein | 4 | 3 | 3 | 3 | 3 | 4 | 5 | 0.27 | 1.00 | 0.78 |
| **OG000765** | hypothetical protein, conserved | 4 | 4 | 4 | 4 | 2 | 1 | 4 | 0.09 | 0.96 | 0.95 |
| **OG000649** | cysteine peptidase C (CPC) | 1 | 1 | 12 | 2 | 1 | 1 | 1 | 0.00 | 0.74 | 1.00 |
| **OG000217** | heat shock protein, putative,HSP70-like protein | 5 | 2 | 3 | 2 | 2 | 3 | 2 | 0.00 | 0.42 | 0.50 |
| **OG000176** | vacuolar type h  ATPase subunit | 2 | 2 | 3 | 2 | 3 | 2 | 2 | 0.00 | 1.00 | 0.93 |
| **OG000081** | hypothetical protein, conserved | 5 | 1 | 2 | 1 | 2 | 2 | 3 | 0.00 | 0.75 | 0.50 |
| **OG000160** | hypothetical protein, conserved | 4 | 1 | 3 | 1 | 1 | 1 | 4 | 0.08 | 0.87 | 0.85 |
| **OG000207** | protein disulfide isomerase | 2 | 2 | 2 | 3 | 1 | 2 | 2 | 0.00 | 0.43 | 0.93 |
| **OG000335** | hypothetical protein, conserved | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0.00 | 0.57 | 0.85 |
| **OG006306** | membrane-bound acid phosphatase | 1 | 2 | 2 | 2 | 1 | 3 | 2 | 0.15 | 1.00 | 0.85 |
| **OG000293** | hypothetical protein, conserved | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 0.08 | 1.00 | 0.73 |
| **OG000276** | hypothetical protein, conserved | 3 | 2 | 2 | 2 | 1 | 2 | 1 | 0.00 | 0.77 | 0.58 |
| **OG000219** | single strand-specific nuclease | 3 | 3 | 3 | 1 | 1 | 1 | 2 | 0.17 | 0.57 | 0.91 |
| **OG000303** | chaperone protein DNAj | 2 | 2 | 2 | 2 | 3 | 1 | 1 | 0.00 | 0.50 | 0.92 |
| **OG000630** | hypothetical protein, conserved | 1 | 2 | 3 | 2 | 2 | 1 | 1 | 0.00 | 1.00 | 0.91 |
| **OG001308** | endonuclease G | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 0.00 | 0.55 | 0.60 |
| **OG001673** | protein kinase | 2 | 1 | 2 | 1 | 2 | 1 | 1 | 0.11 | 0.70 | 0.44 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG000963** | protein kinase | 3 | 1 | 2 | 2 | 1 | 1 | 1 | 0.00 | 0.55 | 0.67 |
| **OG002725** | hypothetical protein, conserved | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 0.00 | 0.60 | 0.89 |
| **OG000551** | hypothetical protein, conserved | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 0.00 | 0.80 | 0.67 |
| **OG000092** | membrane-bound acid phosphatase 1 precursor (MBAP1) | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 0.00 | 0.90 | 0.67 |
| **OG004852** | calreticulin | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 0.00 | 0.90 | 1.00 |
| **OG000787** | hypothetical protein, conserved | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 0.00 | 0.80 | 0.60 |
| **OG002163** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 0.00 | 0.89 | 0.44 |
| **OG001038** | hypothetical protein, conserved | 1 | 1 | 3 | 1 | 1 | 1 | 1 | 0.00 | 0.67 | 1.00 |
| **OG001140** | hypothetical protein, conservedhypothetical protein, | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 0.00 | 0.89 | 0.63 |
| **OG001296** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0.88 | 0.13 | 0.75 |
| **OG006720** | hypothetical protein, conserved | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 0.63 | 0.88 | 1.00 |
| **OG004853** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 0.63 | 0.75 | 0.75 |
| **OG006881** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0.38 | 0.88 | 0.88 |
| **OG003106** | endosomal integral membrane protein | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 0.38 | 1.00 | 1.00 |
| **OG001612** | hypothetical protein, conserved | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 0.25 | 0.56 | 0.43 |
| **OG001381** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0.13 | 1.00 | 0.75 |
| **OG001335** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0.13 | 0.88 | 0.63 |
| **OG000657** | subtilisin-like serine peptidase,serine peptidase, clan SB, family S8-like protein | 4 | 1 | 3 | 1 | 1 | 1 | 1 | 0.00 | 0.58 | 0.63 |
| **OG005372** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0.00 | 0.75 | 0.88 |
| **OG002314** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 1.00 |
| **OG002624** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 0.00 | 1.00 | 0.75 |
| **OG002030** | disulfide isomerase | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 1.00 |
| **OG002031** | lipophosphoglycan biosynthetic protein, putative (LPG3) | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0.00 | 0.50 | 0.63 |
| **OG003041** | hypothetical protein, conserved | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.63 | 1.00 |
| **OG001740** | endoplasmic reticulum oxidoreductin, putative, pol-associated gene 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.75 | 0.86 |
| **OG002642** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0.00 | 0.88 | 0.50 |
| **OG005593** | hypothetical protein, conserved | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.88 | 0.86 |
| **OG002277** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 0.00 | 1.00 | 0.63 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG001249** | hypothetical protein, conserved | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.88 | 0.86 |
| **OG006079** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 1.00 |
| **OG005866** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.00 | 1.00 | 1.00 |
| **OG003507** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.86 | 1.00 | 0.43 |
| **OG005073** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.83 | 1.00 | 0.57 |
| **OG002872** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.43 | 0.43 | 0.86 |
| **OG005527** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.33 | 1.00 | 0.57 |
| **OG004564** | chaperone protein DNAj | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.29 | 0.86 | 1.00 |
| **OG003811** | GPI transamidase component GAA1 (TbGAA1) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.17 | 1.00 | 0.57 |
| **OG003232** | GPI transamidase component Tta2 (TTA2) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.14 | 1.00 | 0.43 |
| **OG003673** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.14 | 1.00 | 0.43 |
| **OG001701** | phosphatidylglycerophosphate synthase | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.14 | 1.00 | 0.71 |
| **OG003814** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0.00 | 0.75 | 1.00 |
| **OG005543** | kinetoplastid-specific phospho-protein phosphatase | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.88 | 1.00 |
| **OG006067** | alpha-1,2-mannosyltransferase | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.86 | 1.00 |
| **OG007510** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.86 | 1.00 |
| **OG006523** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.43 |
| **OG004738** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.86 | 0.83 |
| **OG006865** | GPI inositol deacylase precursor (GPIdeAc) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.43 | 0.83 |
| **OG004921** | FG-GAP repeat protein | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.71 |
| **OG002275** | cytosolic coat protein | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 1.00 |
| **OG003115** | TPR-repeat-containing chaperone protein DNAJ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.71 | 0.86 |
| **OG004026** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.86 |
| **OG002747** | cytosolic coat protein | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 1.00 |
| **OG005745** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.71 |
| **OG003730** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.57 |
| **OG006467** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.86 |
| **OG005764** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 1.00 |
| **OG002557** | thioredoxin | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.71 | 0.71 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG006512** | calcium-dependent lipid binding protein, putative,synaptotagmin, putative (CaLB) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.86 | 0.43 |
| **OG003804** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.71 | 0.71 |
| **OG003062** | GPI transamidase, putative (TbGPI16) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.71 | 0.71 |
| **OG004772** | NADH-cytochrome b5 reductase, putative (B5R) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.71 |
| **OG005645** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.57 | 0.86 |
| **OG002950** | cyclophilin-type peptidyl-prolyl cis-trans isomerase, putative (PPIase) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.43 | 0.86 |
| **OG004631** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.43 |
| **OG001676** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.43 | 0.71 |
| **OG003457** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.43 |
| **OG004662** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.71 |
| **OG004975** | N-acetyl-D-glucosaminylphosphatidylinositol de-N-acetylase (GPI12) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.71 | 1.00 |
| **OG005659** | chaperone protein DNAj | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.43 |
| **OG002194** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.71 |
| **OG005952** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.71 |
| **OG002575** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.86 | 0.71 |
| **OG003262** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.71 | 0.71 |
| **OG007120** | dolichyl-P-Man:GDP-Man7GlcNAc2-PP-dolichyl alpha-1,6-mannosyltransferase, putative (ALG12) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.86 |
| **OG003366** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.57 | 0.43 |
| **OG004589** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 1.00 |
| **OG005414** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.71 | 0.57 |
| **OG004304** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.43 |
| **OG006114** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.57 | 0.83 |
| **OG002105** | acyltransferase | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.71 |
| **OG000444** | protein kinase, putative,eukaryotic translation initiation factor 2-alpha kinase precursor | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.71 | 0.86 |
| **OG002536** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.43 |
| **OG003774** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.86 | 0.86 |
| **OG003928** | short-chain dehydrogenase | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.71 | 0.43 |
| **OG006104** | transporter | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1.00 | 1.00 | 0.83 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG004037** | lysyl-tRNA synthetase | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.67 | 0.00 | 0.67 |
| **OG006355** | rhomboid-like protein,serine peptidase, Clan S-, family S54 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.67 |
| **OG003496** | short-chain dehydrogenase | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.71 | 0.83 |
| **OG003203** | 3-keto-dihydrosphingosine reductase | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.86 | 0.83 |
| **OG005827** | dolichyl-P-Man:GDP-Man5GlcNAc2-PP-dolichyl alpha-1,2-mannosyltransferase | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.86 | 0.67 |
| **OG000821** | hypothetical protein, conserved | 1 | 6 | 4 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.43 |
| **OG000719** | vacuolar-type proton translocating pyrophosphatase 1, putative (PPase1) | 1 | 2 | 5 | 1 | 1 | 0 | 1 | 0.18 | 1.00 | 0.91 |
| **OG000431** | hypothetical protein, conserved (pseudogene),hypothetical protein, conserved truncated | 1 | 3 | 4 | 1 | 0 | 2 | 1 | 0.00 | 0.82 | 0.73 |
| **OG001167** | flagellum-adhesion glycoprotein | 3 | 3 | 3 | 1 | 1 | 0 | 1 | 0.44 | 0.75 | 1.00 |
| **OG000371** | hypothetical protein, conserved | 1 | 2 | 3 | 2 | 1 | 0 | 1 | 0.22 | 1.00 | 0.90 |
| **OG000397** | serine carboxypeptidase III precursor, putative,serine peptidase, Clan SC, Family S10 (CBP1) | 1 | 3 | 2 | 1 | 2 | 0 | 1 | 0.00 | 0.40 | 0.67 |
| **OG002755** | UDP-glucose:glycoprotein glucosyltransferase | 2 | 1 | 3 | 1 | 0 | 1 | 1 | 0.00 | 0.56 | 0.63 |
| **OG001166** | hypothetical protein, conserved | 2 | 1 | 2 | 1 | 0 | 1 | 1 | 0.13 | 0.75 | 0.63 |
| **OG002096** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 3 | 0 | 1 | 0.00 | 0.88 | 1.00 |
| **OG001259** | ABC transporter | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 1.00 | 0.86 | 0.86 |
| **OG006206** | hypothetical protein, conserved | 1 | 1 | 2 | 0 | 1 | 1 | 1 | 0.43 | 0.86 | 0.71 |
| **OG001099** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 2 | 1 | 0.43 | 0.71 | 0.83 |
| **OG002805** | subtilisin-like serine peptidase,serine peptidase, clan SB, family S8-like protein | 3 | 1 | 0 | 1 | 1 | 1 | 1 | 0.00 | 0.75 | 0.67 |
| **OG005530** | quiescin sulfhydryl oxidase (QSOX) | 1 | 1 | 2 | 1 | 0 | 2 | 1 | 0.00 | 0.88 | 0.71 |
| **OG002632** | hypothetical protein, conserved | 2 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 0.71 | 0.50 |
| **OG004422** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.86 |
| **OG004362** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 0 | 1 | 1 | 0.00 | 0.86 | 0.50 |
| **OG005861** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.86 |
| **OG005176** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 0 | 1 | 1 | 0.00 | 0.57 | 0.50 |
| **OG001463** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.86 |
| **OG004514** | carbonic anhydrase-like protein | 1 | 1 | 2 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 1.00 |
| **OG007455** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 0 | 1 | 1 | 0.00 | 1.00 | 0.83 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG005028** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.43 |
| **OG005227** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.83 | 1.00 | 0.80 |
| **OG007320** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0.83 | 1.00 | 0.50 |
| **OG005296** | hypothetical protein, conserved | 1 | 1 | 2 | 0 | 1 | 1 | 1 | 0.67 | 1.00 | 0.50 |
| **OG006155** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.67 | 0.00 | 0.50 |
| **OG006691** | MSP-A | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.50 | 0.50 | 0.83 |
| **OG006445** | hypothetical protein, conserved | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0.33 | 0.67 | 0.60 |
| **OG005003** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0.17 | 1.00 | 0.67 |
| **OG000525** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.17 | 0.83 | 0.67 |
| **OG007440** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.00 | 0.83 | 0.83 |
| **OG005755** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.83 |
| **OG005542** | hypothetical protein, conserved | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0.00 | 0.50 | 1.00 |
| **OG006243** | hypothetical protein, conserved | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0.00 | 1.00 | 0.50 |
| **OG004492** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.00 | 1.00 | 0.83 |
| **OG001206** | lectin, putative,mannose-specific lectin | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 0.67 | 0.67 |
| **OG003069** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.00 | 1.00 | 1.00 |
| **OG004990** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 0.67 | 1.00 |
| **OG005368** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.83 |
| **OG003247** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.00 | 1.00 | 0.83 |
| **OG007288** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.80 |
| **OG005230** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.83 |
| **OG007097** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.67 |
| **OG006437** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.67 |
| **OG006718** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.80 |
| **OG003947** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 0.67 | 0.50 |
| **OG005848** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.80 |
| **OG005017** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.83 |
| **OG003686** | chaperone protein DNAj | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 0.83 | 0.50 |
| **OG004628** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0.00 | 0.67 | 0.83 |
| **OG006156** | hypothetical protein, conserved | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0.00 | 1.00 | 0.80 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG006882** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 0.83 | 0.67 |
| **OG003399** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.83 |
| **OG004946** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 0.83 | 0.40 |
| **OG006271** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.00 | 1.00 | 0.40 |
| **OG003520** | glycosyltransferase family 28 protein | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.00 | 1.00 | 0.83 |
| **OG005567** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.00 | 0.67 | 1.00 |
| **OG004556** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.00 | 1.00 | 0.67 |
| **OG005016** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.80 |
| **OG004033** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 0.67 | 0.83 |
| **OG004904** | hypothetical protein, conserved | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0.00 | 1.00 | 0.50 |
| **OG004477** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.83 |
| **OG003298** | mitochondrial carrier protein (MCP3) | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.40 | 0.50 | 0.40 |
| **OG001931** | hypothetical protein, conserved | 3 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 0.75 | 0.40 |
| **OG005031** | hypothetical protein, conserved | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.40 |
| **OG004938** | CDP-diacylglycerol synthetase | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.40 |
| **OG000621** | FKBP-type peptidyl-prolyl cis-trans isomerase, conservedFKBP-type peptidylprolyl cis-trans isomerase | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.00 | 0.67 | 1.00 |
| **OG000220** | *T. cruzi* surface protease GP63 (pseudogene) putative | 2 | 0 | 1 | 0 | 36 | 1 | 3 | 0.03 | 0.82 | 0.82 |
| **OG000537** | lysosomal/endosomal membrane protein p67 (p67) | 1 | 2 | 6 | 1 | 1 | 0 | 2 | 0.00 | 1.00 | 0.67 |
| **OG000462** | hypothetical protein, conserved | 1 | 3 | 0 | 0 | 2 | 2 | 2 | 0.80 | 1.00 | 0.60 |
| **OG005865** | *T. cruzi* hypothetical protein conserved | 1 | 0 | 0 | 1 | 5 | 1 | 1 | 0.00 | 0.89 | 0.50 |
| **OG006457** | hypothetical protein, conserved | 1 | 1 | 1 | 0 | 2 | 0 | 2 | 0.14 | 1.00 | 0.86 |
| **OG004663** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 0 | 0 | 1 | 0.67 | 0.50 | 1.00 |
| **OG006433** | hypothetical protein, conserved | 1 | 1 | 2 | 0 | 1 | 0 | 1 | 0.00 | 0.50 | 1.00 |
| **OG004856** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 0 | 0 | 1 | 0.00 | 0.83 | 0.83 |
| **OG007309** | SNARE domain-containing protein | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1.00 | 1.00 | 1.00 |
| **OG006226** | hypothetical protein, conserved | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 0.80 | 1.00 | 1.00 |
| **OG003922** | hypothetical protein, conserved | 2 | 1 | 0 | 1 | 1 | 0 | 1 | 0.80 | 0.83 | 0.75 |
| **OG006297** | hypothetical protein, conserved | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0.80 | 0.20 | 0.40 |
| **OG005862** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0.20 | 1.00 | 0.60 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **OG006036** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0.00 | 0.80 | 0.50 |
| **OG005336** | dolichyl-P-Man:GDP-Man5GlcNAc2-PP-dolichyl alpha-1,3-mannosyltransferase, putative (ALG3) | 2 | 1 | 1 | 1 | 0 | 0 | 1 | 0.00 | 1.00 | 0.60 |
| **OG000686** | Golgi/lysosome glycoprotein 1 (tGLP1) | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0.00 | 1.00 | 0.80 |
| **OG004621** | COP-coated vesicle membrane protein erv25 precursor, putative,ER--golgi transport protein erv25 precursor | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0.00 | 1.00 | 1.00 |
| **OG001608** | UDP-galactose transporter | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0.00 | 1.00 | 0.80 |
| **OG005894** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 1.00 |
| **OG004832** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0.00 | 1.00 | 0.75 |
| **OG005100** | *T. cruzi* hypothetical protein conserved | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0.00 | 1.00 | 0.40 |
| **OG005888** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0.75 | 0.00 | 0.50 |
| **OG008378** | hypothetical protein, conserved | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0.25 | 0.80 | 0.75 |
| **OG001376** | hypothetical protein, conserved | 3 | 1 | 1 | 0 | 1 | 0 | 1 | 0.00 | 0.71 | 0.75 |
| **OG008372** | hypothetical protein, conserved | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0.00 | 1.00 | 0.75 |
| **OG000144** | *T. cruzi* surface protease GP63 putative | 1 | 0 | 0 | 1 | 47 | 4 | 0 | 0.46 | 0.63 | 0.54 |
| **OG006115** | hypothetical protein, conserved | 1 | 7 | 5 | 0 | 0 | 0 | 1 | 0.00 | 0.43 | 0.85 |
| **OG007353** | *T. cruzi* hypothetical protein | 1 | 0 | 0 | 1 | 1 | 0 | 2 | 0.00 | 0.80 | 0.80 |
| **OG008103** | *T. cruzi* hypothetical protein | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0.25 | 1.00 | 1.00 |
| **OG006884** | hypothetical protein | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0.00 | 0.75 | 0.75 |
| **OG006776** | | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0.00 | 1.00 | 0.50 |

# *References*

**Abreu Filho, B. A., Dias Filho, B. P., Vermelho, A. B., Jankevicius, S. I., Jankevicius, J. V., and dos Santos, R. L.** (2001). Surface component characterization as taxonomic tools for *Phytomonas* spp identification. Parasitology Research, 87(2), 138–144.

**Acosta-Serrano, A.** (2001). The mucin-like glycoprotein super-family of *Trypanosoma cruzi*: structure and biological roles. Molecular and Biochemical Parasitology, 114(2), 143–150.

**Ahmed, M.** (1994). Trypanosomiasis in common carp (*Cyprinus carpio* L.). PhD Thesis. Leuven: Catholic University.

**Akiyoshi, B., and Gull, K.** (2014). Discovery of unconventional kinetochores in kinetoplastids. Cell, 156(6), 1247–58.

**Allen, C. L., Goulding, D., and Field, M. C.** (2003). Clathrin-mediated endocytosis is essential in *Trypanosoma brucei*. The EMBO Journal, 22(19), 4991–5002.

**Almeida, M. L., Cavazzana, M., Santos, M. A., Jankevicius, S. I., and Jankevicius, J. V.** (1990). Isolation, axenic cultivation and ultrastructural characterization of *Phytomonas* sp. in urucum (*Bixa orellana*). Memorias Do Instituto Oswaldo Cruz, 85, 135.

**Alves e Silva, T. L. A., Vasconcellos, L. R. C., Lopes, A. H., and Souto-Padrón, T.** (2013). The Immune Response of Hemocytes of the Insect *Oncopeltus fasciatus* against the Flagellate *Phytomonas serpens*. PloS One, 8(8), e72076.

**Aragão, H. de B.** (1927). Sur un flagellé du latex de *Manihot palmata*, *Phytomonas françai* nsp. CR Soc. Biol, 97, 1077–1080.

**Aragão, H. de B.** (1931). Untersuchungen über *Phytomonas françai*. Memórias Do Instituto Oswaldo Cruz, 25(4), 299–306.

**Asai, T., Tena, G., Plotnikova, J., Willmann, M. R., Chiu, W.-L., Gomez-Gomez, L., Boller, T., Ausubel, F. M., and Sheen, J.** (2002). MAP kinase signalling cascade in Arabidopsis innate immunity. Nature, 415(6875), 977–83.

**Attias, M., and De Souza, W.** (1986). Axenic Cultivation and Ultrastructural Study of a *Phytomonas* sp. Isolated from the Milkweed Plant *Euphorbia hyssopifolia*. The Journal of Protozoology, 33(1), 84–87.

**Attias, M., Resende, M. L. V., Bezerra, J. L., and De Souza, W.** (1987). Roystonea regia a new host palm for *Phytomonas staheli*. Memórias Do Instituto Oswaldo Cruz, 82, 35.

**Ayala, S., Quintero, O., and Barreto, P.** (1975). Tripanosomkidos de plantas laticiferas y sus insetos transmisores en Colombia y Costa Rica. Revista de Biologia Tropical, 23, 5–15.

**Bancroft, T. L.** (1927). Flagellates in certain Queensland plants. Proceedings of the Royal Society of Queensland, 39, 22.

**Barbieri, L., Falasca, A., Franceschi, C., Licastro, F., Rossi, C. A., and Stirpe, F.** (1983). Purification and properties of two lectins from the latex of the euphorbiaceous plants *Hura crepitans* L. (sand-box tree) and *Euphorbia characias* L. (Mediterranean spurge). The Biochemical Journal, 215(3), 433–9.

**Barry, J.** (1998). VSG gene control and infectivity strategy of metacyclic stage *Trypanosoma brucei*. Molecular and Biochemical Parasitology, 91(1), 93–105.

**Barry, J. D., and McCulloch, R.** (2001). Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite. Advances in Parasitology, 49, 1–70.

**Bastin, P., Bagherzadeh, Z., Matthews, K. R., and Gull, K.** (1996). A novel epitope tag system to study protein targeting and organelle biogenesis in *Trypanosoma brucei*. Molecular and Biochemical Parasitology, 77(2), 235–9.

**Bebber, D. P., Ramotowski, M. A. T., and Gurr, S. J.** (2013). Crop pests and pathogens move polewards in a warming world. Nature Climate Change, 3(11), 985–988.

**Beneteau, J., Renard, D., Marché, L., Douville, E., Lavenant, L., Rahbé, Y., Dupont, D., Vilaine, F., and Dinant, S.** (2010). Binding properties of the *N*-acetylglucosamine and high-mannose N-glycan PP2-A1 phloem lectin in Arabidopsis. Plant Physiology, 153(3), 1345–61.

**Benne, R., Van den Burg, J., Brakenhoff, J. P., Sloof, P., Van Boom, J. H., and Tromp, M. C.** (1986). Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell, 46(6), 819–26.

**Bensaude, M.** (1925). Flagellates in plants: a review of foreign literature. Phytopathology, 15, 273–281.

**Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D. C., Lennard, N. J., Caler, E., Hamlin, N. E., Haas, B., Böhme, U., Hannick, L., Aslett, M. a, Shallom, J., Marcello, L., Hou, L., Wickstead, B., Alsmark, U. C. M., Arrowsmith, C., Atkin, R. J., Barron, A. J., Bringaud, F., Brooks, K., Carrington, M., Cherevach, I., Chillingworth, T.-J., Churcher, C., Clark, L. N., Corton, C. H., Cronin, A., Davies, R. M., Doggett, J., Djikeng, A., Feldblyum, T., Field, M. C., Fraser, A., Goodhead, I., Hance, Z., Harper, D., Harris, B. R., Hauser, H., Hostetler, J., Ivens, A., Jagels, K., Johnson, D., Johnson, J., Jones, K., Kerhornou, A. X., Koo, H., Larke, N., Landfear, S., Larkin, C., Leech, V., Line, A., Lord, A., Macleod, A., Mooney, P. J., Moule, S., Martin, D. M. a, Morgan, G. W., Mungall, K., Norbertczak, H., Ormond, D., Pai, G., Peacock, C. S., Peterson, J., Quail, M. a, Rabbinowitsch, E., Rajandream, M.-A., Reitter, C., Salzberg, S. L., Sanders, M., Schobel, S., Sharp, S., Simmonds, M., Simpson, A. J., Tallon, L., Turner, C. M. R., Tait, A., Tivey, A. R., Van Aken, S., Walker, D., Wanless, D., Wang, S., White, B., White, O., Whitehead, S., Woodward, J., Wortman, J., Adams, M. D., Embley, T. M., Gull, K., Ullu, E., Barry, J. D., Fairlamb, A. H., Opperdoes, F., Barrell, B. G., Donelson, J. E., Hall, N., Fraser, C. M., Melville, S. E., and El-Sayed, N. M.** (2005). The genome of the African trypanosome *Trypanosoma brucei*. Science, 309(5733), 416–22.

**Berrio, C., and Ocho, A.** (1991). La Marchitez Sorpresiva de la Palma Aceitera en la Zona Sur del Lago de Maracaibo. Revista de Difusión de Tecnología Agrícola Y Pesquera Del FONAIAP, 38.

**Bezerra, J. L., Oliveira, D., and Carvalho, A.** (1983). *Phytomonas staheli* em piaçaveira (*Attalea funifera* Mart.) na Bahia, Brasil: CEPLACI CEPEC, Informe Técnico, 226.

**Boetzer, M., and Pirovano, W.** (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics, 15(1), 211.

**Bolger, A. M., Lohse, M., and Usadel, B.** (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England), btu170–.

**Borst, P., and Cross, G. A.** (1982). Molecular basis for trypanosome antigenic variation. Cell, 29(2), 291–303.

**Brazil, R. P., Fiorini, J. E., and Silva, P. M. F. E**. (1990). *Phytomonas* sp., a trypanosomatid parasite of tomato, isolated from salivary glands of *Phthia picta* (Hemiptera: Coreidae) in southeast Brazil. Memórias Do Instituto Oswaldo Cruz, 85(2), 239–240.

**Breitenbach, H. H., Wenig, M., Wittek, F., Jordá, L., Maldonado-Alconada, A. M., Sarioglu, H., Colby, T., Knappe, C., Bichlmeier, M., Pabst, E., Mackey, D., Parker, J. E., and Vlot, A. C.** (2014). Contrasting Roles of the Apoplastic Aspartyl Protease APOPLASTIC, ENHANCED DISEASE SUSCEPTIBILITY1-DEPENDENT1 and LEGUME LECTIN-LIKE PROTEIN1 in Arabidopsis Systemic Acquired Resistance. Plant Physiology, 165(2), 791–809.

**Bringaud, F., Rivière, L., and Coustou, V.** (2006). Energy metabolism of trypanosomatids: adaptation to available carbon sources. Molecular and Biochemical Parasitology.

**Broadhead, R., Dawe, H. R., Farr, H., Griffiths, S., Hart, S. R., Portman, N., Shaw, M. K., Ginger, M. L., Gaskell, S. J., McKean, P. G., and Gull, K.** (2006). Flagellar motility is required for the viability of the bloodstream trypanosome. Nature, 440(7081), 224–7.

**Brown, C. T., Crusoe, M. R., Edvenson, G., Fish, J., Howe, A., McDonald, E., Nahum, J., Nanlohy, K., Ortiz-Zuazaga, H., Pell, J., Simpson, J., Scott, C., Srinivasan, R. R., and Zhang, Q.** (2014). The khmer software package: enabling efficient sequence analysis.

**Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H.** (2012). A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. Genomics.

**Buchan, D., Minneci, F., Nugent, T., Bryson, K., and Jones, D.** (2013). UCL-CS Bioinformatics: PSIPRED citation.

**Burstein, A., and Romero, O.** (1956). Flagelados en el latex de la *Jatropha macrantha* (*Huanarpo hembra*). Archivos Peruanos de Patologia Y Clinica, 10, 1–12.

**Buscaglia, C. A., Campo, V. A., Frasch, A. C. C., and Di Noia, J. M.** (2006). *Trypanosoma cruzi* surface mucins: host-dependent coat diversity. Nature Reviews. Microbiology, 4(3), 229–36.

**Camargo, E. P.** (1999). *Phytomonas* and other trypanosomatid parasites of plants and fruit. Advances in Parasitology, 42, 29–112.

**Camargo, E. P., Kastelein, P., and Roitman, I.** (1990). Trypanosomatid parasites of plants (*Phytomonas*). Parasitology Today, 6(1), 22–25.

**Campo, V., Cremona, L., Ja, A., Noia, J. M. Di, Overath, P., Sa, D. O., and Frasch, A. C.** (2002). Gene Discovery in the Freshwater Fish Parasite *Trypanosoma carassii* : Identification of trans -Sialidase-Like and Mucin-Like Genes, 70(12), 7140–7144.

**Carrara, F. E., Jankevicius, S. I., Cavazzana, M., Catarino, L. M., Ueno, C. T., Jankevicius, J. V., Kaneshima, E. N., and Takihata, J. T.** (1992). Detection and in vitro cultivation of flagellate Protozoa found in orange (*Citrus aurantium* L. cultivar pera) and in grape (*Vitis vinifera*). Memorias Do Instituto Oswaldo Cruz, 87, 241.

**Carrari, F., Baxter, C., Usadel, B., Urbanczyk-Wochniak, E., Zanor, M.-I., Nunes-Nesi, A., Nikiforova, V., Centero, D., Ratzka, A., Pauly, M., Sweetlove, L. J., and Fernie, A. R.**

(2006). Integrated analysis of metabolite and transcript levels reveals the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior. Plant Physiology, 142(4), 1380–96.

**Catarino, L. M., Cavazzana, M., Santos, M. A., Ueno, C. T., Jankevicius, S. I., Jankevicius, J. V., Attias, M., and De Souza, W.** (1991). Axenic cultivation and ultrastructure of *Phytomonas* sp. from pomegranate (*Punica* sp.) and from phytophagous hemipteran Leptoglossus sp. Memórias Do Instituto Oswaldo Cruz1, 86, 271.

**Catarino, L. M., Serrano, M. G., Cavazzana, M., Almeida, M. L., Kaneshina, E. K., Campaner, M., Jankevicius, J. V., Teixeira, M. M. G., and Itow-Jankevicius, S.** (2001). Classification of trypanosomatids from fruits and seeds using morphological, biochemical and molecular markers revealed several genera among fruit isolates. FEMS Microbiology Letters, 201(1), 65–72.

**Cavazzana, M., Almeida, M. L., Garcia, J. F., Catarino, L. M., Ono, C. J., Jankevicius, J. V., and Jankevicius, S. I.** (1993). Isolation and cultivation of trypanosomatids from mulberry (*Morus* sp.) and “Guandu” beans (*Cajanus flaws*). Memorias Do Instituto Oswaldo Cruz, 88, 282.

**Cavazzana, M., Baccan, G. C., Camilo, S. B., Tagata, E., Jankevicius, J. V., and Jankevicius, S. I.** (1995). Isolation and in vitro cultivation of trypanosomatids found in apple (*Malus* sp.). 1Memórias Do Instituto Oswaldo Cruz, 90, 246.

**Cavazzana, M., Ueno, C. T., Kaneshima, E. N., T.G., D., Jankevicius, J. V., and Itow-Jankevicius, S.** (1996). Detection and isolation of protozoa in pitanga fruit (*Eugenia* sp., Myrtaceae). Memórias Do Instituto Oswaldo Cruz, 91, 100.

**Cazzulo, J., Stoka, V., and Turk, V.** (2001). The Major Cysteine Proteinase of *Trypanosoma cruzi*: A Valid Target for Chemotherapy of Chagas Disease. Current Pharmaceutical Design, 7(12), 1143–1156.

**Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S. L., Nuwaysir, L. M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E. W., Moritz, R. L., Katz, J. E., Agus, D. B., MacCoss, M., Tabb, D. L., and Mallick, P.** (2012). A cross-platform toolkit for mass spectrometry and proteomics. Nature Biotechnology, 30(10), 918–20.

**Chang, L., and Karin, M.** (2001). Mammalian MAP kinase signalling cascades. Nature, 410(6824), 37–40.

**Chaumont, F., Schanck, A., Blum, J. J., and Opperdoes, F. R.** (1994). Aerobic and anaerobic glucose metabolism of Phytomonas sp. isolated from Euphorbia characias. Molecular and Biochemical Parasitology, 67(2), 321–331.

**Collingridge, P. W., and Kelly, S.** (2012). MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. BMC Bioinformatics, 13(1), 117.

**Collins, T. J.** (2007). ImageJ for microscopy. BioTechniques, 43(1 Suppl), 25–30.

**Conchon, I., Campaner, M., Sbravat, C., and Camargo, E. P.** (1989). Trypanosomatids, Other Than *Phytomonas* spp., Isolated and Cultured from Fruit. The Journal of Protozoology, 36(4), 412–414.

**D'Avila-Levy, C. M., Santos, L. O., Marinho, F. A., Dias, F. A., Lopes, A. H., Santos, A. L. S., and Branquinha, M. H.** (2006). Gp63-like molecules in *Phytomonas serpens*: possible role in the insect interaction. Current Microbiology, 52(6), 439–44.

**Da Silva, R. V., Malvezi, A. D., Augusto, L. da S., Kian, D., Tatakihara, V. L. H. L. H., Yamauchi, L. M., Yamada-Ogatta, S. F., Rizzo, L. V., Schenkman, S., Pinge-Filho, P., da**

**Augusto, L. S., Kian, D., Tatakihara, V. L. H. L. H., Yamauchi, L. M., Yamada-Ogatta, S. F., Rizzo, L. V., Schenkman, S., and Pinge-Filho, P.** (2013). Oral Exposure to *Phytomonas* serpens Attenuates Thrombocytopenia and Leukopenia during Acute Infection with *Trypanosoma cruzi*. PLoS ONE, 8(7), e68299.

**De Assis, R. R., Ibraim, I. C., Nogueira, P. M., Soares, R. P., and Turco, S. J.** (2012). Glycoconjugates in New World species of *Leishmania*: polymorphisms in lipophosphoglycan and glycoinositolphospholipids and interaction with hosts. Biochimica et Biophysica Acta, 1820(9), 1354–65.

**Dean, S., Sunter, J., Wheeler, R. J., Hodkinson, I., Gluenz, E., and Gull, K.** (2015). A toolkit enabling efficient, scalable and reproducible gene tagging in trypanosomatids. Open Biology, 5(1), 140197–140197.

**De Greef, C., & Hamers, R.** (1994). The serum resistance-associated (SRA) gene of Trypanosoma brucei rhodesiense encodes a variant surface glycoprotein-like protein. *Molecular and Biochemical Parasitology*, *68*(2), 277–84.

**Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J. K., Martin, D. B., Nesvizhskii, A. I., and Aebersold, R.** (2010). A guided tour of the Trans-Proteomic Pipeline. Proteomics, 10(6), 1150–9.

**Di Lucca, A. G. T., Trinidad Chipana, E. F., Talledo Albújar, M. J., Dávila Peralta, W., Montoya Piedra, Y. C., and Arévalo Zelada, J. L.** (2013). Slow wilt: another form of Marchitez in oil palm associated with trypanosomatids in Peru. Tropical Plant Pathology, 38(6), 522–533.

**Dietrich, M. A., Arnold, G. J., Nynca, J., Fröhlich, T., Otte, K., and Ciereszko, A.** (2014). Characterization of carp seminal plasma proteome in relation to blood plasma. Journal of Proteomics, 98, 218–32.

**Dollet, M.** (1984). Plant Diseases Caused by Flagellate Protozoa (*Phytomonas*). Annual Review of Phytopathology, 22(1), 115–132.

**Dollet, M., Cambrony, D., and Gargani, D.** (1982). Culture axéique in vitro de *Phytomonas* sp. (Trypanosomatidae) *d'Euphorbe transmis* par *Stenocephalusagilis scop* (Coreide). Comptes Rendus de L'Academie Des Sciences, Paris, 295, 547–550.

**Dollet, M., Gargani, D., Cambrony, D., and Dzido, J. L.** (1983). Observations of flagellate protozoa (*Phytomonas* sp.) in the xylem of *Euphorbia lasiocarpa* in Ecuador. Canadian Journal of Botany, 61(1), 237–240.

**Dollet, M., Giannotti, J., and Ollagnier, M.** (1977). Observations de protozoaires flagelles dans les tubes cribles de Palmiers a huile malades. Comptes Rendus Academie Des Seances Paris, 284 (Serie D), 643–645.

**Dollet, M., Lopez, G., Genty, P., and Dzido, J. L.** (1979). Récherches actuelles de l'I.R.H.O. sur les dépérissements du cocotier et du palmier a huile en Amérique du Sud, associes aux protozoaires flagelles intraphlohiques (*Phytomonas*). Oléagineux, 34, 238–241.

**Dollet, M., Sturm, N. R., and Campbell, D. A.** (2001). The spliced leader RNA gene array in phloem-restricted plant trypanosomatids (*Phytomonas*) partitions into two major groupings: epidemiological implications. Parasitology, 122(Pt 3), 289–97.

**Dollet, M., Sturm, N. R., and Campbell, D. A.** (2012). The internal transcribed spacer of ribosomal RNA genes in plant trypanosomes (*Phytomonas* spp.) resolves 10 groups. Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases, 12(2), 299–308.

**Dollet, M., Sturm, N. R., Sánchez-Moreno, M., and Campbell, D. A.** (2000). 5S ribosomal RNA gene repeat sequences define at least eight groups of plant trypanosomatids (*Phytomonas* spp.): phloem-restricted pathogens form a distinct section. The Journal of Eukaryotic Microbiology, 47(6), 569–74.

**Donelson, J. E.** (2003). Antigenic variation and the African trypanosome genome. Acta Tropica, 85(3), 391–404.

**Donovan, C.** (1909). Kala-azar in Madras, especially with regard to its connecxion with the dog and the bug (Chonnorrhinus). The Lanclet, 177, 1195–1196.

**Dos Santos, S. L., Freitas, L. M., Lobo, F. P., Rodrigues-Luiz, G. F., Mendes, T. A. de O., Oliveira, A. C. S. and Bartholomeu, D. C.** (2012). The MASP family of *Trypanosoma cruzi*: changes in gene expression and antigenic profile during the acute phase of experimental infection. PLoS Neglected Tropical Diseases, 6(8), e1779.

**Elia, G.** (2008). Biotinylation reagents for the study of cell surface proteins. Proteomics, 8(19), 4012–24.

**Elias, C. G. R., Chagas, M. G., Souza-Gonçalves, A. L., Pascarelli, B. M. O., D'Avila-Levy, C. M., Branquinha, M. H., and Santos, A. L. S.** (2012). Differential expression of cruzipain- and gp63-like molecules in the phytoflagellate trypanosomatid *Phytomonas serpens* induced by exogenous proteins. Experimental Parasitology, 130(1), 13–21.

**Elias, C. G. R., Pereira, F. M., Dias, F. A., Silva, T. L. A., Lopes, A. H. C. S., D'Avila-Levy, C. M., Branquinha, M. H., and Santos, A. L. S.** (2008). Cysteine peptidases in the tomato trypanosomatid *Phytomonas serpens*: influence of growth conditions, similarities with cruzipain and secretion to the extracellular environment. Experimental Parasitology, 120(4), 343–52.

**El-On, J., Bradley, D. J., and Freeman, J. C.** (1980). *Leishmania donovani*: action of excreted factor on hydrolytic enzyme activity of macrophages from mice with genetically different resistance to infection. Experimental Parasitology, 49(2), 167–74.

**El-Sayed, N. M., Myler, P. J., Bartholomeu, D. C., Nilsson, D., Aggarwal, G., Tran, A.-N., Ghedin, E., Worthey, E. A., Delcher, A. L., Blandin, G., Westenberger, S. J., Caler, E., Cerqueira, G. C., Branche, C., Haas, B., Anupama, A., Arner, E., Aslund, L., Attipoe, P.,**

**Bontempi, E., Bringaud, F., Burton, P., Cadag, E., Campbell, D. A., Carrington, M., Crabtree, J., Darban, H., da Silveira, J. F., de Jong, P., Edwards, K., Englund, P. T., Fazelina, G., Feldblyum, T., Ferella, M., Frasch, A. C., Gull, K., Horn, D., Hou, L., Huang, Y., Kindlund, E., Klingbeil, M., Kluge, S., Koo, H., Lacerda, D., Levin, M. J., Lorenzi, H., Louie, T., Machado, C. R., McCulloch, R., McKenna, A., Mizuno, Y., Mottram, J. C., Nelson, S., Ochaya, S., Osoegawa, K., Pai, G., Parsons, M., Pentony, M., Pettersson, U., Pop, M., Ramirez, J. L., Rinta, J., Robertson, L., Salzberg, S. L., Sanchez, D. O., Seyler, A., Sharma, R., Shetty, J., Simpson, A. J., Sisk, E., Tammi, M. T., Tarleton, R., Teixeira, S., Van Aken, S., Vogt, C., Ward, P. N., Wickstead, B., Wortman, J., White, O., Fraser, C. M., Stuart, K. D., and Andersson, B.** (2005). The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science, 309(5733), 409–15.

**Emms, D. M., and Kelly, S.** (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology, 16(1), 157.

**Fantham, H. B.** (1925). Some parasitic Protozoa found in South Africa VIII. South African Journal of Science, 12, 346–354.

**Ferguson, M. A.** (1999). The structure, biosynthesis and functions of glycosylphosphatidylinositol anchors, and the contributions of trypanosome research. Journal of Cell Science, 112, 2799–2809.

**Fiebig, M., Gluenz, E., Carrington, M., and Kelly, S.** (2014). SLaP mapper: a webserver for identifying and quantifying spliced-leader addition and polyadenylation site usage in kinetoplastid genomes. Molecular and Biochemical Parasitology, 196(2), 71–4.

**Field, M. C., and Carrington, M.** (2009). The trypanosome flagellar pocket. Nature Reviews. Microbiology, 7(11), 775–86.

**Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M.**

(2014). Pfam: the protein families database. Nucleic Acids Research, 42(Database issue), D222–30.

**Fiorini, J. E., Cucolichio, G., Rezende, D., Nascimento, L. C., and Faria e Silva, P. M.** (1995). Isolation and cloning of a trypanosomatid with choanomastigote form from mulberry. Memórias Do Instituto Oswaldo Cruz, 90, 247.

**Fiorini, J. E., Fiorini, C. M., DePaula, K. F., Pedreira, R. C., Almeida, D. C., and Brazil, R. P.** (1990). Cultivation of a *Phytomonas* isolated from orange (*Citrus aurantium* L. var. Lumia) and *Leptoglossus stigma* (Herbst, 1784) (Hemiptera Coreidae). Memórias Do Instituto Oswaldo Cruz, 36, 412–414.

**Fiorini, J. E., Silva, F. P. M., Brazil, R. P., Roitman, I., Angluster, J., and De Souza, W.** (1986). Detection of trypanosomatids in *Solanum gilo* and *Solanum lycopersicum* in Alfenas M.G., Brazil. Memórias Do Instituto Oswaldo Cruz, 81, 33.

**Franca, C.** (1911). Sur l'existence en Portugal de "*Leptomonas davidi*" Lafont dans le latex de "*Euphorbia peplus* L." et "*E. segetalis* L.". Bulletin de La Société de Pathologie Exotique, 4, 532–534.

**Franca, C.** (1920). La flagellose des Euphorbes II. Archiv Fur Protistenkunde, 34, 108–132.

**Franchini, G.** (1922a). Amibes et autres protozoaires de plants a latex du museum de Paris. Bulletin de La Societe de Pathologie Exotique, 15, 197–203.

**Franchini, G.** (1922b). Essais d'inoculation aux souris blanches du latex parasite de differentes especes d'euphorbes. Annales de L'instutut Pasteur, 36, 873–881.

**Franchini, G.** (1922c). Flagelles et amibes d'une Urtiacee exotique, *Ficus parietalis*. Bulletin de La Societe de Pathologie Exotique, 15, 399–405.

**Franchini, G.** (1922d). Nouvelles recherches sur les trypanosomes des euphorbes et sur leur culture. Bulletin de La Societe de Pathologie Exotique, 15, 299–303.

**Franchini, G.** (1922e). Sur un flagelle nouveau du latex de deux Apocynees. Bulletin de La Societe de Pathologie Exotique, 15, 109–113.

**Franchini, G.** (1922f). Sur un trypanosome du latex de deux especes d'Euphorbes. Bulletin de La Societe de Pathologie Exotique, 15, 18–23.

**Franchini, G.** (1923a). Nouvelles recherches sur les protozoaires des plants a latex. Bulletin de La Société de Pathologie Exotique, 16, 655–659.

**Franchini, G**. (1923b). Sur un protozoaire d'*Euphorbia cereiformis* et sur sa culture. Bulletin de La Société de Pathologie Exotique1, 16, 642–646.

**Franchini, G.** (1931). Étude sur un flagellé special du latex d'un figuier de l'Erythrée (*Ficus hochstetteri* (MG) A. Rich). Bulletin de La Société de Pathologie Exotique, 24, 843–848.

**Frankenberg, N., Moser, J., and Jahn, D.** (2003). Bacterial heme biosynthesis and its biotechnological application. Applied Microbiology and Biotechnology, 63(2), 115–27.

**Frasch, A. C.** (2000). Functional Diversity in the Trans-sialidase and Mucin Families in *Trypanosoma cruzi*. Parasitology Today, 16(7), 282–286.

**Freymuller, E., Milder, R., Jankevicius, J. V., Jankevicius, S. I., and Camargo, E. P.** (1990). Ultrastructural Studies on the Trypanosomatid *Phytomonas serpens* in the Salivary Glands of a Phytophagous Hemipteran. The Journal of Protozoology, 37(3), 225–229.

**Gadelha, C., Zhang, W., Chamberlain, J. W., Chait, B. T., Wickstead, B., and Field, M. C.** (2015). Architecture of a Host-Parasite Interface: Complex Targeting Mechanisms Revealed Through Proteomics. Molecular & Cellular Proteomics : MCP, 14(7), 1911–26.

**Gallagher, S.** (2006). One-dimensional SDS gel electrophoresis of proteins. Current Protocols in Immunology.

**Galli-Valerio, B.** (1921). La flagelliase des euphoriacees en Suisse. Schweizerische Medizinische Wochenschrift, 2(50), 1154.

**Gargani, D., Menara, A., Segur, C., and Dollet, M.** (1992). In vitro cultivation of *Phytomonas* from latex and phloem-restricted *Phytomonas*. Oléagineux, 47, 596.

**Gaupels, F., and Ghirardo, A.** (2013). The extrafascicular phloem is made for fighting. Frontiers in Plant Science, 4, 187.

**Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H.** (2004). Open mass spectrometry search algorithm. Journal of Proteome Research, 3(5), 958–64.

**Gibbs, A. J.** (1957). *Leptomonas serpens* nsp. parasitic in the digestive tract and salivary glands of *Nezara viridula* (Pentatomidae) and in the sap of *Solanum lycopersicum* (tomato) and other plants. Parasitology, 47(3-4), 297–303.

**Gluenz, E., Höög, J. L., Smith, A. E., Dawe, H. R., Shaw, M. K., and Gull, K.** (2010). Beyond 9+0: noncanonical axoneme structures characterize sensory cilia from protists to humans. FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology, 24(9), 3117–21.

**Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B.** (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America, 108(4), 1513–8.

**Goder, V., and Spiess, M.** (2001). Topogenesis of membrane proteins: determinants and dynamics. FEBS Letters, 504(3), 87–93.

**Gómez-Gómez, L., and Boller, T.** (2000). FLS2: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. Molecular Cell, 5(6), 1003–11.

**Gray, M. W.** (2012). Evolutionary Origin of RNA Editing. *Biochemistry*, *51*(26), 5235–5242.

**Green, N. M.** (1975). Advances in Protein Chemistry Volume 29. Advances in Protein Chemistry (Vol. 29). Elsevier.

**Griffin, N. M., Yu, J., Long, F., Oh, P., Shore, S., Li, Y., Koziol, J. a, and Schnitzer, J. E.** (2010). Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nature Biotechnology, 28(1), 83–9.

**Gupta, G., Oghumu, S., and Satoskar, A. R.** (2013). Mechanisms of immune evasion in leishmaniasis. Advances in Applied Microbiology, 82, 155–84.

**Haag, J., O'HUigin, C., and Overath, P.** (1998). The molecular phylogeny of trypanosomes: evidence for an early divergence of the Salivaria. Molecular and Biochemical Parasitology, 91(1), 37–49.

**Hamilton, P. B., Gibson, W. C., and Stevens, J. R.** (2007). Patterns of co-evolution between trypanosomes and their hosts deduced from ribosomal RNA and protein-coding gene phylogenies. Molecular Phylogenetics and Evolution, 44(1), 15–25.

**Hamilton, P. B., Stevens, J. R., Gidley, J., Holz, P., and Gibson, W. C.** (2005). A new lineage of trypanosomes from Australian vertebrates and terrestrial bloodsucking leeches (Haemadipsidae). International Journal for Parasitology, 35(4), 431–43.

**Harvey, R. B., and Lee, S. B.** (1943). Flagellates of laticiferous plants. Plant Physiology, 18, 633–355.

**Hayashi, F., Smith, K. D., Ozinsky, A., Hawn, T. R., Yi, E. C., Goodlett, D. R., Eng, J. K., Akira, S., Underhill, D. M., and Aderem, A.** (2001). The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. Nature, 410(6832), 1099–103.

**Hecker, H., Schwarzenbach, M., and Rudin, W.** (1990). Development and interactions of *Trypanosoma rangeli* in and with the reduviid bug *Rhodnius prolixus*. Parasitology Research, 76(4), 311–318.

**Hewitt, R.** (1940). Phytomonad Flagellates from Mexican Milkweeds. Journal of Parasitology, 26(2), 35–36.

**Higgins, M. K., and Carrington, M.** (2014). Sequence variation and structural conservation allows development of novel function and immune evasion in parasite surface protein families. Protein Science : A Publication of the Protein Society, 23(4), 354–65.

**Hoare, C. A., and Wallace, F. G.** (1966). Developmental Stages of Trypanosomatid Flagellates: a New Terminology. Nature, 212(5068), 1385–1386.

**Holmes, F.** (1924). Herpetomonad flagellates in the lates of milkweed in Maryland. Phytopathology, 14, 146–151.

**Holmes, F.** (1925). Geographical distribution of the milkweed flagellate, *Herpetomonas elmassiani* (Mignone). Phytopathology, 15, 296–299.

**Holmes, F.** (1931). *Herpetomonas bancrofti* n. sp. from the latex of a ficus in Queensland. Contributions from Boyce Thompson Intitute for Plant Research, 3, 375–384.

http://faostat.fao.org/

http://www.cdc.gov/parasites/sleepingsickness/biology.html. (2015). CDC African Trypanosomiasis.

**Iborra, F., Kimura, H., and Cook, P.** (2004). The functional organization of mitochondrial genomes in human cells. BMC Biology, 2, 2–9.

**Ienne, S., Pappas, G., Benabdellah, K., González, A., and Zingales, B.** (2012). Horizontal gene transfer confers fermentative metabolism in the respiratory-deficient plant trypanosomatid *Phytomonas serpens.* Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases, 12(3), 539–48.

**Ioff, I**. (1927). The finding of *Leptomonas davidi* in the Saratov (Volga basin) (in Russian). Russian Journal of Tropical Medicine, 5, 47–49.

**Ivens, A. C., Peacock, C. S., Worthey, E. a, Murphy, L., Aggarwal, G., Berriman, M., Sisk, E., Rajandream, M.-A., Adlem, E., Aert, R., Anupama, A., Apostolou, Z., Attipoe, P., Bason, N., Bauser, C., Beck, A., Beverley, S. M., Bianchettin, G., Borzym, K., Bothe, G., Bruschi, C. V, Collins, M., Cadag, E., Ciarloni, L., Clayton, C., Coulson, R. M. R., Cronin, A., Cruz, A. K., Davies, R. M., De Gaudenzi, J., Dobson, D. E., Duesterhoeft, A., Fazelina, G., Fosker, N., Frasch, A. C., Fraser, A., Fuchs, M., Gabel, C., Goble, A., Goffeau, A., Harris, D., Hertz-Fowler, C., Hilbert, H., Horn, D., Huang, Y., Klages, S., Knights, A., Kube, M., Larke, N., Litvin, L., Lord, A., Louie, T., Marra, M., Masuy, D., Matthews, K., Michaeli, S., Mottram, J. C., Müller-Auer, S., Munden, H., Nelson, S., Norbertczak, H., Oliver, K., O'neil, S., Pentony, M., Pohl, T. M., Price, C., Purnelle, B., Quail, M. a, Rabbinowitsch, E., Reinhardt, R., Rieger, M., Rinta, J., Robben, J., Robertson, L., Ruiz, J. C., Rutter, S., Saunders, D., Schäfer, M., Schein, J., Schwartz, D. C., Seeger, K., Seyler, A., Sharp, S., Shin, H., Sivam, D., Squares, R., Squares, S., Tosato, V., Vogt, C., Volckaert, G., Wambutt, R., Warren, T., Wedler, H., Woodward, J., Zhou, S., Zimmermann, W., Smith, D. F., Blackwell, J. M., Stuart, K. D., Barrell, B., and Myler, P. J**. (2005). The genome of the kinetoplastid parasite, *Leishmania major*. Science , 309(5733), 436–42.

**Jackson, A. P.** (2010). The evolution of amastin surface glycoproteins in trypanosomatid parasites. Molecular Biology and Evolution, 27(1), 33–45.

**Jackson, A. P., Allison, H. C., Barry, J. D., Field, M. C., Hertz-Fowler, C., and Berriman, M.** (2013). A cell-surface phylome for African trypanosomes. PLoS Neglected Tropical Diseases, 7(3), e2121.

**Jackson, A. P., Berry, A., Aslett, M., Allison, H. C., Burton, P., Vavrova-Anderson, J., Berriman, M.** (2012). Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. Proceedings of the National Academy of Sciences of the United States of America, 109(9), 3416–21.

**Jaffe, C. L., and Dwyer, D. M.** (2003). Extracellular release of the surface metalloprotease, gp63, from *Leishmania* and insect trypanosomatids. Parasitology Research, 91(3), 229–37.

**Jankevicius, J. V., Jankevicius, S. I., Campaner, M., Conchon, I., Madea, L. A., Telxeira, M. M. G., Freymuller, E., and Camargo, E. P.** (1989). Life Cycle and Culturing of *Phytomonas serpens* (Gibbs), a Trypanosomatid Parasite of Tomatoes. The Journal of Protozoology, 36(3), 265–271.

**Jankevicius, S. I., Almeida, I. L., Jankevicius, J. V., Cavazzana, M., Attias, M., and Souza, W.** (1993). Axenic Cultivation of Trypanosomatids Found in Corn (*Zea mays*) and in Phytophagous Hemipterans (*Leptoglossus zonatus* Coreidae) and Their Experimental Transmission. The Journal of Eukaryotic Microbiology, 40(5), 576–581.

**Jankevicius, S. I., Jankevicius, J. V., Menezes, M. C. M. D., Torrezan, H. C., Menezes, J. R., and Rezende, M. I.** (1987). *Phytornonas* sp. found in leguminous crops. Memórias Do Instituto Oswaldo Cruz, 82, 35.

**Jashni, M. K., Mehrabi, R., Collemare, J., Mesarich, C. H., and de Wit, P. J. G. M.** (2015). The battle in the apoplast: further insights into the roles of proteases and their inhibitors in plant-pathogen interactions. Frontiers in Plant Science, 6, 584.

**Jaskowska, E., Butler, C., Preston, G., and Kelly, S.** (2015). *Phytomonas*: Trypanosomatids adapted to plant environments. PLoS Pathogens, 11(1), e1004484.

**Johnson, R. J., McCoy, J. G., Bingman, C. A., Phillips, G. N., and Raines, R. T.** (2007). Inhibition of human pancreatic ribonuclease by the human ribonuclease inhibitor protein. Journal of Molecular Biology, 368(2), 434–49.

**Jones, J. D. G., and Dangl, J. L.** (2006). The plant immune system. Nature, 444(7117), 323–9.

**Kangussu-Marcolino, M. M., de Paiva, R. M. C., Araújo, P. R., de Mendonça-Neto, R. P., Lemos, L., Bartholomeu, D. C., Mortara, R. A., DaRocha, W. D., and Teixeira, S. M. R.**

(2013). Distinct genomic organization, mRNA expression and cellular localization of members of two amastin sub-families present in *Trypanosoma cruzi*. BMC Microbiology, 13(1), 10.

**Kastelein, P.** (1987). Investigations on “Hartrot” of coconut and oilpalms in Suriname. PhD Thesis.

**Kastelein, P., and Camargo, E. P.** (1990). Trypanosomatid protozoa in fruit of Solanaceae in southeastern Brazil. Memorias Do Instituto Oswaldo Cruz, 85, 413–417.

**Kastelein, P., Karyosemito, C., and Segeren, P.** (1984). The role of weeds in the incidence of hartrot or fatal wilt of palms: 111. Enumeration of weeds as hosts of the protozoa Phytomonas (Trypanosomatidae) and attempts of culturing the flagellate. Surinaamse Landbouw, 32, 1542.

**Kehr, J.** (2006). Phloem sap proteins: their identities and potential roles in the interaction between plants and phloem-feeding insects. Journal of Experimental Botany, 57(4), 767–74.

**Keller, O., Kollmar, M., Stanke, M., and Waack, S**. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics, 27(6), 757–63.

**Kelly, S., Ivens, A., Manna, P. T., Gibson, W., and Field, M. C.** (2014). A draft genome for the African crocodilian trypanosome *Trypanosoma grayi.* Scientific Data, 1.

**Kelly, S., Reed, J., Kramer, S., Ellis, L., Webb, H., Sunter, J., Salje, J., Marinsek, N., Gull, K., Wickstead, B., and Carrington, M.** (2007). Functional genomics in *Trypanosoma brucei*: A collection of vectors for the expression of tagged proteins from endogenous and ectopic gene loci. Molecular and Biochemical Parasitology, 154(1), 103–109.

**Kitajima, E. W.** (1986). Flagellate protozoon associated with poor development of the root system of Cassava in the Espirito Santo State, Brazil. Phytopathology, 76(6), 638.

**Kopylova, E., Noé, L., and Touzet, H.** (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics, 28(24), 3211–7.

**Koreny, L., Lukes, J., and Al., E.** (2012). Aerobic kinetoplastid flagellate Phytomonas does not require heme for viability. PNAS, 109(10), 3808–3813.

**Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L.** (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of Molecular Biology, 305(3), 567–80.

**Lafont, A.** (1909). Sur la présence d'un Leptomonas, parasite de la classe des Flagelles dans le lates de l'Euphorbia pilulifera. Comptes Rendus Des Seances de La Societe de Biologie, 66, 1011–1013.

**Lafont, A**. (1910). Sur la présence d'un Leptomonas, parasite de la classe des Flagelles, dans le latex de trois Euphorbiacees. Annales de L'instutut Pasteur, Paris, 24, 205–219.

**Lafont, A.** (1911). Sur la transmission du *Leptomonas davidi* des euphorbes par un hemiptere, *Nysius euphorbiae*. Comptes Rendus Des Seances de La Societe de Biologie, 70, 58–59.

**Langmead, B., and Salzberg, S. L.** (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–9.

**Laveran, A., and Franchini, G.** (1920). Contribution a l'étude de la flagellose des euphorbes. Bulletin de La Société de Pathologie Exotique, 13, 796–800.

**Leeuw, G. T. N. De, and Kastelein, P.** (1987). Observation of flagellate protozoa (*Phytomonas* sp.) in *Cocos nucifera, Cecropia palmata* and *Euphorbia hirta*. Journal of Phytopathology, 120(1), 25–30.

**Lom, J., and Dyková, I**. (1992). Protozoan parasites of fishes. Elsevier.

**Lukeš, J., Paris, Z., Regmi, S., Breitling, R., Mureev, S., Kushnir, S., Pyatkov, K., Jirků, M., and Alexandrov, K. A.** (2006). Translational initiation in Leishmania tarentolae and Phytomonas serpens (Kinetoplastida) is strongly influenced by pre-ATG triplet and its 5′ sequence context. Molecular and Biochemical Parasitology, 148(2), 125–132.

**Lukeš, J., Skalický, T., Týč, J., Votýpka, J., and Yurchenko, V.** (2014). Evolution of parasitism in kinetoplastid flagellates. Molecular and Biochemical Parasitology.

**MacGregor, P., Szöőr, B., Savill, N. J., and Matthews, K. R.** (2012). Trypanosomal immune evasion, chronicity and transmission: an elegant balancing act. Nature Reviews. Microbiology, 10(6), 431–8.

**Magán, R., Marín, C., Salas, J. M., Barrera-Pérez, M., Rosales, M. J., and Sánchez-Moreno, M.** (2004). Cytotoxicity of three new triazolo-pyrimidine derivatives against the plant trypanosomatid: *Phytomonas* sp. isolated from *Euphorbia characias*. Memórias Do Instituto Oswaldo Cruz, 99(6), 651–656.

**Manna, P. T., Kelly, S., & Field, M. C.** (2013). Adaptin evolution in kinetoplastids and emergence of the variant surface glycoprotein coat in African trypanosomatids. Molecular Phylogenetics and Evolution, 67(1), 123–128.

**Marín, C., Alberge, B., Dollet, M., Pagès, M., and Bastien, P.** (2008). First complete chromosomal organization of a protozoan plant parasite (*Phytomonas* spp.). Genomics, 91(1), 88–93.

**Maslov, D. a, Votýpka, J., Yurchenko, V., and Lukeš, J.** (2013). Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. Trends in Parasitology, 29(1), 43–52.

**Maslov, D. A., Nawathean, P., and Scheel, J.** (1999). Partial kinetoplast-mitochondrial gene organization and expression in the respiratory deficient plant trypanosomatid *Phytomonas serpens.* Molecular and Biochemical Parasitology, 99(2), 207–21.

**Mayor, S., and Maxfield, F. R.** (1995). Insolubility and redistribution of GPI-anchored proteins at the cell surface after detergent treatment. Molecular Biology of the Cell, 6(7), 929–944.

**McGhee, R. B., and Hanson, W. L.** (1964). Comparison of the Life Cycle of *Leptomonas oncopelti* and *Phytomonas elmassiani* . The Journal of Protozoology, 11(4), 555–562.

**McGhee, R. B., and McGhee.** (1971). The Relation of Migration of *Oncopeltus fasciatus* to Distribution of *Phytomonas elmassiani* in the Eastern United States. The Journal of Protozoology, 18(2), 344–352.

**McGhee, R. B., and Postell, F. J.** (1976). Axenic cultivation of *Phytomonas davidi* Lafont (Trypanosomatidae), a symbiote of laticiferous plants (Euphorbiaceae). The Journal of Protozoology, 23(2), 238–241.

**McNeely, T. B., and Turco, S. J.** (1990). Requirement of lipophosphoglycan for intracellular survival of *Leishmania donovani* within human monocytes. Journal of Immunology, 144(7), 2745–50.

**Medina, J. M., Rodrigues, J. C. F., Moreira, O. C., Atella, G., Souza, W. de, and Barrabin, H.** (2015). Mechanisms of growth inhibition of *Phytomonas serpens* by the alkaloids tomatine and tomatidine. Memorias Do Instituto Oswaldo Cruz, 110(1), 48–55.

**Mehlert, A., Bond, C. S., and Ferguson, M. A. J.** (2002). The glycoforms of a *Trypanosoma brucei* variant surface glycoprotein and molecular modeling of a glycosylated surface coat. Glycobiology, 12(10), 607–12.

**Menara, A., Dollet, M., Gargani, D., and Louise, C.** (1988). Culture in vitro sur cellules d'invertebres des *Phytomonas* spp. (Trypanosomatidae) associes au hartrot, maladie du cocotier. Comptes Rendus de L'Academie Des Sciences, 307, 597–602.

**Migone, L. E.** (1916). Un nouveau flagellé des plants: *Leptomonas elmassiani*. Bulletin de La Société de Pathologie Exotique, 9, 356–359.

**Miranda, M. R., Sayé, M., Reigada, C., Carrillo, C., and Pereira, C. A.** (2015). *Phytomonas*: A non-pathogenic trypanosomatid model for functional expression of proteins. Protein Expression and Purification, 114, 44–47.

**Molinas, S. M., Altabe, S. G., Opperdoes, F. R., Rider, M. H., Michels, P. A. M., and Uttaro, A. D.** (2003). The multifunctional isopropyl alcohol dehydrogenase of *Phytomonas* sp. could

be the result of a horizontal gene transfer from a bacterium to the trypanosomatid lineage. The Journal of Biological Chemistry, 278(38), 36169–75.

**Muller, E., Gargani, D., Schaeffer, V. V., Stevens, J., Fernandez-Becerra, C., Sanchez-Moreno, M., and Dollet, M.** (1994). Variability in the phloem restricted plant trypanosomes (*Phytomonas* spp) associated with wilts of cultivated crops. European Journal of Plant Pathology, 100(6), 425–434.

**Nawathean, P., and Maslov, D. A.** (2000). The absence of genes for cytochrome c oxidase and reductase subunits in maxicircle kinetoplast DNA of the respiration-deficient plant trypanosomatid *Phytomonas*. Current Genetics, 38(2), 95–103.

**Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G.** (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering, 10(1), 1–6.

**Nieschulz, O.** (1924). Zur Morphologie der Kulturformen einer Herpetomonas aus *Euphorbia cereiformis.* Zentralblatt Für Bakteriologie, Parasitenkunde, Infektionskrankheiten Und Hygiene. Abteilung Originale, 61(31), 1–316.

**Nikolenko, S. I., Korobeynikov, A. I., and Alekseyev, M. A.** (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC Genomics, 14(Suppl 1), S7.

**Noguchi, H., and Tilden, E. B.** (1926). Comparative studies of Herpetomonads and Leishmanias. The Journal of Experimental Medicine, 44, 307–325.

**Oladiran, A., and Belosevic, M.** (2012). Recombinant glycoprotein 63 (Gp63) of *Trypanosoma carassii* suppresses antimicrobial responses of goldfish (*Carassius auratus* L.) monocytes and macrophages. International Journal for Parasitology, 42(7), 621–33.

**Oliveira, A. P., Silva, L. R., Andrade, P. B., Valentão, P., Silva, B. M., Gonçalves, R. F., Pereira, J. A., and Guedes de Pinho, P.** (2010). Further insight into the latex metabolite profile of *Ficus carica*. Journal of Agricultural and Food Chemistry, 58(20), 10855–63.

**Overath, P., Haag, J., Mameza, M. G., and Lischke, A.** (1999). Freshwater fish trypanosomes: definition of two types, host control by antibodies and lack of antigenic variation. Parasitology, 119(06), 591–601.

**Overath, P., Ruoff, J., Stierhof, Y. D., Haag, J., Tichy, H., Dyková, I., and Lom, J.** (1998). Cultivation of bloodstream forms of *Trypanosoma carassii*, a common parasite of freshwater fish. Parasitology Research, 84(5), 343–7.

**Pal, A., Hall, B. S., Jeffries, T. R., and Field, M. C.** (2003). Rab5 and Rab11 mediate transferrin and anti-variant surface glycoprotein antibody recycling in *Trypanosoma brucei*. The Biochemical Journal, 374(Pt 2), 443–51.

**Parthasarathy, M. V, VAN Slobbe, W. G., and Soudant, C.** (1976). Trypanosomatid flagellate in the phloem of diseased coconut palms. Science, 192(4246), 1346–8.

**Pays, E., Vanhamme, L., and Pérez-Morga, D.** (2004). Antigenic variation in *Trypanosoma brucei*: facts, challenges and mysteries. Current Opinion in Microbiology, 7(4), 369–74.

**Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H.** (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods, 8(10), 785–6.

**Podlipaev, S. A.** (1986). *Phytomonas elmassiani* (Mastigophora: Trypanosomadina) from the plant *Cynanchum sibiricum* (Asclepiadaceae) in Central Asia and Kazakhstan (in Russian). Proceedings of the Zoological Institute of the Academy of Sciences, USSR, 144, 61–65.

**Poisson, G., Chauve, C., Chen, X., and Bergeron, A.** (2007). FragAnchor: a large-scale predictor of glycosylphosphatidylinositol anchors in eukaryote protein sequences by qualitative scoring. Genomics, Proteomics & Bioinformatics, 5(2), 121–30.

**Poisson, R.** (1930). *Herpetomonas tortum* nsp., parasite intestinal des *Camptopus lateralis* (Germ.) (Hemiptères Coréidae, Alydaria) des environs de Banyuls. Rôle possible de cet insect comme agent transmetteur de phytoflagellose. Comptes Rendus Des Séances de La Société de Biologie, 103, 1061–1064.

**Porcel, B. M., Denoeud, F., Opperdoes, F., Noel, B., Madoui, M.-A., Hammarton, T. C., Field, M. C., Da Silva, C., Couloux, A., Poulain, J., Katinka, M., Jabbari, K., Aury, J.-M., Campbell, D. A., Cintron, R., Dickens, N. J., Docampo, R., Sturm, N. R., Koumandou, V. L., Fabre, S., Flegontov, P., Lukeš, J., Michaeli, S., Mottram, J. C., Szöőr, B., Zilberstein, D., Bringaud, F., Wincker, P., and Dollet, M.** (2014). The streamlined genome of *Phytomonas* spp. relative to human pathogenic kinetoplastids reveals a parasite tailored for plants. PLoS Genetics, 10(2), e1004007.

**Price, M. N., Dehal, P. S., and Arkin, A. P.** (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS ONE, 5.

**Redman, C. A., Schneider, P., Mehlert, A., and Ferguson, M. A.** (1995). The glycoinositol-phospholipids of *Phytomonas*. Biochemical Journal, 311, 495–503.

**Reichenow, E.** (1940). Ostafrikanische Beobachtungen an Trypanosomiden. Archiv Fur Protistenkunde, 94, 267–287.

**Risso, M. G., Garbarino, G. B., Mocetti, E., Campetella, O., Gonzalez Cappa, S. M., Buscaglia, C. A., and Leguizamon, M. S.** (2004). Differential expression of a virulence factor, the trans-sialidase, by the main *Trypanosoma cruzi* phylogenetic lineages. The Journal of Infectious Diseases, 189(12), 2250–9.

**Rodhain, J.** (1926). Phytoflagellé du latex d'une Asclépediacée congolaise. Annales de La Société Belge de Médecine Tropicale, 6, 271–274.

**Rodhain, J., and Bequaert, J.** (1911). Presence de *Leptomonus* dans le latex d'une euphorbe congolaise. Bulletin de La Société de Pathologie Exotique, 4, 299–303.

**Rooney, H. C. E., Van't Klooster, J. W., van der Hoorn, R. A. L., Joosten, M. H. A. J., Jones, J. D. G., and de Wit, P. J. G. M.** (2005). Cladosporium Avr2 inhibits tomato Rcr3 protease required for Cf-2-dependent disease resistance. Science, 308(5729), 1783–6.

**Ruiz, A.** (1958). Contribuiciòn al estudio del género *Phytomonas* Donovan e Costa Rica. 11. *Phytomonas elmassiani* (Migone, 1916) Wenyon, 1926. Revista de Biologia Tropical, 6, 263–272.

**Ruszczyk, A., Forlenza, M., Savelkoul, H. F. J., and Wiegertjes, G. F.** (2008). Molecular cloning and functional characterisation of a cathepsin L-like proteinase from the fish kinetoplastid parasite *Trypanosoma carassii*. Fish & Shellfish Immunology, 24(2), 205–14.

**Sambrook, J., and Russell, D. W.** (2006). The Inoue Method for Preparation and Transformation of Competent *E. Coli* : "Ultra-Competent" Cells. *Cold Spring Harbor Protocols*, *2006*(1), pdb.prot3944.

**Sanchez-Moreno, M., Fernandez-Becerra, C., Mascaro, C., Rosales, M. J., Dollet, M., and Osuna, A.** (1995). Isolation, in vitro culture, ultrastructure study, and characterization by lectin-agglutination tests of *Phytomonas* isolated from tomatoes (*Lycopersicon esculentum*) and cherimoyas (*Anona cherimolia*) in southeastern Spain. Parasitology Research, 81(7), 575–581.

**Sanchez-Moreno, M., Lasztity, D., Coppens, I., and Opperdoes, F. R.** (1992). Characterization of carbohydrate metabolism and demonstration of glycosomes in a *Phytomonas* sp. isolated from *Euphorbia characias*. Molecular and Biochemical Parasitology, 54(2), 185–199.

**Santoni, V., Molloy, M., and Rabilloud, T.** (2000). Membrane proteins and proteomics: un amour impossible? Electrophoresis, 21(6), 1054–70.

**Santos, A. L. S., d'Avila-Levy, C. M., Elias, C. G. R., Vermelho, A. B., and Branquinha, M. H.** (2007). *Phytomonas serpens*: immunological similarities with the human trypanosomatid pathogens. Microbes and Infection / Institut Pasteur, 9(8), 915–21.

**Santos, D'Avila-Levy, C. M., Dias, F. a, Ribeiro, R. O., Pereira, F. M., Elias, C. G. R., Souto-Padrón, T., Lopes, A. H. C. S., Alviano, C. S., Branquinha, M. H., and Soares, R. M. .** (2006). *Phytomonas serpens*: cysteine peptidase inhibitors interfere with growth, ultrastructure and host adhesion. International Journal for Parasitology, 36(1), 47–56.

**Santos, R. F. ., Romeiro, A., Attias, M., and De Souza, W.** (1994). Reisolation of a *Phytomonas* sp. from *Chamaesyce thymifolia*. Memorias Do Instituto Oswaldo Cruz, 89, 219.

**Schenkman, S., Diaz, C., & Nussenzweig, V.** (1991). Attachment of *Trypanosoma cruzi* trypomastigotes to receptors at restricted cell surface domains. Experimental Parasitology, 72(1), 76–86.

**Sergent, E.** (1921). Existence de *Leptomonas davidi* dans le latex d'Euphorbiacées d'Algérie (*E. peploides*). Archives de l'Institut Pasteur de l'Afrique Du Nord, 1, 58.

**Serrano, M. G., Nunes, L. R., Campaner, M., Buck, G. A., Camargo, E. P., and Teixeira, M. M.** (1999). Trypanosomatidae: *Phytomonas* detection in plants and phytophagous insects by PCR amplification of a genus-specific sequence of the spliced leader gene. Experimental Parasitology, 91(3), 268–79.

**Siddall, M. E., and Desser, S. S.** (1992). Alternative leech vectors for frog and turtle trypanosomes. The Journal of Parasitology, 78(3), 562–3.

**Siegel, T., and Gunasekera, K.** (2011). Gene expression in *Trypanosoma brucei*: lessons from high-throughput RNA sequencing. Trends in Parasitology, 27(10), 434–441.

**Simpson, A. G. B., Stevens, J. R., and Lukes, J.** (2006). The evolution and diversity of kinetoplastid flagellates. Trends in Parasitology, 22(4), 168–74.

**Simpson, J. T., and Durbin, R.** (2010). Efficient construction of an assembly string graph using the FM-index. Bioinformatics, 26(12), i367–73.

**Siqueira-Jaccoud, R. J.** (1955). Contribuição para o estudo da *Euphorbia braziliensis* Lam. Memórias Do Instituto Oswaldo Cruz, 54, 103–113.

**Slobbe, W. G., Parthasarathy, M. V., and Hesen, J. A. J.** (1978). Hartrot or fatal wilt of palms. II. Oil palms (*Elaeis guineensis*) and other palms. Principes, 22, 15–25.

**Späth, G. F., and Beverley, S. M.** (2001). A lipophosphoglycan-independent method for isolation of infective *Leishmania* metacyclic promastigotes by density gradient centrifugation. Experimental Parasitology, 99(2), 97–103.

**Späth, G. F., Epstein, L., Leader, B., Singer, S. M., Avila, H. A., Turco, S. J., and Beverley, S. M.** (2000). Lipophosphoglycan is a virulence factor distinct from related glycoconjugates in the protozoan parasite *Leishmania major*. Proceedings of the National Academy of Sciences of the United States of America, 97(16), 9258–63.

**Späth, G. F., Garraway, L. A., Turco, S. J., and Beverley, S. M.** (2003). The role(s) of lipophosphoglycan (LPG) in the establishment of *Leishmania major* infections in mammalian hosts. Proceedings of the National Academy of Sciences of the United States of America, 100(16), 9536–41.

**Stadler, C., Rexhepaj, E., Singan, V. R., Murphy, R. F., Pepperkok, R., Uhlén, M., Simpson, J. C., and Lundberg, E.** (2013). Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. Nature Methods, 10(4), 315–23.

**Stahel, G.** (1931). Zur Kenntis der Siebrohrenkrankheit (Phloemnekrose) des Kaffeebaumes in Surinam. I. Mikroskopische Untersuchungen und Infektionsversuche. Phytopathologische Zeitschrift, 4, 65–82.

**Stoco, P. H., Wagner, G., Talavera-Lopez, C., Gerber, A., Zaha, A., Thompson, C. E., …**

**Grisard, E. C. (**2014). Genome of the avirulent human-infective trypanosome--Trypanosoma rangeli. *PLoS Neglected Tropical Diseases*, *8*(9), e3176.

**Strong, R. P.** (1924). Investigations upon flagellate infections. American Journal of Tropical Medicine and Hygiene, 4, 345–372.

**Stuart, K., Brun, R., Croft, S., Fairlamb, A., Gürtler, R. E., McKerrow, J., Reed, S., and Tarleton, R.** (2008). Kinetoplastids: related protozoan pathogens, different diseases. The Journal of Clinical Investigation, 118(4), 1301–10.

**Sturm, N. R., Dollet, M., Lukes, J., and Campbell, D. A.** (2007). Rational sub-division of plant trypanosomes (*Phytomonas* spp.) based on minicircle conserved region analysis. Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases, 7(5), 570–6.

**Sutton, R. E., and Boothroyd, J. C.** (1986). Evidence for trans splicing in trypanosomes. Cell, 47(4), 527–35.

**Teixeira, M. M. G., Takata, C. S. A., Conchon, I., Campaner, M., and Camargo, E. P.** (1997). Ribosomal and kDNA markers distinguish two subgroups of *Herpetomonas* among old species and new trypanosomatids isolated from flies. The Journal of Parasitology, 83(1), 58.

**Teixeira, M. M., Campaner, M., and Camargo, E. P.** (1994). Detection of trypanosomatid *Phytomonas* parasitic in plants by polymerase chain reaction amplification of small subunit ribosomal DNA. Parasitology Research, 80(6), 512–6.

**Teixeira, S. M., Russell, D. G., Kirchhoff, L. V, and Donelson, J. E.** (1994). A differentially expressed gene family encoding “amastin,” a surface protein of *Trypanosoma cruzi* amastigotes. The Journal of Biological Chemistry, 269(32), 20509–16.

**Ten Have, A., Espino, J. J., Dekkers, E., Van Sluyter, S. C., Brito, N., Kay, J., González, C., and van Kan, J. A. L.** (2010). The *Botrytis cinerea* aspartic proteinase family. Fungal Genetics and Biology : FG & B, 47(1), 53–65.

**Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., and Borodovsky, M.** (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Research, 18(12), 1979–90.

**Tjalsma, H., Lambooy, L., Hermans, P. W., and Swinkels, D. W.** (2008). Shedding & shaving: Disclosure of proteomic expressions on a bacterial face. *PROTEOMICS*, *8*(7), 1415–1428.

**Tran, K. D., Rodriguez-Contreras, D., Vieira, D. P., Yates, P. A., David, L., Beatty, W., Elferich, J., and Landfear, S. M.** (2013). KHARON1 mediates flagellar targeting of a glucose transporter in *Leishmania mexicana* and is critical for viability of infectious intracellular amastigotes. The Journal of Biological Chemistry, 288(31), 22721–33.

**Trudgian, D. C., Ridlova, G., Fischer, R., Mackeen, M. M., Ternette, N., Acuto, O., Kessler, B. M., and Thomas, B.** (2011). Comparative evaluation of label-free SINQ normalized spectral index quantitation in the central proteomics facilities pipeline. Proteomics, 11(14), 2790–7.

**Trudgian, D. C., Thomas, B., McGowan, S. J., Kessler, B. M., Salek, M., and Acuto, O.** (2010). CPFP: a central proteomics facilities pipeline. Bioinformatics (Oxford, England), 26(8), 1131–2.

**Uttaro, A. D., and Opperdoes, F. R.** (1997a). Characterisation of the two malate dehydrogenases from *Phytomonas* sp. Purification of the glycosomal isoenzyme. Molecular and Biochemical Parasitology, 89(1), 51–9.

**Uttaro, A. D., and Opperdoes, F. R.** (1997b). Purification and characterisation of a novel iso-propanol dehydrogenase from *Phytomonas* sp. Molecular and Biochemical Parasitology, 85(2), 213–9.

**Vainstein, M. H., and Roitman, I.** (1986). Cultivation of *Phytomonas françai* associated with poor development of root system of cassava. Journal of Protozoology, (17), 511–513.

**Vainstein, M. H., Da Silva, J. B. T., De Lima, V. M. Q. G., Roitman, I., De Souza, W., Dollet, M., and Camargo, E. P.** (1987). Electrophoretic analysis of isoenzymes in the identification of Trypanosomatids of the genus *Phytomonas*. The Journal of Protozoology, 34(4), 442–444.

**Vickerman, K.** (1962). Observations on the life cycle of *Phytomonas elmassiani* (Mignone) in East Africa. The Journal of Protozoology, 9, 26–33.

**Visentini, A.** (1914). La flagellosi delle Euforbie in Italia. Rendiconti Della Reale Academia Dei Lincei, 23, 663–666.

**Votýpka, J., Klepetková, H., Jirků, M., Kment, P., and Lukeš, J.** (2012). Phylogenetic relationships of trypanosomatids parasitising true bugs (Insecta: Heteroptera) in sub-Saharan Africa. International Journal for Parasitology, 42(5), 489–500.

**Votýpka, J., Maslov, D. A., Yurchenko, V., Jirků, M., Kment, P., Lun, Z.-R., and Lukes, J.** (2010). Probing into the diversity of trypanosomatid flagellates parasitizing insect hosts in South-West China reveals both endemism and global dispersal. Molecular Phylogenetics and Evolution, 54(1), 243–53.

**Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., and Earl, A. M.** (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One, 9(11), e112963.

**Wheeler, R. J., Gluenz, E., and Gull, K.** (2011). The cell cycle of *Leishmania*: morphogenetic events and their implications for parasite biology. Molecular Microbiology, 79(3), 647–62.

**Wheeler, R. J., Gluenz, E., and Gull, K.** (2013). The limits on trypanosomatid morphological diversity. PloS One, 8(11), e79581.

**Wu, C. C., and Yates, J. R.** (2003). The application of mass spectrometry to membrane proteomics. Nature Biotechnology, 21(3), 262–7.

**Wu, Y., El Fakhry, Y., Sereno, D., Tamar, S., and Papadopoulou, B.** (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.

**Yao, C., Li, Y., Donelson, J. E., and Wilson, M. E.** (2010). Proteomic examination of *Leishmania chagasi* plasma membrane proteins: Contrast between avirulent and virulent (metacyclic) parasite forms. Proteomics. Clinical Applications, 4(1), 4–16.

**Zakai, H. A., Chance, M. L., and Bates, P. A.** (1998). In vitro stimulation of metacyclogenesis in *Leishmania braziliensis*, *L. donovani, L. major* and *L. mexicana*. Parasitology, 116 (Pt 4, 305–9.

**Zotta, G.** (1921). Un Leptomonas du type *L. davidi* Laf. chez des euphorbes de France. Comptes Rendus Des Seances de La Société de Biologie, 85, 226–228.

**Zotta, G.** (1924). La leptomonadiose spontanee chez *Cynanchum acutum*, Asclepiadee autotchone en Roumanie. Comptes Rendus Des Seances de La Societe de Biologie, 90, 141–143.