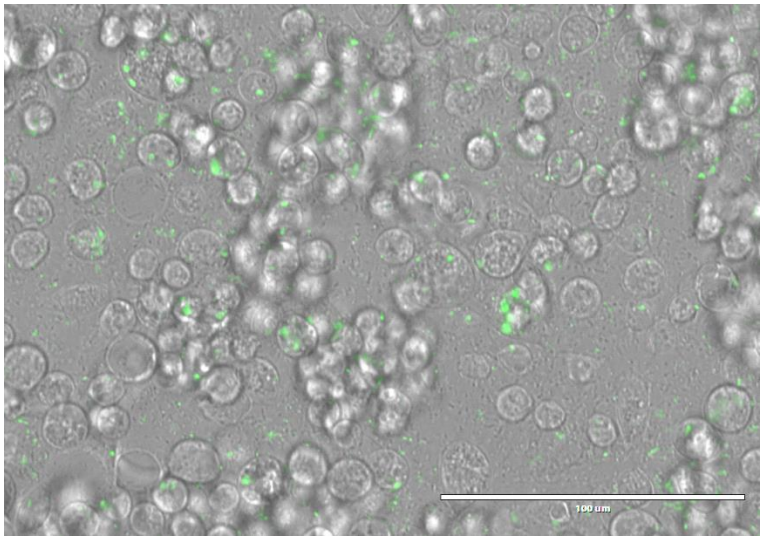
Functional Genomics of the Insect-Vector Symbiont, *Sodalis glossinidius*



Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Lauren Clare Gordon

November 2019

UNIVERSITY OF LIVERPOOL

Contents

Acknowledgements

Firstly, I would like to thank my wonderful family and best friend Gill for their love and support, without which I would not have completed this journey. You all kept me going and for that, I will be forever grateful.

I would also like to thank my friends Chris, Paul, Tom, Mary, Mike, Nomes, Poppy, Johnny, Stefan and Pilgrim for providing much-needed respites through the pastime of RPGs.

Thank you to my supervisor Prof. Alistair Darby for his guidance and expertise throughout this project. I would also like to acknowledge the Darby Group both present and past, for their support: Fran, Dave, Sam, Joe, Gwen and Alex. By extension I would also like to thank Dr. Ian Goodhead for his *Sodalis* wisdom and expertise. A particular thank you to Mark, of whom started his project at the same time and has been there for me throughout.

I would also like to thank my secondary supervisor Prof. Jay Hinton, of whom has always treated me like one of his own group. Within the Hinton Group I would particularly like to thank Rocio and Nico, who's expertise proved invaluable to my project.

Thank you to my tertiary supervisor Prof. Greg Hurst, and Pol and Eva within his research group, who's symbiont and cell line insights were much appreciated.

Stef and Blanca, your love and strength meant the world to me and I shall always treasure our coffee breaks.

I would also like to acknowledge the University of Liverpool and the BBSRC for providing me with this opportunity.

Sam, words will never do justice as a 'thank you'. You pulled me out of the hole I found myself in and I completed this thesis because of your love and unwavering belief in me. I look forward to the rest of our lives, together.

Abbreviations

Aedes albopictus larvae cells (C6/36)

Adenosine triphosphate-binding cassette (ABC)

Adenosine triphosphate (ATP)

Analysis of Variance (ANOVA)

Animal African trypanosomiasis (AAT)

Antimicrobial peptides (AMPs)

Bacterial artificial chromosome (BAC)

Base pair (bp)

Binary alignment map (BAM)

Basic Local Alignment Search Tool (BLAST®)

Cationic antimicrobial peptide (CAMP)

Card agglutination test for trypanosomiasis (CATT)

Centimetre (cm)

Citric acid cycle (CAC)

Clusters of Orthologous Group (COG)

Coding sequence (CDS)

Colony forming units (CFU)

Contour-clamped homogeneous electric field (CHEF)

Cytoplasmic incompatibility (CI)

Deionised water (dH_2O)

Dimethyl sulfoxide (DMSO)

Deoxyribonucleic acid (DNA)

Differentially expressed (DE)

Double-stranded deoxyribonucleic acid (dsDNA)

Drosophila Schneider 2 cells (S2)

Early log phase (ELP)

Ectoperitrophic space (ES)

Electron transport chain (ETC)

Foetal bovine serum (FBS)

Flux Balance Analysis (FBA)

Gram (g)

Green fluorescent protein (GFP)

Henrietta Lacks (HeLa)

Human African trypanosomiasis (HAT)

Hydroxymethyl pyrimidine pyrophosphate (HMP-PP)

Inducible nitric oxide synthase (iNOS)

Insertion sequence (IS)

Inside end (IE)

Kilobase (Kb)

Kyoto Encyclopaedia of Genes and Genomes (KEGG)

Kyoto Encyclopaedia of Genes and Genomes Orthology (KOs)

Late log phase (LLP)

Late stationary phase (LSP)

Likelihood ratios (LR)

Litre (L)

Long slender (LS)

Lysogeny Broth (LB)

Master mix (MM)

Megabase (Mb)

Microgram (µg)

Microlitre (µL)

Milligram (mg)

Millilitre (mL)

Mitsuhashi and Maramorosch Insect Medium (MMI)

'Modified' LB (mLB)

Multi locus sequence typing (MLST)

Multiplicity of infection (MOI)

M9 Minimal Salts Medium (M9)

M9 Minimal Salts Media without glucose ($M9^{g-}$)

Nanogram (ng)

Nicotinamide adenine dinucleotide hydrogen (NADH)

Nicotinamide adenine dinucleotide phosphate hydrogen (NADPH)

N-(3-oxohexanoyl) homoserine lactone (OHHL)

Open reading frame (ORF)

Optical density (OD)

Outside end (OE)

Pan-African Tsetse and Trypanosomiasis Eradication Campaign (PATTEC)

Pathogen-associated molecular patterns (PAMP)

Pentose phosphate pathway (PPP)

Peptidoglycan recognition proteins (PGRPs)

Peritrophic matrix (PM)

Phosphoenolpyruvate (PEP)

Phosphotransferase system (PTS)

Polymerase chain reaction (PCR)

Progressive control pathways (PCPs)

Proventriculus (PV)

Quality control (QC)

Quorum sensing (QS)

Reactive nitrogen intermediates (ROI)

Revolutions per minute (rpm)

Reactive oxygen species (ROS)

Ribonucleic acid (RNA)

Ribonucleic acid sequencing (RNASeq)

Ribosomal ribonucleic acid (rRNA)

Rickettsia-like organisms (RLO)

Salmonella pathogenicity island (SPI)

Sequence alignment map (SAM)

Sequential aerosol technique (SAT)

Short stumpy (ST)

Single nucleotide polymorphism (SNP)

Single-stranded deoxyribonucleic acid (ssDNA)

Sodalis symbiosis regions (SSR)

Sterile insect technique (SIT)

Thiamine phosphate synthase (TPS)

Thiazole phosphate carboxylate (THZ-P)

Transcripts per million (TPM)

Transposon (Tn)

Transposon-directed insertion site sequencing (TraDIS)

Tricarboxylic acid (TCA)

Type III secretion system (TTSS)

Ultraviolet (UV)

Variant surface glycoprotein (VSG)

Wild type (WT)

World Health Organisation (WHO)

Yersinia secretion apparatus (Ysa)

Abstract

Animal- (AAT) and human African trypanosomiasis (HAT) is endemic within sub-Saharan Africa and is caused by *Trypanosoma* spp. parasites vectored by biting tsetse flies. The facultative secondary symbiont, *Sodalis glossinidius*, has been controversially implemented in increased parasite establishment in tsetse. As the role of *S. glossinidius* in tsetse is not fully understood within the literature, the research presented here aimed to utilise a functional genomics approach to elucidate *S. glossinidius* functionality from a genetic context. Initial phenotypic-level *in vitro* media screening experiments confirmed *S. glossinidius* heterotrophy and revealed higher growth levels in the presence of glucose: *S. glossinidius* was unable to grow in a minimal salts medium (M9) devoid of a sufficient organic carbon source, and showed higher growth values in glucose-positive M9 variations compared to equivalent glucose-negative counterparts. This glucose utilisation was also observed with better growth between a complex medium rich in glucose (Mitsuhashi and Maramorosch Insect Medium) versus one with lower concentrations (Schneider's Insect Medium). Subsequent genotypic-level transposon-directed insertion site sequencing (TraDIS) library selection experiments supported *S. glossinidius* glucose utilisation with essential gene candidacy in glycolysis, gluconeogenesis and the pentose phosphate pathway. These results, in combination within essentiality in the citric acid cycle, a wide range of carbon source metabolism pathways, and virulence-associated genes (Omp porins, flagellar components and type III secretion system constituents), experimentally confirm the sequence-inferred literature consensus that *S. glossinidius* has retained a functional repertoire more aligned with free-living organisms. Many of the essential gene candidates were pseudogenes, which when considered with the literature evidence that *S. glossinidius* is actively maintaining a core pseudogene set across lineages, experimentally supports the theory that symbionts in early stages of genome degradation associated with the free-living to symbiont lifestyle switch preference pseudogene retention. The novel TraDIS library presented here provides the currently missing tool for subsequent targeted functionality *in vivo* experiments, aimed at fully understanding the *S. glossinidius* role in the tsetse system.

Chapter 1; Introduction

1.1. Vector Epidemiology

1.1.1. Human African Trypanosomiasis

Trypanosomiasis is caused by protozoan parasites in the genus *Trypanosoma*; *T. brucei* sub-species *gambiense* is the causative agent for the chronic form of human African trypanosomiasis (HAT) [or "sleeping sickness"] (Simarro *et al*., 2010). HAT can be anthroponotic and is endemic in the western and central parts of Africa (Simarro *et al*., 2010). The zoonotic *T. b. rhodesiense* is responsible for the acute form of HAT in eastern and southern Africa (Simarro *et al*., 2010). HAT is distributed across 36 sub-Saharan countries and risks an estimated 69.3 million people; the rurality of a proportion of individuals affected results in undiagnosed cases, meaning that the number of officially-reported cases offer a skewed prediction of the [estimated] actual disease burden (Simarro *et al*., 2010; Simarro *et al*., 2012; WHO, 2019a). Control and surveillance strategies have resulted in massive reductions in both the number of reported and estimated cases of HAT, dropping from approximately 40,000 reported and 300,000 estimated cases in 1998 to 1,446 reported and 10,000 estimated cases in 2017, to 977 reported cases in 2018 (WHO, 1998; WHO, 2019a; WHO 2019b). Despite this HAT remains endemic, reasons ranging from emergence of trypanosome resistance to drugs, patients refusing treatment, to the inability to create a vaccine (Brun *et al.,* 2010; Jamonneau *et al*., 2012). It has also been shown that trypanosomes utilise a variant surface glycoprotein (VSG) switching mechanism to avoid the mammalian immune system (reviewed in Horn, 2014).

HAT is diagnosed using a serological test known as the card agglutination test for trypanosomiasis (CATT) [*T. b. gambiense*], microscopic evaluation of blood or lymph node(s) aspirates for parasite load and staging criteria such as cerebral spinal fluid examination from lumbar puncture and white blood cell counts (Magnus, *et al*. 1978; WHO, 1998; Kennedy, 2006). Clinical manifestations of HAT occur in two stages; the early [haemolymphatic] and late [encephalitic] stage

(Kennedy, 2004; Kennedy, 2013). The early stage presents with flu-like symptoms as the trypanosomes spread in the bloodstream, endocrine and systemic organs, and the eyes; such as headache, fatigue, fever and malaise, in addition to weight loss and arthralgia (Kennedy, 2004; Kennedy, 2013). Enlargement of organs such as the spleen and/or liver, in addition to lymphadenopathy, can follow the initial flu-like symptoms, in addition to cardiac [myocarditis and congestive cardiac failure], ophthalmological [iritis and keratitis], endocrine [alopecia and impotence] and fertility [sterility and stillbirths] problems (Atouguia and Kennedy, 2000; Kennedy, 2005; Kennedy, 2013). The late stage occurs as the trypanosomes cross the blood-brain barrier and affect the central nervous system, pathologically identified when meningoencephalitis occurs with immune cells infiltrating the brain tissue (Adams *et al.*, 1986; Atouguia and Kennedy, 2000; Kennedy, 2013). Other symptoms of the late stage involve motor and sensory system problems, and mental and sleep disturbances, from which the disease gained its name (Kennedy, 2005; Kennedy, 2013). The sleep disturbances involve disruption of the sleep/wake cycle, nocturnal insomnia, uncontrollable onset of sleep and daytime somnolence; 74.4% of 2541 encephalitic-stage patients had sleeping disorder (Buguet *et al*., 2005; Blum *et al*., 2006). If left untreated, HAT is mostly always fatal; a study following a small collection of patients [11] that continued to refuse treatment found that they became asymptomatic (Jamonneau *et al*., 2012). The treatments are expensive and themselves offer relatively serious side-affects; early stage drugs [suramin and/or pentamidine] can result in hyper- or hypoglycaemia, anaphylactic shock, renal failure and bone marrow toxicity and late stage drugs [melarsoprol or eflornithine plus nifurtimox] are painful and toxic, with 10% of patients experiencing reactive encephalopathy (Kennedy 2008; Brun *et al*., 2010; Kennedy, 2013). A new drug, fexinidazole, has recently been approved under the most recent WHO guidelines, offering new avenues for further disease management (Chappuis, 2018; WHO, 2019b).

1.1.2. Animal African Trypanosomiasis

Animal African trypanosomiasis (AAT) [or "nagana"] is caused by *T. brucei*, *T. congolense* and *T. vivax* (WHO, 2019c). AAT is most known for affecting cattle but can also infect small ruminants and pigs (Ebhodaghe *et al*., 2018). It has been estimated that at least 48 million cattle are at risk of AAT, jeopardising an estimated $1321 million and $830 million worth of meat and milk produce, respectively; a major socio-economical concern for rural populations relying on animal husbandry (Kristjanson *et al*., 1998; von Wissmann *et al*., 2011). AAT affecting cattle also has a knock-on implication within agriculture, as disease burden means that there are less animals for traction and less availability of manure for arable crop production (Kristjanson *et al*., 1998; FITCA, 2005). In impoverished communities where a mix of livestock and crop-livestock farming are the main source(s) of income, AAT is devastating; 87% of cattle-owning communities surveyed reported livestock as their primary income source amongst a continuous AAT burden (Holt *et al*., 2016). Also adding to the cost implications of AAT is cattle disease treatments, with an estimated $35 million a year to both government and livestock producers (Kristjanson *et al*., 1998). In comparison to HAT, AAT is relatively under-considered but progressive control pathways (PCPs) for AAT are more recently being driven; PCPs provide frameworks for understanding, controlling and eliminating [human and] animal diseases (Diall *et al*., 2017).

The clinical symptoms of AAT include wasting within the animal, where weight is lost, the skin becomes tighter and a lack of energy is apparent (LSTM, 2019). Discharge may also be observed from the eyes, including photophobia and continual weeping (LSTM, 2019). Disease signs appear 11 – 21 days post-infective bite with repeating fever and without treatment, death is common within one to three months (LSTM, 2019).

1.1.3. Parasite Lifecycle in Insect Vector

The trypanosomatids causing the African trypanosomiases are spread by the bite of blood-feeding tsetse flies [Dipteran: *Glossina* spp.] (WHO, 2019a). All 22

species within the *fusca, palpalis* and *morsitans* subgenera of *Glossina* are vectors with varying success; the *palpalis* group [e.g. *Glossina palpalis gambiensis*] are known to transmit *T.b. gambiense*, whereas the *morsitans* group [e.g. *Glossina morsitans morsitans*] are vectors to *T.b. rhodesiense* (Hoare, 1972; Aksoy *et al*., 2014). This geographical separation is seen from habitat preference between groups, where riverine tsetse are associated with vegetation lining water bodies in central and western areas, compared to savannah tsetse that establish in thickets in eastern Africa (Rogers, 1977; Hargrove, 1981; CDC, 2019). Tsetse physiology is somewhat unique amongst insects as they are viviparous and feed their intrauterine offspring through milk glands. The offspring are birthed as 3rd instar larvae and pupate immediately after deposition, and this typically only occurs 8 – 10 times per female (Wang *et al*., 2013).

The majority of *Trypanosoma* spp. are transmitted in the tsetse saliva as mammalian-infective metacyclic forms, but *Trypanosoma* spp. undergo multiple metamorphic changes throughout their lifecycle, enabling different characteristics in different host environments (Van den Abbeele *et al*., 1999; Dyer *et al*., 2013). This introduces bottlenecks affecting *Trypanosoma* spp. infection and establishment in tsetse; the long and proliferative slender (LS) forms ingested by the tsetse within the mammalian blood meal will not survive in the mid-gut, but the short stumpy (ST) forms will and further transform into the proliferative procyclic form (Dyer *et al*., 2013). These procyclic forms cross the peritrophic matrix (PM) [a physical barrier to invasive components including pathogens, composed of glycoproteins and chitin] and proliferate in the ectoperitrophic space (ES) [in between the PM and the mid-gut], three – six days post-infection (Gibson and Bailey, 2003; Hegedus *et al.*, 2009; Dyer *et al*., 2013). After six days, the procyclic forms migrate to the proventriculus (PV) [a muscular organ] and transform into long trypomastigotes, which then divide into both long and short epimastigotes (Van den Abbeele *et al*., 1999; Sharma *et al*., 2008; Dyer *et al*., 2013). These forms are then thought to re-enter the lumen of the gut and migrate to the salivary glands, forming into metacyclic forms for mammalian infection (Van den Abbeele *et*

al., 1999; Dyer *et al*., 2013). The described lifecycle(s) of African trypanosomiasis-associated *Trypanosoma* spp. are summarised in Figure 1.1.

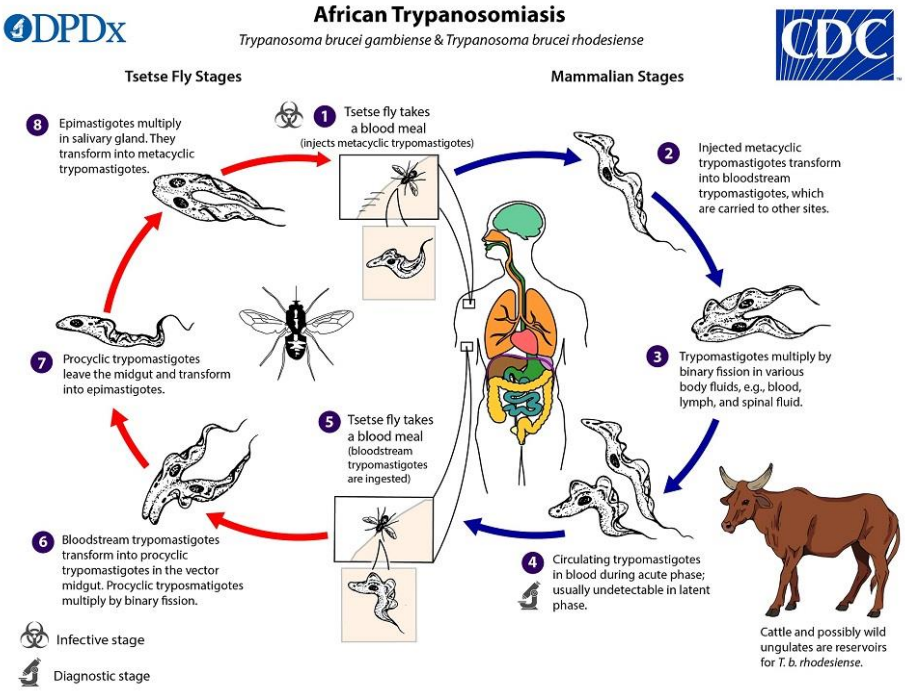


Figure 1.1. Graphical depiction of *Trypanosoma brucei gambiense* and *T. b. rhodesiense* lifecycles in both insect vector and mammalian host. Taken from Division of Parasitic Diseases and Malaria (DPDM), Centers for Disease Control and Prevention (CDC).

Despite tsetse being the sole vector of [African] *Trypanosoma* spp., percentage infection within wild populations is low; < 1% infection rate has been reported in *G. fuscipes fuscipes* and more recently, 16.8% in *G. morsitans submorsitans* and 2.2% and 6.9% in Guinea and Mali *G. palpalis gambiensis*, respectively (Okoth and Kapaata, 1986; Ouedraogo *et al.*, 2018). Alongside the bottlenecking experienced by *Trypanosome* spp. within-host, this low infection rate is also linked to the tsetse's immune system, which involves a combination of physical and innate systems such as epithelia providing a barrier, haemocytes partaking in phagocytosis and the production of peptidoglycan recognition proteins (PGRPs) and AMPs via signalling pathways such as the Toll, Imd and JAK/STAT (reviewed in Lemaitre and Hoffman, 2007; reviewed in Matetovici *et al*., 2019). It

has been shown that reactive oxygen species (ROS) induce trypanosome cell death in tsetse; 1.5 hours post-ROS production lead to an excess of Ca^{2+} in the mitochondrial nucleus and after 5 hours, over 90% of *T. b. brucei* cells showed extensive nuclear DNA fragmentation (Ridgley *et al*., 1999). Evidence supporting this was presented by MacLeod *et al.* (2007), where *T. b. brucei* infection in *G. m. morsitans* was increased [compared to a control] by adding antioxidants such as glutathione and *N*-acetylcysteine. Glutathione constitutes [plus spermidine] trypanothione, the main trypanosome reducing agent (Fairlamb *et al.,* 1985). Another factor affecting tsetse refractoriness to trypanosome infection includes the age of the fly; 38.1% of teneral [unfed] *G. m. centralis* developed midgut *T. congolense* infections, versus 8.1% in 30-day old non-teneral *G. m. centralis* (Moloo and Shaw, 1989). Welburn and Maudlin (1992) also reported the same phenomenon, whereby teneral laboratory *G. m. morsitans* were significantly more likely to develop mid-gut *T. congolense* establishment, than non-teneral flies. The sex of the fly also appears to influence trypanosome infection; it was found that *Trypanosome* sp. were less likely to mature in female *G. m. morsitans* than male, potentially as a result of X-linked gene products killing or preventing parasites from migrating and thus establishing (Milligan *et al*., 1995).

1.1.4. Vector Control Strategies

As previously mentioned, the reported cases of African trypanosomiases have dropped drastically over the last two decades, one of the main reasons being parasite-vector control; the implication of the tsetse control programme utilising pour-on insecticide for cattle and insecticide impregnated targets resulted in a 90% reduction in *G. tachinoides* and *G. m. submorsitans* populations and AAT reports to below 5%, in Burkina Faso (Bauer *et al.*, 1999). The success of these types of programmes relies heavily on local community upkeep and it was found that 37% of 261 household surveys reported individuals would contribute labour for trap and target maintenance, but it was predicted that only 56% actually did (Kamuanga *et al*., 2001). Other techniques including clearing tsetse habitat [bush clearing]

resulted in suppression of tsetse population but failed to eradicate them completely; clearings ranged from 440 yards to several miles, resulting in anywhere between 60 – 90% fly reduction, depending on the time of year (Morris, 1946). The sequential aerosol technique (SAT) as part of the Pan-African Tsetse and Trypanosomiasis Eradication Campaign (PATTEC) utilised aerial and ground spraying, in conjunction with the aforementioned insecticide-treated targets and cattle, and has seen a 98% reduction in the targeted riverine tsetse species [*G. tachinoides* and *G. p. gambiensis*] in the upper west areas of Ghana (Adam *et al*., 2013). The sterile insect technique (SIT) utilises ionizing radiation or chemicals to induce infertility in either sex and has also been a vital tool in the control of tsetse; the release of > 8.5 million sterile male *G. austeni* over 3 years on the Island of Unguja [Zanzibar] resulted in the complete collapse of the indigenous fly population (Knipling, 1959; Vreysen *et al*., 2000).

These approaches have been vital in suppressing the threat of HAT and AAT across the endemic countries of sub-Saharan Africa, but the disease, causative parasite(s) and insect vector still remain as both mortality and economical risks. New opportunities for disease control have been proposed manipulating maternally inherited contributors to host infection burden (covered below).

1.1.5. Symbiont Involvement in Vector-Parasite Interactions

Early studies involving tsetse refractoriness to trypanosomes noted maternal inheritance to infection susceptibility, whereby later studies implicated the roles of symbiotic organisms; Rickettsia-like organisms (RLO) were found in the ovaries, and female and male mid-gut in 85%, 100% and 95% of susceptible *G. m. morsitans*, compared to 15%, 10% and 25% in the refractory line, respectively (Maudlin, 1982; Maudlin and Ellis, 1985). Reasons behind why and how the RLO were related to this reduction in refractoriness was unknown, but an early study proposed an RLO mechanism involving D(+)glucosamine where inhibition of tsetse mid-gut lectins by *N*-acetylglucosamine produced via the action, most likely on the PM, of a bacterially-produced exochitinase (Welburn *et al.*, 1993). These tsetse

lectins were linked in early studies to agglutination and lysis of *T. congolense* and *T. b. brucei* (Stiles *et al*., 1990; Abubakar *et al.*, 1995). Abubakar *et al.* (1995) also reported inhibition of the *G. morsitans* mid-gut agglutinin action against *T. brucei* in the presence of D-glucosamine, which supports the preceding early work of Maudlin and Welburn (1987) where laboratory tsetse given infective feeds of *T. congolense* and *T. b. rhodesiense* and fed D(+)glucosamine showed average 94% and 85% mid-gut infections, respectively. However, a more recent study has shown that *N*-acetylglucosamine is an ROS scavenger and the infection increase seen when feeding tsetse *N*-acetylglucosamine could be due to the decrease in ROS causing trypanosome cell death, rather than inhibiting tsetse mid-gut lectins (Ridgley *et al*., 1999; Xing *et al.,* 2005).

The RLO were re-classified in 1999 as *Sodalis glossinidius* [strain $M1^T$] and are secondary symbionts to tsetse (Dale and Maudlin, 1999). Tsetse house two other symbionts, including the obligate primary symbiont *Wigglesworthia glossinidia* and secondary *Wolbachia pipientis* (Wigglesworth, 1929; O'Neill *et al.,* 1993; Aksoy, 1995a). It was assumed for many years that tsetse only play host to these three symbionts, but more recently it has been found by 16S rRNA gene amplicon sequencing that *G. f. fuscipes* and *G. tachinoides* harbour *Spiroplasma* (Doudoumis *et al*., 2017). Multi locus sequence typing (MLST) identified two strains of *Spiroplasma*, the highest densities being located in larval guts versus teneral or 15-day old adults, in testes versus ovaries in adults, and in live versus prematurely deceased adults (Doudoumis *et al*., 2017). Housing specific locations within the fly (Figure 1.2), tsetse symbionts are maternally inherited due to the host's viviparous nature, through the reproductive routes. *Wigglesworthia* inhabit bacteriocytes [differentiated epithelial cells forming the bacteriome organ] and are transmitted through the milk glands, *Sodalis* intra- and intercellularly inhabit the gut lumen, fat body, salivary glands, spermathecae and uterus and are transmitted through [and inhabit] the milk glands, and *Wolbachia* both inhabit and transmit via the ovaries (Wigglesworth, 1929; Reinhardt *et al*., 1972; Denlinger and Ma, 1975; Southwood *et al*., 1975; Shaw and Moloo, 1991; Aksoy *et al.*, 1997; Attardo *et al*., 2008; Balmand *et al*., 2013). More recently, it has been shown that *Wigglesworthia* exist

as two distinct populations either localised within the bacteriome or 'free-living' within the milk glands, the latter being responsible for transmission to the intrauterine larva (Pais *et al.*, 2008). Without their symbionts, specifically *Wigglesworthia* for tsetse, insect hosts become infertile and have been shown to die, following bacteriome destruction (Hill *et al*., 1973; Nogge, 1976; Schlein, 1976).

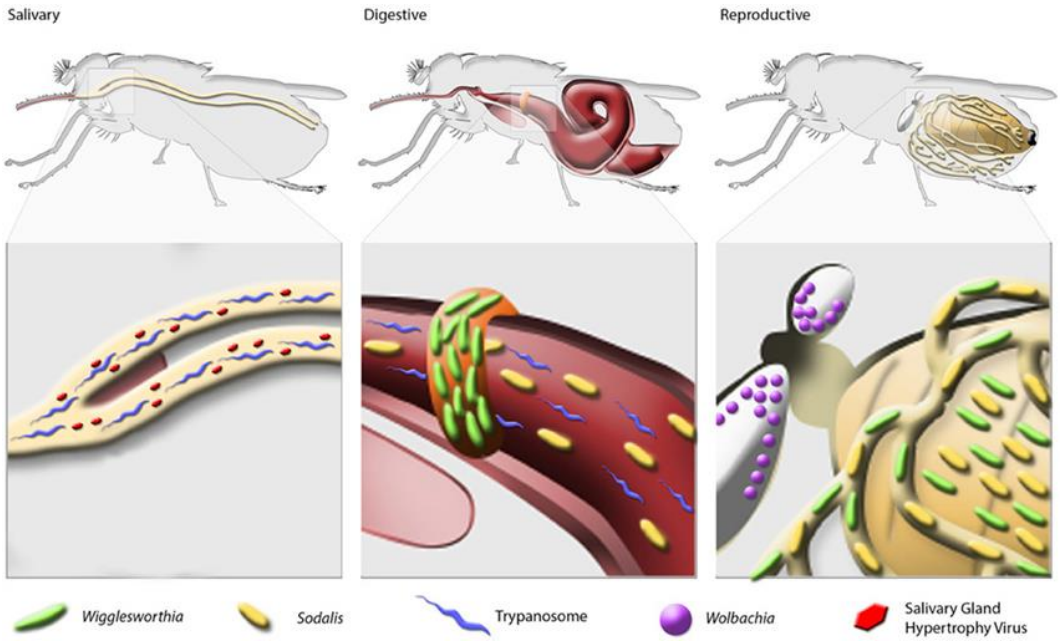


Figure 1.2. Graphical depiction of organism location within tsetse. *Wigglesworthia* (green) shown within the bacteriome. *Sodalis* (yellow) shown within gut lumen and reproductive tracts. *Wolbachia* (purple) shown within ovaries. Salivary Gland Hypertrophy Virus (SGHV) (red) shown within salivary glands. Trypanosome (blue) shown within salivary glands and digestive tract. **Taken from Wang *et al.* (2013)**

Primary symbionts are known to have long ancestral histories with their hosts, resulting in reduced genomes [*Wigglesworthia* genome at 0.70 Mb] that restrict survival to purely within-host and concordant phylogenies that strongly support ancient evolutionary host-symbiont cospeciation (Moran *et al.*, 1993; Aksoy *et al*., 1995b; Chen *et al*., 1999; Akman *et al*., 2002). This ability to successfully maintain long-term colonisation is fascinating when considering the aforementioned insect immune system. It has been found that *S. glossinidius* possesses an as-of-yet undefined immune downregulation mechanism in tsetse,

offering one explanation for this symbiotic capability of immune bypassing; [*Sodalis*-cleared] tsetse RNASeq expression levels of immune response genes were low when [re-]exposed to *S. glossinidius*, especially when compared to *Escherichia coli* levels (Trappeniers *et al*., 2019). Further to this, inactivation of the *phoP* response regulator in *S. glossinidius* has been shown to produce an AMP-sensitive phenotype (Pontes *et al.,* 2011). The PhoP-PhoQ two-component system regulates virulence determinants and senses environmental pH, $Mg2^{+}$ and $Ca2^{+}$ (Groisman, 2001; Clayton *et al.*, 2017). *S. glossinidius* quorum sensing (QS) has been found to prime its oxidative stress response in relation to cell density, the results of which could be extended to explain survival within the internal conditions of the host; N-(3-oxohexanoyl) homoserine lactone (OHHL) [a QS signalling molecule] was found to increase the expression of genes producing proteins implicated in detoxifying ROS (Pontes *et al*., 2008). Specifically relating to *W. glossinidia*, one could infer an ability to avoid attack from the host immune system based on research utilising antibiotic clearance; particular antibiotics were ineffective at clearing *W. glossinidia* from the bacteriome, which could potentially be extended to deeming the organ a physical barrier against host immune factors (Pais *et al.*, 2008). One long-established evolutionary reason for a lack of host attack on symbionts is that primary symbionts offer nutritional components, such as thiamine, pantothenic acid, pyridoxine, folic acid and biotin [B vitamins], that are lacking in blood-feeding insect host diets, which was confirmed by the genome sequencing of *Wigglesworthia* finding the retention of gene sets involving the biosynthesis of the cofactors (Nogge, 1981; Akman *et al*., 2002). In addition to supplying its tsetse host with B vitamins, *Wigglesworthia* have also been implicated in tsetse digestion efficiency, where *G. m. morsitans* cleared of *Wigglesworthia* did not digest their blood meal as well as WT *G. m. morsitans* [using haemoglobin levels as proxy] (Pais *et al.*, 2008). It was also found that non-teneral adult *G. m. morsitans* cleared of *Wigglesworthia* showed higher *T. b. rhodesiense* infection rates than WT non-teneral adult *G. m. morsitans*, suggesting a protective *Wigglesworthia* phenotype (Pais *et al.*, 2008). Interestingly, the same significant difference in *T. b. rhodesiense* infection rates was not observed in cleared teneral flies (Pais *et al.*, 2008). Pais *et al.*

(2008) also found that *G. m. morsitans* maintained at 3°C above normal colony temperature showed higher percentages of mortality when cleared of *Wigglesworthia* [versus not], suggesting a *Wigglesworthia* influence on host heat protection.

The secondary symbiont *W. pipientis* is known as the 'parasitic symbiont' and although beneficial mechanisms, such as *Wolbachia*-mediated apoptosis prevention, have been presented in other insects, no apparent benefit has been documented for tsetse (Pannebakker *et al*., 2007). For decades tsetse reproducibility was documented to be influenced by a trait involving maternal inheritability, with *Wolbachia* being theorised as the most likely candidate. *Wolbachia* has been experimentally confirmed as the agent using aposymbiotic *G. m. morsitans*, where it was found that mating *Wolbachia*-infected males with *Wolbachia*-free females resulted in strong cytoplasmic incompatibility (CI), during embryogenesis (Alam *et al*., 2011). CI in this context refers to the phenomenon where if the mating female tsetse is not infected with *Wolbachia*, the embryo does not survive (Schneider *et al*., 2013). As mentioned previously, novel attempts to control the HAT and AAT vector are underway; given a large enough release of *Wolbachia*-infected flies, Alam *et al*. (2011) reported that exploitation of the CI mechanism to drive a refractory tsetse phenotype into natural fly populations could be successful.

1.2. *Sodalis glossinidius*

1.2.1. Overview

S. glossinidius are known as a 'facultative' secondary symbiont as their infection rates in wild tsetse populations fluctuate. One study found that 4.6% [*n* = 196] of Kenyan *G. austeni* samples were infected, in comparison to 70.8% [*n* = 24] of south African samples (Wamwiri *et al*., 2013). Also noted was the link between *Sodalis* presence and trypanosome infection, where a higher proportion of trypanosome infections were found in *Sodalis*-positive *G. pallidipes*, than *Sodalis*-negative (Wamwiri *et al*., 2013). The idea that *Sodalis* interferes with tsetse

refractoriness to trypanosomes has been documented (as discussed in 1.1.5.), but the exact reason(s), mechanism(s) or gene(s) behind this remains elusive. The theory is also controversial within the literature, as some studies have not been able to experimentally confirm the link. For example, no significant difference in *T. b. brucei* establishment in the mid-gut or development into metacyclic forms in the salivary glands was found between *G. m. morsitans* cleared of *S. glossinidius* versus *S. glossinidius*-positive *G. m. morsitans* (Trappeniers *et al*., 2019). Despite this, as *S. glossinidius* has been shown to also transmit horizontally, utilising *S. glossinidius* for vector control via paratransgenesis (referring to genetically manipulating the symbiont to express foreign gene products, such as anti-trypanosomal agents) could prove pivotal in the fight against HAT and AAT (Aksoy *et al*., 2008; De Vooght *et al*., 2015). For example, a recombinant *Sodalis* expressing significant amounts of an anti-trypanosomal nanobody, Nb_An46, was shown to colonise midgut tissues [a location vital in trypanosome vector establishment] of *G. m. morsitans* transiently cleared of WT *Sodalis* (De Vooght *et al*., 2014).

Sodalis is not unique to tsetse; *Candidatus S. melophagi* symbiont was isolated from *Melophagus ovinus* [sheep ked] (Diptera: Hippoboscidae), *Candidatus S. pierantonius* str. SOPE from *Sitophilus oryzae* [rice weevil] (Coleoptera: Curculionidae), and multiple other *Sodalis*-allied symbionts from aphids, beetles, stinkbugs and ants (Sameshima *et al*., 1999; Burke *et al*., 2009; Grünwald *et al*., 2010; Kaiwa *et al*., 2010; Chrudimský *et al*., 2012; Oakeson *et al*., 2014). As can be discerned from this, it would appear that *Sodalis* is restricted to a symbiotic lifestyle. However, in 2012 a new species was discovered within the *Sodalis* genus; "strain HS", later re-named *S. praecaptivus*, was isolated from a patient's arm wound (Clayton *et al*., 2012; Chari *et al*., 2015). The fact that the wound was environmental in origin [branch impalement], combined with how this species has not been previously isolated from an arthropod host, strongly suggests a free-living phenotype for this bacterium. As sequence alignments between *S. glossinidius* and *S. praecaptivus* revealed a high level of gene conservation, it could be theorised that *S. praecaptivus* acts as a free-living genetic model for the *S. glossinidius* genome prior to its genome degradation associated with becoming a symbiont

(Clayton *et al*., 2012; Oakeson *et al*., 2014). The fact that *S. praecaptivus* is potentially pathogenic, in combination with the *S. glossinidius*-retention of virulence-associated gene sets [detailed in 1.2.2.1.], could suggest that *S. glossinidius* had virulent capabilities before its lifestyle switch to symbiotic. Due to the relatively recent discovery of *S. praecaptivus*, there is still a lot of research that needs to be done to fully establish the pheno- and genotypic story. Of-note research that has been done so far includes Enomoto *et al.* (2017)'s discovery that QS is important in *S. praecaptivus* establishment in an insect host; *S. praecaptivus* mutants deficient in QS displayed a phenotype that was insect-killing when microinjected into weevil hosts, showing that *S. praecaptivus* utilise QS systems to maintain a prolonged presence *in vivo*. It was theorised from this that bacteria initially utilise virulence factors to invade an insect host and then use QS for the repression of toxin expression, for long-term symbioses (Enomoto *et al.*, 2017). Further to this, it has been shown that like *S. glossinidius*, *S. praecaptivus* possesses a PhoP-PhoQ system required for AMP-resistance *in vivo* (Clayton *et al.*, 2017). It was detailed when first discovered that incubating *S. praecaptivus* on MacConkey plates at 35°C [and 5% CO_2] produced "wet, mucoid" colonies (Clayton *et al*., 2012). However as shown in Figure 1.3, *S. praecaptivus* displays a phenotypic change with temperature contrary to this assessment. When incubated [on LB agar] at 37°C the colonies are distinct and circular, compared to when being incubated at 25°C where the bacterium displays a mucoidy phenotype (Gordon L., unpublished data).

Figure 1.3. Photographs of *Sodalis praecaptivus* colonies after being incubated on LB plates at either 37°C or 25°C. The left-hand plate was incubated at 37°C and shows distinct circular colonies. The right-hand plate was incubated at 25°C and shows mucoidy colonies.

1.2.2. Genetics

1.2.2.1. Annotation History

The first construction of the *S. glossinidius* genome was reported in 2001, which used contour-clamped homogeneous electric field (CHEF) gel electrophoresis analysis to estimate the genome to be [average] 2.07 Mb in size (Akman *et al*., 2001). It was also reported that *S. glossinidius* had an A+T content of 45%, a 134 Kb extrachromosomal plasmid, and by use of hybridising DNA to an *E. coli* macroarray, a predicted 1,800 gene orthologues (Akman *et al*., 2001). In 2005, a *S. glossinidius* bacterial artificial chromosome (BAC) library was created and two distinct symbiosis regions (SSR) encoding type-III secretion systems (TTSSs) characterised; SSR-1 and SSR-2 (Dale *et al*., 2005). TTSSs are used by pathogens to translocate [inject] virulence proteins into the target cell cytosol, resulting in the interference of cellular process such as host cell signal transduction (reviewed in Hueck, 1998). TTSSs have also been linked to the invasion of cells; using a murine model showed *Salmonella* Typhimurium invasion and disruption of small intestine M cells was reliant on the TTSS products [such as *invC*] encoded for by *Salmonella* pathogenicity island 1 (SPI-1) (Jones *et al*., 1994; Penheiter *et al*., 1997). This invasion strategy has also been observed in *S. glossinidius*; using the *Aedes albopictus* cell line C6/36, miniTn5 *S. glossinidius* clones lacking the ability to attach and/or invade were selected (Dale *et al*., 2001). The DNA of a single clone that was deemed able to attach but unable to invade the cell was cloned and sequenced, finding that the miniTn5 had integrated into a region with a sequence similarity to the TTSS product *invC* from *S. enterica* (Dale *et al*., 2001). In relation to gene content, SSR-1 was reported to be most similar to the TTSS *Yersinia* secretion apparatus (Ysa) from *Yersinia enterolitica*; Ysa is closely related to SPI-1 and the *Shigella* TTSS, Mxi-Spa (Foultier *et al*., 2002; Dale *et al*., 2005). SSR-1 possesses the genetic components required for a fully functional TTSS apparatus and proteins facilitating entry into host cells; YspB and YspC [named based on genetically similar Ysa counterparts] effector proteins are homologous to *S. enterica* Sip proteins and have been shown to be involved with host cell invasion (Kaniga *et al*., 1995; Dale *et al*., 2005). SSR-2 was reported to be most similar to SPI-1 and the gene inventory retains syringe

components, but does not contain functional homologs for the needle substructure nor effector protein genes (Dale *et al*., 2005). SSR-2 was implicated in *S. glossinidius* ability to persist intracellularly by expression assays and knockout mutants of TTSS proteins; expression of *invA* [located on SSR-2] was only seen in substantial levels at a time point [48 hours] associated after invasion and SSR-2 *orgA* [homolog] mutants showed impaired replication in C6/36 compared to WT (Dale *et al*., 2001; Dale *et al*., 2005).

The complete genome sequence of *S. glossinidius* was published in 2006, reporting a genome size of 4.17 Mb, 2,432 protein coding domain sequences (CDSs) [a coding capacity of 51%], a G+C content of 54.7% and 972 pseudogenes (Toh *et al*., 2006). Pseudogenes are described as genes that were originally functional but have been fragmented and silenced due to frameshift, nonsense or missense mutations (reviewed in Goodhead and Darby, 2015). Toh *et al*. (2006) also reported that 60% of the *S. glossinidius* CDSs were homologous to other known proteins, 20% had unknown functions but were conversed between other bacteria, 11% were related to phage and 9% were unassigned [based on homology to public database entries]. Further to the previously characterised SSRs, a third region was identified; SSR-3 was reported to be the most similar to SPI-2 [and has retained all of the SPI-2-associated virulence genes] and the *Y. pestis* island (Toh *et al*., 2006). SSR-3 has a gene inventory containing various secretion system components, including apparatus, effectors, chaperons and an operon for the two-component regulator, SseAB (Toh *et al*., 2006). Three of the extrachromosomal plasmids previously characterised were noted, including plasmid-like elements pSG1, pSG2 and pSG4 [all with putative RepA protein], and bacterio-phage like element pSG3 [no RepA protein homolog and 56% ORF homology to bacteriophage] (Darby *et al*., 2005; Toh et al., 2006). A heavy proportion (38.69%) of pSG1 is pseudogenised. A phage was also reported: ϕSG1. The aforementioned number of pseudogenes [972] is unusually high for free-living bacteria; *S. enterica* serovar Typhi CT18 has 204 and *Photorhabdus luminescens* has 157 (Parkhill *et al*., 2001; Duchaud *et al*., 2003). Large proportions of the *S. glossinidius* pseudogenes are involved with defence mechanisms and the transport and metabolism of carbohydrate and inorganic ion

[using Clusters of Orthologous Group (COG) categorisations of putative functions based on homologues] (Toh *et al.*, 2006).

In 2010, the *S. glossinidius* genome was re-annotated, involving a re-characterisation of the number of pseudogenes, re-analysis of insertion sequence (IS) types and prophage elements, and a more detailed re-construction of the metabolic profile (Belda *et al*., 2010). The re-characterisation of the originally reported 972 pseudogenes found a new total of 1501 [filtered from 1724 potential] (Toh *et al.*, 2006; Belda *et al*., 2010). The re-annotated 3932 CDSs were assigned functionality using the functional classification scheme [Sanger Institute] and it was found that the highest numbers were categorised into surface proteins (770 CDSs/19.6%) and mobile genetic elements (831 CDSs/21.1%) (Belda *et al*., 2010) (Figure 1.4).

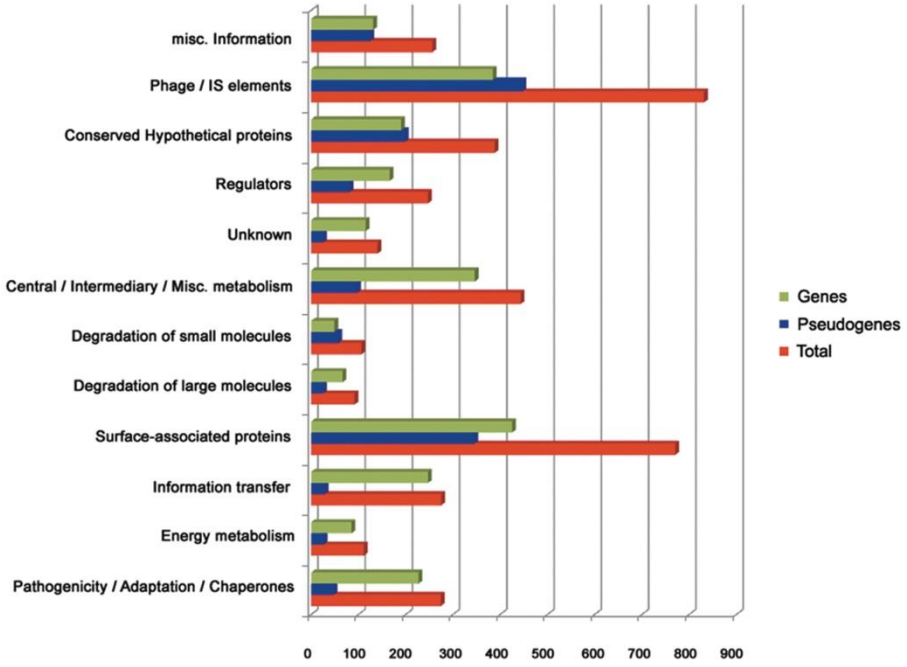


Figure 1.4. Re-annotated *Sodalis glossinidius* CDSs assigned functional categories. Red: number of CDSs (genes and pseudogenes). Green: number of genes. Blue: number of pseudogenes. **Taken from Belda *et al.* (2010)**

These two categories also have the highest number of pseudogenes; 345 and 447, representing 44.8% and 53.8% of their whole category, respectively (Belda *et al*.,

2010). There was a significant reduction in the number of hypothetical proteins following re-annotation compared to the original genome, involving 787 down to 190 conserved hypothetical and 115 unknown [to public databases] proteins, the latter originally described as 221 (Toh *et al.*, 2006; Belda *et al*., 2010). It was reported that the majority of the re-annotated 787 original hypothetical proteins were re-assigned to surface proteins (117 genes) and mobile DNA (245 genes) functional categories (Belda *et al*., 2010). A large proportion of the aforementioned mobile genetic elements category CDS assignments consisted of phage-related genes, constituting 17.8% of the entire genome [CDSs] (Belda *et al*., 2010). Two *S. glossinidius* genome regions were reported to be homologous [whole genome level] to Mu-family dsDNA bacteriophages, including regions named SGLp1 [enterobacterial phage Mu] and SGLp2 [*Burkholderia phage BcepMu*] (Belda *et al*., 2010). The prophage element in SGLp1 was deemed inactive due to the high proportion of gene inactivation in tail assembly, integration and transposition, lytic/lysogeny regulation, and protease and scaffolding proteins (Belda *et al*., 2010). The prophage element in SGLp2 was deemed active due to the retention of functionality and only a small proportion of pseudogenisation [5 out of 47 CDSs] (Belda *et al*., 2010). The cell adhesion and invasion lysis gene cassette was fully functional, in addition to genes encoding for DNA packaging, capsid formation and contractile tail biosynthesis (Belda *et al*., 2010). However, Belda *et al*. (2010) noted that this prophage element is most likely an uninducible cryptic prophage due to a lack of a phage control region. The re-analysis of IS found five distinct types, named ISSgl1 – 5 and all with homology to IS families known to γ-proteobacteria, but only ISSgl1 and ISSgl2 were deemed to be functional (Belda *et al*., 2010).

In 2018, a new annotation was uploaded into the public repository referencing a different strain designation: *S. glossinidius* strain SgGMMB4, referring to a *S. glossinidius* [Sg] isolate [B4] from *G. m. morsitans* [GMM] (Goodhead *et al*., 2018). SgGMMB4 was reported to have a 4.1 Mb chromosome, a mean G+C content of 54.4%, 3,336 putative CDSs, 2,286 putative pseudogenes and the same three previously annotated plasmids (Darby *et al*., 2005; Goodhead *et al*., 2018). It was reported that pseudogene enrichment focused within transmembrane

transport such as metal ions, receptors, transposition and glycerol metabolic processes (Goodhead *et al*., 2018). Accompanying the annotation was transcript- and proteomic data, predominantly to elucidate the potential importance of the unusually high number of pseudogenes within *S. glossinidius* (Goodhead *et al*., 2018). The transcriptomic data revealed that within any given condition, 850 and 644 pseudogenes exhibited active and putatively active sense transcription, respectively (Goodhead *et al*., 2018). Minimum activity thresholds were set using transcripts per million (TPM) ≥ 1 [putatively active] and ≥ 10 [active] from all [three] biological replicates in at least one culture condition [early- (ELP), late log phase (LLP), and late stationary phase (LSP)] (Figure 1.5) (Goodhead *et al*., 2018). Further to this, 385 and 364 pseudogenes exhibited active and putatively active antisense transcription, respectively (Goodhead *et al*., 2018).

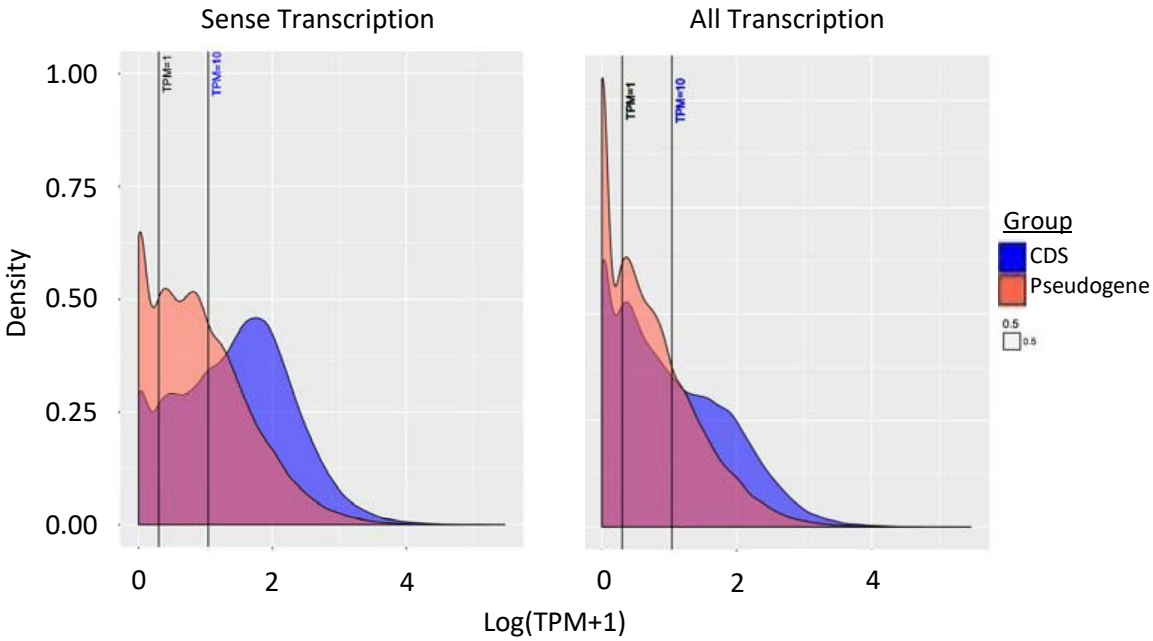


Figure 1.5. Density plots of *S. glossinidius* expression across three culture conditions. Left plot: sense transcription only. Right plot: all transcription. Blue: coding domain sequence (CDS). Orange: pseudogene. Two minimum thresholds for activity represented by the lines; left-most (per plot) represents transcripts per million (TPM) ≥ 1 [putatively active] and right-most (per plot), TPM ≥ 10 [active]. **Taken [and re-labelled] from Goodhead *et al*. (2018).**

As can be seen from Figure 1.5 [specifically sense transcription], higher levels of greater expression are observed within CDSs, compared to pseudogenes (Goodhead *et al*., 2018). Having said this, there is a relatively high level of overlap between CDS and pseudogene expression within the lower ranges, highlighting the difficulty in characterising pseudogenes based on expression profiles (Goodhead *et al*., 2018). It was reported that the pseudogenes being actively expressed [TPM ≥ 10] focused towards organonitrogen compound metabolic processes, compared to transposase activity amongst the pseudogenes being putatively expressed [TPM ≥ 1] (Goodhead *et al*., 2018). Differential expression analysis within the active transcription set [TPM ≥ 10] revealed that there were 219 DE pseudogenes, involved in the binding activities of nucleic acid and organic cyclic compounds, between LLP and LSP [compared to ELP growth] (Goodhead *et al*., 2018). It was also reported that up-regulation of SSR-1 was observed in ELP and within LLP and LSP, for SSR-2 and 3 (Goodhead *et al*., 2018). Comparatively, down-regulation is observed in LSP and ELP for SSR-1 and SSRs 2 and 3, respectively (Goodhead *et al*., 2018). The up-regulation of SSR-1 [containing homologous genes for functional TTSS apparatus and proteins associated with host cell invasion] in ELP potentially links in with the previously mentioned *S. glossinidius* utilisation of TTSS for host-cell invasion; expression of SSR-1 has been observed at a time point [24 hours] associated with invasion of insect cells (Kaniga *et al*., 1995; Dale *et al*., 2005; Goodhead *et al*., 2018). Using the same logic, this extends to the up-regulation of SSR-2 within LLP and LSP supporting the observation that SSR-2 is expressed post-invasion (Dale *et al*., 2005; Goodhead *et al*., 2018). The proteomics data revealed the presence of protein products corresponding to 34 pseudogenes from previous annotations; 27 were re-annotated in SgGMMB4 due to the detection of peptides (Goodhead *et al*., 2018). Comparison [based on single nucleotide polymorphism (SNP) rates] of the SgGMMB4 genome to six other *S. glossinidius* isolates from different, phylogenetically distant tsetse species revealed that there were 2,729 CDSs and 1,796 pseudogenes in the core genome [all seven genomes], strongly suggesting that the pseudogenes are both stable and actively maintained (Goodhead *et al*., 2018).

1.2.2.2. Metabolic Predictions

The original *S. glossinidius* genome annotation only reported three intact phosphotransferase systems (PTSs); *N*-acetylglucosamine, mannose, and mannitol (Toh *et al.*, 2006). Pathways typical of free-living bacteria were also reported to be functional, including the citric acid cycle (CAC), the pentose phosphate pathway (PPP), glycolysis and gluconeogenesis (Toh *et al.*, 2006). It was also the case that components of all [bar alanine] the amino acid biosynthesis pathways were covered by the chromosome, but amino acid degradation content was lacking in terms of gene inventory (Toh *et al.*, 2006). A variety of energy production and conversion through anaerobic respiration genes and anaerobic fermentation of lactate and butyrate are completely lacking, in addition to a loss in the glyoxylate bypass to the CAC (Toh *et al.*, 2006). As one would expect from adaptation to the iron-rich gut of the tsetse fly, *S. glossinidius* is reliant on the ability to transport heme/haemoglobin and utilises a TonB-dependent ATP-binding cassette (ABC) transporter homologous to *Y. enterolitica* HutB/HemU/HmuV and a TonB-independent iron transporter homologous to SitABCD of *Salmonella* [located on SPI-1] (Zhou *et al*., 1999; Toh *et al.*, 2006). Other *S. glossinidius* siderophore systems were also reported, including the intact FeoA [ferrous iron transporter] and homologous achromobactin-like gene clusters on pSG1 encoding for biosynthesis and its ABC transporter (Darby *et al.*, 2005; Toh *et al.*, 2006).

The more detailed metabolic re-construction from the re-annotated genome of *S. glossinidius* was achieved using a combination of the Kyoto Encyclopaedia of Genes and Genomes (KEGG) Automated Annotation Server, Blast2GO and literature searches involving the metabolic databases, ECOCYC and METACYC (Kanehisa and Goto, 2000; Conesa *et al*., 2005; Belda *et al*., 2010; Kanehisa *et al*., 2016, 2017; Keseler *et al.,* 2017; Caspi *et al*., 2018). In line with the original annotation conclusions, it was reported that *S. glossinidius* possesses a metabolic capability more associated with free-living bacteria than symbiotic obligates based on the energy utilisation from a carbon source range, and metabolite and macromolecule biosynthesis competency (Toh *et al*., 2006; Belda *et al*., 2010). One of the key findings from this, contradictory from the original

annotation, was that *S. glossinidius* is not able to biosynthesise L-arginine [from glutamate]; gene inactivation affecting *argA*, *argC*, *argD* and *argG*, corresponding with the first, third, fourth and seventh steps of the L-arginine biosynthetic pathway (Toh *et al*., 2006; Belda *et al*., 2010). This pathway has eight steps, involving L-ornithine production from L-glutamate in five of them and L-arginine production from L-ornithine and carbamoyl phosphate, in three (Cunin *et al*., 1986). It was theorised that *S. glossinidius* thus acquires arginine from tsetse utilising an *artPIQMJ* ABC transport system previously characterised in *E. coli* (Wissenbach *et al*., 1995; Belda *et al*., 2010). The inactive biosynthetic pathway for L-arginine including the inability to produce the ornithine intermediate, in conjunction with an inactive Lysine/Arginine/Ornithine (LAO) ABC transport system, also negatively implicates the biosynthesis of putrescine; a metabolite involved with cell development synthesised from arginine or ornithine (Belda *et al*., 2010; ChEBI, 2019). Further to this, it was also reported that the *S. glossinidius speC* gene encoding ornithine decarboxylase is inactivated (Belda *et al*., 2010). Ornithine decarboxylase is required for the biosynthesis of putrescine from L-ornithine (Belda *et al*., 2010). It was also reported that *S. glossinidius* is unable to biosynthesise thiamine as it lacks the essential sulfur-carrier protein ThiI [encoded by *thiI*], and *thiF* [thiamin (thiazole moiety) biosynthesis protein ThiF] is pseudogenised (Belda *et al*., 2010). However, it may be able to utilise thiamine through a *yjbQ*-encoded salvage pathway, in combination with an intact thiamine ABC transporter (Belda *et al*., 2010). *yjbQ* [designated SG2130 in the re-annotated genome, and GMMB4_05121 in *S. glossinidius* strain SgGMMB4] encodes for the YjbQ enzyme, which has been shown to complement thiamine auxotrophy in thiamine phosphate synthase (TPS)-deficient *E. coli* mutants (Morett *et al.*, 2008). Belda *et al*. (2010) reported that *W. glossinidia* also lacks *thiF*, and as a result, postulated that pathway intermediates produced by one symbiont may complement gaps in the defunct thiamine pathway for the other. *S. glossinidius* retains the ability to synthesise thiazole phosphate carboxylate (THZ-P) and *W. glossinidia* is able to produce hydroxymethyl pyrimidine pyrophosphate (HMP-PP), both required for producing functional thiamine diphosphate (Belda *et al*., 2010).

S. glossinidius L-arginine auxotrophy was further confirmed using the *in silico* tool of Flux Balance Analysis (FBA), which is a mathematical approach that calculates metabolite flow through reconstructed, genome-scale metabolic networks (Orth *et al*., 2010; Belda *et al*., 2012). Using *E. coli* K-12 as the closely-related reference organism for orthologue obtainment, Belda *et al*. (2012) constructed the *i*EB458 network [using FBA]; initial biomass equations found that to ensure *in silico S. glossinidius* survival, basic nutritional requirements [amino acids, cofactors, phospholipids and nucleotides] expected for prokaryotes, were required (Belda *et al*., 2012). When using glucose as the primary carbon source, equivalent biomasses were produced between *S. glossinidius* and *E. coli* K-12 within the ancestral metabolic network, but failed to within the [*S. glossinidius*] extant network (Belda *et al*., 2012). Gene pseudogenisation resulted in a failure of glycogen biosynthesis from ATP and α-D-glucose 1-phosphate [*glgA* and *glgB*], and L-arginine from L-glutamate [*argA, argG, argD* and *argC*]. Both glycogen and L-arginine were biomass components set within the equation thresholds and upon their removal as such, the glycogen biosynthesis pathway was restored in the extant network (Belda *et al*., 2012). However, the elimination of L-arginine as a biomass constituent had no impact on the lethality of that phenotype, so it was concluded that only by externally introducing L-arginine that viable *S. glossinidius* biomass phenotypes were produced [*in silico*] as the biosynthesis of putrescine and spermidine were restored (Belda *et al*., 2012). As previously mentioned, these polyamines are involved in cell development and can only be produced from L-arginine, as the L-orthinine alternative for biosynthesis [*speC*] has been pseudogenised (Belda *et al.*, 2010; Belda *et al*., 2012). Further to the essentiality of externally-supplied L-arginine, both *ppc* [encodes for phosphoenolpyruvate (PEP) carboxylase] pseudogenisation, resulting in the failure of oxaloacetate production, and the previously mentioned absence of glyoxylate bypass genes, meant severe reduction in biomass within the entire metabolic system (Toh *et al*., 2006; Belda *et al*., 2012). Putrescine produced from L-arginine was required in the CAC for the production of succinate, which is converted into oxaloacetate in the absence of PEP carboxylase (Belda *et al*., 2012). A more recent FBA-based paper using *S.*

glossinidius strain SgGMMB4 and *S. praecaptivus* as the ancestral reference for gene orthologues claimed that *S. glossinidius* preferred *N*-acetylglucosamine over glucose as a carbon source, supporting the results of assimilation tests conducted closer to the time when *S. glossinidius* was first cultured *in vitro*; stronger growth was observed when using *N*-acetylglucosamine and raffinose as primary carbon sources, in comparison to glucose (Welburn *et al.*, 1987; Dale and Maudlin, 1999; Hall *et al*., 2019). However, Hall *et al*. (2019)'s conclusion was partly based on the pseudogenisation of the glucose-specific PTS transporter [encoded for by *ptsG*] and could potentially be incorrect as even though the *ptsG* _1 - _5 gene copies are pseudogenised, *ptsG*_6 is intact, 1.28 Kb in length and has RNASeq evidence of transcription (Goodhead *et al*., 2018). The intensity of the *S. glossinidius* L-arginine auxotrophic phenotype was disputed within this study, where *S. glossinidius* cultures in a defined medium named SGM11 lacking L-arginine showed vastly reduced but not totally eradicated growth, and were subsequently rescued by the addition of L-glutamate (Hall *et al*., 2019). Of-note for the FBA studies discussed: FBA uses genome annotations for its *in silico* metabolic predictions, which for well-annotated genomes does not pose much of an issue (Orth *et al*., 2010). However, heavily pseudogenised genomes with relatively high numbers of mobile genetic elements, such as that of *S. glossinidius*, can lead to 'untrustworthy' annotations from the inability of assembly software to differentiate genetic components as pseudogene fragments or genuine short genes (Orth *et al*., 2010).

1.2.3. Evolutionary Genomics of Symbionts

As has been previously mentioned, it has been long established that primary symbionts have long ancestral histories with their hosts, one of the evidences being reduced genomes. Genome reduction in bacteria takes place when a lifestyle switch from free-living to symbiotic is made; the idea of 'minimal genomes' defines the remaining set of genes as being the minimum amount required for survival in a lenient environment, for example genes that are required for contributing to host wellbeing (reviewed in McCutcheon and Moran, 2012). This relaxed selection

introduces a maintenance gap, whereby the 'unnecessary' genes are lost due to large deletional events (Mira *et al*., 2001). Where large free-living bacterial populations can account for random deleterious mutations by horizontal gene transfer and active maintenance of functional gene sets, the small population sizes of symbionts imposed by the bottlenecking of host vertical transmission routes exacerbates vulnerability to genetic drift; imposing the full effect of Muller's ratchet and driving dramatic genome evolution (Moran, 1996; Mira *et al*., 2001; Mira and Moran, 2002; reviewed in McCutcheon and Moran, 2012). Further to genomic decay, lifestyle switching sees a surge in mobile genetic elements, an abundance of pseudogene formation, rearrangements within the genome and chromosomal fragment deletion (reviewed in McCutcheon and Moran, 2012). These types of genetic issues also present problems when trying to decipher phylogenetic histories, but multiple studies have used genetic techniques including 16S comparisons to relate *S. glossinidius* to *W. glossinidia* amongst other Proteobacteria [specifically Enterobacteriaceae]; supported by the cladogram created by Husník *et al.* (2011) placing *S. glossinidius* in the same clade as *W. glossinidia* (Figure 1.6) (O'Neill *et al*., 1993; Aksoy *et al*., 1997; Toh *et al.,* 2006).

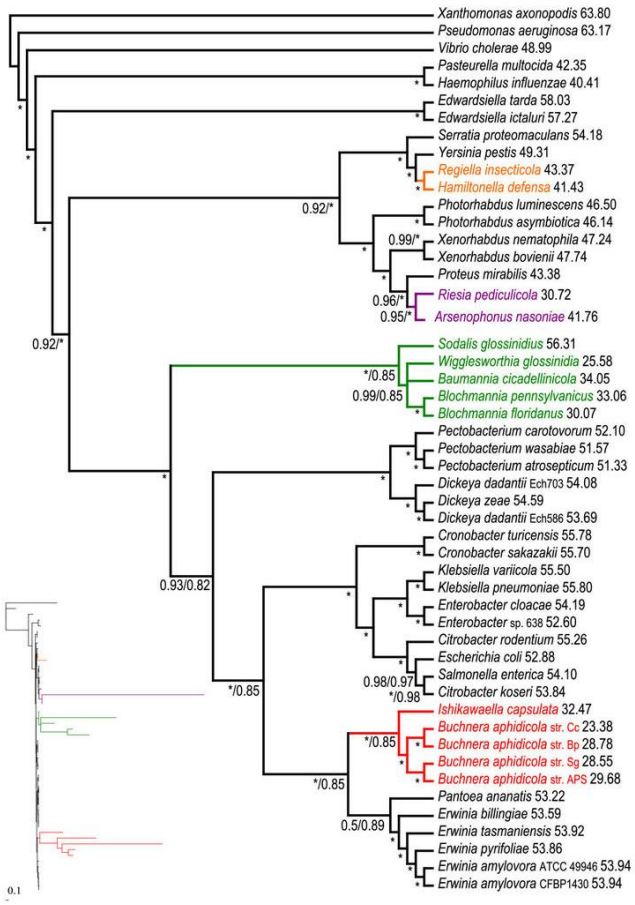


Figure 1.6. PhyloBayes cladogram using 69 genes. The cladogram was constructed using the CAT+GTR and CAT model from amino acid matrix recorded with Dayhoff6. **Taken from Husník *et al*. (2011).**

As has been eluded to throughout this writing, *S. glossinidius* is the epitome of genome degradation associated with switching from free-living to symbiotic, with a small coding capacity (only approximately 50%), large proportions of the genome encoding mobile genetic elements and an abnormally high number of pseudogenes. This facultative secondary symbiont offers a unique insight into the switching process, as multiple evidences have been presented that *S. glossinidius* is a new symbiont still undergoing the adaptation. Evidences at the genotypic level include *S. glossinidius* retention of a functional repertoire more associated with free-living bacteria, involving fully functional metabolic pathways able to utilise a range of carbon sources, virulence-associated gene sets including TTSS and siderophore biosynthesis and transport [in the context of pathogenic organism in-

host survival], not to mention the genome size itself (Ratledge and Dover, 2000; Dale *et al*., 2005; Darby *et al*., 2005; Toh *et al*., 2006; Belda *et al*., 2010). Multiple studies have utilised the tsetse host to confirm recent *S. glossinidius* association. For example, using *S. glossinidius* populations from different tsetse species, one study found near-identical 16S sequences, and another low SNP rates with a high number of shared pseudogenes (Aksoy *et al*., 1997; Goodhead *et al*., 2018). Phenotypic evidence includes the ability to culture *S. glossinidius in vitro* devoid of insect cell lines, confirming it as one of few currently known symbionts able to grow in laboratory media outside of its host (Welburn *et al.*, 1987; Dale and Maudlin, 1999; Matthew *et al*., 2005). Having said this, *S. glossinidius* is highly fastidious, most likely due to the level of genome degradation. Laboratory plating requires rich agar supplementation with blood, a microaerophilic atmosphere and incubation at 25°C in order to achieve colonies, after 5 – 10 days. Laboratory liquid culturing requires rich media statically incubated at 25°C in order to achieve mid-log phase growth, after 3 – 5 days (Matthew *et al*., 2005).

1.2.4. Genetic Manipulation for Functional Understanding

Since the introduction of next generation sequencing, the discovery of new genes and gene sets within prokaryotes has increased rapidly. Within 14 years after the first fully-sequenced prokaryote genome of *Haemophilus influenzae* Rd (Fleischmann *et al.*, 1995), 1052 more had been made freely accessible within online databases (Galperin and Koonin, 2010). Discovering a new gene has applications, such as potentially being able to track bacterial ancestry, but without understanding what function the gene has, the usefulness of knowing it exists becomes questionable. The new knowledge gained from the explosion in genome sequencing meant that researchers could then be more selective in their targets, as knowing which genes exist and where within a bacterial genome allows for specificity within experiments such as knockout mutant creations and the cloning and expression of desired genes. Experiments such as these are aimed at elucidating gene function, as even with the vast amount of sequence data, the

collective knowledge of linking phenotype to genotype remains sparse. As an example, the use of cloning, recombineering and genome editing in *S.* Typhimurium ST_{313} allowed for the discovery of a novel prophage ("Salmonella virus BTP_1") (Owen *et al.*, 2017). BTP_1 was theorised as having the potential to be associated with virulence in the epidemic bacterium; the BTP_1 gene, *st_{313}-td*, was previously reported as being linked with virulence in mice (Owen *et al.*, 2017; Herrero-Fresno *et al.*, 2014).

One of the most successful and widely-used methods aimed at bridging the genotype-phenotype gap is transposon sequencing. In the context of deliberate laboratory introduction, transposons are short pieces of DNA that can randomly integrate into a genome with the action of the enzyme, transposase. An example of a transposon that can be utilised *in vitro* includes Tn5, which is known to operate under the so-called 'cut and paste' mechanism of transposition. There are three components involved in Tn5 transposition: transposon DNA (can be embedded within donor DNA, such as a plasmid vector), the transposase enzyme and the target DNA for integration (reviewed in Reznikoff, 2003). Tn5 is composed of two sections of inverted transposable IS *50*, both containing an outward-facing, 19 bp-long outside end (OE) and an inwards-facing, 19 bp-long inside end (IE) (reviewed in Reznikoff, 2003). IEs are important transposase binding sites (Figure 1.7).

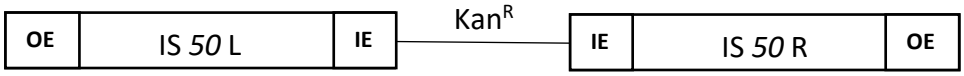


Figure 1.7. Graphical representation of the Tn5 transposon. Two sections of inverted transposable insertion sequences (IS *50* L and R) flank internal DNA such as antibiotic resistance genes, represented here with kanamycin as an example (Kan^R). Both IS *50* L and R comprise of outside end (OE) and inside end (IE) 19 bp-long sequences (that vary at seven positions), which are important transposase binding sites. **Figure derived from Reznikoff (2003).**

Specifically for Tn5 transposition, it is the OE sections that are inverted (reviewed in Reznikoff, 2003). These IS sections encompass internal DNA including antibiotic resistance genes, such as kanamycin resistance (reviwed in Reznikoff, 2003).

The transposase that partakes in the Tn5 system has been classed within the IS4 family of transposases and its catalytic core within the RnaseH protein superfamily, which encompasses proteins such as retroviral integrases and Mu transposases (reviewed in Reznikoff, 2003). It can either be transcribed from the IS *50* R section of the Tn5 itself, or alternatively supplemented into the system (reviewed in Reznikoff, 2003). The mode-of-action of Tn5 integration is graphically depicted in Figure 1.8, but briefly: a synaptic complex is formed when two monomers of the transposase bind to each OE region of the Tn5 (where positions 1 – 9 are inserted into each active site) and then dimerise, forming a loop-like structure (reviewed in Reznikoff, 2003). This is followed by transposon DNA cleavage from donor DNA (in the presence of Mg++ or Mn++), involving 3'-strand nicking, hairpin formation and hairpin cleavage: nucleophiles (the oxygen from H_2O, in this case) donate electrons that nick the 3'(transferred)-strands at both ends (Figure 1.8 A and B); hairpins are formed when the 3'OH group on the nicked transferred strand undergoes nucleophilic attack onto the 5' strand, which frees the transposon DNA from the flanking donor DNA (Figure 1.8 B and C); the hairpins are hypothesised to be nicked and move away from the active site, allowing for the docking of the target DNA (Figure 1.8 D) (reviewed in Reznikoff, 2003). This is followed by target capture and transfer, which little is known, but it is theorised that target DNA awaits strand transfer whilst bound to a basic cleft within the synaptic complex (reviewed in Reznikoff, 2003). The phosphates within the held target DNA backbone are attacked at their phosphodiester bonds (9 bp apart) by two 3'OH group ends (which are embedded within the basic cleft of the synaptic complex), which causes the transposon to integrate (reviewed in Reznikoff, 2003).

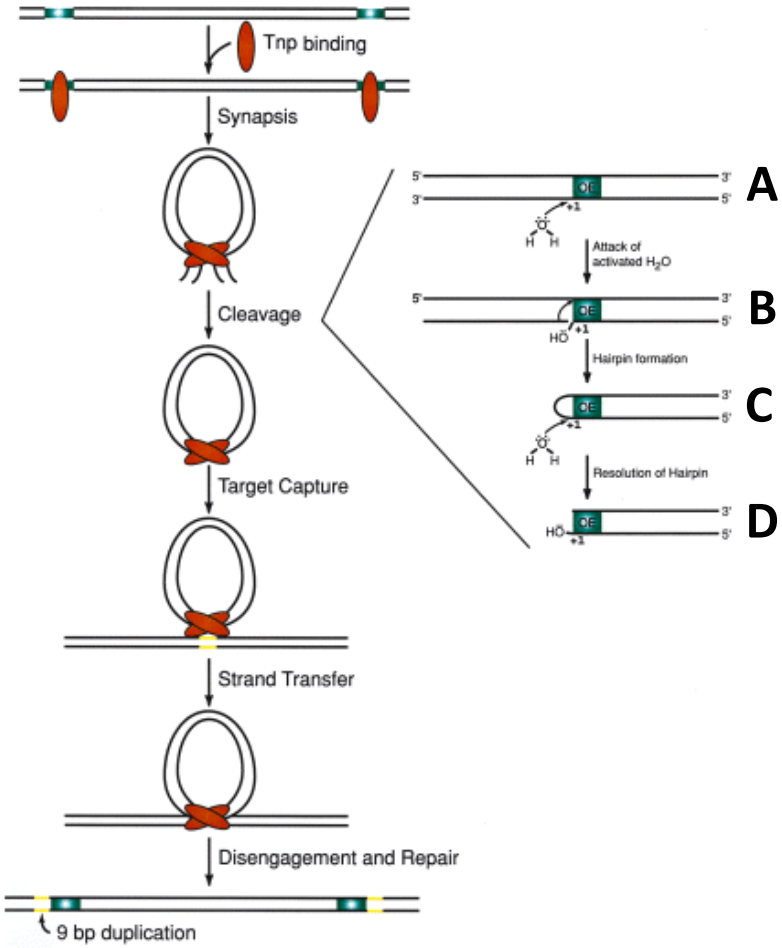


Figure 1.8. Graphical representation of the Tn5 'cut and paste' mechanism of integration. The transposase binds to the inverted insertion sequences on the transposon and forms a synaptic complex. Transposon DNA is cleaved from donor DNA by the steps of 3'-strand nicking, hairpin formation and hairpin cleavage. The target DNA is captured within a basic cleft in the synaptic complex, which is attacked by within-cleft 3'OH groups, and then integrates. The complex is removed *in vitro* by the addition of compounds such as SDS or phenol. **Figure taken from Reznikoff (2003).**

Once the organism-of-interest has been transformed to carry the transposon, it can be exposed to selection pressures and the mutants that are recoverable can be sequenced to see which genes contain insertions. The underlying notion of transposon sequencing is that the genes without insertions can be considered essential for the particular selection pressure. 382 genes were deemed essential in *Mycoplasma genitalium*, possessing the smallest known genome size for culturable prokaryotes, for growth in SP4 medium (Glass *et al.*,

2006). Transposon sequencing has also successfully been used upon *E. coli* K-12 and *S. enterica* serovar Typhi, whereby 355 and 356 [laboratory conditions] growth-associated essential genes were identified, respectively (Goodall *et al*., 2018; Langridge *et al*., 2009). The two aforementioned essential gene data sets were obtained using a particular type of transposon sequencing, known as transposon-directed insertion site sequencing (TraDIS), where the organism is transformed with a transposon, such as the previously mentioned Tn5, and the pooled mutants selected for within a desired environment (for example *in vitro* growth in the laboratory) (Langridge *et al*., 2009). The DNA is extracted, fragmented and poly(A) overhangs introduced, which allows for ligation of the T overhang Illumina® (Illumina, Inc.©) sequencing adaptors (Langridge *et al*., 2009). The samples undergo PCR using primers that enrich for both the Illumina® (Illumina, Inc.©) adaptors and part of the transposon, creating amplicons of transposon-flanking sequences (Langridge *et al*., 2009). These are sequenced and the reads are mapped back to a reference genome (Langridge *et al*., 2009).

The difficulty in *in vitro* cultivation has contributed to the scarcity in *S. glossinidius* functional publications and thus a lack of full understanding of the symbiont's role within its tsetse host. The previously mentioned recombinant *Sodalis* expressing a trypanolytic nanobody is one an example of the successful transformation, displaying its scope for genetic manipulation (De Vooght *et al*., 2014). Other examples include the miniTn5 mutants created by Dale *et al*. (2001) to study the invasion dynamics of *S. glossinidius*, in addition to the reported creation of *S. glossinidius* λ Red recombination knock-out mutants (Pontes and Dale, 2011), but despite this, no whole-genome functional genomics experiments have been published.

1.3. Aims and Objectives

As has been discussed within this writing, the fight against HAT and AAT remains prevalent within sub-Saharan communities. The concept of utilising the disease vector symbionts as parasite control strategies is proving to be an exciting new avenue in disease control, with particular focus on the secondary symbiont *S. glossinidius*; the ability to culture *S. glossinidius in vitro* and its scope for genetic manipulation makes for a near-perfect model. However, a distinct lack of functional data within the [*S. glossinidius*] field introduces potential caveats into the full implementation of *S. glossinidius* as a standard control strategy, stemming from a general absence of understanding of its biological role within tsetse. The main objective of this research project was to fill the functional genotype-phenotype gap for *S. glossinidius*, presenting:

- Broad-range metabolite cultivation experiments aimed at further clarifying the *S. glossinidius* metabolic profile, in the context of confirming gene functionality within the genome (Chapter 2)
- The initial adaptation, development and final optimisation of genetic transformation methods used for elucidating functionality within an organism, aimed at specific use on *S. glossinidius* (Chapter 3)
- The development of a cell line infection protocol adapted for *S. glossinidius*, aimed specifically for subsequent use in *in vivo* functional studies (Chapter 4)
- A *S. glossinidius* TraDIS library utilising different selection pressures aimed at identifying essential gene candidacy, and thus functionality within specific conditions (Chapter 5)

Chapter 2; Metabolic Profiling of *Sodalis glossinidius*

2.1. Introduction

Growth conditions needed by bacterial species have been fully established for many years. Factors such as oxygen, pH, temperature and salinity need to be at appropriate levels for individual bacterial species, to a degree where these factors have been known to maintain unique niches. For example, *Halomonas* spp. and *Dunaliella salina* have adapted to thrive in the extreme ends of the salinity range, allowing them to proliferate without competition, but additionally meaning that ranges considered 'normal' for the majority of bacteria could be uninhabitable and potentially toxic (Xiao-Ran *et al*., 2018; Oren, 2005). Another example being thermophiles such as *Pyrolobus fumarii* that can survive on deep sea steam vents reaching 113°C, but now cannot survive below 90°C (Blöchl *et al*., 1997).

The direct impacts of these growth conditions, for example too high of a salinity value causing lysis in a bacterial species not adapted for such a range, act in conjunction with the indirect impacts. Metabolic pathway routes can change based on the surrounding environment, an example including the glycolysis pathway being a fermentation process in the absence of oxygen versus a respiration process when in abundance (Madigan and Martinko, 2006a). Glycolysis is one of the major energy production pathways and ultimately leads to the production of pyruvate from glucose. Pyruvate enters another major energy pathway, the CAC: a pathway that [during respiration] ultimately results in the production of CO_2 and [38 molecules of] ATP, the latter being produced as part of the electron transport chain (ETC) when NADH donates the electrons taken from the CAC carbon compounds to oxygen molecules (Madigan and Martinko, 2006a, b). The aforementioned vitally important catabolic CAC, also known by tricarboxylic acid (TCA) cycle or Krebs' cycle (Krebs and Johnson, 1937a, b), is a series of oxidation-reduction reactions, where high-energy carbon compounds are readily oxidised and coenzymes are freely reduced (Madigan and Martinko, 2006b; Berg *et al*., 2002). In short, pyruvate (produced during glycolysis) enters the cycle and is oxidatively decarboxylated into acetyl-CoA, which combines with oxaloacetate to form citrate (a tricarboxylic acid, hence TCA cycle) (Madigan and Martinko, 2006b; Berg *et al*., 2002). Citrate is

oxidatively decarboxylated to form α-ketoglutarate, which is oxidatively decarboxylated to form succinate (Berg *et al*., 2002). Succinate is regenerated back into oxaloacetate and thus closes the cycle (Berg *et al*., 2002). The molecules within the CAC are also important for many other reactions, such as the creation of amino acids from the precursor molecules α-ketoglutarate and oxaloacetate, the latter of which can additionally be converted into the glucose precursor, phosphoenolpyruvate (Madigan and Martinko, 2006a, b). Intermediates from the glycolysis reaction(s) also feed into the PPP, which is known to produce NADPH and ribose 5-phosphate, used in nucleotide synthesis (Madigan and Martinko, 2006c). The PPP also produces compounds that feed back into the glycolysis pathway; glyceraldehyde 3-phosphate is the glycolysis intermediate that is ultimately converted into pyruvate (Madigan and Martinko, 2006a, c). These types of energy production pathways are commonly shared, but different organisms utilise the products differently; heterotrophs use organic compounds such as sugars and amino acids for energy sources, whereas autotrophs utilise inorganic compounds such as CO_2 (Madigan and Martinko, 2006a).

Culturing bacteria from different origins, for example medically important pathogens, *in vitro* is possible and relatively easy for the majority of bacterial species using complex culture media offering an abundant range of carbon and nitrogen sources required for growth. But due to the rich nature of complex media, understanding exactly which components were required was difficult to decipher, so the idea of defined minimal media for highlighting exactly which components were necessary was constructed. Amongst the earliest examples are minimal media 1 [MM1] for *Escherichia coli*, CH, A and B medium in *Cellulomonas* spp. and *Bacillus cereus*, and X_1-medium and MMG for thermophilic *Bacillus caldotenax* (Kuhn *et al.*, 1979; Summers *et al*., 1979; Reiling *et al.*, 1985). The use of M63 minimal media has highlighted some functionality within the *E. coli* stress response, where in minimal media (M63) versus complex media (Luria-Bertani), it was revealed that expression for metabolic pathways and regulatory factors genes were regulated by RpoS [sigma factor] (Dong and Schellhorn, 2009). Not only did the use of these minimal media reveal particular mechanisms and functions required for individual bacterial

species, but its concept of identifying required components has also been successfully utilised for previously-unculturable organisms; “uncultivable” *Psychrobacter* sp. strain MSC33 was successfully grown *in vitro* upon the addition of a specific short peptide (Nichols *et al*., 2008).

Sodalis glossinidius is the secondary symbiont of tsetse flies and has been controversially linked to reducing host refractoriness to trypanosomes. Despite *S. glossinidius* being one of few known symbiotic organisms able to be cultured, it’s a highly fastidious organism whereby mid-log phase is only achieved after a minimum of 3 days post-inoculation only if the culture is stably maintained statically at 25°C (Matthew *et al*., 2005). This fastidious nature introduces problems such as contamination and difficulty in producing reliable biological replicates. There is a distinct lack of full understanding involving the metabolic profile of *S. glossinidius*, stemming from the above-mentioned laboratory issues in combination with its highly degraded genome. Studies attempting to elucidate the metabolic landscape have either involved assimilation tests using complex media, predictions based on the genome sequence, or *in silico* models (Dale and Maudlin, 1999; Toh *et al.*, 2006; Belda *et al*., 2010; Belda *et al*., 2012; Hall *et al*., 2019). Factors that have been agreed between the studies attempting to understand the *S. glossinidius* metabolic profile include (a.) the retention of pathways such as glycolysis, the CAC, the PPP, and gluconeogenesis (b.) biosynthetic pathways, including amino acids and siderophores, remain intact whilst degradative pathways are less functional/missing (c.) the ability to utilise a range of carbon sources (d.) the inability to biosynthesise thiamine or L-arginine, the latter impacting on L-ornithine, putrescine and ultimately the CAC (Dale *et al*., 2005; Darby *et al*., 2005; Toh *et al*., 2006; Belda *et al*., 2010; Belda *et al*., 2012; Hall *et al*., 2019). The original genome publication created a metabolic map (Figure 2.1) based on the sequence data, showing the few intact PTS transporters, including *N*-acetylglucosamine, mannose and mannitol, and the aforementioned functional pathways [(a.)] (Toh *et al*., 2006).

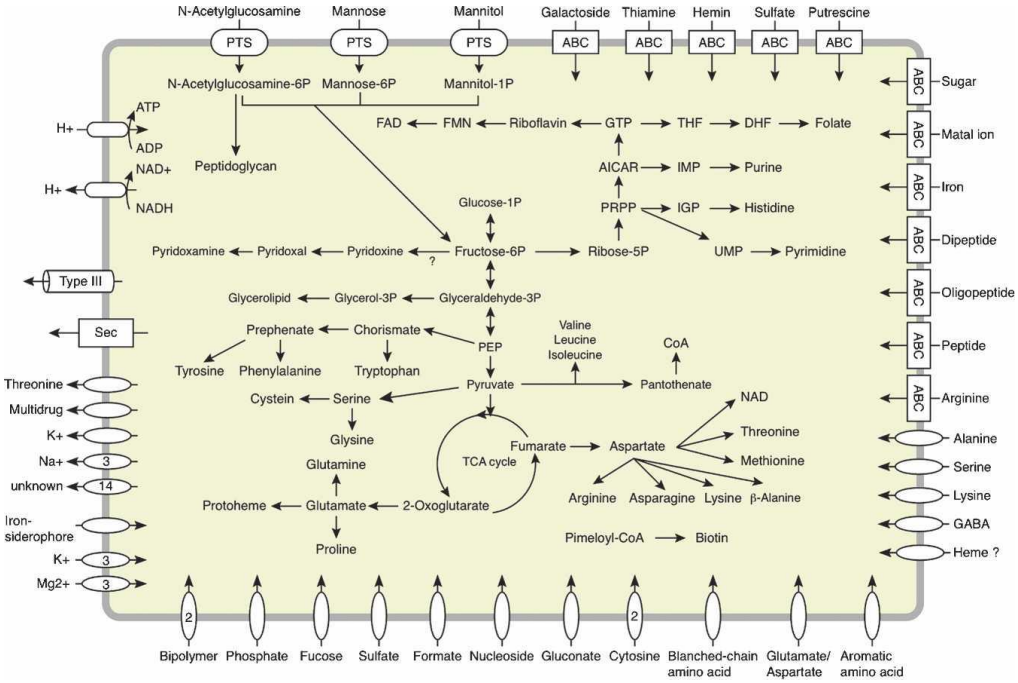


Figure 2.1. A map depicting metabolic pathways utilised in *Sodalis glossinidius*. Constructed using the data obtained from chromosomal genes (incl. putative functions). **Taken from Toh *et al.* (2006).**

In order to elucidate some functionality within the *S. glossinidius* genome, experiments attempting to discern more details regarding the metabolic capabilities were conducted screening a variety of minimal and complex growth media. The hypothesis for the initial experiment (2.2.1.) was that *S. glossinidius* would not be able to grow in minimal media offering a comparatively limited selection of growth components such as Lysogeny Broth (LB), but would in media containing a complex range of components such as Mitsuhashi and Maramorosch Insect Medium (MMI) and Schneider's Insect Medium (Table 2.1). The hypothesis for the supplementation experiment (2.2.2.) was that *S. glossinidius* wouldn't be able to grow in M9 due to a lack of usable growth components for a fastidious organism, but would be able to in the M9 variations that contained supplemented amino acid, vitamin and carbohydrate [namely glucose] sources (Table 2.1). The hypothesis for the carbon screening experiment (2.2.3.) was that *S. glossinidius* would grow best with glucose, raffinose, *N*-acetylglucosamine, mannitol and mannose as these are either the transport systems identified as intact or sources previously revealed as preferred (Dale and Maudlin, 1999).

2.2. Methods

2.2.1. Minimalistic versus Complex Media

LB (Merck Millipore), a variation of LB referred to here as 'modified' LB (mLB), Mitsuhashi and Maramorosch Insect Medium (MMI) and Schneider's Insect Medium (Sigma-Aldrich) and counterparts of all of these media containing 10% foetal bovine serum (FBS) (Gibco) were tested (Table 2.1). 900 μL of each media variation was pipetted into individual wells of a 24 well plate (Greiner Bio-One), along with 100 μL of *S. glossinidius* strain SgGMMB4 (GenBank accession no. LN854557) suspension seeded directly from a mid-log phase Schneider's Insect Medium + 10% FBS culture. The 24 well plates were statically incubated at 25°C for 5 days. Optical density (OD) at A_{600nm} of each well was then measured (using relevant blank media as a zero) and the data was analysed with Analysis of Variance (ANOVA) and Tukey HSD post-hoc tests using R (R Core Team, 2013).

2.2.2. Component-Supplemented Minimal Medium

Experiments were conducted testing individual amino acid, vitamin and carbon sources from complex media, using minimal media as a base. M9 Minimal Salts Medium (M9) (Miller, 1972), M9 + sodium bicarbonate [120 mg/L], M9 + tryptone [10,000 mg/L], M9 + lactalbumin hydrolysate [6500 mg/L], M9 + yeast extract [5000 mg/L], combinations of M9 + yeast extract with either tryptone or lactalbumin hydrolysate, all of the mentioned but in a variation of M9 without glucose ($M9^{g-}$) and Schneider's Insect Medium (Sigma-Aldrich) + 10% FBS (Gibco) were tested. 900 μL of each media variation was pipetted into individual wells of a 24 well plate (Greiner Bio-One), along with 100 μL of *S. glossinidius* strain SgGMMB4 (GenBank accession no. LN854557) suspension seeded directly from a mid-log phase Schneider's Insect Medium + 10% FBS culture. The 24 well plates were statically incubated at 25°C for 5 days. OD at A_{600nm} of each well was then measured (using relevant blank media as a zero) and the data was analysed with ANOVA and Tukey HSD post-hoc tests using R (R Core Team, 2013).

Table 2.1. A table showing the individual components within different growth media. Amounts of each component are listed as mg/L.

Media	Components	Concentration (mg/L)	Volume in 1 L
Lysogeny Broth (LB)	Tryptone	10000	
	Sodium chloride	10000	
	Yeast extract	5000	
'Modified' LB (mLB)	Lactalbumin hydrolysate	10000	
	Sodium chloride	10000	
	Yeast extract	5000	
Mitsuhashi and Maramorosch Insect Medium (MMI)	Calcium chloride dehydrate	190	
	Magnesium chloride anhydrous	46.9	
	Potassium chloride	200	
	Sodium bicarbonate	120	
	Sodium chloride	7000	
	Sodium phosphate monobasic	173.9	
	D(+) glucose	4000	
	Lactalbumin hydrolysate	6500	
	Yeast extract	5000	
Schneider's Insect Medium	Calcium chloride anhydrous	600	
	Magnesium sulfate anhydrous	1810	
	Potassium chloride	1600	
	Sodium bicarbonate	400	
	Sodium chloride	2100	
	Disodium phosphate	700	
	Fumaric acid	60	
	D(+) glucose	2000	
	α-ketoglutaric acid	350	
	L(-) malic acid	600	
	Succinic acid	60	
	D(+) trehalose	2000	
	Yeast extract	2000	
	β-alanine	500	
	L-arginine	600	
	L-aspartic acid	400	
	L-cystine dihydrochloride	27	
	L-cysteine	60	
	L-glutamic acid	800	
	L-glutamine	1800	
	Glycine	250	
	L-histidine	400	
	L-isoleucine	150	
	L-leucine	150	
	L-lysine	1650	
	L-methionine	150	
	L-proline	1700	
	L-serine	250	
	L-threonine	350	
	L-tryptophan	100	
	L-tyrosine disodium salt hydrate	720	
	L-valine	300	
Minimal Salts Medium (M9)	20% glucose		20 mL
	1 M magnesium sulfate		2 mL
	1 M calcium chloride		0.1 mL
	(Deionized water)		(up to 1 L)
	5 x M9 salts stock solution		200 mL
For 5 x M9 salts stock:	Sodium phosphate dibasic heptahydrate	64000	
	Monopotassium phosphate	15000	
	Sodium chloride	2500	
	Ammonium chloride	5000	

2.2.3. Carbon Source Screen

A variety of individual carbon sources were selected based on the available metabolic data, including the metabolic map based on the original genome annotation (Figure 2.1), the latest strain SgGMMB4 genome (GenBank accession no. LN854557) and pathways identified using the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000; Kanehisa *et al*., 2016, 2017). Stock solutions of 30 mg/mL glucose, fructose, sucrose, mannose, raffinose, trehalose, maltose, *N*-acetylglucosamine, proline or mannitol in $M9^{g-}$ (Miller, 1972) + yeast extract + lactalbumin hydrolysate were created. The volume needed of a 30 mg/mL carbon source stock to make final well carbon source concentrations of 0 mg/mL (negative control), 5 mg/mL, 10 mg/mL, 15 mg/mL, 20 mg/mL or 25 mg/mL was added to each well, in a 24 well plate (Greiner Bio-One). Included within the calculation were the volumes needed of sterile $M9^{g-}$ + yeast extract + lactalbumin hydrolysate and *S. glossinidius* strain SgGMMB4 mid-log phase Schneider's Insect Medium (Sigma-Aldrich) + 10% FBS (Gibco) culture to make a final well volume of 1 mL (Table 2.2). Three wells per plate contained Schneider's Insect Medium + 10% FBS for use as a positive control. OD at A_{600nm} of each well was then measured (using relevant blank media as a zero) and the data was analysed with ANOVA and Tukey HSD post-hoc tests using R (R Core Team, 2013).

Table 2.2. A table showing the volumes needed per well for each concentration of carbon source tested. "Carbon source" (column 2) referring to glucose, fructose, sucrose, mannose, raffinose, trehalose, maltose, *N*-acetylglucosamine, proline or mannitol. "Sterile media" (column 3) referring to M9 Minimal Salts Medium (M9) (Miller, 1972) minus glucose + yeast extract + lactalbumin hydrolysate.

For a final well concentration of (mg/mL):	**Add:**		
	30 mg/mL carbon source stock (μL)	**Sterile media (μL)**	**Mid-log phase *S. glossinidius* culture (μL)**
0	0	900	100
5	167	733	100
10	333	567	100
15	500	400	100
20	667	233	100
25	833	67	100

2.3. Results

2.3.1. Minimalistic versus Complex Media

The results from the initial minimal versus complex media screen aimed at clarifying *S. glossinidius* metabolic capabilities are displayed in Figure 2.2.

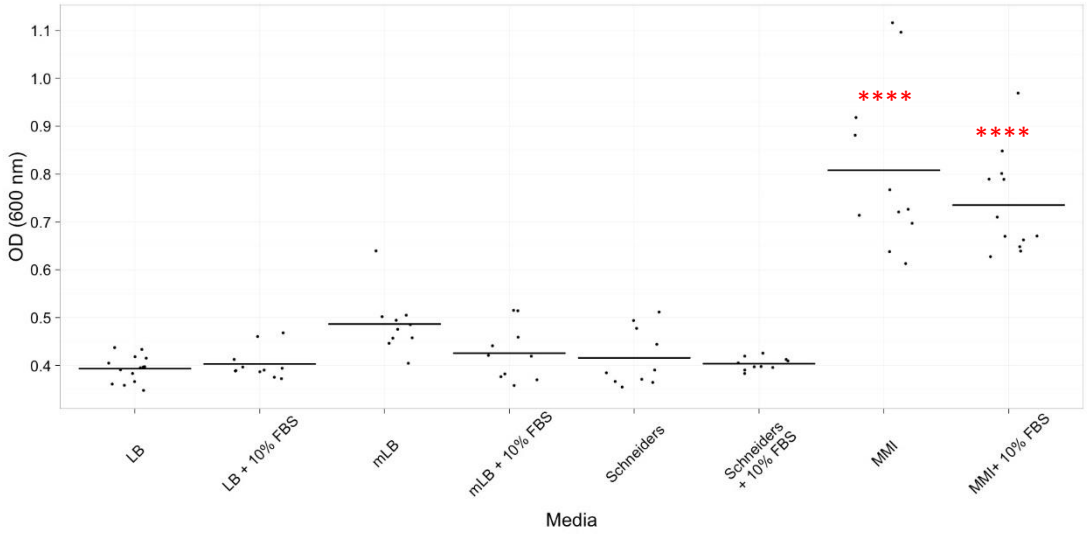


Figure 2.2. Data showing the growth of *Sodalis glossinidius* in different media. 100 μL of mid-log phase Schneider's Insect Medium + 10% foetal bovine serum (FBS) *S. glossinidius* suspension was seeded into 900 μL of either LB (*N* = 14), LB + 10% FBS (*N* = 11), mLB ('modified LB' - lactalbumin hydrolysate instead of tryptone) (*N* = 10), mLB + 10% FBS (*N* = 10), Mitsuhashi and Maramorosch Insect Medium (MMI) (*N* = 11), MMI + 10% FBS (*N* = 12), Schneider's Insect Medium (*N* = 10) or Schneider's Insect Medium + 10% FBS (*N* = 10). OD at A_{600nm} was measured after 5 days. ANOVA: $F_{7,80} = 46.38$, $p = < 2 \times 10^{-16}$. Tukey-HSD: "****" indicating $p = 0$.

As highlighted by Figure 2.2, there was a significant difference between MMI and MMI + 10% FBS [$p = 0$] compared to all other media [but not compared to each other]. Interestingly, the same significant result did not apply to the arguably broader [to MMI] Schneider's Insect Medium, with or without 10% FBS. The most notable differences between MMI and Schneider's Insect Medium are the higher concentrations of glucose and yeast extract, and the presence of lactalbumin hydrolysate, in MMI (Table 2.1). Yeast extract provides sources of carbon, nitrogen and vitamins and lactalbumin hydrolysate provides sources of amino acids, peptides and carbohydrates. As *S. glossinidius* is a fastidious symbiont, these higher

concentrations of vital growth factors could be a reason for the difference in growth levels seen between MMI and Schneider's Insect Medium.

Also apparent from Figure 2.2 is the lack of statistical significance between the 10% FBS variations, suggesting that the growth factor commonly used in cultivating *S. glossinidius in vitro* provides no growth benefit.

2.3.2. Component-Supplemented Minimal Medium

The results from deconstructing complex media and supplementing minimal media with the individual constituents, aimed at identifying exactly which components *S. glossinidius* utilises when grown in complex media, are displayed in Figure 2.3.

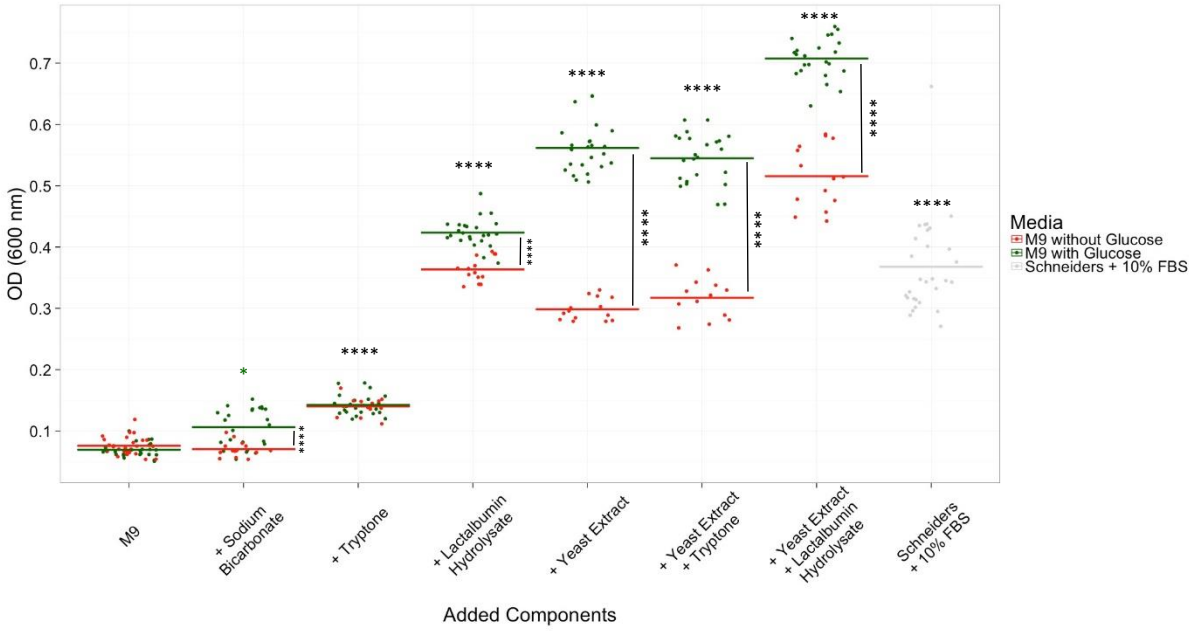


Figure 2.3. Data showing the growth of *Sodalis glossinidius* in different variations of M9 Minimal Salts Media (M9). 100 μL of mid-log phase Schneider's Insect Medium + 10% foetal bovine serum (FBS) *S. glossinidius* suspension was seeded into 900 μL of either M9 (*N* = 28), M9 minus glucose ($M9^{g-}$) (*N* = 27), M9 + sodium bicarbonate (*N* = 24), $M9^{g-}$ + sodium bicarbonate (*N* = 15), M9 + tryptone (*N* = 24), $M9^{g-}$ + tryptone (*N* = 14), M9 + lactalbumin hydrolysate (*N* = 24), $M9^{g-}$ + lactalbumin hydrolysate (*N* = 14), M9 + yeast extract (*N* = 24), $M9^{g-}$ + yeast extract (*N* = 14), M9 + yeast extract + tryptone (*N* = 24), $M9^{g-}$ + yeast extract + tryptone (*N* = 13), M9 + yeast extract + lactalbumin hydrolysate (*N* = 23), $M9^{g-}$ + yeast extract + lactalbumin hydrolysate (*N* = 14) or Schneider's Insect Medium + 10% FBS (*N* = 29). OD at A_{600nm} was measured after 5 days. Glucose positive ANOVA: $F_{7,192} = 854.7$, $p = < 2 \times 10^{-16}$. Glucose positive Tukey-HSD: "****" indicating $p = 0$ and "*" indicating $p = 0.05$, from negative control [M9]. Glucose negative ANOVA: $F_{7,132} = 249.1$, $p = < 2 \times 10^{-16}$. Glucose negative Tukey-HSD: "****" indicating $p = 0$, from negative control [$M9^{g-}$].

As highlighted by Figure 2.3, there was a significant difference to the negative control [M9 or $M9^{g-}$, depending] for all of the component variations tested, excluding $M9^{g-}$ versus $M9^{g-}$ + sodium bicarbonate and M9 versus $M9^{g-}$, confirming *S. glossinidius* heterotrophy; *S. glossinidius* cannot grow in minimal media providing a basic carbon source [M9 versus $M9^{g-}$, $p > 0.05$], but can when supplemented with sufficient amino acid [tryptone, lactalbumin hydrolysate, Schneider's Insect Medium], carbon [yeast extract, glucose, lactalbumin hydrolysate, Schneider's Insect Medium] or vitamin [yeast extract, Schneider's Insect Medium] sources, confirming the experiment's hypothesis. There was also a significant difference [$p = 0$] to the positive control [Schneider's Insect Medium + 10% FBS] for all of the glucose-positive variations tested, suggesting that *S. glossinidius* can grow to the same levels and/or higher as a complex medium rich in amino acids, vitamins, carbon sources and carbohydrates using just salts, water, glucose and lactalbumin hydrolysate, yeast extract and/or tryptone. This positive difference also applied to $M9^{g-}$ + yeast extract + lactalbumin hydrolysate [$p = 0$], implying that *S. glossinidius* is able to gain sufficient carbon from the supplements alone (although adding a distinct carbon source [glucose] resulted in comparatively significant higher growth values). There was a significant difference [$p = 0$] between tryptone [bacterial media amino acid source] and lactalbumin hydrolysate [insect cell line media amino acid source] for both $\pm$ glucose variations, suggesting that *S. glossinidius* appears to be able to more freely utilise insect cell media amino acid sources than bacterial media, perhaps suggesting higher bioavailability within the hydrolysate. The presence of yeast extract significantly [$p = 0$] increased the growth rates for both $\pm$ glucose variations of both tryptone and lactalbumin hydrolysate variations, suggesting that the source of nitrogen and vitamins are important for *S. glossinidius*. A significant difference [$p = 0$] was observed between the presence and absence of glucose for all component variations tested, excluding the negative controls [M9 versus $M9^{g-}$], and the tryptone $\pm$ glucose variation, suggesting that *S. glossinidius* grows better in the presence of glucose, even when there are other sources of carbon present (such as yeast extract or lactalbumin hydrolysate).

2.3.3. Carbon Source Screen

The results from the carbon source screen, aimed at experimentally confirming the functionality of *S. glossinidius* metabolic systems identified as intact within the literature using the genetic sequence, are displayed in Figure 2.4.

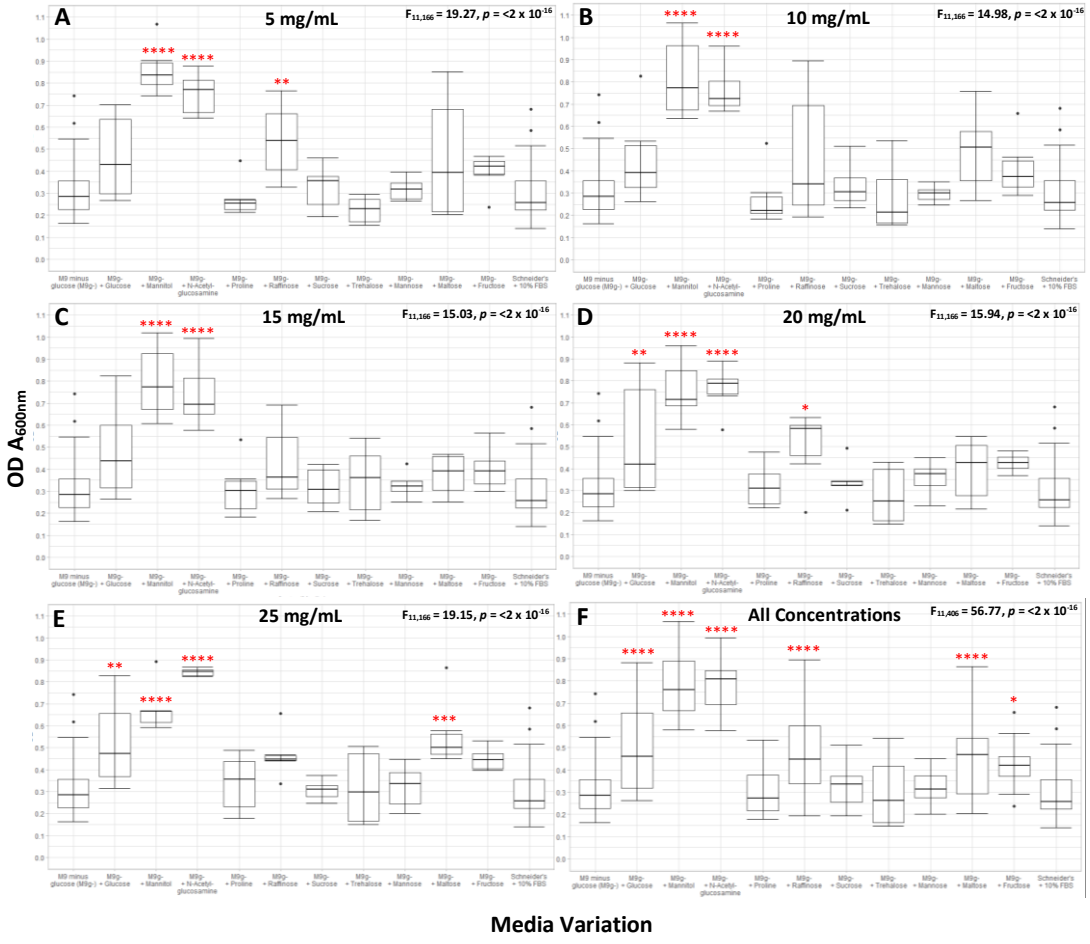


Figure 2.4. Data showing the growth of *Sodalis glossinidius* in different variations of M9 Minimal Salts Media (M9). Mid-log phase Schneider's Insect Medium + 10% foetal bovine serum (FBS) *S. glossinidius* suspension was seeded into either M9 minus glucose ($M9^{g-}$) + yeast extract (YE) + lactalbumin hydrolysate (LH) [negative control] (*N* = 54), $M9^{g-}$ + YE + LH + glucose (*N* = 30 total, *N* = 6 per concentration), $M9^{g-}$ + YE + LH + mannitol (*N* = 30 total, *N* = 6 per concentration), $M9^{g-}$ + YE + LH + *N*-acetylglucosamine (*N* = 30 total, *N* = 6 per concentration), $M9^{g-}$ + YE + LH + proline (*N* = 30 total, *N* = 6 per concentration), $M9^{g-}$ + YE + LH + raffinose (*N* = 30 total, *N* = 6 per concentration), $M9^{g-}$ + YE + LH + sucrose (*N* = 30 total, *N* = 6 per concentration), $M9^{g-}$ + YE + LH + trehalose (*N* = 30 total, *N* = 6 per concentration), $M9^{g-}$ + YE + LH + mannose (*N* = 30 total, *N* = 6 per concentration), $M9^{g-}$ + YE + LH + maltose (*N* = 30 total, *N* = 6 per concentration), $M9^{g-}$ + YE + LH + fructose (*N* = 30 total, *N* = 6 per concentration) or Schneider's Insect Medium + 10% FBS [positive control] (*N* = 58). Concentrations of the different carbon sources included 5 mg/mL (**A**), 10 mg/mL (**B**), 15 mg/mL (**C**), 20 mg/mL (**D**) or 25 mg/mL (**E**) or (**F**) all concentrations combined. OD at A_{600nm} was measured after 5 days. ANOVA: displayed on graphs. Tukey-HSD: "****" indicating $p = 0$, "***" indicating $p = 0.001$, "**" indicating $p = 0.01$ and "*" indicating $p = 0.05$, from negative control [$M9^{g-}$].

As can be seen from Figure 2.4, all the hypothesised carbon sources except for mannose aided in statistically significant higher growth values compared to the negative control. All concentrations tested for mannitol and *N*-acetylglucosamine were significantly higher than the negative control, but glucose was only significantly different in the higher concentrations (Figure 2.4 D and E), fitting the observations of higher growth values in comparatively glucose-rich MMI $\pm$ 10% FBS versus Schneider's Insect Medium $\pm$ 10% FBS (Figure 2.2). As the associated transporters for these carbon sources are all intact (excluding raffinose), the conclusion that *S. glossinidius* is able to utilise a range of carbon sources, an ability similar to that of free-living bacteria, is robust (Toh *et al.*, 2006). *S. glossinidius* appears to be lacking a specific system for raffinose transport (such as the raffinose ABC transport system encoded by the *raf* operon, in *Streptococcus pneumoniae*), so how *S. glossinidius* is transporting raffinose is unclear (Toh *et al.*, 2006; Tyx *et al*., 2011). It could be theorised that raffinose is sharing a PTS system, for example the glucose transporter as raffinose is made up of monosaccharides of similar structure. Glucose could also be sharing the mannose ABC transporter, based on the same structural similarity logic. At a concentration of 25 mg/mL maltose was also significantly different from the negative control (Figure 2.4 E), however as with raffinose, the maltose/trehalose transporter is not reported to be present/intact so its utilisation is unclear but the glucose-based disaccharide could theoretically be sharing the glucose and/or mannose transporter(s) (Toh *et al.*, 2006). When ignoring concentrations (Figure 2.4 F), it can be concluded that glucose, mannitol, *N*-acetylglucosamine, raffinose, maltose and fructose generally aid in better *S. glossinidius in vitro* growth compared to the negative control.

The data presented in Figure 2.4 is re-represented in Figure 2.5 A, B C and D.

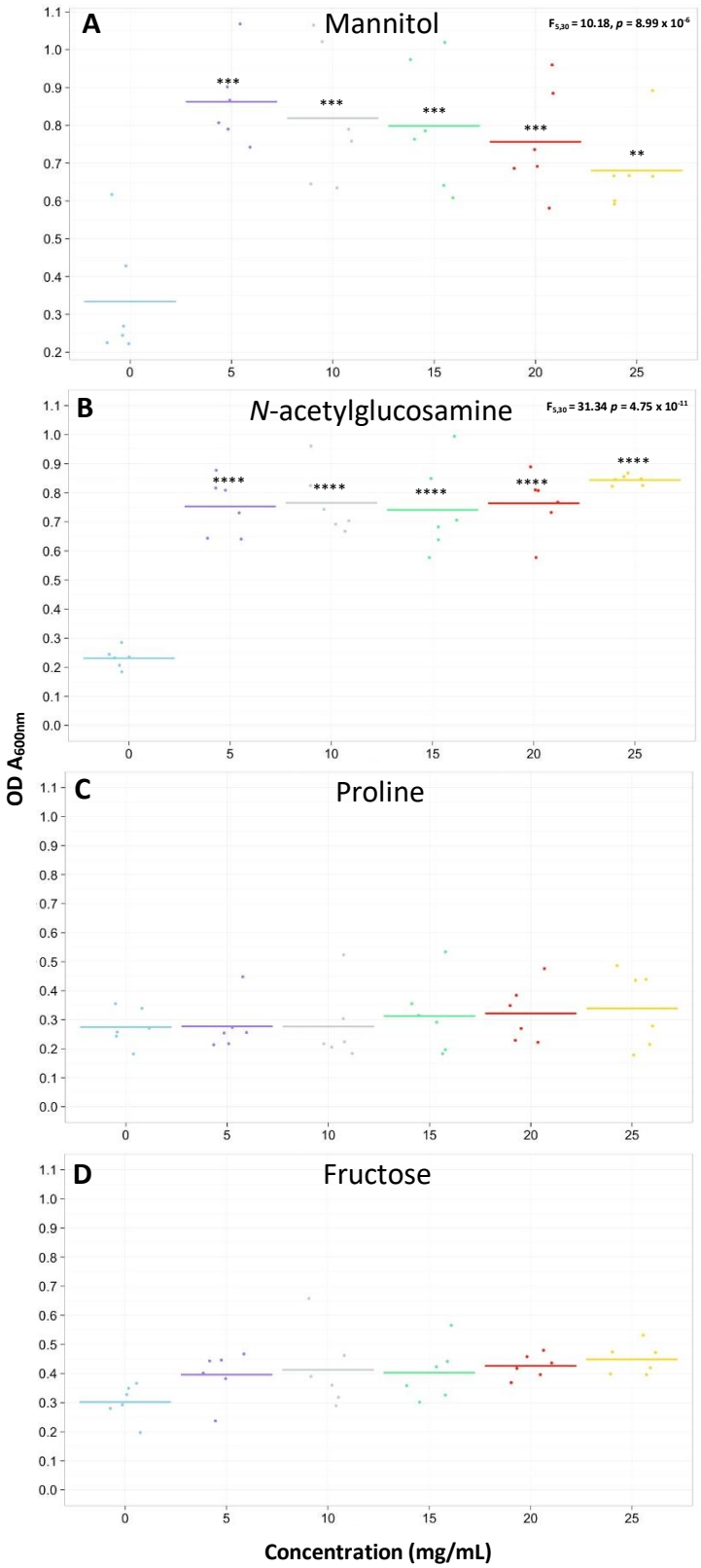


Figure 2.5. Data showing the growth of *Sodalis glossinidius* in different variations of M9 Minimal Salts Media (M9). Mid-log Schneider's Insect Medium + 10% foetal bovine serum (FBS) *S. glossinidius* suspension was seeded into either M9 minus glucose ($M9^{g-}$) + yeast extract (YE) + lactalbumin hydrolysate (LH) [negative control], $M9^{g-}$ + YE + LH + mannitol (**A**), *N*-acetylglucosamine (**B**), proline (**C**) or fructose (**D**). Concentrations included 5 mg/mL, 10 mg/mL, 15 mg/mL, 20 mg/mL or 25 mg/mL. OD at A_{600nm} was measured after 5 days. ANOVA: displayed on graphs. Tukey-HSD: "****" indicating $p = 0$, "***" indicating $p = 0.001$ and "**" indicating $p = 0.01$, from negative control.

As has already been discussed [Figure 2.4] and can be seen from Figure 2.5 A and B, mannitol and *N*-acetylglucosamine display significantly greater growth values than the negative control. Interestingly, Figure 2.5 A reveals a negative growth trend with increasing mannitol concentrations, suggesting that high concentrations of this monosaccharide [comparatively] hinder growth. The opposite trend is observed with *N*-acetylglucosamine (Figure 2.5 B), suggesting that higher concentrations aid in higher *S. glossinidius* growth rates. This same positive growth trend can be observed with proline (Figure 2.5 C) and fructose (Figure 2.6 D), although not statistically significant. Proline has been shown to be a vital energy source for tsetse, specifically important as the metabolite fuelling the flight process (Bursell, 1966). The positive trend observed between *S. glossinidius* growth and proline concentration (Figure 2.5 C) could suggest that the symbiont is adapting its metabolic processes based on environmental metabolite availability.

2.4. Discussion

The concept of using minimal media in the laboratory has long been used to identify specific components, and thus putative mechanisms and pathways involved, required for individual organisms that complex media masks. If applied to organisms traditionally restricted to a host organism, this simple technique can contextualise genetic predictions of metabolic functionality to observable *in vitro* phenotypes. As it currently stands, the functional understanding of *S. glossinidius* metabolic capabilities relies on data obtained from assimilation tests using complex media, predictions based on the genome sequence, or *in silico* models (Dale and Maudlin, 1999; Toh *et al.*, 2006; Belda *et al*., 2010; Belda *et al*., 2012; Hall *et al*., 2019). However, whilst this chapter was being written, a study using similar techniques was published in conjunction with FBA model predictions; supplementing M9 with selected components to create a defined medium, SGM11 (Miller, 1972; Hall *et al*., 2019). However, not only were their growth results not significantly different to a media comparatively rich in carbon, amino acid and peptide sources, but trehalose was added to SGM11, which has been observed to inhibit *S. glossinidius* growth (Parsons, G. [Darby Group] unpublished data, not shown; Hall *et al*., 2019). The lack of publications within the literature attempting to understand *S. glossinidius* metabolic functionality most likely stems from its fastidious nature; a highly degraded genome makes for a laboratory phenotype with complex nutritional requirements and long growth phases, introducing difficulty in manipulation, high contamination risks and arduous experiments.

The metabolic experiments conducted here utilised broad-range metabolite cultivation experiments with the aim of further clarifying the *S. glossinidius* metabolic profile, in the context of confirming gene functionality within the genome. Initial minimal versus complex media screens confirmed the hypothesis that the fastidious symbiont requires a range of nutritional supplements to grow *in vitro*, as *S. glossinidius* growth values were significantly higher in MMI $\pm$ 10% FBS than comparatively minimal media such as LB (Figure 2.2). As Schneider's Insect Medium contains more defined available amino acids than MMI, one could theorise that Schneider's Insect Medium would aid in higher growth values, but the results

shown in Figure 2.2 dispute this however, showing that MMI ± 10% FBS supported comparatively higher growth. Possible reasons for this observation include the presence of growth-inhibitory trehalose in Schneider's Insect Medium, in addition to the higher concentrations of glucose and yeast extract and presence of lactalbumin hydrolysate, in MMI (Table 2.1) (Parsons, G. [Darby Group] unpublished data, not shown). Schneider's Insect Medium provided similar growth values to LB, which disputes the theory that a growth media containing more complex components aids in higher growth.

In order to isolate which of the components in the complex media were important, the media was deconstructed and the individual components tested in a minimal medium [M9]. The results confirm *S. glossinidius* heterotrophy, as *S. glossinidius* was only able to grow when provided with sufficient amino acid, carbon, nitrogen and vitamin sources (Figure 2.3). The nitrogen and vitamin source was shown to be important for *S. glossinidius*, as growth rates in the presence of yeast extract significantly increased for the tryptone and lactalbumin hydrolysate ± glucose variations (Figure 2.3). M9 supplemented with lactalbumin hydrolysate, yeast extract, the combination of both ± glucose, or yeast extract with tryptone supported statistically greater *S. glossinidius* growth than Schneider's Insect Medium (Figure 2.3), suggesting that using targeted growth media rather than a broad range complex medium is sufficient for *S. glossinidius in vitro* cultivation. The results also clearly show that *S. glossinidius* favours glucose, even when other carbohydrate sources are available (Figure 2.3).

S. glossinidius glucose utilisation was also supported by the carbon source screen, testing growth in a range of glucose, fructose, sucrose, mannose, raffinose, trehalose, maltose, *N*-acetylglucosamine, proline or mannitol concentrations in $M9^{g-}$ + yeast extract + lactalbumin hydrolysate; $M9^{g-}$ + yeast extract + lactalbumin hydrolysate with 20 or 25 mg/mL supplemented glucose aided in significantly higher growth values than $M9^{g-}$ + yeast extract + lactalbumin hydrolysate (Figure 2.4). The appearance of this significance only at higher concentrations links in with the earlier result of greater *S. glossinidius* growth in the comparatively glucose-rich MMI, than in Schneider's Insect Medium (Figure 2.2). All hypothesised [based on

available genetic data predictions] carbon sources theorised to support *S. glossinidius* growth were experimentally confirmed, excluding mannose (Figure 2.4). There are conflicting conclusions amongst studies regarding glucose or *N*-acetylglucosamine as the optimal *S. glossinidius* carbon source, but from the results presented here it would appear that at higher concentrations [20 mg/mL] there is no significant difference between the carbon sources and both aid in significantly greater growth than a negative control (Figure 2.4). The data presented in Figure 2.4 also suggests that the remaining intact *S. glossinidius* transporters are lenient to similarly-structured compounds as both raffinose and maltose aided in significantly higher growth levels than the control, but there are no reported intact specific transport systems for them (Toh *et al*., 2006).

The results presented here experimentally confirm the genome- and model-based predictions within the literature that *S. glossinidius* is able to utilise a range of carbon sources, displaying metabolic capabilities more associated with free-living bacteria. This ability is characteristic of its [theorised] current evolutionary placing; *S. glossinidius* has not fully established as obligate symbiont, so still retains its ancestral functional repertoire. The work presented here provides the opportunity to contextualise *in vitro* phenotypes *in vivo*, as studying the effects of targeted knock-out *S. glossinidius* mutants within the tsetse system could confirm any metabolic influence and thus *S. glossinidius* functionality within its host.

Chapter 3; Method Development for the Transformation of *Sodalis glossinidius*

3.1. Introduction

Reproducibility of experimental methods within scientific literature has always been considered a pillar of 'good science'. If the results from an experiment cannot be reproduced, whether by the same research group or by another, then the results are more likely to be considered unreliable. A survey of 1,576 researchers conducted by *Nature* showed than over 70% were unable to reproduce other research groups' results, in addition to greater than 50% not being able to reproduce their own (Baker, 2016). The reasons behind said percentages included several factors: participants agreed that the pressure to publish lead to less time to perform replications; hurdles such as differences within reagents being used or difficulty in repeating specialised or arduous protocols, more than once; the issue of selective reporting, such as researchers not including statistically insignificant method details/results that came from poor oversight during the experiments (Baker, 2016). Many of these problems are accentuated when the researchers are under time constraints, as optimising an experiment prior to conducting the 'final' version for publication can eliminate a vast amount of the experimental time, so one is forced to implement a cut-off point at which to cease optimisations for fear of compromising both the 'final' experiment and the subsequent analysis.

There is also the issue of when one is attempting to conduct novel experiments on a fastidious/under-researched organism. Where 'standard' protocols do not exist for an organism, researchers have to utilise protocols optimised for closely-related organisms and conduct multiple more optimisation experiments. Not only is this process time-consuming, but a vast amount of the time the optimisations could prove unsuccessful from the lack of full understanding for the organism. *Sodalis glossinidius* is a prime example of an under-researched organism, issues stemming from a massively degraded genome introducing issues in genetic manipulation, and a fastidious laboratory phenotype offering high contamination risks and time-consuming experiments. Examples of published

experiments involving the genetic manipulation of *S. glossinidius* include a recombinant mutant expressing a trypanolytic nanobody and a λ Red *S. glossinidius* mutant (Pontes and Dale, 2011; De Vooght *et al*., 2014). The λ Red *S. glossinidius* mutant was created via the Datsenko and Wanner (2000) system, which uses homologous recombination via the action of phage λ Red recombinase, utilising a helper and template plasmid (detailed in 3.2.2.). One of the main benefits of this system is the ability to target the integration of a gene of interest, not only allowing the user to create a specific gene knock-out mutant, but also allowing for freedom of selection for the marker (Datsenko and Wanner, 2000). This system offers ease-of-use, as the only adaptations that need to be introduced into the protocol are the amplification primer sequences in the first and second PCR of the pKD3 step(s), and the external primers used within the fusion PCR (Datsenko and Wanner, 2000). These studies show the *S. glossinidius* scope for genetic manipulation, but despite this, there is still no functional genomics data available for the fastidious symbiont. With this in mind, the development of established methods for transforming an organism and thus gaining this knowledge needed to be conducted, as there is a distinct lack of 'standard manipulation protocols' for *S. glossinidius* in the literature. One such example of an established transformation method for elucidating specific gene function in an organism is that of utilising transposons, which are short pieces of DNA that, through the action of a transposase enzyme, can randomly integrate into the chromosome (described in 1.2.4.). This method was adopted by Canals *et al*. (2019) to reveal the genomic functionality of *Salmonella* Typhimurium ST_{313} strain D23580 for survival and growth both in *in vitro* conditions and murine macrophages. The particular method used in this study was a final optimised version of their earlier method adaptation using the Epicentre® (Illumina, Inc.©) transformation protocol, which uses the 'cut and paste' Tn5 mechanism (described in 1.2.4.) (Canals *et al*., 2012; Canals *et al*., 2019). Regarding the Epicentre® standard issue protocol, it was released in the form of a kit and used *Escherichia coli* as the development template (Illumina, Inc.©). Briefly, the Epicentre® EZ-Tn5™ <KAN-2> Insertion Kit requires the transformation [electroporation] of target-organism competent cells with their reaction mixture [EZ-Tn5TM 10 X Reaction

Buffer, EZ-Tn5TM <KAN-2> Transposon, EZ-Tn5TM <KAN-2> Transposase, and water], following incubation steps involving their EZ-Tn5^{TM} 10 X Stop Solution (Illumina, Inc.©).

This chapter presents the initial *S. glossinidius* adaptation, development and then final optimisation of genetic transformation methods used for elucidating functionality within an organism, such as transposon transformation and knock-out mutants. The hypothesis was that by refining and/or reducing the number and/or intensity of the protocol steps, that genetic manipulation designed for 'hardy' organisms could be applied to a highly fastidious symbiont.

3.2. Methods

3.2.1. Transposon Transformation Using the Epicentre® EZ-Tn5™ <KAN-2> Insertion Kit

3.2.1.1. *Sodalis glossinidius*

The method used to transform *S. glossinidius* strain SgGMMB4 (GenBank accession no. LN854557) was adapted from that outlined in Canals *et al*. (2019). The final *S. glossinidius*-optimised protocol is as follows: prior to chilling 200 mL *S. glossinidius* mid-log phase Schneider's Insect Medium (Sigma-Aldrich) + 10% FBS (Gibco) culture (split across four lots of 50 mL falcon tubes) on ice, the OD of the culture was measured at A_{600nm} and a dilution series (from 10^{-0} to 10^{-6}) was spot-plated on horse blood (TCS Biosciences) Columbia agar (Sigma-Aldrich) plates. Following a 10 min ice-chill step, the 50 mL cultures underwent their first wash; cultures were centrifuged at 8000 rpm at 4°C for 10 min and the supernatant discarded, the pellets gently re-suspended in 1 mL of ice-cold, sterile 10% glycerol, the number of falcon tubes reduced by half by transferring two of the 1 mL re-suspensions into the remaining two and lastly, the final volumes made up to 40 mL using ice-cold sterile 10% glycerol (80 mL final volume split across two 40 mL falcon tubes). The second wash commenced in the same way as the first, including the halving of the re-suspended pellet volume; the two 1 mL pellet re-suspensions were made up to 40 mL using ice-cold sterile 10% glycerol and combined into one falcon tube. The third and final wash commenced in the same way as the first and second, except that the 1 mL initial re-suspension was not made up to 40 mL and instead transferred into a pre-chilled 1.5 mL Eppendorf tube (Eppendorf Ltd.) and then centrifuged at 8,000 rpm at 4°C for 10 min. The supernatant was removed and the pellet re-suspended in 400 µL of ice-cold, sterile 10% glycerol. The 400 µL suspension was then aliquoted into eight lots of 45 µL and one 40 µL (for use as the negative control) in pre-chilled, 2 mm gap electroporation cuvettes (Geneflow©). For the production of Epicentre® EZ-Tn5™ <KAN-2> Transposomes (Illumina, Inc.©) *in vitro*, both the transposome reaction and negative control mixtures were prepared by adding the reagents listed in Table 3.1, in order.

Table 3.1. List of components to add (in order) for the creation of the transposome reaction mixture and the negative control reaction mixture. The mixtures originate from the Canals *et al*. (2019) protocol, which is an adaptation from the Epicentre® EZ-Tn5™ <KAN-2> Insertion Kit (Illumina, Inc.©).

Transposome reaction mixture components	Negative control reaction mixture components
1 µL of glycerol	1 µL of glycerol
1 µL of TypeOne™ Restriction Inhibitor	5 µL of dH_2O
2 µL of the EZ-Tn5<KAN-2> Transposon (Illumina, Inc.©)	
2 µL of EZ-Tn5 Transposase (Illumina, Inc.©)	

The mixtures were mixed by pipetting the solution up and down, briefly spun in a microfuge and then incubated for 2 hours at room temperature. A petri dish approximately half filled with water was used as the base for the dialysis step: 0.025 µm MF-Millipore™ Membrane Filter paper was carefully placed onto the water's surface using forceps, ensuring that no water came into contact with the top of the paper, and two lots of 3 µL (post-incubation) transposome reaction mixture was pipetted at polar ends of the paper. A second 0.025 µm MF-Millipore™ Membrane Filter paper was also placed into the same petri dish for the negative control reaction mixture to undergo dialysis in the same way. After 30 min of dialysis, the transposome mixture was carefully aspirated from the paper and transferred into a 1.5 mL Eppendorf tube, in addition to the negative control reaction mixture into a second 1.5 mL Eppendorf tube. A pipette was used to estimate the new, post-dialysis transposome mixture volume and this was divided equally between the electroporation cuvettes. The same calculated volume was applied to the negative control reaction mixture and the equivalent dH_2O volume added to the 40 µL negative control electroporation cuvette. The electroporation cuvettes were chilled on ice for 10 min and then transformed individually using a Bio-Rad© Gene Pulser™ by electroporation (2.5 kV, 25 µFD, 200 Ω). Immediately after each transformation, the samples were re-suspended in 1 mL room-temperature Schneider's Insect Medium + 10% FBS and transferred to 1.5 mL Eppendorf tubes. 20 µL from each recovery suspension, including the negative

control, were spot-plated on horse blood Columbia agar plates both with and without 50 µg/mL kanamycin. In addition to this, dilution series' (from 10^{-0} to 10^{-6}) from three randomly selected, transposon-positive recovery suspensions were spot-plated on horse blood Columbia agar plates with 50 µg/mL kanamycin. After a 1 hour [statically at 25°C] incubation time, 20 µL from each post-incubation suspension, including the negative control, were spot-plated on horse blood Columbia agar plates with 50 µg/mL kanamycin. In addition to this, dilution series' (from 10^{-0} to 10^{-6}) from the same initial three randomly selected, transposon-positive recovery suspensions, now post-incubation suspensions, were spot-plated on horse blood Columbia agar plates with 50 µg/mL kanamycin. This allowed for representative enumeration of the number of transposon mutants, in triplicate. All the post-incubation suspensions were transferred into individual 15 mL falcon tubes with 9 mL Schneider's Insect Medium + 10% FBS with 50 µg/mL kanamycin and allowed to grow to mid-log phase (3 – 5 days) statically at 25°C. A separate 9 mL Schneider's Insect + 10% FBS culture seeded with 1 mL mid-log phase *S. glossinidius* culture was set up at the same time, to act as a positive control for growth rate. OD A_{600nm} was measured for all of the cultures upon reaching mid-log phase, which were then stocked at -80°C in individual 1 mL aliquots of 600 µL + 400 µL 50% glycerol. Once visible colonies were present on all of the plates used during the experiment (after approximately 7 – 10 days), the estimated number of mutants within the library could be calculated: if the number of colonies on the pre-1-hour-incubation plates are half of the total on the post-1-hour-incubation plates, then the cells have doubled and the total number of mutants assumed within the library (as enumerated by the post-incubation-1-hour plates) needs to be divided by two.

Based on the biology behind the Canals *et al.* (2019) *S.* Typhimurium ST_{313} strain D23580-adaptions on the Epicentre® EZ-Tn5™ <KAN-2> Insertion Kit (Illumina, Inc.©) standard issue protocol, some of the steps were immediately changed for use on the vastly more fastidious *S. glossinidius*: the medium used was Schneider's Insect Medium + 10% FBS instead of SOC medium; the plates used were horse blood Columbia agar [instead of LB] plates and were incubated for a 5 – 10

days under microaerophilic conditions at 25°C; the recovery cultures were statically incubated for 1 hour at 25°C, instead of 45 – 60 min on a shaker at 37°C; the use of plates to collect mutants was replaced with liquid culture, whereby plates were only used for enumeration of mutant numbers; the final Schneider's Insect Medium + 10% FBS 400 mL mutant pool culture was statically incubated for 3 – 5 days at 25°C instead of incubating on a shaker at 37°C overnight. Regarding replacing the use of plates to collect mutants to liquid culture: firstly, the fastidious nature of *S. glossinidius* results in lengthy time scales for colony formation on a plate, often allowing for contamination to halt experimental progress. Secondly, observations from multiple experiments involving the growth of *S. glossinidius* both on plates and in liquid culture strongly suggests a phenotypic preference for liquid suspension; the most likely reason for this being the vast amounts of bioavailable nutrients available in the Schneider's Insect Medium + 10% FBS liquid culture versus the solid medium.

Other optimisations included: electroporation assays to ascertain voltage strength on cell survival; kanamycin assays to detect any effect of varying concentration on growth rate.

Electroporation assay: mid-log phase *S. glossinidius* cultures in Schneider's Insect Medium + 10% FBS underwent washes in order to create electrocompetent cells, as outlined in the *S. glossinidius*-optimised transformation protocol above. However, half of the culture volume used was kept at room temperature throughout the experiment (as opposed to the cultures being chilled throughout). 50 µL aliquots were electroporated with 1 µL (chilled, in the case of half of the aliquots) glycerol, using 25 µFD and 200 Ω, but under five different voltage intensities: 0 kV, 1 kV, 1.5 kV, 2 kV or 2.5 kV. 1 mL Schneider's Insect Medium + 10% FBS was added to the post-electroporation suspensions and dilution series' were created, which were plated on horse blood Columbia agar plates and grown under microaerophilic conditions for 5 – 10 days.

Kanamycin assay: dilution series (10^0 – 10^{-6}) of mid-log phase Schneider's Insect Medium + 10% FBS *S. glossinidius* cultures were plated on a range of kanamycin concentrations (0 µg/mL, 2 µg/mL, 5 µg/mL, 10 µg/mL, 25 µg/mL or 50 µg/mL). Further experiments were subsequently conducted to determine *S. glossinidius* kanamycin sensitivity in liquid culture; 100 µL mid-log phase *S. glossinidius* Schneider's Insect Medium + 10% FBS culture was passaged into 10 mL Schneider's Insect Medium + 10% FBS containing a range of kanamycin concentrations (0 µg/mL, 2 µg/mL, 5 µg/mL, 10 µg/mL, 25 µg/mL or 50 µg/mL). OD A_{600nm} was measured on day 3 and 4.

A protocol efficacy experiment was conducted using the same method described above [3.2.1.1.], except for using a green fluorescent protein (GFP) named pZEP17.1 extracted from either *E. coli* strain K-12 substrain MG1655 or transformed *S. glossinidius* strain SgGMMB4 [previously-transformed using *E. coli* strain K-12 substrain MG1655 pZEP17.1 GFP] in place of the transposon.

3.2.1.2. *Sodalis praecaptivus*

The adapted Epicentre® EZ-Tn5™ <KAN-2> Insertion Kit (Illumina, Inc.©) method used for transformation outlined in Canals *et al.* (2019) was also tested on *S. praecaptivus*. Optimisations included the electroporation and kanamycin assays detailed in 3.2.1.1., in addition to a voltage-antibiotic transformation assay using the Epicentre® EZ-Tn5™ <KAN-2> Insertion Kit (Illumina, Inc.©). *S. praecaptivus* adaptations for the electroporation and kanamycin assays used on *S. glossinidius* [3.2.1.1.] included the use of LB agar and incubations at 37°C for 1 – 2 days.

Voltage-antibiotic transformation assay using the Epicentre® EZ-Tn5™ <KAN-2> Insertion Kit (Illumina, Inc.©): two-day *S. praecaptivus* LB cultures were washed to create electrocompetent cells, as outlined in Canals *et al*. (2019). These were transformed with the Epicentre® EZ-Tn5™ <KAN-2> Insertion Kit (Illumina, Inc.©) using the protocol outlined in Canals *et al*. (2019), save for the following adaptations: the electrocompetent cells were split into aliquots and transformed at 0 kV, 1 kV, 1.5 kV, 2 kV or 2.5 kV, followed by being plated on LB agar with either 0

μg/mL, 10 μg/mL or 20 μg/mL kanamycin. These variants were conducted in triplicate, resulting in 231 plates (not including the negative controls). The negative controls used glycerol instead of the transposon, were transformed at 2.5 kV and plated on LB agar plates with 0 μg/mL kanamycin.

3.2.2. λ Red Homologous Recombination of *Sodalis glossinidius* Using the pKD46-pKD3 System

For the Datsenko and Wanner (2000) system, homologous recombination occurs via the action of phage λ Red recombinase utilising a helper and template plasmid. The helper plasmid, otherwise known as pKD46 (Figure 3.1 A), contains the three λ Red genes that are required for homologous recombination of the desired DNA into the target sequence: γ encodes the product Gam, which inhibits host RecBCD exonuclease V (degrades DNA), so that Bet and Exo can promote recombination; β encodes the single-stranded DNA (ssDNA) binding protein Bet, which anneals complementary ssDNA strands during the homologous recombination process; *exo* encodes the exonuclease Exo, which acts on dsDNA during the homologous recombination process (Murphy, 1998; Datsenko and Wanner, 2000). The three aforementioned λ Red genes are expressed under an arabinose-inducible P_{araB} promotor (Datsenko and Wanner, 2000). pKD46 is transformed into the organism-of-interest and should be maintained under ampicillin selection (pKD46 encodes for ampicillin resistance) (Datsenko and Wanner, 2000).

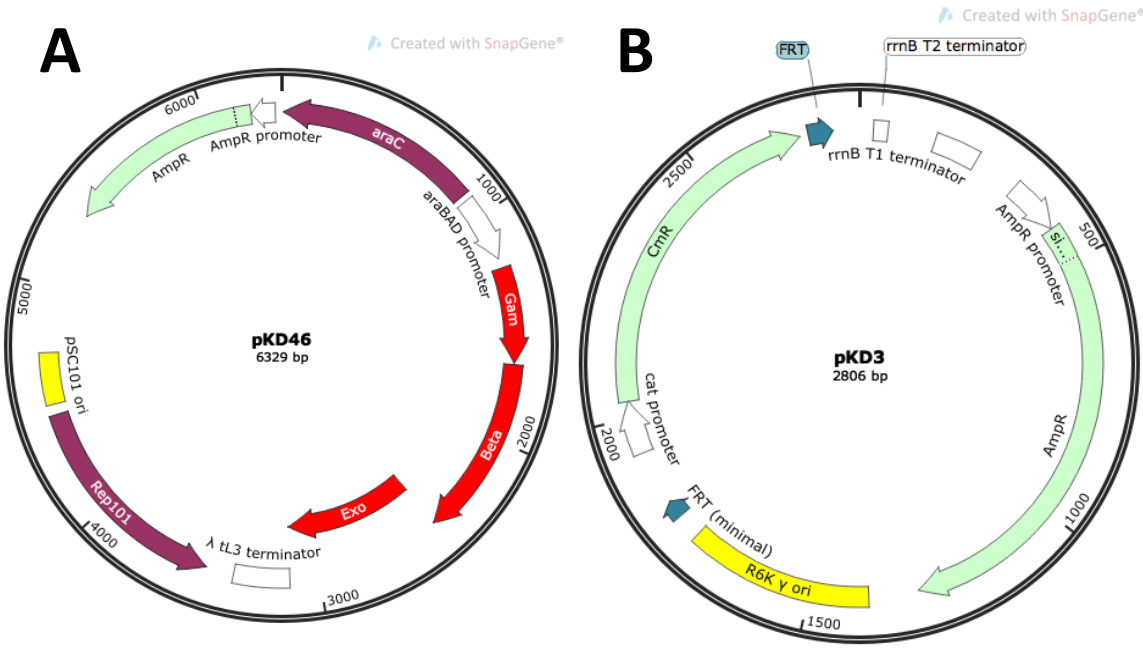


Figure 3.1. Graphical representations of the pKD46 'helper plasmid' (AY048746.1) [A] and the pKD3 'template plasmid' (AY048742.1) [B]. A: araBAD promotor (arabinose-inducible promoter) controls the expression of 3 λ Red genes; γ (produces Gam, which inhibits host RecBCD exonuclease V, so that Bet and Exo can promote recombination), β (produces Bet) and *exo* (produces Exo). AmpR (ampicillin resistance) is the selection for pKD46. **B:** FRT (flippase recognition targets) regions are shown in grey, flanking CmR (chloramphenicol acetyltransferase). The CmR sequence was edited manually as the published sequence (AY048742.1) contains errors meaning the chloramphenicol resistance cassette isn't annotated. **Made using SnapGene®. System created by Datsenko & Wanner (2000).**

The template plasmid, otherwise known as pKD3 (Figure 3.1 B), is the component within the Datsenko and Wanner (2000) helper-template plasmid system of recombination that contains the desired DNA for integration into the target sequence. PCR is used in order to create three fragments from pKD3: two of the fragments, each approximately 1 Kb in size, are created from primers that amplify sequence surrounding the location for integration of the pKD3-derived chloramphenicol resistance cassette; the third fragment, approximately 1.2 Kb in size, is created from primers that contain two priming sites targeting the outer-edges of the pKD3 chloramphenicol resistance cassette, in addition to 36 nucleotide homology extensions complementary to the ends of the first and second fragments (Datsenko and Wanner, 2000). The three fragments undergo a fusion PCR in order

to create one approximately 3.2 Kb linear DNA fragment (using external primers that amplify into the combined fragment), 10 – 100 ng of which is transformed into 25 µL electrocompetent (i.e. washed in 10% glycerol) pKD46-carrying bacterial aliquots via electroporation (Datsenko and Wanner, 2000). The transformed aliquots are re-suspended into 1 mL of appropriate liquid media and incubated at 37°C for one hour (Datsenko and Wanner, 2000). Half of the post-incubation cultures are plated onto appropriate agar medium plates supplemented with chloramphenicol and the remaining volumes left to stand overnight (Datsenko and Wanner, 2000). If no chloramphenicol resistant mutants develop on the plates within 24 hours, the overnight standing cultures can be spread onto appropriate agar medium plates supplemented with chloramphenicol (Datsenko and Wanner, 2000).

Following the Datsenko and Wanner (2000) system in conjunction with the protocol published by Pontes and Dale (2011), λ Red homologous recombination on *S. glossinidius* strain SgGMMB4 was attempted. The *S. glossinidius* strain SgGMMB4 genomic region selected to insert the chloramphenicol resistance cassette was chosen due to its overall hypothetical nature and the relative lack of coding sequences both up- and down-stream (Figure 3.2). Primer sets were designed (Table 3.2) to amplify genes PROKKA_03540 and PROKKA_03541 with ~1 kB overlaps into the rest of the genome sequence (Figure 3.4).

Figure 3.2. Screenshot of genomic region chosen for cassette insertion, via Artemis (Wellcome Sanger Institute). Highlighted in red boxes: *S. glossinidius* strain SgGMMB4 genes chosen for chloramphenicol cassette insertion by λ Red recombination, PROKKA_03540 and PROKKA_03541.

An overview of the methods tested and optimisations conducted throughout 3.2.2. and 3.3.2. are displayed in Figure 3.3.

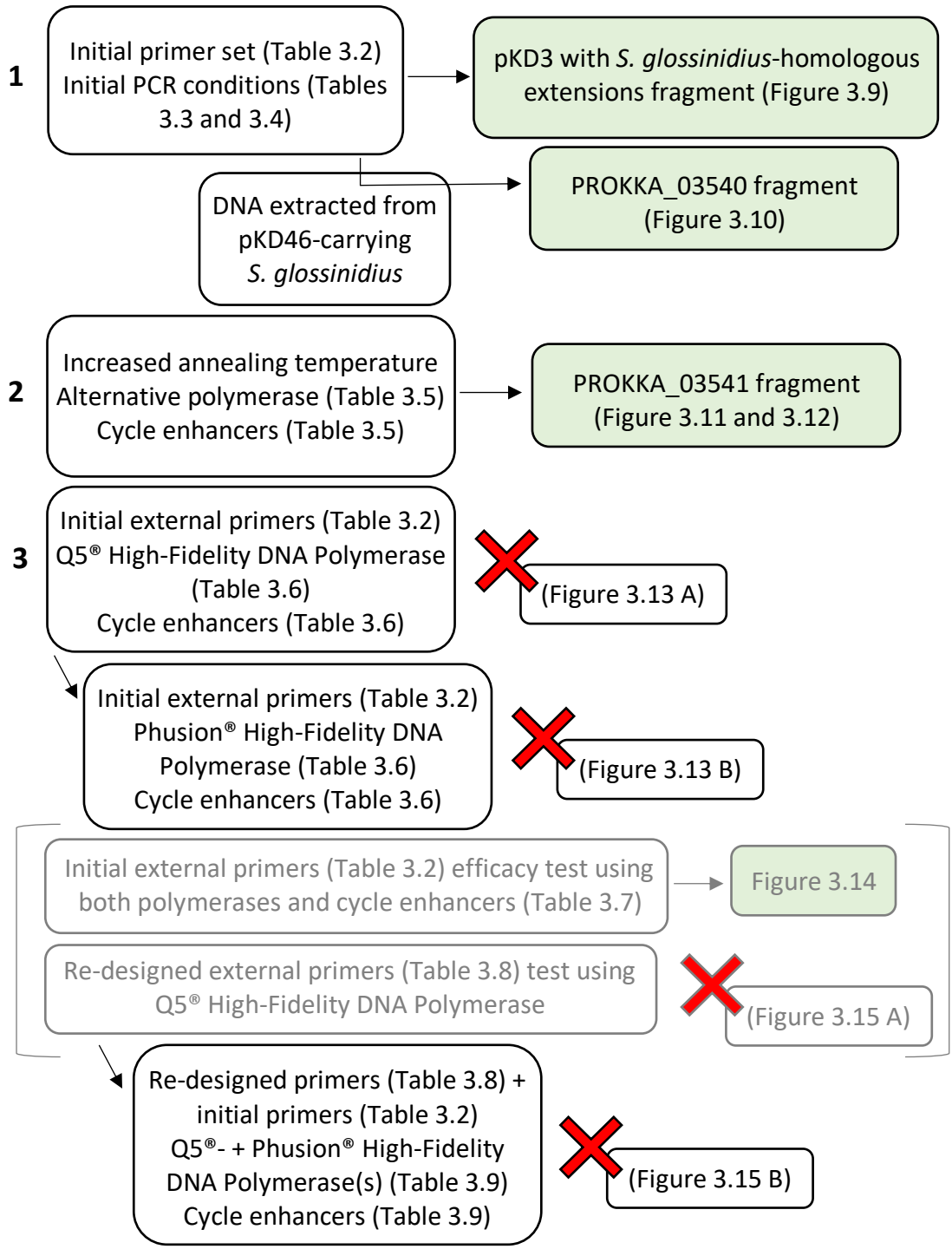


Figure 3.3. Graphical overview of the methods tested and optimisations conducted throughout 3.2.2. and 3.3.2. (1.) initial experiments conducted to create the fragments **(2.)** optimisations conducted when (1.) did not produce products **(3.)** various optimisations conducted for the creation of the final fused product (i.e. *Sodalis*-specific pKD3 fragment). Green indicates successful fragment creation.

Table 3.2. Table of the initial *Sodalis glossinidius* strain SgGMMB4 λ Red recombination primers. The first column lists the primer descriptions, where the second column lists the associated primer sequence. Underlined sequence refers to the pKD3 priming sites: GTGTAGGCTGGAGCTGCTTC is the first priming site and is downstream of the pKD3 chloramphenicol resistance cassette; CATATGAATATCCTCCTTAG is the second priming site and is upstream of the pKD3 chloramphenicol resistance cassette.

Fragment/gene description	**Primer sequence (5' – 3')**
External forward primer	AGCAGGAGCAATTGCACCCG
PROKKA_03540 reverse primer	CATATACCCGCTGTGCAACTTC
pKD3 with *S. glossinidius* strain SgGMMB4 sequence homology forward primer	AAGTTGCACAGCGGGTATATGACGCGATACCGGCCTGGCAATTGAGCATATGAATATCCTCCTTAG
pKD3 with *S. glossinidius* strain SgGMMB4 sequence homology reverse primer	GACGCCGATGCTTGGCGTAAGGGTACCGGTAGGACTGGCCTGATGGTGTAGGCTGGAGCTGCTTC
PROKKA_03541 forward primer	CTTACGCCAAGCATCGGCGTC
External reverse primer	CTGAATGGCTGGGGCGTTGGCG

TACAGCAGGAGCAATTGCACCCGCTTTTCACCGCAAAAATGCGCGACCTGCCTCAGAACCGAGGTATTTTCCT
CTATAACGCCGCGGGACAGATGGTCGCTACCTCCGACACTGCGCTGCCGACTAATACCGATTTTGCCGATCGC
GACTATTACCGCTACCATCGCGACCACAATGACAGTAGTTTGCATATTGGCAAAGTGATCCGCAACCGGCTCT
CCGGCGGGCTCATTTTGCCGGTTTCGCTCCGTATTAATCATCCCGATAGCAGCTTTGCGGGAGTGCTTTTGGAA
AGCGTGGTTATCGATTATTTCCGCCACTTTTACGCCCGCTTCGTGATAGACGACGACTCCACGTTAATGATGAT
GCTCAACAATGGCACCGTTTTGTATCGTCACCCTTACAACGAGCGGGCAATCGGCAGTTCCATTTCAGCCTCGC
CGCTCTATCAGCAACATATCGGCCAATAAACCTCAGGCGTCATGACGGGCTCCCTCAGCAGCCAGGATCCGCA
GAAAATCTACAGCTTCACCCATTTGCAGCAGTTTCCGGTGGCGATCACGACCGCTATGTCGTTGAATCAGGCA
CTGGTGGATTGGCGTCAGGATGCGATAAGTCATGTGATCCTGGTCGTTATTTTTCTGGCGCTGCTCAGCCTGC
TGGCTATCTCGTTTATCCGCCAGATACACCGCCGCCGCATGCGTGTGGAAGAAGAGTTACGCTACGCCCAGCG
TGAGCTACGCAAGTTGAACCAATCACTGGAACATTGGCGCGCCGTGATGACCTCACCGGCCTGTATAATCGCC
GCCATTTCGATTTGGCCCTTACCGATGAGTTCAACCGTGCGGTAGTTAGCCAGGATTCACTGGGTATAATCTTG
CTGGACATCGACTTTTTCAAATAATATAACGATCTCTACGGCCACGTCGCCGGCGATGATTGCCTGAAACGCAT
TGGCGAAGCGCTGAAAGCCCTTCCATTGCGACACCGGATCAGGTGGCGCGCTACGGTGGCGAGGAGTTTATT
ATTCTGCTACCGCAAACCGATGCGGCAGGCGCGCGGGAAGTTGCACAGCGGGTATATGACGCGATACCGGCC
TGGCAATTGAGCATCAGGCCAGTCCTACCGGTACCCTTACGCCAAGCATCGGCGTCTATGTGGGGCAGCCGG
CGACCCGTGACGACTCTCCGTCGAGGATGGTCAGCCGGGCCGATACTGCTCTCTATCAGGCCAAGGGCGAGG
GACGAAACCGCATTTGCATCGCCTGACACGGAGCGCCGAGCTTCTCTGCCAGCGCTAACCCGTCAGGATGCGT
ATCCACCACCAGCGCTTACGGTGGGCCCTTGATGACCGCCGCATGTGCCGCCATTAACCCGACGGGACCGGC
GCCATAGTTCGCCACGCTATCGCCGGGTCGCAAACCCGCCAGTTCGGTGACATGCCAACCGGTCGGGGAGAT
GTCGGACAGCATAACATCGCCATCTTCTTTCTCCGTGGCGTCCCTTAGCAGCACAAGACAGTTGGAATCCTCAT
AGGGAACGCAAAGTCATTCCGCCTGCCCCCATCCCACGGCTCCATTTCGACAAAGCCATAGGCCGTCCCCGGC
GAAGGAGAATACGTTCATGGGCTGATCATTAGTGGCAACACCCAGCCGAACGCTAAAATCAGCTGGGCGTTT
TCCGTCAGTTAAGACCTGCGCTTAGCGCTGGGTATCGACAATGATACGACCCCGTATTTGATCCGCCATCAGC
GCTTTTGCGGTGTCTATCGCCTCTGACAGCCCGATTTCGCGGCTTATCCGTTCCAGTAGCGGCTCATCGACGAG
TTCAGCAATACGCTGCCAGGCGGCGCGACGATCCGCCTCGGGCCGCATGACGCTGTCGATACCTACCAGCGTT
ATGCCACGCAAAATAAACGGCGCGACGCTTGCGGGAAACTCCATGCCCTAGGCCAGACCACAGGCCGTCACG
ACGCCCTGATAACATAATCCGGCGCAAACATTGGCCAGCGTATGGCTCCCGACGCTGTCAACGGCCGCCGCCC
ATTGCCGAGTGGGCATCCAGTCCGGCAGGCAGCGGGATTAGCCAGTCGCCGTTGACGCGGGCTTTTGCGCCA
GCCCGCCCCAGTGCTTTTCGCCAACGCCCCAGCCATTCAGCACAACCCGATCGCCGGCGCGAAAACCGGCATG
GCGACTGTGTTCTAACGTACCGGCTAAATCGATTCCAGGTACCATCGGGAAGTGGTGCACCACCGGCCCTTCT

Figure 3.4. Graphical representation of the *Sodalis glossinidius* strain SgGMMB4 genomic region chosen for primer design. The sequence highlighted in yellow is the gene PROKKA_03540 (encodes hypothetical protein) and highlighted in green is PROKKA_03541 (encodes hypothetical protein). The red sequence is the external forward primer, the brown underlined sequence is the PROKKA_03540 reverse primer location, the blue underlined sequence is the PROKKA_03541 forward primer and the black underlined sequence is the external reverse primer location.

Two mid-log phase *S. glossinidius* strain SgGMMB4 Schneider's Insect Medium (Sigma-Aldrich) + 10% FBS (Gibco) cultures, one that separately underwent a 10 min 35°C heat shock step, were made electrocompetent by pelleting and washing with 10% glycerol under chilled conditions. Both suspensions of electrocompetent cells were corrected to be at an OD A_{600nm} value of 50 then transformed with pKD46 (AY048746.1) at a final concentration of 500 ng/μL. The transformed cells were re-suspended in 1 mL Schneider's Insect Medium (Sigma-Aldrich) + 10% FBS (Gibco) and underwent a two-hour rest period, then transferred into 9 mL Schneider's Insect Medium (Sigma-Aldrich) + 10% FBS (Gibco) at an ampicillin concentration of 75 μg/mL. The 10 mL cultures were statically incubated at 25°C. Once mid-log phase was reached, the cultures were glycerol-stocked at -80°C. Three (plus one negative control) separate PCR experiments were used to create the three *S. glossinidius*-gene-specific fragments from pKD3: two corresponding to the two chosen genes (~1 Kb) and then the third corresponding to pKD3 (AY048742.1) with *S. glossinidius*-homologous extensions (~1.2 Kb). The PCR cycle used was as follows: 30 second denaturing step at 98°C, 35 cycles of denaturing-annealing-elongation steps of 10 seconds at 98°C-30 seconds at 55°C-60 seconds at 72°C, followed by a final elongation step of five min at 72°C. The PCR mix used is outlined in Table 3.3, where "Primer A", "Primer B" and "DNA" were adapted for each individual fragment (Table 3.4).

Table 3.3. A table outlining the list of components used within the master mix for a PCR reaction creating *Sodalis glossinidius*-specific fragments from pKD3. The first column lists the components used within the master mix. The second column lists the volumes that would be added into one sample.

Master Mix Component	Volume per Sample (μL)
DNA	2
Primer A (10 μM)	5
Primer B (10 μM)	5
dNTP (10 μM)	2
5 X Q5® Reaction Buffer	20
Q5® High-Fidelity DNA Polymerase	1
H_2O	65

Table 3.4. Table of which PCR component was used for the creation of which fragment. The first column lists the fragment to be created. The second column lists which forward primer was used to create the equivalent fragment. The third column lists which reverse primer was used to create the equivalent fragment. The fourth column lists which DNA input was used for which fragment. See Table 3.2 for the sequences of the primers listed here.

Fragment	Primer A	Primer B	DNA
PROKKA_03540	external forward primer	PROKKA_03540 reverse primer	*S. glossinidius* strain SgGMMB4 colony in 500 µL Schneider's Insect Medium + 10% FBS
PROKKA_03541	PROKKA_03541 forward primer	external reverse primer	*S. glossinidius* strain SgGMMB4 colony in 500 µL Schneider's Insect Medium + 10% FBS
pKD3 with *S. glossinidius*-homologous extensions	pKD3 with *S. glossinidius* sequence homology forward primer	pKD3 with *S. glossinidius* sequence homology reverse primer	pKD3
Negative control	PROKKA_03541 forward primer	external reverse primer	H_2O

The amplicons were run on a 1% agarose gel (Figure 3.9) and the pKD3 with *S. glossinidius*-homologous extensions fragment was extracted from the gel using the BioLine Isolate II PCR and Gel Kit. DNA was extracted using the Zymo Quick-DNA™ Miniprep Plus Kit. Repeats for the creation of the PROKKA_03540 and PROKKA_03541 fragments were conducted, but DNA extracted from pKD46-carrying *S. glossinidius* was used instead of an *S. glossinidius* colony in 500 µL Schneider's Insect Medium + 10% FBS. The PROKKA_03540 fragment was successfully amplified and the amplicon product (Figure 3.10) was extracted from a 1% agarose gel using the BioLine Isolate II PCR and Gel Kit and DNA was extracted using the Zymo Quick-DNA™ Miniprep Plus Kit.

As the successful amplification of the PROKKA_03541 fragment was not achieved, optimisation involved increasing the annealing temperature within the PCR cycle from 55°C to 60°C. The amplicons were run on a 1% agarose gel (Figure 3.11). Phusion® High-Fidelity DNA Polymerase (New England BioLabs®) was tested as an alternative polymerase, in conjunction with cycle amplifiers including DMSO, betaine and a combination of both as PCR reaction mix variants (Table 3.5), using the methods previously described. These amplifiers break down any secondary structures forming within the DNA that could block the polymerase from amplifying the strand(s).

Table 3.5. Table of DMSO, betaine, a combination of both and different enzyme PCR reaction mix variants. The first column lists the sample number/group name associated with the components listed in column two. The second column lists the PCR components. The third column lists the volumes of the components used for each sample/group listed in column one. See Table 3.2 for the sequences of the primers listed here.

Sample number	PCR reaction mix	Volume per Sample (µL)
Q5® mix	pKD46-carrying *S. glossinidius* DNA	2
	PROKKA_03541 forward primer (10 µM)	5
	External reverse primer (10 µM)	5
	dNTP (10 µM)	2
	5 X Q5® Reaction Buffer	20
	Q5® High-Fidelity DNA Polymerase	1
Phusion® mix	pKD46-carrying *S. glossinidius* DNA	2
	PROKKA_03541 forward primer (10 µM)	5
	External reverse primer (10 µM)	5
	dNTP (10 µM)	2
	Phusion® Master (Reaction Buffer and Polymerase)	50
1	Q5® mix	
	H_2O	60
	DMSO	5
2	Q5® mix	
	H_2O	45
	Betaine	20
3	Q5® mix	
	H_2O	40
	DMSO	5
	Betaine	20
4	Q5® mix	
	H_2O	65
5	Phusion® mix	
	H_2O	31
	DMSO	5
6	Phusion® mix	
	H_2O	16
	Betaine	20
7	Phusion® mix	
	H_2O	11
	DMSO	5
	Betaine	20
8 (negative control)	Same as sample 4, but with H_2O instead of DNA	

Successful samples 3 and 7 amplicon products (Figure 3.12) were excised from a 1% agarose gel using the BioLine Isolate II PCR and Gel Kit and DNA was extracted from sample 3 using the Zymo Quick-DNA™ Miniprep Plus Kit. Overlap fusion PCR was conducted on the three fragments using an initial PCR cycle with no primers (Table 3.6), the products of which were immediately put into a second PCR cycle with the addition of 5 µL of each of the two external primers (Table 3.2) and an additional 0.5 µL of the Q5® High-Fidelity DNA Polymerase (New England BioLabs®). The cycle used for the initial PCR was as follows: 30 second denaturing step at 98°C, 10 cycles of denaturing-annealing-elongation steps of 10 seconds at 98°C-30 seconds at 50°C-60 seconds at 72°C, followed by a final elongation step of five min at 72°C. The cycle used for the second PCR step was the same, except for: 25 cycles of denaturing-annealing-elongation steps of 10 seconds at 98°C-30 seconds at 60°C-two min at 72°C. The amplicons were run on a 1% agarose gel (Figure 3.13 A). Repeats were conducted using Phusion® High-Fidelity DNA Polymerase (New England BioLabs®). The amplicons were run on a 1% agarose gel (Figure 3.13 B).

Table 3.6. Table listing the components used in the fusion PCR. The first column lists the sample number/group name associated with the components listed in column two. The second column lists the PCR components. The third column lists the volumes of the components used for each sample/group listed in column one. For "1.[x]" samples, see Figure 3.13 A. For "2.[x]" samples, see Figure 3.13 B.

Sample number	PCR reaction mix	Volume per Sample (µL)
	PROKKA_03540 fragment	x for 100 ng
	pKD3 with *S. glossinidius*-homologous extensions fragment	x for 120 ng
	PROKKA_03541 fragment	x for 100 ng
	dNTP (10 µM)	2
	H_2O	x up to 100
Q5® mix	5 X Q5® Reaction Buffer	20
	Q5® High-Fidelity DNA Polymerase	1
Phusion® mix	Phusion® Master (Reaction Buffer and Polymerase)	50
	DMSO	5
	Betaine	20
1.1	Q5® mix	
1.2 & 1.3	Q5® mix	
	DMSO	5
	Betaine	20
1.4 (negative control)	Same as sample 1.1, but with 6 µL H_2O instead of the fragments	
2.1	Phusion® mix	
2.2	Same as sample 2.1, but with H_2O instead of the fragments	

Following the unsuccessful fusion PCRs, positive controls were tested in order to confirm the efficacy of the external primers, whereby DNA extracted from pKD46-carrying *S. glossinidius* underwent a PCR with the components listed in Table 3.7. The PCR cycle used was as follows: 30 second denaturing step at 98°C, 25 cycles of denaturing-annealing-elongation steps of 10 seconds at 98°C-30 seconds at 60°C-two min at 72°C, followed by a final elongation step of five min at 72°C. The amplicons were run on a 1% agarose gel (Figure 3.14).

Table 3.7. Table listing the components used in primer efficacy test PCR. The first column lists the sample number/group name associated with the components listed in column two. The second column lists the PCR components. The third column lists the volumes of the components used for each sample/group listed in column one. See Table 3.2 for the sequences of the primers listed here.

Sample number	PCR reaction mix	Volume per Sample (µL)
Q5® mix	pKD46-carrying *S. glossinidius* DNA	2
	External forward primer (10 µM)	5
	External reverse primer (10 µM)	5
	dNTP (10 µM)	2
	5 X Q5® Reaction Buffer	20
	Q5® High-Fidelity DNA Polymerase	1
Phusion® mix	pKD46-carrying *S. glossinidius* DNA	2
	External forward primer (10 µM)	5
	External reverse primer (10 µM)	5
	dNTP (10 µM)	2
	Phusion® Master (Reaction Buffer and Polymerase)	50
1	Q5® mix	
	H_2O	65
2	Q5® mix	
	Betaine	20
	H_2O	40
	DMSO	5
3	Phusion® mix	
	H_2O	36
4	Phusion® mix	
	H_2O	11
	DMSO	5
	Betaine	20
5 (negative control)	Same as sample 1, but with H_2O instead of DNA	

The initial primer sets (Table 3.2) designed to amplify genes PROKKA_03540 and PROKKA_03541 with ~1 kB overlaps into the rest of the genome sequence were re-designed (Table 3.8) to include longer overhangs, in addition to removing the potential hairpin within the original PROKKA_03541 forward primer.

Table 3.8. Table of the re-designed *Sodalis glossinidius* λ Red recombination primers. The first column lists the primer descriptions, where the second column lists the associated primer sequence. Underlined sequence refers to the pKD3 priming site: GTGTAGGCTGGAGCTGCTTC is the first priming site and is downstream of the pKD3 chloramphenicol resistance cassette.

Fragment/gene description	**Primer sequence (5' – 3')**
pKD3 with *S. glossinidius* sequence homology reverse primer	ACATAGACGCCGATGCTTGGCGTAAGGGTACCGGTAGGACTGGCCTGATGGTGTAGGCTGGAGCTGCTTC
PROKKA_03541 forward primer	GCCAAGCATCGGCGTCTATGT
External reverse primer	CGTTAGAACACAGTCGCCATG

These re-designed primers underwent PCR as previously described [using Q5® High-Fidelity DNA Polymerase (New England BioLabs®)] and the results visualised on a 1% agarose gel (Figure 3.15 A). Subsequent optimisation experiments involved the following: DMSO was and was not added to each of the fragment-creation samples, using both original [Table 3.2] and re-designed [Table 3.8] primer sets (Table 3.9). This was in addition to using DNA extracted from a fresh *S. glossinidius* Schneider's Insect Medium + 10% FBS mid-log phase culture. The amplicons were run on a 1% agarose gel (Figure 3.15 B).

Table 3.9. Table listing the components used in original and re-designed primer PCR test. The first column lists the sample number/group name associated with the components listed in column two. The second column lists the PCR components. The third column lists the volumes of the components used for each sample/group listed in column one. See Tables 3.2 and 3.8 for the sequences of the primers listed here.

Sample number	PCR reaction mix	Volume per Sample (µL)
Standard mix	Phusion® HF Buffer	100
	dNTP (10 µM)	10
	H_2O	325
	Phusion® High-Fidelity DNA Polymerase	1
DMSO mix	Phusion® HF Buffer	80
	dNTP (10 µM)	8
	H_2O	248
	DMSO	12
	Phusion® High-Fidelity DNA Polymerase	1
1	Standard mix	87
	pKD3 with *S. glossinidius* sequence homology forward primer (Table 3.2)	5
	pKD3 with *S. glossinidius* sequence homology reverse primer (Table 3.2)	5
	pKD3 DNA	2
2	Same as 1, but with DMSO mix	
3	Standard mix	87
	PROKKA_03541 forward primer (Table 3.8)	5
	External reverse primer (Table 3.8)	5
	S. glossinidius strain SgGMMB4 DNA	2
4	Same as 3, but with DMSO mix	
5	Standard mix	87
	External forward primer (Table 3.2)	5
	External reverse primer (Table 3.8)	5
	S. glossinidius strain SgGMMB4 DNA	2
6	Same as 5, but with DMSO mix	
7	Standard mix	87
	External forward primer (Table 3.2)	5
	External reverse primer (Table 3.2)	5
	S. glossinidius strain SgGMMB4 DNA	2
8	Same as 7, but with DMSO mix	
9 (negative control)	Same as 7, but with 2 µL of H_2O instead of *S. glossinidius* strain SgGMMB4 DNA	

3.3. Results

3.3.1. Transposon Transformation Using the Epicentre® EZ-Tn5™ <KAN-2> Insertion Kit

The results from the [multiple] electroporation assays, aimed at optimising a genetic transformation method for specific use on *S. glossinidius,* were unsuccessful: the plates were either blank, contaminated or involved only a single [so unreliable] replicate voltage variant displaying countable colonies. The only *S. praecaptivus* electroporation assay results obtained [repeats yielded blank plates] are displayed in Figure 3.5.

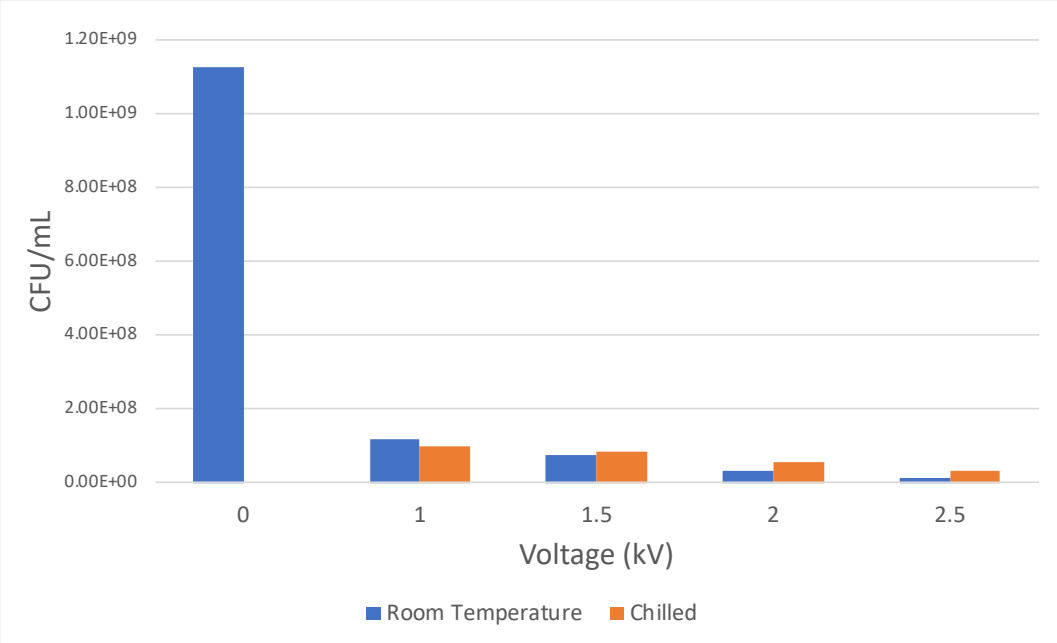


Figure 3.5. A graph displaying CFU/mL of *Sodalis praecaptivus* post-electroporation using a range of voltages (kV), either at room-temperature or ice-chilled. Blue indicates the CFU/mL values of the *S. praecaptivus* cultures 'blankly' electroporated at room temperature (*N* = 1 per voltage variant). Orange indicates the CFU/mL values of the *S. praecaptivus* cultures 'blankly' electroporated when kept on ice (*N* = 1 per voltage variant).

As *S. praecaptivus* is able to grow on minimal solid [and liquid] media and shows countable colonies overnight, displaying functional laboratory characteristics more

in line with free-living, easily-manipulatable bacteria, this inability to obtain repeats was unexpected. This symbiont-level difficulty in laboratory manipulation observed in *S. praecaptivus* is further highlighted by the results shown in Figure 3.5: electroporating *S. praecaptivus* results in a dramatic reduction in cell survival. The chilled samples had slightly higher cell survival following transformation at higher voltages, but due to the results only consisting of one replicate, this result is not statistically significant.

The results from the kanamycin assays, aimed at optimising a genetic transformation method for specific use on *S. glossinidius,* for *S. glossinidius* and *S. praecaptivus* are shown in Figure 3.6 A and B, respectively.

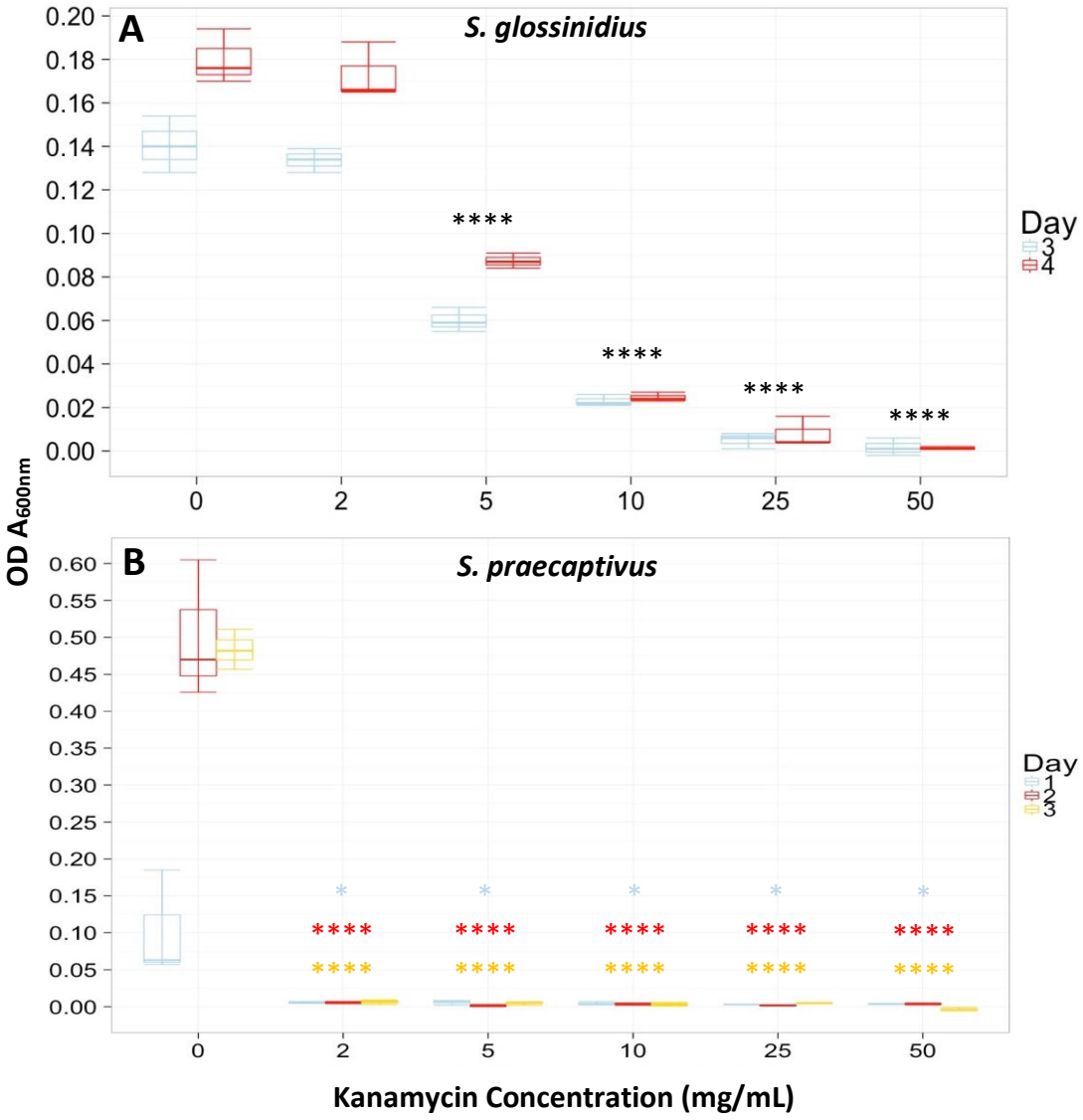


Figure 3.6. A graph showing the OD A_{600nm} readings of (A) *Sodalis glossinidius* and (B) *S. praecaptivus* cultures growing under a range of kanamycin concentrations. (A): blue plots - three days post-inoculation; red plots - four days post-inoculation. "****" indicating $p = 0$, from negative control [0 µg/mL kanamycin]. Three days post-inoculation ANOVA: $F_{5,12} = 266$, $p = 7.41 \times 10^{-12}$. Four days post-inoculation ANOVA: $F_{5,12} = 305.8$, $p = 3.42 \times 10^{-12}$. **(B):** blue plots – one day post-inoculation; red plots - two days post-inoculation; gold plots - three days post-inoculation. "*" indicating $p = 0.05$, from negative control [0 µg/mL kanamycin]. One day post-inoculation ANOVA: $F_{5,12} = 5.418$, $p = 0.00777$. Two days post-inoculation ANOVA: $F_{5,12} = 85.14$, $p = 6.01 \times 10^{-9}$. Three days post-inoculation ANOVA: $F_{5,12} = 914.4$, $p = 4.7 \times 10^{-15}$.

When comparing the tolerance of a free-living bacteria possessing a genome not undergoing genome degradation associated with becoming a symbiont [*S. praecaptivus*] to a symbiont with a vastly reduced functional capability [*S.*

glossinidius], the levels of *S. glossinidius* growth in lower concentrations of kanamycin is an unusual result (Figure 3.6 A). This natural resistance was not experimentally tested any further and the reason behind the tolerance is unknown. As can be seen from Figure 3.6 B, *S. praecaptivus* displays a negative growth trend that is more typical of free-living prokaryotes facing a traditionally absent antibiotic.

The only antibiotic variation from the *S. praecaptivus* voltage-antibiotic transformation assay using the Epicentre® EZ-Tn5^{TM} <KAN-2> Insertion Kit (Illumina, Inc.©) that yielded countable colonies was the 0 µg/mL concentration (Figure 3.7). The other antibiotic variant plates were either blank or did not display countable colonies, suggesting that the Tn5 did not successfully integrate into the chromosome of *S. praecaptivus*.

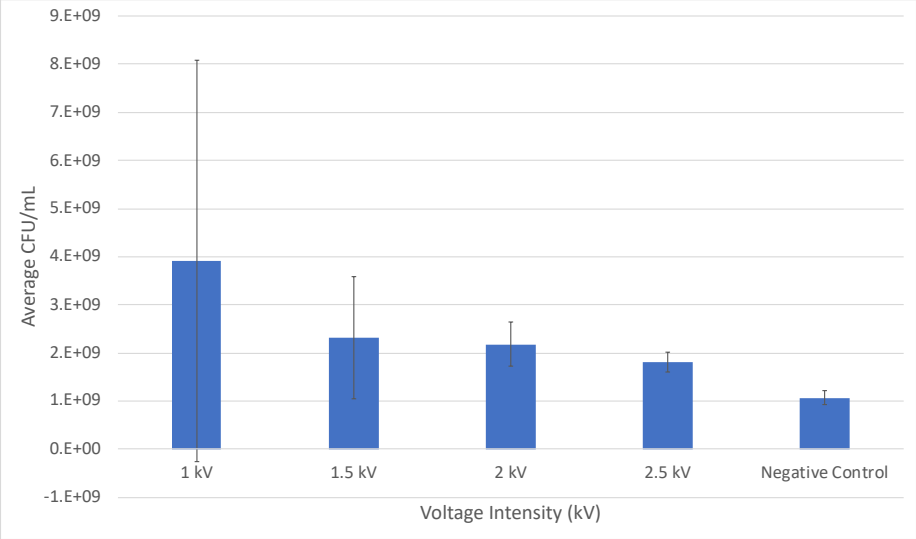


Figure 3.7. Plate observations two days after electroporating *Sodalis praecaptivus* cultures with Tn5, using different voltage intensities. Negative control: glycerol instead of the transposon, 2.5 kV and 0 µg/mL LB. Antibiotic variant: LB plate with 0 µg/mL kanamycin. CFU/mL: average of 10^{-6} triplicates per voltage variant (except for 2 kV variant: average of duplicate). Error bars as standard deviation.

As can be seen from Figure 3.7, the negative control is similar to the variant equivalent: both electroporated at 2.5 kV and plated on 0 µg/mL kanamycin LB, but

the negative control was transformed with glycerol instead of transposon (Table 3.1). The observations indicate that *S. praecaptivus* cell survival decreases with increasing electroporation voltage intensities (Figure 3.7).

The protocol efficacy tests using GFP results are displayed in Figure 3.8.

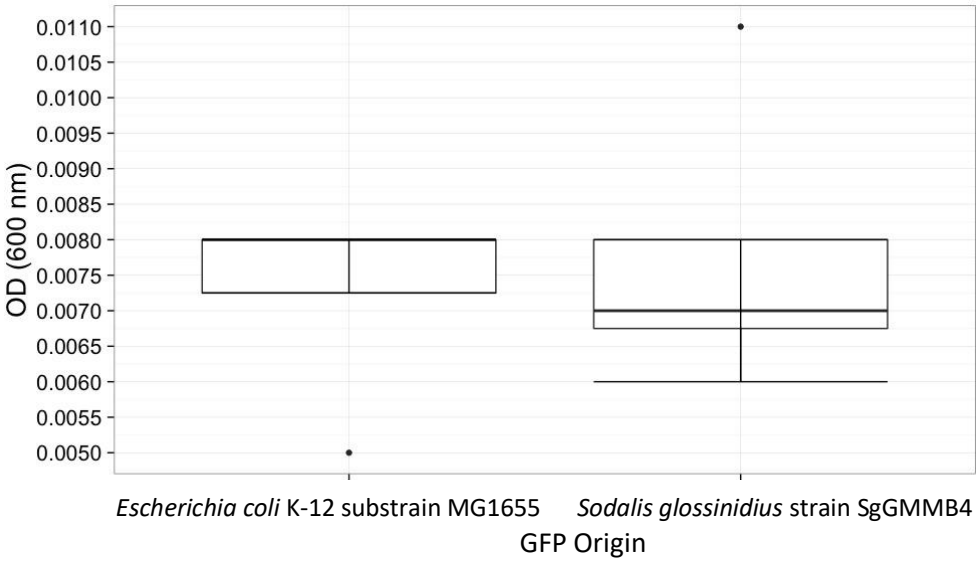


Figure 3.8. Growth values of survival cultures four days post-transformation using GFP from different origins: *Escherichia coli* K-12 substrain MG1655 or *Sodalis glossinidius* strain SgGMMB4. Survival cultures: Schneider's Insect Medium + 10% foetal bovine serum containing 50 µg/mL kanamycin. ANOVA: $F_{1,6} = 0.14$, $p = 0.722$

As can be seen from Figure 3.8, the growth values were too low to be deemed as a successful transformation; a lack of growth in cultures containing kanamycin suggests an unsuccessful integration of the GFP. This is further supported by the multiple check plates from throughout the protocol, whereby none of the kanamycin plates showed mutants, fluorescent or none. The fact that *S. glossinidius* has been previously successfully transformed using GFP [named pZEP17.1, extracted from *E. coli* K-12 substrain MG1655], but utilising a different protocol than that used here, would suggest that the Epicentre® EZ-Tn5^{TM} <KAN-2> Insertion Kit (Illumina, Inc.©) is not viable for the use on highly fastidious organisms such as *S. glossinidius*. This is greatly supported by the fact that no successful *S. glossinidius* transposon mutants were created using the Epicentre® EZ-Tn5^{TM} <KAN-2> Insertion Kit (Illumina, Inc.©), despite multiple repeats and the optimisations detailed above.

3.3.2. λ Red Homologous Recombination of *Sodalis glossinidius* Using the pKD46-pKD3 System

The pKD3 with *S. glossinidius*-homologous extensions fragment amplified as expected, but the other fragments did not display expected products (Figure 3.9).

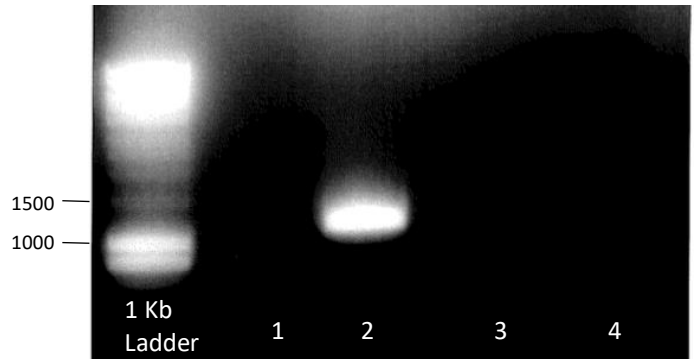


Figure 3.9. A 1% agarose gel displaying amplicon products following the PCR outlined in Tables 3.3 and 3.4. Lane 1 held the PROKKA_03540 fragment. Lane 2 held the pKD3 with *S. glossinidius*-homologous extensions fragment (product approximately 1.2 Kb). Lane 3 held the PROKKA_03541 fragment. Lane 4 held the negative control. Relevant ladder labels in bp.

The repeats conducted for the creation of the PROKKA_03540 and PROKKA_03541 fragments, but using DNA extracted from pKD46-carrying *S. glossinidius* instead of an *S. glossinidius* colony in 500 µL Schneider's Insect Medium + 10% FBS, successfully created the PROKKA_03540 fragment (Figure 3.10).

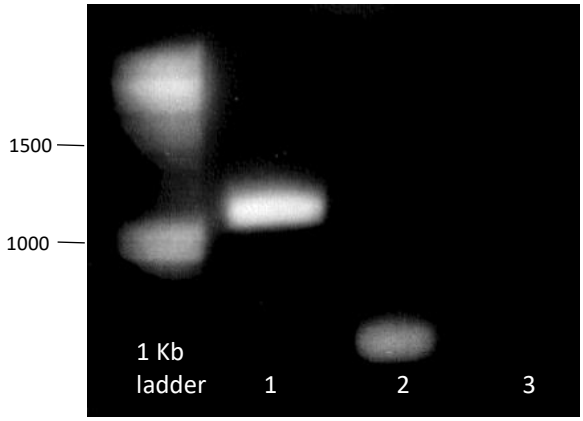


Figure 3.10. A 1% agarose gel displaying amplicon products. Lane 1 held the PROKKA_03540 fragment (product approximately 1.2 Kb). Lane 2 held the PROKKA_03541 fragment. Lane 3 held the negative control. Relevant ladder labels in bp.

Regarding the annealing temperature increase, the gel images revealed a faint amplicon band at the expected size (Figure 3.11). This low amount of product was still not sufficient, however.

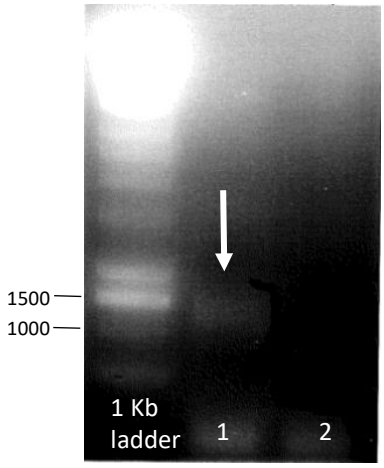


Figure 3.11. A 1% agarose gel displaying amplicon products. Lane 1 held the PROKKA_03541 fragment (product approximately 1.2 Kb, as shown by white arrow). Lane 2 held the negative control. Relevant ladder labels in bp.

In regard to the alternative use of Phusion® High-Fidelity DNA Polymerase (New England BioLabs®), in conjunction with cycle enhancers, the 1% agarose gel revealed that samples 3 [Q5® mix with both DMSO and betaine] and 7 [Phusion® mix with both DMSO and betaine] were the only variants that produced a fragment of the expected size (approximately 1.2 Kb) (Figure 3.12).

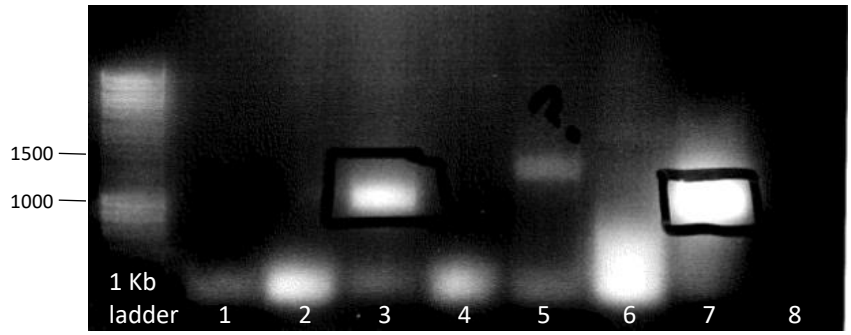


Figure 3.12. A 1% agarose gel displaying amplicon products. Lanes 1 – 8 are detailed in Table 3.5 as samples 1 – 8. The products in lanes 3 and 7 are approximately 1.2 Kb. The product in lane 5 is approximately 1.5 Kb. Relevant ladder labels in bp.

The overlap extension cycle from the fusion PCR failed to produce a fragment of the desired size (~3.2 Kb), despite the addition of cycle amplifiers such as betaine and DMSO and the additional attempts using Phusion® High-Fidelity DNA Polymerase (New England BioLabs®) (Figure 3.13 A and B).

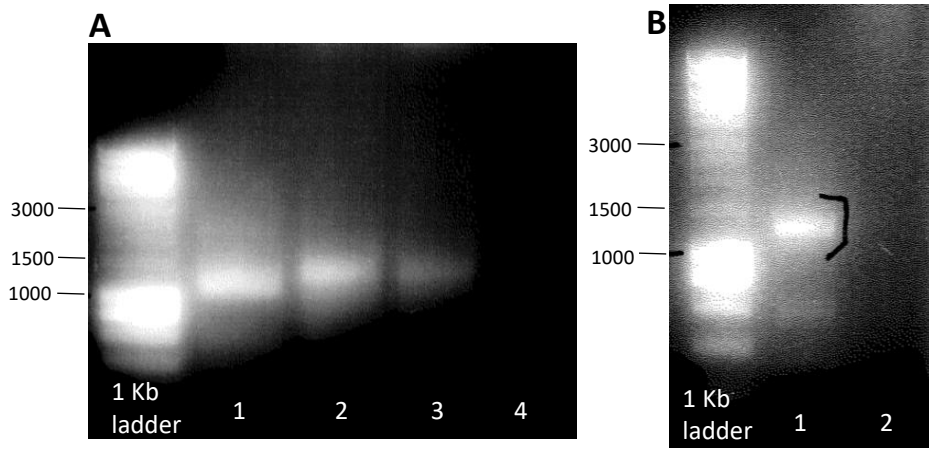


Figure 3.13. 1% agarose gels displaying amplicon products from the fusion PCR. A [Q5®]: Lanes 1 – 4 are detailed in Table 3.6 as samples 1.1 – 1.4. **B [Phusion®]:** Lanes 1 – 2 are detailed in Table 3.6 as samples 2.1 and 2.2. Relevant ladder labels in bp.

The external primers efficacy test amplicon products can be seen in Figure 3.14.

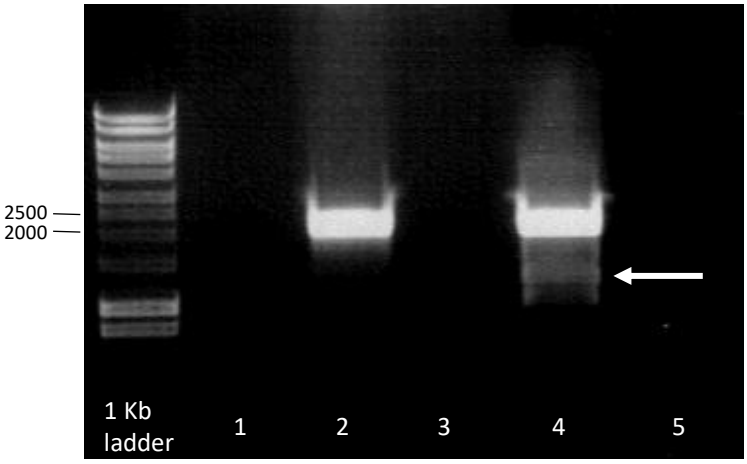


Figure 3.14. A 1% agarose gel displaying amplicon products from the primer efficacy test PCR. Lanes 1 – 5 are detailed in Table 3.7 as samples 1 – 5. The white arrow indicates non-specific binding. Relevant ladder labels in bp.

As can be seen in the 1% agarose gel (Figure 3.14), the only successful (i.e. correct size of 2 – 2.5 Kb) amplicon products were from the samples with the cycle amplifiers. The use of the Phusion® High-Fidelity DNA Polymerase (New England BioLabs®) enzyme also introduced non-specific binding, as can be seen by the multiple bands within the gel (white arrow). The ability of the primers to produce the expected product was confirmed, so the reason for the unsuccessful fusion of the fragments is unknown.

The amplicons from the re-designed primer PCR tests without cycle enhancers can be seen in Figure 3.15 A.

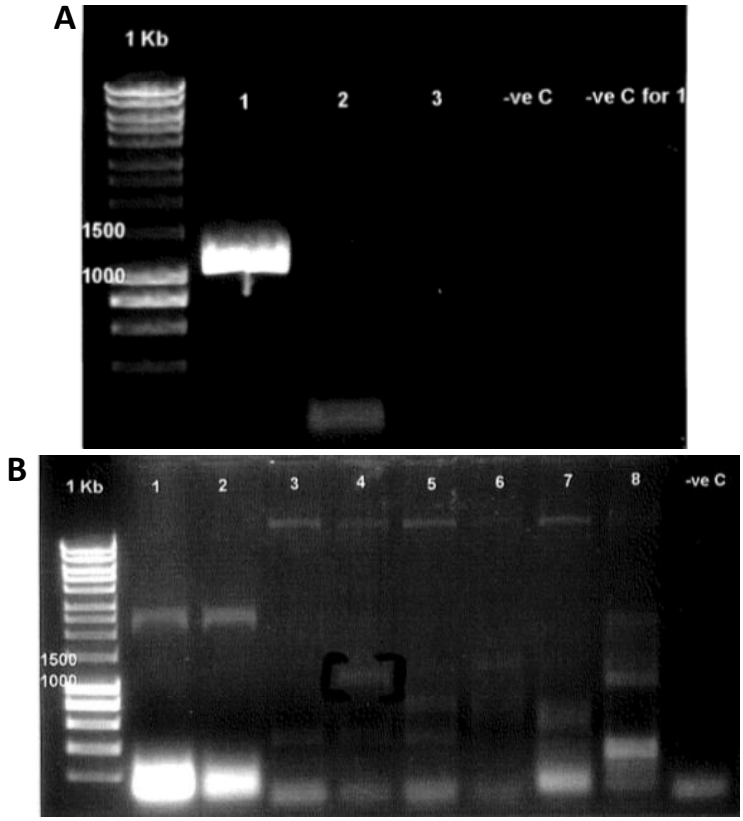


Figure 3.15. 1% agarose gels displaying amplicon products.
A: Lane 1 held the PROKKA_03540 fragment (product approximately 1.2 Kb) (Table 3.2). Lane 2 held the pKD3 with *S. glossinidius*-homologous extensions fragment (Table 3.8). Lane 3 held the PROKKA_03541 fragment (Table 3.8). Lane 4 ("-ve C") held the negative control. Lane 5 ("-ve C for 1") held the negative control specific to the sample in lane 1. **B:** Lanes 1 – 9 are detailed in Table 3.9 as samples 1 – 9. Relevant ladder labels in bp.

As can be seen from Figure 3.15 A, the re-designed primers also failed to produce fragments of the correct size. The resultant optimisation experiment [i.e. with cycle enhancers] amplicons can be seen in Figure 3.15 B; bands showing non-specific binding and low yield products. The only samples producing an amplicon product of the correct size (1.2 Kb) were 4 and 8: re-designed primers for the PROKKA_03541 fragment without cycle enhancers, and initial external primers with cycle enhancers, respectively. However, the bands did not contain enough product to be extracted.

Attempts to transform *S. glossinidius* using the Datsenko and Wanner (2000) pKD46-pKD3 system proved unsuccessful, despite multiple optimisation experiments.

3.4. Discussion

Reproducibility within scientific literature has been an issue amongst researchers for a long time; the ability, or inability, to repeat other research groups' experiments pulls into question the reliability of the original data conclusions made. Reasons ranging from the pressure to submit causing time constraints leading to poor oversight in experiments, to inconsistency within reagents, have been given by the scientific community when asked if they considered there to be a "reproducibility crisis" (Baker, 2016). Attempting novel experiments on a fastidious organism also falls under this reproducibility issue, as one has to initially use an existing and presumably-optimised protocol for an 'easy' organism and then perform further fastidious-organism-based optimisation experiments. When other scientists within the same 'easy' organism research community are unable to reproduce the experiment, the undertaking then involved in the fastidious-organism optimisation becomes substantially more arduous.

S. glossinidius is a fastidious, facultative symbiont of tsetse that requires rich growth media, takes 5 – 10 days under microaerophilic conditions to form countable colonies, and needs to be statically grown at 25°C. It is the epitome of genome degradation, whereby only approximately 50% of its genetic content has a coding capacity and high percentages of its genes are pseudogenised (Toh *et al.*, 2006; Belda *et al*., 2010; Goodhead *et al*., 2018). Despite the coverage of research involving the *S. glossinidius* genome, there is a distinct lack of functional genomics data, resulting in little being known of its role(s) within tsetse. In order to fix this, genetic manipulation experiments to elucidate functionality within bacteria, such as selecting for transposon mutants in specific environments or using knock-out mutants to confirm specific gene sets' functionality, need to be conducted. Examples of studies successfully using transposon mutants in specific selection environments to elucidate functionality include the identification of 356 growth-essential genes in *S. enterica* serovar Typhi and 724 in *S.* Typhimurium ST_{313} strain D23580 using Tn5 (Langridge *et al*., 2009; Canals *et al*., 2019). Chakravortty *et al*. (2002) successfully used the Datsenko and Wanner (2000) system to confirm specific gene sets' functionality, finding that SPI-2 is involved in protecting

intracellular *S.* Typhimurium from macrophage reactive nitrogen intermediates (ROI) by interfering with inducible nitric oxide synthase (iNOS) localisation (Chakravortty *et al*., 2002). The Datsenko and Wanner (2000) system has also been implemented on a large scale to create the Keio collection, which involved selective targeting of single genes and their precise deletion to create an *E. coli* K-12 library of 3985 mutants, cumulatively possessing only the [theoretically] essential gene set (Baba *et al*., 2006). As previously discussed, the system has been used by Pontes and Dale (2011) in *S. glossinidius* and was followed here, in combination with the original authors' publication [Datsenko and Wanner (2000)]. But a lack of fully detailed methodology, for example exactly how the authors constructed their *S. glossinidius*-specific pKD3/"*fliM*::*cat* allele", lead to difficulty in protocol re-creation, lengthy experimental repeats based on 'guess work' of exact conditions needed for particular steps, and multiple optimisations; a prime example of the reproducibility issues discussed within this writing (Pontes and Dale, 2011). Ultimately, no successful creation of λ Red *S. glossinidius* mutants was achieved. The creation of *S. glossinidius* [and *S. praecaptivus*] Tn5 mutants using the Epicentre® EZ-Tn5^{TM} <KAN-2> Insertion Kit (Illumina, Inc.©) also proved unsuccessful, again despite multiple optimisation experiments. This in conjunction with the negative results from the protocol efficacy tests using GFP in place of the transposon for *S. glossinidius*, which has been previously successfully transformed with GFP using a different protocol, demonstrates a failing in the system. However, due to the fact that the Epicentre® EZ-Tn5^{TM} <KAN-2> [and similar] Insertion Kit (Illumina, Inc.©) has been successfully utilised for other organisms, the evidence would suggest that any questions of kit reproducibility can only be implicated in *S. glossinidius* fastidiousness, in this case (Langridge *et al*., 2009; Canals *et al*., 2019).

As can be taken from the points covered in this writing, reproducibility issues within scientific literature can introduce arduous optimisation experiments for researchers attempting experimental recreations, which is only enhanced when attempting the protocol on highly fastidious organisms. When 'standard' protocols for fastidious organisms are difficult to follow due to missing protocol details, or simply do not exist, the production of negative results becomes the norm.

Chapter 4; Method Development for the *Sodalis glossinidius* Infection of Insect Cell Lines

4.1. Introduction

The concept of culturing eukaryotic cells *in vitro* was first introduced in the early 1900's, the most infamous of which being HeLa cells, which were taken from patient Henrietta Lacks' epidermoid carcinoma of the cervix (Harrison *et al.*, 1907; Carrel and Burrows, 1911a, b; Scherer *et al*., 1953). Immortalised cell lines can stem from a variety of eukaryotic organisms, ranging from the aforementioned human, to animals such as mice and insects. These lines are traditionally used for the study of pathogens' intracellular behaviour, for example the use of peritoneal macrophages from BALB/c mice to identify that *Salmonella* Typhimurium virulence is dependent on intracellular [macrophage] survival (Fields *et al*., 1986). It was soon realised that cell lines could also be used to culture prokaryotes observed to be restricted within host organisms. For example, RLO observed within *Glossina* spp. were successfully cultured *in vitro* using cells from the mosquito, *Aedes albopictus* (Welburn *et al*., 1987). These RLO were later re-classified as *Sodalis glossinidius* and are secondary, facultative symbionts of tsetse flies, which are vectors for human and animal [African] *Trypanosoma* spp. (Dale and Maudlin, 1999; WHO, 2019a). *S. glossinidius* are relatively unique within the symbiont classification, as they are one of few symbionts currently known to be successfully cultivated *in vitro* in cell-free media (Dale and Maudlin, 1999). Despite this useful phenotype, the functional understanding of *S. glossinidius* within its host remains sparse with a lack of publications in the field because of issues regarding a highly degraded genome and an arduous laboratory phenotype, introducing regular contamination and difficulty in genetic manipulation. With the novel concept of exploiting disease vectors' internal symbionts as control strategies, this lack of functional understanding has stayed the implementation of recombinant symbionts as standard. The proof-of-concept for tsetse being presented in the form of recombinant *S. glossinidius* producing a trypanolytic nanobody successfully colonising the midgut of *G. morsitans morsitans* (De Vooght *et al*., 2014). The utilisation of genetic

manipulation methods for the elucidation of genome functionality has been successfully applied across many prokaryote spp., including for example in *S. enterica* serovar Typhi where 356 [laboratory conditions] growth-associated essential genes were identified using a method harnessing transposon mutants in specific selection environments (detailed in 1.2.4.) (Langridge *et al*., 2009). Applying this method to transposon symbiont mutants within insect cell lines would provide essential gene candidates and functionality knowledge in a down-scaled insect environment [cell line] and elucidate functionality in the context of a whole system [tsetse]. Examples of insect cell lines includes those obtained from the previously-mentioned *A. albopictus* [Dipteran: Culicidae], and *Drosophila melanogaster* [Dipteran: Drosophilidae]. C6/36 cells were originally obtained from *A. albopictus* larvae and *Drosophila* Schneider 2 (S2) cells were originally obtained from late stage embryos, respectively (Singh, 1967; Schneider, 1972).

As previously mentioned, *S. glossinidius* was originally cultured *in vitro* using insect cells, the protocol and subsequent protocols developed in the context of maintaining *S. glossinidius*: mixing and centrifuging tsetse haemolymph with C6/36 cells, then passaging 1:10 every three days with fresh C6/36 culture (Welburn *et al*., 1987; Dale and Maudlin, 1999). A protocol in the context of studying *S. glossinidius* infection dynamics [as opposed to just *in vitro* maintenance] was published by Dale *et al*. (2001) as an adaptation to the previous protocols: [mutant] *S. glossinidius* was mixed at a multiplicity of infection (MOI) of ~10 with C6/36 cells for 24 h at 25°C, after which the invaded insect cells were pelleted [1,500 x g for five min at 25°C]. However, as this was aimed at studying *S. glossinidius* 'natural' invasion dynamics of insect cells, the protocol (as with the previous protocols) is relatively simplistic and not developed for optimal bacterial invasion and maintenance within host cells. Canals *et al*. (2019) utilised a comparatively more-detailed infection protocol adapted for optimal *S.* Typhimurium [ST_{313} strain D23580] infection of RAW264.7 macrophages to identify 68 essential genes for intra-macrophage replication.

The data presented here has examined an array of protocol variations for the development of a *S. glossinidius*-adapted infection protocol of insect cell lines, aimed specifically for subsequent use in *in vivo* functional studies [Chapter 5].

4.2. Methods

4.2.1. *Sodalis glossinidius*-Development of Cell Line Infection Protocols

The protocol presented here was adapted for *S. glossinidius* infection of S2 and C6/36 cells from those published by Dale *et al*. (2001) and Canals *et al*. (2019), using *S. glossinidius* strain SgGMMB4 (GenBank accession no. LN854557) transformed with GFP [named pZEP17.1, extracted from *Escherichia coli* strain K-12 substrain MG1655]. The final protocol is as follows: prior to infection, all wells [from a 6-well plate (Sarstedt)] were incubated with 0.5 mg/mL concanavalin A [a carbohydrate-binding protein often used as a sample adherent for microscope slides] for 10 min and then air-dried with UV exposure. Cells were then transferred to the wells at 10^6 [based on haemocytometer measurements] in 2 mL Schneider's Insect Medium (Sigma-Aldrich) + 10% FBS (Gibco) and left to settle in the concanavalin A-treated wells for 3 days. The six-well cell culture plate was spun in a centrifuge at 1000 rpm for five min and the cell medium was removed. The adhered cells were combined with cell-number-adjusted [for MOI 100], washed GFP *S. glossinidius* cells, where the final well volumes were 2 mL [Schneider's Insect Medium + 10% FBS]. These were mixed by gently rotating the plate by hand, spun in a centrifuge at 1000 rpm for five min and then statically incubated at 25°C for 1 hour. After the incubation, 2 μL 100 mg/mL gentamicin was added to the 2 mL well volumes for a working concentration of 100 μg/mL, and the plates gently mixed by hand rotation. The plate was then spun in a centrifuge at 1000 rpm for five min and statically incubated at 25°C for 1 hour. After the second incubation, the plate was spun in a centrifuge at 1000 rpm for five min and the medium removed from all wells. 2 mL Schneider's Insect Medium + 10% FBS with 10 μg/mL gentamicin was added to each well and the plate was incubated statically at 25°C for 48 hours. Cell scrapers (Sigma) were then used to dislodge the cells from the wells and all six well volumes were pooled into a 15 mL falcon tube. Following vigorous vortexing, the plate pool was then spun in a centrifuge at 8000 rpm for 10 min. The pool supernatant was removed, the pellet vigorously re-suspended in Schneider's Insect Medium + 10% FBS via a pipette and the suspension vortexed. Serial dilutions were

created from the pool (10^0 – 10^6), were spot-plated on horse blood (TCS Biosciences) Columbia agar (Sigma-Aldrich) plates and statically incubated at 25°C for up to 10 days.

The immediate adaptations applied to the Canals *et al*. (2019) protocol included: the use of Schneider's Insect Medium + 10% FBS; the initial GFP *S. glossinidius* Schneider's Insect Medium + 10% FBS at 50 µg/mL kanamycin cultures were washed with Schneider's Insect Medium + 10% FBS to remove residual kanamycin; incubation at 25°C; final minimum incubation of 24 hours. Optimisations of the protocol included: reductions in washing steps and addition of concanavalin A [due to cell loss]; MOI variants at 5, 10, 50, 100 or 200; final incubation increments of 24, 48, 72 or 96 hours; extraction variants of 0.1% Triton X-100 (Sigma), 1% Triton X-100, 4000 rpm centrifugation or 8000 rpm centrifugation.

MOI variants: the protocol was followed as detailed above in 4.2.1., but the adhered insect cells were combined with washed GFP *S. glossinidius* cells at MOI at 5, 10, 50, 100 or 200 [for a final well volume of 2 mL].

Final incubation increment variants: the protocol was followed as detailed above in 4.2.1., but the final incubation was either at 24, 48, 72 or 96 hours.

Extraction variants: the protocol was followed using S2 cells as detailed above in 4.2.1., but three 6-well plates underwent the variations post-final incubations as follows: (1.) the plate was spun in a centrifuge at 1000 rpm for 5 min and the medium was removed. 1 mL Schneider's Insect Medium + 10% FBS with 0.1% Triton X-100 was added to each of the top three wells and (2.) 1 mL Schneider's Insect Medium + 10% FBS with 1% Triton X-100 was added to each of the bottom three wells. Cell scrapers were used to dislodge the cells from the wells and each Triton X-100 3-well/row variation was separately pooled into a 15 mL falcon tube. Both pools were vigorously vortexed and left to statically incubate at room temperature for 10 min. (3.) Following the final incubation, the S2 cells were dislodged from their wells using a cell scraper and all six wells were pooled into a 15

mL falcon tube. The plate pool was spun in a centrifuge at 4000 rpm (4.) or at 8000 rpm for 10 min. Post-centrifugation, both pools had their supernatants reduced from 6 mL to 2 mL and the pellets vigorously re-suspended via a pipette. All four pools created had serial dilutions' created from them (10^0 – 10^6), were spot-plated on horse blood Columbia agar plates and incubated at 25°C for up to 10 days.

Each well was photographed using a Life Technologies EVOS® FL Cell Imaging System, using the GFP overlay function. In total [excluding quality control images], 82 S2 and 74 C6/36 images were manually counted using ImageJ (National Institutes of Health) with the Cell Counter plugin (Kurt De Vos, University of Sheffield), whereby each GFP *S. glossinidius*-infected insect cell and all insect cells were marked. Additional photographs were taken using the Life Technologies EVOS® FL Cell Imaging System before and after the experiment to visualise any cell loss, but these were quality control and so were not counted or included in the data set.

4.3. Results

4.3.1. *Sodalis glossinidius*-Development of Cell Line Infection Protocols

The results from the protocol step-reduction optimisation experiment, aimed at developing a cell line infection method for specific use on *S. glossinidius* within *in vivo* functional studies, are displayed in Figure 4.1 A and B.

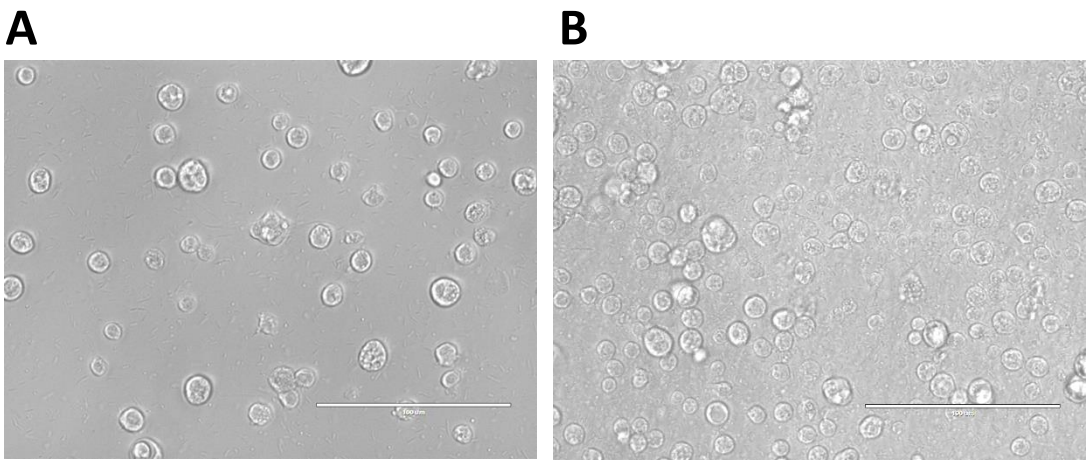


Figure 4.1 Microscope images of S2 cells. A: S2 cells infected with GFP-*S. glossinidius* in Schneider's Insect Medium + 10% FBS, after experimentation but before the final incubation(s) **B:** S2 cells infected with GFP-*S. glossinidius* in Schneider's Insect Medium + 10% FBS, after experimentation but before the final incubation(s) implementing the reduction in protocol steps and concanavalin A.

As can be seen from Figure 4.1 B versus A, the retention of S2 cells increased after experimentation but before the final incubation step by reducing the number of wash steps and the use of concanavalin A-treatment of wells.

The [average] results from the MOI and final incubation increment variant experiments, aimed at developing a cell line infection method for specific use on *S. glossinidius* within *in vivo* functional studies, can be seen in Figure 4.2 A and B.

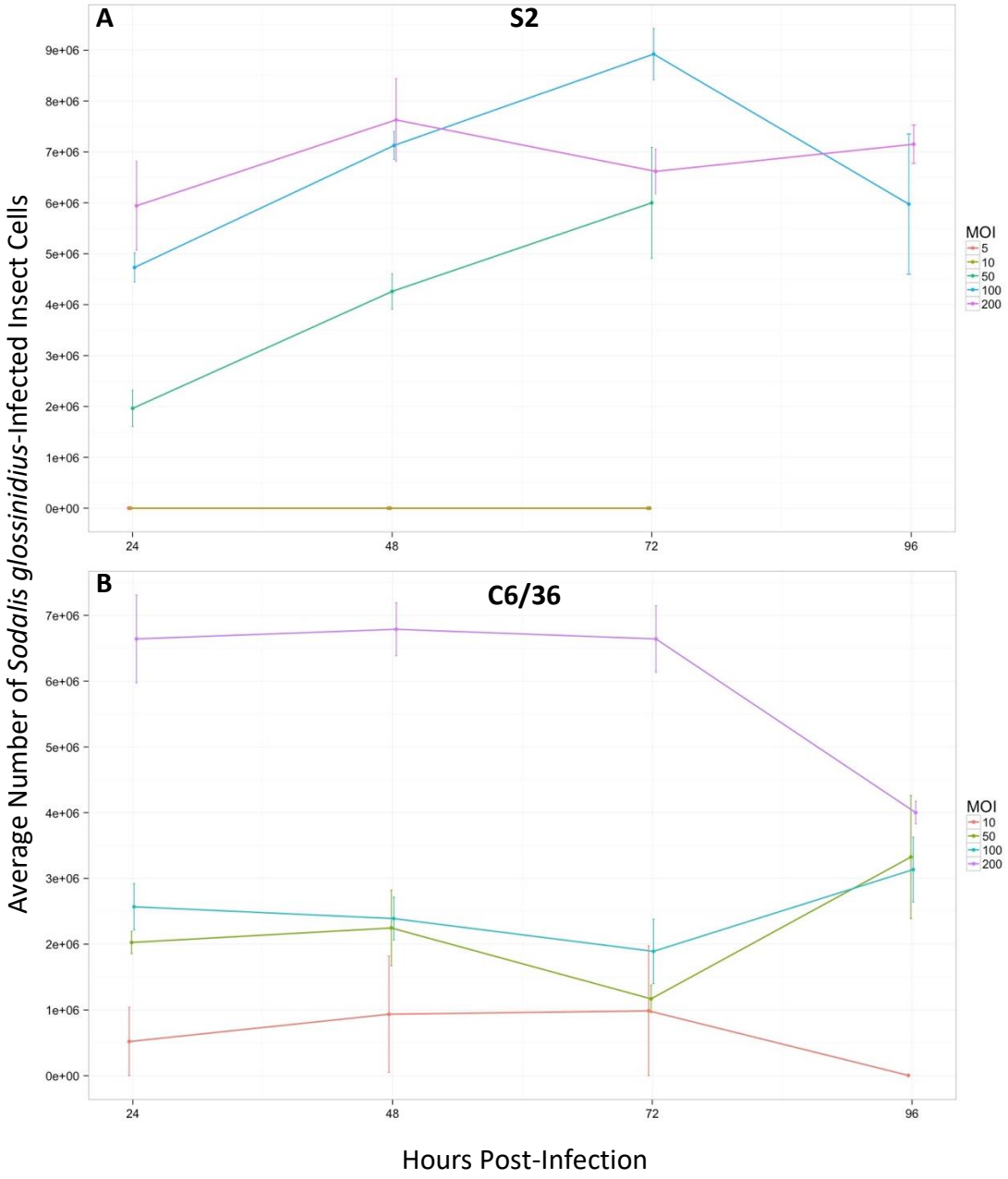


Figure 4.2. The average number of GFP *Sodalis glossinidius*-infected insect cells over a range of 24-hour time increments, using a range of MOIs. A: insect cell line - S2. MOI 5 *N* = 3. MOI 10 *N* = 3. MOI 50 *N* = 12. MOI 100 *N* = 36. MOI 200 *N* = 30. **B:** insect cell line – C6/36. MOI 10 *N* = 7. MOI 50 *N* = 14. MOI 100 *N* = 30. MOI 200 *N* = 24. The [average] numbers of GFP-*Sodalis glossinidius*-infected insect cells have been corrected for the whole well from the microscope image.

As can be seen from Figure 4.2 A and B, the general trends show higher numbers of GFP *S. glossinidius*-infected insect cells when higher MOI values are used. S2 cells show slightly greater overall average numbers of infection compared to C6/36 cells,

per variant [minus MOI 10 and 200]. Interestingly, the difference between S2 MOI 100 and MOI 200 (Figure 4.2 A) was not as large as expected, demonstrating a potential saturation point in the bacterial-uptake capability of S2, especially when compared to the C6/36 MOI 100 versus MOI 200 difference (Figure 4.2 B). A slight decrease was expected in the proportion of GFP *S. glossinidius*-infected insect cells over the 24-hour increments, for two reasons: firstly, the cells were in 'maintenance culture' [10 μg/mL gentamicin] meaning that the GFP *S. glossinidius* cells were not permitted to re-infect more insect cells after their initial host; secondly, the cells have a relatively fast proliferation rate [more so for S2], meaning any uninfected cells should have increased in number over the time increments used. This expectation can be observed where an S2 MOI of 100 and C6/36 MOI of 10 and 200 were used, but was not observed within the other MOI variants (Figure 4.2 A and B). The C6/36 MOI 50 and 100 variants (Figure 4.2 B) displayed an unexpected loss-recovery trend after 72 hours post-infection, where infected insect cell numbers are greater than those observed over the previous two time-increments. From obversions during numerous experimental repeats, C6/36 proliferates relatively slowly [compared to S2], so the ability to generate the difference seen in cell number within the 24 hours remains unclear. The data shown in Figure 4.2 A and B has been re-represented in Figure 4.3 A – D.

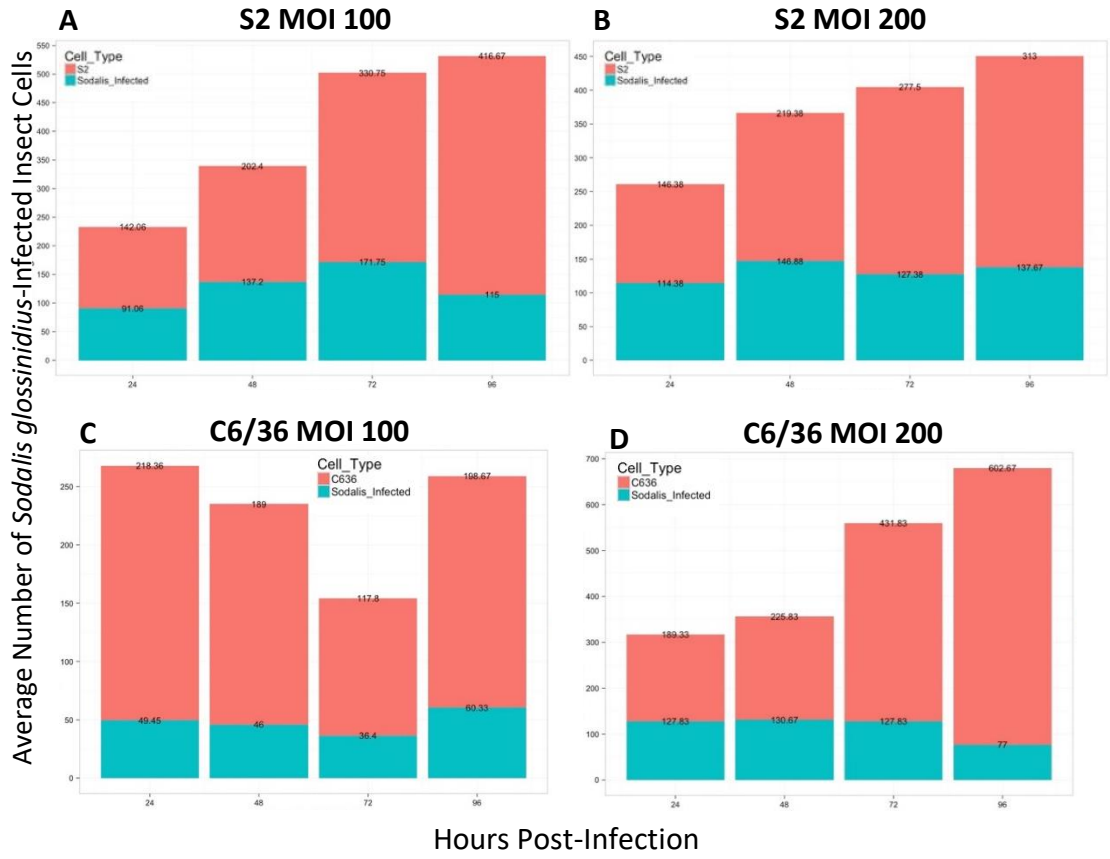


Figure 4.3. The relative average numbers of GFP *Sodalis glossinidius*-infected insect cells out of average total insect cells, over 24-hour time increments. A: infected S2 cells at MOI 100. 24 hours *N* = 16. 48 hours *N* = 10. 72 hours *N* = 4. 96 hours *N* = 6 **B:** infected S2 cells at MOI 200. 24 hours *N* = 8. 48 hours *N* = 8. 72 hours *N* = 8. 96 hours *N* = 6 **C:** infected C6/36 cells at MOI 100. 24 hours *N* = 11. 48 hours *N* = 11. 72 hours *N* = 5. 96 hours *N* = 3. **D:** infected C6/36 cells at MOI 200. 24 hours *N* = 6. 48 hours *N* = 6. 72 hours *N* = 6. 96 hours *N* = 6. The [average] numbers used here are from the microscope images (i.e. not corrected for the whole well).

The previously-described expectation of an increase in the number of uninfected insect cells can be more clearly seen in the re-represented data, newly displaying the trend in S2 MOI 200 [and C6/36 MOI 50 (Appendix Figure 2 B)] as a steady infection number with increasing uninfected cells (Figure 4.3 A - D). The re-representation of data more clearly shows the vastly lower numbers of GFP *S. glossinidius*-infected insect cells in the MOI 5, 10 and 50 variants, re-confirming the conclusion drawn above [from Figure 4.2 A and B] that higher levels of infection are seen when higher MOIs [100 and 200] are used (Appendix Figures 1 – 2). The

greater numbers of infected cells seen in S2 compared to C6/36 as discussed above [from Figure 4.2 A and B] are also more clearly seen in the re-represented data (Figure 4.3 A - D); S2 MOI 100 infected cell numbers similar or higher than C6/36 numbers at MOI 200. Levels of variation not observed in S2 were seen in C6/36; further to the loss-recovery observation discussed above [from Figure 4.2 B], the numbers of GFP *S. glossinidius*-infected C6/36 cells appeared sporadic across MOI variants [compared to trends seen in S2] (Figure 4.3 C and D and Appendix Figure 2).

The results from the extraction variant experiment, aimed at developing a cell line infection method for specific use on *S. glossinidius* within *in vivo* functional studies, are outlined in Figure 4.4.

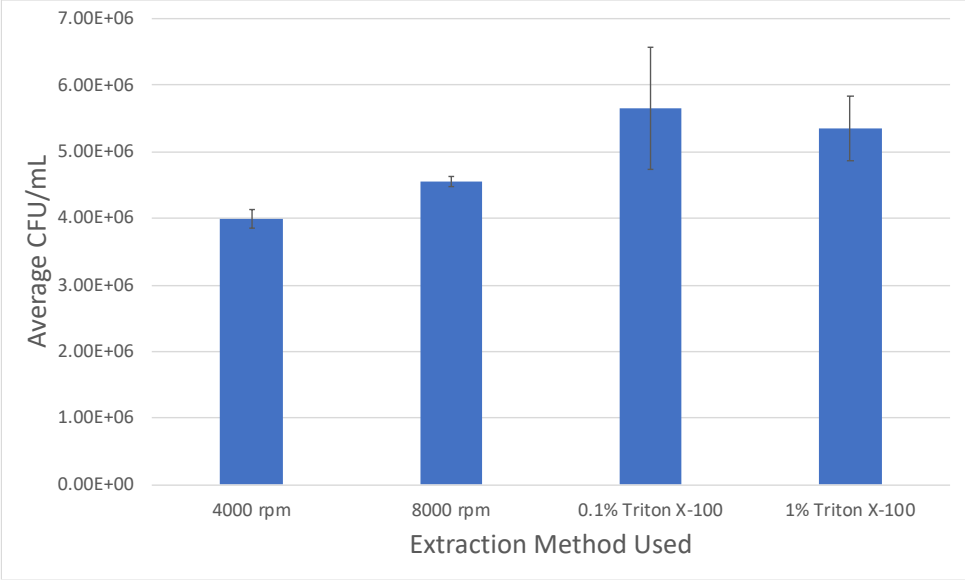


Figure 4.4. Average CFU/mL of *Sodalis glossinidius* after being extracted from S2 cells using Triton X-100 or centrifugation. Average [from duplicates – *N* = 2, per variant] CFU/mL calculated after 10 days growth on horse blood Columbia agar. Error bars as standard deviation.

As can be seen from Figure 4.4, the [average] CFU/mL difference between variants was minimal [all in the same factor], suggesting little beneficial reason for using

Triton X-100 over centrifugation for the lysing of the S2 host cells. This is furthered when considering the observations on the horse blood Columbia agar plates that *S. glossinidius* colonies from the Triton X-100 variants were smaller than those from the centrifugation plates, suggesting a potentially detrimental effect on *S. glossinidius* growth.

With the aim of the experiments presented here as developing an *S. glossinidius*-adapted insect cell line infection protocol specifically for subsequent use in *in vivo* functional studies [TraDIS, covered in Chapter 5], optimal conditions were selected as follows: infection of the S2 cell line at an MOI of 100 for 48 hours, after which extraction of bacteria from host cells via centrifugation at 8000 rpm. The reasons for selecting the S2 cell line over the C6/36 line included: comparatively faster growth rates; higher [average] numbers of cell infection (for example, after 48 hours 40.76% infection in S2 cells at MOI 100 versus 26.10% of C6/36 at MOI 100 or 36.81% at MOI 200); C6/36 displayed unexplainable variation trends, including a loss-recovery phenomenon at 72 hours and sporadic infection numbers across MOI variants; C6/36 has a comparatively [to S2] difficult laboratory behaviour as it's naturally-adherent, so has an increased tendency of cell bunching and forming multiple confluent layers on top of one another. The selection of MOI 100 was as follows: MOI 100 showed higher infection rates than MOIs 5, 10 or 50, in addition to showing similar infection rates than those seen at MOI 200, suggesting an uptake saturation point; microscope observations showed higher rates of multiple bacterial infections in one insect cell (example seen in Appendix Figure 3) at MOI 200 versus MOI 100, where ideally one would want a 1:1 infection. In relation to choosing 48 hours: the Canals *et al*. (2019) protocol cultured the bacteria post-cell line extraction in order to generate enough DNA yield for sequencing. For TraDIS experiments, it is important not to mix selections within one sample, otherwise one would not be able to confidently relate dictated essential genes to one selection pressure. With this in mind, culturing the extracted *S. glossinidius* cells prior to DNA extraction would introduce a second selection pressure to the pre-existing infection of an insect cell condition: selection under

Schneider's Insect Medium + 10% FBS. This means that to generate enough genetic material for sequencing, one would have to allow for *S. glossinidius* proliferation within the insect cells. As *S. glossinidius* doubling time has been calculated at approximately 26 hours (Matthew *et al.*, 2005), extracting them from the insect cell line at 48 hours post-infection would have allowed for approximately two doubling events (depending on which division stage the cells were in at the point of infection). Allowing for longer infection [72 or 96 hours] of the insect cells risks *S. glossinidius* over-infection and potential host cell lysis, which also introduces the possibility of gene selection for host cell exit in place of host cell infection; important for TraDIS experiments investigating individual selection pressures. From observations during the numerous experimental repeats, it's also the case that at longer time periods the S2 cells have surpassed a reasonable level of confluence, meaning that a relatively large amount of insect cell components would be included within the DNA extractions aimed at *S. glossinidius*. Allowing for a shorter infection [24 hours] of the insect cells risks a low intracellular *S. glossinidius* load and so a small DNA yield, for sequencing. Further to this, the variance at 48 hours is tighter than that observed at 24 hours within the same range of percentage infection [MOI 100], dictating 48 hours at the statistically favourable choice (Figure 4.4).

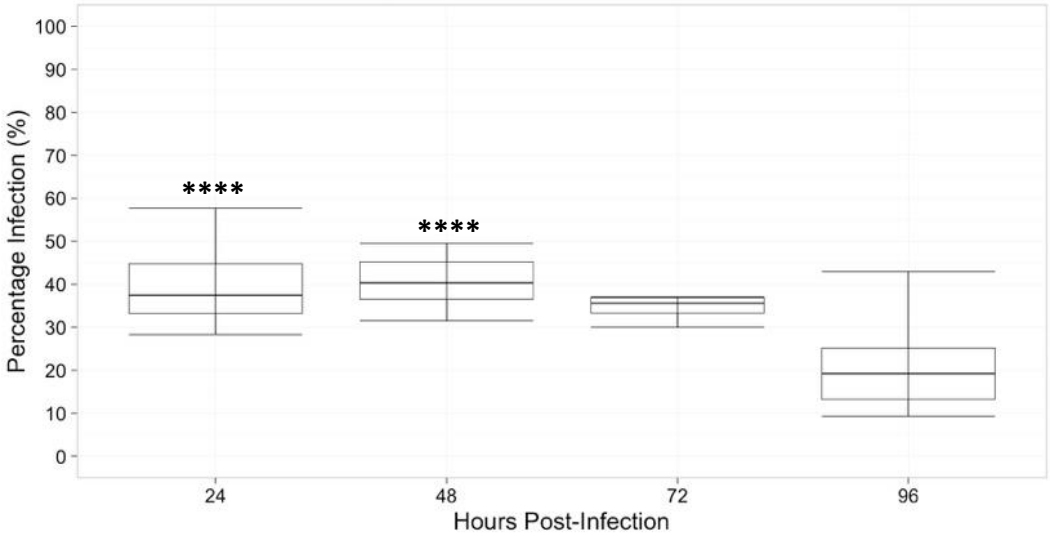


Figure 4.4. The percentage infection of S2 cells with GFP-*S. glossinidius* at an MOI of 100, over 24-hour time increments. Percentages as averages from *N* = 16 for 24 hours, *N* = 10 for 48 hours, *N* = 4 for 72 hours and *N* = 6 for 96 hours. ANOVA: $F_{3,32} = 8.425$, $p = 0.000287$. "****" indicating $p = 0$, from 96 hours.

The selection of 8000 rpm for extraction of *S. glossinidius* was as follows: similar *S. glossinidius* cell recovery numbers between variants; potential negative growth effects of Triton X-100 on *S. glossinidius* seen by comparatively smaller colony sizes; specifically 8000 rpm over 4000 rpm to reduce as much as possible the loss of *S. glossinidius* cells in the discarded supernatant.

4.4. Discussion

Utilising cell lines *in vitro* to study the dynamics of prokaryotes has been successfully used to clarify functionality in the genomic context. For example, 68 essential genes specifically for intra-macrophage replication were identified for *S.* Typhimurium [ST_{313} strain D23580] using RAW264.7 macrophages (Canals *et al*., 2019). This level of functional knowledge depth is absent for *S. glossinidius*; a fastidious secondary symbiont of tsetse flies implicated in its host's fastidiousness to parasites (detailed in 1.1.5.). Reasons for this stem from issues regarding a highly degraded genome and an arduous laboratory phenotype, introducing regular contamination, difficulty in genetic manipulation and thus, a lack of functional genomics publications. Developing genetic manipulation methods used for elucidating gene functionality within a eukaryotic line optimised for other prokaryotes for *S. glossinidius* would bridge this knowledge-gap and potentially clarify the symbiont's role in its host. Identifying essential gene candidates in an insect cell line would provide viable targets for functionality studies within the teste system. Protocols for *S. glossinidius in vitro* maintenance within cell lines have been published, in addition to simplistic infection protocols aimed at studying 'natural' invasion dynamics, but a fully optimised infection protocol is absent (Welburn *et al*., 1987; Dale and Maudlin, 1999; Dale *et al*., 2001). With this in mind, the infection assays presented here aimed to develop an *S. glossinidius*-adapted infection protocol of insect cell lines specifically for subsequent use in *in vivo* functional studies [TraDIS, Chapter 5].

Using the results, the optimal conditions for the TraDIS cell line selection pressure [Chapter 5] include infection of the S2 cell line at an MOI of 100 for 48 hours, after which extraction of bacteria from host cells via centrifugation at 8000 rpm. Justifications for this are fully detailed in 4.3.1., but of-note reasons include: higher infection rates of the S2 cell line compared to the C6/36 cell line across all [bar MOI 10 and 200] MOI variants tested; MOI 100 offering higher infection rates than the lower MOI variant ranges, and similar rates to the MOI 200 variant; 48 hours post-infection offering the ideal point for optimal bacterial DNA yield, in between too little DNA [24 hours] and a change in selection from host cell infection

to exit [72 – 96 hours]; the use of physical [centrifugation at 8000 rpm] over chemical [Triton X-100] lysis of host cells for bacterial extraction due to minimal cell recovery number differences and a detrimental growth effect observed when using the detergent. The successful development of an *S. glossinidius*-adapted infection protocol of insect cell lines was achieved here and was subsequently applied as a standard protocol for the TraDIS cell line selection pressure [Chapter 5].

Chapter 5; Transposon-Directed Insertion Site Sequencing (TraDIS) of *Sodalis glossinidius*

5.1. Introduction

Methods aimed at genetically manipulating an organism-of-interest to clarify the function of specific genes have long been established. Two examples of this including the use of homologous recombination to knock-out/-in target genes, and harnessing naturally mutagenic transposons within selected environments to identify which genes are essential for survival within the pressure [fully detailed in 1.2.4.] (Datsenko and Wanner, 2000; Langridge *et al*., 2009). The latter example has been successfully used to identify essential genes in bacterial species such as *Salmonella* spp. and *Escherichia* spp. for environments ranging from laboratory growth, to cell line [macrophage] infection, to colonisation of animals (Langridge *et al*., 2009; Chaudhuri *et al*., 2013; Goodall *et al*., 2018; Canals *et al*., 2019). These studies utilised a particular type of transposon-based genetical manipulation known as TraDIS (Langridge *et al*., 2009; Barquist *et al*., 2016). Detailed fully in 1.2.4., but briefly TraDIS utilises the random insertion of transposons, selection of the mutants within specific pressures, and amplification of transposon-flanking sequences in the surviving mutants (Langridge *et al*., 2009; Barquist *et al*., 2016). Experimental transposon insertion can be achieved through various techniques, ranging from using manufactured kits such as the Epicentre® EZ-Tn5™ <KAN-2> Insertion Kit (Illumina, Inc.©), to constructed plasmid delivery vectors ["plasposons"] such as pRL27 (Dennis and Zylstra, 1998; Larsen *et al*., 2002). The elucidation of specific gene functionality within relevant environments has proved vital in furthering the understanding of phenotypes for organisms like the previously mentioned *Salmonella* spp. and *Escherichia* spp., knowledge that is currently lacking for the highly fastidious, secondary symbiont [of tsetse flies], *Sodalis glossinidius* (Dale and Maudlin, 1999). Issues stemming from a massively degraded genome introducing difficulty in genetic manipulation, and a fastidious laboratory phenotype offering high contamination risks and time-consuming experiments have contributed to this lack of knowledge. Also contributing are field studies unable to replicate the finding

that *S. glossinidius* implicates host fastidiousness to trypanosomes from both other field studies and those reproducing similar experiments in a laboratory setting (Maudlin, 1982; Maudlin and Ellis, 1985; Welburn *et al.*, 1993; Trappeniers *et al*., 2019). The genetic scope of *S. glossinidius* has been shown by studies that have published successful genetic manipulation results (Beard *et al*., 1993; Pontes and Dale, 2011; De Vooght *et al*., 2014). One study introduced a miniTn5 using pUTkm1 DNA into *S. glossinidius* strain $M1^T$, the mutants of which were subjected to insect cells to study the invasion mechanisms (de Lorenzo *et al*., 1990; Dale *et al*., 2001). This study represents one of the few that have transformed *S. glossinidius* in an attempt to understand its genetic functionality, but its limited approach [i.e. only focusing on *S. glossinidius* infection of insect cells], lack of corresponding functional genomics sequence data, and no further publications utilising Tn5 *S. glossinidius* mutants feeds into the knowledge gap. Other studies aimed at clarifying *S. glossinidius* functionality have discerned that it is able to utilise a broad range of carbon sources, but its preferred primary source is in dispute; early studies found a preference for *N*-acetylglucosamine over glucose, supported by a more recent *in silico* study, but others have found no significant difference between growth levels when using *N*-acetylglucosamine versus glucose [at 20 mg/mL, 2.3.3.], and [*in silico*] levels matching those seen in *Escherichia coli* when using glucose (Dale and Maudlin, 1999; Toh *et al*., 2006; Belda *et al*., 2010, 2012; Hall *et al*., 2019). *In vitro* preferencing studies utilise minimal media supplemented with test components to identify which source provides better growth [as shown in Chapter 2], the construct of which providing an ideal TraDIS selection pressure for confirming specific gene functionality in a targeted environment. Identifying which genes are essential in media deliberately restricting access to chosen sources, such as only *N*-acetylglucosamine or glucose, could reveal novel *S. glossinidius* metabolic transport mechanisms or utilisation preferences. This also applies to using cell lines as a TraDIS selection pressure, which would identify essential genes for *S. glossinidius* infection and maintenance in insect cells and could be applied to understanding functionality within the larger insect system [tsetse host].

The *S. glossinidius* TraDIS library presented here utilised different selection pressures [complex medium, supplemented minimal medium, and an insect cell line] with the aim of identifying essential gene candidacy, and thus functionality within specific conditions. It represents a novel insight into *S. glossinidius* gene essentiality and functionality, providing the currently absent functional genomics data within a field presently inferring phenotypic characteristics based on genetic sequence information and observational studies involving host species. The hypothesis for the complex medium selection pressure was that the *S. glossinidius* essential genes would cover the widest range of transport and utilisation pathways, as the medium provides a rich source of amino acids, carbohydrates and vitamins. The hypothesis for the supplemented minimal medium selection pressure restricted specifically to *N*-acetylglucosamine or glucose availability, was that the *S. glossinidius* essential gene range would be more defined and confirm the separate hypothesis [based on the data presented in Chapter 2] that the remaining intact *S. glossinidius* transporters are lenient. The hypothesis for the insect cell line selection pressure was that the *S. glossinidius* essential genes would cover virulence- and free-living associated genes, and pseudogenes, as these sets have not been removed from the genome but instead actively maintained, suggesting that they are important for the symbiont's ability to maintain symbiosis. The essential genes identified in this pressure could be extended to the larger insect system [tsetse] and potentially clarify *S. glossinidius* functionality within its host (Goodhead *et al*., 2018).

5.2. Methods

5.2.1. Transposon Transformation of *Sodalis glossinidius*

5.2.1.1. Extraction of the pRL27 Construct

pRL27 (Figure 5.1) was constructed by Larsen *et al*. (2002) and contains a *tetA*-promotor-controlled transposase gene [*tnp*] from plasmid RP4 encoding a hyperactive transposase, a mini-Tn5 element with kanamycin as its selectable marker [*KmR*], and an origin of replication from plasmid R6K [*oriR6K*] to allow for subsequent transposon insertion site cloning.

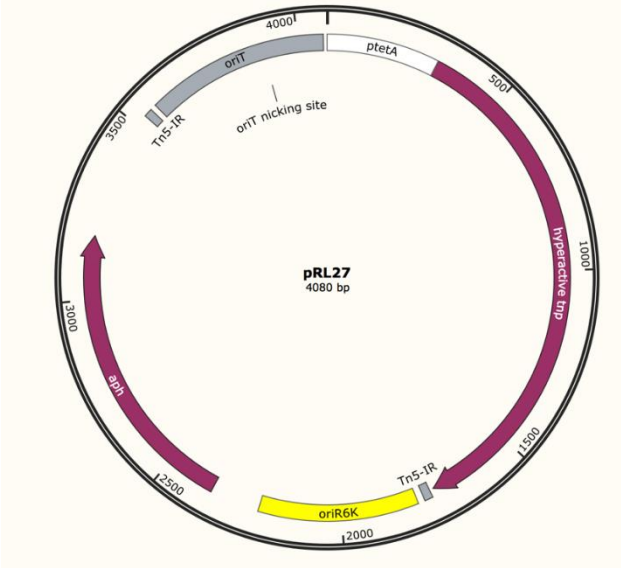


Figure 5.1. The pRL27 construct, taken from Chegg© Inc. Plasmid contains: origin of transfer [*oriT*]; hyperactive transposase [*tnp*] controlled by a *tet*A promotor [p*tetA*]; mini-Tn5 element ("Tn5-RL27") encoding an origin of replication [*oriR6K*], a selectable kanamycin marker [*aph*], encompassed by transposon inverted repeats [Tn5-IR].

The pRL27 construct also has the origin of transfer [*oriT*] from plasmid RP4, which allows for the option of mobilisation from donor to target via conjugation (Larsen *et al*., 2002). Both *tnp* and the transposon (Tn5-RL27) are mutated to increase activity by 1000-fold, compared to wild-type Tn5; *tnp* mutations EK54 and LP372 increase OE binding of the transposon and dimerization respectively, and the Tn5-RL27

inverted repeats have been optimised for EK54 with the creation of a OE/IE mosaic end sequence containing specific OE base-pairs (Weinreich *et al*., 1994; Zhou and Reznikoff, 1997; Zhou *et al.*, 1998). The *tnp* gene is lost after transposition as the p*tetA*::*tnp* gene construct is outside of the transposon, meaning that the insertions are stable (Larsen *et al*., 2002).

pRL27 (Appendix Figure 4) was extracted from the *E. coli* strain BW20767 using an Invitrogen™ PureLink™ HiPure Plasmid Midiprep Kit (Larsen *et al*., 2002). Overnight cultures of BW20767/pRL27 were grown in 100 mL LB with 50 μg/mL kanamycin, shaking at 37°C. The extraction was conducted using the supplier-provided protocol, but was modified from step 9 as follows: the pellet was re-suspended in 500 μL pure H_2O and transferred to a new 1.5 mL Eppendorf tube (Eppendorf Ltd.). 50 μL 3 M sodium acetate at pH 5.2 and 370 μL isopropanol was added. The mixture was spun in a microfuge at maximum speed for 15 min at 4°C, and the supernatant discarded. 1 mL cold 70% ethanol was added, being careful not to disturb the pellet, and spun again in a microfuge at maximum speed for 15 min at 4°C. The supernatant was discarded, the pellet air-dried and then re-suspended in 100 μL pure H_2O. If needed, excess was stored at -20°C.

5.2.1.2. Transformation of *Sodalis glossinidius* with pRL27

200 mL [split across four 50 mL volumes in 25 cm^2 cell culture flasks (Greiner)] Schneider's Insect Medium (Sigma-Aldrich) + 10% FBS (Gibco) was inoculated with *S. glossinidius* strain SgGMMB1 and statically incubated at 25°C to mid-log phase. Prior to a 10-min ice-chill step, the four lots of 50 mL cultures were transferred into four 50 mL falcon tubes. 100 μL from each culture was spread-plated on horse blood (TCS Biosciences) Columbia agar (Sigma-Aldrich) plates [to check for contamination]. Following the 10-min ice-chill step, each 50 mL culture underwent their first wash: cultures were centrifuged at 8000 rpm at 4°C for 10 min and the supernatant discarded, the pellets gently re-suspended in 1 mL of ice-cold [sterile] 10% glycerol, the number of falcon tubes reduced by half via transferring two of the 1 mL re-suspensions into the remaining two and lastly, the final volumes

made up to 50 mL using ice-cold [sterile] 10% glycerol (100 mL final volume split across two 50 mL falcon tubes). The second wash commenced in the same way as the first, including the halving of the re-suspended pellet volume: the two 1 mL pellet re-suspensions were made up to 50 mL using ice-cold [sterile] 10% glycerol in one falcon tube. The third and final wash commenced in the same way as the first and second, except that the 1 mL initial re-suspension was not made up to 50 mL and instead transferred into a pre-chilled 1.5 mL Eppendorf tube and then centrifuged at 8,000 rpm at 4°C for 10 min. The supernatant was removed, and the pellet re-suspended in 500 µL of ice-cold [sterile] 10% glycerol. The 500 µL suspension was then aliquoted into 50 µL in pre-chilled, 2 mm gap electroporation cuvettes (Geneflow©). The appropriate volume of pRL27 was added to each 50 µL cuvette for it to be at a plasmid concentration of 3 µg, per transformation; a NanoDrop® ND-1000 Spectrophotometer concentration [pRL27] reading of 0.8 µg/µL would mean adding 3.75 µL plasmid, for example. One cuvette per transformation did not contain plasmid, but instead the same volume of 10% glycerol [negative control]. The electroporation cuvettes were chilled on ice for 10 min and then individually transformed using a Bio-Rad© Gene Pulser™ by electroporation (2.5 kV, 25 µFD, 200 Ω). Immediately after each transformation, the samples were re-suspended in 1 mL room-temperature Schneider's Insect Medium + 10% FBS, transferred to 1.5 mL Eppendorf tubes and statically incubated at 25°C ['recovery culture(s)']. 10 µL from each post-electroporation suspension, including the negative control, were spot-plated on horse blood Columbia agar plates with 50 µg/mL kanamycin. After 20 hours incubation, 10 µL from each recovery culture, including the negative control, were spot-plated on horse blood Columbia agar plates with 50 µg/mL kanamycin. The negative control suspensions were additionally plated in the same way, but on horse blood Columbia agar plates without antibiotics. The remaining volumes were seeded into individual 9 mL Schneider's Insect Medium + 10% FBS at 50 µg/mL kanamycin. After mid-log phase was reached, the individual cultures were spun down in a microfuge (maximum speed for five min) and re-suspended, where 500 µL was aspirated, spun down in a microfuge (maximum speed for five min) and stocked in -80°C as a pellet, and 1 mL

for 10% glycerol stocks for -80°C storage. 10 µL from each culture was also spot-plated on horse blood Columbia agar plates to check for contamination, whereby if positive, the stocks were discarded.

Based on counting colonies, 10,600 mutants were created. This does not include one of the transformation events that failed to produce countable colonies, so the actual mutant input number is larger.

5.2.1.3. Confirmation of Transposon Insertion

In order to confirm transposon insertion in the *S. glossinidius* mutants, PCR was conducted on the portion of the transposon that expresses kanamycin resistance, using the stocked pellets (detailed in 5.2.1.2.) as input. A master mix was created for the samples as outlined in Table 5.1, where 20 µL was used per 1 µL sample pellet.

Table 5.1. Components used within the master mix for a PCR reaction targeting the kanamycin resistance cassette within the transposon. The first column lists the components used within the master mix. The second column lists the volumes that would be added into one sample. The third column lists the volumes (adjusted for the number of samples to be included in the PCR reaction) of the master mix components added together to create the master mix, 20 µL of which was used per sample.

Master Mix Component	Volume per Sample (µL)	Volume Adjusted for Number of Samples (µL)
2 X MyTaq™ Red Mix (Bioline)	10	330
Forward Primer (at 10µM)	1	33
Reverse Primer (at 10µM)	1	33
H_2O	8	264

The samples selected for the PCR reaction (Table 5.2) were random and used as representatives for all the samples electroporated within the same event; multiple transformation events occurred on different dates in order to generate a minimum of 10,000 *S. glossinidius* Tn5 mutants.

Table 5.2. The representative samples used within the PCR reaction targeting the kanamycin resistance cassette within the transposon. The first column lists the various transformation events. The second column lists the label associated with each sample; the pre-decimal integer refers to the washed, electrocompetent *Sodalis glossinidius* cell aliquot used and the post-decimal integer refers to the cuvette replicate number of the associated aliquot.

Transformation Event	Sample Label
1	n/a
2	2.1
	2.4
	2.8
	3.1
	3.3
	3.6
	4.3
	4.5
	4.8
3	2.2
4	1.2
	1.4
	1.6
	2.4
	2.6
	2.8
5	3.3
	3.5
	3.7
	4.5
	4.7
	4.9
6	1.3
	1.7
	1.9
	2.2
	2.5
	2.8
7	- ve C (WT *Sodalis glossinidius* DNA)
	+ ve C (pre-confirmed, successfully-transformed, pRL27-donated-Tn5 *Sodalis glossinidius*)

The PCR cycle used involved an initial two min denaturing step at 95°C, 35 cycles of denaturing-annealing-elongation steps of 15 seconds at 95°C-30 seconds at 55°C-30 seconds at 72°C, followed by a final elongation step of five min at 72°C. The

amplicons were run on a 1% agarose gel with 4 µL Midori Green and run for 30 min at 100 V (Figure 5.2).

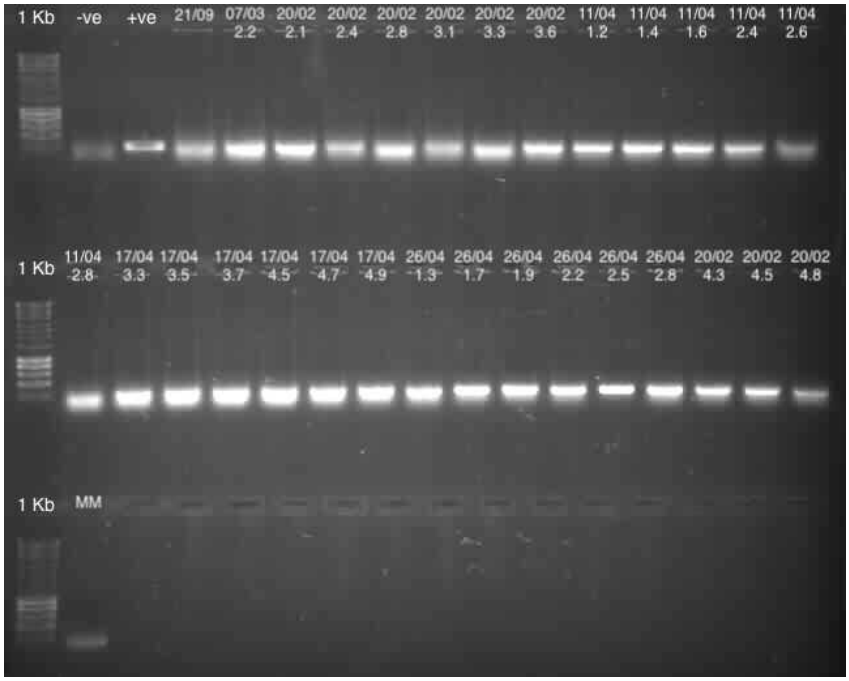


Figure 5.2. A gel image showing the representative sample amplicons used within the PCR reaction targeting the kanamycin resistance cassette within the transposon. A 1 Kb ladder was used in order to visualise the sample amplicon bands at the expected 200 bp size. As can be seen, all the transformation samples (outlined in Table 5.2) and the positive control display a band at the 200 bp size, indicating positive amplification of the targeted transposon region. The negative control and master mix (MM)-only sample display a faint, smeared band below the 200 bp level, indicating a failed amplification of the targeted transposon region.

5.2.2. Selection Pressures for the Transposon-Directed Insertion Site Sequencing (TraDIS) *Sodalis glossinidius* Library

A graphical overview of the TraDIS experimental design is displayed in Figure 5.3, but fully detailed in 5.2.2.1. – 5.2.2.4..

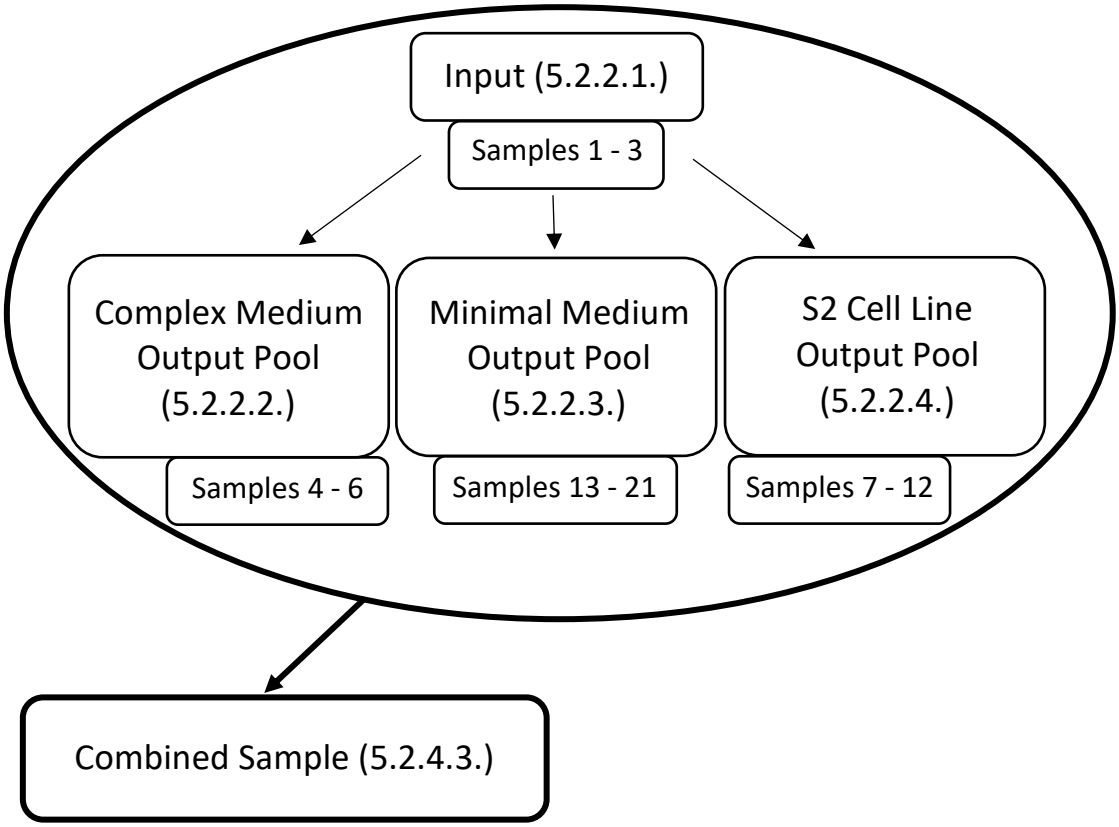


Figure 5.3. A graphical overview of the *Sodalis glossinidius* transposon-directed insertion site sequencing (TraDIS) experimental design. The input samples were used as inoculant for the complex medium-, minimal medium-, and S2 cell line output pools. The combined sample was computationally created by combining all experimental samples. The bracketed numbers refer to the chapter sections where the methodology is fully detailed for the corresponding pool(s).

5.2.2.1. Input Pool

Single 1 mL stock replicates from all transformations were pooled together (72 mL total from 72 stocks) in a 175 cm^2 cell culture flask (Sarstedt) and made up to 150 mL using Schneider's Insect Medium (Sigma-Aldrich) + 10% FBS (Gibco). 50 mL of the pool was transferred into a 175 cm^2 cell culture flask with 50 mL

Schneider's Insect Medium + 10% FBS at 50 μg/mL kanamycin [arbitrarily named "Mixed Culture 1" here for protocol-following ease] and statically incubated at 25°C to mid-log phase. 9 mL was aspirated from the pool in order to make 15 lots of 1 mL stocks (600 μL pool culture + 400 μL 50% glycerol) to store at -80°C. All 15 were kept as back-ups in the case of contamination, which were split and arbitrarily pre-designated for the various selections [re-culturing the pool, the media-based outputs and the S2 outputs]. A further 1 mL was used to create a $10^0 – 10^6$ dilution series, which was spot-plated (10 μL per spot) on horse blood (TCS Biosciences) Columbia agar (Sigma-Aldrich) plates both with and without 50 μg/mL kanamycin. The remaining pool culture was split across two 50 mL falcon tubes and spun in a centrifuge at 4000 rpm for 15 min. The supernatant was removed and the pellets re-suspended and combined into one 600 μL volume of Schneider's Insect Medium + 10% FBS, which was combined with 400 μL 50% glycerol and stocked at -80°C ["Sample 1"].

50 mL of Mixed Culture 1 was passaged into a 175 cm^2 cell culture flask with 50 mL Schneider's Insect Medium + 10% FBS [arbitrarily named "Mixed Culture 2" here for protocol-following ease]. 1 mL of the 100 mL was used to create a $10^0 – 10^6$ dilution series, which was spot-plated (10 μL per spot) on a horse blood Columbia agar plate. The remaining 49 mL [of Mixed Culture 1] was spun in a centrifuge at 4000 rpm for 15 min. The supernatant was removed, the pellet re-suspended in 600 μL Schneider's Insect Medium + 10% FBS, combined with 400 μL 50% glycerol and stocked at -80°C ["Sample 2"].

Once Mixed Culture 2 reached mid-log phase, 1 mL was aspirated for use in the complex medium pool outputs [5.2.2.2.], 3 mL for the minimal medium pool outputs [5.2.2.3.], and 36 mL for use in the S2 cell line infection outputs [5.2.2.4.]. A further 1 mL was used to create a $10^0 – 10^6$ dilution series, which was spot-plated (10 μL per spot) on a horse blood Columbia agar plate. The remaining 59 mL [of Mixed Culture 2] was spun in a centrifuge at 4000 rpm for 15 min. The supernatant was removed, the pellet re-suspended in 600 μL Schneider's Insect Medium + 10% FBS, combined with 400 μL 50% glycerol and stocked at -80°C ["Sample 3"].

5.2.2.2. Complex Medium Output Pool

The designated 1 mL [5.2.2.1.] was passaged into a 175 cm^2 cell culture flask (Sarstedt) with 99 mL Schneider's Insect Medium (Sigma-Aldrich) + 10% FBS (Gibco) and statically incubated at 25°C to mid-log phase [arbitrarily named "Mixed Culture 3" here for protocol-following ease]. 1 mL [of Mixed Culture 3] was passaged into a 175 cm^2 cell culture flask with 99 mL Schneider's Insect Medium + 10% FBS [arbitrarily named "Mixed Culture 4" here for protocol-following ease]. A further 1 mL [of Mixed Culture 3] was used to create a $10^0 – 10^6$ dilution series, which was spot-plated (10 μL per spot) on a horse blood (TCS Biosciences) Columbia agar (Sigma-Aldrich) plate. The remaining 98 mL [of Mixed Culture 3] was spun in a centrifuge at 4000 rpm for 15 min. The supernatant was removed, the pellets re-suspended and combined into one 600 μL volume of Schneider's Insect Medium + 10% FBS, which was combined with 400 μL 50% glycerol and stocked at -80°C ["Sample 4"].

1 mL mid-log phase Mixed Culture 4 was passaged into a 175 cm^2 cell culture flask with 99 mL Schneider's Insect Medium + 10% FBS [arbitrarily named "Mixed Culture 5" here for protocol-following ease]. 1 mL [of Mixed Culture 4] was used to create a $10^0 – 10^6$ dilution series, which was spot-plated (10 μL per spot) on a horse blood Columbia agar plate. The remaining 98 mL [of Mixed Culture 4] was spun in a centrifuge at 4000 rpm for 15 min. The supernatant was removed, the pellets re-suspended and combined into one 600 μL volume of Schneider's Insect Medium + 10% FBS, which was combined with 400 μL 50% glycerol and stocked at -80°C ["Sample 5"].

1 mL of Mixed Culture 5 was used to create a $10^0 – 10^6$ dilution series, which was spot-plated (10 μL per spot) on a horse blood Columbia agar plate. The remaining 99 mL was spun in a centrifuge at 4000 rpm for 15 min. The supernatant was removed, the pellets re-suspended and combined into one 600 μL volume of Schneider's Insect Medium + 10% FBS, which was combined with 400 μL 50% glycerol and stocked at -80°C ["Sample 6"].

5.2.2.3. Minimal Medium Output Pool

The designated 3 mL [5.2.2.1.] were separately passaged into three separate 175 cm^2 cell culture flasks (Sarstedt) containing 99 mL M9 (Miller, 1972) + yeast extract [5 g/L], $M9^{g-}$ + yeast extract [5 g/L], or $M9^{g-}$+ yeast extract [5 g/L] + *N*-acetylglucosamine [5 g/L] respectively and statically incubated at 25°C to mid-log phase [arbitrarily named "Mixed Cultures 6 - 8" here for protocol-following ease]. 1 mL [of Mixed Cultures 6 – 8] were separately passaged into individual 175 cm^2 cell culture flasks with 99 mL of the corresponding M9 variation [arbitrarily named "Mixed Cultures 9 - 11" here for protocol-following ease]. 1 mL volumes [of Mixed Cultures 6 – 8] were aspirated to use in the creation of individual variation 10^0 – 10^6 dilution series', which were spot-plated (10 µL per spot) on separate horse blood (TCS Biosciences) Columbia agar (Sigma-Aldrich) plates. The remaining 98 mL volumes [of Mixed Cultures 6 – 8] were split across two falcon tubes per M9 variation and spun in a centrifuge at 4000 rpm for 15 min. Per each M9 variation, the supernatants were removed and the two pellets re-suspended and combined into one 600 µL volume of Schneider's Insect Medium (Sigma-Aldrich) + 10% FBS (Gibco), and combined with 400 µL 50% glycerol for -80°C stocking ["Samples 13, 14 and 15", respectively].

1 mL of mid-log phase Mixed Cultures 9 - 11 were separately passaged into 175 cm^2 cell culture flasks with 99 mL of the corresponding M9 variation and statically incubated at 25°C to mid-log phase [arbitrarily named "Mixed Cultures 12 - 14" here for protocol-following ease]. 1 mL [of Mixed Cultures 9 – 11] were aspirated to use in the creation of individual variation 10^0 – 10^6 dilution series', which were spot-plated (10 µL per spot) on separate horse blood Columbia agar plates. The remaining 98 mL volumes [of Mixed Cultures 9 – 11] were split across two falcon tubes per M9 variation and spun in a centrifuge at 4000 rpm for 15 min. Per each M9 variation, the supernatants were removed and the two pellets re-suspended and combined into one 600 µL volume of Schneider's Insect Medium + 10% FBS, and combined with 400 µL 50% glycerol for -80°C stocking ["Samples 16, 17 and 18", respectively].

1 mL of mid-log phase Mixed Cultures 12 - 14 were aspirated to use in the creation of individual variation 10^0 – 10^6 dilution series', which were spot-plated (10 μL per spot) on separate horse blood Columbia agar plates. The remaining 98 mL volumes from each M9 variation were split across two falcon tubes per M9 variation and spun in a centrifuge at 4000 rpm for 15 min. Per each M9 variation, the supernatants were removed and the two pellets re-suspended and combined into one 600 μL volume of Schneider's Insect Medium + 10% FBS, and combined with 400 μL 50% glycerol for -80°C stocking ["Samples 19, 20 and 21", respectively].

5.2.2.4. S2 Cell Line Output Pool

Six replicate 6-well cell culture plates (Sarstedt) were infected with the designated 36 mL [5.2.2.1.]. Prior to infection, all 36 wells were incubated with 0.5 mg/mL concanavalin A for 10 min and then air-dried with UV exposure. S2 cells were then transferred to the wells at 10^6 [based on haemocytometer measurements] in 2 mL Schneider's Insect Medium (Sigma-Aldrich) + 10% FBS (Gibco) and left to settle in the concanavalin A-treated wells for 3 days. The infected 6-well cell culture plates were spun in a centrifuge at 1000 rpm for five min and the cell medium was removed. Schneider's Insect Medium + 10% FBS was combined with transformed *S. glossinidius* at an MOI of 100 in a 2 mL final volume. The volumes were mixed by gently rotating the plates by hand, spun in a centrifuge at 1000 rpm for five min and then incubated at 25°C for 1 hour. After the incubation, 2 μL 100 mg/mL gentamicin was added to the 2 mL well volumes for a working concentration of 100 μg/mL and the plates gently mixed by hand rotation. The plates were then spun in a centrifuge at 1000 rpm for five min and incubated at 25°C for 1 hour. After the second incubation, the plates were spun in a centrifuge at 1000 rpm for five min and the medium removed from all 36 wells. 2 mL Schneider's Insect Medium + 10% FBS with 10 μg/mL gentamicin was added to each well and the plates were incubated statically at 25°C for 48 hours. Cell scrapers were then used to dislodge the cells from the wells and the individual well cultures pooled per plate in individual 15 mL falcon tubes [6 total falcon tube 12 mL culture

pools for the 6 replicate plates]. The plate pools were vortexed and then spun in a centrifuge at 8000 rpm for ten min. The supernatants were removed, the pellets vigorously re-suspended in 1 mL Schneider's Insect Medium + 10% FBS and then vortexed. 600 μL from each of the 6 plate pools were transferred into induvial 400 μL 50% glycerol suspensions and stocked at -80°C ["Samples 7, 8, 9, 10, 11 and 12", respectively]. The remaining volumes were used to create individual $10^0 – 10^6$ dilution series', which were spot-plated (10 μL per spot) on separate horse blood Columbia agar plates.

5.2.3. Library Preparation for Sequencing

DNA from the 21 samples was individually extracted using the Zymo *Quick*-DNA™ Miniprep Plus Kit following the manufacturer's instructions, except for eluting the DNA in 103 μL pure H_2O. 3 μL from each sample was used for quantification using both a NanoDrop™ 1000 Spectrophotometer, in addition to a Qubit™ dsDNA HS Assay Kit and Invitrogen™ Qubit™ 4 Fluorometer. Based on the readings, the appropriate volume was taken from the remaining 100 μL and diluted to a concentration of 1 μg in pure H_2O for a final volume of 100 μL. A Diagenode Bioruptor® Pico was used to shear the adjusted-concentration samples to 300 bp using an on-off cycle of 15 seconds-90 seconds at 4°C. The sheered samples were transferred to a 96-well plate and underwent a 1.8 X AMPure XP bead size selection, as follows: 180 μL of beads were added to each of the 100 μL samples and incubated at room temperature for five min. The plate was then placed onto a magnet, where once the beads had cleared from the solution, the supernatant was removed. 200 μL 80% ethanol was added to each sample, incubated for 30 seconds and then removed. A further 200 μL 80% ethanol was added to each sample and incubated for 30 seconds. The supernatant was removed and the beads allowed to air-dry for exactly five min. The plate was removed from the magnet, the beads were re-suspended in 57.5 μL H_2O and incubated for two min. The plate was then placed back onto the magnet, where once the beads had cleared from the solution, the inline indices were added as outlined in Table 5.3.

Table 5.3. A table showing which inline index was used for which sample during the library preparation stage of the *Sodalis glossinidius* transposon library data set

Inline Index 1 (GTACGT)	Inline Index 2 (AGCTAGC)	Inline Index 3 (CATGCATG)	Inline Index 4 (TCGATCGAT)	Inline Index 5 (GGTAGGTCTA)
Samples 1, 6, 11, 16, 21	Samples 2, 7, 12, 17	Samples 3, 8, 13, 18	Samples 4, 9, 14, 19	Samples 5, 10, 15, 20

55 μL of each sample was then transferred into a 96-well PCR plate to go forward into the NEBNext® Ultra™ DNA Library Prep Kit for Illumina® (New England Biolabs, Inc.©), which was followed up to step 1.3A, as follows: the 21 size-selected samples first underwent NEBNext® End Prep (step 1.1), where they were each mixed with 3 μL (green) End Prep Enzyme Mix and 6.5 μL (green) End Repair Reaction Buffer (10 X) (New England Biolabs, Inc.©). The 96-well PCR plate was run in a thermocycler first at 20°C for 30 min, then at 65°C for 30 min. The samples then moved onto Adapter Ligation (step 1.2), where 15 μL (red) Blunt/TA Ligase Master Mix, 2.5 μL (red) NEBNext Adaptor for Illumina® and 1 μL (red) Ligation Enhancer were mixed into each sample in the 96-well PCR plate (New England Biolabs, Inc.©). This was run in a thermocycler at 20°C for 15 min, where 3 μL (red) USER™ Enzyme (New England Biolabs, Inc.©) was then added and then the plate run for a further 15 min at 37°C. The samples were transferred to a 96-well plate and underwent Size Selection of Adapter Ligated DNA (step 1.3A): 13.5 μL H_2O was added to each sample to make the volume up to 100 μL, in addition to 55 μL AMPure XP beads. The plate was incubated at room temperature for five min and then placed onto a magnet. Once the solutions were clear of beads, the supernatant was transferred to a new 96-well plate, where 25 μL beads were added. The new plate was incubated for five min and then placed onto a magnet, where once the beads had cleared from the solution, the supernatant was removed. 200 μL 80% ethanol was added to each sample, incubated for 30 seconds and then removed. A further 200 μL 80% ethanol was added to each sample and incubated for 30 seconds. The supernatant was removed, and the beads allowed to air-dry for exactly five min. The plate was

removed from the magnet, the beads were re-suspended in 16 µL H_2O and incubated for two min. The plate was then placed back onto the magnet, where once the beads had cleared from the solution, 14 µL of each sample was transferred into a 96-well PCR plate. The samples then underwent two rounds of PCR, as follows: the first round included 25 µL KAPA HiFi, 6 µL H_2O and 2.5 µL both forward and reverse primer (Table 5.4) to every 14 µL sample.

Table 5.4. A table listing the primers used during library preparation (1st round PCR), prior to next-generation sequencing on the Illumina® HiSeq 4000 System. The forward primer (Nadal_For) used to amplify fragments containing the transposon sequence (in green) and the reverse primer (Nadal_Rev) used to amplify fragments containing part of the NEBNext® Adaptor (in *blue italics*) and add in part of the read 2 priming site (in blue).

Primer Label	Primer Sequence
Nadal_For	ATCCTCTAGAGTCGACCTGCAGGCATG
Nadal_Rev	GACTGGAGTTCAGACGTGT*GCTCTTCCGATC*

The first round PCR cycle used involved an initial 48 second denaturing step at 98°C, 10 cycles of denaturing-annealing-elongation steps of 15 seconds at 98°C-30 seconds at 65°C-30 seconds at 72°C, followed by a final elongation step of 60 seconds at 72°C. The amplicons from the first round PCR underwent a 0.75 X AMPure XP bead size selection (as previously detailed) and were subsequently used an input for the second round PCR: 25 µL KAPA HiFi, 5 µL H_2O, 2.5 µL both forward and reverse primer (Table 5.5) to every 15 µL sample were run in a thermocycler using an initial 48 second denaturing step at 98°C, 15 cycles of denaturing-annealing-elongation steps of 15 seconds at 98°C-30 seconds at 65°C-30 seconds at 72°C, followed by a final elongation step of 60 seconds at 72°C. The amplicons from the second round PCR underwent a 0.75 X AMPure XP bead size selection (as previously detailed) and were eluted in 20 µL H_2O. The samples were quantified using a Qubit™ dsDNA HS Assay Kit and Invitrogen™ Qubit™ 4 Fluorometer and the library size ascertained on the Agilent 5200 Fragment Analyzer using the Agilent HS

NGS Kit (Agilent Technologies, Inc.©). The libraries were pooled in equimolar amounts using both the Qubit™ and Fragments Analyzer data, the pool size-selected for 200 – 800 bp using the Sage Pippin Prep (Sage Science, Inc.©) and then treated to remove free adaptors using the Illumina® Free Adaptor Blocking Reagent (Illumina, Inc.©). Both the quantity and quality of the resultant final pool was again assessed on the Agilent 5200 Fragment Analyzer (Agilent Technologies, Inc.©) and subsequently by qPCR using the KAPA Library Quantification Kit on a LightCycler® 480 (Roche Molecular Systems, Inc.©), according to manufacturer's instructions. The pool was sequenced on the Illumina® HiSeq 4000 System (Illumina, Inc.©).

Table 5.5. A table listing the primers used during library preparation (2nd round PCR), prior to next-generation sequencing on the Illumina® HiSeq 4000 System. The forward primers (Nadal_Enrich_F1 – 5) amplify fragments containing the same part of the NEBNext® Adaptor (in *italics*) and transposon sequence (in green) amplified in the 1st round (Table 5.4), but also incorporate the P5 adaptor (in **bold**) and inline indices 1 – 5 (in red) ("NEBNext® Universal Primer"). The reverse primers (Rev_UDI0001 – 0021) amplify fragments containing the same part of the NEBNext Adaptor (in *blue italics*) amplified in the 1st round (Table 5.4), but also incorporate the P7 adaptor (in orange), unique sequencing indices (in red) and the rest of the read 2 priming site (in blue) ("NEBNext® Indexed primers").

Nadal_Enrich_F1	**AATGATACGGCGACCACCGAGATCT***ACACTCTTTCCCTACACGACGCTCTTCCGATCT*GTACGTGACCTGCAGGCATGCAAGCTTCAG
Nadal_Enrich_F2	**AATGATACGGCGACCACCGAGATCT***ACACTCTTTCCCTACACGACGCTCTTCCGATCT*AGCTAGCGACCTGCAGGCATGCAAGCTTCAG
Nadal_Enrich_F3	**AATGATACGGCGACCACCGAGATCT***ACACTCTTTCCCTACACGACGCTCTTCCGATCT*CATGCATGGACCTGCAGGCATGCAAGCTTCAG
Nadal_Enrich_F4	**AATGATACGGCGACCACCGAGATCT***ACACTCTTTCCCTACACGACGCTCTTCCGATCT*TCGATCGATGACCTGCAGGCATGCAAGCTTCAG
Nadal_Enrich_F5	**AATGATACGGCGACCACCGAGATCT***ACACTCTTTCCCTACACGACGCTCTTCCGATCT*GGTAGGTCTAGACCTGCAGGCATGCAAGCTTCAG
Rev_UDI0001	CAAGCAGAAGACGGCATACGAGATAACCGCGGGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0002	CAAGCAGAAGACGGCATACGAGATGGTTATAAGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0003	CAAGCAGAAGACGGCATACGAGATCCAAGTCCGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0004	CAAGCAGAAGACGGCATACGAGATTTGGACTTGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0005	CAAGCAGAAGACGGCATACGAGATCAGTGGATGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0006	CAAGCAGAAGACGGCATACGAGATTGACAAGCGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0007	CAAGCAGAAGACGGCATACGAGATCTAGCTTGGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0008	CAAGCAGAAGACGGCATACGAGATTCGATCCAGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0009	CAAGCAGAAGACGGCATACGAGATCCTGAACTGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0010	CAAGCAGAAGACGGCATACGAGATTTCAGGTCGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0011	CAAGCAGAAGACGGCATACGAGATAGTAGAGAGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0012	CAAGCAGAAGACGGCATACGAGATGACGAGAGGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0013	CAAGCAGAAGACGGCATACGAGATAGACTTGGGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0014	CAAGCAGAAGACGGCATACGAGATGAGTCCAAGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0015	CAAGCAGAAGACGGCATACGAGATCTTAAGCCGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0016	CAAGCAGAAGACGGCATACGAGATTCCGGATTGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0017	CAAGCAGAAGACGGCATACGAGATCTGTATTAGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0018	CAAGCAGAAGACGGCATACGAGATTCACGCCGGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0019	CAAGCAGAAGACGGCATACGAGATACTTACATGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0020	CAAGCAGAAGACGGCATACGAGATGTCCGTGCGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*
Rev_UDI0021	CAAGCAGAAGACGGCATACGAGATAAGGTACCGTGACTGGAGTTCAGACGTGTG*CTCTTCCGATC*T*

5.2.4. Bioinformatic Analysis

5.2.4.1. Initial Quality Control

Prior to analysis, reads underwent demultiplexing based on sequencing indices, then sequencing adapter and quality trimming: briefly, raw FASTQ files were trimmed to remove Illumina® adapter sequences using Cutadapt version 1.2.1 (Martin, 2011). The option "-O 3" was set, so the 3' end of any reads which matched the adapter sequence over at least 3 bp was trimmed off. The reads were further trimmed to remove low quality bases, using Sickle version 1.200 with a minimum window quality score of 20. After trimming, reads shorter than 20 bp were removed. If both reads from a pair passed this filter, each was included in the R1 (forward reads) or R2 (reverse reads) file. Unpaired reads were not used in further analysis. The above steps conducted by the Centre for Genomic Research.

5.2.4.2. Bio-TraDIS Pipeline Overview

The Bio-TraDIS pipeline is designed to identify, with statistical significance [where possible], essential genes within a bacterial transposon library (Barquist *et al*., 2016). There are multiple steps within said pipeline, including preparation of the reads, alignment of the reads to a reference genome, followed by statistical scoring of genes based on the frequency of detected insertion sites. The Bio-TraDIS pipeline comprises a set of scripts for processing FASTQ sequence data through to determining putative essential genes (Barquist *et al*., 2016). During the PCR stage of the library preparation, inline indices are incorporated using primers that target the 3' end of the transposon and prime into chromosome sequence. The analysis pipeline carries out the following steps: the raw reads are filtered with the *filter_tradis_tags* script, to retain only those that start with the transposon tag and specific inline index (Barquist *et al*., 2016). The tag and index are subsequently trimmed from the reads with the *remove_tradis_tags* script (Barquist *et al*., 2016). Both within the *filter_tradis_tags* and *remove_tradis_tags* scripts, the user specifies the number of mismatches allowed within the transposon tag. Filtered and trimmed reads are then mapped to the target organism reference genome using

SMALT (Genome Research Ltd.), to generate an alignment BAM file. The positions of alignments with mapping quality of 30+ are ascertained using the *tradis_plot* script (Barquist *et al*., 2016). Based on the reference genome annotations, gene insertion sites are counted for each gene with the *tradis_gene_insertion_sites* script, which outputs tables showing the number and frequency of gene insertions, based on clustering of nearby mapping sites (Barquist *et al*., 2016). Finally, the *tradis_essentiality.R* script is used to estimate essential genes (Barquist *et al*., 2016). All the detailed scripts are run through a wrapper script, *bacteria_tradis* (Barquist *et al*., 2016). Running the *bacteria_tradis* wrapper script directly from the command line, having not selected for any of the additional parameters, has only 3 essential parameters: the file containing a list of FASTQ files (transposon library sequencing forward reads; reverse reads are not utilised); the sequence of the transposon tag to search for, and the reference genome FASTA file (Barquist *et al*., 2016). There are several parameters that are set to a default: the first FASTQ file read length governs the subsequent mapping parameters for SMALT (Ponstingl and Ning, 2010), such as kmer value, step size and percent identity; the number of mismatches allowed within the transposon tag is set to 0; the quality score used for downstream mapping analysis is set to a minimum of 30 [99.9% confidence in the mapping position]. The pipeline also offers the options to ignore the last 10% of the gene when counting insertions. The Bio-TraDIS pipeline predicts essentiality using bi-modal (essential (mode at 0) and non-essential models) observations of insertion site distribution, after the genes have been normalised based on length; dividing the number of gene insertion sites by the gene length provides an "insertion index" (Langridge *et al*., 2009; Barquist *et al*., 2016). Bio-TraDIS calculates the $\log_2$-likelihood ratios (LR) between the two models and subsequently dictates a gene as statistically essential if the $\log_2$-LR is less than -2; indication of the likelihood that the gene is essential as four times greater to the essential model than the non-essential model (Langridge *et al*., 2009; Barquist *et al*., 2016). If a gene has a $\log_2$-LR greater than 2, then Bio-TraDIS classes it as non-essential (Langridge *et al*., 2009; Barquist *et al*., 2016). Gamma distributions are fitted to the two modes using the R

MASS library (Figure 5.4) (Langridge *et al*., 2009; R Core Team, 2013; Barquist *et al*., 2016).

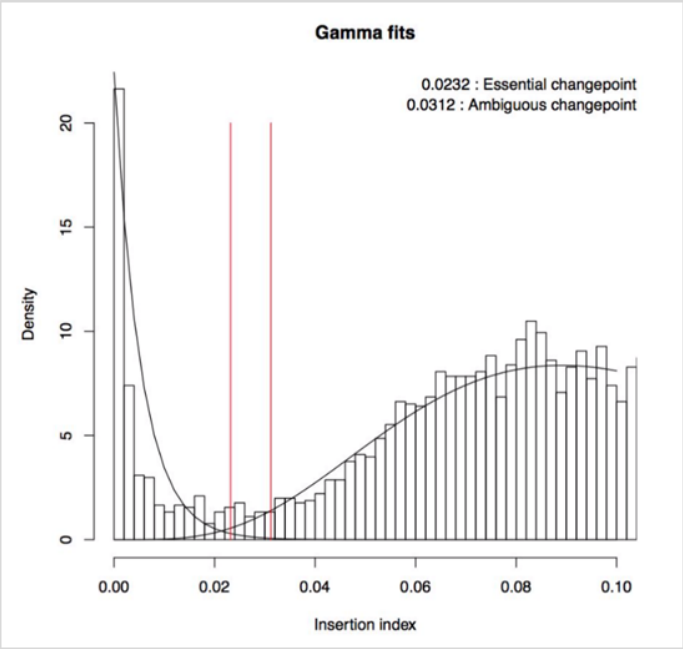


Figure 5.4. Example gamma distribution plot produced within the Bio-TraDIS pipeline graphically displaying the essentiality range of genes. Any genes falling between the 0.0232 – 0.0311 changepoint range [indicated by the red lines] are statistically dictated as essential, according to Bio-TraDIS. Each range is unique to each sample, according to bi-model observations of insertion site distribution. Langridge *et al*. (2009); R Core Team (2013); Barquist *et al*. (2016). **Taken from the GitHub© manual (Barquist *et al*., 2016).**

There are various final output files: essential gene list(s), ambiguous gene list(s), full gene list(s) and QC plots. Additionally produced are: a mapping statistics file, including data such as total number of reads, number and percentage of reads that mapped and the total number of unique insertion sites; insertion site plots containing insertion counts per nucleotide position; BAM files containing the mapped reads.

5.2.4.3. Bio-TraDIS Adaptation for the *Sodalis glossinidius* TraDIS Library

While convenient, the *bacteria_tradis* wrapper was not used as it lacked flexibility; only single-end reads can be processed using this script, meaning that the

information for the mate read is not utilised (Barquist *et al*., 2016). Therefore, individual modules were selected and an adapted method was created, as follows: the *filter_tradis_tags* and *remove_tradis_tags* scripts were used as previously detailed [5.2.4.2.] (Barquist *et al*., 2016). Then, any selected R1 reads shorter than 30 bp were removed, as these were more likely to ambiguously map to the reference genome and skew the data set. The filtered R1 reads were re-paired with their corresponding read mates and aligned to the reference genome, using SMALT (Ponstingl and Ning, 2010). Unpaired reads were not used in this analysis. The R1 read mate [R2 read] alignments were removed using SAMtools (Li *et al*., 2009; Li, 2011). The SMALT 'sample' functionality was used to estimate the library insert size, read pairs were aligned to the reference genome using SMALT with provided insert size information, and the *tradis_plot* script was used as previously detailed [5.2.4.2.] to identify alignment positions with quality 30+ (Ponstingl and Ning, 2010; Barquist *et al*., 2016). The *tradis_gene_insertion_sites* script was used as previously detailed [5.2.4.2.] to count gene insertion sites [the last 10% of the gene was ignored, when counting] (Barquist *et al*., 2016).

Paired-end reads were used because the alignment of the second read would help disambiguate alignment locations of any forward reads, which would otherwise be discarded due to mapping to multiple loci (and therefore having a very low mapping quality score), as correct alignment positions are determined by the close mate proximity. 0 – 2 mismatches were allowed within the detection of the transposon tag and sequence because cases were observed where reads did not start with the expected first base, but contained the rest of the expected sequence. The reason for this remains unknown, but by not allowing for mismatches, these reads are then excluded from analysis. Further to this, polymerases can have sequencing- [base read incorrectly] and amplification errors. The last 10% of the gene was ignored, rather than counting insertions within the entire length of the gene, as it has been reported that many genes deemed essential were able to tolerate transposon insertions within the last part of their coding sequence. This would mean that by not ignoring the last 10%, in cases where

there were insertions at the end of a gene, the Bio-TraDIS pipeline would not class them as essential (Barquist *et al*., 2016).

An additional sample was created by combining all 21 individual samples ["combined sample"] and running this through the adapted pipeline, as detailed above [5.2.4.3.]. However, the differences were as follows: after processing the individual 21 sample FASTQ data through the initial QC steps previously detailed, all read pairs from all samples were binned based on matching the first 5 bp of an inline index. The *filter_tradis_tags* and *remove_tradis_tags* scripts were then run on these files, based on the [five] separate inline indices (Barquist *et al*., 2016). The reason for this was to retain the small percentage of reads that were incorrectly assigned to a sample based on sequencing index misassignment (for example, through index-hopping), binning them instead based on inline index assignment only. The resultant files from the *filter_tradis_tags* and *remove_tradis_tags* script runs were then merged back into one combined file and run through the rest of the adapted pipeline, as previously detailed [5.2.4.3.] (Barquist *et al*., 2016).

5.2.4.4. Further Analysis

Using the gene lists produced by Bio-TraDIS [following the adaptations outlined in 5.2.4.3.], further analysis was conducted outside of the pipeline. In this analysis, genes detected allowing for 2 transposon tag mismatches with 0 transposon insertions are referred to as "essential gene candidates", with emphasis on "candidates" considering the issues detailed in 5.3.1.. Initial screening of the essential gene candidates included analysing genes with lengths $\geq$ 1 Kb, followed by the addition of genes between 0.99 – 0.50 Kb; the likelihood of transposons 'missing' a gene when randomly integrating into the genome increases with decreasing gene length, which when considering the *S. glossinidius* strain SgGMMB4 annotation that contains many small 'genes' that more likely to be pseudogenised components of one larger gene, would mean including highly-likely false positives. In this analysis, any genes with lengths shorter than 0.50 Kb were considered too high risk for false positives and not included. Comparisons of

essential gene candidates were conducted between samples, using sample two as the input and combining replicates of the same selection pressure into groups; for example, samples four, five and six were grouped in a comparison as they are replicates of the complex medium output selection. The comparison included identifying which essential gene candidates were shared both within and between groups, thus allowing for the identification of candidates unique to each group. Subsequent analysis included comparing output pool groups, instead of using the input.

Functional annotation of essential gene candidates was carried out using the eggNOG-mapper (emapper) (Huerta-Cepas *et al*., 2017) and the bacterial database (bactNOG), to obtain Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000; Kanehisa *et al*., 2016, 2017) orthologs and Clusters of Orthologous Groups (COG) (Tatusov *et al*., 1997; Kristensen *et al*., 2010; Galperin *et al*., 2015) annotations. KEGG Orthology (KOs) were uploaded into the KEGG Pathway Mapper [which uses the originally annotated genome (Toh *et al*., 2006)] to identify gene essentiality within the KEGG database (Kanehisa and Goto, 2000; Kanehisa *et al*., 2016, 2017). Artemis (Carver *et al*., 2012) was used in order to visualise genes [location, size, counterparts], in addition to RNASeq evidence of transcription (Goodhead *et al.*, 2018). In regard to the RNASeq data, three replicates were taken from three time-points [12, 60, and 160 hours post-inoculation] from axenic Schneider's Insect Medium + 10% FBS *S. glossinidius* strain SgGMMB4 culture(s) and the extracted RNA run on an Illumina® MiSeq System (Goodhead *et al.*, 2018). Reciprocal Basic Local Alignment Search Tool (BLAST®) (National Library of Medicine) searches were conducted in order to compare *S. glossinidius* essential gene candidates to *E. coli* K-12 (Goodall *et al*., 2018), *S. enterica* serovar Typhi (Langridge *et al*., 2009), and *Mycoplasma genitalium* (Glass *et al*., 2006) essential genes; initial reciprocal BLAST®s were conducted upon whole genomes, subsequently filtered based on given essential gene candidate lists. An e-value cut-off threshold was set to $\leq$ 0.0001.

5.3. Results

5.3.1. Bioinformatic Analysis; Bio-TraDIS Adaptation for the *Sodalis glossinidius* Transposon-Directed Insertion Site Sequencing (TraDIS) Library

The *tradis_essentiality.R* script failed to identify essentiality based on its methods (Barquist *et al*., 2016). The most likely reason for this is that the library was too low quality for its bi-modal method of analysing insertion site distribution, after calculating the insertion indices; a relatively low transposon coverage of the genome, in combination with low percentages of retained reads (Appendix Figure 5), would lead to very few insertion sites (Appendix Figure 6) and ultimately low insertion index values, which would be indistinguishable from indices associated with true essential genes. Whilst the percentage of retained reads mapping to the genome does not drop below 62.48% (Appendix Figure 7), the number of those retained reads varies greatly between samples (Figure 5.5).

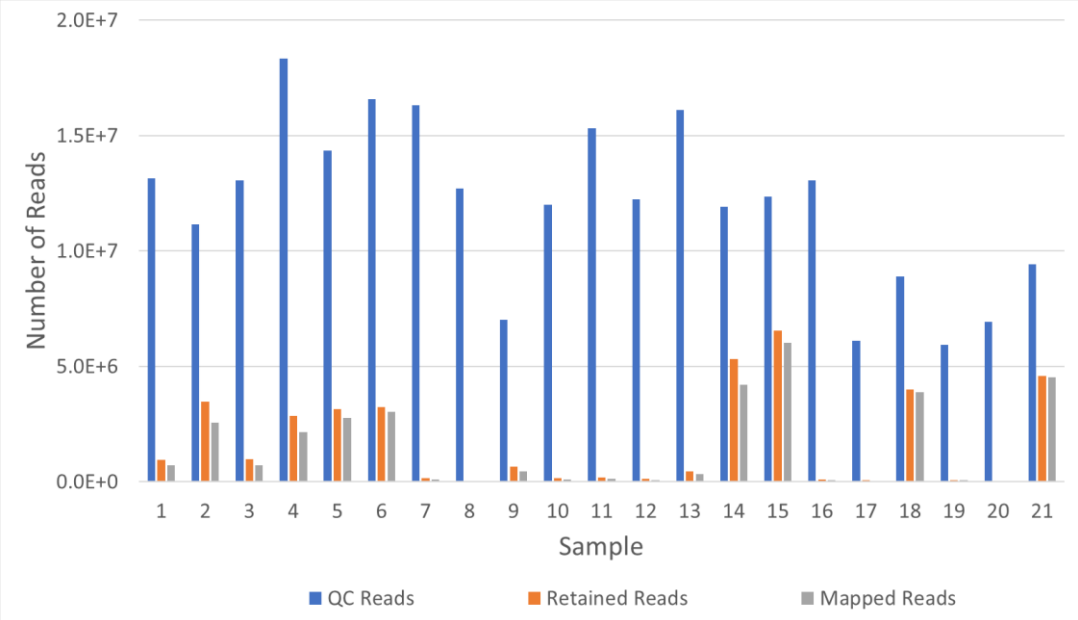


Figure 5.5. *Sodalis glossinidius* library metrics following the Bio-TraDIS pipeline analysis. Blue: number of reads following QC [detailed in 5.2.4.1]. **Orange:** number of retained [i.e. are filtered for the transposon tag] reads. **Grey:** number of [retained] reads mapped to the reference genome.

For example, as can be seen from Figure 5.5, sample 15 retained 6,545,459 [52.97%] of its reads post-QC filtering, and 6,008,208 [91.79%] of those retained reads were mappable to the reference genome. For sample 8, 16,815 [67.4%] of its 24,949 retained reads were mappable to the reference genome, but those retained reads are only 0.2% of the post-QC filtered reads. For samples such as sample 8, the mapped reads cover very little of the genome, making it statistically more likely that genes containing no insertion sites are due to the low complexity of the library [i.e. false positives], rather than being genuinely essential due to containing no transposons. Based on the given reasons, samples that did not retain more than 1 million of their reads post-filtering were deemed low complexity samples and were not included for further analysis; samples 2, 4, 5, 6, 15, 18 and 21 were included within this library analysis (Figure 5.3 and 5.5.).

5.3.2. Essential Gene Candidates for *Sodalis glossinidius in vitro* Growth in a Complex Medium

The results from the adapted Bio-TraDIS pipeline [5.2.4.3.] for the complex medium pool showed an average of 3,301 insertion sites; theoretically one transposon insertion for every 1,272 nucleotides. Average 3,082,185 [18.78%] of the reads were retained post-filtering, and average 2,651,842 [86.04%] of these retained reads mapped to the reference genome. The results from the further analysis [5.2.4.4.] showed that 75.17% (4368/5811) overall genes did not have any transposon insertions in sample four, 79.47% (4618/5811) in sample five, and 78.59% (4567/5811) in sample six (Figure 5.3); 4,121 of these essential gene candidates were shared between the samples (i.e. forming the [complex medium] pool).

5.3.2.1. Essential Gene Candidates for *Sodalis glossinidius* Long-Term Establishment in a Complex Medium

342 essential gene candidates were unique to sample 6, when comparing to the input sample [2] (Figure 5.3); these genes representing essentiality for long-term establishment in Schneider's Insect Medium + 10% FBS (Figure 5.6).

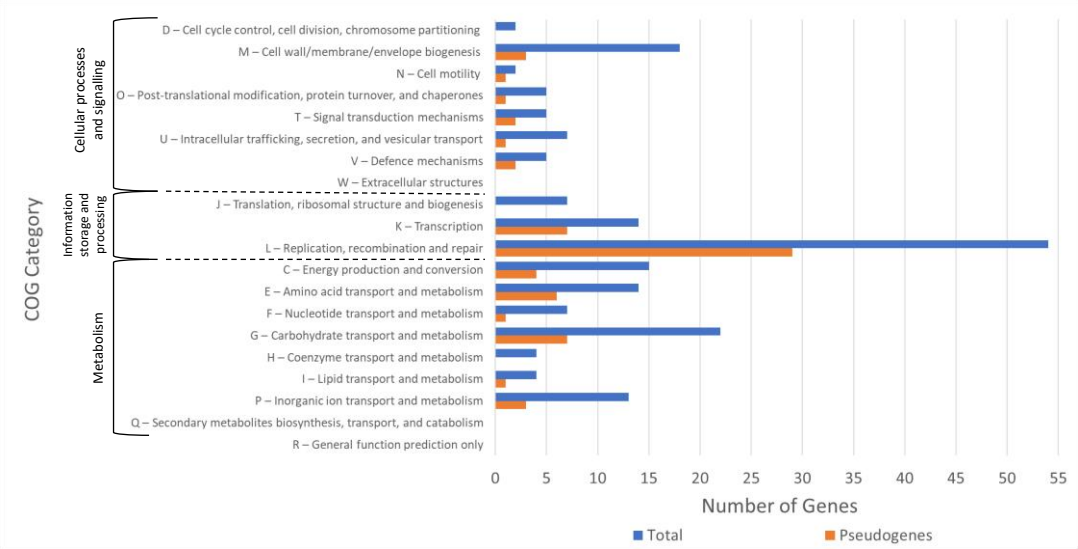


Figure 5.6. Graphical representation of the essential gene candidate COG annotations for *Sodalis glossinidius* long-term establishment in a complex medium. Blue: all genes [including pseudogenes] with 0 insertions. **Orange**: pseudogenes with 0 insertions. COG S [Function Unknown] contained 78 genes, including 21 pseudogenes. 70 genes, including 21 pseudogenes, were unannotated.

As can be seen from Figure 5.6, 43.27% (148/342) of the genes were either unannotated or were assigned to COG S [Function Unknown], highlighting the issues involved in annotating degraded symbiont genomes. 12.87% (44/342) of the unique essential gene candidates were categorised under Cellular Processes and Signalling, 21.93% (75/342) under Information Storage and Processing, and 23.10% (79/342) under Metabolism (Figure 5.6). 22.73% (10/44) of the unique essential gene candidates categorised under Cellular Processes and Signalling were pseudogenised, 48% (36/75) within Information Storage and Processing, and 27.85% (22/79) within Metabolism (Figure 5.6).

The 1 – 0.50 Kb essential gene candidate KO mapper inputs revealed pathway hits expected from mutant selection in such a nutrient rich environment; the complex medium used for this selection [Schneider's Insect Medium + 10% FBS] contains abundant mineral sources, acids important for the CAC, two defined [glucose and trehalose] and multiple undefined carbon sources, and a large range of amino acids (Table 2.1). Hits within glycolysis/gluconeogenesis (sgl00010) [also including pyruvate metabolism (sgl00620)], the PPP (sgl00030), metabolism pathways involving epimers [fructose and mannose (sgl00051)]/poly- and disaccharide constituents [starch and sucrose (sgl00500)], all support the *S. glossinidius* utilisation of glucose. Other known *S. glossinidius*-retained pathways hit included the citrate cycle (TCA cycle) (sgl00020) [also including glyoxylate and dicarboxylate metabolism (sgl00630)] and oxidative phosphorylation (sgl00190). The utilisation of the broad range of amino acids within the medium was supported by hits within the glycine, serine and threonine (sgl00260), alanine, aspartate and glutamate (sgl00250), tyrosine (sgl00350), cysteine [supported by hits within the TcyB component of the cysteine ABC transporter] and methionine (sgl00270), and vitamin B6 (sgl00750) metabolism pathways. Two porin proteins within the osmoregulatory control [two-component] system, OmpC and OmpF were hit, linking into the potential *S. glossinidius* ancestral virulence phenotype, as OmpC (also present in the potentially pathogenic *S. praecaptivus*) and OmpF porins are immunogenic adjuvants in virulent *S.* Typhi, inducing pathogen-associated molecular patterns (PAMP)-like host responses during infection (Isibasi *et al*., 1992; Isibasi *et al*., 1994; Clayton *et al*., 2012; Chari *et al*., 2015; Pérez-Toledo *et al*., 2017). An example of a PAMP includes flagellin, which is a protein that makes up bacterial flagellar (Doetsch and Sjoblad, 1980). Flagellar have strong virulence links as they are vital for pathogen formation of biofilms, and adhesion and invasion of host cells (reviewed in McCarter, 2006). Essentiality hits were observed in flagellar assembly (sgl02040) and the fact that both these and the porins genes were elected as essential gene candidates for long-term establishment in a nutrient-rich environment, supports their virulence link; intracellular environments offer pathogens a relatively abundant source of growth necessities. This concept is

furthered by the hits observed within cationic antimicrobial peptide (CAMP) resistance (sgl01503), as CAMPs are produced by the host immune response to infection. The potential *S. glossinidius* ancestral virulence phenotype is also supported by essentiality hits involved with TTSS [detailed in 1.2.2.1.] (in bacterial secretion system (sgl03070)).

Associated unique genes detailed in Appendix Table 1.

5.3.2.2. Essential Gene Candidates for *Sodalis glossinidius* Growth in a Complex Medium versus a Minimal Medium

812 of the [4,121] shared essential gene candidates were unique to the complex medium pool, when comparing the group to the minimal medium pool [samples 15, 18, and 21] (Figure 5.3); these genes representing essentiality specifically for *S. glossinidius* growth in complex over minimal media (Figure 5.7).

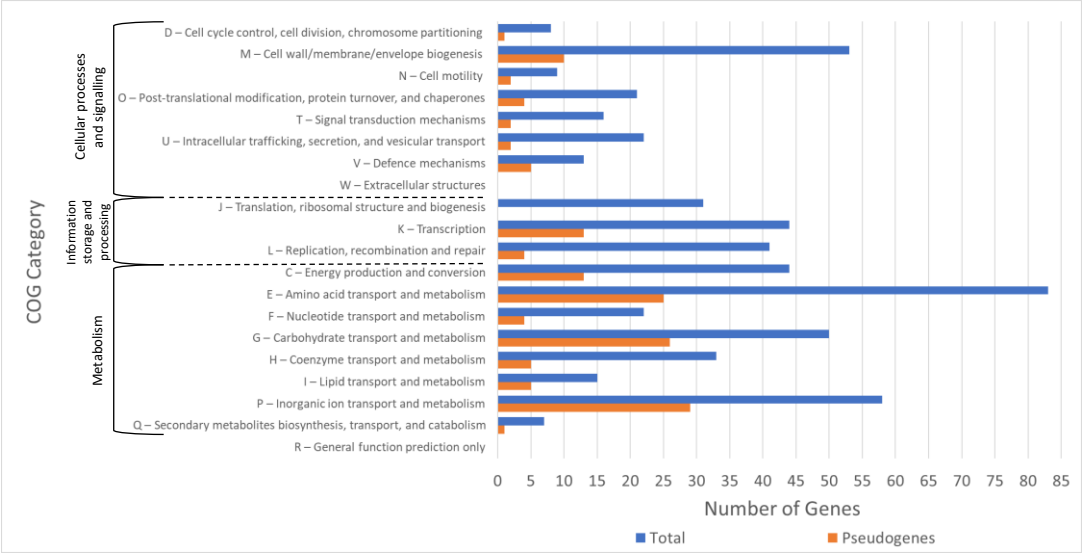


Figure 5.7. Graphical representation of the COG annotations for the essential gene candidates unique to the complex medium pool, using the minimal medium pool as input. Blue: all genes [including pseudogenes] with 0 insertions. **Orange**: pseudogenes with 0 insertions. COG S [Function Unknown] contained 150 genes, including 54 pseudogenes. 128 genes, including 50 pseudogenes, were unannotated.

As can be seen from Figure 5.7, 34.24% (278/812) of the genes were either unannotated or were assigned to COG S [Function Unknown]. 17.49% (142/812) of the unique essential gene candidates were categorised under Cellular Processes and Signalling, 14.29% (116/812) under Information Storage and Processing, and 38.42% (312/812) under Metabolism (Figure 5.7). 18.31% (26/142) of the unique essential gene candidates categorised under Cellular Processes and Signalling were pseudogenised, 14.66% (17/116) within Information Storage and Processing, and 34.62% (108/312) within Metabolism (Figure 5.7).

The 1 – 0.50 Kb unique essential gene candidate KO mapper inputs revealed essentiality within glycolysis/gluconeogenesis (sgl00010) [also including pyruvate metabolism (sgl00620)], oxidative phosphorylation (sgl00190), the citrate cycle (TCA cycle) (sgl00020) [also including glyoxylate and dicarboxylate metabolism (sgl00630), pantothenate and CoA biosynthesis (sgl00770), and arginine biosynthesis (sgl00220)], and the PPP (sgl00030). These are pathways that *S. glossinidius* has retained and the associated genes are revealed here as essential when subjected specifically to a growth medium that offers nutrient components complementary to the pathways [such as fumaric, α-ketoglutaric, malic, and succinic acid for the CAC, and glucose for glycolysis and the PPP]. This active utilisation of glucose is supported by the carbohydrate metabolism pathways identified [using the KEGG mapper] using the uniquely essential genes to *S. glossinidius* growth in the complex versus minimal medium; including glucose epimers [fructose and mannose (sgl00051) and galactose (sgl00052)] and poly-/disaccharide constituents [starch and sucrose (sgl00500)]. There was also essentiality for the metabolism pathways involving the vast range of amino acids that the complex medium provides versus the minimal medium, including cysteine and methionine (sgl00270), alanine, aspartate and glutamate (sgl00250), glycine, serine and threonine (sgl00260), histidine (sgl00340), arginine and proline (sgl00330), tyrosine (sgl00350), and vitamin B6 (sgl00750). Amino acid biosynthesis pathways were also included: hits within phenylalanine, tyrosine and tryptophan biosynthesis (sgl00400) involving the shikimate pathway and ultimately leading to tyrosine and phenylalanine metabolism; lysine biosynthesis (sgl00300) involving

parts of the pathway feeding into glycine, serine and threonine metabolism (sgl00260) and ultimately leading to its degradation (sgl00310); valine, leucine and isoleucine biosynthesis (sgl00290) stemming from glycine, serine and threonine (sgl00260) and pyruvate (sgl00620) metabolism, ultimately leading to their degradation (sgl00280). A variety of ABC transporters and two-component systems were included in the essentiality hits; of-note within amino acid ABC transporters [the ArtI component of arginine, GltL component of glutamate/aspartate, TcyA component of cystine, and the MetN component of D-methionine], the OppB, OppC, and OppD components of the oligopeptide ABC transporter [supported by Opp (oligopeptide ABC transporter periplasmic component(s)) hits across multiple quorum sensing systems (sgl02024)], the TbpA and ThiQ components of the thiamine ABC transporter [supported by thiamine metabolism (sgl00730)], two-component systems involved with glutamate [the GlnG (two-component nitrogen regulation protein), GlnA (glutamine synthetase), and AatP (glutamate/aspartate transport system ATP-binding protein) components], the RcsD component involved with capsular polysaccharide synthesis from glucose, and virulence-associated two-component systems [the EpsC (UDP-N-acetylglucosamine 2-epimerase) component involved in the autoregulation of virulence genes, the CheB (response regulator for chemotaxis) component involved in flagellar assembly, the Clp (cyclic AMP-regulatory protein) component, and the PhoP (two-component response regulator protein) component from the PhoQ-PhoP system involved in regulating virulence determinants]. The coverage of the CAMP resistance (sgl01503) pathway is extended from that seen in long-term establishment in the complex medium [5.3.2.1.]. The essential candidacy within oligopeptide transport systems supports the observations that oligopeptide levels within liquid medium reduce with increasing *S. glossinidius* growth levels, suggesting that *S. glossinidius* actively transports and uses available oligopeptides (Darby, A.C., personal correspondence).

Associated unique genes detailed in Appendix Table 2.

5.3.3. Essential Gene Candidates for *Sodalis glossinidius in vitro* Growth in a Minimal Medium versus a Complex Medium

The results from the adapted Bio-TraDIS pipeline [5.2.4.3.] for the minimal medium pool showed an average of 2,948 insertion sites; theoretically one transposon insertion for every 1,425 nucleotides. Average 5,042,657 [49.28%] of the reads were retained post-filtering, and average 4,809,339 [95.37%] of these retained reads mapped to the reference genome. The results from the further analysis [5.2.4.4.] showed that 72.54% (4215/5811) overall genes did not have any transposon insertions in sample 15, 75.22% (4371/5811) in sample 18, and 81.28% (4723/5811) in sample 21 (Figure 5.3); 3,638 of these essential gene candidates were shared between the samples (i.e. forming the [minimal medium] pool).

444 of the [3,638] shared essential gene candidates were unique to the minimal medium pool, when comparing the group to the input sample and the complex medium pool [samples 2, 4, 5, and 6] (Figure 5.3); these genes representing essentiality specifically for *S. glossinidius* growth in minimal media over complex (Figure 5.8).

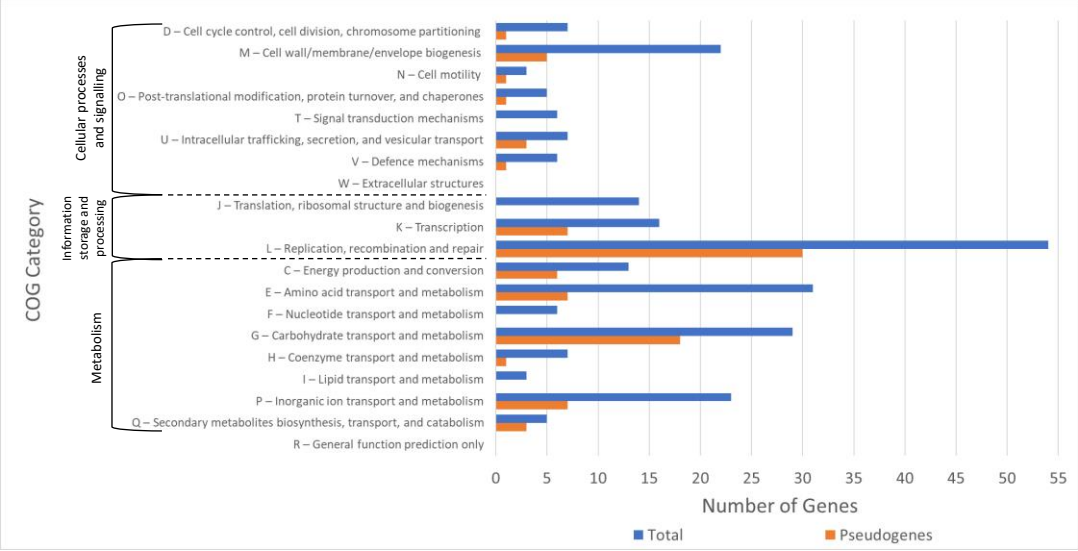


Figure 5.8. Graphical representation of the COG annotations for the essential gene candidates unique to the minimal medium pool, using sample 2 and the complex medium pool as input. Blue: all genes [including pseudogenes] with 0 insertions. **Orange**: pseudogenes with 0 insertions. COG S [Function Unknown] contained 114 genes, including 41 pseudogenes. 79 genes, including 26 pseudogenes, were unannotated.

As can be seen from Figure 5.8, 43.47% (193/444) of the genes were either unannotated or were assigned to COG S [Function Unknown]. 12.61% (56/444) of the unique essential gene candidates were categorised under Cellular Processes and Signalling, 18.92% (84/444) under Information Storage and Processing, and 26.35% (117/444) under Metabolism (Figure 5.8). 21.43% (12/56) of the unique essential gene candidates categorised under Cellular Processes and Signalling were pseudogenised, 44.05% (37/84) within Information Storage and Processing, and 35.90% (42/117) within Metabolism (Figure 5.8).

The 1 – 0.50 Kb KO mapper inputs from the unique minimal medium essential gene candidates revealed pathway crossovers with the unique inputs from within the complex medium, confirming the pathways' general essentiality within *S. glossinidius*; those involved with energy production (glycolysis/gluconeogenesis (sgl00010) [also including pyruvate metabolism (sgl00620)], the PPP (sgl00030), oxidative phosphorylation (sgl00190), and CAC representation [glyoxylate and dicarboxylate metabolism (sgl00630), pantothenate and CoA biosynthesis (sgl00770), and arginine biosynthesis (sgl00220)]), and virulence (CAMP resistance (sgl01503), flagellar assembly (sgl02040) [including flagellar transcriptional activator FlhC within the quorum sensing (sgl02024) and two-component (sgl02020) systems], and two-component systems (sgl02020) involving the OmpF porin, the PhoQ component of the PhoQ-PhoP system, and the EpsD (UDP-N-acetyl-D-mannosaminuronic acid dehydrogenase) component involved in the autoregulation of virulence genes). The minimal medium pool restricted mutant *S. glossinidius* [defined] carbon access to *N*-acetylglucosamine, which is a derivative of glucose formed from the glycolysis intermediate fructose-6-phoshpate. Unique [to the minimal medium pool] essentiality candidacy in the PTS (sgl02060) component involving a glucose epimer (PTS system mannose-specific IID component ManZ), in conjunction with the *N*-acetylglucosamine-specific PTS [incorrectly annotated as *ptsG_1*] containing transposon insertions, confirms the hypothesis [based on the data presented in Chapter 2] that the remaining intact *S. glossinidius* transporters are lenient; *N*-acetylglucosamine is most likely transporting via the mannose PTS system, in this selection (Toh *et al*., 2006; Goodhead *et al.*, 2018). A stronger

presence of essentiality candidacy was observed within sulfur metabolism (sgl00920) within the minimal medium pool, versus the complex medium pool, which is an expected result considering the minimal medium is entirely salts-based aside from 20% glucose [replaced with yeast extract and *N*-acetylglucosamine, in this case] (Miller, 1972).

Associated unique genes detailed in Appendix Table 3.

5.3.4. Overall Essential Gene Candidates for *Sodalis glossinidius*

The results from the adapted Bio-TraDIS pipeline [5.2.4.3.] for the combined sample [all 21 samples combined: 5.2.4.3.] showed 12,804 insertion sites; theoretically one transposon insertion for every 328 nucleotides. 37,738,993 [15.92%] of the reads were retained post-filtering, and 34,220,483 [90.68%] of these retained reads mapped to the reference genome. The results from the further analysis [5.2.4.4.] showed that 46.34% (2693/5811) overall genes did not have any transposon insertions, with only 6.16% (166/2693) of these genes over 1 Kb in length. Of-note within these 2693 genes, 43.74% (1178/2693) are pseudogenes; 17 of these are $\geq$ 1 Kb in length, all of which have RNASeq evidence of transcription (Goodhead *et al.*, 2018).

Of the 137 pSG1 genes, 23.36% (32/137) did not have any transposons insertions; 59.38% (19/32) of these were pseudogenes and 28.13% (9/32) encoded for hypothetical proteins. The remaining pSG1 genes without transposon insertions (4/32) included virulence- and free-living associated genes, including bacterial conjugation TrbI-like protein associated with type IV secretion systems, haemolysin expression-modulating protein Hha shown to regulate haemolysin levels in virulent *E. coli*, IS66 Orf2 like protein homologous to insertion sequence elements found in enteropathogenic *E. coli* B171, and a [putative] siderophore-binding lipoprotein YfiY precursor (Nieto *et al.*, 1991; Han *et al*., 2001; Lawley *et al*., 2003). Of the 56 pSG2 genes, 32.14% (18/56) did not have any transposons insertions; 11.11% (2/18) of these were pseudogenes and 44.44% (8/18) encoded for hypothetical proteins. The remaining pSG2 genes without transposon insertions (8/18) include those involved

in genetic maintenance, such as single-stranded- and DNA-binding proteins, plasmid copy number regulatory proteins, and mobilisation proteins involved in plasmid conjugation (Castagnoli *et al*., 1989; Zhang and Meyer, 1997; Marceau, 2012). All 12 pSG3 genes, 91.67% (11/12) of which encode for hypothetical proteins, had transposon insertions [ranging from 1 – 85]. When comparing the number of genes with transposon insertions in the chromosome versus the plasmids [pSG1, pSG2 and pSG3], it could be suggested that the plasmids are less essential for *S. glossinidius*; 52.85% (2963/5606) of the chromosomal genes had insertions, whereas 75.61% (155/205) of the combined plasmid genes had insertions.

5.3.4.1. Clusters of Orthologous Groups (COG) Analysis

The results from the COG assignments are displayed in Figure 5.9.

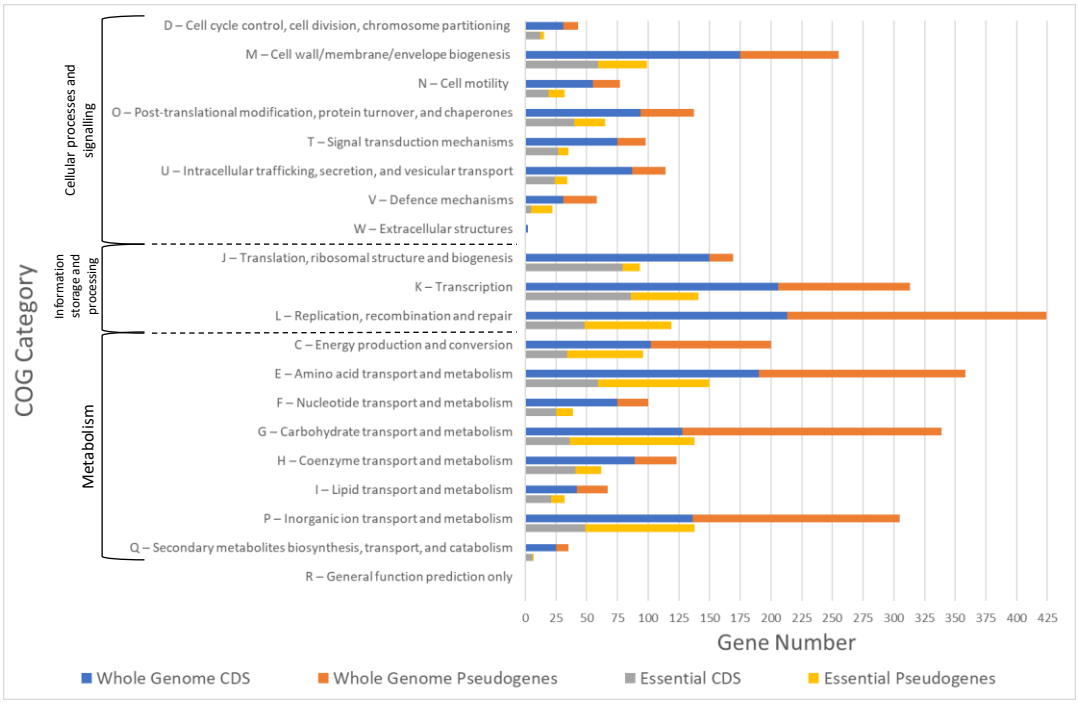


Figure 5.9. Graphical representation of the COG annotations for the whole genome compared to the combined sample essential gene candidates. Blue: CDSs from the whole genome. **Orange**: pseudogenes from the whole genome. **Grey:** all CDSs with 0 insertions in the combined sample. **Yellow:** all pseudogenes with 0 insertions in the combined sample. For the whole genome, COG S [Function Unknown] contained 788 CDSs and 551 pseudogenes. For the combined sample essential gene candidates, COG S contained 326 CDSs and 278 pseudogenes. 723 CDSs and 428 pseudogenes were unannotated within the whole genome, and 531 CDSs and 263 pseudogenes within the combined sample essential gene candidates.

As can be seen from Figure 5.9, 51.91% (1398/2693) of the genes [CDSs + pseudogenes] were either unannotated or were assigned to COG S [Function Unknown]. 6.91% (186/2693) of the essential CDS candidates were categorised under Cellular Processes and Signalling, 7.91% (213/2693) under Information Storage and Processing, and 10.06% (271/2693) under Metabolism (Figure 5.9). 38.41% (116/302) of the unique essential gene candidates categorised under Cellular Processes and Signalling were pseudogenised, 39.66% (140/353) under Information Storage and Processing, and 59.06% (391/662) under Metabolism (Figure 5.9).

Figure 5.9 shows that the proportions of CDSs to pseudogenes within each COG subcategory do not change substantially between unselected [whole genome] versus selected [combined sample essential gene candidates] environments. What does change however are the proportions between COG subcategories; COG J [Translation, Ribosomal Structure and Biogenesis] shows higher numbers of CDSs for both the whole genome and the combined sample essential gene candidates selections, whereas COG L [Replication, Recombination and Repair] shows high numbers of pseudogenes for both selections. It could be theorised that the genes within COG subcategories containing higher proportions of CDSs are more essential for *S. glossinidius*, as they are not undergoing removal in the form of pseudogenisation associated with establishing as host associated (reviewed in McCutcheon and Moran, 2012). Of particular interest is the high pseudogene numbers within the Metabolism category of COGs, which would suggest that the free-living associated range of metabolism pathways reported within the literature are being phased-out as *S. glossinidius* establishes as full symbiont (Toh *et al*., 2006; Belda *et al*., 2010).

5.3.4.1.1. Metabolism

5.3.4.1.1.1. Amino Acid Transport and Metabolism (COG E)

Within the Metabolism category, the Amino Acid Transport and Metabolism (COG E) subcategory contained 22.66% (150/662) of the essential gene candidates; 12.67% (19/150) of these with gene lengths $\geq$ 1 Kb and 60.67% (91/150) being pseudogenised. The COG E $\geq$ 1 Kb essential gene candidate KO mapper inputs revealed that the amino acid biosynthesis pathway (sgl01230) contained one of the highest number of hits (Figure 5.10).

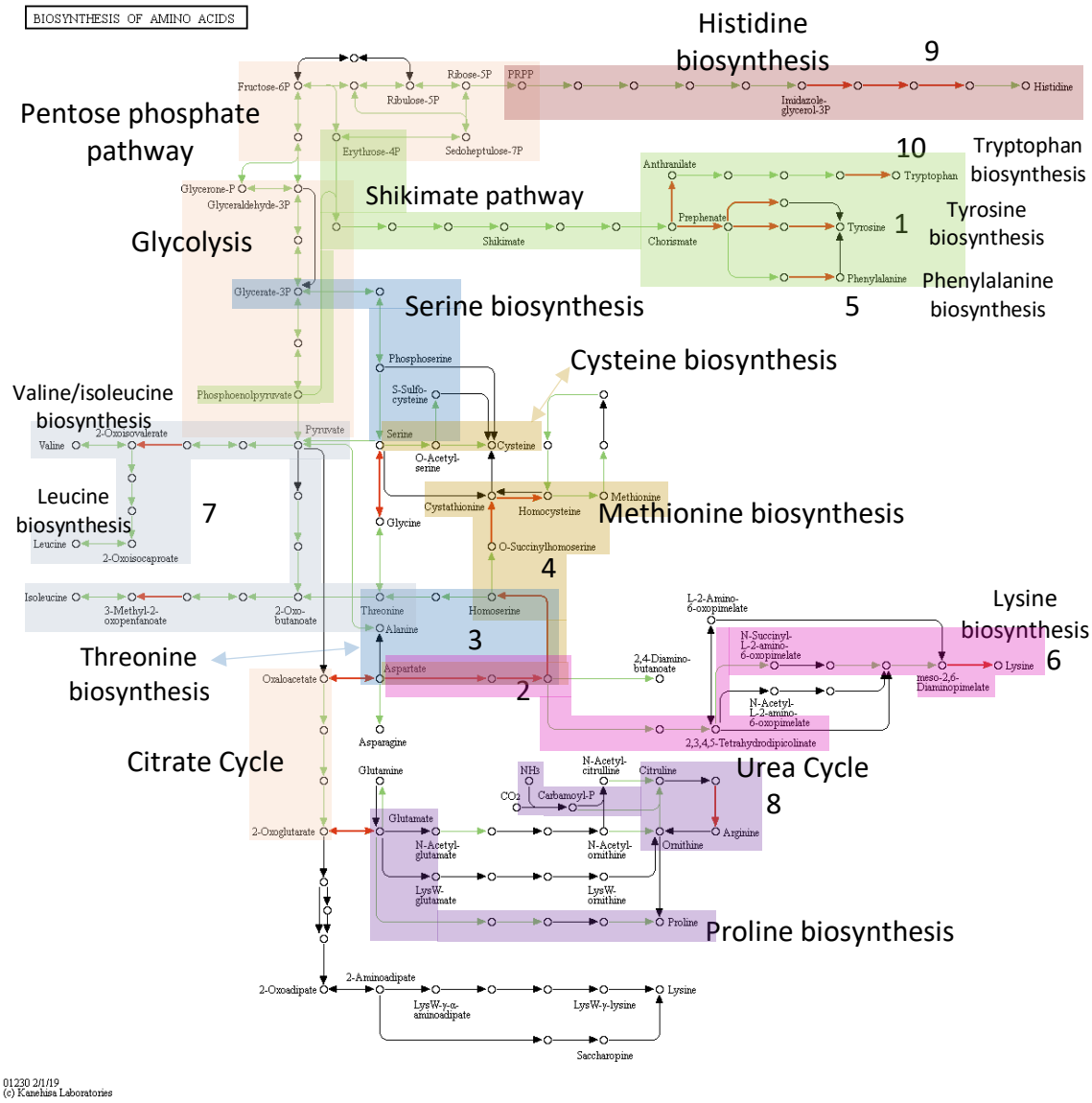


Figure 5.10. Kyoto Encyclopaedia of Genes and Genomes (KEGG) amino acid biosynthesis pathway of *Sodalis glossinidius*, using the KEGG Orthology (KOs) from COG E ≥ 1 Kb essential gene candidates as input. Green lines: intact pathways. **Black lines**: pathways/genes that are not intact. **Red lines:** pathways/genes that have been input as KOs. The colours represent pathway categories, which have been appropriately labelled. 16 KOs hit the KEGG amino acid biosynthesis pathway (sgl01230). Kanehisa and Goto (2000); Kanehisa *et al*. (2016), (2017).

As can be seen from the KEGG predicted amino acid biosynthesis pathway (Figure 5.10), the COG E ≥ 1 Kb essential gene candidates are involved in a variety of metabolic pathways.

The tyrosine biosynthesis pathway, as part of aromatic amino acid metabolism, using chorismate (M00025), as highlighted by Figure 5.10 "1" and Figure 5.11, is fully covered by *tyrA* (T-protein) and *tyrB* (aromatic-amino-acid aminotransferase). Tyrosine is an amino acid important for the production of proteins and its phosphorylation [in bacteria] has been associated with the metabolism of DNA, virulence, the formation of exopolysaccharides, and stress response (Grangeasse *et al.*, 2012). The fact that this pathway has been fully covered [by *tyrA* and *tyrB*] further confirms both the pathway's and the genes' potential essentiality; the essential gene candidates are not random and have a relationship to one another.

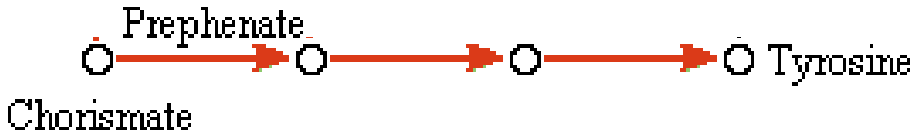


Figure 5.11. Kyoto Encyclopaedia of Genes and Genomes (KEGG) tyrosine biosynthesis pathway of *Sodalis glossinidius*, using the KEGG Orthology (KOs) from *tyrA* and *tyrB* as input. Green lines: intact pathways. **Black lines**: pathways/genes that are not intact. **Red lines:** pathways/genes that have been input as KOs. Kanehisa and Goto (2000); Kanehisa *et al*. (2016), (2017).

The ectoine biosynthesis pathway, as part of amino acid metabolism, using aspartate (M00033), as highlighted by Figure 5.10 "2" and Figure 5.12, is partially covered by *usg* (USG-1 protein), but not to completion; 2,4-diaminobutanoate is an intermediate compound in the production of ectoine from aspartate, suggesting a premature end to the pathway and an incomplete KEGG entry. Ectoine is a compound that is used by prokaryotes for protection against osmotic stress (Widderich *et al*., 2014).

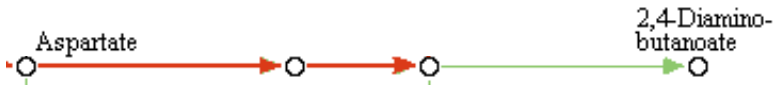


Figure 5.12. Kyoto Encyclopaedia of Genes and Genomes (KEGG) ectoine biosynthesis pathway of *Sodalis glossinidius*, using the KEGG Orthology (KOs) from *usg* as input. Green lines: intact pathways. **Black lines**: pathways/genes that are not intact. **Red lines:** pathways/genes that have been input as KOs. Kanehisa and Goto (2000); Kanehisa *et al*. (2016), (2017).

The remaining intact [according to KEGG] pathways were all partially hit by the input KOs from the COG E ≥ 1 Kb essential gene candidates and are summarised in Table 5.6.

Table 5.6. Kyoto Encyclopaedia of Genes and Genomes (KEGG) intact pathways partially hit by the COG E ≥ 1 Kb essential gene candidates KO inputs. Kanehisa and Goto (2000); Kanehisa *et al*. (2016), (2017).

Pathway	Category	KO Input Genes	Figure 5.7 Label
Threonine biosynthesis (M00018)	Amino acid metabolism	*metL_1, lysA, usg*	3
Methionine biosynthesis (M00017)	Cysteine and methionine metabolism	*metL_1, lysA, usg, metB*	4
Phenylalanine biosynthesis (M00024)	Aromatic amino acid metabolism	*tyrA, tyrB*	5
Histidine biosynthesis (M00026)	Histidine metabolism	*hisB, hisC*	9
Tryptophan biosynthesis (M00023)	Aromatic amino acid metabolism	*trpE, trpB*	10

The pathways not fully intact [according to KEGG] included the valine/isoleucine (M00019) [Figure 5.10 “7”], lysine [succinyl-DAP] (M00016) [Figure 5.10 “6”], and the arginine biosynthesis pathway. The latter is part of

arginine and proline metabolism using ornithine (M00844), as highlighted by Figure 5.10 “8” and Figure 5.13, and is partially covered by *argH* (argininosuccinate lyase).

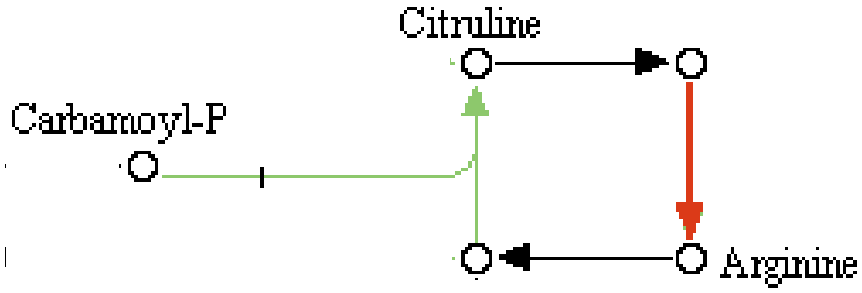


Figure 5.13. Kyoto Encyclopaedia of Genes and Genomes (KEGG) arginine biosynthesis pathway of *Sodalis glossinidius*, using the KEGG Orthology (KOs) from *argH* as input. Green lines: intact pathways. **Black lines**: pathways/genes that are not intact. **Red lines:** pathways/genes that have been input as KOs.Kanehisa and Goto (2000); Kanehisa *et al*. (2016), (2017).

As *S. glossinidius* has been previously shown to be [using the re-annotated genome] unable to biosynthesise arginine, the essential candidacy of *argH* was unexpected (Belda *et al*., 2010; Belda *et al*., 2012). However, the *S. glossinidius* strain SgGMMB4 genome shows *argH* as 1.4 Kb in length and possessing RNASeq evidence of transcription, supporting the more recent conclusions that *S. glossinidius* might not be fully arginine auxotrophic (Goodhead *et al.*, 2018; Hall *et al*., 2019). However, where only a part of a pathway is hit/intact, it is less likely that the pathway section is genuinely essential when the rest of it is not hit/intact; *argB* [0.78 Kb] (acetylglutamate kinase) and *argE* [1.15 Kb] (acetylornithine deacetylase) had 2 and 3 transposon insertions, respectively. Contrary to this, the disparity between the KEGG pathways and the *S. glossinidius* strain SgGMMB4 genome could be masking part of the pathway, as *argC* (N-acetyl-gamma-glutamyl-phosphate reductase) was also included within the COG E $\geq$ 1 Kb essential gene candidate KO input and has RNASeq evidence of transcription, but was missing from the KEGG pathway prediction [Figure 5.13] (Goodhead *et al.*, 2018).

The KEGG pathway predictions following the COG E 0.99 – 0.50 Kb essential gene candidate KO mapper inputs looked the same, except for the changes outlined in Table 5.7.

Table 5.7. Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway prediction changes when submitting COG E ≥ 1 Kb essential gene candidates versus 0.99 – 0.50 Kb KOs. RNASeq evidence of transcription obtained from Goodhead *et al.* (2018). Kanehisa and Goto (2000); Kanehisa *et al*. (2016), (2017).

Existing Pathway	Existing KO Input Gene (≥ 1 Kb)	Additional KO Input Gene, Function and Size (Kb)	RNASeq Evidence?
Threonine biosynthesis (M00018)	*metL_1, lysA, usg*	*thrB_1* (homoserine kinase), 0.93	Yes
Histidine biosynthesis (M00026)	*hisB, hisC*	1. *hisA* (1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino] imidazole-4-carboxamide isomerase), 0.74 2. *hisH* (imidazole glycerol phosphate synthase subunit HisH), 0.59	1. Yes 2. Yes
Lysine [succinyl-DAP] biosynthesis (M00016)	*metL_1, lysA, usg*	*dapA_1* (4-hydroxy-tetrahydrodipicolinate synthase), 0.54	Yes
New Pathway	**Existing KO Input Gene (≥ 1 Kb)**	**New KO Input Gene, Function and Size (Kb)**	**RNASeq Evidence?**
Serine biosynthesis (M00020)	-	GMMB4_01435 (glycerate dehydrogenase), 0.62	Yes
Shikimate pathway (M00022)	-	1. *aroL* (shikimate kinase 2), 0.53 2. *aroK_2* (shikimate kinase 1), 0.52	1. Yes 2. Yes

The addition of *thrB_1* (0.93 Kb) [Table 5.7] to the input KOs covers the threonine biosynthesis pathway to near completion, further suggesting that it is essential for

S. glossinidius (Figure 5.14 A). This essentiality conclusion potentially extends to the lysine [succinyl-DAP] biosynthesis pathway, as with the inclusion of *dapA_1* (0.54 Kb) [Table 5.7], the amount of candidate essentiality (and intact sections) dominates over the amount of the pathway not intact [according to KEGG] (Figure 5.14 B).

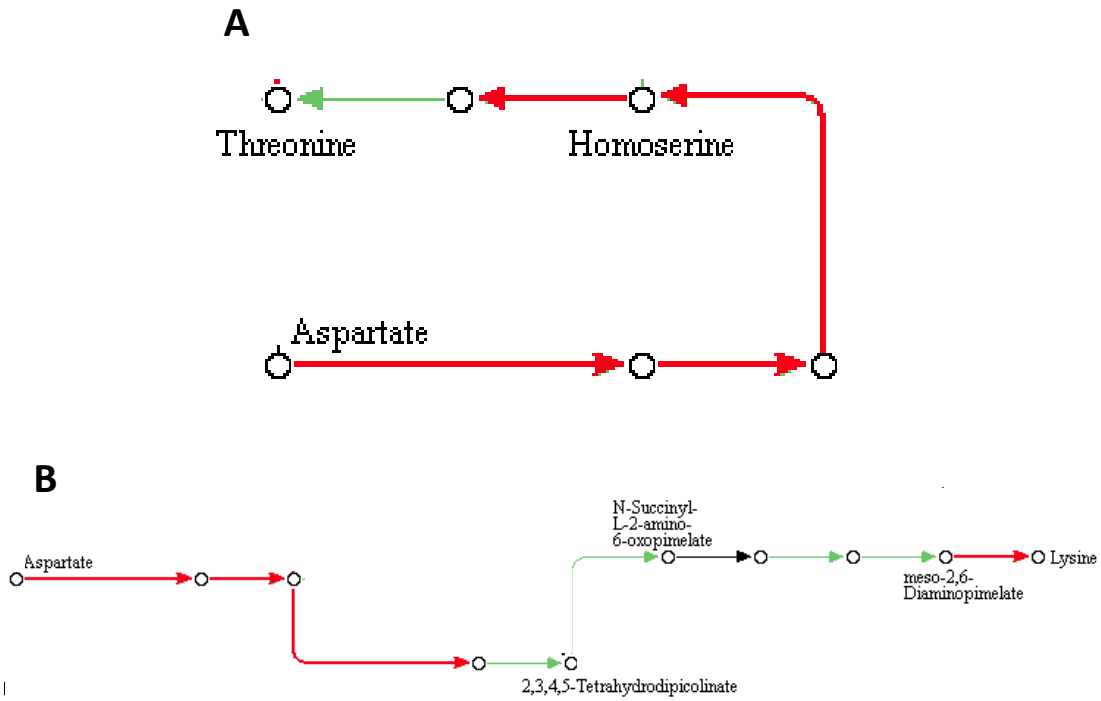


Figure 5.14. Kyoto Encyclopaedia of Genes and Genomes (KEGG) amino acid biosynthesis pathways of *Sodalis glossinidius*, using KEGG Orthology (KOs) as input. Green lines: intact pathways. **Black lines**: pathways/genes that are not intact. **Red lines:** pathways/genes that have been input as KOs. **A:** Threonine biosynthesis pathway (M00018) using *metL_1, lysA, usg* and *thrB_1* KOs **B:** Lysine [succinyl-DAP] biosynthesis pathway (M00016) using *metL_1, lysA, usg* and *dapA_1* KOs. Kanehisa and Goto (2000); Kanehisa *et al*. (2016), (2017).

The COG E 0.99 – 0.50 Kb essential gene candidate KO submission also saw a hit-increase in the number of ABC transporters, from the 1 [potG, putrescine transport] within the ≥ 1 Kb submission. Of-note, the arginine transporter has two [out of 5] components as essential candidates, including artM and artQ, both of which have RNASeq evidence of transcription (Goodhead *et al*., 2018). This level of essential candidacy for a transport system dedicated to an amino acid for which *S.*

glossinidius is [controversially] auxotrophic is of scientific interest for further functionality studies.

5.3.4.1.1.2. Carbohydrate Transport and Metabolism (COG G)

Within the Metabolism category, the Carbohydrate Transport and Metabolism (COG G) subcategory contained 20.85% (138/662) of the essential gene candidates; 3.62% (5/138) of these with gene lengths ≥ 1 Kb and 73.91% (102/138) being pseudogenised. COG G has the highest number of essential pseudogene candidates within the Metabolism category. Of-note hits within the COG G ≥ 1 Kb essential gene candidate KO mapper inputs were within the PPP (sgl00030), glycolysis/gluconeogenesis (sgl00010), and fructose and mannose metabolism (sgl00051), supporting both the literature conclusions that *S. glossinidius* has retained these pathways and also the notion of them being actively used (Toh *et al*., 2006; Belda *et al*., 2010; Belda *et al*., 2012). Surprisingly, one of the essential gene candidate KO inputs hit within thiamine metabolism (sgl00730); it has been previously reported, and subsequently reinforced, that *S. glossinidius* is unable to biosynthesise thiamine using the 'conventional' pathway (Toh *et al*., 2006; Belda *et al*., 2010; Hall *et al.*, 2019). Belda *et al*. (2010) proposed that *S. glossinidius* could instead be utilising thiamine through a salvage pathway-ABC transporter work-around [detailed in 1.2.2.2.]. This inability to biosynthesise thiamine via the 'conventional' pathway was reported due to the apparent lack of *thiI* [sulfur-carrying protein], and a pseudogenised *thiF* [thiamin (thiazole moiety) biosynthesis protein], only the latter of which is true for *S. glossinidius* strain SgGMMB4 as the present *thiI* gene is 1.45 Kb in length and has RNASeq evidence of transcription (Belda *et al*., 2010; Goodhead *et al*., 2018). It is the case that all the thiamine-related *S. glossinidius* strain SgGMMB4 genes [excluding *thiH_1* and *thiH_2*], including the pseudogenised *thiF,* have RNASeq evidence of transcription (Table 5.8) (Goodhead *et al*., 2018).

Table 5.8. Thiamine-related genes in *Sodalis glossinidius* strain SgGMMB4. All genes have RNASeq evidence of transcription, except for *thiH_1* and *thiH_2* (Goodhead *et al*., 2018). Number of insertions identified using the *S. glossinidius*-adapted Bio-TraDIS pipeline [5.2.4.3.] (Barquist *et al*., 2016). Of-note, *thiB and thiQ* are not included in the thiamine metabolism (map00703) Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway map. Kanehisa and Goto (2000); Kanehisa *et al*. (2016), (2017).

Gene	Function	Length (Kb)	Number of Insertions	COG
dxs	1-deoxy-D-xylulose-5-phosphate synthase	1.86	2	H
thiI	tRNA sulfurtransferase	1.45	1	H
iscS	Cysteine desulfurase	1.20	0	E
rsgA	Putative ribosome biogenesis GTPase RsgA	1.04	0	G
thiL	Thiamine-monophosphate kinase	0.99	1	H
thiB	Thiamine-binding periplasmic protein precursor	0.98	1	H
thiK	Thiamine kinase	0.86	0	H
thiH_2	pseudogene	0.84	1	H
thiQ	Thiamine import ATP-binding protein ThiQ	0.70	1	H
adk	Adenylate kinase	0.65	0	F
thiM	Hydroxyethylthiazole kinase	0.58	0	H
thiF	pseudogene	0.38	1	H
thiH_1	pseudogene	0.29	0	H

As can be seen from Table 5.8, 84.62% (11/13) of the thiamine-related genes are within the length ranges explored within this analysis [1 – 0.50 Kb]. There are 36 [including the 10 thiamine biosynthesis (M00127) entries] KEGG entries involved with thiamine metabolism (map00703) and *S. glossinidius* strain SgGMMB4 covers 25% (9/36) of these; this low amount of pathway gene content would suggest that the pathway is non-functional and thus supports the conclusion that *S. glossinidius* in unable to biosynthesis thiamine using the 'conventional' pathway (Belda *et al*., 2010; Hall *et al*., 2019). The number of insertions in the thiamine-related genes within the combined sample range from 0 – 2, but the numbers vary between the other pools (Figure 5.15).

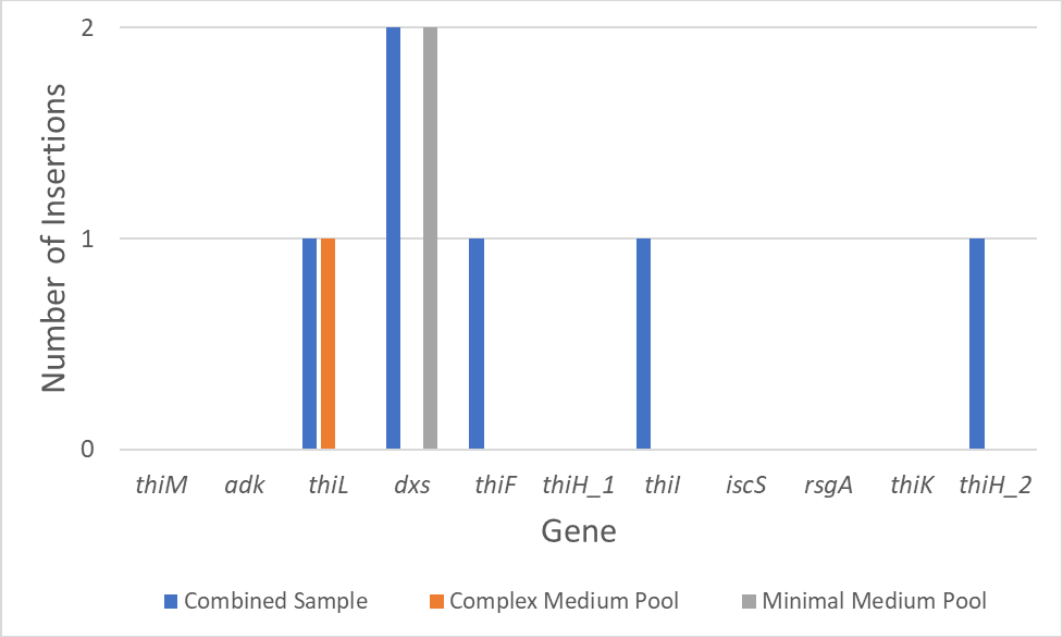


Figure 5.15. Number of transposon insertions in thiamine-related genes from different TraDIS pools, analysed by the *Sodalis glossinidius*-adapted Bio-TraDIS pipeline. Blue: number of insertions in the combined sample. Orange: number of insertions in the complex medium pool [Schneider's Insect Medium + 10% foetal bovine serum]. Grey: number of insertions in the minimal medium pool [M9 minimal salts medium + yeast extract + *N*-acetylglucosamine].

As can be seen from Figure 5.15, 45.45% (5/11) of thiamine-related genes have transposon insertions from the combined sample; as detailed in 5.2.4.3., this sample is a combination of all 21 samples. However, three of these five genes do not contain insertions in the complex medium pool or minimal medium pool, suggesting that *thiF* and *thiI* hold essentiality candidacy for *S. glossinidius in vitro* growth specifically in both these conditions. Figure 5.15 would also suggest that *thiL* is an essential gene candidate specifically for *S. glossinidius in vitro* growth in minimal medium with restricted carbon, nitrogen and vitamin sources, and that *dxs* is an essential gene candidate specifically for *S. glossinidius in vitro* growth in complex medium abundant in a variety of nutrients. As it is likely that the associated pathway for the genes discussed here [Table 5.8 and Figure 5.15] is non-functional, it could be theorised that the essential gene candidates are being used for a different function in *S. glossinidius*; either the thiamine salvage pathway-ABC

transporter work-around presented by Belda *et al*. (2010), or potentially novel interactions yet to be defined.

The COG G 0.99 – 0.50 Kb essential gene candidate KO mapper inputs increased the coverage from the ≥ 1 Kb inputs, and additionally covered other carbohydrate metabolism pathways [galactose (sgl00052), starch and sucrose (sgl00500), and inositol phosphate (sgl00562) metabolism]. This supports the previously discussed conclusion that *S. glossinidius* is able to utilise a range of carbon sources (Toh *et al*., 2006; Belda *et al*., 2010; Belda *et al*., 2012; 2.3.2. and 2.3.3).

5.3.4.1.1.3. Inorganic Ion Transport and Metabolism (COG P)

Within the Metabolism category, the Inorganic Ion Transport and Metabolism (COG P) subcategory contained 20.85% (138/662) of the essential gene candidates; 5.07% (7/138) of these with gene lengths ≥ 1 Kb and 64.49% (89/138) being pseudogenised. Of-note hits within the COG P ≥ 1 Kb essential gene candidate KO mapper inputs were within selenocompound (sgl00450), sulfur (sgl00920) and purine (sgl00230) metabolism pathways. The COG P 0.99 – 0.50 Kb essential gene candidate KO mapper inputs increased the coverage from the ≥ 1 Kb inputs, but additionally covered ABC transporters (sgl02010) and two-component systems (sgl02020): the intact [according to KEGG] ABC transporter hits included the monosaccharide transporters [methyl-galactoside, and xylitol], di- and oligopeptide transporters, phosphate and amino acid transporters [phosphate, and D-methionine], and the iron(II)/manganese transporter. The essential candidacy of a range of monosaccharide transporters, in conjunction with them being fully intact, again supports the conclusion(s) that *S. glossinidius* is able to utilise a range of carbon sources (Toh *et al*., 2006; Belda *et al*., 2010; Belda *et al*., 2012; 2.3.2. and 2.3.3).

5.3.4.1.1.4. Energy Production and Conversion (COG C)

Within the Metabolism category, the Energy Production and Conversion (COG C) subcategory contained 14.50% (96/662) of the essential gene candidates; 14.58% (14/96) of these with gene lengths ≥ 1 Kb and 48.96% (47/96) being pseudogenised. The COG C ≥ 1 Kb essential gene candidate KO mapper inputs revealed that the citrate cycle (TCA cycle) (sgl00020) contained one of the highest number of hits (Figure 5.16).

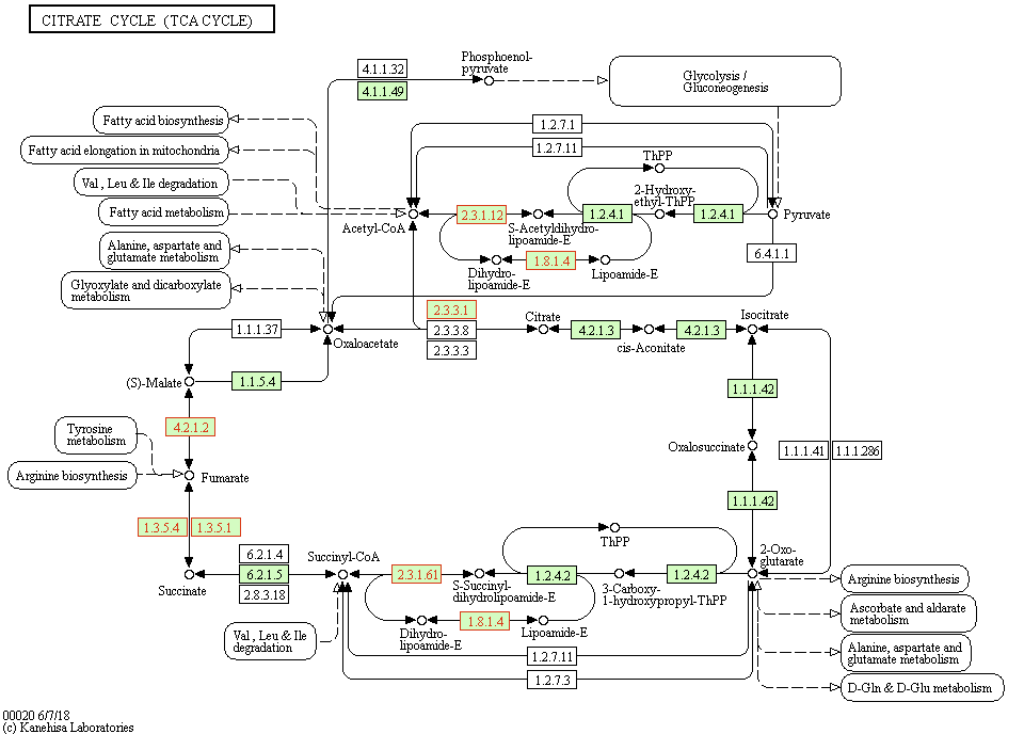


Figure 5.16. Kyoto Encyclopaedia of Genes and Genomes (KEGG) citrate cycle (TCA cycle) of *Sodalis glossinidius*, using the KEGG Orthology (KOs) from COG C ≥ 1 Kb essential gene candidates as input. Green: present genes **White**: absent genes **Red:** genes that have been input as KOs. 7 KOs hit the KEGG citrate cycle (TCA cycle). Kanehisa and Goto (2000); Kanehisa *et al*. (2016), (2017).

The level of essentiality candidacy within the citrate cycle (Figure 5.16) [also including glyoxylate and dicarboxylate metabolism (sgl00630)], in conjunction with the KO input hits within glycolysis/gluconeogenesis (sgl00010) [also including 2-oxocarboxylic acid (sgl01210) and pyruvate (sgl00620) metabolism] and the PPP (sgl00030) both here [COG C] and within COG G, supports the literature conclusions

that *S. glossinidius* has retained a range of energy pathways; a trait more associated with free-living organisms (Toh *et al*., 2006; Belda *et al*., 2010). This trait conclusion is further supported by the essentiality hits within oxidative phosphorylation (sgl00190), in addition to a range of metabolism pathways, including fructose and mannose (sgl00051), propanoate (sgl00640), butanoate (sgl00650), glycerolipid (sgl00561), glycerophospholipid (sgl00564), glycine, serine and threonine (sgl00260), tryptophan (sgl00380), and nicotinate and nicotinamide (sgl00760). The essentiality data here confirms that *S. glossinidius* is actively utilising these retained pathways, suggesting that the retention is not a redundant artefact of genome degradation associated with establishing with a host.

The COG C 0.99 – 0.50 Kb essential gene candidate KO mapper inputs increased the coverage from the ≥ 1 Kb inputs, but additionally increased the metabolism pathway range: C5-branched dibasic acid (sgl00660), methane (sgl00680), and taurine and hypotaurine (sgl00430).

5.3.4.1.2. Information Storage and Processing

Within the Information Storage and Processing category, the Transcription (COG K) subcategory contained the highest number (141/353) of essential gene candidates; 1.42% (2/141) with gene lengths ≥ 1 Kb and 39.01% (55/141) being pseudogenised. Of-note COG K 1 – 0.50 Kb essential gene candidate KO mapper input hits included expected pathways associated with the COG subcategory, including RNA degradation (sgl03018) and RNA polymerase (sgl03020) [specifically the DNA-directed RNA polymerase α subunit].

The Replication, Recombination and Repair (COG L) subcategory also contained a relatively high number (119/353) of essential gene candidates; 10.08% (12/119) with gene lengths ≥ 1 Kb and 59.66% (71/119) being pseudogenised. The COG L 1 ≥ Kb essential gene candidate KO mapper input hits included expected pathways associated with the COG subcategory, including homologous recombination (sgl03440), RNA degradation (sgl03018), DNA replication (sgl03030), mismatch repair (sgl03430), and a two-component system (sgl02020) [specifically

involving the chromosomal replication initiator protein (*dnaA*)]. The COG L 0.99 – 0.50 Kb essential gene candidate KO mapper inputs increased the coverage from the ≥ 1 Kb inputs, but additionally covered base excision repair (sgl03410).

The Translation, Ribosomal Structure and Biogenesis (COG J) subcategory also contained essential gene candidates (93/353); 12.90% (12/93) with gene lengths ≥ 1 Kb and 15.05% (14/93) being pseudogenised. Of-note COG J 1 – 0.50 Kb essential gene candidate KO mapper input hits included expected pathways associated with the COG subcategory, including aminoacyl-tRNA biosynthesis (sgl00970), sulfur relay system (sgl04122), RNA degradation (sgl03018), and ribosome [structure] (sgl03010).

5.3.4.1.3. Cellular Processes and Signalling

Within the Cellular Processes and Signalling category, the Cell Wall/Membrane/Envelope Biogenesis (COG M) subcategory contained the highest number (99/302) of essential gene candidates; 25.25% (25/99) with gene lengths ≥ 1 Kb and 40.40% (40/99) being pseudogenised. Of-note hits from the COG M ≥ 1 - 0.50 Kb essential gene candidate KO mapper inputs included expected pathways associated with the COG subcategory, including the lipopolysaccharide (sgl00540) and peptidoglycan (sgl00550) biosynthesis, and the vancomycin (sgl01502) and β-lactam (sgl01501) resistance [protection against cell well synthesis interference], pathways.

5.3.4.2. Reciprocal Basic Local Alignment Search Tool (BLAST®) Analysis

The reciprocal BLAST® (National Library of Medicine) comparison results showed that there were 91 essential gene candidates shared between *S. glossinidius* and *E. coli* K-12 (Goodall *et al*., 2018), 46 with *S. enterica* serovar Typhi (Langridge *et al*., 2009), and 0 with *M. genitalium* (Glass *et al*., 2006). The COG categories are summarised in Figure 5.17 A and B.

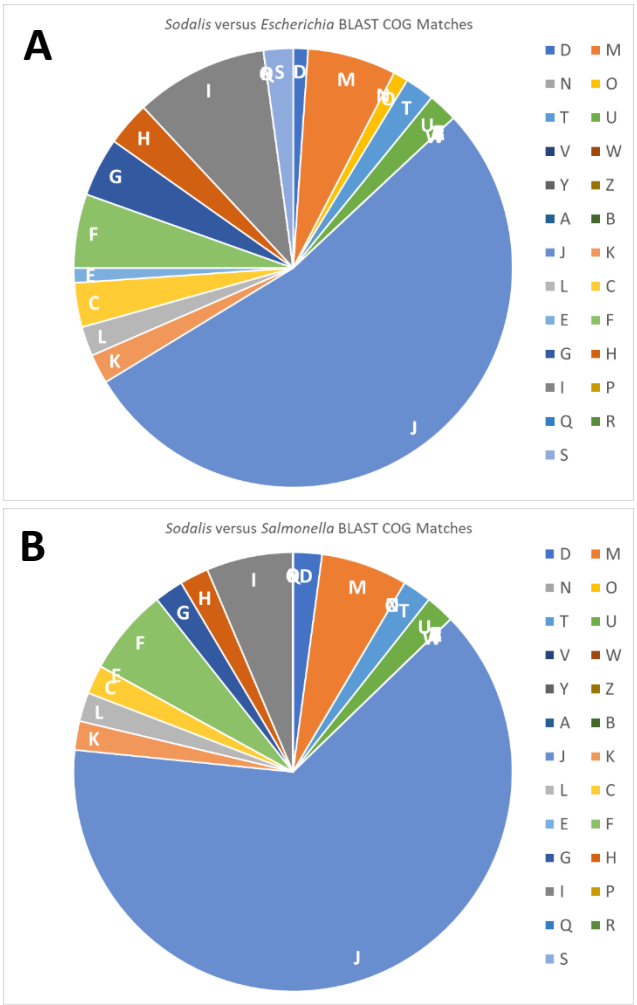


Figure 5.17. Reciprocal Basic Local Alignment Search Tool (BLAST®) essential gene matches between *Sodalis glossinidius* versus (A) *Escherichia coli* K-12 and (B) *Salmonella enterica* serovar Typhi. The data is visualised based on Clusters of Orthologous Group (COGs) annotations. Goodall *et al*. (2018); Langridge *et al*. (2009).

As can be seen from Figure 5.17 A and B, COG J [Translation, Ribosomal Structure and Biogenesis] holds the highest proportion of essential gene candidates that match between *S. glossinidius* and *E. coli* K-12 or *S. enterica* serovar Typhi; 30S and 50S ribosomal proteins constituting the highest percentage. A proportion of the COG category-expected gene sets detailed in 5.3.4.1.2. and 5.3.4.1.3. matched between the transposon libraries; within RNA polymerase (sgl03020), two-component system (sgl02020), aminoacyl-tRNA biosynthesis (sgl00970), sulfur relay system (sgl04122), lipopolysaccharide biosynthesis (sgl00540), peptidoglycan

biosynthesis (sgl00550), and vancomycin resistance (sgl01502). Particularly of interest were the library matches within thiamine metabolism (sgl00730), oxidative phosphorylation (sgl00190), the citrate cycle (TCA cycle) (sgl00020) and glycolysis/gluconeogenesis (sgl00010). The matches in higher-coverage transposon libraries, in combination with all matched *Sodalis* genes having RNASeq evidence of transcription, increases the confidence of the genes' essentiality verdict in *S. glossinidius* (Langridge *et al.*, 2009; Goodall *et al.*, 2018; Goodhead *et al*., 2018). Interestingly, this did not apply to the presumably main [intact] gene involved in glucose transport, as *ptsG_6* [1.28 Kb] had 10 transposon insertions, suggesting that it is not essential and thus glucose is not essential for *S. glossinidius*. However, the transposon library matches involving *ptsG_2* [0.53 Kb] (pseudogene), *ptsG_3* [0.86 Kb] (pseudogene) and *crr* [0.51 Kb] (glucose-specific phosphotransferase system (PTS) enzyme IIA component)], in conjunction with their RNASeq evidence of transcription, confirms these genes' essentiality verdict, suggesting that *crr*, and potentially some interaction involving *ptsG_2* and *ptsG_3*, could instead [of *ptsG_6*] be transporting glucose. This supports both the literature conclusions and the metabolic data presented in this writing [2.3.2. and 2.3.3.] that *S. glossinidius* is able to utilise glucose as a carbon source (Belda *et al*., 2012; Goodhead *et al*., 2018). Visualising *ptsG_2* and *ptsG_3* in Artemis (Carver *et al*., 2012), there is 53 bp overlap [starting with ATG and ending with TAG] between the two genes, suggesting that they are most likely one pseudogene, but have been fragmented in the annotation due to the software detecting the codons.

5.3.5. Validation

The *S. glossinidius* strain SgGMMB1 isolate was sequenced using the Oxford Nanopore Technonologies© MinION™ [1D run on a single flowcell], in order to identify sequence differences to *S. glossinidius* strain SgGMMB4. The genome was assembled using Flye (Kolmogorov *et al*., 2019) and annotated using Prokka (Seemann, 2014). The sequences were near-identical, save for a prophage region missing in SgGMMB1 comparative to SgGMMB4 (Appendix Figure 8).

5.4. Discussion

Functional genomics data, and thus a full functionality understanding, is minimal for the tsetse facultative symbiont, *S. glossinidius*. Previous research has linked the symbiont's presence with reduced host refractoriness to trypanosomes, which are the causative agents for human- and animal African trypanosomiasis (Maudlin, 1982; Maudlin and Ellis, 1985; Welburn *et al.*, 1993; Wamwiri *et al*., 2013; WHO, 2019a). Unlike *S. praecaptivus* that has only been observed environmentally, *S. glossinidius* has only been isolated from its host organism, confirming its identification as 'symbiont' (Clayton *et al*., 2012; Chari *et al*., 2015). This is supported by its heavily degraded genome, which a trait resulting from a lifestyle switch from free-living to host-associated, where bacterial gene selection relaxes due host-compensation (reviewed in McCutcheon and Moran, 2012). However, the unique [to symbionts] ability to culture *S. glossinidius in vitro* devoid of a cell line is suggestive of a progressing evolutionary phenotype; the genetic degradation typically rendering symbiont genomes unable to support independent laboratory cultivation is not fully observed *S. glossinidius*, as it has retained a functional gene repertoire more associated with free-living organisms (Toh *et al.*, 2006; Belda *et al.*, 2010; reviewed in McCutcheon and Moran, 2012). This phenotype offers a unique opportunity to study the functionality of an organism that has not yet fully established as a host-restricted symbiont, but its degraded genetic content has resulted in a highly fastidious model and a distinct lack of functionality publications. Laboratory plating requires rich agar supplementation with blood, a microaerophilic atmosphere and incubation at 25°C in order to achieve colonies, after 5 – 10 days. Laboratory liquid culturing requires rich media statically incubated at 25°C in order to achieve mid-log phase growth, after 3 – 5 days (Matthew *et al*., 2005). Publications that have explored *S. glossinidius* functionality include the *in vitro*-based study identifying that a TTSS is used for host-cell invasion, and sequence- and *in silico*-based studies that have aimed to elucidate its metabolic landscape (Dale *et al*., 2001; Toh *et al*., 2006; Belda *et al.*, 2010; Belda *et al*., 2012; Hall *et al.*, 2019). The invasion-dynamics study [that utilised a transformed mutant], in conjunction with additional studies that have transformed *S. glossinidius*, demonstrate the

symbiont's scope for genetic manipulation (Beard *et al.*, 1993; Dale *et al*., 2001; Pontes and Dale, 2011; De Vooght *et al*., 2014). This manipulation advantage was utilised in the data presented here, where *S. glossinidius* was transformed with the Tn5 transposon in order to create a TraDIS library. TraDIS has been successfully used to clarify functionality within specific environments for organisms such as *Escherichia* spp. and *Salmonella* spp., seeing library sizes ranging from 1 million to 8,550 mutants (Langridge *et al*., 2009; Chaudhuri *et al*., 2013; Goodall *et al*., 2018; Canals *et al*., 2019). The *S. glossinidius* library presented here comprising of > 10,600 mutants is novel for the organism and represents new insights into the symbiont's functionality within targeted selection environments.

The TraDIS library presented here revealed essential gene candidates for selection environments including long-term establishment in a complex growth medium, complex versus minimal medium survival, and overall essentiality using a [computationally] combined sample. As previous studies have identified that *S. glossinidius* has retained a relatively broad range of pathways and associated mechanisms, the hypothesis for these experiments assumed that the genes involved in these would be deemed essential. This extends to the virulence-associated genes and unusually high number of pseudogenes that have not only been retained during the degradation process, but actively maintained across lineages, suggesting that these gene sets are important for *S. glossinidius* maintaining symbiosis (Toh *et al.*, 2006; Belda *et al.*, 2010; Goodhead *et al*., 2018). The essential gene candidates identified as generally essential [i.e. shared between selections, and within the combined sample] for *S. glossinidius* confirm the literature conclusions that the symbiont has retained a range of utilisation pathways more associated with free-living organisms, and subsequently strengthens these conclusions by demonstrating that *S. glossinidius* is actively using them (Toh *et al.*, 2006; Belda *et al.*, 2010). These pathways include the CAC, glycolysis, gluconeogenesis, the PPP, oxidation phosphorylation, siderophore utilisation, and those involved with a range of carbon- and amino acid metabolism. The essentiality candidacy within the metabolism [including biosynthesis] pathways supports the literature conclusions that *S. glossinidius* is able to utilise a broad

range of carbon sources (Toh *et al.*, 2006; Belda *et al.*, 2010). Also of-interest within these pathways was the essentiality candidacy specifically concerning arginine and thiamine, as both these pathways have been previously deemed as non-functional in *S. glossinidius* (Toh *et al*., 2006; Belda *et al*., 2010; Belda *et al*., 2012; Hall *et al*., 2019). Essential candidates partially covering the biosynthesis/metabolism and transport of arginine are potentially more in-line with recent conclusions that *S. glossinidius* is not a true arginine auxotroph (Hall *et al*., 2019). The low level of essentiality within the thiamine pathway supports the previous conclusions that the thiamine pathway is non-functional, and that *S. glossinidius* could be using a salvage pathway to compensate (Belda *et al*., 2010; Belda *et al*., 2012). The few genes that do hold essential candidacy could potentially be involved in the proposed *yjbQ*-encoded work-around, although unconfirmed (Belda *et al*., 2010; Belda *et al*., 2012). The essentiality within various pathways involving glucose [for example, glycolysis] not only supports the data presented in this writing (Chapter 2) that *S. glossinidius* actively utilises the carbohydrate, but in particular selections could preference it; for long-term establishment in a complex medium offering both glucose and trehalose, only systems involving glucose were amongst the essential candidates. Also present [amongst the essential gene candidates] for long-term establishment in a complex medium offering levels of nutrients one could expect from an intercellular environment were virulence-associated gene sets, including those that have been identified as PAMPs/PAMP-like (Omp porins and flagellar components), or those vital for host infection (TTSS, CAMP resistance), in pathogenic organisms (Isibasi *et al*., 1992; Isibasi *et al*., 1994; Jones *et al*., 1994; Penheiter *et al*., 1997; reviewed in McCarter, 2006; Pérez-Toledo *et al*., 2017). The essentiality not only confirms a potential ancestral virulence phenotype for *S. glossinidius*, but also that the active retention of the gene sets in the recent host-restricted environment advocates importance for symbiosis. Observed within the minimal medium selection pressure restricted to the glucose derivative *N*-acetylglucosamine, was essentiality within the [glucose epimer] mannose PTS, confirming the hypothesis based on the data presented in Chapter 2 that the remaining intact *S. glossinidius* transporters are lenient to similar compounds.

Specific transporters reported as intact within the literature include the *N*-acetylglucosamine, mannose and mannitol PTSs, and the galactoside, thiamine, hemin, sulfate, putrescine and arginine ABC transporters (Figure 2.1) (Toh *et al*., 2006). More general intact transport systems include the sugar, metal ion, iron, di-, oligo-, and peptide ABC transporters (Figure 2.1) (Toh *et al*., 2006). Further to the mannose PTS, the gene essentiality data presented here also reveals that the galactoside, thiamine, putrescine, arginine, iron, di-, and oligopeptide ABC transporters are particularly important for *S. glossinidius*. Many of the essential gene candidates discussed here are pseudogenised, which is a typical feature of genome degradation associated with the lifestyle switch from free-living to symbiotic (reviewed in McCutcheon and Moran, 2012). However, obligate symbionts have minimal numbers of pseudogenes, the current opinion being that genes being pseudogenised in this way are no longer needed and eventually removed entirely from the genome; for example the *Wigglesworthia glossinidia* genome contains only 0.87% (6/686) pseudogenes (Aksoy, 1995a; Akman *et al*., 2002; reviewed in McCutcheon and Moran, 2012). The fact that *S. glossinidius* is actively retaining its pseudogenes would suggest not only that they are important for symbiosis, but also that organisms in the early stages of switching bias their [pseudogenes] retention (Goodhead *et al*., 2018). Further to this, COG annotations revealed a pseudogene bias within Metabolism subcategories, suggesting that the involved genes are of particular importance at the current evolutionary staging of *S. glossinidius*, but could be undergoing removal as *S. glossinidius* fully establishes as a symbiont.

The data presented here provides the missing functional genomics data currently surrounding *S. glossinidius*. The essential gene candidates provide defined targets for further functionality studies, such as *in vivo* host experiments utilising targeted knock-out *S. glossinidius* mutants, and the missing tool for fully understanding the *S. glossinidius* role in tsetse.

Chapter 6; General Discussion

Sodalis glossinidius is a facultative symbiont of the trypanosome disease vector, the tsetse fly [Dipteran: *Glossina* spp.]. Evidence suggests that *S. glossinidius* has only recently made the lifestyle switch from free-living to host-associated, including isolates from phylogenetically distant *Glossina* species' showing high levels of sequence similarity and low SNP rates (Aksoy *et al*., 1997; Goodhead *et al*., 2018). This recent association offers the ability to study an organism currently in the process of establishing as a symbiont, as genomic degradation resulting from the lifestyle switch has not yet fully eliminated the functional repertoire possessed by free-living organisms (Toh *et al*., 2006; Belda *et al*., 2010; reviewed in McCutcheon and Moran, 2012; Goodhead *et al*., 2018). This results in the ability to culture and study *S. glossinidius in vitro* devoid of a cell line (Dale and Maudlin, 1999). Other evidences supporting the current evolutionary stage of *S. glossinidius* include the retention of a variety of energy production and metabolism pathways, the retention of virulence-associated gene sets, a genome size [4.2 Mb] in-line with free-living bacteria, a heavily degraded genome with an unusually high number of pseudogenes, and the fact that it has only been isolated from its host [i.e. not found in the environment] (Darby *et al*., 2005; Toh *et al*., 2006; Belda *et al*., 2010; Goodhead *et al*., 2018). The culturability of *S. glossinidius* comes with a caveat however, as the symbiont's *in vitro* phenotype is highly fastidious in nature, with long growth times, the need for rich solid- and liquid media, and a very high tendency for contamination (Matthew *et al*., 2005). This has fed into the knowledge gap currently within the *S. glossinidius* field, as difficulty in genetic manipulation experiments due to both the fastidiousness and the degraded genome of the symbiont have led to very few publications successfully studying its genetic functionality. Contradictory results from studies unable to replicate previous work has also created controversiality within conclusions of the *S. glossinidius* role in tsetse. One study found higher proportions of trypanosome infections in *Sodalis*-positive *G. pallidipes* than *Sodalis*-negative, whilst another found no significant difference in *T. b. brucei* mid-gut establishment or development in the salivary glands between *S. glossinidius*-negative *G. m. morsitans* versus *S. glossinidius*-

positive *G. m. morsitans* (Wamwiri *et al*., 2013; Trappeniers *et al*., 2019). This link between trypanosome infection rate and *S. glossinidius* presence was noted early on in studies linking maternally inherited traits [RLO, later re-classified as *S. glossinidius*] to a lowered tsetse refractoriness to trypanosomes (Maudlin, 1982; Maudlin and Ellis, 1985; Welburn and Maudlin, 1990; Dale and Maudlin, 1999). The array of studies that have explored *S. glossinidius* functionality have involved assimilation tests using complex media, predictions based on the genome sequence, *in silico* models, or limited-scope genetic manipulation studies; miniTn5 *S. glossinidius* mutants were used only to study the invasion dynamics of insect cells, and the creation of λ Red recombination knock-out mutants was reported and these subsequently used to identify a *S. glossinidius* AMP-sensitive phenotype with the inactivation of the *phoP* response regulator (Dale and Maudlin, 1999; Dale *et al*., 2001; Toh *et al.*, 2006; Belda *et al*., 2010; Pontes and Dale, 2011; Pontes *et al*., 2011; Belda *et al*., 2012; Hall *et al*., 2019). These genetics studies, in conjunction with the recombinant *Sodalis* expressing a trypanolytic nanobody, demonstrate that the symbiont can be successfully transformed, but a functional genomics sequence data set remains missing within the literature (Dale *et al*., 2001; Pontes and Dale, 2011; De Vooght *et al*., 2014). Examples of such data sets include the transposon libraries created in related organisms, which have been successfully utilised to clarify genetic functionality in chosen environments, where 355 genes were found to be essential for *Escherichia coli* K-12 *in vitro* growth, 356 genes were essential for *Salmonella enterica* serovar Typhi *in vitro* growth, 68 essential genes for intra-macrophage *S.* Typhimurium ST_{313} strain D23580 replication, and 611 genes classed as "potentially important" for *S.* Typhimurium ST4/74 colonisation in chickens, pigs and calves (Langridge *et al*., 2009; Chaudhuri *et al*., 2013; Goodall *et al*., 2018; Canals *et al*., 2019). Many of these studies have used TraDIS to clarify which genes are important for a given selection pressure environment; fully detailed in 1.2.4, but briefly the method pins on the random insertion of transposons, selection of the mutants within specific pressures, and amplification of transposon-flanking sequences in the surviving mutants (Langridge *et al*., 2009; Chaudhuri *et al*., 2013; Barquist *et al*., 2016; Goodall *et al*., 2018).

The work presented within this writing was conducted with the aim of providing the currently missing functional genomics information within the *S. glossinidius* field, by adapting and optimising the transformation methods successfully used in the related free-living organisms. Many of the optimisation experiments conducted throughout this project lead to the unsuccessful transformation of *S. glossinidius*, not only confirming the symbiont's fastidiousness phenotype in a genetic context, but also the "reproducibility crisis" voiced within the scientific community; transformation methods previously published as successful in the creation of *S. glossinidius* mutants could not be reproduced here [3.3.2.] (Pontes and Dale, 2011; Baker, 2016). The unsuccessful utilisation of a kit-based (Illumina, Inc.©) method of transformation that has been used for related organisms can only be attributed to *S. glossinidius* fastidiousness, in this case [3.3.1.] (Langridge *et al*., 2009; Canals *et al*., 2012; Chaudhuri *et al*., 2013; Goodall *et al*., 2018; Canals *et al*., 2019). The successful creation of the *S. glossinidius* Tn5 library was achieved using a plasposon delivery system; pRL27 containing a kanamycin-marked mini-Tn5 element, and a hyperactive transposase [5.2.1.1.] (Dennis and Zylstra, 1998; Larsen *et al*., 2002). The library contains > 10,600 *S. glossinidius* Tn5 mutants and has been selected-for in multiple pressured environments, including general and long-term *S. glossinidius* establishment in a complex medium [5.2.2.2.], *in vitro* growth in a nutrient-restricted environment [5.2.2.3.], and intracellular development within an insect cell host [5.2.2.4.]. With the current literature conclusions that *S. glossinidius* has retained a range of energy and metabolism pathways, the hypothesis from experiments utilising a complex medium abundant in growth nutrients [5.2.2.2.] was that the genes involved in a range of pathways would be deemed essential (Toh *et al*., 2006; Belda *et al*., 2010; Goodhead *et al*., 2018). The *S. glossinidius*-adapted analysis [5.2.4.3. - 4.] conducted on the mutants selected-for in this way confirmed the hypothesis, as essentiality was observed within energy pathways such as glycolysis, gluconeogenesis, the PPP, the CAC, and oxidative phosphorylation, and a range of [amino acid and carbon] metabolism pathways [5.3.2.1. – 2.]. Within these results was a strong focus on *S. glossinidius* utilisation of glucose, which is a disputed [primary] carbon source

within the literature. Early studies using assimilation tests found stronger growth when using *N*-acetylglucosamine and raffinose as primary carbon sources, in comparison to glucose (Dale and Maudlin, 1999). This *N*-acetylglucosamine preferencing in *S. glossinidius* was also observed in a more recent *in silico* study, but this conclusion was partly based on an incorrect assessment that a *S. glossinidius* glucose-specific transport system is pseudogenised (Hall *et al*., 2019). Whilst some of the PtsG gene copies are pseudogenised, *ptsG_6* is intact, 1.28 Kb in length and has RNASeq evidence of transcription (Goodhead *et al*., 2018). Contrary to these glucose conclusions, Belda *et al*. (2012) found similar [*in silico*] growth levels between *S. glossinidius* and *E. coli* when using glucose, which is supported by the data presented here in Chapter 2, where *S. glossinidius* growth is significantly increased in the presence of glucose [2.3.2.], and at higher concentrations there is no significant difference between glucose and *N*-acetylglucosamine [2.3.3.]. This data suggests not only that *S. glossinidius* does not preference *N*-acetylglucosamine over glucose, but also confirms [in conjunction with the other sources tested in 2.3.3.] the literature conclusions that *S. glossinidius* has retained the ability to utilise a range of carbon sources; glucose, raffinose, *N*-acetylglucosamine, mannitol, and maltose all aided in significantly greater growth values [than a negative control] [2.3.3.] (Toh *et al.*, 2006; Belda *et al*., 2010). Other of-note essentiality within this analysis involves the arginine and thiamine pathways previously deemed non-functional in *S. glossinidius* (Belda *et al*., 2010; Belda *et al*., 2012; Hall *et al*., 2019). Partial essentiality coverage of the biosynthesis/metabolism and transport of arginine is more in-line with recent conclusions that *S. glossinidius* is not a true arginine auxotroph, and the low level of essentiality within the thiamine pathway supports the previous non-functional conclusions (Toh *et al*., 2006; Belda *et al*., 2010; Hall *et al*., 2019). Also observed within the TraDIS selection pressures involving a rich medium offering levels of nutrients one could expect from an intercellular environment was essentiality within virulence-associated gene sets, including those that have been identified as PAMPs/PAMP-like (Omp porins and flagellar components), or those vital for host infection (TTSS, CAMP resistance), in pathogenic organisms (Isibasi *et al*., 1992; Isibasi *et al*., 1994; Jones *et al*., 1994;

Penheiter *et al*., 1997; reviewed in McCarter, 2006; Pérez-Toledo *et al*., 2017). When considering the current evolutionary placement of *S. glossinidius*, both the active retention and essentiality of virulence-associated gene sets not only confirms a potential ancestral virulent phenotype, but also that these genes are likely important for symbiosis (Toh *et al*., 2006; Belda *et al.*, 2010; Goodhead *et al*., 2018). The hypothesis concerning the nutrient-restricted TraDIS selection pressures [5.2.2.3.] predicted that the essential gene range would be more defined, and additionally confirm the separate hypothesis that the remaining intact *S. glossinidius* transporters are lenient; results from 2.3.3. suggested that raffinose and maltose could be sharing the intact transporters specific to structurally-similar compounds, as they are without specific functional transport systems. The results within the minimal medium selection pressure restricted to the glucose derivative *N*-acetylglucosamine [5.3.3.] confirm the associated hypotheses, via observations that the [glucose epimer] mannose PTS was essential, suggesting that *S. glossinidius* was leniently allowing passage of *N*-acetylglucosamine through this PTS. As the samples involved with the *S. glossinidius* intracellular development within an insect cell host [5.2.2.4.] did not pass the minimal quality thresholds set within this analysis [5.3.1.], the hypothesis that virulence- and free-living associated genes, and pseudogenes, are important for symbiosis specifically within this condition remains unconfirmed. However, this hypothesis been separately confirmed if one considers the nutrient-rich medium as proxy for intracellular environments; virulence- and free-living associated genes were essential candidates in 5.3.2.1. – 2.. Many of the essential gene candidates were pseudogenised, which when considering *S. glossinidius* is actively retaining its pseudogenes, would suggest not only that they are important for symbiosis, but also that organisms in the early stages of switching bias their retention (Goodhead *et al*., 2018).

The data presented here provides the missing functional genomics data currently surrounding *S. glossinidius*, with the essential gene candidates providing defined targets for further functionality studies, such as *in vivo* host experiments utilising targeted knock-out *S. glossinidius* mutants, and the missing tool for fully understanding the *S. glossinidius* role in tsetse.

References

Abubakar, L. *et al.* (1995). Properties of a Blood-Meal-Induced Midgut Lectin from the Tsetse Fly *Glossina morsitans. Parasitology Research*. 81(4), pp. 271 - 275

Adam, Y. *et al*. (2013). The Sequential Aerosol Technique: A Major Component in an Integrated Strategy of Intervention against Riverine Tsetse in Ghana. *PLoS Neglected Tropical Diseases*. 7(3): e2135

Adams, J.H. *et al.* (1986). Human African Trypanosomiasis (*T.b. gambiense*): A Study of 16 Fatal Cases of Sleeping Sickness with Some Observations on Acute Reactive Arsenical Encephalopathy. *Neuropathol Appl Neurobiol*. 1, pp. 81 - 94

Akman, L. *et al*. (2001). Genome Size Determination and Coding Capacity of *Sodalis glossinidius*, an Enteric Symbiont of Tsetse Flies, as Revealed by Hybridization to *Escherichia coli* Gene Arrays. *Journal of Bacteriology*. 183(15), pp. 4517 - 4525

Akman, L. *et al*. (2002). Genome Sequence of the Endocellular Obligate Symbiont of Tsetse Flies, *Wigglesworthia glossinidia. Nature Genetics.* 32, pp. 402 - 407

Aksoy, S. (1995)a. *Wigglesworthia* gen. nov. and *Wigglesworthia glossinidia* sp. nov., Taxa Consisting of the Mycetocyte-Associated, Primary Endosymbionts of Tsetse Flies. *Int J Syst Bacteriol*. 45(4), pp. 848 – 851

Aksoy, S. *et al*. (1995)b. Mycetome Endosymbionts of Tsetse Flies Constitute a Distinct Lineage Related to Enterobacteriaceae. *Insect Molecular Biology*. 4(1), pp. 15 - 22

Aksoy, S. *et al.* (1997). Phylogeny and Potential Transmission Routes of Midgut-Associated Endosymbionts of Tsetse (Diptera: Glossinidae). *Insect Molecular Biology.* 6(2), pp. 183 - 190

Aksoy, S. *et al*. (2008). Paratransgenesis Applied for Control of Tsetse Transmitted Sleeping Sickness. *Adv Exp Med Biol.* 627, pp. 35 - 48

Aksoy, S. *et al*., (2014). Trypanosome Transmission Dynamics in Tsetse. *Curr Opin Insect Sci.* 3, pp. 43 – 49

Alam, U. *et al*. (2011). *Wolbachia* Symbiont Infections Induce Strong Cytoplasmic Incompatibility in the Tsetse Fly *Glossina morsitans. PLoS Pathog*. 7(12):e100241

Atouguia, J.L.M. and Kennedy, P.G.E. (2000). Neurological Aspects of Human African Trypanosomiasis. *Infectious Diseases of the Nervous System*. pp. 321–72. Butterworth-Heinemann, Oxford

Attardo, G.M. *et al*. (2008). Analysis of Milk Gland Structure and Function in *Glossina morsitans*: Milk Protein Production, Symbiont Populations and Fecundity. *Journal of Insect Physiology*. 54(8), pp. 1236 - 1242

Baba, T. *et al.* (2006). Construction of *Escherichia coli* K-12 In-Frame, Single-Gene Knockout Mutants: The Keio Collection. *Mol Syst Biol*. 2: 2006

Baker, M. (2016). 1,500 Scientists Lift the Lid on Reproducibility. *Nature* (News Feature). 533(7604)

Balmand, S. *et al*. (2013). Tissue Distribution and Transmission Routes for the Tsetse Fly Endosymbionts. *Journal of Invertebrate Pathology*. 112(S1), pp. S116 – S122

Barquist, L. *et al*. (2016). The TraDIS Toolkit: Sequencing and Analysis for Dense Transposon Mutant Libraries. 32(7), pp. 1109 - 1111

Bauer, B. *et al* (1999). Improvement of Cattle Productivity through Rapid Alleviation of African Animal Trypanosomosis by Integrated Disease Management Practices in the Agropastoral Zone of Yalé, Burkina Faso. *Tropical Animal Health and Production.* 31(2), pp. 89 – 102

Beard, C.B. *et al*. (1993). Genetic Transformation and Phylogeny of Bacterial Symbionts from Tsetse. *Insect Molecular Biology*. 1(3), pp. 123 - 131

Belda, E. *et al*. (2010). Mobile Genetic Element Proliferation and Gene Inactivation Impact Over the Genome Structure and Metabolic Capabilities of *Sodalis glossinidius*, the Secondary Endosymbiont of Tsetse Flies. *BMC Genomics*. 11:449

Belda, E. *et al.* (2012). Metabolic Networks of *Sodalis glossinidius*: A Systems Biology Approach to Reductive Evolution. *PLoS One*. 7(1):e30652

Berg, J.M. *et al*. (2002). Biochemistry, 5th Edition. *W. H. Freeman and Company*. Chapter 17 (The Citric Acid Cycle)

Blöchl, E. *et al*. (1997). *Pyrolobus fumarii*, gen. and sp. nov., Represents a Novel Group of Archaea, Extending the Upper Temperature Limit for Life to 113 Degrees C. *Extremophiles*. 1(1), pp. 14 – 21

Blum, J. *et al*. (2006). Clinical Aspects of 2541 Patients with Second Stage Human African Trypanosomiasis. *Acta Tropica.* 97(1), pp. 55 - 64

Brun, R. *et al.* (2010). Human African Trypanosomiasis. *The Lancet*. 375(9709), pp. 148 - 159

Buguet, A. *et al*. (2005). Sleep Structure: A New Diagnostic Tool for Stage Determination in Sleeping Sickness. *Acta Tropica.* 93(1), pp. 107 - 117

Burke, G.R. *et al*. (2009). Evolution and Diversity of Facultative Symbionts from the Aphid Subfamily Lachninae. *Appl Environ Microbiol*. 75(16), pp. 5328 - 5335

Bursell, E. (1966). Aspects of the Flight Metabolism of Tsetse Flies (*Glossina*). *Comparative Biochemistry and Physiology*. 19(4), pp. 809 - 818

Canals, R. *et al*. (2012). High-Throughput Comparison of Gene Fitness Among Related Bacteria. *BMC Genomics*. 13:212

Canals, R. *et al*. (2019). The Fitness Landscape of the African *Salmonella* Typhimurium ST_{313} Strain D23580 Reveals Unique Properties of the pBT1 Plasmid. *PLoS Pathogens.* 15(9):e1007948

Carrel, A. and Burrows, M.T. (1911)a. Cultivation of Tissues *in vitro* and its Technique. *J Exp Med.* 13(3), pp. 387 – 396

Carrel, A. and Burrows, M.T. (1911)b. An Addition to the Technique of the Cultivation of Tissues *in vitro. J Exp Med.* 14(3), pp. 244 – 247

Carver, T. *et al*. (2012). Artemis: An Integrated Platform for Visualization and Analysis of High-Throughput Sequence-Based Experimental Data. *Bioinformatics.* 28(4), pp. 464 - 469

Caspi, R. *et al.* (2018). The MetaCyc Database of Metabolic Pathways and Enzymes. *Nucleic Acids Research.* 46(D1), pp. D633 - D639

Castagnoli, L. *et al*. (1989). Genetic and Structural Analysis of the ColE1 Rop (Rom) Protein. *EMBO J.* 8(2), pp. 621 – 629

Centers for Disease Control (CDC) (2019). Epidemiology & Risk Factors article, accessed 2019 at https://www.cdc.gov/parasites/sleepingsickness/epi.html

Chakravortty, D. *et al*. (2002). Salmonella Pathogenicity Island 2 Mediates Protection of Intracellular Salmonella from Reactive Nitrogen Intermediates. *Journal of Experimental Medicine*. 195(9), pp. 1155 - 1166

Chappuis, F. (2018). Oral Fexinidazole for Human African Trypanosomiasis. *The Lancet.* 391(10116), pp. 100 – 102

Chari, A. *et al*. (2015). Phenotypic Characterization of *Sodalis praecaptivus* sp. nov., a Close Non-Insect-Associated Member of the *Sodalis*-Allied Lineage of Insect Endosymbionts. *International Journal of Systematic and Evolutionary Microbiology*. 65, pp. 1400 – 1405

Chaudhuri, R.R. *et al*. (2013). Comprehensive Assignment of Roles for *Salmonella* Typhimurium Genes in Intestinal Colonization of Food-Producing Animals. *PLoS Genet*. 9(4):e1003456

Chemical Entities of Biological Interest (ChEBI) (2019). The European Bioinformatics Institute (EMBL-EBI). Accessed in 2019 via https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:17148

Chen, X. *et al*. (1999). Concordant Evolution of a Symbiont with its Host Insect Species: Molecular Phylogeny of Genus *Glossina* and its Bacteriome-Associated Endosymbiont, *Wigglesworthia glossinidia. Journal of Molecular Evolution*. 48(1), pp. 49 - 58

Chrudimský, T. *et al*. (2012). Candidatus *Sodalis melophagi* sp. nov.: Phylogenetically Independent Comparative Model to the Tsetse Fly Symbiont *Sodalis glossinidius. PLoS One.* 7(7):e40354

Clayton, A.L. *et al*. (2012). A Novel Human-Infection-Derived Bacterium Provides Insights into the Evolutionary Origins of Mutualistic Insect–Bacterial Symbioses. *PLoS Genetics.* 8(11):e1002990

Clayton, A.L. *et al.* (2017). The Regulation of Antimicrobial Peptide Resistance in the Transition to Insect Symbiosis. *Mol Microbiol*. 103(6), pp. 958 – 972

Conesa, A. *et al*. (2005). Blast2GO: A Universal Tool for Annotation, Visualization and Analysis in Functional Genomics Research. *Bioinformatics.* 21(18), pp. 3674 - 3676

Cunin, R. *et al*. (1986). Biosynthesis and Metabolism of Arginine in Bacteria. *Microbiol Rev*. 50(3), pp. 314 – 352

Dale, C. and Maudlin, I. (1999). *Sodalis* gen. nov. and *Sodalis glossinidius* sp. nov., a Microaerophilic Secondary Endosymbiont of the Tsetse Fly *Glossina morsitans morsitans*. *International Journal of Systematic Bacteriology*. 49, 267 – 275.

Dale, C. *et al*. (2001). The Insect Endosymbiont *Sodalis glossinidius* Utilizes a Type III Secretion System for Cell Invasion. *PNAS*. 98(4), pp. 1883 – 1888

Dale, C. *et al*. (2005). Degenerative Evolution and Functional Diversification of Type-III Secretion Systems in the Insect Endosymbiont *Sodalis glossinidius. Molecular Biology and Evolution.* 22(3), pp. 758 – 766

Darby, A.C. *et al*. (2005). Extrachromosomal DNA of the Symbiont *Sodalis glossinidius. J. Bacteriol.* 187(14), pp. 5003 - 5007

Darling, A.C.E. *et al*. (2004). Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Res*. 14(7), pp. 1394 – 1403

Datsenko, K.A. and Wanner,B.L. (2000). One-Step Inactivation of Chromosomal Genes in *Escherichia coli* K-12 Using PCR Products. *Proceedings of the National Academy of Sciences of the United States of America*. 97(12), pp. 6640 – 6645

De Felice, M. *et al*. (1977). Growth Inhibition of *Escherichia coli* K-12 by L-valine: A Consequence of a Regulatory Pattern. *Molecular and General Genetics*. 156(1), pp 1 – 7

de Lorenzo, V. *et al*. (1990). Mini-Tn5 Transposon Derivatives for Insertion Mutagenesis, Promoter Probing, and Chromosomal Insertion of Cloned DNA in Gram-Negative Eubacteria. *J Bacteriol*. 172(11), pp. 6568 – 6572

De Vooght, L. *et al*. (2014). Delivery of a Functional Anti-Trypanosome Nanobody in Different Tsetse Fly Tissues via a Bacterial Symbiont, *Sodalis glossinidius. Microb Cell Fact.* 13: 156

De Vooght, L. *et al*. (2015). Paternal Transmission of a Secondary Symbiont during Mating in the Viviparous Tsetse Fly. *Molecular Biology and Evolution.* 32(8), pp. 1977 - 1980

Denlinger, D.L. and Ma, W. (1975). Maternal Nutritive Secretions as Possible Channels for Vertical Transmission of Microorganisms in Insects: The Tsetse Fly Example. *Pathobiology of Invertebrate Vectors of Disease.* 226(1), pp. 162 - 165

Dennis, J.J. and Zylstra, G.J. (1998). Plasposons: Modular Self-Cloning Minitransposon Derivatives for Rapid Genetic Analysis of Gram-Negative Bacterial Genomes. *Appl Environ Microbiol*. 64(7), pp. 2710 - 2715

Diall, O. *et al*. (2017). Developing a Progressive Control Pathway for African Animal Trypanosomiasis. *Trends in Parasitology*. 33(7), pp. 499 - 509

Dong, T. and Schellhorn, H.E. (2009). Control of RpoS in Global Gene Expression of *Escherichia coli* in Minimal Media. *Molecular Genetics and Genomics.* 281(1), pp. 19 – 33

Doudoumis, V. *et al*. (2017). Challenging the *Wigglesworthia, Sodalis, Wolbachia* Symbiosis Dogma in Tsetse Flies: *Spiroplasma* is Present in Both Laboratory and Natural Populations. *Scientific Reports*. 7:4699

Duchaud, E. *et al*. (2003). The Genome Sequence of the Entomopathogenic Bacterium *Photorhabdus luminescens. Nature Biotechnology*. 21, pp. 1307 - 1313

Dyer, N. *et al*. (2013). Flying Tryps: Survival and Maturation of Trypanosomes in Tsetse Flies. *Trends in Parasitology.* 29(4), pp. 188 - 196

Ebhodaghe, F. *et al*. (2018). A Systematic Review and Meta-Analysis of Small Ruminant and Porcine Trypanosomiasis Prevalence in Sub-Saharan Africa (1986 – 2018). *Acta Tropica.* 188, pp. 118 - 131

Fairlamb, A.H. *et al.* (1985). Trypanothione: A Novel Bis(Glutathionyl)Spermidine Cofactor for Glutathione Reductase in Trypanosomatids. *Science*. 227(4693), pp. 1485 - 1487

Farming In Tsetse Controlled Areas (FITCA) (2005). Kenya Project, 1999 – 2004: Lessons Learned

Fields, P.I. *et al*. (1986). Mutants of *Salmonella typhimurium* That Cannot Survive Within the Macrophage are Avirulent. *PNAS*. 83(14), pp. 5189 - 5193

Fleischmann, R.D. *et al.* (1995). Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd. *Science*. 269(5223), pp. 496 - 512

Foultier, B. *et al*. (2002). Characterization of the *ysa* Pathogenicity Locus in the Chromosome of *Yersinia enterocolitica* and Phylogeny Analysis of Type III Secretion Systems. *J. Mol. Evol.* 55, pp. 37 - 51

Galperin, M.Y. and Koonin, E.V. (2010). From Complete Genome Sequence to 'Complete' Understanding? *Trends in Biotechnology*. 28(8), pp. 398 – 406

Galperin, M.Y. *et al*. (2015). Expanded Microbial Genome Coverage and Improved Protein Family Annotation in the COG Database. *Nucleic Acids Res*. 43(D1), D261 – D269

Gibson, W. and Bailey, M. (2003). The Development of *Trypanosoma brucei* Within the Tsetse Fly Midgut Observed Using Green Fluorescent Trypanosomes. *Kinetoplastid Biology and Disease*. 2:1

Gladstone, G. P. (1939). Inter-Relationships Between Amino-Acids in the Nutrition of *B. anthracis*. *Br J Exp Pathol.* 20(2), pp. 189 – 200

Glass, J.I. *et al.* (2006). Essential Genes of a Minimal Bacterium. *Proceedings of the National Academy of Sciences of the United States of America*. 103(2), pp. 425 – 430

Goodall, E.C.A. *et al.* (2018). The Essential Genome of *Escherichia coli* K-12. *MBIO*. 9(1), e02096-17

Goodhead, I. and Darby, A.C. (2015). Taking the Pseudo out of Pseudogenes. *Current Opinion in Microbiology*. 23, pp. 102 - 109

Goodhead, I. *et al*. (2018). Large Scale and Significant Expression from Pseudogenes in *Sodalis glossinidius* - A Facultative Bacterial Endosymbiont. *bioRxiv.* doi: https://doi.org/10.1101/124388

Grangeasse, C. *et al.* (2012). Bacterial Tyrosine Kinases: Evolution, Biological Function and Structural Insights. *Philosophical Transactions of the Royal Society B-Biological Sciences*. 367(1602), pp. 2640 – 2655

Groisman, E.A. (2001). The Pleiotropic Two-Component Regulatory System PhoP-PhoQ. *J Bacteriol.* 183(6), pp. 1835 – 1842

Grünwald, S. *et al*. (2010). Microbial Associations in Gut Systems of Wood- and Bark-Inhabiting Longhorned Beetles [Coleoptera: Cerambycidae]. *Systematic and Applied Microbiology.* 33(1), pp. 25 - 34

Hall, R. *et al*. (2019). A Tale of Three Species: Adaptation of *Sodalis glossinidius* to Tsetse Biology, *Wigglesworthia* Metabolism, and Host Diet. *MBio*. 10(1):e02106-18

Han, C.G. *et al*. (2001). Structural and Functional Characterization of IS679 and IS66-Family Elements. *J Bacteriol.* 183(14), pp. 4296 - 4304

Hargrove, J.W. (1981). Tsetse Dispersal Reconsidered. *Journal of Animal Ecology.* 50, pp. 351 – 373

Harrison, R.G. (1907). Observations of the Living Developing Nerve Fiber. *Society for Experimental Biology and Medicine*. 1(5), pp. 116 - 128

Hegedus, D. *et al*. (2009). New Insights into Peritrophic Matrix Synthesis, Architecture and Function. *Annual Review of Entomology*. 54, pp. 285 - 302

Herrero-Fresno, A. *et al.* (2014). The Role of the *st_{313}-td* Gene in Virulence of *Salmonella* Typhimurium ST_{313}. *PLOS One.* 9(1), e84566

Hill, P. *et al*. (1973). The Production of "Symbiont-Free" *Glossina morsitans* and an Associated Loss of Female Fertility. *Transactions of The Royal Society of Tropical Medicine and Hygiene.* 67(5), pp. 727 – 728

Hoare, C.A. (1972). The Trypanosomes of Mammals, a Zoological Monograph. Oxford: Blackwell Scientific Publications

Holt, H.R. *et al*. (2016). Assessment of Animal African Trypanosomiasis (AAT) Vulnerability in Cattle-Owning Communities of Sub-Saharan Africa. *Parasites & Vectors*. 9:53

Horn, D. (2014). Antigenic Variation in African Trypanosomes. *Mol Biochem Parasitol.* 195(2), pp. 123 – 129

Hueck, C.J. (1998). Type III Protein Secretion Systems in Bacterial Pathogens of Animals and Plants. *Microbiology and Molecular Biology Reviews*. 62(2), pp. 379 – 433

Huerta-Cepas, J. *et al*. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper*. Mol Biol Evol*. 34(8), pp. 2115 - 2122

Isibasi, A. *et al*. (1992). Active Protection of Mice Against *Salmonella typhi* by Immunization with Strain-Specific Porins. *Vaccine*. 10(12), pp. 811 – 813

Isibasi, A. *et al*. (1994). Role of Porins From *Salmonella typhi* in the Induction of Protective Immunity. *Microbial Pathogenesis and Immune Responses.* 730, pp. 350 - 352

Jamonneau, V. *et al*. (2012). Untreated Human Infections by *Trypanosoma brucei gambiense* are not 100% Fatal. *PLoS Neglected Tropical Diseases*. 6(6): e1691

Jones, B.D. *et al*. (1994). *Salmonella typhimurium* Initiates Murine Infection by Penetrating and Destroying the Specialized Epithelial M Cells of the Peyer's Patches. *J Exp Med.* 180(1), pp. 15 - 23

Kaiwa, N. *et al.* (2010). Primary Gut Symbiont and Secondary, *Sodalis*-allied Symbiont of the Scutellerid Stinkbug *Cantao ocellatus*. *Appl Environ Microbiol.* 76(11), pp. 3486 – 3494

Kamuanga, M. *et al.* (2001). Evaluating Contingent and Actual Contributions to a Local Public Good: Tsetse Control in the Yale Agro-Pastoral Zone, Burkina Faso. *Ecological Economics*. 39(1), pp. 115 - 130

Kanehisa, M. and Goto, S. (2000). Kyoto Encyclopaedia of Genes and Genomes. *Nucleic Acids Res*. 28, pp. 27 - 30

Kanehisa, M. *et al.* (2016). KEGG as a Reference Resource for Gene and Protein Annotation. *Nucleic Acids Res*. 44, D457 - D462

Kanehisa, M. *et al.* (2017). KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res*. 45, D353 - D361

Kaniga, K. *et al*. (1995). Homologs of the *Shigella* IpaB and IpaC Invasins Are Required for *Salmonella typhimurium* Entry into Cultured Epithelial Cells. *Journal of Bacteriology*. 177(14), pp. 3965 - 3971

Kennedy, P.G.E. (2004). Human African Trypanosomiasis of the CNS: Current Issues and Challenges. *J Clin Invest.* 113(4), pp. 496 – 504

Kennedy, P.G.E. (2005). Sleeping Sickness - Human African Trypanosomiasis. *Pract Neurol*. 5, pp. 260 – 267

Kennedy, P.G.E. (2006). Diagnostic and Neuropathogenesis Issues in Human African Trypanosomiasis. *International Journal for Parasitology.* 36(5), pp. 505 – 512

Kennedy, P.G.E. (2008). The Continuing Problem of Human African Trypanosomiasis (Sleeping Sickness). *Ann Neurol.* 64(2), pp. 116 - 126

Kennedy, P.G.E. (2013). Clinical Features, Diagnosis, and Treatment of Human African Trypanosomiasis (Sleeping Sickness). *The Lancet; Neurology.* 12(2), pp. 186 - 194

Keseler, I.M. *et al.* (2017). The EcoCyc Database: Reflecting New Knowledge About Escherichia coli K-12. *Nucleic Acids Research*. 45(D1), pp. D543 – D550

Knipling, E.F. (1959). Sterile-Male Method of Population Control. *Science.* 130(3380), pp. 902 - 904

Kolmogorov, M. *et al*. (2019). Assembly of Long, Error-Prone Reads Using Repeat Graphs. *Nature Biotechnology.* 37, pp. 540 – 546

Krebs, H.A. and Johnson, W.A. (1937)a. Metabolism of Ketonic Acids in Animal Tissues. *Biochem J*. 31(4), pp. 645 - 660

Krebs, H.A. and Johnson, W.A. (1937)b. Acetopyruvic Acid (Alphagamma-Diketovaleric Acid) as an Intermediate Metabolite in Animal Tissues. *Biochem J*. 31(5), pp. 772 - 779

Kristjanson, P.M. *et al*. (1998). Measuring the Costs of African Animal Trypanosomiasis, the Potential Benefits of Control and Returns to Research. *Agricultural Systems*. 59, pp. 79 - 98

Kristensen, D.M. *et al*. (2010). A Low-Polynomial Algorithm for Assembling Clusters of Orthologous Groups from Intergenomic Symmetric Best Matches. *Bioinformatics.* 26(12), pp. 1481 - 1487

Kuhn, H. *et al.* (1979). Defined Minimal Medium for a Thermophilic *Bacillus* sp. Developed by a Chemostat Pulse and Shift Technique. *European Journal of Applied Microbiology and Biotechnology.* 6(4), pp. 341 – 349

Langridge, G.C. *et al.* (2009). Simultaneous Assay of Every *Salmonella* Typhi Gene Using One Million Transposon Mutants. *Genome Research*. 19(12), pp. 2308 – 2316

Larsen, R.A. *et al*. (2002). Genetic Analysis of Pigment Biosynthesis in *Xanthobacter autotrophicus* Py2 Using a New, Highly Efficient Transposon Mutagenesis System that is Functional in a Wide Variety of Bacteria. *Archives of Microbiology*. 178(3), pp 193 – 201

Lawley, T.D. *et al*. (2003). F Factor Conjugation is a True Type IV Secretion System. *FEMS Microbiology Letters*. 224(1), pp. 1 – 15

Lemaitre, B. and Hoffman, J. (2007). The Host Defence of *Drosophila melanogaster*. *Annual Review of Immunology*. 25, pp. 697 - 743

Li, H. *et al*. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics*. 25(16), pp. 2078 - 2079

Li, H. (2011). A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data. *Bioinformatics*. 27(21), pp. 2987 - 2993

Liverpool School of Tropical Medicine (LSTM) (2019). What are the Symptoms of Animal Trypanosomiasis? article, accessed in 2019 at https://www.tsetse.org/tsetse-faq/tsetse-and-trypanosomiasis/what-are-the-symptoms-of-animal-trypanosomiasis

MacLeod, E.T. *et al.* (2007). Antioxidants Promote Establishment of Trypanosome Infections in Tsetse. *Parasitology*. 134, pp. 827 - 831

Madigan, M.T. & Martinko, J.M. (2006)a. Brock Biology of Microorganisms Eleventh Edition. *Pearson Education, Inc., Pearson Prentice Hall™.* Unit I (Principles of Microbiology), Chapter 5 (Nutrition, Laboratory Culture and Metabolism of Microorganisms), Section IV (Major Catabolic Pathways, Electron Transport and the Proton Motive Force)

Madigan, M.T. & Martinko, J.M. (2006)b. Brock Biology of Microorganisms Eleventh Edition. *Pearson Education, Inc., Pearson Prentice Hall™.* Unit I (Principles of Microbiology), Chapter 5 (Nutrition, Laboratory Culture and Metabolism of Microorganisms), Section V (Carbon Flow in Respiration and Catabolic Alternatives)

Madigan, M.T. & Martinko, J.M. (2006)c. Brock Biology of Microorganisms Eleventh Edition. *Pearson Education, Inc., Pearson Prentice Hall™.* Unit III (Metabolic Diversity and Microbial Ecology), Chapter 17 (Metabolic Diversity), Section V (Hydrocarbon Oxidation and the Role of O_2 in the Catabolism of Organic Compounds)

Magnus, E. *et al*. (1978). A Card Agglutination Test with Stained Trypanosomes (CATT) for the Serological Diagnosis of *T. b. gambiense* Trypanosomiasis. *Ann Soc Belge Med Trop*. 58, pp. 169-176

Marceau, A.H. (2012). Functions of single-strand DNA-binding proteins in DNA replication, recombination, and repair. *Methods Mol Biol.* 922, pp. 1 - 21

Martin, M. (2011). Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. *EMBnet.journal*. 17(1), pp. 10 - 12

Matetovici, I. *et al*. (2019). Innate Immunity in the Tsetse Fly (*Glossina*), Vector of African Trypanosomes. *Developmental and Comparative Immunology*. 98, pp. 181 - 188

Matthew, C.Z. *et al*. (2005). The Rapid Isolation and Growth Dynamics of the Tsetse Symbiont *Sodalis glossinidius. FEMS Microbiology Letters.* 248(1), pp. 69 - 74

Maudlin, I. (1982). Inheritance of Susceptibility to *Trypanosoma congolense* Infection in *Glossina morsitans*. *Annals of Tropical Medicine & Parasitology*. 76(2), pp. 225 - 227

Maudlin, I. and Ellis, D.S. (1985). Association Between Intracellular Rickettsial-like Infections of Midgut Cells and Susceptibility to Trypanosome Infection in *Glossina* spp.. *Parasitology*. 71(50, pp. 683 - 687

Maudlin, I. and Welburn, S.C. (1987). Lectin Mediated Establishment of Midgut Infections of *Trypanosoma congolense* and *Trypanosoma brucei* in *Glossina morsitans*. *Tropical Medicine & Parasitology*. 38, pp. 167 - 170

McCutcheon, J.P. and Moran, N.A. (2012). Extreme Genome Reduction in Symbiotic Bacteria. *Nature Reviews Microbiology*. 10, pp. 13 – 26

Miller, J.H. (1972). Experiments in Molecular Genetics. *Cold Spring Harbor Laboratory, Cold Spring Harbor, New York*. pp. 431 - 432.

Milligan, P.J.M *et al*. (1995). Trypanozoon: Infectivity to Humans is Linked to Reduced Transmissibility in Tsetse: II. Genetic Mechanisms. *Experimental Parasitology*. 81(3), pp. 409 - 415

Mira, A. *et al*. (2001). Deletional Bias and The Evolution of Bacterial Genomes. *Trends in Genetics*. 17(10), pp. 589 – 596

Mira, A. and Moran, N. (2002). Estimating Population Size and Transmission Bottlenecks in Maternally Transmitted Endosymbiotic Bacteria. *Microbial Ecology*. 44(2), pp. 137 - 143

Moloo, S.K. and Shaw, M.K. (1989). Rickettsial Infections of Midgut Cells are Not Associated with Susceptibility of *Glossina morsitans centralis* to *Trypanosoma congolense* Infection. *Acta Tropica*. 46(4), pp. 223 - 227

Moran, N. *et al.* (1993). A Molecular Clock in Endosymbiotic Bacteria is Calibrated Using the Insect Hosts. *Proc. R. Soc. Lond.* B. 253, pp. 167 - 171

Moran, A. (1996). Accelerated Evolution and Muller's Rachet in Endosymbiotic Bacteria. *PNAS*. 93(7), pp. 2873 - 2878

Morett, E. *et al*. (2008). Sensitive Genome-Wide Screen for Low Secondary Enzymatic Activities: The YjbQ Family Shows Thiamin Phosphate Synthase Activity. *Journal of Molecular Biology*. 376(3), pp. 839 - 853

Morris, K.R.S. (1946). The Control of Trypanosomiasis by Entomological Means. *Bulletin of Entomological Research.* 37(2), pp. 201 - 250

Murphy, K.C. (1998). Use of Bacteriophage Lambda Recombination Functions to Promote Gene Replacement in *Escherichia coli. Journal of Bacteriology*. 180(8), pp. 2063 – 2071

Nichols, D. *et al.* (2008). Short Peptide Induces an "Uncultivable" Microorganism to Grow *In Vitro. Applied and Environmental Microbiology*. 74(15), pp. 4889 - 4897

Nieto, J.M. *et al.* (1991). The *hha* Gene Modulates Haemolysin Expression in *Escherichia coli*. *Mol Microbiol*. 5(5), pp. 1285 - 1293

Nogge, G. (1976). Sterility in Tsetse Flies (*Glossina morsitans* Westwood) Caused by Loss of Symbionts. *Experientia*. 32(8), pp. 995 - 996

Nogge, G. (1981). Significance of Symbionts for the Maintenance of an Optimal Nutritional State for Successful Reproduction in Hematophagous Arthropods. *Parasitology*. 81(4), pp. 101 – 104

Oakeson, K.F. *et al*. (2014). Genome Degeneration and Adaptation in a Nascent Stage of Symbiosis. *Genome Biol Evol*. 6(1), pp. 76 – 93

Okoth, J.O. and Kapaata, R. (1986). Trypanosome Infection Rates in *Glossina fuscipes fuscipes* Newst. in the Busoga Sleeping Sickness Focus, Uganda. *Annals of Tropical Medicine & Parasitology*. 80(4), pp. 459 - 461

O'Neill, S.L. *et al.* (1993). Phylogenetically Distant Symbiotic Microorganisms Reside in *Glossina* Midgut and Ovary Tissues. *Medical and Veterinary Entomology*. 7(4), pp. 377 - 383

Orth, J.D. *et al*. (2010). What is Flux Balance Analysis? *Nature Biotechnology*. 28, pp. 245 - 248

Ouedraogo, G.M.S. *et al.* (2018). Prevalence of Trypanosomes, Salivary Gland Hypertrophy Virus and *Wolbachia* in Wild Populations of Tsetse Flies from West Africa. *BMC Microbiology*. 18:153

Oren, A. (2005). A Hundred Years of *Dunaliella* Research: 1905 – 2005. *Saline Systems*. 1:2

Owen, S. *et al.* (2017). Characterization of the Prophage Repertoire of African *Salmonella* Typhimurium ST_{313} Reveals High Levels of Spontaneous Induction of Novel Phage BTP1. *Frontiers in Microbiology*. 8, 235

Pais, R. *et al.* (2008). The Obligate Mutualist *Wigglesworthia glossinidia* Influences Reproduction, Digestion, and Immunity Processes of its Host, the Tsetse Fly. *Appl Environ Microbiol.* 74(19), pp. 5965 - 5974

Pannebakker, B.A. *et al*. (2007). Parasitic Inhibition of Cell Death Facilitates Symbiosis. *PNAS*. 104(1), pp. 213 - 215

Parkhill, J. *et al*. (2001). Complete Genome Sequence of a Multiple Drug Resistant *Salmonella enterica* serovar Typhi CT18. *Nature*. 413, pp. 848 -852

Penheiter, K.L. *et al*. (1997). Non-Invasive *Salmonella typhimurium* Mutants are Avirulent Because of an Inability to Enter and Destroy M Cells of Ileal Peyer's Patches. *Molecular Microbiology*. 24(4), pp. 697 - 709

Pérez-Toledo, M. *et al*. (2017). *Salmonella* Typhi Porins OmpC and OmpF Are Potent Adjuvants for T-Dependent and T-Independent Antigens. *Front Immunol*. 8:230

Ponstingl, H. and Ning, Z. (2010). SMALT – A New Mapper for DNA Sequencing Reads. *F1000Posters*. 1:313 (poster).

Pontes, M.H. *et al*. (2008). Quorum Sensing Primes the Oxidative Stress Response in the Insect Endosymbiont, *Sodalis glossinidius. PLoS ONE.* 3(10):e3541

Pontes, M.H. *et al.* (2011). Attenuation of the Sensing Capabilities of PhoQ in Transition to Obligate Insect–Bacterial Association. *PLoS Genetics.* 7(11):e1002349

Pontes, M.H. and Dale, C. (2011). Lambda Red-Mediated Genetic Modification of the Insect Endosymbiont *Sodalis glossinidius. Applied and Environmental Microbiology.* 77(5), pp. 1918 – 1920

R Core Team (2013). R: A language and Environment for Statistical Somputing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org

Ratledge, C. and Dover, L.G. (2000). Iron Metabolism in Pathogenic Bacteria. *Annual Review of Microbiology*. 54, pp. 881 - 941

Reiling, H.E. *et al.* (1985). Mass culture of *Escherichia coli*: Medium Development for Low and High Density Cultivation of *Escherichia coli* B/r in Minimal and Complex Media. *Journal of Biotechnology*. 2(3–4), pp 191 - 206

Reinhardt *et al*. (1972). Ultrastructural Study of the Midgut Mycetome Bacteroides of the Tsetse fly *Glossina morsitans*, *G. fuscipes* and *G. brevipalpis* (Diptera: Brachycera). *Acta Trop*. 29, pp. 280 - 288

Reznikoff, W.S. (2003). Tn5 as a Model for Understanding DNA Transposition. *Molecular Microbiology*. 47(5), pp. 1199 – 1206

Ridgley, E.L. *et al*. (1999). Reactive Oxygen Species Activate a Ca^{2+}-Dependent Cell Death Pathway in the Unicellular Organism *Trypanosoma brucei brucei*. *Biochem J.* 340, pp. 33 – 40

Rogers, D. (1977). Study of a Natural Population of *Glossina fuscipes fuscipes* Newstead and a Model of Fly Movement. *Journal of Animal Ecology.* 46, pp 309 – 330

Sameshima, S. *et al*. (1999). Phylogenetic Comparison of Endosymbionts with Their Host Ants Based on Molecular Evidence. *Zoological Science*. 16(6), pp. 993 - 1000

Scherer, W. *et al*. (1953). Studies on the Propagation *in vitro* of Poliomyelitis Viruses. *J Exp Med*. 97(5), pp. 695 – 710

Schlein, Y. (1976). Lethal Effect of Tetracycline on Tsetse Flies Following Damage to Bacteroid Symbionts. *Experientia*. 33(4), pp. 450 - 451

Schneider, I. (1972). Cell Lines Derived from Late Embryonic Stages of *Drosophila melanogaster. J Embryol Exp Morphol.* 27(2), pp. 353 - 365

Schneider, D.I. *et al*. (2013). Global *Wolbachia* Prevalence, Titer Fluctuations and their Potential of Causing Cytoplasmic Incompatibilities in Tsetse Flies and Hybrids of *Glossina morsitans* Subgroup Species. *Journal of Invertebrate Pathology*. 112(S1), pp. S104 - S115

Seemann, T. (2014). Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics*. 30(14), pp. 2068 - 2069

Sharma, R. *et al*. (2008). Asymmetric Cell Division as a Route to Reduction in Cell Length and Change in Cell Morphology in Trypanosomes. *Protist*. 159(1), pp. 137 – 151

Shaw, M.K. and Moloo, S.K. (1991). Comparative Study on Rickettsia-Like Organisms in the Midgut Epithelial Cells of Different *Glossina* Species. *Parasitology*. 102(2), pp. 193 - 199

Simarro, P.P. *et al*. (2010). The Atlas of Human African Trypanosomiasis: A Contribution to Global Mapping of Neglected Tropical Diseases. *International Journal of Health Geographics*. 9:57

Simarro, P.P. *et al*. (2012). Estimating and Mapping the Population at Risk of Sleeping Sickness. *PLoS Neglected Tropical Diseases*. 6(10): e1859

Singh, K.R.P. (1967). Cell Cultures Derived from Larvae of *Aedes albopictus* (Skuse) and *Aedes aegypti* (L.). *Current Science*. 36(19), pp. 506 - 508

Southwood, T.R.E. *et al*. (1975). The Micro-Organisms of Tsetse Flies. *Acta Trop*. 32, pp. 259 - 266

Stiles, J.K. *et al*. (1990). Identification of Midgut Trypanolysin and Trypanoagglutinin in *Glossina palpalis* spp. (Diptera: Glossinidae). *Parasitology*. 101(3), pp. 369 - 376

Summers, R.J. *et al*. (1979). Continuous Cultivation for Apparent Optimization of Defined Media for *Cellulomonas* sp. and *Bacillus cereus. Applied and Environmental Microbiology*. 38(1), pp. 66 - 71

Tatusov, R.L. *et al*. (1997). A Genomic Perspective on Protein Families. *Science*. 278(5338), pp. 631 - 637

Toh, H. *et al*. (2006). Massive Genome Erosion and Functional Adaptations Provide Insights into the Symbiotic Lifestyle of *Sodalis glossinidius* in the Tsetse Host. *Genome Res*. 16, pp. 149 - 156

Trappeniers, K. *et al*. (2019). The Tsetse Fly Displays an Attenuated Immune Response to Its Secondary Symbiont, *Sodalis glossinidius. Front. Microbiol.* 10:1650

Tyx, R.E. *et al*. (2011). Role of Dihydrolipoamide Dehydrogenase in Regulation of Raffinose Transport in *Streptococcus pneumoniae*. *J Bacteriol*. 193(14), pp. 3512 - 3524

Van den Abbeele, J. *et al*. (1999). *Trypanosoma brucei* spp. Development in the Tsetse Fly: Characterization of the Post-Mesocyclic Stages in the Foregut and Proboscis. *Parasitology.* 118(Pt 5), pp. 469 – 478

von Wissmann, B. *et al.* (2011). Factors Associated with Acquisition of Human Infective and Animal Infective Trypanosome Infections in Domestic Livestock in Western Kenya. *PLoS Neglected Tropical Diseases*. 5:1, e941

Vreysen, M.J.B. *et al*. (2000). *Glossina austeni* (Diptera: Glossinidae) Eradicated on The Island of Unguja, Zanzibar, Using the Sterile Insect Technique. *J. Econ. Entomol.* 93(1), pp. 123 - 135

Wamwiri, F.N. *et al*. (2013). *Wolbachia, Sodalis* and Trypanosome Co-Infections in Natural Populations of *Glossina austeni* and *Glossina pallidipes. Parasites & Vectors.* 6:232

Wang, J. *et al.* (2013). Tsetse Fly Microbiota: Form and Function. *Front. Cell. Infect. Microbiol.* 3:69

Weinreich, M.D. *et al*. (1994). Evidence that the *cis* Preference of the Tn5 Transposase is Caused by Nonproductive Multimerization. *Genes Dev*. 8(19), pp. 2363 - 2374

Welburn, S.C. *et al.* (1987). *In vitro* Cultivation of Rickettsia-Like-Organisms from *Glossina* spp.. *Annals of Tropical Medicine & Parasitology*. 81(3), pp. 331 - 335

Welburn, S.C. and Maudlin, I. (1990). Rickettsia-Like Organisms, Puaparial Temperature and Susceptibility to Trypanosome Infection in *Glossina morsitans*. *Parasitology*. 102(2), pp. 201 - 206

Welburn, S.C. and Maudlin, I. (1992). The Nature of the Teneral State in *Glossina* and its Role in the Acquisition of Trypanosome Infection in Tsetse. *Annals of Tropical Medicine & Parasitology*. 86(5), pp. 529 - 536

Welburn, S.C. *et al.* (1993). Rickettsia-Like Organisms and Chitinase Production in Relation to Transmission of Trypanosomes by Tsetse Flies. *Parasitology*. 107, pp. 141 - 145

Widderich, N. *et al*. (2014). Biochemical Properties of Ectoine Hydroxylases from Extremophiles and Their Wider Taxonomic Distribution Among Microorganisms. *PLoS ONE.* 9(4), e93809

Wigglesworth, V.B. (1929). Digestion in the Tsetse-Fly: A Study of Structure and Function. *Parasitology*. 21(3), pp. 288 - 321

Wissenbach, U. *et al*. (1995). A Third Periplasmic Transport System for L-Arginine in *Escherichia coli*: Molecular Characterization of the *artPIQMJ* Genes, Arginine Binding and Transport. *Molecular Microbiology*. 17(4), pp. 675 - 686

World Health Organisation (WHO) (1998). Control and Surveillance of African Trypanosomiasis. Report of a WHO Expert Committee. WHO Technical Report Series 881, World Health Organization, Geneva

World Health Organisation (WHO) (2019)a. Trypanosomiasis, Human African (Sleeping Sickness) article, updated 12/04/2019. https://www.who.int/news-room/fact-sheets/detail/trypanosomiasis-human-african-(sleeping-sickness)

World Health Organisation (WHO) (2019)b. WHO Interim Guidelines for the Treatment of Gambiense Human African Trypanosomiasis. WHO Interim Guidelines. Geneva

World Health Organisation (WHO) (2019)c. The Parasite article, accessed in 2019 at https://www.who.int/trypanosomiasis_african/disease/parasite/en/

Xiao-Ran *et al*. (2018). Chapter Eleven – *Halomonas* and Pathway Engineering for Bioplastics Production. *Methods in Enzymology*. 608, pp. 309 – 328

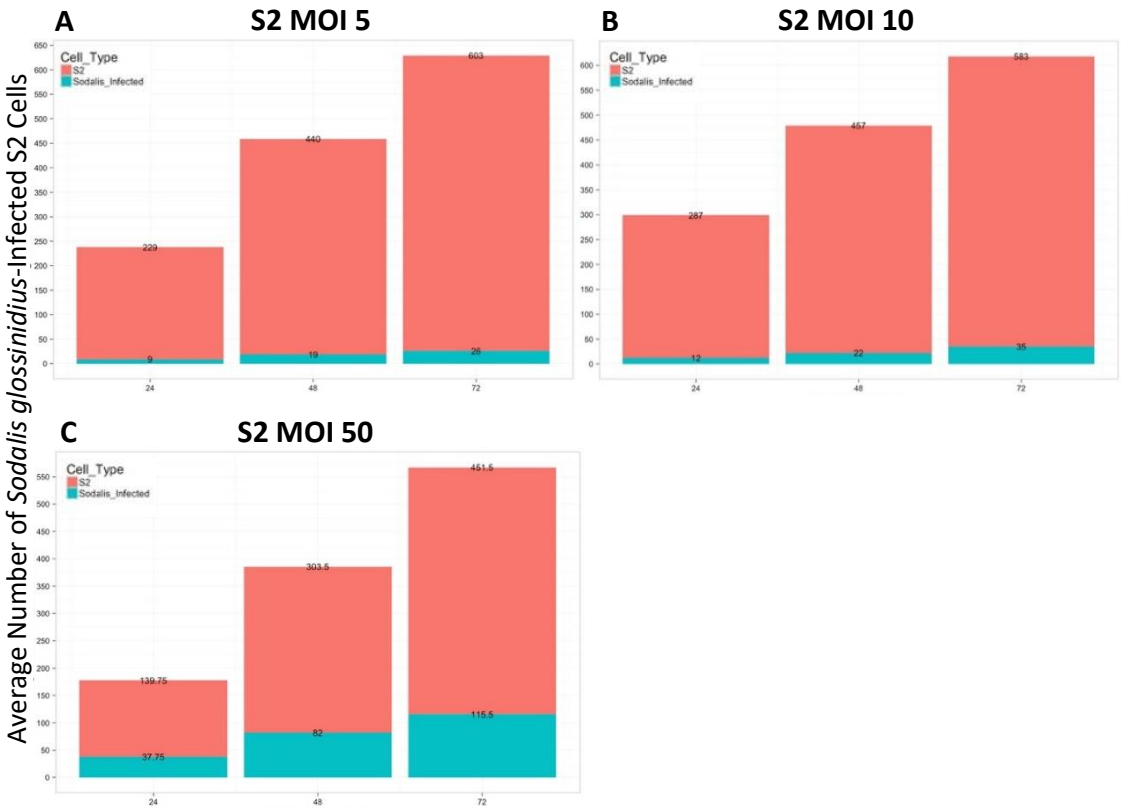
Zhang, S. and Meyer, R. (1997). The Relaxosome Protein MobC Promotes Conjugal Plasmid Mobilization by Extending DNA Strand Separation to the Nick Site at the Origin of Transfer. *Mol Microbiol*. 25(3), pp. 509 - 516

Zhou, M. and Reznikoff, W.S. (1997). Tn5 Transposase Mutants That Alter DNA Binding Specificity. *J Mol Biol*. 271(3), pp. 362 - 373

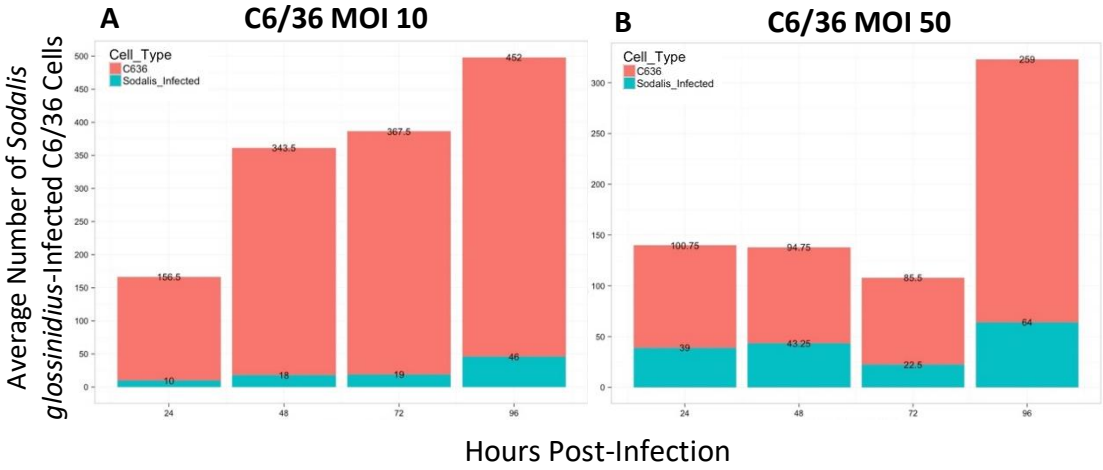
Zhou, M. *et al.* (1998). Molecular Genetic Analysis of Transposase-End DNA Sequence Recognition: Cooperativity of Three Adjacent Base-Pairs in Specific Interaction with a Mutant Tn5 Transposase. *J Mol Biol*. 276(5), pp. 913 - 925

Zhou, D. *et al*. (1999). *Salmonella typhimurium* Encodes a Putative Iron Transport System within the Centisome 63 Pathogenicity Island. *Infect Immun*. 67(4), pp. 1974 - 1981

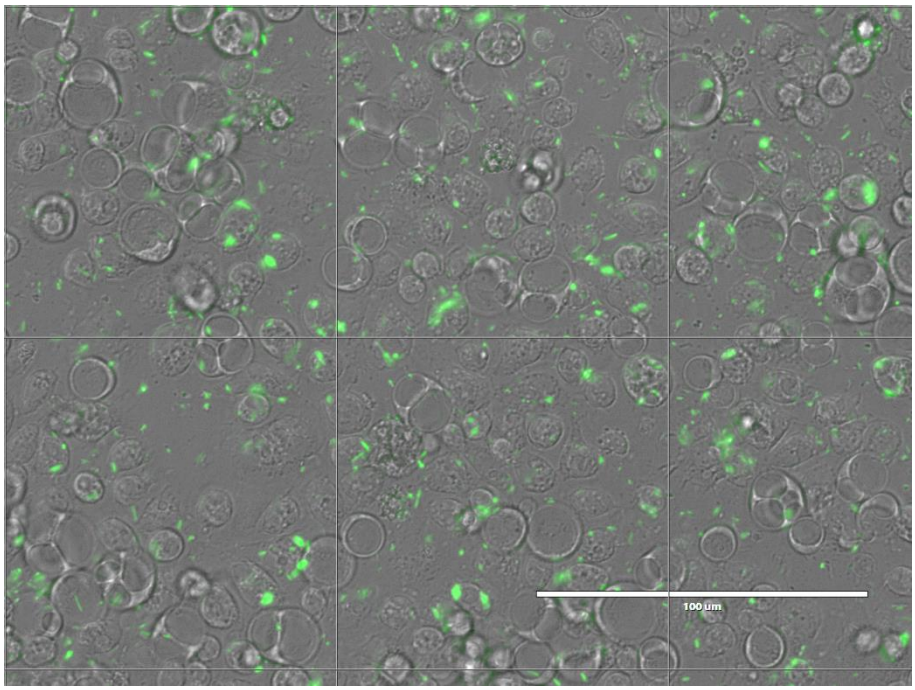
Appendix



Appendix Figure 1. The proportions of average GFP *Sodalis glossinidius*-infected S2 cells out of average total S2 cells, over 24-hour time increments. A: infected S2 cells at MOI 5. 24 hours *N* = 1. 48 hours *N* = 1. 72 hours *N* = 1.**B:** infected S2 cells at MOI 10. 24 hours *N* = 1. 48 hours *N* = 1. 72 hours *N* = 1. **C:** infected S2 cells at MOI 50. 24 hours *N* = 4. 48 hours *N* = 4. 72 hours *N* = 4. The [average] numbers used here are from the microscope images (i.e. not corrected for the whole well).

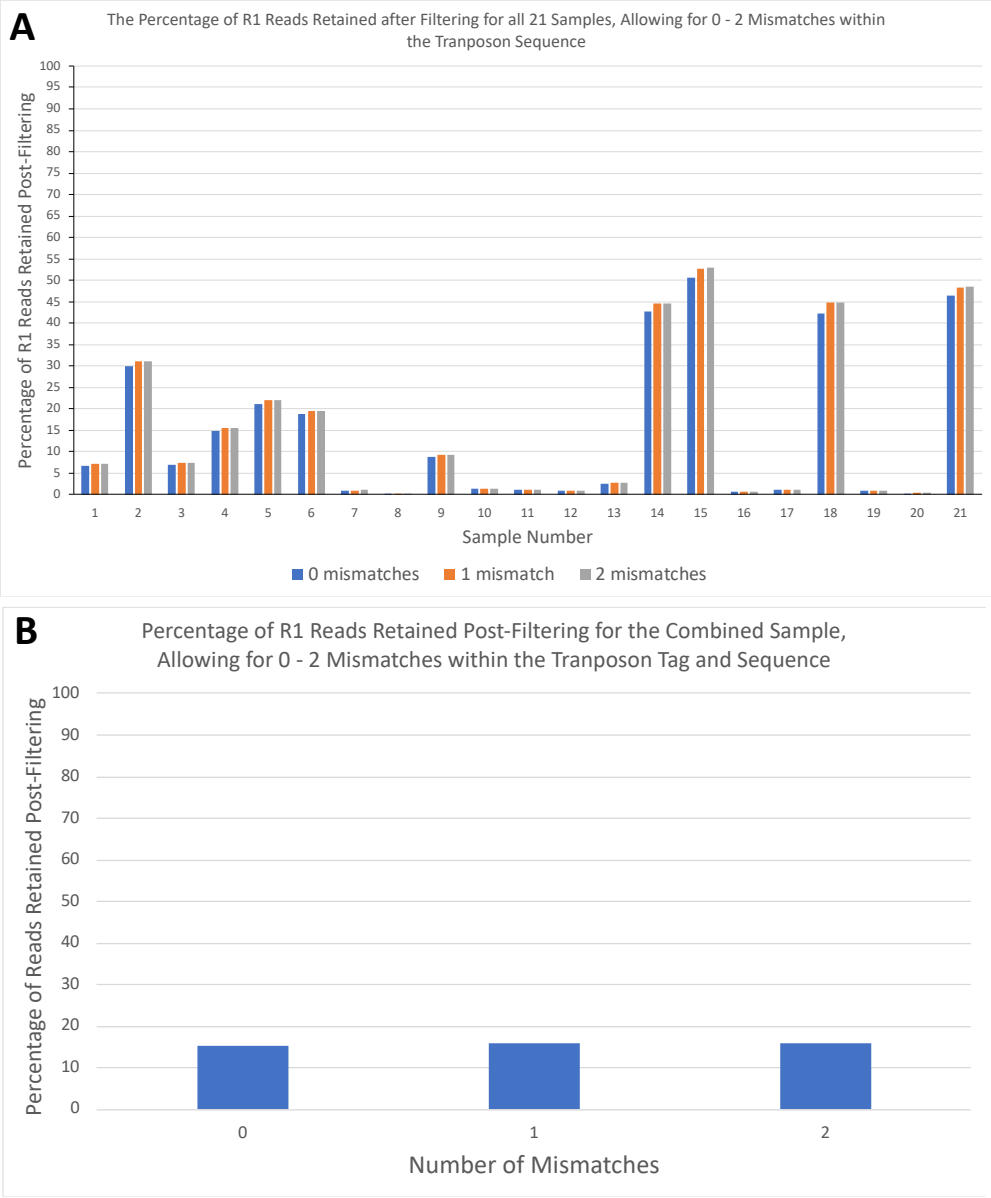


Appendix Figure 2. The proportions of average GFP *Sodalis glossinidius*-infected C6/36 cells out of average total C6/36 cells, over 24-hour time increments. A: infected C6/36 cells at MOI 10. 24 hours *N* = 2. 48 hours *N* = 2. 72 hours *N* = 2. 96 hours *N* = 1. **B:** infected C6/36 cells at MOI 50. 24 hours *N* = 4. 48 hours *N* = 4. 72 hours *N* = 4. 96 hours *N* = 2. The [average] numbers used here are from the microscope images (i.e. not corrected for the whole well).

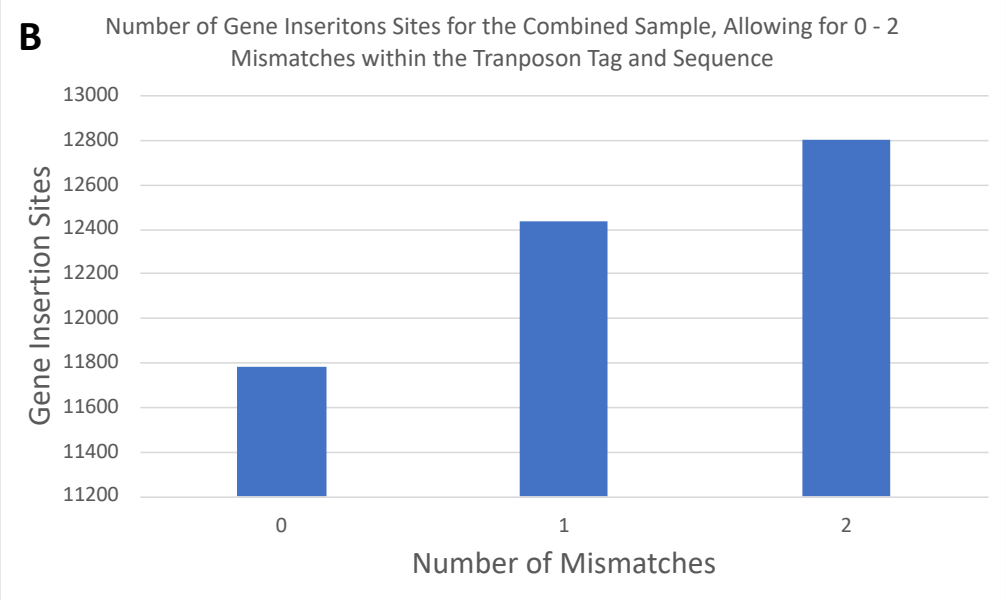


Appendix Figure 3. Microscope image of multiple GFP-*Sodalis glossinidius* infections, per S2 cell. Schneider's Insect Medium + 10% foetal bovine serum (FBS) at an MOI of 200, 24 hours post-infection.

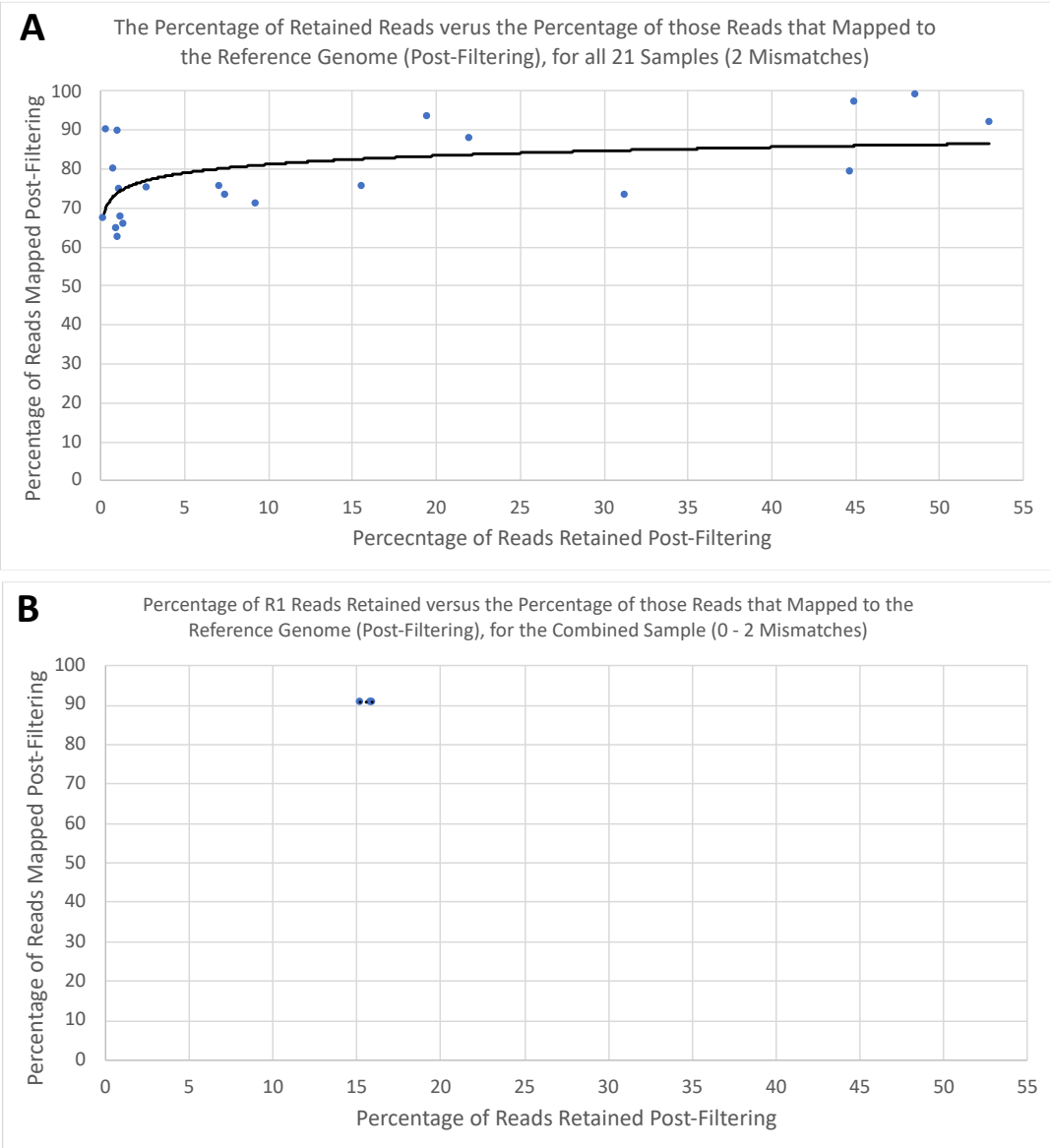
Appendix Figure 4. Template strand sequence of pRL27 [4080 bp]. In green: tnP (transposase) [1431 bp]. In red: IRs (Tn5 inverted repeats) [19 bp each]. In peach: oriVR6K (Lamda pir dependent origin of replication) [440 bp]. In blue: KmR (Km resistance gene) [816 bp]. In purple: traJ (conjugal transfer gene expression regulator) [234 bp]. In orange: oriTRP4 (origin of transfer, starting at G) [481 bp].



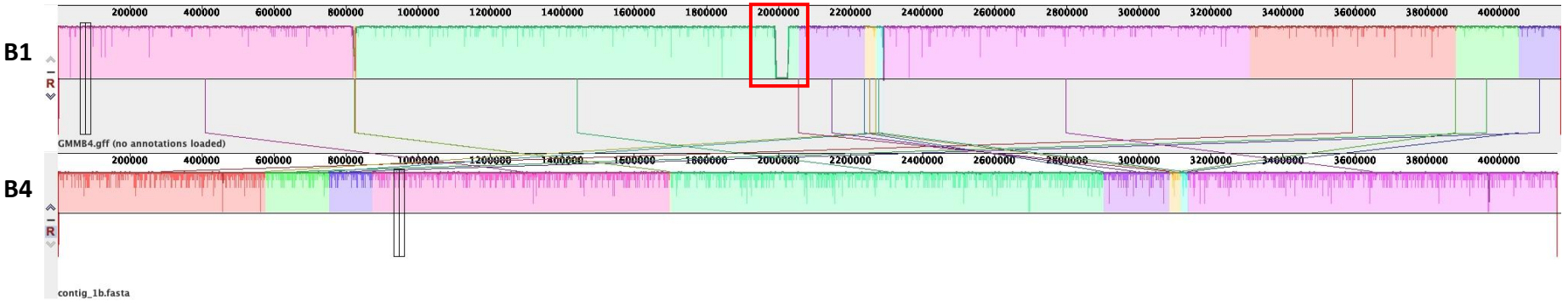
Appendix Figure 5. Graphical representation(s) of the percentage of retained reads, after filtering for the transposon tag. A: all 21 samples individually shown. Blue: 0 mismatches in the transposon tag. Orange: 1 mismatch in the transposon tag. Grey: 2 mismatches in the transposon tag. **B:** the combined sample [all 21 samples, in one].



Appendix Figure 6. Graphical representation(s) of the numbers of gene insertions sites. A: all 21 samples individually shown. Blue: 0 mismatches in the transposon tag. Orange: 1 mismatch in the transposon tag. Grey: 2 mismatches in the transposon tag. **B:** the combined sample [all 21 samples, in one].



Appendix Figure 7. Graphical representation(s) of the percentage of retained reads versus mapped to the reference genome, after filtering for the transposon tag. A: all 21 samples individually shown. 2 mismatches were allowed in the transposon tag. **B:** the combined sample [all 21 samples, in one], with the separate points showing a 0 – 2 mismatch range.



Appendix Figure 8. Mauve multiple sequence alignment of *Sodalis glossinidius* strain SgGMMB1 (top, "B1") to *S. glossinidius* strain SgGMMB4 (bottom, "B4"). Sequence synteny is represented by solid colour blocks. The red box outlines the region missing in strain SgGMMB1, compared to strain SgGMMB4. **Darling *et al*. (2004).**

Appendix Table 1. The essential gene candidates and their associated Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway hits unique to the *Sodalis glossinidius* long-term establishment in a complex medium (5.3.3.1.) selection. Kanehisa and Goto (2000); Kanehisa *et al*. (2016), (2017).

KEGG Pathway Hit	**Gene Name(s)**
Glycolysis/gluconeogenesis (sgl00010)	*pgi gpmA acs fbp*
Pyruvate metabolism (sgl00620)	*mqo acs*
Pentose phosphate pathway (sgl00030)	*pgi fbp*
Fructose and mannose metabolism (sgl00051)	*fucA rhaD fbp*
Starch and sucrose metabolism (sgl00500)	*pgi*
Citrate cycle (TCA cycle) (sgl00020)	*mqo sucA*
Glyoxylate and dicarboxylate metabolism (sgl00630)	*acs*
Oxidative phosphorylation (sgl00190)	*nuoM*
Glycine, serine and threonine metabolism (sgl00260)	*dat thrC gpmA*
Alanine, aspartate and glutamate metabolism (sgl00250)	*gabD*
Tyrosine metabolism (sgl00350)	*gabD*
Cysteine and methionine metabolism (sgl00270)	*metC*
Cysteine ABC transporter (sgl02010)	*tcyB*
Vitamin B6 metabolism (sgl00750)	*thrC*
Two-component systems (sgl02020)	*ompC ompF*
Flagellar assembly (sgl02040)	*flgI flgL_2 flhB_1 fliI_1 fliM_2 fliR_2*
Cationic antimicrobial peptide (CAMP) resistance (sgl01503)	*acrA arnT_3 acrB_2 sapB*
Bacterial secretion system (sgl03070)	*yscN yscU*

Appendix Table 2. The essential gene candidates and their associated Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway hits unique to the complex medium pool (5.3.3.2.) selection. Kanehisa and Goto (2000); Kanehisa *et al*. (2016), (2017).

KEGG Pathway Hit	Gene Name(s)
Glycolysis/gluconeogenesis (sgl00010)	*aceE aceF pykF pykA pgk pckA yeaD pgi pgm*
Pyruvate metabolism (sgl00620)	*mqo aceE aceF pflB pykF pykA gloB pckA pta*
Oxidative phosphorylation (sgl00190)	*nuoF nuoH nuoL nuoN atpB atpA atpG cyoB*
Citrate cycle (TCA cycle) (sgl00020)	*icd mqo aceE sucA aceF pckA sucC*
Glyoxylate and dicarboxylate metabolism (sgl00630)	*gcvT glnA*
Pantothenate and CoA biosynthesis (sgl00770)	*ilvC ilvE coaE mazG acpH*
Arginine biosynthesis (sgl00220)	*gdhA argI aspC argB glnA*
Pentose phosphate pathway (sgl00030)	*prs rpe pgi pgm*
Fructose and mannose metabolism (sgl00051)	*mtlD algC*
Galactose metabolism (sgl00052)	*galE_1 pgm kbaZ gatZ_3*
Starch and sucrose metabolism (sgl00500)	*pgi pgm*
Cysteine and methionine metabolism (sgl00270)	*serA asd metE_1 hhaIM csyE metK ilvE serC aspC metC sdaA gshA dcyD mtnB*
Alanine, aspartate and glutamate metabolism (sgl00250)	*gdhA pyrB aspC glmS aspA_2 purB glnA asnB carB*
Glycine, serine and threonine metabolism (sgl00260)	*serA asd gcvT serC sdaA ilvA*
Histidine metabolism (sgl00340)	*hisD hisG hisF*
Arginine and proline metabolism (sgl00330)	*proA_1 proC aspC*
Tyrosine metabolism (sgl00350)	*aspC*
Vitamin B6 metabolism (sgl00750)	*serC pdxB*
Phenylalanine, tyrosine and tryptophan biosynthesis (sgl00400)	*aroE_3 trpGD_2 aroA aspC aroC*
Lysine biosynthesis (sgl00300)	*asd dapD dapE dapA_2 dapA_4 dapF*
Valine, leucine and isoleucine biosynthesis (sgl00290)	*leuB ilvC ilvE leuC leuB leuD1 ilvA*
Valine, leucine and isoleucine degradation (sgl00280)	*ilvE*
Arginine ABC transporter (sgl02010)	*artI_1*
Glutamate/aspartate ABC transporter (sgl02010)	*glnQ_4*
Cystine ABC transporter (sgl02010)	*fliY*
D-methionine ABC transporter (sgl02010)	*metN*
Oligopeptide ABC transporter (sgl02010)	*oppB_1 oppC oppD_1 oppD_2*
Quorum sensing (sgl02024)	*oppB_1 oppC oppD_1 oppD_2*
Thiamine ABC transporter (sgl02010)	*thiB thiQ*
Thiamine metabolism (sgl00730)	*dxs thiH_2 thiI*
Two-component systems (sgl02020)	*ntrC glnA glnQ_4 wecB crp phoP cheB*
Cationic antimicrobial peptide (CAMP) resistance (sgl01503)	*amiB ppiA cpxA phoP arnA acrB_1*
Sulfur metabolism (sgl00920)	*cysD cysE sbp_2*

Appendix Table 3. The essential gene candidates and their associated Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway hits unique to the minimal medium pool (5.3.4.) selection. Kanehisa and Goto (2000); Kanehisa *et al*. (2016), (2017).

KEGG Pathway Hit	**Gene Name(s)**
Glycolysis/gluconeogenesis (sgl00010)	*gatY_2 tpiA acs fbp*
Pyruvate metabolism (sgl00620)	*acs*
Pentose phosphate pathway (sgl00030)	*tal gatY_2 fbp*
Oxidative phosphorylation (sgl00190)	*nuoI*
Glyoxylate and dicarboxylate metabolism (sgl00630)	*acs*
Pantothenate and CoA biosynthesis (sgl00770)	*ilvG ilvI_2*
Arginine biosynthesis (sgl00220)	*argE alaA*
Cationic antimicrobial peptide (CAMP) resistance (sgl01503)	*arcA arnT_3 phoQ acrB_2 sapB ddpD*
Flagellar assembly (sgl02040)	*flgK_3 flhB_2 flhC_2 fliM_2*
Quorum sensing (sgl02024)	*flhC_2*
Two-component systems (sgl02020)	*flhC_2 ompF phoQ wbpA*
Phosphotransferase system (PTS) (sgl02060)	*manZ_2*
Sulfur metabolism (sgl00920)	*cysI cysH metA cysK cysW_3 cysP*