University of Salford MANCHESTER

Characterising the Bacterial Microbiota of *Phlebotomus* and *Glossina* insects: Vectors of Parasitic Diseases

A thesis submitted for the degree of
Doctor of Philosophy
April 2021

Poppy Pescod

School of Science, Engineering and the Environment
University of Salford

So, Nat'ralists observe, a Flea

Hath smaller Fleas that on him prey,

And these have smaller yet to bite 'em;

And so proceed ad infinitum.

- Jonathan Swift (1733)

Contents

List of figures

List of tables

Acknowledgements

The support I've had from everyone around me made this thesis possible. I'd like to thank my supervisors Ian and Chloe first and foremost for being my mentors, friends, cheerleaders and role models – I owe you both so many beers! I can't thank Rachael Antwis enough for guiding me through the murky world of microbiomes, and for being so patient and generous. Thanks to Judy Mwangi, who made the fieldwork in Uganda possible, worked so hard and kept me going! Every single person I met at Salford made being there a great experience: the 212 dream team Alex and Terence; Marion and the teaching technicians; Sarah, Joe, Richard and all the other lecturers and admin staff.

This work is built on the back of some amazing collaborations with other institutions. Thanks to Alistair Darby and Lauren Gordon at the University of Liverpool, and Lee Haines and Alvaro Acosta-Serrano at the LSTM, for always letting me tap their wealth of *Sodalis* knowledge and for their generous donations of cultures, resources and time. Thanks also to Geraldine Foster, Gala Garrod and Mike Coleman at the LSTM, and the Rajendra Memorial Research Institute, for all of their time and help facilitating the sand fly work. Thanks to everyone at the CGR for their hard work on the *Sodalis* transcription data. Finally, thank you to Patrick Vudriko and Deborah Amulen at Makerere University, and Patrick Abila at NaLIRRI, for being so open to working with us and for making me feel so welcome in Uganda.

A lot of the coherency and clarity in the final version of this thesis is down to my examiners Ben Makepeace and Richard Birtles; thanks for giving me such a good grilling in the viva, your comments and ideas have made every chapter something I feel really proud of. I also had a lot of fun!

To my parents and my siblings, I'm so grateful for your love and support over the last few years – bet you're glad I didn't make you read it though! Thanks to my parents-in-law for always supporting us and always letting me cuddle their animals. To my amazing friends, the Connaught Road crew, thank you for managing to make me laugh while I was writing up and for all of the animal sitting you've done for me over the years!

Lastly, to the person who has stuck by my side through eight years of university, work, unemployment, multiple house moves, several ill-advised animal adoptions, planning a wedding mid-PhD, AND a global pandemic – my husband and my best friend, Johnny. Sorry you've had to do so much washing up for the last few years.

Candidate Declaration

All the work presented in this thesis is my own and was performed by me unless otherwise stated. It involved collaborations with multiple colleagues:

Chapter 2: Judy Mwangi (JM), Patrick Vudriko and Patrick Abila assisted with trapping wild tsetse, and JM assisted with *Sodalis* field culturing. Lee Haines (LH) allowed the use and adaptation of her *Sodalis* isolation protocol.

Chapter 3: LH, under the supervision of Alvaro Acosta-Serrano (AAS), maintained a *Sodalis* culture for 10 years to produce the *Sg*GMMC1* strain. AAS and LH provided us with the strain and allowed the use of their lab space and equipment for culturing at the LSTM. Alistair Darby and Lauren Gordon provided the ancestral *Sg*GMMB4 strain and allowed use of their lab space and equipment for culturing at the University of Liverpool.

Chapter 4: AAS and LH facilitated access to the LSTM tsetse fly colony; LH provided trypanosome stabilites and taught me how to infect tsetse with trypanosomes, how to identify a successful trypanosome infection, and how to dissect tsetse for tissue collection and *Sodalis* isolation. RNA sequencing was performed by Margaret Hughes and Charlotte Nelson at the Centre for Genomic Research; Sam Haldenby and Liu Xuan at the CGR processed the transcription read data.

Chapters 5 and 6: Geraldine Foster (GF) and Mike Coleman coordinated the collection and transport of sand fly samples with the Rajendra Memorial Research Institute in India. GF and Gala Garrod assisted with DNA extraction of sand fly samples and allowed the use of their lab space at LSTM. Rachael Antwis allowed the use of her sequencing primers, protocol and assisted with the microbiome analysis.

Abbreviations

Abbreviation	Description
AAT	Animal African trypanosomiasis
abs logFC	Absolute logarithmic fold change
ASV	Amplicon sequence variant
bp	Base pair
CDS	Coding DNA sequence
CI	Cytoplasmic incompatibility
CL	Cutaneous leishmaniasis
COG	Conserved orthologous group
CPM	Counts per million
DALY	Disability-adjusted life year
DDT	Dichlorodiphenyltrichloroethane
DE	Differential expression
dsRNA	Double-stranded RNA
FDR	False discovery rate
GlcNAc	N-acetylglucosamine
Gmm	*Glossina morsitans morsitans*
GO	Gene ontology
Gpal	*Glossina palidipes*
HAT	Human African trypanosomiasis
HGT	Horizontal gene transfer
indel	insertion/deletion
IVM	Integrated vector management
kb	kilobases (1,000s bases)
logFC	logarithmic fold change
LSTM	Liverpool School of Tropical Medicine
Mbp	Million base pairs
MDS	Multi-dimensional scaling
MGE	Mobile genetic elements
MLST	Multi-locus sequence typing
MMI	Mitsuhashi and Maramorosch insect growth medium
MU	Makerere University
NALIRRI	National Livestock Resources Research Institute
NGS	Next-generation sequencing
NMDS	Non-metric multi-dimensional scaling
nt	Nucleotide
NTD	Neglected tropical disease
OTU	Operational taxonomic unit
PBS	Phosphate-buffer solution

PERMANOVA	Permutational multivariate analysis of variance
PHC	Primary Health Centre
PKDL	Post-kala-azar dermal leishmaniasis
RT	Room temperature
SBS	Sequencing by synthesis
SMRT	Single molecule real time
SNP	Single nucleotide polymorphism
SPRI	Solid phase reversible immobilisation
TTSS	Type III secretion system
VBD	Vector-borne disease
VL	Visceral leishmaniasis
WGS	Whole genome sequencing
WHO	World Health Organization

Abstract

As the technology used to examine bacteria using culture-independent methods has advanced, so too has our knowledge of the bacterial microbiome of insect vectors and its effect on the pathogens they carry. Trypanosomiasis and leishmaniasis are neglected tropical diseases caused by kinetoplast parasites and transmitted between human hosts by biting flies. In this thesis the bacterial microbiota of tsetse flies and sand flies, the vectors of trypanosomiasis and leishmaniasis respectively, were examined using molecular biology and next generation sequencing to understand their distribution in wild populations and their impact on pathogen transmission.

An experimentally evolved isolate of the tsetse secondary endosymbiont *Sodalis glossinidius* (*Sg*GMMC1*), which is believed to increase tsetse susceptibility to trypanosome infection, was sequenced and compared to an ancestor *S. glossinidius* strain (*Sg*GMMB4). *Sg*GMMC1* contained a single large 17 kb deletion as well as 57 small indels and SNPs resulting in 12 putatively pseudogenised genes, supporting the theory that the *S. glossinidius* genome is undergoing evolutionary reduction.

The distribution of known tsetse symbionts including *Sodalis* and *Spiroplasma,* as well as the trypanosome parasite, was explored in wild Ugandan tsetse populations. *Sodalis* was detected in 39% of wild-caught tsetse, *Wolbachia* in 1.7%, *Spiroplasma* in 88% and *Trypanosoma* in 12%, providing essential information about the ecology of these symbiotic organisms in wild populations.

Lab-reared tsetse flies were experimentally infected with trypanosomes to examine metatranscriptional differences of commensal bacteria between trypanosome-refractory and successfully infected flies. *Wigglesworthia* expression differed over time but not between refractory and infected tsetse; in *Sodalis* virulence genes were upregulated in trypanosome infected flies compared to refractory flies; and little effect of trypanosome infection on *Serratia* or *Wolbachia* transcription was observed. These results suggest that increased *Sodalis* virulence may increase tsetse susceptibility to trypanosome infection.

Finally, a novel 16S sequencing pipeline for low-biomass individual sand flies was designed and implemented on wild *Phlebotomus argentipes*, the vector of visceral leishmaniasis in the Bihar region of India, to discover variation in individual sand fly microbiota within and between houses, villages, and larger districts. A pipeline was optimised to reduce the effect of contamination in low-biomass samples, allowing the description of a core microbiome of seven genera including *Corynebacterium*, *Acinetobacter,* and *Staphylococcus*. A variable peripheral microbiome was discovered with no effect of geographical location, and included the important human pathogen *Orientia,* implying a potential role for *P. argentipes* in *Orientia* transmission.

Chapter 1 Introduction

1.1 Abstract

Neglected tropical diseases are major causes of continuing poverty for billions of people around the world. Two of these diseases, trypanosomiasis and leishmaniasis, are parasitic illnesses spread by the bite of infected vectors, tsetse flies and sand flies respectively. Modern genetic sequencing technology has rapidly advanced the study of vector-borne diseases, and in particular the study of the vector's bacterial microbiota and their potential uses in control strategies. Many disease vectors harbour symbiotic bacteria, ranging from obligate mutualists to parasites, which can affect vector competence. Due to their close association with the vector these symbiotic bacteria have the potential to be used in disease control strategies, either through exploitation of their natural effects or by genetic modification. The microbiomes of tsetse flies and sand flies are differently structured: in addition to a recently discovered peripheral microbiome tsetse flies have an obligate endosymbiont *Wigglesworthia*, which is present in all members of the genus, as well as secondary endosymbiont *Sodalis* and parasitic symbiont *Wolbachia.* Sand flies have no known mutualistic bacteria, instead containing a diverse microbiome heavily influenced by their diet and environment. Further study of the role of bacterial microbiota in both vectors has been identified as an important research avenue for control of their respective vectored diseases.

1.2 Vector-borne neglected tropical diseases

1.2.1 Neglected tropical diseases

Neglected tropical diseases (NTDs) are a group of twenty communicable diseases that affect some of the world's most impoverished and vulnerable populations, primarily in tropical areas. One in three people is at risk of contracting an NTD, with one in six being already infected by at least one NTD [6]. The loss of income associated with contracting an NTD, combined with the immense social stigma, contribute significantly to the continued poverty of affected populations. All NTDs as defined by the World Health Organization (WHO) would be immediately amenable to broad control, elimination (no longer a public health concern) or eradication (extinction of the pathogen in the wild) through the improvement of living conditions, novel treatments and control strategies [7]. A current list of NTDs defined by WHO can be found in Table 1-1.

Table 1-1 - Neglected tropical diseases classified by WHO as of 2020 [8]. * - vector-borne diseases.

Neglected tropical diseases (2020)		
Buruli ulcer	Human African trypanosomiasis*	Soil-transmitted helminthiasis
Chagas disease*	Leishmaniasis*	Taeniasis/cystercosis
Dengue and chikungunya fever*	Chromoblastomycosis and other deep mycoses	Endemic treponematoses (Yaws)
Dracunculiasis	Lymphatic filariasis*	Leprosy
Echinococcosis	Onchocerciasis*	Scabies
Trachoma	Rabies	Snakebite envenoming
Foodborne trematodiases	Schistosomiasis*	

1.2.2 Introduction to vector-borne diseases

Many infectious diseases, including seven of the twenty NTDs listed in Table 1-1, are transmitted between hosts by vectors. The term “vector” is defined as an organism capable of carrying and transmitting pathogens to hosts or to the environment [9]; the pathogens that vectors carry between hosts cause vector borne diseases (VBDs). VBDs account for more than 17% of infectious diseases, and cause more than 700,000 deaths every year [10]. There are four main taxa of VBD pathogens: filaria, protozoa, bacteria and viruses [11].

All of the vectors in Table 1-2 are haematophagous (blood-feeding) arthropods; all except *Rhodnius* (Order: Hemiptera) are flies (Order: Diptera). Blood-feeding is common to many vectors as pathogens are easily transmitted during a blood meal, either by direct injection of the pathogen or by creating a bite wound through which excreted pathogens can enter the host bloodstream. Flying insects are well adapted to spread pathogens: flight allows the wide dispersal of insect populations and rapid invasion of novel environments, and most insects are resistant to desiccation in harsh climates [12]. This allows for rapid and widespread transmission of pathogens between hosts by haematophagous Dipterans.

As insects and pathogens both have temperature and humidity ranges that determine suitable habitats and seasons for VBD transmission, climate change is extending the ranges of both factors outside of areas and seasons historically known for VBD transmission. VBDs are therefore an emerging, growing global health concern [13]. VBDs are associated with warm, tropical areas, where the insect life

Table 1-2 - Ten of the most important VBDs to infect humans, and details on their morbidity and mortality as of 2019, vector species, and causative agents. DALY – Disability-adjusted life year; Asterisks (*) denote diseases classified as neglected tropical diseases [14].

Disease	Causative agent	Vector	Yearly deaths	DALYs (million)
Malaria	Protozoan: *Plasmodium*	Mosquito: *Anopheles*	643,000	46.400
Lymphatic filariasis *	Nematodes: *Wuchereria, Brugia*	Mosquito: *Culex, Aedes, Anopheles, Mansonia*	None	1.630
Human African Trypanosomiasis *	Protozoan: *Trypanosoma*	Tsetse fly: *Glossina*	1,360	0.083
Chagas disease *	Protozoan: *Trypanosoma*	Reduviid bug: *Rhodnius, Triatoma*	9,490	0.275
Leishmaniasis *	Protozoan: *Leishmania*	Sand fly: *Phlebotomus, Lutzomyia*	5,710	0.697
Onchocerciasis *	Nematode: *Onchocerca*	Black fly: *Simulium*	None	1.230
Dengue fever *	Flavivirus	Mosquito: *Aedes*	36,100	2.380
Yellow fever	Flavivirus	Mosquito: *Aedes, Haemogogus*	4,280	0.290
Zika	Flavivirus	Mosquito: *Aedes*	None	0.000347
Chikungunya *	Togavirus	Mosquito: *Aedes*	230 [15]	0.106 [15]

cycle is accelerated and where vector populations increase seasonally. This increases VBD burden primarily in tropical areas with low-income populations due to their high levels of contact with insects, low provision of medicine and sanitation, and high levels of malnutrition [16]. Human VBDs are also increasing outside of tropical areas; for example, tick-borne diseases such as Lyme disease, anaplasmosis and Rocky Mountain spotted fever are increasing in the United States due to higher tick densities and increased ranges [17].

1.2.3 Vector-borne pathogen biology and transmission dynamics

Vector-borne diseases are balancing acts between three separate organisms, with complex interactions and competition dynamics between each [18]:

- A pathogen capable of developing in multiple, limited, hosts;
- A vertebrate host that develops an infection with the pathogen that can be transmitted to a vector;
- A vector that can acquire an infection from the vertebrate host and transmit it to a second host.

Both the vertebrate host and vector categories may contain multiple species, depending on the host specificity of the pathogen being transmitted. Of the pathogens of humans, 61% are also found in non-human animals and can be transmitted between different host species; these pathogens are known as zoonoses, and cause the majority of emerging diseases in humans [19]. Many haematophagous insects are able to feed on multiple host species, facilitating transmission of zoonotic pathogens between humans and animals; Figure 1-1 illustrates the transmission pathways of zoonotic VBDs between hosts. Transmission to, and reinfection of, human populations from animals is an important aspect of disease epidemiology: incidence of a disease in humans may be low, but if reservoirs of the pathogen exist in animal populations, transmission is difficult to disrupt sufficiently for elimination of the pathogen. Zoonotic VBDs are therefore important in terms of current disease control and emerging diseases of the future.

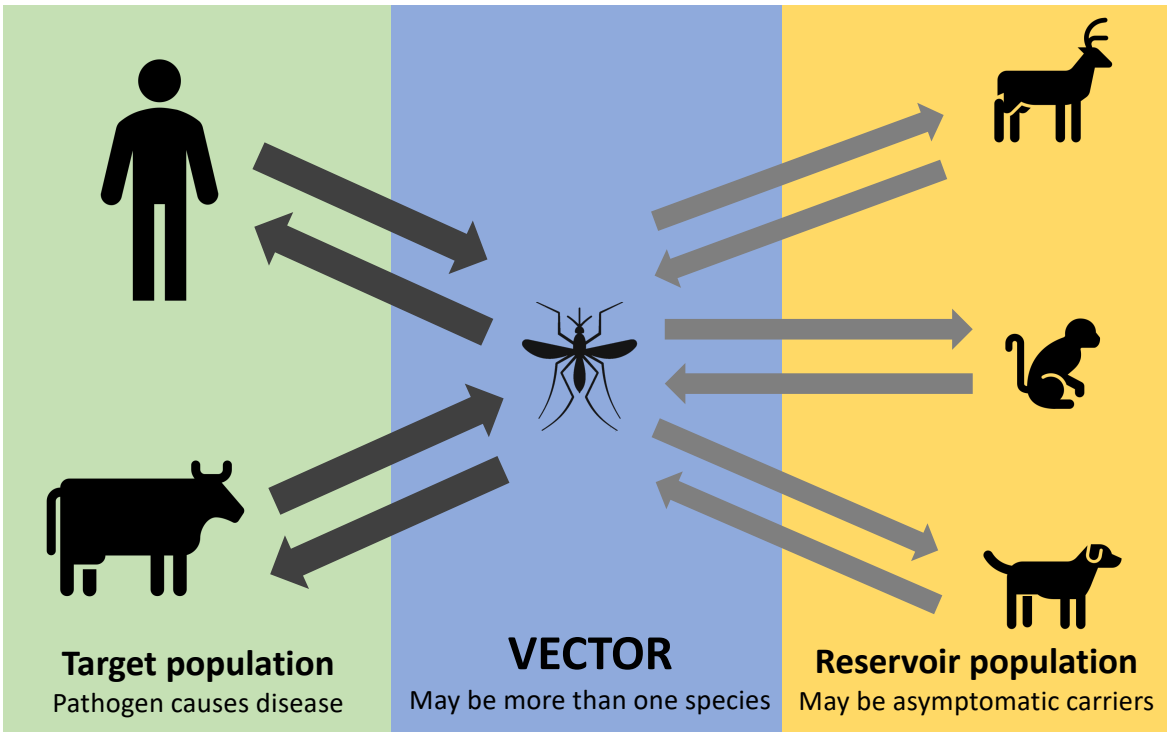


Figure 1-1 - Vector-borne pathogens are often capable of infecting not only multiple target hosts where the parasite life cycle can be completed, but also secondary reservoir hosts where the parasite can exist but not necessarily develop. Vectors are capable of transmitting pathogens between target (or spillover) and reservoir populations, allowing the reinfection of the target population even if clinical cases are controlled effectively.

Parasites with a sexual reproductive cycle, as is the case in protozoa and filaria, have more complicated host dynamics than those that only reproduce asexually. Any host in which intrinsic incubation – essential life cycle stages such as gametocyte union – occurs is known as a definitive host. Hosts in which only extrinsic incubation occurs without the sexual component of the life cycle, usually to amplify population size or for larval development, are known as intermediate hosts. Definitive hosts can be either the vector species or the vertebrate host: for example, although gametocytes (sexual cells produced by meiosis) of the malaria agent *Plasmodium* are produced in the human vertebrate host, they do not fuse to form oocytes until they infect the mosquito vector, making the mosquito the definitive host [20].

154

155 To complete its life cycle, a vector-borne pathogen will need to negotiate a range of
156 environments with different available nutrients and challenges. Developing parasites
157 are often required to move through different insect tissues – for example,
158 *Plasmodium* parasites are ingested via a mosquito blood meal and move through the
159 mouthparts and foregut to the midgut, where ingested gametocytes fertilise and
160 transform to ookinetes. Ookinetes must then exit the midgut through the peritrophic
161 membrane in order to traverse the hemocoel as oocysts to the salivary glands,
162 where mammal-infective sporozoites are released from oocysts to enter the salivary
163 glands for injection into the mosquito's next host [21]. On the other hand, the
164 causative parasitic agent of trypanosomiasis *Trypanosoma brucei* undergoes
165 differentiation in different areas of their tsetse fly host. *T. brucei* bloodstream forms
166 are ingested by tsetse in a blood meal and undergo differentiation into procyclic
167 trypomastigotes which are capable of multiplication within the midgut. *T. brucei*
168 parasites then cross the peritrophic membrane in order to differentiate within the
169 ectoperitrophic space between the peritrophic membrane and epithelium into a non-
170 multiplicative metacyclic form, before moving back through to the salivary glands and
171 being injected into the next host by the tsetse fly [22].

172

173 Moving around the host brings parasites into contact with a range of host defences,
174 requiring different avoidance strategies; Figure 1-2 shows the different types of direct
175 and indirect barriers to infection faced by parasites. Insects have an innate immune
176 system comprising of physical barriers, humoral responses (circulating proteins in
177 the haemolymph and organs that detect and neutralise pathogens) and cellular
178 responses (immune cells in the haemolymph that phagocytose or encapsulate

pathogens) [23]. Parasites must be able to avoid or overcome these responses sufficiently to complete their life cycle within the vector and transmit to a novel host. For example, *Plasmodium* parasites are able to mitigate the effects of mosquito immune system molecules, such as free radicals and anti-microbial peptides, by the presence of PIMMS43 proteins in the *Plasmodium* outer membrane which counteract these immune responses [24]. A physical barrier between parasites and the gut epithelia, called the peritrophic membrane, provides protection from the contents of blood meals. Digestive enzymes secreted in response to a blood meal can also present a barrier to parasite infection: in sand flies up to 50% of the initial inoculum of *Leishmania* parasites are killed by digestive enzymes within the first day [25, 26].

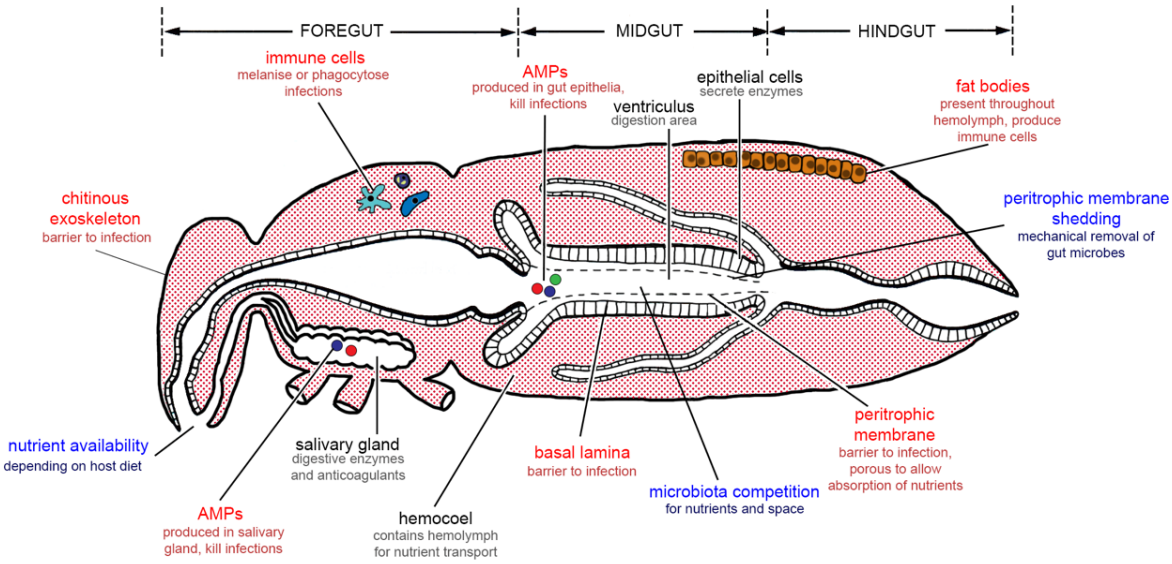


Figure 1-2 - A diagram of the generalised insect digestive system, with barriers to infection labelled. Red labels indicate immune responses mounted by the insect to clear infections; blue labels indicate indirect barriers to infection; black labels show important anatomy of the digestive system. AMPs = antimicrobial peptides. Modified from [2].

The metabolic needs of parasites and their hosts greatly overlap, and so the parasite must strike a balance between maximising its own fitness without overly reducing the fitness of the host. Host and parasite will co-evolve over the course of their relationship in an evolutionary arms race to maximise their own fitness, which requires the parasite to carefully balance virulence with transmissibility [27]. Parasites must be able to adapt to changes in host diet and stress responses, and will require different nutrients and conditions at different life stages within the vector host [28]. Parasites often show a close dependence on host metabolites: for example, *Plasmodium* parasites have lost the ability to synthesise amino acids *de novo*, opting instead to scavenge them from the catabolism of haemoglobin present in the mammalian host and the mosquito blood meal [29, 30]. This balancing act of surviving within a host while maximising fitness is also seen in commensal organisms; the microbiome of vectors, and their interaction with the host and the parasite, will be discussed in Section 1.3.

1.2.4 VBD control strategies

Vector control is the principal method used to control many VBDs, both historically and currently. Many VBDs have no safe or effective vaccine, including malaria [31], human African trypanosomiasis (HAT) [32], dengue [33] and chikungunya [34]. In some cases, the parasite's ability to evade the host immune system has slowed or prevented vaccine development: malaria parasites sequester within host cells to avoid the host immune response [35]; trypanosomes have highly variable antigenic coatings on their surface, reducing the host immune cells ability to mount a lasting defence [36]. A vaccine for one out of four serotypes of dengue virus has been developed, but is limited by antibody-dependent enhancement: immunity to one

dengue serotype can cause more severe disease when infected with others [33]. While some neglected VBDs are controllable through mass drug administration, such as onchocerciasis and lymphatic filariasis [37, 38], most lack the treatment options necessary for this approach. For example, Chikungunya has no treatment at all [34] and a novel oral HAT treatment has recently been developed to replace current intensive intravenous treatment regimens, but can cause severe adverse reactions [39, 40].

Vector control strategies interrupt transmission of pathogens by reducing contact between the vector and the host. There are many forms of vector control that have been used historically and are still used in some form today. Table 1-3 lists the most common forms of vector control. In the past, reduction of the vector habitat and destruction of reservoir hosts (see Figure 1-1) has been used to great effect: during the construction of the Panama Canal in 1904-1914 both yellow fever and malaria were effectively eliminated from the region through an intense integrated anti-mosquito program of habitat control, housing improvement, rudimentary chemical larvicides, trapping and prophylactic treatments [41]. Trypanosomiasis was effectively controlled in discreet areas in the first half of the 20th century using a combination of tsetse trapping, reservoir host destruction (mainly through the slaughter of game animals), and habitat spraying. However, whenever these control activities were stopped tsetse and trypanosomes re-infested the areas [42].

238

239 Table 1-3 - Types of control commonly employed to reduce or eliminate insect vector
240 species. Modified from [43, 44].
241

Type of control	Targeted life stage	Intervention
Chemical	Immature	Chemical larvicides
	Adult	Insecticide-treated bednets and clothing Indoor residual spraying Habitat spraying from aircraft or vehicles Insecticide treatment for cattle Insecticide-treated traps and targets Topical repellent Spatial repellent
Non-chemical	Immature	Microbial larvicides Predator species (fish or invertebrates) Permanent habitat modification (e.g. drainage of surface water) Regulatory measures (e.g. drain clearance, rubbish disposal)
	Adult	Housing improvement (e.g. screens for windows and doors) Removal trapping

242

243

244 Vector control has since evolved to become more integrated, widespread, and
245 sustainable. Modern vector control takes into account the impact on the
246 environment, ruling out the large-scale slaughter of animals where possible, and
247 limiting the use of pesticides which persist in the environment with adverse effects on
248 wildlife, such as dichlorodiphenyltrichloroethane (DDT) [45]. In 2004 the WHO
249 published a Global Strategic Framework on Integrated Vector Management (IVM)
250 [46], stating the need for collaboration between the health sector, private and public
251 agencies at local and global levels, and at-risk communities. IVM uses combinations
252 of the control strategies outlined in Table 1-3, implemented by multiple groups

simultaneously, and aims to control vectors broadly rather than focusing on a single vector species or pathogen. However, insecticides still form a large part of IVM, despite the fact that there is now documented insecticide resistance in all insect vector groups [47]. In 2017 the WHO published a global vector control response [48] outlining the steps necessary to achieve a 75% reduction in global VBD mortality by 2030. Innovation, particularly in the development of novel control tools, and an increase in basic research were identified as key aspects of meeting those goals. A major area of IVM innovation over the last decade has involved the use of bacteria in vector control strategies; elucidating the potential roles of bacteria native to insect vectors in VBD control will be the focus of this thesis.

1.3 Microbiota of vectors

All complex organisms are host to communities of microorganisms whose effect on the host range from beneficial to harmful – these communities are known as an organism's microbiota, or microbiome [49]. Bacteria with a close relationship with the host, whether beneficial or deleterious, are known as symbionts [9]. While it is acknowledged that some bacteria are transient members of the microbiome, introduced through the diet or environment with no real positive or negative effect on the host, the presence of a microbiome will have an impact on the host regardless of the effects of individual species.

While the microbiome is made up of a mixture of viruses, bacteria and fungi, bacterial symbionts are the most well-studied and well-understood, and will be the main clade discussed in this work. It is also worth noting that the nomenclature of insect bacteria symbionts does not always follow standard bacterial nomenclature, as few symbiotic bacteria are culturable to the standard necessary for classification under current rules. Many related symbionts isolated from different host species are given a single species designation despite in all likelihood being different species [50]; for this reason, many of the bacteria discussed in this work will be referred to using their genus name rather than a binomial.

The individual species that make up a microbiome have context-dependent effects on the host: they can play an important role in both disease and its prevention, causing disease states in a host by their presence, overabundance, or absence. These organisms are classified by their effects on the host, their tissue tropism within

the host, whether they can exist outside of the host and whether they persist in the host population and through subsequent generations.

The most traditional characteristic used to classify a symbiont within a host is whether the microbe has a net positive or negative effect on host fitness (the host's ability to survive and reproduce). This tends to span a linear spectrum from parasite to mutualist; parasitic microbes cause a net reduction in host fitness, and mutualistic bacteria increase host fitness. However, classifying a symbiont definitively as either a parasite or a mutualist is usually impossible; its effect will vary depending on the density of the microbe within the host, the environment, and the host species. Here, three types of behaviour exhibited by host-associated microbiota will be outlined: parasitic, mutualistic, and commensal.

1.3.1 Parasitic bacteria and the insect host

Symbionts that have a deleterious effect on the host's survival and fitness are referred to as parasites, or as having a parasitic relationship with the host. These deleterious effects may constitute mechanical or physiological injury to host tissues, stimulation of damaging immune responses, or nutrient deprivation; most parasites will cause more than one of these effects to facilitate infection of the host [9]. The degree of damage a parasite can cause to its host is referred to as the parasite's virulence; the mechanisms by which damage is done are known as virulence factors. Examples of virulence factors are the Cry proteins produced by *Bacillus thuringiensis,* which can compromise the midguts of a wide range of insects by acting on gut epithelial cells to facilitate infection [51] or cause death in insect brain cells by forming pores in the cell membrane [52]. Genes encoding virulence factors

312 are commonly located on mobile genetic elements which can be transferred between
313 bacteria of the same species, different species or to other clades of life by horizontal
314 gene transfer (HGT), allowing for the rapid spread of virulence factors within
315 bacterial populations [53].
316
317 Toxins like Cry proteins have pathogenic effects on a wide range of insect hosts, but
318 not all toxins are universally deleterious, making the categorisation of parasitic
319 organisms complex. Some toxins are beneficial for promoting the infection and
320 transmission of other microbes but are detrimental to the insect host. For example,
321 virulence factors produced by *Xenorhabdus* and *Photorhabdus* species, commensal
322 bacteria in insect-invading nematodes, promote establishment of nematodes in the
323 host by inhibiting the insect immune response [54]. In tsetse flies, *Sodalis* increases
324 host susceptibility to infection with parasitic trypanosomes, by an unknown
325 mechanism [55, 56]; however, *Sodalis* has also been shown to increase longevity in
326 trypanosome-naïve tsetse [57], making their role in the host more complicated than
327 simply parasitic or beneficial. Closely-related bacteria can also have different effects
328 on an insect host due to differences in virulence factors: the genus *Serratia* includes
329 *S. marcescens,* a pathogen of diverse insect clades [58], and *S. symbiotica*, which
330 synthesises essential nutrients for its aphid host [59]. Notably, both *S. symbiotica*
331 and the tsetse fly-associated bacterium *Sodalis* have recently evolved insect
332 associations compared to more long-term associated bacteria [60, 61]; the effect of
333 host association longevity on the role of bacteria in insects will be discussed in
334 Section 1.3.3.
335

An important parasitic capability of bacteria, which can be utilised in control strategies, is the manipulation of the host's reproduction to ensure transmission of the bacterium. This strategy is used by some bacteria that are vertically transmitted from mother to offspring, making the females of the host species valuable for increased transmission. The selection pressure is therefore for reproductive manipulation to increase the number of females in a population, or to increase the proportion of infected females in the population.

The symbiont *Wolbachia*, which is present in an estimated 50% of arthropod species as well as some filarial nematodes [62, 63], has highly varied relationships from parasitic reproductive manipulation to essential nutrient provision depending on the host and bacterial strain [64, 65]. Due to the diversity of evolutionary lineages within the *Wolbachia* genus and the difficulties in classifying them, rather than being divided into species and subspecies strains are divided into at least 16 supergroups – A-Q except for G, which is a combination of A and B [66-73] – based on a multi locus sequence typing system (MLST) [74]. In many arthropod species, including tsetse flies (*Glossina* sp.) and fruit flies (*Drosophila* sp.), *Wolbachia* causes cytoplasmic incompatibility (CI) between infected males and uninfected females, or between males and females infected with non-compatible *Wolbachia* strains, resulting in embryonic termination [75-78]. *Wolbachia* causes different types of reproductive manipulation in other insects: for example, *Wolbachia* feminises genetic males during pupation in the butterfly *Eurema hecabe* [79]; kills males during embryogenesis in the ladybird species *Adalia bipunctata* and the butterfly species *Acraea encedon* [80]; and induces parthenogenesis in several species of wasp (Order: Hymenoptera) [81-83]. The widespread host associations and manipulative

effects of *Wolbachia*, as well as its ability to block the development of certain pathogens [84], have led to its use in multiple VBD control strategies which will be discussed in Section 1.4.1.

Reproductive manipulation may not confer fitness costs to the host, and associations may change over a relatively short period, as has been demonstrated in the whitefly species *Bemisia tabaci*. A *Rickettsia* species successfully invaded a whitefly population from 2001-2007 using a combination of reproductive manipulation and facultative mutualism, by improving the survival and fecundity of the host as well as forcing the host to produce a higher proportion of daughters [85]; this was followed by a decline of the *Rickettsia* from near-fixation in the whitefly population in 2011 to 36% prevalence in 2016, ostensibly due to the loss of conferred fitness benefits to the host [86].

1.3.2 Commensal bacteria and the insect host

Commensal bacteria, while not directly benefiting or reducing host fitness, have a more general effect on the host environment. Commensal bacteria can increase or decrease host susceptibility to infection with other organisms. For example, the presence of a normal microbiome in mosquitoes increases the virulence of the pathogenic fungus *Beauveria bassiana*, possibly by the fungus facilitating translocation of bacteria which normally reside mutualistically in the gut, to the hemocoel where they become pathogenic [87]. A similar effect is seen in sand flies; the protozoan parasite *Leishmania* – which causes leishmaniasis in humans and animals, and reduces the sand fly's ability to take full blood meals [25, 88] – cannot establish an infection in sand flies whose bacterial microbiota has been removed

with antibiotics [89]. This effect has been attributed to the bacterial microbiota creating a permissible environment within the insect host for the invading parasite to successfully establish – in the case of the sand fly and *Leishmania*, it is hypothesised that utilisation of sucrose by the microbiome was necessary to prevent the *Leishmania* parasites dying from osmotic stress [89]. Commensal resident bacteria will also present a challenge to invading microbes by occupying the available niches within the host and competing for resources and space [90].

1.3.3 Mutualistic bacteria and the insect host

Many clades have independently evolved close relationships with individual bacterial species, whereby the bacteria increase the fitness of the insect host. These essential microbial partners are characterised by a long-term association with a host species and frequent transmission within the host population, resulting in a high level of specificity and co-evolution between host and microbe to the mutual benefit of each [91]. In some cases, with single-celled organisms and associated bacteria, these symbiotic relationships become so close that the microbe is absorbed into the host cell and becomes an integral part of the cell, in a process known as symbiogenesis. An outstanding example of this is the hypothesis that eukaryogenesis involved assimilation of an archaeon by an alpha-Proteobacterium to become an intracellular energy production unit, giving rise to modern mitochondria in eukaryotic cells [92].

Insects have evolved closer relationships with specific bacteria than most higher organisms – providing nutrients lacking in a restricted diet, or contributing to the host's defences against parasites – and their small body size results in a less complex microbial community than larger animals, making insects ideal models for

411 the study of host-symbiont interactions [93]. In insects, beneficial bacterial partners
412 tend to be divided into two groups – primary and secondary symbionts, which will be
413 discussed further here.
414
415 1.3.3.1 Primary symbionts
416 Primary symbionts are obligate (essential) microbial partners for their host, found in
417 all individuals in a species; most have become entirely specialised to live
418 intracellularly in a host: they can no longer invade other species or live freely in the
419 environment [94]. Most primary symbionts rely entirely on the host for transmission,
420 which occurs vertically – from mother to offspring – and acquires all nutrients
421 necessary for its growth and development from the host diet. As with most symbiont
422 descriptions, there are exceptions: for example, *Rhodococcus rhodnii* is a primary
423 symbiont of rejuviid bugs (*Rhodnius prolixus*), essential for their development and
424 reproduction, but is transmitted through contaminated egg surfaces or faeces [95].
425 Primary symbiont-host relationships have evolved independently in different bacterial
426 clades, usually in insects with restricted diets such as blood or phloem: for example,
427 the primary symbiont *Buchnera aphidicola* has been associated with its aphid host
428 for 160-280 million years, providing the host with essential amino acids lacking in its
429 phloem sap diet [96, 97].
430
431 Insects are unique in that approximately 15% of insects have specialised cells for
432 containing primary endosymbionts (intracellular symbionts), known as bacteriocytes
433 [98, 99]. These cells, which often constitute an organ-like structure known as a
434 bacteriome, facilitate the growth of endosymbionts and allow the exchange of

metabolic products essential to the host and the endosymbiont. Almost all insects with restricted diets have a primary bacterial endosymbiont providing essential nutrients not present in their diet, which suggests that the evolution of endosymbiosis was a route for evolutionary innovation that allowed for the great diversity and niche adaptation present in the insect clade [100].

1.3.3.2 Secondary symbionts

Secondary symbionts are also known as facultative symbionts, as their presence is not a requirement for host survival or reproduction; as a result, they tend to be erratically distributed within host populations [101, 102]. They usually show less tissue specificity than primary symbionts, and are often found both extracellularly in the haemolymph, within reproductive organs and in the gut, as well as intracellularly in the bacteriome [103, 104]. This can be attributed to retaining several virulence genes that are common in bacterial parasites and are used for invading novel hosts, such as the type III secretion system in the tsetse secondary symbiont *Sodalis* which is used to invade host cells [105].

To persist and spread in an insect population, secondary symbionts must affect the insect phenotype either by conferring a benefit to the host or promoting their own transmission over the host fitness. Several bacteria that are classified as secondary symbionts in the latter category, promoting their own transmission at the expense of the host, have been discussed in the previous section on parasites. For example, in associations where *Wolbachia* fulfils a secondary symbiont role it can manipulate

host reproduction to increase its transmission in a host population (see Section 1.3.1).

Secondary symbionts can provide nutritional support to their hosts by metabolising components of the host's diet that would otherwise be inaccessible, or by complementing the production of essential nutrients by primary symbionts. For example, the higher termite family Termitidae diet consists of indigestible lignocellulose complexes, and a 'bioreactor' community of bacteria within the gut is responsible for digesting the lignocellulose into compounds usable by the host; this has allowed higher termites to invade new ecological niches and has contributed significantly to the success of the clade [106]. Secondary symbionts can also be known to help the insect host endure heat stress, possibly by rescuing functions usually performed by a primary symbiont, which has specialised to function only within the strict normal parameters of the host environment [107].

There are several instances of interactions between primary and secondary symbionts that occur regularly to increase host fitness, rather than in response to stress. In tsetse flies, the primary symbiont *Wigglesworthia* is responsible for the production of metabolites which facilitate tsetse reproduction; however, there is evidence to suggest that biosynthesis of the amino acid arginine is achieved by complementary metabolism between *Wigglesworthia* and the secondary symbiont, *Sodalis* [108]. This complementary metabolism between primary and secondary symbionts is taken further in the mealybug *Planococcus citri*, in which a matryoshka doll-style symbiont system exists. The primary symbiont *Candidatus* Tremblaya princeps lives intracellularly within the insect bacteriocytes, with the secondary

symbiont *Candidatus* Moranella endobia living intracellularly within the primary symbiont [109-111]. Genes from the host and both nested bacteria form a network of interdependent metabolic pathways to synthesise essential amino acids lacking in the host diet [112].

Secondary symbionts can also confer direct protection against parasites: the most outstanding example of this is the protection of aphids against parasitoid wasps conferred by *Hamiltonella defensa*. Parasitoid wasps pierce a host's carapace with an ovipositor and lay eggs in the aphid body cavity, where the larva will consume the aphid as its food source. One of the many mobile genetic elements in the *H. defensa* genome is APSE-1, which produces a toxin that prevents the development of parasite larvae in the aphid, rescuing the host [113-118]. *H. defensa* also confers protective advantages to whitefly, by expressing factors which suppress plant defences induced by the whitefly feeding on a tomato plant [119].

1.4 VBD control strategies involving bacteria

The WHO have identified vector control strategies involving bacteria as an important area for research to work towards VBD control targets [48]. Although most advances in the field of microbiota-based vector control have occurred in the last decade, pathogenic bacteria have been used in vector control strategies since the early 1970s. The first microbe to be regularly used in insect control strategies, *B. thuringiensis*, is still used for its parasitic effect on plant-feeding insects to protect crops and can be used as a mosquito larvicide [120, 121]. The use of insect symbionts in VBD control is now advancing quickly due to advances in genetic

technology and our understanding of VBD host-parasite dynamics. Pathogen transmission can be reduced either by targeting the vectorial capacity of the insect or by suppression of the insect's fitness. Three main strategies for vector control using insect microbiota have emerged [122]:

1. The formation of heterologous associations - where an insect-associated microbe is introduced to a new host species
2. The disruption of the vector's naturally associated microbes
3. The genetic modification of insect-associated microbes.

1.4.1 Heterologous associations

1.4.1.1 Vector competency reduction by heterologous associations

Heterologous associations are used when the introduced microbe is compatible enough with the new host that it can establish and spread in the population, whilst also conferring a disadvantage either to the host or the transmitted pathogen. This has been most extensively trialled with the mosquito species *Aedes aegypti* transinfected with *Wolbachia*. Two main strategies are used to control VBDs using *Wolbachia*: population replacement strategy and the incompatible insect technique. Some strains of *Wolbachia* can reduce the infection of their hosts with certain pathogens including bacterial, filarial nematodes and the malaria parasite *Plasmodium* [84, 123-127]. Mechanisms for this pathogen resistance vary depending on the strain and host species, including competition with the pathogen for resources, inhibition of pathogen entry into the host cell, and induction of the host immune response [127-134]. *Ae. aegypti* is a vector for several viruses including chikungunya, Zika, yellow fever and dengue, which affect millions of people around

the world [135] (see Table 1-2), and is not naturally infected with *Wolbachia* [136]. In a population replacement strategy, CI causes the rapid replacement of natural mosquito populations with *Wolbachia*-infected populations; the natural pathogen-blocking activity of *Wolbachia* reduces disease transmission [137-142]. This has been successfully trialled by introducing the *Wolbachia* strain *w*AlbB into wild *Ae. aegypti* in Kuala Lumpur, Malaysia, which resulted in reduced incidence of dengue at release sites [139]. The incompatible insect technique uses *Wolbachia*-induced CI to directly suppress wild mosquito population size, and has been demonstrated to work in multiple trials with *w*Pip and *w*AlbB in *Ae. aegypti* and *w*Pip in *Ae. albopictus* [143-147].

1.4.1.2 Issues with heterologous associations in vector control

Although releasing transmissible symbionts like *Wolbachia* into novel host populations appears to allow for self-sustaining control, there are still issues to be addressed by future research. Fluctuation in environmental conditions, such as temperature, can reduce the burden, CI and vertical transmission of *w*Mel in *Ae. aegypti* [148]. Other members of the *Ae. aegypti* microbiota may be affected by the introduction of *Wolbachia* – for example, a study showed increased insect-specific flavivirus loads and infection rates in *Ae. aegypti* infected with *w*Mel, demonstrating that *Wolbachia* does not inhibit all flaviviruses [149]. As with most control methods there is the potential for the evolution of resistance: in *Drosophila*, anti-viral function depends on high *Wolbachia* load, which confers fitness costs to the host – whereas CI only requires low *Wolbachia* load [150]. In mosquitoes this would mean low *Wolbachia* load would be selected for, which would reduce the efficacy of

Wolbachia-associated protection against viral infections [151]. Further investigation into the physiological interactions between hosts and their microbiota, as well as long-term field studies, are required.

1.4.2 Genetically modified microbial partners

1.4.2.1 Vector competency reduction using paratransgenesis

Paratransgenesis is the alteration of vector competency by genetically manipulating the vector's associated microbiota [152]. To be a candidate for paratransgenesis the microorganism must be culturable, amenable to genetic manipulation and must be horizontally or vertically transmissible to spread throughout a population [153]. The bacteria are then genetically manipulated to produce an agent that is specifically toxic to the parasite, without reducing host fitness, fecundity, or the ability of the bacterial agent to propagate through the insect population enough to prevent its fixation in the population. This concept has been demonstrated in *Rhodnius prolixus*, the hemipteran vector of *Trypanosoma cruzi* (Chagas disease). *Rhodococcus rhodnii* is an obligate symbiont of *R. prolixus* and is found in the midgut, where it comes into contact with *T. cruzi* taken up in a blood meal. A shuttle plasmid was used to insert a gene encoding the peptide cecropin A, which is lethal to *T. cruzi*, into *R. rhodnii* which were then fed to *R. prolixus* first-instar nymphs. Transgenic *R. rhodnii* were shown to persist through all *R. prolixus* developmental stages and reduced *T. cruzi* abundance in the insect midgut to almost undetectable levels [154-156]. Symbiotic bacteria of *R. prolixus* are transferred between individuals via their faeces; a simulated faecal paste called CRUZIGARD was developed to exploit this

578 transmission route to spread paratransgenic *R. rhodnii* in *R. prolixus* populations
579 [157].
580
581 Paratransgenesis has also been demonstrated to reduce infection of *Plasmodium*
582 *sp.* in mosquitoes. A range of bacteria including *Pantoea agglomerans* and *Asaia*
583 *bogorensis* have been transformed to express antibody fragments or anti-microbial
584 peptides, reducing successful infection of *Plasmodium* in the mosquito midgut [158,
585 159]. Most insect vectors do not share a symbiont transmission pathway that is as
586 easy to exploit as that of *R. prolixus*, so transmissibility of potential paratransgenesis
587 candidates is vital. The *Asaia* genus is also being investigated as a potential
588 paratransgenesis candidate: it is transmitted vertically as well as horizontally and is
589 widespread among insects including mosquitoes [160].
590
591 Genetically-modified mutualistic symbionts can be used as a 'Trojan horse' for
592 introduction of lethal agents into vector hosts – either double-stranded RNA (dsRNA)
593 that are specific to essential genes of the insect, or protein toxins. The use of dsRNA
594 has been demonstrated in *R. prolixus*, using *R. rhodnii* to deliver dsRNA that
595 reduced the insect host fecundity [161]. The gene encoding the Tc protein toxin,
596 produced naturally by *Photorhabdus* bacteria, has been inserted into the bacterial
597 symbiont *Enterobacter cloacae* of the pest termite *Coptotermes formosanus*; within
598 12 hours 75% of insects were colonised, and within a month >90% were dead
599 compared to <11% mortality in control groups [162]. This method of insect control
600 also relies on bacteria which are readily culturable and readily spread throughout
601 host populations – with ready transmission even more important, as acquisition of
602 the virulent bacteria will naturally be selected against.

1.4.2.2 Issues with GMOs in insect pest control

The use of paratransgenic bacteria in insect pest control clearly has great potential, but no systems using them have yet been developed commercially. The use of GMOs carries a stigma based on fear that the GMO will be disseminated to non-target hosts, or that genetically modified constructs will be transmitted to other bacteria in the environment by horizontal gene transfer. These issues are combated by making the active agents highly specific to the desired host system, or by including bio-containment strategies within the modification – such as a reliance on non-standard amino acids, resistance to mutagenesis and HGT, or expression of a 'suicide gene' in habitats other than the desired host [162, 163]. Progress in field trials has been limited, and extra difficulties are encountered when international regulatory agencies must cooperate to approve trials.

1.4.3 Targeting of obligate microbial partners

Another method of vector control involving microbial partners is to eliminate them from the host, damaging host growth, reproduction, and survival. Antibiotics are commonly used to eliminate obligate symbionts during research [164], but are not suitable for use in large-scale control efforts due to the likelihood of resistance developing and spreading to become a public health issue [165]. Instead, physiological interactions between the host and the symbiont can be targeted – such as the exchange of nutrients or the host's immune system [166, 167].

A control strategy targeting symbionts is ideally suited to bacteria that are so specialised to their host environment that they are not horizontally transmitted and are consistently localised in one location – the best example being bacteria that live in specialised structures within their host, such as bacteriocyte cells [99]. This makes them simple to target and removes the chance that the targeted bacteria will reinvade from another host. It is possible to target horizontally transmitted symbionts, but as they will be reintroduced from other insects or the environment the removal pressure needs to be constant; this can be achieved by using other microbes, genetically modified or not, which can be horizontally transmitted between hosts to target the obligate symbionts. This strategy was used to control termites that damage wooden structures – protists living in the termite hindgut and are essential for wood digestion were targeted by a commercially-available genetically-modified yeast *Kluyveromyces lactis* [168]. Targeting of a primary symbiont is also being deployed to control filarial nematodes, the vector-borne causative agents of neglected diseases such as lymphatic filariasis and onchocerciasis (Table 1-1). In several human-infecting filarial species *Wolbachia* has a primary symbiont role, in contrast to its parasitic role in many insects (see Section 1.3.1), and patient treatments involving anti-*Wolbachia* agents are currently being developed [169].

The genetic study and manipulation of microbial partners, which are resistant to current culture methods, has only been feasible since the development of genomic technologies in the last few decades, making this a broad and novel area of research. The technological advances making the study of insect symbionts possible will be discussed in Section 1.7.

1.5 Tsetse, trypanosomes and *Sodalis glossinidius*

1.5.1 Human and animal trypanosomiasis

Protozoan parasites of the genus *Trypanosoma* (Class: Kinetoplasta) cause VBDs in both animals and humans, with significant socio-economic impacts. HAT, also known as sleeping sickness, is a vector borne NTD of sub-Saharan Africa that presents in two clinical forms, caused by two subspecies of *Trypanosoma brucei: T. b. gambiense* and *T. b. rhodesiense*. Animal African trypanosomiasis (AAT) is primarily caused by *T. b. brucei*, *T. vivax*, *T. congolense, T. evansi* and *T. simiae* [170]. In cattle AAT is known as nagana, a wasting disease with enormous economic impact on largely rural and impoverished areas [171]. All forms of African trypanosomiasis are vectored between hosts by the bite of the tsetse fly, *Glossina sp.* (Family: Glossinidae), whose distribution largely defines the areas at risk of HAT and AAT [172].

Also of note is American trypanosomiasis, or Chagas disease, caused by *T. cruzi* and vectored by the bite of an infected triatomine bug (Family: Reduviidae). Endemic in 21 Latin American countries, where it affects predominantly impoverished and immunosuppressed populations, Chagas disease causes cardiomyopathy and is responsible for a global loss of 806,170 disability-adjusted life years (DALYs) [173].

1.5.1.1 Human African trypanosomiasis (sleeping sickness)

Both *T. b. gambiense*- and *T. b.* rhodesiense-caused diseases are included in the term HAT and grouped together for control strategies, despite being pathologically and geographically distinct. *T. b. gambiense* is widespread in western and central

Africa, endemic in 24 countries and accounts for 98% of reported cases of sleeping sickness [174]. Infection with *T. b. gambiense* is almost always fatal if left untreated, although a small proportion of infected individuals demonstrate trypanotolerance: whereby infection with trypanosomes does not result in clinical disease or high parasitaemia, and does not require intervention [175]. Both *gambiense* and *rhodesiense* forms of HAT present in two stages: an initial haemolymphatic phase in the blood and lymph, and a subsequent meningoencephalitic phase when the parasites enter the central nervous system [39]. In *gambiense* HAT the initial phase presents with reoccurring fevers over several months, often going unnoticed until the disease has progressed to the second phase [176]. The progression of these symptoms, if untreated, will cause death on average within three years [177]. Rats, primates and pigs are known reservoirs of *T. b. gambiense,* although the effect of animal reservoirs on the epidemiology of *T. b. gambiense* is believed to be minimal [178-180]. The *rhodesiense* form of HAT has similar symptoms to the *gambiense* form, but symptoms manifest over a much shorter time period: the initial phase lasts a few weeks, followed by death within six months [181]. Animal reservoirs are important in the epidemiology of *rhodesiense* HAT, especially domestic cattle [182, 183].

Diagnosis of the early stages of HAT is difficult and expensive, making surveillance an ineffective tool to control HAT on its own. As symptoms of both forms of HAT are non-specific, they are frequently mistaken for other diseases and misdiagnosed. All treatment courses for both forms and stages of HAT, except fexinidazole, require intramuscular injection or intravenous drip application; the main treatment melarsoprol causes encephalopathy which kills up to 12% of patients [39]. Due to the

699 ability of trypanosomes to vary their antigen surface glycoproteins to evade the host
700 immune system a vaccine has proved difficult to produce [184]. However, a vaccine
701 candidate against the AAT pathogen *T. vivax*, which has a more limited antigen
702 repertoire, has recently been developed to target a conserved antigen located on the
703 flagellum [185, 186].
704
705 Both forms of HAT have had, and continue to have, important socio-economic
706 consequences for the development of endemic countries. Human forms of
707 trypanosomiasis are thought to have played an important role in the evolution of
708 early hominids [187] and the migration of humans to settle in non-infested areas until
709 the modern age [188]. In the 19th century, colonial displacement across sub-Saharan
710 Africa forced populations to move into tsetse-endemic areas which had historically
711 been avoided; this, combined with the famine and poverty caused by colonisation,
712 caused several HAT epidemics in the 19th and early 20th centuries [189].
713
714 Twice throughout history – in the 1960s and the 1990s – sustained control efforts
715 have greatly reduced HAT cases across sub-Saharan Africa; after the first control
716 effort reduced cases to less than 5,000 control was relaxed and the disease
717 reappeared in force in the 1970s [190]. Since the 1990s the collaboration of the
718 WHO, national control programmes and non-governmental organisations to control
719 HAT has reduced cases dramatically, as shown in Figure 1-3, and HAT is now
720 targeted for elimination as a public health problem by 2020 [190].
721

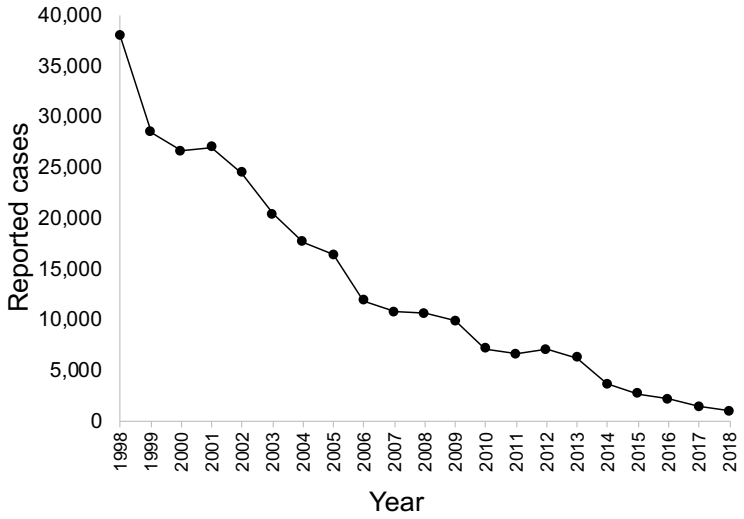


Figure 1-3 - The number of reported cases of both gambiense and rhodesiense African trypanosomiasis, from 1998 to 2018. Data from World Health Organization [3].

1.5.1.2 Animal African trypanosomiasis (nagana)

AAT is endemic in 37 out of 54 African countries and causes the death of millions of livestock each year, as well as loss of productivity in many more [191, 192]. AAT is caused by the zoonotic *T. b. rhodesiense* and the animal-specific *T. vivax* and *T. congolense* [193]. AAT symptoms result in a loss of milk production and reduced ability to plow land for planting, and as a result of AAT approximately 8.7 million km^2 of land is unsuitable for mixed crop and livestock agriculture [194]. As the communities in the AAT range are mostly rural, farming communities, AAT causes huge economic losses in sub-Saharan Africa and is credited as a major reason for the lower economic development of AAT-endemic areas [16].

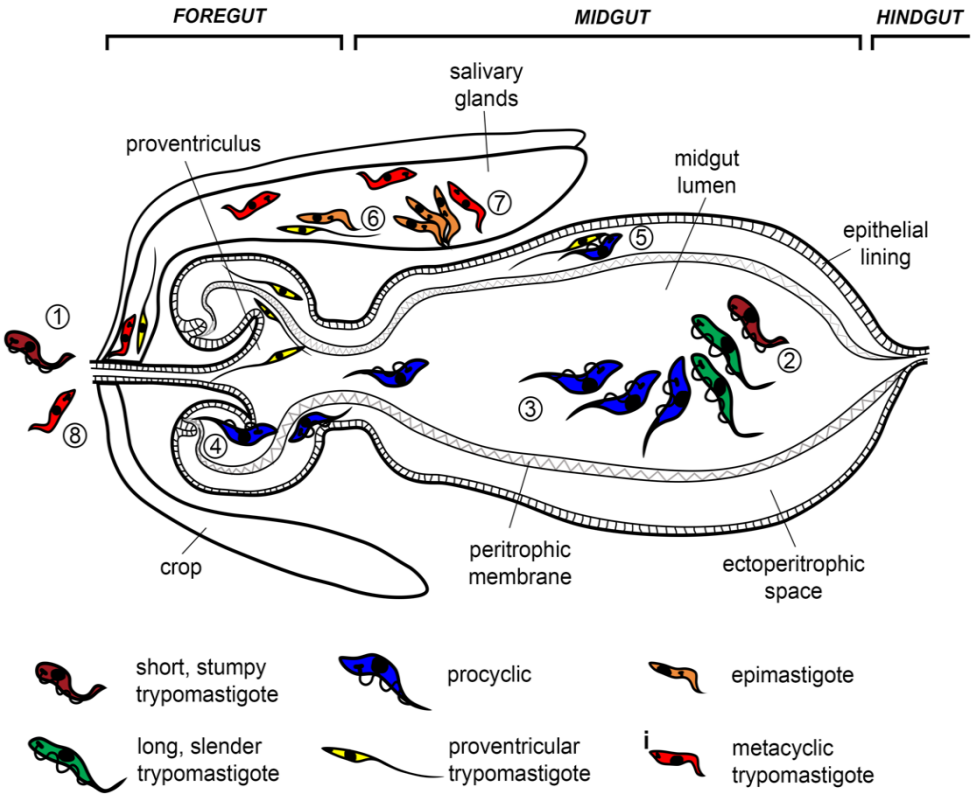
Although the primary vector of trypanosomiasis is the tsetse fly, it is worth noting that some trypanosomes, in particular *T. vivax,* have also been shown to be transmitted between hosts mechanically. In Africa mechanical transmission occurs through other

biting flies such as *Stomoxys sp.* and *Atylotus sp.* [195, 196], and in the Americas *T. cruzi* is primarily vectored by infected *Triatoma sp.* faeces entering the host through cuts or mucous membranes [197]. Multiple transmission methods through several vector species add a complex dimension to vector-mediated control of trypanosomiasis.

1.5.2 The tsetse fly vector

Tsetse flies include all flies in the genus *Glossina* (Class: *Diptera*) (Figure 1-4); there are 31 species and subspecies of tsetse, all of which are considered capable of transmitting trypanosomes [198]. Tsetse are categorised into three subgroups, determined by their habitat: riverine (*palpalis* group), savanna (*morsitans* group) and forest-dwelling (*fusca* group). The main species involved in *rhodesiense* HAT transmission are *G. palpalis s. l.* and *G. fuscipes s. l.* in the *palpalis* group [199]; *gambiense* HAT is transmitted mainly by the *morsitans* group, including *G. morsitans, G. swynnertoni* and *G. pallidipes* [172]. AAT is transmitted by a wide range of tsetse species in all three of the tsetse subgroups [200-204]. Figure 1-5 shows the life cycle of *Trypanosoma brucei s.l.* in tsetse flies.

Figure 1-4 - Lab-reared unfed male *Glossina morsitans morsitans*, taken at the Liverpool School of Tropical Medicine.



1) Short, stumpy trypomastigotes are ingested from a mammalian host during a blood meal, migrating to the anterior end of the midgut
2) Short, stumpy trypomastigotes differentiate to slender forms and multiply asexually for around five days
3) Slender form trypomastigotes differentiate into procyclics in the midgut lumen, and travel into the proventriculus while multiplying
4) Procyclics cross the forming peritrophic membrane and enter the ectoperitrophic space
5) Procyclics differentiate into long, motile proventricular trypomastigotes. Unable to multiply, proventricular trypomastigotes migrate forwards out of the ectoperitrophic space in the proventriculus and into the salivary glands
6) In the salivary glands trypomastigotes differentiate into epimastigotes, which attach to host cells or multiply sexually in the lumen
7) Epimastigotes differentiate into mammal-infective metacyclic trypomastigotes
8) Metacyclic trypomastigotes are injected into a mammalian host via the tsetse saliva during a blood meal

Figure 1-5 - The life cycle of *Trypanosoma brucei s.l.* in the tsetse fly vector. Adapted from Aksoy et al. (2003), Awuoche et al. (2017), Peacock et al. (2019) and Rose et al. (2020).

Both male and female tsetse are obligate haematophages, meaning they feed solely on blood from a living host – although there is evidence to suggest that tsetse may feed on plant sugars when blood meals are scarce [205]. Tsetse flies reproduce by

761 adenotrophic viviparity: a single egg is produced and retained within the female
762 uterus, where it hatches and is raised through four larval instar stages fed by
763 specialised milk glands, before being birthed by the female fly [1]. Birthed larvae
764 immediately bury underground and pupate for one month before emerging as an
765 adult tsetse fly; tsetse females can produce a single offspring every 10-20 days
766 [206]. Interruption of the tsetse life cycle by vector control the most efficient method
767 for suppressing outbreaks of trypanosomiasis [207, 208].
768
769 Despite being the primary vector species for trypanosomes, tsetse are naturally
770 refractory to trypanosome infection; in the wild usually fewer than 10% of tsetse are
771 found to be infected with trypanosomes despite high infection rates in mammalian
772 hosts in the area [209]. The degree of refractoriness is attributed to a complex
773 combination of factors, including the species of tsetse and strain of trypanosome, the
774 temperature, tsetse age and sex, degree of tsetse starvation, number of inoculating
775 trypanosomes, composition of the blood meal and interactions with tsetse symbionts
776 [209]. Through their long evolutionary history, tsetse and trypanosomes have
777 developed a complex system of barriers to trypanosome infection and responses to
778 these barriers [57]. The main barrier to the initial establishment of trypanosomes in
779 the midgut are tsetse-produced lectins – carbohydrate-binding proteins that attach to
780 pathogens and mediate the host immune response. When fed a trypanosome-
781 infected blood meal containing lectin inhibitors, 100% infection rates of tsetse have
782 been produced [210].
783
784 A major factor affecting the susceptibility of tsetse to trypanosome infection is tsetse
785 age. Newly emerged, unfed tsetse adults have distinct appearances and behaviours

786 from older flies and are referred to as “teneral” [211]. Teneral flies have soft
787 exoskeletons, a protruding head pouch (ptilinum) used to break out of the puparial
788 case, and are still maturing physiologically; in the wild, newly emerged tsetse adults
789 rest for two days instead of feeding or mate-seeking [212]. Teneral tsetse are more
790 susceptible to trypanosome infection when given an infected bloodmeal than older,
791 previously-fed tsetse, in what is described as the teneral phenomenon [211, 213].
792 This effect is not caused by ingestion of fewer trypanosomes or any trypanosome-
793 mediated interaction [214], but a complex combination of host factors which are still
794 being investigated. In teneral flies the protective peritrophic membrane is still
795 forming, requiring 84 hours post-eclosion to be fully extended along the length of the
796 midgut [214, 215]. Peritrophic membrane length has been found to correlate
797 positively with tsetse refractoriness to the establishment of a trypanosome infection
798 [214, 215], and recent elucidation of the trypanosome life cycle involving crossing the
799 peritrophic membrane while it is still being formed in the proventriculus supports the
800 idea of trypanosomes exploiting the immature peritrophic membrane (see Figure 1-5)
801 [216]. Additionally, tsetse EP protein [217] – which is mostly found in the peritrophic
802 membrane and midgut lumen and has a role in preventing trypanosome
803 establishment [218] – is expressed less in teneral tsetse compared to mature tsetse
804 [219]. As well as host-mediated effects, microbiome-mediated effects could also be
805 contributing to the teneral phenomenon; this will be explored in Chapter 4 .
806

807 ### 1.5.3 Microbiota of tsetse flies

808 Insects with extremely restricted diets all possess closely associated bacteria that
809 provide nutrients absent from the host diet [220]; tsetse are obligate
810 haematophages, feeding solely on blood in the adult life stage and on interuterine

milk from specialised glands during the larval stage [1]. Food is the primary source of commensal microbiota for animals, with most insects containing around 20-30 different bacterial taxa [221-223]; the largely sterile nature of tsetse food sources has resulted in a comparatively low number of associated commensal bacteria [93]. Three species of bacteria play important roles in the nutrition, fecundity and vectorial capacity of tsetse flies: *Wigglesworthia glossinidia*, an obligate primary symbiont; *Wolbachia*, a parasitic reproductive manipulator; and *Sodalis glossinidius*, a secondary endosymbiont [224]. These, as well as other bacteria associated with tsetse flies, are discussed in this section.

1.5.3.1 Wigglesworthia glossinidia

Wigglesworthia [225] has evolved an obligate symbiosis with tsetse flies for 50-80 million years; it is now only able to exist within the tsetse host and is entirely dependent on tsetse for its transmission [226]. This is evident by the reduced genome of *Wigglesworthia*, which is 700 kilobases (kb) long and highly specialised to life in the tsetse host environment [227]. *Wigglesworthia* is present in all tsetse and constitutes 71-99% of the tsetse microbiome [228]. Two separate populations of *Wigglesworthia* exist within the host: intracellular populations within bacteriocytes, specialised cells making up the bacteriome [225]; and extracellular populations in milk gland secretions, where they are transmitted vertically to tsetse offspring [229]. *Wigglesworthia* plays two essential roles within the tsetse host: providing essential nutrients by *Wigglesworthia* that are lacking in the blood diet of tsetse, including thiamine (vitamin B1), pyridoxine (B6) and folate [224]; and activating the

development of the tsetse immune system [219, 230-232]. When *Wigglesworthia* is removed, tsetse are unable to produce full-term larvae [233-236].

1.5.3.2 Wolbachia pipientis

The widespread insect associations of *Wolbachia* [237] and its manipulation of insect reproductive systems are discussed in Section 1.3.1. In tsetse *Wolbachia* is found intracellularly in the germ line, where it is vertically transmitted transovarially, and causes CI [238]. The abundance of *Wolbachia* in wild tsetse varies by tsetse species and population, from completely absent in *G. p. palpalis*, to present in 100% of individuals in a population of *G. m. morsitans* [239]. Tsetse-infecting *Wolbachia* belong to supergroups A and B [239]. The *Wolbachia* genome is small at 0.86-1.8 million base pairs (Mbp) and contains the bacteriophage WO as well as a large number of mobile elements [240, 241]. *Wolbachia* has been identified as having potential uses in tsetse control strategies [240], using CI in conjunction with other control strategies to reduce tsetse population densities, as discussed in Sections 1.3.1 and 1.4.1.1.

1.5.3.3 Sodalis glossinidius

Sodalis [242] is the most recently evolved facultative endosymbiont of tsetse flies [60]. The recent evolution of *Sodalis*' relationship as a symbiont of tsetse is evidenced by its flexibility: *Sodalis* exhibits broad tissue tropism within the tsetse host, and can be found in the midgut, fat body, salivary gland, milk glands and freely in the haemolymph [104, 243]. Unlike most insect-associated symbionts, *Sodalis* can be cultured outside of the host in cell-free medium [244, 245], demonstrating that it

has not had a long evolutionary history in the tsetse niche habitat [246]. Culture of *Sodalis in vitro* is possible due to its retention of a functional repertoire of genes: the genome of *Sodalis* is 4.1 Mbp, similar in size to the genome of free-living *Escherichia coli* [247], and large in comparison to the 700 kb streamlined, specialised genome of *Wigglesworthia* [227]. Several bacterial isolates discovered in other insects, such as stink bugs, have been designated close allies of *Sodalis* [248-250], and one member of the *Sodalis* genus, *S. praecaptivus*, is free-living [251, 252]. The evolution of *Sodalis* from free-living to obligate symbiont will be examined experimentally in Chapter 3.

Sodalis prevalence in wild tsetse varies by tsetse species and location: studies of tsetse across sub-Saharan Africa have found prevalence rates of *Sodalis* to range between 1-94% [102, 253], although it is completely absent in some populations [254, 255], suggesting a truly facultative symbiosis. However, for unknown reasons, in lab-reared tsetse colonies *Sodalis* becomes fixed at 100% of individuals [256]; this suggests that *Sodalis* may confer an advantage when selection pressures found in the wild are removed, or that transmission is more efficient in a lab environment. *Sodalis* is transmitted both vertically, via milk secretions and transovarial transmission, and horizontally via sexual contact [104, 257], allowing for its fixation in a closed lab tsetse population.

The effect of *Sodalis* on the tsetse host is yet to be clearly defined. *Sodalis* does not appear to have a pathogenic phenotype in tsetse, even increasing tsetse longevity in some studies [57], and does not stimulate a full immune response in the tsetse [258]. However, *Sodalis* has retained a type III secretion system (TTSS) [105]: TTSS

883 deliver effector proteins from bacteria into eukaryotic cells, are commonly used by
884 pathogenic bacteria for modulation of the host immune system and are often present
885 on pathogenicity islands associated with mobile genetic elements [259]. The *Sodalis*
886 TTSS is similar to those of pathogenic bacteria *Yersinia enterocolitica* and
887 *Salmonella*, but has modified or lost some of the effectors associated with
888 pathogenicity, retaining the needle-like TTSS structure on the cell surface for
889 infecting host cells [60]. *Sodalis* has also retained a complete set of genes encoding
890 a flagellum apparatus and expresses these genes in immature tsetse, despite no
891 evidence of flagellar motility *in vitro*. This is hypothesised to allow *Sodalis* to colonise
892 multiple tissues of the tsetse host, and in particular to facilitate infection of the milk
893 secretions for vertical transmission [60].
894
895 While the effect of *Sodalis* on tsetse fecundity is still not clearly defined [224], the
896 presence of *Sodalis* in tsetse appears to correlate positively with trypanosome
897 infection, suggesting a role for *Sodalis* in increasing tsetse susceptibility to
898 trypanosome infection [55]. The broad tissue tropism of *Sodalis,* and particularly its
899 presence throughout the digestive system, allows it to interact with trypanosomes at
900 multiple points in the parasite life cycle: *T. congolense* is found at various stages in
901 the proboscis, foregut, midgut and proventriculus, *T. vivax* in the proboscis and
902 oesophagus, and *T. brucei* in the foregut, midgut, proventriculus and salivary glands
903 [260, 261]. The positive correlation between *Sodalis* and trypanosome infection has
904 been observed in multiple field studies of tsetse, with different tsetse and
905 trypanosome species [55]. *Sodalis*-mediated increased susceptibility to
906 trypanosomes in tsetse has also been observed experimentally: lab-bred tsetse with

907 *Sodalis* removed using streptozotocin, which does not affect *Wigglesworthia*, were
908 much less likely to contract established trypanosome infections [262].
909
910 However, the mechanism by which *Sodalis* increases trypanosome infection rates in
911 tsetse is yet to be determined. Currently, the prevailing hypothesis is that *Sodalis*
912 indirectly reduces the activity of tsetse-produced trypanocidal lectins; *Sodalis*
913 produces chitinase which breaks down chitin in the tsetse peritrophic membrane to
914 produce *N*-Acetylglucosamine (GlcNAc), the primary source of carbon for *Sodalis*
915 growth [242]. GlcNAc inhibits trypanocidal lectins, increasing the tsetse fly
916 susceptibility to trypanosome infection [210, 263, 264]. When adult tsetse emerge
917 from the pupa their guts still contain the remains of their last larval milk meal
918 containing maternally-inherited *Sodalis* [214]. *Sodalis* populations increase within the
919 pupa, producing an accumulation of factors including GlcNAc in the newly-eclosed
920 teneral fly. The GlcNAc concentration in teneral flies is sufficient to inhibit tsetse
921 lectins and make the host susceptible to trypanosome establishment from the first
922 blood meal, perhaps partially explaining the increased trypanosome susceptibility
923 observed in teneral tsetse [264] (see Section 1.5.2).
924
925 *Sodalis* has several attributes that make it a promising microorganism for use in
926 tsetse and trypanosome control strategies. Revisiting the important criteria for
927 bacteria-based vector control strategies discussed in Section 1.4, the best bacteria
928 to use in VBD control strategies are amenable to culture *in vitro*, have a
929 demonstrable ability to infect the target vector without causing excessive mortality,
930 and can be genetically manipulated to express control factors such as modification of
931 the host environment or destruction of the pathogen. *Sodalis* is an ideal candidate for

use in a paratransgenesis-based tsetse control strategy: its broad tissue tropism allows contact with trypanosomes at multiple stages of their life cycle [104], is both maternally and paternally transmissible [229, 257], and can be cultured and genetically modified [244, 245]. A recombinant strain of *Sodalis* expressing functional anti-trypanosome Nanobodies® has been used to deliver the Nanobodies to tsetse in an *in vivo* lab-based proof of concept study [265, 266]. While promising, the effector molecule was only stable within the tsetse host when its native *Sodalis* population was significantly suppressed, indicating the need for further study.

1.5.3.4 The peripheral microbiome of tsetse

The tsetse microbiome is dominated by the above three bacteria, often referred to as the *Wigglesworthia-Sodalis-Wolbachia* dogma; *Wigglesworthia* on its own constitutes between 71-99% of the bacterial community in individual tsetse [228]. The difficulty of detecting bacteria at low prevalence has prevented the exploration of peripheral bacteria in the tsetse microbiota until recently. A variety of bacteria commonly associated with insects have since been discovered in tsetse, including *Acinetobacter, Enterococcus, Enterobacter, Sphingobacterium, Lactococcus, Staphylococcus, Klebsiella, Rickettsia* and a novel species *Serratia glossinae* (see Table 1-4) [228, 255, 267-270]. *Spiroplasma* has recently been identified as a potential tsetse symbiont: with a high larval titre reducing throughout adulthood, indicating maternal transmission, *Spiroplasma* may play a role in enhancing larval fitness and could be a target for bacteria-based tsetse control [228].

Table 1-4 - Peripheral microbiota observed in wild tsetse flies (*Glossina* sp.), excluding the three key symbionts *Wigglesworthia*, *Sodalis*, *Wolbachia*.

MORSITANS		**PALPALIS**			**FUSCA**
G. palidipes [271]	*G. m. morsitans [272]*	*G. f. fuscipes [255, 273]*	*G. p. palpalis [268, 274]*	*G. p. gambiensis [228, 269]*	*G. fuscipleuris [271]*
Acinetobacter	*Acinetobacter*	*Arthrobacter*	*Acinetobacter*	*Klebsiella*	*Bacillus*
Agrococcus	*Aquabacterium*	*Bacillus*	*Chryseo-bacterium*	*Serratia*	*Exiguo-bacterium*
Anaerococcus	*Burkholderiales*	*Burkholderia*	*Enterobacter*		*Kocuria*
Arthrobacter	*Cloacibacterium*	*Erwinia*	*Enterococcus*	*G. pallicera [268]*	*Microbacterium*
Bacillus	*Comamonas*	*Exiguobacterium*	*Lactococcus*	*Acineto-bacter*	*Sinomonas*
Cloacibacterium	*Coryne-bacterium*	*Lysinibacillus*	*Providencia*	*Enterobacter*	*Staphylococcus*
Enhydrobacter	*Delftia*	*Morganella*	*Sphingo-bacterium*	*Pseudo-monas*	
Enterobacter	*Ellin6067*	*Paenibacillus*	*Staphylo-coccus*		*G. nigrofusca [268]*
Erwinia	*Halomonas*	*Pantoea*			*Enterobacter*
Exiguobacterium	*Mycoplasma*	*Providencia*	*G. tachinoides [228]*		
Pseudo-xanthomonas	*Pseudomonas*	*Pseudomonas*	*Spiroplasma*		*G. medicorum [228]*
Pseudomonas	*Serratia*	*Rummelibacillus*			*Klebsiella*
Serratia	*Shewanella*	*Serratia*			
Sphingo-bacterium	*Rickettsia [275]*	*Solibacillus*			*G. brevipalpis [271]*
Staphylocccus	*G. submorsitans [228]*	*Spiroplasma*			*Bacillus*
Xylella	*Klebsiella*	*Staphylococcus*			
	Rickettsiella	*Terribacillus*			

1.6 Sand flies, leishmania and associated microbiota

1.6.1 The leishmaniases

Leishmaniasis is a complex of NTDs with a global distribution, caused by at least 20 species of protozoan parasites in the *Leishmania* genus (Class: Kinetoplasta, Order: Trypanosomatida) [9, 276]. *Leishmania* parasites and vector species are grouped by whether they are present in the Old World (Mediterranean, Middle East, Africa and India) or the New World (the Americas). Leishmaniasis is associated with poverty, malnutrition, population displacement, and poor infrastructure [277]. The true extent of *Leishmania* parasite infections is difficult to estimate due to the lack of public health records in endemic areas [278], but over 1 million new cases of leishmaniasis are estimated to occur each year [277]. All forms of leishmaniasis are vectored by the bite of female sand flies (Family: Psychodidae, Subfamily: Phlebotominae).

Leishmaniasis is divided into three main clinical forms: cutaneous, mucocutaneous and visceral leishmaniasis (also known as kala-azar). Cutaneous leishmaniasis (CL) is the most common form with an estimated 0.6-1 million cases annually worldwide; around 85% of these cases were concentrated in Afghanistan, Algeria, Bolivia, Brazil, Colombia, Iran, Iraq, Pakistan, Syria and Tunisia in 2018 [277]. In the Old World CL is caused mainly by *Leishmania major, L. tropica* and *L. aethipica*; New World CL is mainly caused by *L. amazonensis, L. mexicana, L. braziliensis* and *L. guyanensis* [279]. CL is not life threatening but leads to severe disfiguration, social stigma, and lasting psychological effects, causing the loss of nearly 45,000 DALYs in 2016 [276, 280-282]. Mucocutaneous leishmaniasis is caused by a strong immune

response to CL parasites, causing severe lesions on mucous membrane surfaces such as the nasal passages and lips [283], and can be fatal [276].

1.6.1.1 Visceral leishmaniasis

The most lethal form of leishmaniasis, visceral leishmaniasis (VL), is caused by *L. donovani* in Asia and Africa and *L. infantum* in the Mediterranean, the Middle East, Central Asia, South America and Central America [276] (see Figure 1-6). Characterised by a persistent but irregular fever and splenomegaly (enlarged spleen), weight loss and acute malnutrition are also common symptoms of VL. On the Indian subcontinent patients often present with skin hyperpigmentation, which

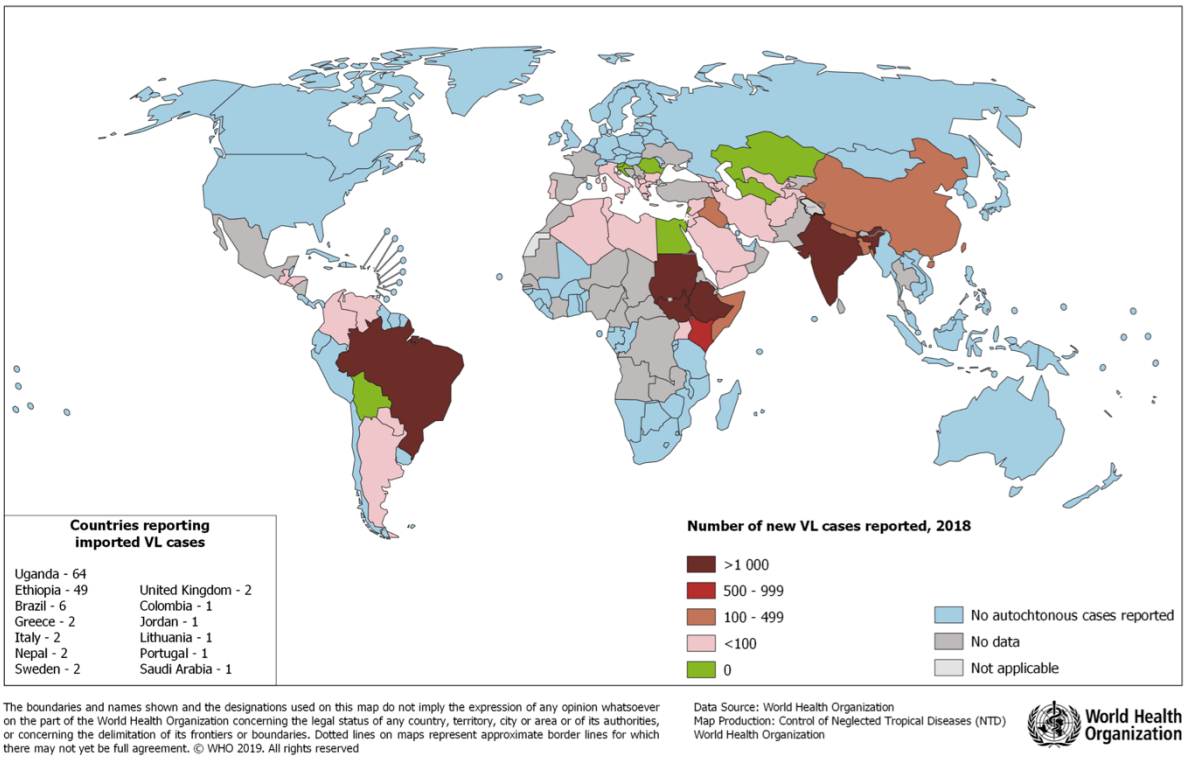


Figure 1-6 - Countries which are endemic for visceral leishmaniasis and the number of reported visceral leishmaniasis cases reported to the World Health Organization in 2018.

has led to the name *kala-azar* for VL, which roughly translates from Hindi as “the black fever” [276, 284].

VL onset can either be acute, occurring within two weeks of an infected bite, or can incubate for up to 8 months before symptoms show; if symptoms are present VL is fatal within two years, with patients dying either from secondary infections or severe anaemia [276]. Asymptomatic carriers of VL parasites are common, with up to 50 asymptomatic carriers per symptomatic person in some areas [285], although it’s possible for an asymptomatic person to develop symptoms years later if they become immunocompromised [286]. Co-infection with HIV is also common, with up to 18% of VL patients in Ethiopia also infected with HIV [287], due to the diseases using the same pathological mechanisms synergistically [288]. This presents a major challenge for VL control – HIV outbreaks were responsible for the re-emergence of VL in southern Europe in the 1990s [276], and in one Indian study half of VL and HIV co-infected patients were unaware of their HIV status [289]. Overall, just under 15,000 deaths registered in 2018 were caused by VL, with an estimated total of 20,000 – 40,000 deaths [4, 277].

Leishmania donovani, the causative agent of VL in Asia and Africa, can also cause a unique skin condition that occurs after VL has been treated, known as post-kala-azar dermal leishmaniasis (PKDL) [276]. Occurrence of PKDL is common in East Africa, with 50% of VL patients developing PKDL within one year of VL treatment, but rarer in Asia, where it is seen in 5-10% of VL patients. PKDL is of epidemiological significance, as lesions are infectious to sand flies and can remain so for decades [290], and can therefore be reservoirs for VL between epidemics [291].

Asymptomatic infections may also have an impact on epidemiology; a large proportion of people infected with *L. infantum* or *L. donovani* will not show any symptoms despite testing positive for the parasite [292].

VL can also afflict animals, predominantly domestic animals such as dogs and cats, in whom it causes visceral and cutaneous symptoms [293]. *L. donovani* is anthroponotic – the disease is maintained by human to human transmission, with a minimal impact of human-animal transmission [294]. *L. infantum*, however, is zoonotic – animals are a major source of human infection, in particular dogs, which can contract canine visceral leishmaniasis or be asymptomatic carriers, and are considered to be the main zoonotic reservoir of *L. infantum* [295].

The primary diagnostic tool for VL is the microscopic identification of the parasite in blood or tissue; the gold standard tissue for this is a splenic aspiration, although the procedure causes haemorrhage in one in every 1,000 patients [276, 296]. VL treatment involves intravenous medication with severe side effects [276, 297]; no vaccine for human leishmaniasis has been developed, although there is a commercially available vaccine against *L. infantum* in dogs [298].

The WHO road map for NTDs 2021-2030, which sets out the objectives to be met over the next decade in order to control NTDs, has targeted VL for elimination as a public health problem in 85% of endemic countries by 2030 [299]. Leishmaniases are particularly difficult to control due to the large number of asymptomatic infections and PDKL cases that can act as disease reservoirs in the absence of clinical cases [276]. Due to the lack of effective surveillance and medicine for leishmaniases

control strategies rely on control of the sand fly vector [300]. The current national control program for VL in India relies primarily on indoor residual spraying using DDT, but has issues with consistent underspraying of households, shorter than expected residual action, and growing DDT resistance [301]. Vector control has dramatically reduced VL cases in India over the last decade [4], but innovative vector control strategies are required to complete the last stage of the WHO elimination target [302].

1.6.2 The sand fly vector

There are almost a thousand species of sand fly (Figure 1-7) in six genera, but only two genera are capable of vectoring leishmaniasis: *Phlebotomus* and *Lutzomyia. Phlebotomus* sand flies are found in the Old World, and *Lutzomyia* are confined to the New World, with no overlap of distribution between the two genera. They are the only known vectors of leishmaniasis, and are also responsible for vectoring *Bartonella bacilliformis* (the causative agent of bartonellosis, or Carrión's disease) and viruses responsible for sand fly fevers [303]. Figure 1-8 gives a full description of the *Leishmania* life cycle within the sand fly, which takes place entirely within the midgut. The adaptability of sand flies to changing environments has contributed to the increasing urbanisation of sand fly-vectored diseases [304].

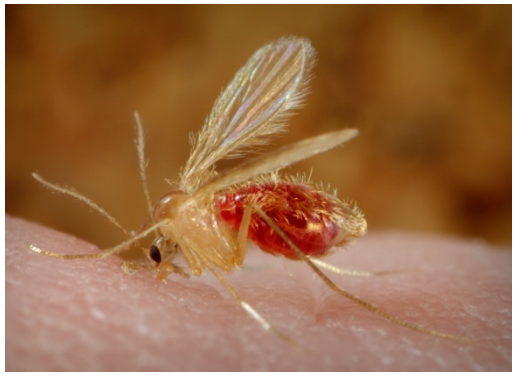
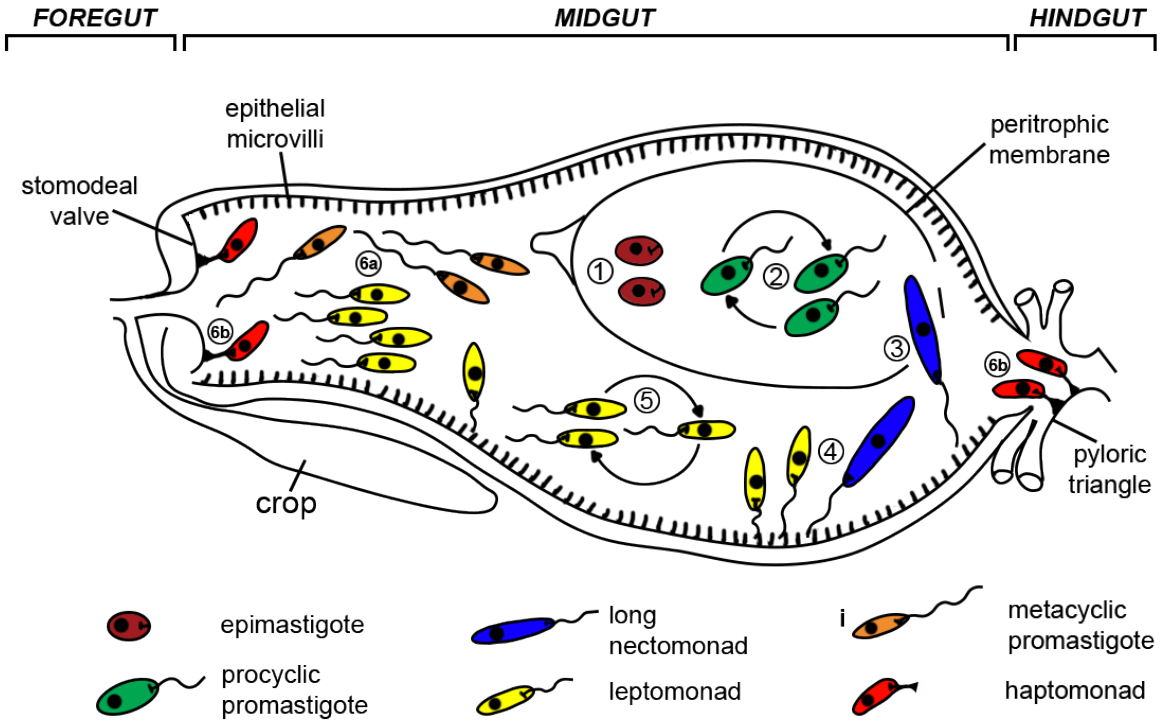


Figure 1-7 - A *Phlebotomus papatasi* sand fly after ingesting a blood meal. Image credit: CDC/Frank Collins, Creative Commons Licence.



1) *Amastigotes* are ingested with a blood meal into the midgut, where they are contained in the peritrophic membrane.
2) Amastigotes differentiate into *procyclic promastigotes* and replicate.
3) Procyclic promastigotes differentiate into *long nectomonads*, which escape through the posterior opening in the peritrophic membrane created by sand fly enzymes.
4) Long nectomonads attach to the midgut microvilli and differentiate into short nectomonads called *leptomonads*
5) Leptomonads replicate and build up in the midgut lumen, then either:
6a) Differentiate into infective *metacyclic promastigotes*; or
6b) Differentiate into *haptomonads* and attach to the chitin lining of the stomodeal valve or pyloric triangle, depending on the *Leishmania* species
Replicating leptomonads obstruct the midgut and haptomonads damage the stomodeal valve, causing reflux of the metacyclic promastigotes during the sand fly's next blood meal, infecting a new host.

Figure 1-8 - A generalised life cycle of *Leishmania* parasites within the sand fly vector. Adapted from Dostálová & Volf (2012).

Sand flies a more classical insect lifecycle than tsetse. Both males and females feed on sugary secretions from plants, but female sand flies will take blood meals as supplementary nutrition necessary for producing a batch of eggs. Eggs are laid in small cracks in masonry or the ground, 30-70 eggs per oviposition, and will hatch after 4-20 days depending on the species and the ambient temperature. Larvae

develop through four instars over 20-30 days, living in leaf litter and feeding on organic matter including fungi, animal faeces and rotting vegetation. Fourth instar larvae can diapause overwinter in some areas and pupate in the ground for 6-13 days before an adult sand fly emerges. *Phlebotomus* and *Lutzomyia* females feed on a wide range of mammals and biting usually occurs outside and at dawn and dusk, or overnight [303]. Adult sand flies are weak fliers, rarely dispersing more than 100m from their breeding places, and even light wind can prevent flying and so prevent biting. Additionally, female sand fly mouthparts are too short to pierce clothing, therefore biting can be avoided by keeping skin covered and using fine mesh bed nets [303, 305]. During the day several important leishmaniasis sand fly species rest in human dwellings in cracks in the walls, including *Phlebotomus papatasi, P. argentipes* and *Lutzomyia longipalpis*; other species rest outside in vegetation or cattle sheds [303].

1.6.3 The microbiota of sand flies

Although sand fly microbiomes are less well studied than other insect vectors like tsetse or mosquitoes, sand flies are known to have a less well-defined, more fluid microbiome than tsetse flies, with higher biodiversity and no known primary symbionts. Understanding the interactions between the host, the parasite and the native microbiota may lead to the development of new control strategies [306]. The *Leishmania*-sand fly system has the potential to be interrupted using paratransgenesis of native bacteria to control leishmaniases, as described in Section 1.6.2. This approach requires the identification of suitable microorganisms in the vector microbiome.

The composition of the sand fly microbiome is largely determined by the environment: sand flies lay eggs in tree trunks, animal burrows or soil, where hatched larvae will feed on the organic matter present [307-309]. Ingested and environmental microbes colonise the larvae, persisting to adulthood to form the primary sand fly microbiota which is supplemented by the adult diet of sugary plant secretions and, in females, blood [306, 310-312]. The sand fly microbiome can change based on an adult female's diet throughout her life history: microbial diversity has been shown to decrease after blood feeding, recovering once the blood meal was fully digested, although overall bacterial abundance increased due to the nutrient-rich blood meal [311-314].

Figure 1-9 illustrates the wide diversity of bacteria found so far in *Leishmania*-vectoring sand flies. Many of the genera in Figure 1-9 are commonly found in soil, including *Enterobacter, Pseudomonas* and *Bacillus* [315, 316]. Other genera are commonly associated with insects as beneficial or pathogenic symbionts including *Serratia*, which contains both the aphid secondary symbiont *S. symbiotica* [317] and the insect pathogen *S. marcescens* [318]. The reproductive manipulator *Wolbachia* has also been documented in two species of *Phlebotomus*. Plant-associated bacteria are present in most sand fly species, including the nitrogen-fixing bacteria *Rhizobium* and *Bradyrhizobium* [319, 320], and the plant pathogen *Ralstonia* [321].

Despite the evidence suggesting that the sand fly environment and food sources are the primary factors dictating the microbiome composition [322], field-caught and lab-reared sand flies of the same species have similar microbiomes, implying other factors contribute to bacterial diversity in sand flies [323-326]. Several genera in

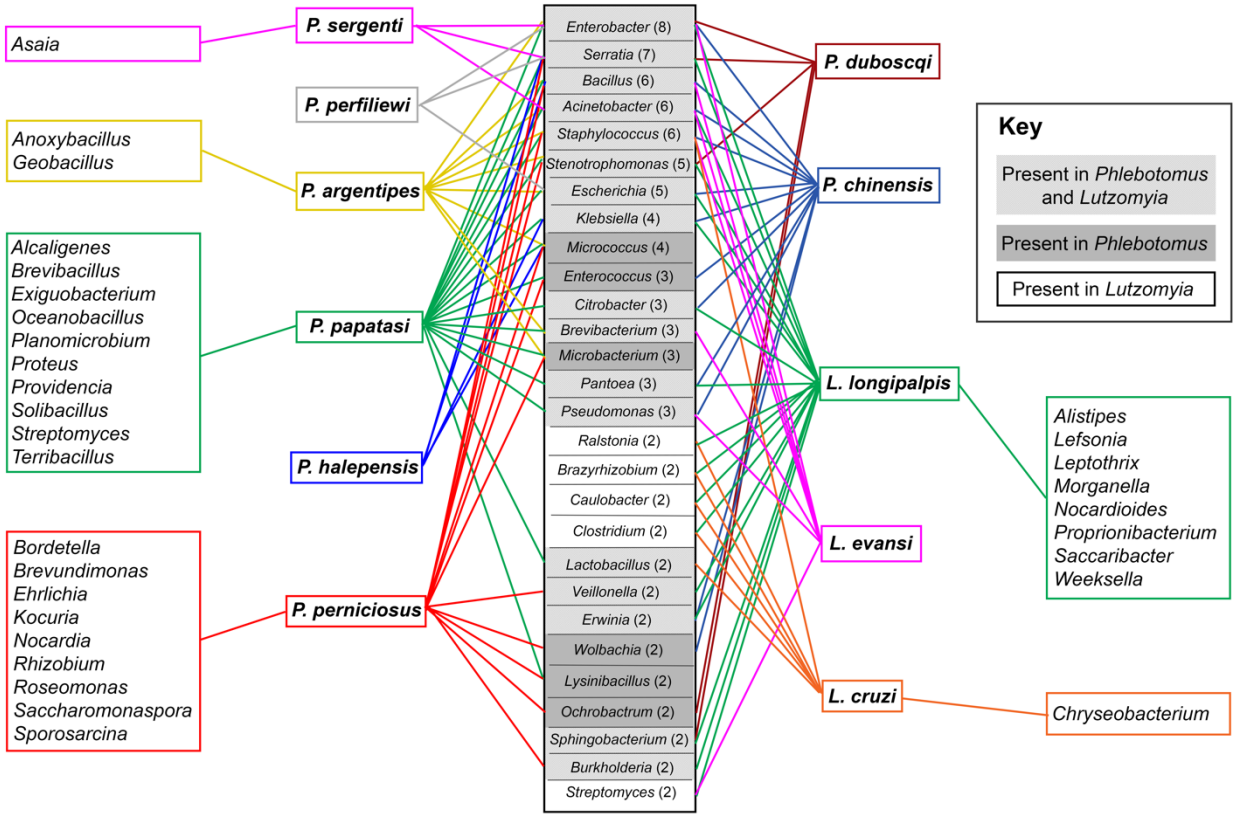


Figure 1-9 - Bacterial genera previously discovered in sand fly species known to vector *Leishmania sp.*, described using a mixture of culture-dependent and -independent methods. Genera in the middle column were found in multiple sand fly species (number of sand fly species in brackets). Genera found in only one sand fly species are shown in coloured boxes. Data from Telleria et al. (2018).

Figure 1-9 are found in both *Phlebotomus* and *Lutzomyia* sand flies, despite wide geographical distances between their habitats, suggesting there may be a common role of these bacteria in sand flies [326].

The effects of the sand fly microbiome on its competence as a *Leishmania* vector are complex and not fully understood. The presence of microbial communities in sand flies appears to impact sand fly development: *L. longipalpis* fed raw rabbit faeces were more likely to lay eggs than flies fed on sterilised faeces [327]. Sand fly guts contain enzymes for digesting microbial cell walls, suggesting that bacteria

1131 themselves may be used as a nutrient source [328]. The microbiome of sand flies
1132 can also stimulate and prime the immune system against *Leishmania* parasites:
1133 production of immune effector molecules such as reactive oxygen species is
1134 stimulated by native microbiota, helping to control *Leishmania* infection
1135 establishment in the sand fly midgut [329, 330]. A competitive relationship between
1136 *Leishmania* and the bacteria *Serratia* has also been observed, whereby infection
1137 with one reduces the ability of the other to infect the sand fly host [330]. However,
1138 when sand flies were treated with antibiotics to remove the native microbiota before
1139 *Leishmania* exposure the parasites were unable to replicate or develop in the sand
1140 fly [89, 311]. The leishmaniacidal effect of increasing the sugar concentrations used
1141 to maintain the axenic (bacteria-free) sand fly colonies in these studies suggests that
1142 the microbiota act as an osmolar buffer for *Leishmania* parasites in the midgut [89,
1143 306, 311].
1144
1145 Previous studies have identified candidate bacteria in sand flies for paratransgenic
1146 control of leishmaniasis. Hillesland *et al.* [322] identified *Bacillus megaterium* and
1147 *Brevibacterium linens* in wild Indian *P. argentipes* by aerobic culture; both bacteria
1148 are used extensively in industry already, indicating their potential for scalable use in
1149 paratransgenesis control strategies. *B. megaterium* is used as a biofertilizer
1150 extensively in India, promoting growth and reducing disease in plants [331]. *B. lin*ens
1151 is used to ripen cheese [332]; many sand flies in the Hillesland study were collected
1152 from areas with household dairy operations. Hillesland *et al.* proposed introducing
1153 paratransgenic *B. megaterium* and *B. linens* into sand fly breeding habitats in large
1154 numbers to colonise sand fly guts and deliver effector molecules [322]. This could be
1155 aided by the natural production of methanethiol and isovaleric acid by *B. linens*,

which have been demonstrated to attract mosquitoes [333, 334]. Other suggested methods of application of paratransgenic bacteria to sand flies include as a cattle feed probiotic, or as biofertilizer in areas where this is already done with wild-type *B. megaterium* [322]. Several proof-of-concept studies have been conducted or are underway, including the introduction of sand fly-associated *B. subtilis* expressing green fluorescent protein into *P. argentipes* [335], and identification of potential effector molecules such as single-chain antibodies [336]. However, more work characterising the sand fly microbiome, and the potential for interactions between paratransgenic and native bacteria within the sand fly host, is needed.

1.7 Methodologies used to study insect microbiota

In the past, insect microbiome research relied on microscopic visualisation of bacteria within insect tissues, characterising bacteria by their phenotype [337]. Primary insect symbionts have evolved to become reliant on the host environment, losing the metabolic plasticity necessary to be readily cultured outside of the host; as a result, none have yet been cultured axenically [338]. Without the ability to culture important insect microbiota understanding their function within the host relied on experimental evidence, and manipulation of symbiotic bacteria was impossible. The advent of genomic technology marked the beginning of in-depth insect microbiome studies [93].

1.7.1 Culture-dependent methods

The estimated proportion of total bacteria that are currently culturable *in vitro* is around 1% [246]. On top of this, insect symbionts are extremely difficult to culture: a lack of selection pressure for genes necessary for life outside of the host environment results in reduced, specialised genomes (which will be discussed further in Chapter 3). As a result, insect symbionts are fastidious in their nutritional and environmental requirements, slow growing and easily outcompeted by contaminants [338]. Despite these difficulties several facultative symbionts have been successfully cultured *in vitro,* either in axenic culture or within cultured insect cell lines - Table 1-5 lists all facultative symbionts that have been successfully cultured outside of insect cells.

Table 1-5 - Insect-associated facultative symbionts that have been successfully cultured in vitro in cell-free media. Ca – Candidatus.

Bacteria species	Host species	Year	Reference
Rhodococcus rhodnii	*Rhodnius prolixus*	1992	[154]
Sodalis glossinidius	*Glossina sp.*	1993	[244]
Ca. Arsenophonus arthropodicus	*Pseudolynchia canariensi*	2006	[339]
Corynebacterium sp.	*Triatoma infestans*	2008	[340]
Acetobacter tropicalis	*Bactrocera oleae*	2009	[341]
Asaia sp.	*Anopheles sp. and Aedes sp.*	2010	[342]
Arsenophonus nasoniae	*Nasonia vitripennis*	2010	[343]
Ca. Sodalis melophagi	*Melophagus ovinus*	2012	[344]
Burkholderia sp.	*Blissus insularis*	2016	[345]
Hamiltonella defensa	*Acyrthosiphon pisum*	2017	[346]
Serratia sp.	*Anopheles sp. and Aedes sp.*	2017	[347]
Spiroplasma poulsonii	*Drosophila sp.*	2018	[348]

The insect gut microbiome is still commonly studied using bacterial culture [349]. While studies using only culture-dependent methods will not give a complete picture of bacterial diversity they can identify bacterial strains that can be readily isolated and studied in depth, as well as used in biotechnology such as paratransgenic control strategies (see Section 1.4). Bacterial culture is readily available in settings with reduced resources, requires less formal training and specialist equipment, and is less expensive than molecular methods [350]. However, it has several drawbacks in addition to the lack of diversity resolution: fresh specimens must be used requiring rapid processing of samples, with any delay or change in the insect environment causing large changes in the gut microbiome. Culture media and conditions must mimic the natural insect environment as closely as possible, which can make the process more difficult and expensive. The slow growth of many insect-associated bacteria increases the study time required and increases the chance of contamination. While culture-dependent techniques are necessary for

biotechnological applications they are less useful for initial exploration of insect microbiota and are of little use for the study of primary symbiosis.

1.7.2 Culture-independent methods

Over the last 20 years molecular studies of insect-associated bacteria have dramatically increased our understanding of their diversity and their effect on the host, as well as changing our perception of the architecture of bacterial genomes [50]. The development of cheaper whole-genome sequencing (WGS) has resulted in readily-available genomes for symbionts that can't be cultured *in vitro*, allowing for in depth study of their metabolic processes and evolution [227, 351, 352]. PCR can be used to amplify genes for detection and allele identification, and to study differential gene expression of insect-associated bacteria within the host under experimental conditions [349]. Next-generation sequencing (NGS) using Illumina, PacBio or Oxford Nanopore platforms allow high-throughput, rapid sequencing of gene fragments, which can be used for WGS or for sequencing of a specific amplicon. Detailed methodologies for both WGS and amplicon sequencing using NGS will be covered in Chapters 3 and 5 respectively. Briefly: DNA is extracted from samples often using a column and centrifuge-based kit, the target amplicon or whole fragmented genome (known as shotgun sequencing) is amplified using PCR, tagged with a barcode sequence to identify the sample, before samples are pooled and run on a sequencer. Sequences are then demultiplexed using the barcodes into their original samples, error rates are calculated and corrected, and sequences are used in downstream analyses depending on the experiment.

A common use of amplicon sequencing in the study of insect microbiomes uses the 16S gene, which codes part of the small ribosomal subunit of bacteria and archaea, to determine the taxonomic composition of a sample. This approach gives better sequencing depth (and consequently can identify rarer species) than shotgun sequencing. The use of 16S as a genetic marker was developed to determine the phylogenetic origins of prokaryotes, and led to the first description of the three domains of life [353, 354]. The 16S gene is present in all known bacteria and archaea, and consists of nine hypervariable regions (V1-V9) enclosed by highly conserved regions essential to gene function [355, 356]. Genetic drift in the hypervariable regions happen at a predictable rate, allowing their use as a measurement of the time since divergence between two bacteria [353]. The 16S sequences of unknown bacteria can be compared to databases of known sequences, allowing the identification of multiple bacterial species from a single sample. Sequencing of the entire ~1,500 base-pair (bp) 16S gene gives the best taxonomic resolution, but in order to increase sequencing depth one to four variable regions are generally used, with primers designed to match the conserved regions on either side.

16S sequencing reads can be clustered into operational taxonomic units (OTUs), whereby sequences with a set similarity threshold (often ≥97%) are clustered together and treated as being the same, reducing the impact of sequencing errors on the taxonomic classification of sequencing reads [357]. Despite the common threshold of ≥97%, clusters of OTUs with 99% similarity also show significant diversity, with potentially important phenotypic effects on ribosome functionality

[358]. However, as sequencing technologies have improved the use of OTUs is no longer the most accurate option. New methods control sequencing errors sufficiently that each read can be taken as accurate, allowing the clustering of reads simply by amplicon sequence variants (ASVs), whereby a single nucleotide difference between reads will classify them as separate ASVs. This gives better resolution of the bacterial taxa in a sequenced sample, without drastically increased computing costs [359].

The main drawback of using the 16S gene for bacterial identification is the lack of comparability of methods between studies. This method of classification hinges on being able to compare sequences within and between studies, but there are no widely used guidelines for methodology, which can have a significant impact on the results [360]. There are several points in the methodology of 16S sequencing for bacterial community analysis where diversity of methods can reduce comparability between studies. Samples must be treated and stored in the same way, for example frozen at -80°C or stored in ethanol, as this has been documented to affect the ratios of certain bacteria [361-363]. DNA must be extracted with the same protocol for sequences to be directly comparable between samples as some bacterial species are less likely to be effectively lysed with certain protocols [364-372]. The choice of variable region to sequence, including the length of the sequenced fragment, can have a significant impact on bacterial identification: longer fragments will give better resolution, but this introduces methodological constraints based on the sequencing platforms used [362, 373-378]. There is no current consensus on which 16S variable regions are the most appropriate for insect microbiome studies. Increased number of cycles in library preparation PCRs can increase bacterial richness but has no

significant effect on community structure and can increase the rate of chimeras [379-381]. Sequencing error rates are high when sequencing amplicons of high similarity like the 16S gene due to overclustering of the sequencing chip, but this can be mitigated by overlapping paired reads and adding a completely different type of DNA (e.g. viral) to the run as a heterogeneity spacer to improve diversity [382, 383]. There are also several different bioinformatics pipelines used to analyse data, across different sequencing platforms, and different databases of 16S sequences are used to assign taxonomy to sequence reads. All of these parameters must be optimised depending on the study – this process will be discussed in depth in Chapter 5.

Since the advent of high-throughput sequencing for microbiome studies the issue of contamination has been widely discussed, but nonetheless is often neglected during experimental design. In both 16S amplicon sequencing and shotgun sequencing methods for bacterial identification, and particularly in studies where low-biomass samples are used, reagent and laboratory contamination often make a large enough proportion of reads to mask true microbiota in the samples [384]. Accounting for contaminating bacteria by sequencing mock communities and reagent-only extraction controls alongside each sequencing run of true samples is essential to mitigate for this issue. Despite controls being recommended for all studies there are still limited tools to use these controls effectively for removal of contaminants *in silico*. Many studies simply subtract all contaminants identified in negative controls from the samples, but as contaminants tend to be soil microbiota this can remove genuine taxa [384]. A novel method of identifying contaminants by examining the prevalence of taxa in negative controls compared to true samples will be described in Chapters 5 and 6.

1.8 Aims and objectives

The aim of this work is to explore the microbiota of two insect vectors of kinetoplast parasites: tsetse flies (*Glossina sp.*), which transmit *Trypanosoma sp.*; and phlebotomine sand flies (*P. argentipes)*, which transmit *L. donovani* in the Indian subcontinent. Both parasites are the causative agents of NTDs, together causing the loss of millions of DALYs (Table 1-2), and both are amenable to microbiota-based control strategies. Despite the close relatedness of *Leishmania* and *Trypanosoma* parasites the tsetse and sand fly vectors are very different in terms of life history, nutritional requirements, morphology and behaviour, resulting in contrasting microbiome structures. Tsetse have a small, well-studied and functional microbiome including a primary endosymbiont, several facultative endosymbionts and very few peripheral bacteria (Section 1.5.3). In contrast, in-depth study of the microbiome of the much smaller sand flies has only been possible since the improvement of genetic technology in the last few decades, and has described a complex network of transient bacteria with little overlap between sand fly species and no known symbionts (Section 1.6.3 and Figure 1-9). By studying these insects in parallel we can not only decipher the specifics of each microbiome that will allow for its manipulation for vector control, but also increase our understanding of the nature of insect microbiome form and function.

These aims will be achieved with the following experiments:

1. Collection of prevalence data of *Sodalis, Spiroplasma, Serratia, Wolbachia* and *Trypanosoma* in wild tsetse populations, looking for any correlation between *Sodalis* and *Trypanosoma* incidence. This will be done by trapping

tsetse in northern Uganda and using PCR probes to detect each microorganism, as well as culture techniques to isolate a wild strain of *Sodalis* for further study (Chapter 2).

2. Comparative genomics of a strain of *Sodalis* that has been cultured *in vitro* for 10 years with an ancestral strain, to assess the evolutionary trajectory of the symbiont (Chapter 3).
3. Use of NGS transcriptomics to identify host and symbiont genes involved in the refractoriness or susceptibility of *Sodalis*-infected tsetse to experimental infection with *T. congolense* (Chapter 4).
4. Development of an NGS amplicon sequencing pipeline to accurately describe the microbiome of individual *P. argentipes*, to assess the impact of small-scale geographical variation on their microbiome and the differences between individual microbiomes (Chapters 5 and 6).

Chapter 2 Prevalence of native bacteria associated with wild tsetse fly populations in Uganda

2.1 Abstract

Two subspecies of *Trypanosoma brucei* cause chronic and acute human African trypanosomiasis; these two subspecies only coexist in Uganda, which also suffers severe economic and social losses caused by animal trypanosomiasis in cattle. The vectors of trypanosomiasis in Africa are tsetse flies (*Glossina sp.*). Tsetse have a limited but functional microbiome, including the secondary endosymbiont *Sodalis glossinidius* which is believed to increase tsetse susceptibility to trypanosomiasis but only studied using isolates from lab-reared tsetse. The recent discovery of further tsetse-associated bacteria with unknown roles within the host includes: *Spiroplasma*, which seems to have a negative association with trypanosome infection; *Wolbachia;* and *Serratia*. Accurate knowledge of the microbiota of tsetse and their association with trypanosome infection is required to ensure the efficacy of microbiota-based control strategies, and potentially to create novel control strategies. In this chapter, a pilot study of wild-caught tsetse flies in Uganda is conducted to assess the presence and trypanosome association of *Sodalis*, *Spiroplasma, Wolbachia* and *Serratia* in the three sympatric tsetse fly species. Due to COVID-19 disruption of lab work a limited sample set of 57 samples was analysed for the presence of three bacteria and trypanosomes: *Sodalis* was found in 39% of flies, *Wolbachia* in 1.7%, *Spiroplasma* in 88% and trypanosomes in 12%, with no significant correlation between any of the four organisms to indicate interactions between them. While more data is needed to elucidate the relationships between these bacteria and trypanosomes, the high

frequency of *Spiroplasma* in the region and its previous identification as a potential disruptor of trypanosome infection in insects makes this a significant dataset for Ugandan trypanosome control strategies.

2.2 Introduction

2.2.1 Trypanosomiasis in Uganda

The two types of HAT are geographically separated by the Rift Valley; both diseases occur together only in Uganda (Figure 2-1). HAT is found in two separate foci in Uganda, with *rhodesiense* HAT found in the south and *gambiense* HAT in the north; these foci do not overlap and so the two diseases are not truly sympatric, but the foci are expanding towards each other and in 2005 were only 150km apart [385]. This expansion is caused by the influx of refugees from South Sudan carrying *gambiense* HAT further into Uganda, and movement of cattle northwards from the *rhodesiense* HAT focus in southern Uganda [385]. The two diseases are predicted to eventually overlap and become sympatric in Uganda, complicating the diagnosis and treatment of both [385-387].

In most of the tsetse-infested regions of Uganda cattle-based agriculture is the primary source of income and is greatly restricted by AAT [388, 389]. In endemic areas cattle AAT infection rates are almost 100% and alongside tick-borne diseases are the highest

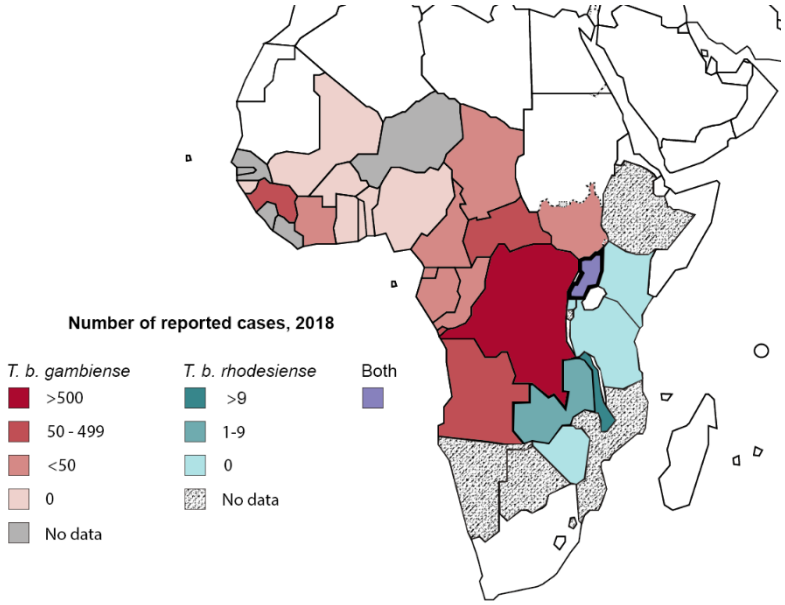


Figure 2-1 - Distribution of gambiense and rhodesiense HAT cases reported in 2018 in endemic countries. In Uganda (highlighted in bolded black), <50 gambiense HAT and 1-9 cases of rhodesiense HAT were reported. Adapted from [5].

cause of financial and production loss for cattle farmers and draughters [390, 391]. AAT in Uganda is caused by the zoonotic *T. b. rhodesiense* as well as *T. vivax* and *T. congolense* [193]. As veterinary services are fully decentralised in Uganda the costs of chemoprophylaxis and chemotherapy of trypanosomiasis in cattle are met by the farmers, constituting a major loss of income [392]. This has had the added effect of deregulating the marketing and application of trypanocidal drugs, leading to the distribution of off-brand alternatives with low efficacy against AAT which facilitate widespread drug resistance [393, 394]. The development of novel drugs for AAT is slow due to the limited financial return, and all drugs currently on the market were developed over 40 years ago, increasing the urgency of alternative AAT prevention strategies [394].

Control of the tsetse vector, either by elimination or ongoing control strategies, has been assessed as an alternative method for HAT and AAT control in Uganda. However, the high initial costs of tsetse control strategies and the ability of tsetse to reinfest large areas quickly mean that novel tsetse-trypanosome ongoing control strategies are being sought [395].

2.2.2 Tsetse microbiota in wild populations

Tsetse flies harbour a limited bacterial microbiome (see Section 1.5.3). While *Wigglesworthia* is present in all individuals of the *Glossina* genus, *Wolbachia* and *Sodalis* are found at varying rates in wild tsetse populations. *Sodalis*, which is a recently-evolved symbiont thought to increase tsetse susceptibility to trypanosome infection [60, 262, 396, 397] and has potential for use in paratransgenesis control strategies [266] (see Section 1.5.3.3), is present in 0-85% of flies within a population

[254, 255, 396, 398]. *Wolbachia* is also present in varying amounts in wild tsetse populations – for example, *Wolbachia* prevalence has been observed from 10 to 100% in *G. morsitans morsitans* (*Gmm*), in low densities in *G. fuscipes* and other *G. morsitans* subspecies, and not at all in *G. tachinoides* [239, 254, 399]. Removing the tsetse microbiome during larval development results in a structurally impaired peritrophic matrix, which lines the tsetse midgut and presents a physical and immunological barrier to trypanosome infection (see Figure 1-2), resulting in increased trypanosome infection rates of aposymbiotic adult flies [219]. The effect of a microbial community as a whole, as well as the individual members themselves, on vectorial capacity makes accurate knowledge of wild tsetse microbial partners an important aspect of understanding HAT and AAT epidemiology.

The development of accessible sequencing technology has enabled more detailed descriptions of the peripheral tsetse microbiome. Other potential tsetse symbionts have been described in wild tsetse, most notably *Spiroplasma* in *Glossina fuscipes fuscipes* (*Gff)* and *G. tachinoides* [228]. *Spiroplasma* has a broad range of hosts including arthropods, mammals and plants, living both intracellularly and extracellularly [400-402], with a range of host effects from protection against parasites [403-405] to pathogenicity [406-408] and male-killing [409-411]. A study of wild *Gff* found *Spiroplasma* infection rates between 3-34%, varying by region and season, with a non-significant negative correlation between *Spiroplasma* and trypanosome infection [273]. However, experimental infection of a laboratory *Gff* line with *Spiroplasma* and trypanosomes resulted in significantly lower trypanosome infections in tsetse with *Spiroplasma* [273]. While the role of *Spiroplasma* in tsetse flies is still unclear, no sex bias was observed in experimentally-infected *Gff*

indicating that male-killing is unlikely; the negative association between *Spiroplasma* and trypanosome infection suggests a protective role [273].

Another potential symbiont of tsetse that has been described in the last decade is *Serratia*, which are found in a range of environments from water and soil to plants and animals, including insects [412]. *Serratia marcescens* was isolated from the gut of the Chagas disease vector *Rhodnius prolixus* in which it produces the red pigment prodigiosin that is toxic to the Chagas disease causative agent *Trypanosoma cruzi* [413, 414]. Members of the *Serratia* genus have been identified as insect symbionts, inhibiting *Plasmodium* infections in *Anopheles gambiae* and improving fecundity and heat stress tolerance in aphids [317, 415]. An isolate of *Serratia* from lab-reared *Glossina palpalis gambiensis* was proposed as the new species *Serratia glossinae*; this isolate was later proposed to be a synonym of *Serratia fonticola* first isolated from water in 1979 [416, 417]. *S. marcescens* is a known pathogen of tsetse flies [418], suggesting that tsetse-associated *Serratia* may be undergoing an evolutionary change from pathogen to symbiont in a similar way to *Sodalis* [60, 251].

The microbiomes of wild and laboratory-reared tsetse have key differences, making the study of wild tsetse microbiota imperative for accurate application of microbiome-based control strategies. For example, while *Sodalis* incidence varies between and within wild populations as described above, in laboratory populations it is either entirely absent or fixed in 100% of the population [256]. *Sodalis* can be cultured readily *in vitro* and isolated from laboratory tsetse populations in which it has 100% incidence [419]. As a result, all key studies examining the *Sodalis* genome, the effect of *Sodalis* on tsetse vector competency or producing paratransgenic *Sodalis* for

control strategies use strains sourced from laboratory tsetse populations [60, 105, 265, 419-421]. As genetic drift may have produced key differences in wild *Sodalis* populations in comparison to *Sodalis* from laboratory-reared tsetse flies, which are frequently subjected to bottlenecks [422, 423], it is imperative that wild *Sodalis* is studied and compared to lab strains to ensure the validity of lab-based studies.

Tsetse flies also possess a limited peripheral microbiome, made up of largely transient infections from the environment (see Table 1-4); the peripheral microbiome of wild tsetse is not comparable with laboratory-reared tsetse which show a more limited microbial diversity [228]. Developing microbiota-based control strategies based on laboratory data may reduce their effectiveness when applied in the field, making the study of wild tsetse microbiota essential. An increasingly close association of peripheral bacteria to an insect host can evolve to become a facultative or obligate symbiosis, potentially increasing the host's vector capacity for pathogens or producing a potential target for vector control – see Section 1.4 for an overview of the role of symbionts in vector control.

2.2.3 Tsetse microbiota in Ugandan populations

Despite the importance of Uganda as a focus of trypanosomiasis, and the potential of tsetse microbiota to help or hinder control strategies, the incidence and distribution of the microbiota of wild Ugandan tsetse flies is still under investigation. In one study of *G. pallidipes* (*Gpal*) in northern Uganda *Sodalis* was seen in 8/45 flies, with other identified genera including *Serratia*, *Acinetobacter* and *Staphylococcus* present infrequently and at low titres [271]. In a larger study of *Spiroplasma* and *Wolbachia* presence in *Gff* across Uganda, *Spiroplasma* infection rates from 0 to 34% were

found in different regions depending on the season, and *Wolbachia* was found in 8% of flies [273]. However, another study of Ugandan *Gff* found populations with up to 90% *Wolbachia* prevalence [424]. Schneider *et al.* (2019) also found a negative association between *Spiroplasma* presence and the probability of trypanosome infection in Ugandan *Gff*, suggesting a possible mechanism of interaction whereby *Spiroplasma* infection may reduce vector capacity [273]. *Serratia* has also been identified in Ugandan *Gpal* at low titres [272].

2.2.4 Methodology for studying tsetse microbiota

Studies of tsetse microbiota either use culture-dependent or culture-independent methods; see Section 1.7 for an overview of these two approaches. Culture-independent studies usually involve either species-specific PCRs to identify individual bacteria, or 16S and/or metagenomic sequencing to identify all bacterial species in the sample. While species-specific PCRs are less sensitive they also circumvent the issue of *Wigglesworthia,* which makes up 99% of the tsetse bacterial community [272], drowning out the signal from other less abundant bacteria and preventing an accurate assessment of the population structure of bacterial species in wild tsetse.

Given the importance of studying the microbiota of wild tsetse rather than laboratory populations, and the potential use of *Sodalis* as both a target for trypanosome control and a model organism for the study of symbiosis evolution, appropriate field methodologies are required for the genomic and biological study of symbionts. In order to sequence a bacterial genome or investigate gene transcription from tsetse microbiota, either a sufficient amount of bacterial DNA needs to be present in the

sample to be detectable among the host DNA or a suitable host-depletion strategy must be used; see Chapter 3 for whole-genome sequencing of *Sodalis*, and Chapter 4 for transcriptome sequencing of *Sodalis* in trypanosome-infected tsetse. In some cases, bacteria may be cultured from the insect first to increase the bacterial biomass for DNA extraction: isolated live cultures of insect-associated bacterial further also allows their study in microbiological laboratory conditions, giving insights into their biology and genome evolution (see Chapter 3). Most of the bacteria commonly associated with insects, in particular the symbiotic bacteria that have evolved to live obligately within the insect host, are difficult or currently impossible to culture – see Section 1.3. However, *Sodalis* can be cultured outside of the host in cell-free media [419]. Culture methods are continuously being improved, and a culture method for the insect endosymbiont *Spiroplasma poulsonii* was released just after the grant application for this project was written [348]; see Chapter 7 for a discussion of potential future work with wild culture of tsetse-associated *Spiroplasma*.

Tsetse are susceptible to desiccation in temperatures above 35°C and spend most of their time resting in the shade and shelter of foliage to avoid this [1]. This results in the rapid death of tsetse in most traps, making the culture of live bacteria from wild-caught tsetse difficult. *Sodalis* is slow-growing and requires enriched media [419], making cultures susceptible to overgrowth by contaminating bacteria and fungi. *Sodalis* shares the heat sensitivity of its host, dying at temperatures above 30°C, and requires a microaerophilic atmosphere (5% oxygen) during the early growth phase [419]. These characteristics of both tsetse and *Sodalis* present methodological barriers to the culture of *Sodalis* from wild tsetse.

2.3 Aims and objectives

This study was funded by a Research Travel Grant from the Microbiology Society (to Poppy Pescod), in association with the School of Veterinary Medicine and Animal Resources at Makerere University (MU) and the National Livestock Resources Research Institute (NALIRRI) of Uganda. This work formed part of a knowledge exchange between MU and the University of Salford, to lead to future partnerships between the two institutes.

The study aimed to assess the abundance and distribution of microbiota of interest in wild tsetse flies and isolate a wild strain of *Sodalis glossinidius* for genotype and phenotype comparison to lab strains. Tsetse were collected from the Karuma Falls Catchment Area, an important human-wildlife interface in northern Uganda with three sympatric species of tsetse – *Gmm, Gff* and *Gpal*. Wild tsetse were trapped and tested by PCR for the presence of *Sodalis*, *Wolbachia*, *Spiroplasma* and *Trypanosoma sp.*, to determine whether there is evidence to support correlations between these microbes in any of the three species of tsetse present in the area.

A detailed methodology for the collection of live *Sodalis* cultures from wild-caught tsetse flies was developed and trialled in the field. These cultures were compared with laboratory *Sodalis* strains and strains from laboratory-reared tsetse. Comparisons between *Sodalis* isolates were made morphologically in culture and by whole-genome sequencing sequenced for comparative analysis, to enable a better understanding of the evolution of *Sodalis* in the wild and in the laboratory, with a

view towards the development of a paratransgenesis-based control strategy for trypanosomes in tsetse flies using *Sodalis*.

2.4 Methods

2.4.1 Tsetse trapping

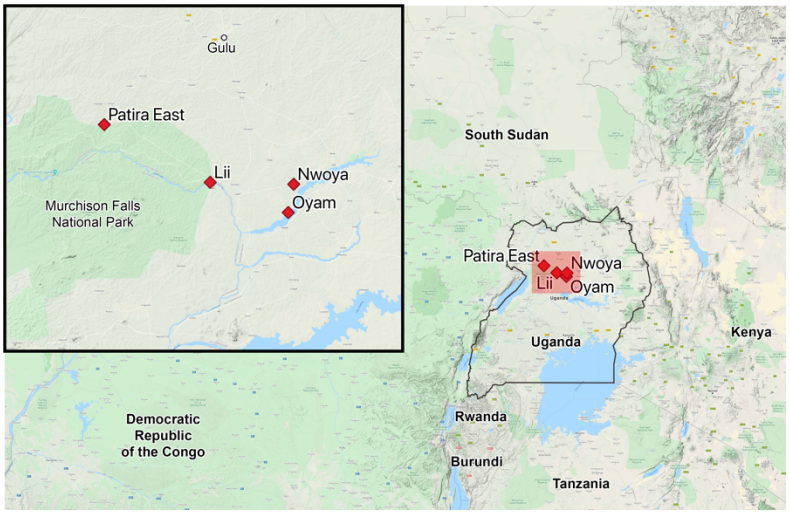


Figure 2-2 - Tsetse fly sampling sites in northern Uganda; Patira East and Lii are within the boundaries of Murchison Falls Naitonal Park, Nwoya and Oyam are on nearby community grazing land.

Trapping was conducted in collaboration with Dr Judy Mwangi (JM) from the University of Salford, Dr Patrick Vudriko (PV) from MU and Patrick Abila (PA) from the NALIRRI. Four tsetse trapping sites were chosen in the Karuma Falls Catchment Area in northern Uganda in a *T. b. gambiense*- and nagana-endemic area, based on the recommendation of the district entomologist and collaborators at MU, in areas known to be infested with three tsetse species – *Gmm, Gpal* and *Gff* (Figure 2-2). Two sites, Patira East and Lii, were within the boundaries of Murchison Falls National Park and therefore were free from livestock grazing or game hunting; Nwoya and Oyam sites were on community grazed land. Each site had a water source nearby and 5-10 traps spaced at least 10 m apart; traps were placed in shaded areas to prevent flies dying from heat exhaustion or desiccating. Trapping was conducted from 15-19 January 2019, with temperatures during the day remaining stable at 28°C and no rainy days. Each trap was put up in the morning, checked that evening and again the following morning before being dismantled, with the sites staggered throughout the week. Unbaited biconical traps were used (Figure 2-3), with netted cages on top that were removed and replaced when each trap was checked.

Tsetse flies collected from the traps were kept in the mesh cages in a cool, humid environment and transported back to the guest house in Gulu each night. Any live tsetse were separated and dissected for bacterial culture (see Figure 2-3, procedure detailed below). Dead tsetse were identified morphologically by sex and reproductive status if female (pregnant or non-pregnant), before being stored at room temperature in 70% ethanol for transport. Dissection and identification of flies was performed with assistance from PA and PV.

Figure 2-3 - Fieldwork for tsetse surveying in Uganda. Left: a biconical trap set up on site in Lii. Right: field dissection of tsetse for *Sodalis* culture.

2.4.2 Tsetse dissection and *Sodalis* culture

A sterile dissection environment was created at the Gulu guesthouse by laying sheets of blue roll soaked in 70% ethanol over a plastic tray, allowing the ethanol to evaporate before lighting a portable Bunsen burner (Figure 2-3). All equipment was thoroughly rinsed in 70% ethanol and dried under flame before use. Bacteria were extracted using the dissection method detailed by Matthew *et al.* [419] and modified by Dr. Lee Haines [425] (unpublished data) (Figure 2-3). Extracted bacteria were transferred to 1 ml of Mitsuhashi-Maramorosch Insect Medium (MMI) (Sigma Aldrich) in 1.5 ml Eppendorf tubes and stored at ambient temperature (approximately 25-28°C) for up to a week. Cultures were transported to a laboratory at MU for further culture work. In a Category 2 pathogen hood, cultures were streak plated onto Columbia agar plates supplemented with 10% defibrinated horse blood (Thermo Fisher) using disposable 5 µL plastic loops. Plates were parafilmed and stored in a sealed container with a CampyGen Sachet (Oxoid) to create a microaerophilic atmosphere (5% oxygen, 10% carbon dioxide and 85% nitrogen) for at least three days at ambient temperature (25°C). Single colonies morphologically identified as *Sodalis* were picked from grown plates and re-streaked onto Columbia agar and blood plates before being grown for at least three days in the same conditions as previously to produce pure *Sodalis* cultures.

2.4.3 DNA extraction

DNA extraction of whole tsetse was performed at the University of Salford. Tsetse were removed from the storage ethanol and washed in fresh 70% ethanol for around 30 seconds before being air dried on a sterile surface. The wings and head were removed from each fly to prevent blocking of the extraction column membranes and

PCR inhibition respectively [426]. DNA was then extracted using the QIAamp DSP DNA Whole Blood kit with an additional proteinase K step. All steps were performed at room temperature unless otherwise stated. In a 1.5 µl Eppendorf 20 µl proteinase K, 100 µl molecular biology-grade water, 200 µl Lysis buffer and the sterile dried tsetse fly were added together. The mixture was homogenised with a glass pestle that had been sterilised with 70% ethanol and air dried. Homogenised samples were incubated at 56°C for two hours. Samples were briefly centrifuged to collect condensation; 200 µl of absolute ethanol was added and mixed thoroughly by pulse vortexing. Samples were then briefly centrifuged and the supernatant was applied to the QIAamp mini spin column, avoiding any unhomogenised carapace pieces that would block the column membrane. Columns were centrifuged at 6,000 x g for one minute before discarding the filtrate and placing the column in a clean collection tube. Columns had 500 µl Wash Buffer 1 added before centrifuging at 6,000 x g for one minute. The filtrate was discarded and 500 µl of Wash Buffer 2 was added to the column before centrifuging at 20,000 x g for one minute. The filtrate was discarded and the column was centrifuged again at 20,000 x g for three minutes to ensure the membrane was completely dry. The column was placed in a 1.5 ml Eppendorf collection tube; 50 µl of Elution Buffer was added directly to the column membrane and incubated for one minute before centrifuging at 6,000 x g for one minute. This step was repeated to leave a final elution of 100 µl of extracted DNA per sample.

2.4.4 Polymerase chain reaction and gel electrophoresis for species detection

PCRs were performed to identify six members of the tsetse microbiota: *Sodalis, Wigglesworthia, Wolbachia, Spiroplasma, Serratia* and *Trypanosoma,* as well as a

PCR to determine the tsetse species. All PCRs were performed with the same conditions and reagents, separately for each primer pair (Table 2-1) and each sample. Reaction mixtures were made up of 12.5 µl MyTaq Red Mix (Bioline), 5 µl of molecular biology-grade water, 1.25 µl of a forward and reverse primer pair at 10 nM and 5 µl of template DNA. Reaction conditions were: 95°C for one minute, 35 cycles of 95°C for 15 seconds, the relative annealing temperature for 15 seconds and 72°C for 10 seconds, followed by a final extension step at 72°C for one minute. PCR products were run on a 1.5% agar gel with GelRed Nucleic Acid stain (Cambridge Biosciences) at 80 W for 90 minutes with a HyperLadder 50 bp for DNA weight comparison and visualised under UV light.

Table 2-1 - Primers used in PCRs to check tsetse fly samples for five associated microbes and host species. Annealing temperatures were calculated by subtracting 2-5°C from the average melting temperature of the primer pair.

Target	Primer	Primer sequence (5'-3')	Anneal temp. (°C)	Product length (nt)	Ref
Sodalis glossinidius	GroELf GroELr	CCA AAG CTA TCG CTC AGG TAG G TTC TTT GCC CAC TTT CGC CAT A	57	95	[427]
Wigglesworthia glossinidia	ThiCF ThiCR	TGA AAA CAT TTG CAA AAT TTG GGT GTT ACA TAG CAT AAC AT	48	366	[236]
Wolbachia sp.	wsp81F wsp691R	TGG TCC AAT AAG TGA TGA AGA AAC AAA AAT TAA ACG CTA CTC CA	51	605	[428]
Spiroplasma sp.	63F TKSSsp	GCC TAA TAC ATG CAA GTC GAA C TAG CCG TGG CTT TCT GGT AA	55	~500	[429] [430]
Serratia sp.	Fpfs1 Rpfs2	CCG GCA TCG GCA AAG TCT ATC TGG CCC GGC TCG TAG CC	67	193	[431]
Trypanosoma sp.	ITS1CF ITS1BR	CCG GAA GTT CAC CGA TAT TG TTG CTG CGT TCT TCA ACG AA	53	200-3700	[432]
Glossina sp.	GlosITS1f GlosITS1r	GTG ATC CAC CGC TTA GAG TGA GCA AAA GTT GAC CGA ACT TGA	60	618-920	[433, 434]

2.4.5 Statistical analysis

Samples were classified as either positive or negative for each of the five organisms tested for based on the presence or absence of an amplified DNA band. Chi-squared tests were performed in R (version 3.6.1) [435] using RStudio (version 1.2.5019) [436] and barplots were created in Microsoft Excel.

2.5 Results

2.5.1 Wild caught tsetse fly demography

DNA was extracted from a total of 175 tsetse flies. The sex ratios of the sample population as determined morphologically in the field can be seen in Figure 2-4. More than twice as many females as males were caught, with 31% of captured females visibly gravid. Of the four sites sampled tsetse flies were only found in Patira East and Lii, with 93% of flies caught in Lii.

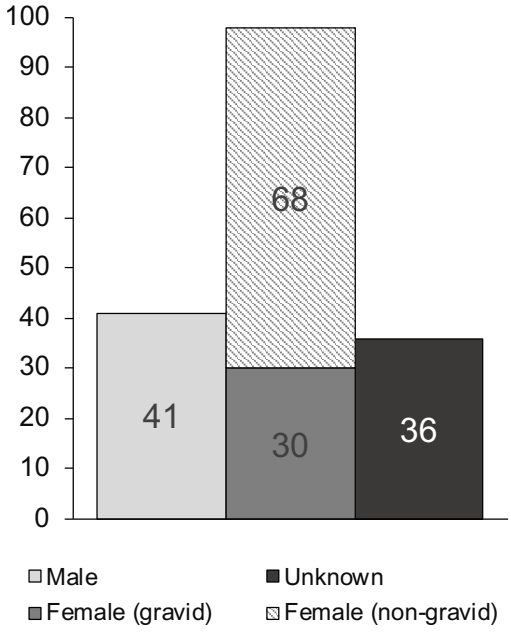


Figure 2-4 – Sex and fecundity information of the wild-caught tsetse that were analysed in this experiment. Including sex as identified morphologically, and the number of females which were visibly gravid. Unknown: samples were too desiccated or degraded to identify sex.

2.5.2 Wild *Sodalis* culture

A total of 68 live tsetse flies were cultured from and grown on blood agar plates. No pure cultures of *Sodalis* were successfully isolated due to the large amount of bacterial and fungal contamination on the plates (Figure 2-5).

Figure 2-5 - Bacterial cultures from two wild-caught tsetse flies, streaked from fly lysate in MMI liquid medium onto Columbia agar supplemented with horse blood. Both bacterial and fungal contaminants are visible.

2.5.3 Symbiont prevalence in wild tsetse populations

Due to disruption from the Covid-19 pandemic, PCR assays were not finished. The following dataset encompasses the work completed to

date: 57 tsetse fly samples were analysed for *Sodalis*, *Wolbachia*, *Spiroplasma* and *Trypanosoma* presence (Figure 2-6). *Sodalis* was identified in 39% of tsetse, whereas *Wolbachia* was only seen in one sample (1.7%). *Spiroplasma* was found in the majority of samples (88%), whereas trypanosomes were only seen in 12% of samples. No correlations between sex, *Spiroplasma, Trypanosoma* or *Sodalis* presence were found (Table 2-2); *Wolbachia* was only present in one sample and so was not used in statistical analyses. A detailed breakdown of the target microorganism incidence by sex and gravid or non-gravid females can be seen in Figure 2-7. Six out of seven *Trypanosoma*-positive samples were also positive for *Spiroplasma*, and four *Trypanosoma*-positive samples also contained *Sodalis* (Table 2-3).

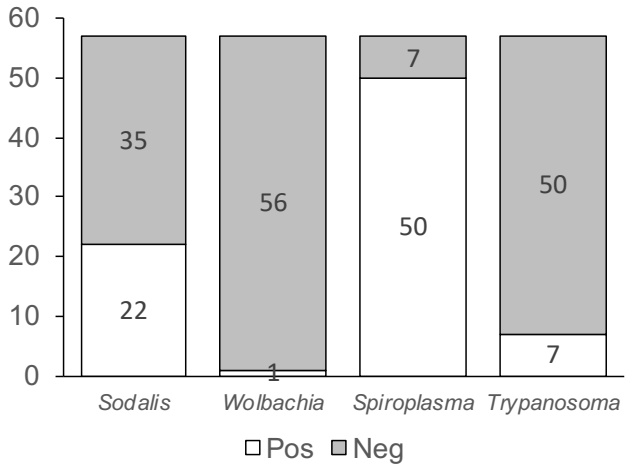


Figure 2-6 - Incidence of four symbiotic organisms in 57 wild-caught tsetse flies from northern Uganda.

Table 2-2 - Chi-squared analysis of presence of *Sodalis*, *Spiroplasma* and *Trypanosoma*, and the correlation with sex (male, female or unknown). No p-values are significant, indicating no correlations. df – degrees of freedom.

Variable 1	Variable 2	χ^2 value	P value	df
Sex	*Sodalis*	1.8484	0.397	2
Sex	*Trypanosoma*	1.5279	0.466	2
Trypanosoma	*Sodalis*	0.43786	0.508	1
Sex	*Spiroplasma*	0.70375	0.704	2
Trypanosoma	*Spiroplasma*	<0.001	1	1
Sodalis	*Spiroplasma*	<0.001	1	1

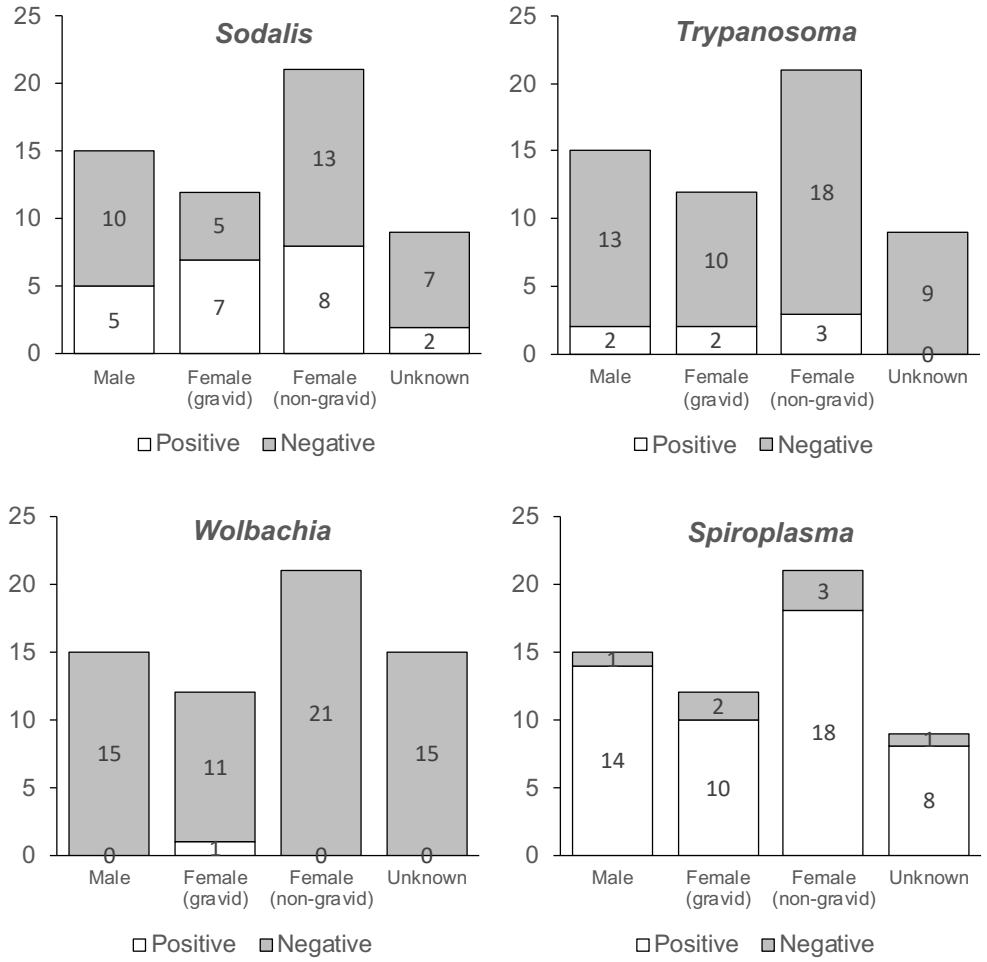


Figure 2-7 - Incidence of *Sodalis, Trypanosoma, Wolbachia and Spiroplasma* in male, gravid and non-gravid females, and unknown sex tsetse fly samples

Table 2-3 - A table of the incidence of: (left) *Trypanosoma* and *Spiroplasma*; and (right) *Trypanosoma* and *Sodalis*.

	Tryp +	Tryp -	**Total**
Spiro +	6	44	50
Spiro -	1	6	7
Total	7	50	**57**

	Tryp +	Tryp -	**Total**
Sod +	4	18	22
Sod -	3	32	35
Total	7	50	**57**

2.6 Discussion

2.6.1 Tsetse fly distribution and demography

The capture of twice as many female tsetse flies as males has been previously described [437, 438], but may be an artefact of the trap type or trapping location: different trapping strategies result in the capture of different ratios of males to females as well as visibly gravid to non-gravid females [439]. A predominance of females will be found in hunting sites as females have a need for more frequent blood meals to produce offspring, whereas males congregate around breeding foci, leading to different sex ratios in trapping sites within 5km of each other [1, 439]. Despite their ability to fly long distances and reinfest cleared areas [440], the tendency of tsetse to congregate in specific areas for feeding, resting or breeding also explains the wide variation in capture rate between our sites despite their close proximity to each other. Whether the lack of trapped tsetse on community grazing land was due to brush clearance or control strategies used to protect grazing cattle, or other unrelated factors, is unclear. This pilot study confirmed the need for wide preliminary trapping to ensure infestation of target areas before conducting wider research projects to survey wild tsetse microbiota.

2.6.2 Culturing bacteria from wild tsetse

The culturing techniques for *Sodalis*, which is an inherently difficult microbe to culture due to its slow growth and enriched media requirements, were developed for use in a laboratory environment with little air movement and a clean work surface. While every possible precaution was taken to ensure the cleanliness of the work, it was insufficient to produce the exceptionally clean cultures necessary to isolate

Sodalis. However, the field trial of this process allows the identification of areas that need improvement and helps with designing strategies for field isolation of symbiotic bacteria in the future. Although only *Sodalis* was targeted for isolation and culture in this study, the less well-characterised *Spiroplasma* and *Serratia* require similar culture strategies [269, 348]. Isolation of all three bacteria from wild tsetse should be attempted in the future, to facilitate the *in vitro* study of their characteristics and their effects on tsetse vectorial capacity.

2.6.3 Prevalence of tsetse fly microbiota

Disruption of lab work due to the Covid-19 pandemic resulted in a reduced dataset, limiting the statistical power of the chi-squared tests and thus the conclusions that can be drawn. Due to the low incidence of *Trypanosoma* in the tsetse flies surveyed (7/57), no significant correlations could be seen between the parasite and any of the three other microorganisms tested for. Low incidence of *Wolbachia* (1/57) meant likewise that no statistical inferences could be made, and no data on the incidence of *Serratia* was collected. It was also not possible to distinguish between tsetse species, which limits the comparability of this study with other studies of tsetse microbiota and obscures any effect of host species. However, the discovery of a large proportion of tsetse flies infected with *Spiroplasma* raises interesting questions about its impact on trypanosome transmission. Although we did not see a significant correlation between *Spiroplasma* and *Trypanosoma* incidence, other studies have seen a negative correlation implying that *Spiroplasma* may play a role in suppressing trypanosome transmission [273].

The microbiota prevalence in this study is not entirely similar to the findings of other studies. A much larger proportion of *Spiroplasma*-infected tsetse were described here than in another Ugandan study by Schneider *et al.*, which described fluctuating prevalence between 0-34% in northern Uganda compared to the 88% seen in this dataset [273]. Schneider *et al.* observed that *Spiroplasma* prevalence was correlated strongly with geographical location, seasonality and tsetse species, potentially explaining the discrepancy in findings. The results from Chapter 2 also confirm the presence of *Sodalis* in the tsetse fly populations in Murchison Falls as has been previously described by Aksoy *et al.* [272], although they found *Sodalis* at low titres in 99% of samples across Uganda, compared to the 39% seen here. The prevalence of *Wolbachia* found in the tsetse survey in Chapter 2 was also lower than in previous studies: Schneider *et al.* found a *Wolbachia* infection rate of 8% in northern Uganda [273], whereas Symula *et al.* found *Wolbachia* present in 90% of tsetse flies, sampled from the same region as this study – Murchison Falls [424].

This study adds evidence to support previous assertations that there is no sexual bias for *Spiroplasma* or *Sodalis* incidence in tsetse flies [273, 397, 441]. As both sexes of tsetse flies take blood meals and transmit trypanosomiasis, sex bias of bacteria with the potential to alter vector capacity may be an important factor in pathogen transmission dynamics. The transmission of symbiotic bacteria can occur horizontally (from individual to individual, during mating or close contact) or vertically (from mother to offspring), or bacteria may be acquired from the environment (see Section 1.3). *Sodalis* is transmitted both horizontally and vertically [257]; *Spiroplasma* transmission dynamics are unknown, although theorised to either be horizontal or by environmental acquisition [273]. Possible transmission routes are an

important aspect of assessing the potential of symbiotic bacteria for use in vector control strategies (see Section 1.4).

Despite the limited dataset produced in this study, it adds valuable information about the heterogeneous presence of *Sodalis* in tsetse populations and helps target areas for future work. Analysis of the remaining samples, and identifying the host species of all samples, will build on the important groundwork laid by this study in examining the microbiota of a heterogenous population of sympatric tsetse species.

Chapter 3 Comparative genome analysis of an experimentally evolved strain of *Sodalis glossinidius* *Sg*GMMC1*

3.1 Abstract

The tsetse bacterial endosymbiont *Sodalis glossinidius* is used as a model for studying recent symbiosis and evolution due to its large genome containing 49% pseudogenes. A strain of *Sodalis* was isolated from lab reared tsetse (*Gmm*) and serially passaged in vitro for 10 years to produce strain *Sg*GMMC1*. Whole-genome sequencing was performed using Illumina MiSeq and Oxford Nanopore MinION sequencers to generate a complete hybrid genome assembly. Genome data for the *Sg*GMMC1* strain is presented and compared to strain *Sg*GMMB4, which was isolated at the same time from the same colony and sequenced immediately. One large deletion of 17 kb is discovered, containing 19 pseudogenes and 14 CDS, impacting thiamine biosynthesis, quorum sensing and transmembrane protein production. Additionally 57 small polymorphisms were observed, 37 affecting coding regions, 9 of which were already classified as pseudogenes and 12 of which were predicted to be newly pseudogenised. Novel pseudogenes included those involved in quorum sensing, carbohydrate metabolism and transmembrane proteins. This provides evidence of genome erosion in *Sodalis* environmental interaction systems, supporting the hypothesis of pseudogenisation and deletion events in symbiont evolution. The raw read sequences have been deposited with the European Nucleotide Archive under accession number PRJEB32321, and the annotated genome is available on figshare at DOI: 10.17866/rd.salford.8052437.

3.2 Introduction

3.2.1 Genome characteristics of symbiotic bacteria

Most insect symbionts follow a common pattern of evolution: from a free-living ancestor which infects a host insect and successfully establishes within the host, persisting in the population through horizontal and eventually vertical transmission [50, 442]. Less genetic flexibility is required in order to persist in a host environment that has low variation of abiotic factors and a consistent supply of limited nutrients than in a free-living environment. Negative selection pressure against the maintenance of disposable or costly genes results in genome reduction of the symbiont: only genes that provide an essential resource not available in the host environment, or that provide a resource whose benefit to the host environment offsets its cost to the producer, are selected for and retained. The trend for genome reduction of symbiotic organisms towards interdependence with their hosts is known as the Black Queen Hypothesis [443].

Genome reduction is driven by accumulated mutations along with a lack of DNA repair genes, which may be lost by genetic drift early on in the adaptation of free-living bacteria to a symbiotic lifestyle [444]. The vertical or horizontal transmission of symbionts reduces their population size producing a genetic bottleneck; bottlenecks are known to amplify the population-wide effects of genetic drift [445], increasing the accumulation of both deleterious and advantageous mutations [446, 447]. Mutations easily become fixed through genetic drift in smaller populations, in a process known as Muller’s ratchet [448]. As mutations accumulate in non-essential genes, frameshifts and large numbers of insertions or deletions (indels) result in the gene

losing its original function in a process known as pseudogenisation [449]. Although originally considered to be DNA with little to no function, pseudogenes are often transcribed to produce functional RNA with roles in gene expression regulation [450]. Pseudogenes constitute a considerable proportion of eukaryote genomes, sometimes matching the number of true genes [451], whereas in most bacterial genomes pseudogenes make up only 1-5% of their total genes [452]. Exceptions to this rule are seen mostly in host-associated bacteria, both pathogenic and beneficial: *Sodalis* has one of the highest proportions of pseudogenes to coding genes of any bacteria, at around 52% pseudogenes [60]; the causative agent of leprosy, *Mycobacterium leprae*, has a similar percentage of pseudogenes [453].

Due to the genome reduction associated with adapting to the host environment, bacterial symbionts have some of the smallest genomes of single-cellular organisms present in nature. Most free-living bacteria have genomes of 2-8 million base pairs (Mbp) [50, 247]; the largest known bacterial genome, belonging to the myxobacterium *Sorangium cellulosum*, comprises 13 Mbp [454]. In comparison, most obligate insect symbionts have genomes of 0.6-0.7 Mbp [50]. In the case of the phloem-feeding insect *Macrosteles quadrilineatus*, a co-dependent partnership of two obligate symbionts includes *Sulcia muelleri* with the smallest bacterial genome yet sequenced, at just 112 kb (0.1 Mbp) [455]. Symbionts with a long association with their host insects also have nucleotide compositions highly biased towards adenine and thymine (A+T); while most bacteria have a guanine and cytosine (G+C) content of around 50%, obligate symbiont G+C contents can be as low as 17% [456]. This bias is characteristic of highly reduced genomes, and can be caused by many common categories of mutation that are usually corrected by the DNA repair

genes often lost early in the genome reduction process [457]. Bias towards A+T results in proteins consisting predominantly of amino acids that are encoded by A+T-rich codons, which can have a deleterious effect on protein thermal stability, creating an evolutionary balance between mutation and selective constraint of low G+C content in symbionts [457, 458].

While many insect symbionts follow this trend of linear genome reduction, there are exceptions. The *Rickettsia* endosymbiont of *Ixodes scapularis* (REIS) has the largest *Rickettsia* genome sequenced so far at around 2 Mbp; this is due to the large-scale proliferation of mobile genetic elements (MGEs) within the genome, products of multiple genomic invasions [459]. MGEs are susceptible to insertions and deletions, resulting in large scale pseudogenisation of MGE regions, or providing hot spots for integration of mobile genes from other bacteria containing important components for a symbiotic lifestyle [459, 460]. In a similar intracellular system, the *Wolbachia* strain *w*Fol found in the parthenogenic springtail *Folsomia candida* has the largest sequenced *Wolbachia* genome to date at 1.8 Mbp. *w*Fol has integrated multiple phage sequences into its genome, resulting in a quarter of its genes consisting of phage regions [461]. This process of genome expansion through MGE incorporation constitutes an ‘escape’ from the Muller’s ratchet of genome reduction, although the cost of accumulating MGEs to gain certain functions may result in a less streamlined and functional genome overall due to their high susceptibility to pseudogenisation.

The line between pathogenic and beneficial symbionts is often blurred; both pathogenic and beneficial bacteria need to successfully establish within host cells or tissues, and therefore often share mechanisms to do so. These shared mechanisms

1901 are often classed as virulence factors – genes or systems that increase the virulence
1902 of a pathogen within a host – despite being shared by both pathogens and
1903 mutualistic symbionts. Virulence factors can be horizontally acquired via HGT by
1904 mobile genetic elements such as plasmids and bacteriophages; some have been
1905 documented to transform non-pathogenic strains of bacteria into pathogens [462-
1906 464]. This suggests that the reverse could be true, or that free-living, parasitic
1907 bacteria may use virulence genes to establish within a host before transitioning to a
1908 facultative lifestyle, retaining the virulence mechanisms for invading host cells
1909 without causing harm. Indeed, virulence factors present in parasitic bacteria
1910 transitioning to a beneficial symbiotic lifestyle can be attenuated by the bacterium, in
1911 order to facilitate establishment in the host. The retention of virulence genes for cell
1912 invasion in beneficial symbionts has been documented in multiple host-symbiont
1913 systems. Both the tsetse fly symbiont *Sodalis glossinidius* and the aphid symbiont
1914 *Hamiltonella defensa* retain two TTSS for host cell invasion [105, 118, 465], as do
1915 multiple pathogens including *Yersinia pestis* and *Salmonella enterica* [466, 467]. In
1916 the aphid-*H. defensa* system the virulence factors used to invade host cells are also
1917 directly beneficial to the host by being deployed for use against parasitoid wasps to
1918 prevent parasitism of the aphid [113, 118, 468].
1919
1920 Quorum sensing is a type of communication between bacterial cells that allows the
1921 determination of local population density and the co-ordination of group behaviour
1922 through the production small molecules called autoinducers, which diffuse across
1923 bacterial cell membranes and modulate the transcription of target genes [469-471].
1924 Quorum sensing is hypothesised to modulate the use of virulence genes in beneficial
1925 symbionts; at low bacterial densities virulence genes are expressed to increase

infection of host cells, and as bacterial density increases quorum sensing will reduce the expression of these virulence genes to attenuate negative effects on the host [472]. This effect has been observed in tsetse flies artificially infected with *Sodalis praecaptivus*, a free-living close relative of the tsetse symbiont *S. glossinidius* [472].

3.2.2 Studying the evolution of symbiont genomes

Insect endosymbionts have among the smallest genomes of all bacteria, with the highest A+T bias and highest pseudogene content; their study has provided valuable information into the evolutionary potential of bacterial genomes for adaptation [473]. Due to their rapid development and reproduction the evolution of bacterial genomes can be observed and influenced in real time; experimental evolution is a research framework whereby a selection pressure is applied to a population of study organisms over sufficient generations to observe an adaptation [474, 475]. By directly observing the dynamic evolutionary process with complete control over selection pressures, many questions about evolution can be investigated and mechanisms of evolution can be demonstrated [475]. Using a combination of experimental evolution and modern genomics, both phenotypic adaptations and whole genome changes can be studied.

Recent advances in genomic technology have accelerated the study of symbiotic bacteria and their evolution from free-living ancestors. The development of ‘omics technologies over the last two decades has allowed the close examination of symbiont genomes, their expression and translation, and the metabolic networks produced by symbionts and their hosts (Section 1.7.2) [476, 477]. Databases of bacterial genomes are freely available, allowing the study of symbiont evolution by

comparative genomics with ancestor strains [478]. In order to best study the evolution of symbionts by comparison with closely-related ancestors, the genomes must be of high quality with no unknown bases (i.e. closed, circular chromosomes), with well-annotated genes and pseudogenes [479]. Early high-throughput sequencing technologies such as Roche 454 and Illumina MiSeq were only capable of producing short reads of <500bp [480]. Assembly of complete genomes from short fragments is made difficult by >500bp repetitive regions scattered throughout bacterial genomes, resulting in genome assemblies consisting of short contiguous sequences (contigs) with breaks of unknown bases between them [480]. The latest next-generation technologies, such as Oxford Nanopore Technologies and Pacific Biosciences, are capable of consistently producing long reads of over 10 kb [480-482]. Although these long reads are sufficient to produce whole, contiguous genomes with no breaks, the error rate of long-read technologies is still much higher than short-read technologies (5-15% vs <1%) [483]. To resolve this a hybrid assembly approach has been developed, using short-read sequencing to produce accurate contigs and long-read sequencing of the same sample to scaffold the short reads onto, producing highly accurate, closed, single-contig bacterial genomes [483, 484]. This has facilitated the study of fine-scale bacterial genome evolution, down to single nucleotide polymorphisms (SNPs), as well as accurate gene annotation. As pseudogenes play an important role in the evolution of symbiotic bacteria, identifying them within the genome is essential; accurate knowledge of bacterial genes and their products allows the classification of pseudogenes by comparison with homologous ancestral gene precursors, with frameshifts or truncations [452].

3.2.3 The *Sodalis glossinidius* genome

Sodalis glossinidius is a Gram-negative secondary bacterial endosymbiont of the trypanosome vector *Glossina sp.*, and its most recently-evolved bacterial partner (see Section 1.5.3.3). Although removal of *Sodalis* reduces host longevity suggesting a mutualistic role [262], the presence of *Sodalis* has been linked to an increase in tsetse fly susceptibility to trypanosome infection by an unknown mechanism [55] (Section 1.5.3.3 and Chapter 4). *Sodalis* has also lost the ability to independently synthesise thiamine, an essential vitamin for tsetse reproduction produced by *Wigglesworthia* [224, 485, 486]. *Sodalis* is hypothesised to contribute co-factors to the thiamine biosynthesis pathway, potentially facilitating or increasing the production of thiamine by *Wigglesworthia* [108].

Sodalis has a 4 Mbp genome [60], similar in size to a free-living bacterium [50, 247], and can be cultured *in vitro* in cell-free media with relative ease [419]; both of these features suggest a recent host association and symbiosis [60, 419]. The only other confirmed member of the *Sodalis* genus and closest relative of *S. glossinidius* is *S. praecaptivus*, a free-living bacteria isolated from a human wound [251]. *S. praecaptivus*, the species name of which translates as ‘not yet taken prisoner’ [251], represents a close evolutionary precursor of *S. glossinidius* and can be used experimentally as a free-living ancestor strain to explore the dynamics of infection and establishment of symbiont ancestors [487].

The evolution of bacteria from free-living to endosymbiosis is thought to involve inactivation of functional genes via small indels to produce pseudogenes, which are subsequently lost in large deletion events [488]. *Sodalis* represents a mid-way point

in the evolutionary transition from a free-living organism towards an obligate endosymbiosis, evidenced by the high proportion of pseudogenes in its genome: 49% compared to the usual 1% present in free-living bacteria [60, 489]. The current state of the *Sodalis* genome provides a preview of the degeneration and streamlining of a developing symbiosis. As well as its use as a study organism for the evolutionary processes of symbiosis, studying *Sodalis* pseudogenes has demonstrated that they can be significantly expressed and may have a regulatory role in cell functions [489].

3.3 Aims and objectives

The study of the genome architecture and evolution of *Sodalis* may yield important information about its unusual transitionary nature, as well as its unknown role in trypanosome transmission. By subjecting *Sodalis* to frequent population bottlenecks *in vitro* over a long period of time, we aimed to observe its trajectory towards a smaller, more streamlined symbiont-like genome via the deletion of pseudogenes and pseudogenisation of coding genes. Here we present a whole-genome sequence assembly of a strain of *Sodalis* isolated from a lab colony of *Glossina morsitans morsitans* and serially passaged *in vitro* for 10 years, thereby representing an "experimentally evolved" strain that can be used to study the evolutionary trajectory of the species.

3.4 Methods

3.4.1 *In vitro* culture of *Sodalis*

In work leading up to this study, *Sodalis* was isolated from two male *G. m. morsitans* from the colony at the Liverpool School of Tropical Medicine in 2007 following the methods in Matthew *et al*. [419]. The strains were named *Sg*GMMC1 and *Sg*GMMB4, and were stored at -80°C. Beginning in 2009, Dr. Lee Haines (LH) at the Liverpool School of Tropical Medicine serially cultured *Sg*GMMC1 at 27°C in a microaerophilic atmosphere and passaged every 10-14 days for 10 years, alternating between solid-phase growth on Columbia agar base supplemented with defibrinated horse blood (Oxoid/Thermo Fisher) and liquid-phase growth in serum-free insect media (Sigma Aldrich), to create the strain *Sg*GMMC1*. Cultures of this strain were then acquired by the author from LH for use in the current study.

3.4.2 Genome sequencing

DNA was extracted from a cell pellet of *Sg*GMMC1* using the Zymo Quick DNA kit and quantified using a Qubit™ 3.0. The *Sg*GMMC1* genome was sequenced using Oxford Nanopore MinION and Illumina MiSeq platforms. A library for the MinION was generated using a SQK-LSK108 1D ligation kit and sequenced on a R9.4 flowcell, producing 149,847 reads of mean length 3,477.62 bp and length N50 of 14,468 bp. A total of 521,111,234 bases were sequenced giving an estimated 118x genome coverage.

The same DNA sample was sequenced on the MiSeq; a QIAseq FX DNA library prep kit and v2 150 bp sequencing cartridge were used producing a total of 755,405

paired-end reads. MiSeq reads were assessed and adaptors trimmed using FastQC (v0.11.4) [490]. MiSeq reads were trimmed for quality using fastp with default settings (v0.12.5) [491].

3.4.3 Genome assembly and comparative analysis

A hybrid genome for *Sg*GMMC1* was built using Unicycler (v0.4.8-beta) [483] by assembling the MiSeq reads *de novo* and mapping the long MinION reads onto this assembly for structural correction. The hybrid assembly was annotated using PROKKA (v1.13.3) [492] using the reference strain *Sg*GMMB4 genome (GenBank Accession LN854557.1), which was deposited in 2016. The annotated *Sg*GMMC1* assembly was compared to *Sg*GMMB4 using the Artemis Comparison Tool (ACT) (v13.0.0) [493] to identify structural changes and large deletions, and Snippy (v3.1) [494] to identify SNPs and indels.

The amino acid sequences of mutated genes were searched in the STRING database [495] to identify conserved orthologous groups (COGs), and the InterPro database [496] to identify disrupted conserved regions and gene ontology (GO) terms. A graph of the *Sg*GMMC1* assembly was produced using Bandage (v0.8.1) [497] to assess its quality, and genome comparison figures were produced using easyfig (v2.2.2) [498], DNAplotter [499], and Addgene [500].

The Snippy output was examined for potential gene changes, removing any mutations in non-coding regions or synonymous SNPs before generating a table of Snippy impacts. Potential novel pseudogenes caused by the polymorphisms identified by Snippy were classified in three tiers using the following criteria:

Tier 1 (most likely pseudogenisation): more than two-thirds of the gene as annotated in *Sg*GMMB4 is removed;

Tier 2: disruption of a conserved region by truncation;

Tier 3 (least likely pseudogenisation): disruption of a conserved region by a missense mutation.

Protein tertiary structures for a selection of Tier 1 and 2 genes were produced using PHYRE2 [501]; the tertiary structures of Tier 3 genes were not examined as mapping with PHYRE2 was not sensitive enough to distinguish changes caused by single missense mutations.

3.5 Results

3.5.1 Genome assembly of *Sg*GMMC1*

The assembled *Sg*GMMC1* genome was 4,288,975 bp in size with 54.43% GC and no unidentified base pairs (Ns). The assembly consisted of six contigs with the N50 also representing the largest contig at 4,149,326 bp. One circular chromosome, four circular plasmids (pSG1, 81,522 bp; pSG2, 26,425 bp; pSG3, 19,512 bp; and pSG4/3, 10,816 bp) and one plasmid fragment (pSG2 plasmid frag, 1,374 bp) were produced. *Sg*GMMC1* was identified with 99.99% similarity to the reference strain *Sg*GMMB4 using BLASTN (v2.8.1) (Figure 3-1) [502].

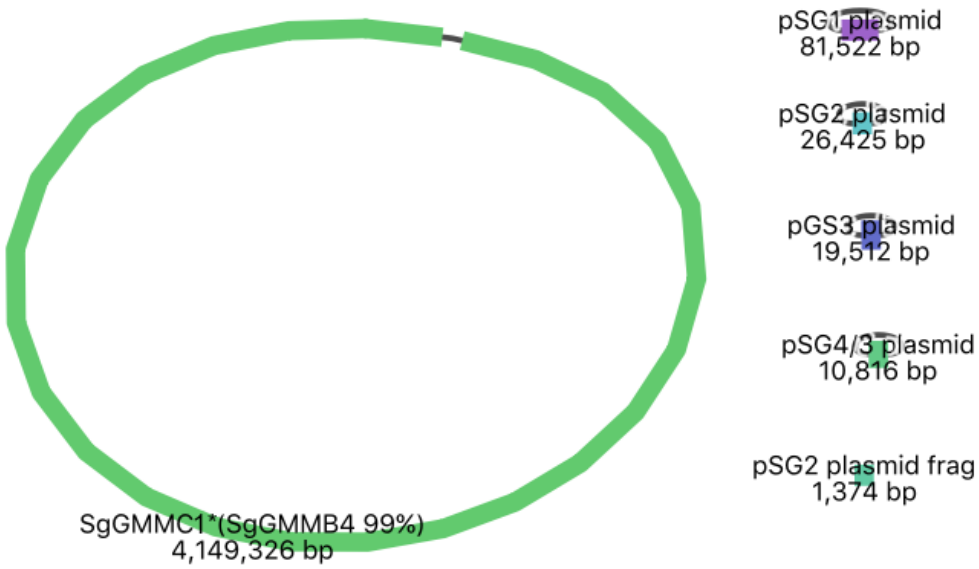


Figure 3-1 - Hybrid genome assembly of *Sodalis glossinidius* strain *Sg*GMMC1*, illustrating the single contiguous, circular chromosome with no ambiguous sections, as well as four circular plasmids and one plasmid fragment. The genome has a 99% similarity to ancestral reference strain *Sg*GMMB4. Figure created in Bandage using BLAST for reference genome percentage.

3.5.2 Comparative genomics of *Sg*GMMC1* and *Sg*GMMB4

One large deletion of 17,209 bp was identified, containing 19 pseudogenes and 14 coding DNA sequences (CDS) (Figure 3-2, Figure 3-3 and Table 3-1). The CDS in this deleted region included: *thiM*, which encodes a protein involved in thiamine biosynthesis; one of two copies of the quorum-sensing transcription regulator *esaR*; three sulfur and one sulfite transmembrane transporter proteins; one copy of the peptidase *yxeP*, the other copy of which is pseudogenised elsewhere in the genome; and multiple sugar phosphotransferases.

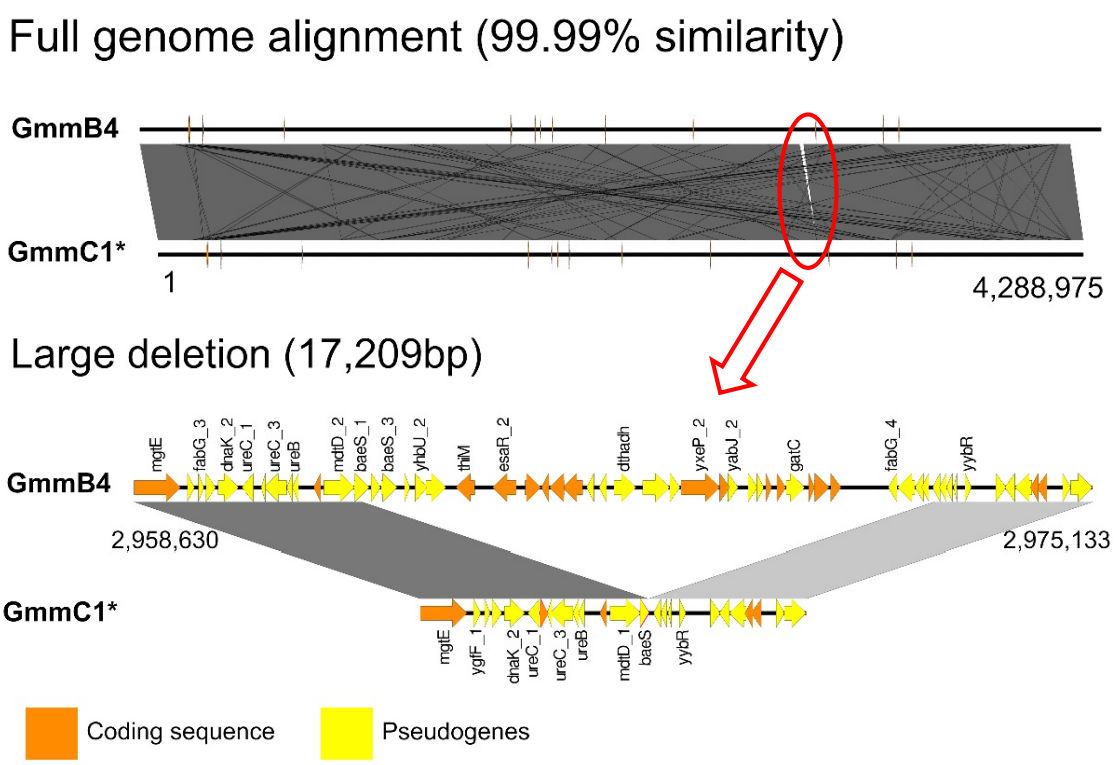


Figure 3-2 - Full genome alignment of *Sodalis glossinidius* evolved strain *Sg*GMMC1* against ancestor strain *Sg*GMMB4 showing 99.99% similarity, with the single large deletion expanded to show details of the genes deleted. The top figure shows the alignment of the entire chromosome of each strain, showing the high level of similarity between the sequencies, with the large deletion circled. The bottom diagram shows the coding genes (orange) and pseudogenes (yellow) flanking and included in the deletion. Genome positions (bp) are relative to the *Sg*GMMB4 reference genome. Both figures were produced using easyfig.

2097

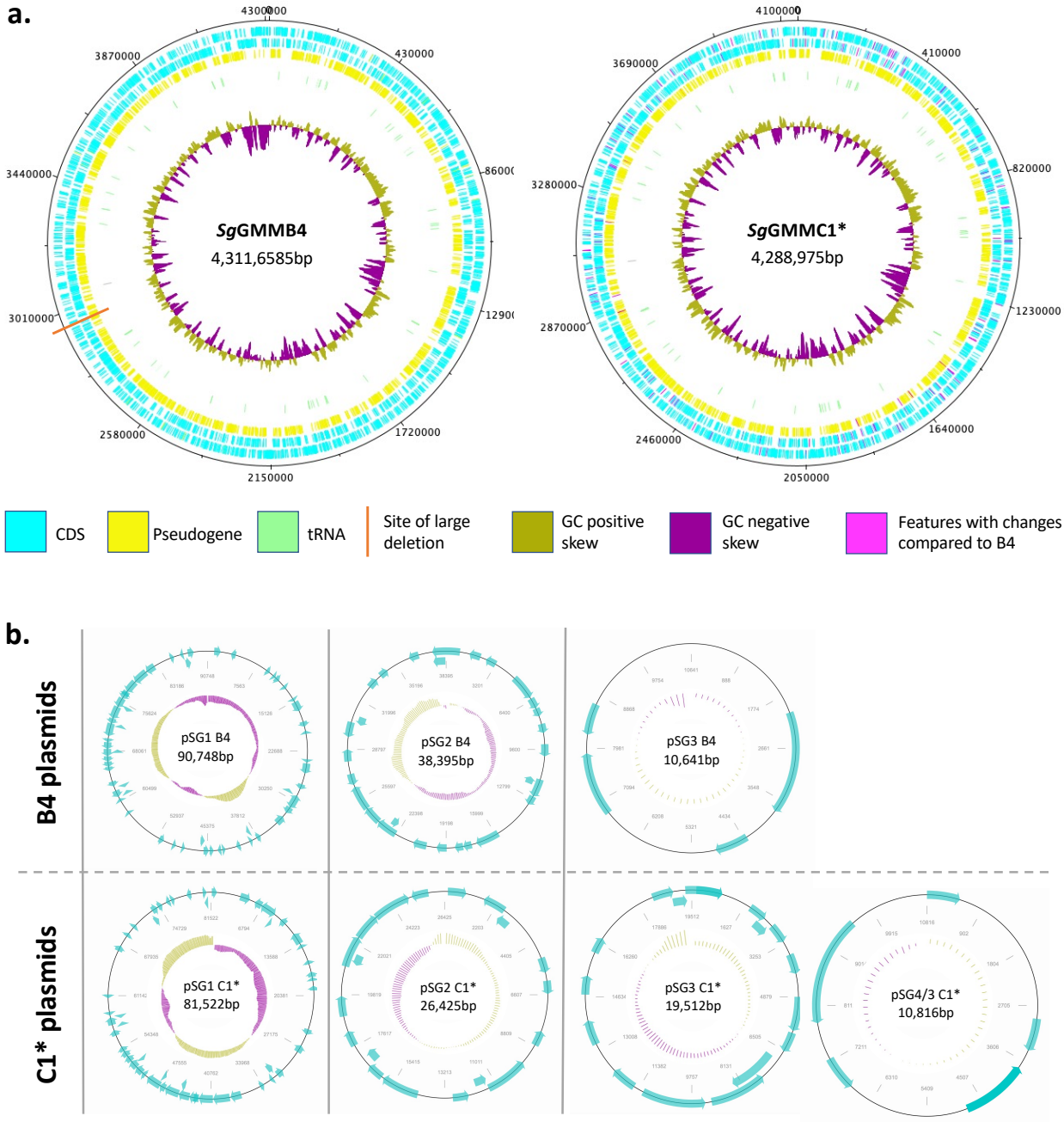


Figure 3-3 – Circular plots of *Sg*GMMB4 and *Sg*GMMC1* showing open reading frames and GC skews; a. chromosomes showing CDS, pseudogenes and tRNA, as well as the location of the 17 kb deletion; and b. plasmids showing open reading frames.

Snippy also identified 8 small deletions (range of 1-27 bp, mean 11.6 bp, standard deviation 6.08), 39 insertions (1-8 bp, mean 3.1 bp, standard deviation 4.72) and 10 SNPs for a total of 57 polymorphisms. Of these polymorphisms 32 resulted in changes to a total of 37 coding regions, as in some cases multiple genes were merged or frameshifted by a single mutation; 9/37 of affected genes were previously identified in *Sg*GMMB4 as pseudogenes. A full output table of polymorphisms in coding regions from Snippy can be found in Table 3-2.

In total 12/28 CDS genes affected by SNPs were predicted to be pseudogenised, with three in Tier 1, three in Tier 2, and six in Tier 3 (Table 3-2). Of the Tier 1 putative pseudogenes one was a hypothetical protein with no homology to known COGs, GO or conserved regions; the second was the transmembrane transport protein gene *hsrA_1* involved in multidrug resistance and quorum sensing, which has another pseudogenised copy elsewhere in the genome; and the third was a carbohydrate metabolism and membrane biosynthesis gene *manC1*. Putative pseudogenes classed as Tier 2 were the carbohydrate metabolism gene *pgl_2*, which has two other copies in the genome: the transmembrane transport gene *ydcV*, and an unknown gene homologous to cell division gene *DamX*. Finally, Tier 3 contained two putative pseudogenes involved in carbohydrate metabolism (*nagC_2* and *ptsG_1*) both of which have multiple other copies in the genome; one lytic phage gene; an energy production gene *sucB*; a membrane biosynthesis gene *wcaJ*; and a ribosomal structure gene *pheS*.

The likelihood of a mutation resulting in inactivation of the gene can be assessed by examining the predicted tertiary structure of the protein (Figure 3-4) [449]. This was performed for one Tier 1 gene *manC1* and two Tier 2 genes *pgl_2* and *ydcV*. In the Tier 1 gene *manC1* the first four alpha helix structures were removed and multiple beta strands were altered in the predicted protein structure, increasing the likelihood of this gene no longer producing a viable protein and thus being inactivated or pseudogenised. In both Tier 2 proteins examined the tertiary structure changes were less pronounced; in *pgl_2* there was an increase in disorder at the C terminus and truncation of one beta strand, but little alteration of the tertiary structure. In *ydcV* an alpha helix at the C terminus was truncated, leaving the remaining tertiary structure unaffected.

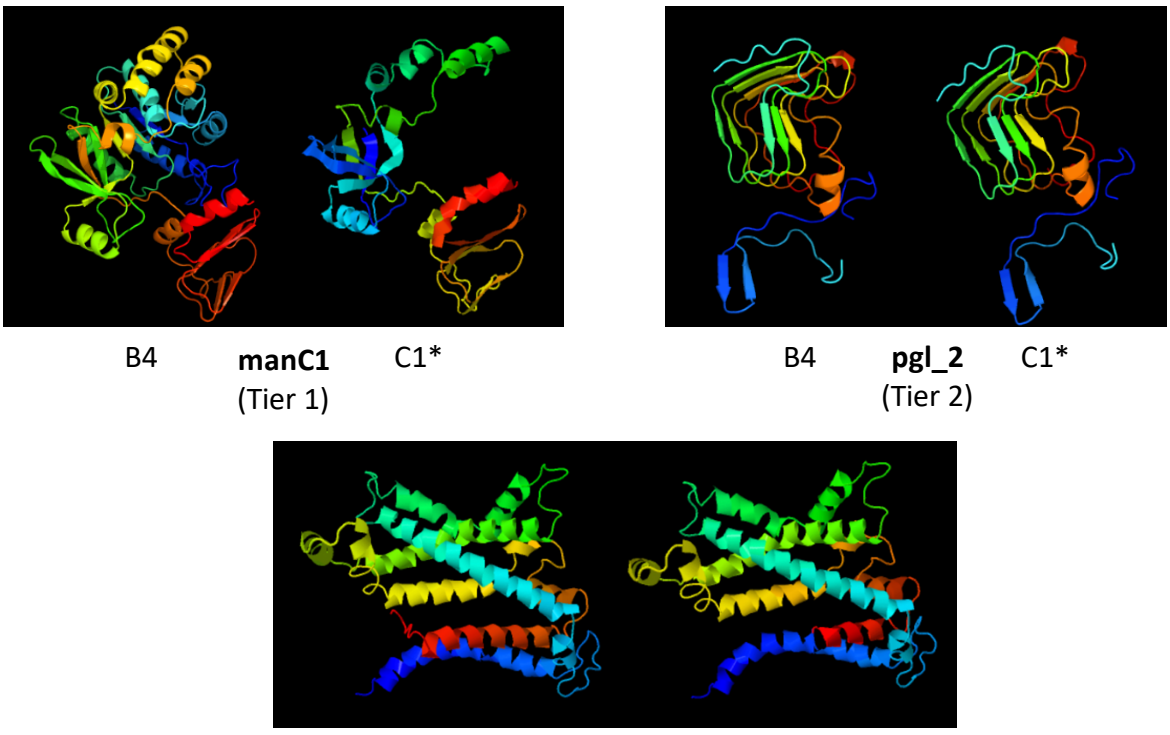


Figure 3-4 - Tertiary structures of predicted proteins from one Tier 1 (*manC*) and Tier 2 (*pgl_2* and *ydcV*) pseudogenised genes in *Sg*GMMB4 and *Sg*GMMC1*, illustrating the structural changes caused by SNPs. Coloured by rainbow from N to C terminus.

Table 3-1 – Genes within the 17,209 bp deleted region of *Sg*GMMC1*. Gene names and products were annotated using PROKKA; conserved orthologous groups were assigned using the STRING database; gene ontology and conserved regions were assigned using the Interpro database.
CDS – coding DNA sequence; pseudo – pseudogene; COG – clustered orthologous group; GO – gene ontology. Yellow – pseudogenes; blue – CDS.

CDS/ pseudo	COG		GO terms			Conserved regions
	Code	Description	Biological	Molecular	Cellular	
baeS_1 Signal transduction histidine-protein kinase BaeS						
pseudo	COG0642	Signal transduction histidine kinase	-	-	-	-
baeS_2 Signal transduction histidine-protein kinase BaeS						
pseudo	COG0642	Signal transduction histidine kinase	**GO:0007165** signal transduction	**GO:0000155** phosphorelay sensor kinase activity	**GO:0016021** integral component of membrane	**IPR003661** Signal transduction histidine kinase, dimerisation/phosphoacceptor domain **IPR003660** HAMP domain (sensor and chemotaxis)
baeS_3 Signal transduction histidine-protein kinase BaeS						
pseudo	COG0642	Signal transduction histidine kinase	-	-	-	**IPR005467** Histidine kinase domain
PROKKA_03996 Hypothetical protein						
pseudo	COG0826	Collagenase-like protease, PrtC family	-	-	-	-
yhbU_2 Putative protease YhbU precursor						
pseudo	COG0826	Collagenase-like protease, PrtC family	-	-	-	**IPR001539** Peptidase U32
PROKKA_03998 Peptidase family U32						
pseudo	COG0826	Collagenase-like protease, PrtC family	-	-	-	**IPR001539** Peptidase U32
thiM Hydroxyethylthiazole kinase						
CDS	COG2145	Hydroxyethylthiazole kinase, sugar kinase family	**GO:0009228** thiamine biosynthetic process	**GO:0004417** hydroxyethylthiazole kinase activity	-	**IPR000417** Hydroxyethylthiazole kinase
esaR_2 Transcriptional activator protein EsaR						

CDS/ pseudo	COG		GO terms			Conserved regions
	Code	Description	Biological	Molecular	Cellular	
CDS	COG2197	DNA-binding response regulator, NarL/FixJ family, contains REC and HTH domains	**GO:0006355** regulation of transcription, DNA-templated	**GO:0003677** DNA binding	-	**IPR000792** Transcription regulator LuxR, C-terminal
PROKKA_04001 Putative acetyltransferase						
CDS	COG0454	N-acetyltransferase, GNAT superfamily (includes histone acetyltransferase HPA2)	-	**GO:0008080** N-acetyltransferase activity	-	**IPR000182** GNAT domain
PROKKA_04002 Hypothetical protein						
CDS	COG0425	TusA-related sulfurtransferase	-	-	-	**IPR001455** TusA-like domain
PROKKA_04003 Hypothetical protein						
CDS	COG2391	Uncharacterized membrane protein YedE/YeeE, contains two sulfur transport domains	-	-	-	**IPR007272** Sulphur transport domain
PROKKA_04004 Putative inner membrane protein						
CDS	COG2391	Uncharacterized membrane protein YedE/YeeE, contains two sulfur transport domains	-	-	-	**IPR007272** Sulphur transport domain
yeeE_1 LysR substrate binding domain protein						
pseudo	COG0583	DNA-binding transcriptional regulator, LysR family	-	-	-	**IPR005119** LysR, substrate-binding
yeeE_2 LysR substrate binding domain protein						
pseudo	COG0583	DNA-binding transcriptional regulator, LysR family	-	-	-	**G3DSA:3.40.190.10** Periplasmic binding protein-like II
dthadh D-threo-3-hydroxyaspartate dehydratase						
pseudo	COG3616	D-serine deaminase, pyridoxal phosphate-dependent	-	-	-	**IPR042208** D-serine dehydratase-like domain
PROKKA_04008 Hypothetical protein						
pseudo	COG1288	Uncharacterized membrane protein YfcC, ion transporter superfamily	-	-	**GO:0016021** integral component of membrane	**IPR018385** C4-dicarboxylate anaerobic carrier-like **IPR018387** Uncharacterised protein YcgA/YfcC
PROKKA_04009 Hypothetical protein						
CDS	COG1288	Uncharacterized membrane protein YfcC, ion transporter superfamily	-	-	**GO:0016021** integral component of membrane	**IPR018385** C4-dicarboxylate anaerobic carrier-like

CDS/ pseudo	COG		GO terms			Conserved regions
	Code	Description	Biological	Molecular	Cellular	
						IPR018387 Uncharacterised protein YcgA/YfcC
yxeP_2 Putative hydrolase YxeP						
CDS	COG1473	Metal-dependent amidase/aminoacylase/carboxypeptidase	-	**GO:0016787** hydrolase activity	-	**IPR002933** Peptidase M20 **IPR017439** Aminohydrolase
PROKKA_04011 Hypothetical protein						
CDS	-	-	-	-	-	-
yabJ_2 Enamine/imine deaminase						
pseudo	COG0251	Enamine deaminase RidA, house cleaning of reactive enamine intermediates, YjgF/YER057c/UK114 family	-	-	-	**IPR006175** YjgF/YER057c/UK114 family
PROKKA_04013 Hypothetical protein						
pseudo	-	-	-	-	-	-
PROKKA_04014 Sulfite exporter TauE/SafE						
pseudo	COG0730	Uncharacterized membrane protein YfcA	-	-	**GO:0016021** integral component of membrane	**IPR002781** Transmembrane protein TauE-like
PROKKA_04015 Sulfite exporter TauE/SafE						
CDS	COG0730	Uncharacterized membrane protein YfcA	-	-	**GO:0016021** integral component of membrane	**IPR002781** Transmembrane protein TauE-like
PROKKA_04016 Hypothetical protein						
CDS	-	-	-	-	-	-
PROKKA_04017 PTS system, lactose/cellobiose specific IIB subunit						
CDS	COG3414	Phosphotransferases system, galactitol-specific IIB component	**GO:0009401** phosphoenolpyruvate-dependent sugar phosphotransferase system	**GO:0008982** protein-N(PI)-phosphohistidine-sugar phosphotransferase activity	-	**IPR003501** Phosphotransferase system, EIIB component, type 2/3
gatC Galactitol permease IIC component						

CDS/ pseudo	COG		GO terms			Conserved regions
	Code	Description	Biological	Molecular	Cellular	
pseudo	COG3775	Phosphotransferases system, galactitol-specific IIC component	-	-	-	**IPR004703** Phosphotransferase system, sugar-specific permease component **IPR013853** Galactitol permease IIC component
PROKKA_04019 Hypothetical protein						
CDS	-	-	-	-	-	-
PROKKA_04020 PTS system galactitol-specific transporter subunit IIA						
CDS	COG1762	Phosphotransferase system mannitol/fructose-specific IIA domain (Ntr-type)	-	-	-	**IPR002178** PTS EIIA type-2 domain
PROKKA_04021 Hypothetical protein						
CDS	NOG33103	non supervised orthologous group	-	-	-	-
fabG_4 3-oxoacyl-[acyl-carrier-protein] reductase FabG						
pseudo	COG1028	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family	-	**GO:0016491** oxidoreductase activity	-	**IPR002347** Short-chain dehydrogenase/reductase SDR
PROKKA_04023 Transposase IS116/IS110/IS902 family protein						
pseudo	COG3547	Transposase	**GO:0006313** transposition, DNA-mediated	**GO:0003677** DNA binding **GO:0004803** transposase activity	-	**IPR003346** Transposase, IS116/IS110/IS902
PROKKA_04024 Transposase						
pseudo	COG3547	Transposase	**GO:0006313** transposition, DNA-mediated	**GO:0003677** DNA binding **GO:0004803** transposase activity	-	**IPR003346** Transposase, IS116/IS110/IS902
PROKKA_04025 Hypothetical protein						
pseudo	COG3547	Transposase	-	-	-	-

Table 3-2 - Snippy output with all insertions, deletions and single-nucleotide polymorphisms in *Sodalis glossinidius* strain *Sg*GMMC1*, in comparison with ancestor strain *Sg*GMMB4. Gene names and products were annotated using PROKKA; conserved orthologous groups were assigned using the STRING database; gene ontology and conserved regions were assigned using the Interpro database.
*** - Tier 1 pseudogenisation; ** - Tier 2 pseudogenisation; * - Tier 3 pseudogenisation.
CDS – Coding DNA sequence. Pseudo – pseudogene. Ins – insertion. Del – deletion. Com – complex polymorphism. Snp – single nucleotide polymorphism. AA – amino acid. B – biological process. M – molecular process. C – cellular process. Yellow – pseudogenes; blue – CDS; orange – potential novel pseudogenes, shaded by tier.

B4 position	Type	CDS/ pseudo	GO Terms		COGs		SNP effect	Disrupted conserved regions
— Chromosome —								
***** **PROKKA_00037** hypothetical protein								
26,577	ins	CDS	-	-	-	-	Frameshift truncates at AA 59/199	-
***** **hsrA_1** putative transport protein HsrA								
351,610	ins	CDS	GO:0055085	B: transmembrane transport	COG0477	MFS family permease	Frameshift removes hsrA_1 on antisense strand, replaces with two overlapping hypotheticals on sense strand (no conserved regions on new genes)	**IPR011701**; **PF07690** (major facilitator superfamily) **IPR020846**; **PS50850** (major facilitator superfamily profile) **IPR036259**; **SSF103473** (MFS general substrate transporter) **PTHR23501:SF180** (multidrug resistance protein B homolog) **G3DSA:1.20.1720.10** (multidrug resistance protein D)
			GO:0022857	M: transmembrane transporter activity				
PROKKA_00810 Tail fiber protein gp37 C terminal								
588,917	ins	pseudo	-	-	COG4675	Microcystin-dependent protein (function unknown)	Frameshift truncates at AA 171/260	**IPR022246; PF12604** (bacteriophage tail fibre protein gp37 C terminal family)
nnr_3 NAD(P)H-hydrate repair enzyme Nnr								
619,593	ins	pseudo	GO:0052855	M: ADP-dependent NAD(P)H-hydrate dehydratase activity	COG0062	Carbohydrate transport and metabolism; NADHX epimerase activity	Frameshift removes first 53/119 AAs and alters AAs 54-73	**IPR000631; PF01256** (carbohydrate kinase family) **PS51383** (YjeF C-terminal domain) **IPR029056** (ribokinase-like superfamily)
					COG0063	Carbohydrate transport and metabolism		

B4 position	Type	CDS/ pseudo	GO Terms		COGs		SNP effect	Disrupted conserved regions
proP_2/proP_1 Proline/betaine transporter								
869,745	ins	CDS	GO:0055085	M: ADP-dependent NAD(P)H-hydrate dehydratase activity	COG0477	Carbohydrate transport and metabolism, major facilitator superfamily	Removes prop_2 stop codon, merging prop_2 and prop_1	All conserved regions intact
			GO:0022857	M: transmembrane transporter activity				
			GO:0016021	C: integral component of membrane				
** **pgl_2** Polygalacturonase								
1,048,509	ins	CDS	GO:0005975	B: carbohydrate metabolic process	COG2706	Carbohydrate transport and metabolism, 6-phosphogluconolactonase	Frameshift truncates at AA 171/182	**IPR000743; PF00295** (glycoside hydrolase family 28) **IPR011050; SSF51126** (pectin lyase-like superfamily)
			GO:0004650	M: polygalacturonase activity				
* **PROKKA_01603** Bacteriophage lysis protein								
1,215,324	com	CDS	-	-	NOG26972	DNA-packaging protein gp3	Missense at AA 23, R to K	**IPR032066; PF16677** (DNA-packaging protein gp3)
1,215,387	snp	CDS	-	-	NOG26972	DNA-packaging protein gp3	Synonymous	-
* **nagC_2** N-acetylglucosamine repressor								
1,424,493	snp	CDS	GO:0006355	B: regulation of transcription, DNA-templated	COG1940	Carbohydrate transport and metabolism, ROK family	Missense at AA 263/419, C to F	**IPR000600; PF00480** (ROK family) **IPR043129; SSF53067** (actin-like ATPase domain)
			GO:0003700	M: DNA-binding transcription factor activity				
* **ptsG_1** PTS system glucose-specific EIICBA component								
1,427,676	snp	CDS	GO:0009401	B: phosphoenolpyruvate-dependent sugar phosphotransferase system	COG1263	carbohydrate transport and metabolism, PTS system	Missense at AA 28/678, M to I	**IPR010974; TIGR01998** (N-acetylglucosamine-specific PTS transporter subunit IIBC) **IPR013013; PS51103** (phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS)) **IPR003352; PF02378** (phosphotransferase system, EIIC) **PTHR30009** (cytochrome c-type synthesis protein and PTS transmembrane component)
			GO:0008982	M: protein-N(PI)-phosphohistidine-sugar phosphotransferase activity	COG1264	carbohydrate transport and metabolism, PTS system		
			GO:0015572	M: N-acetylglucosamine transmembrane transporter activity	COG2190	carbohydrate transport and metabolism, PTS system		
			GO:0019866	C: organelle inner membrane				
			GO:0016021	C: integral component of membrane				
			GO:0016020	C: membrane				

B4 position	Type	CDS/ pseudo	GO Terms		COGs		SNP effect	Disrupted conserved regions
* **sucB** Component of 2-oxoglutarate dehydrogenase								
1,452,053	snp	CDS	GO:0006099	B: tricarboxylic acid cycle	COG0508	Energy production and conversion, dehydrogenase	Missense at AA 274/396, P to S	**IPR006255; TIGR01347** (dihydrolipoyllysine-residue succinyltransferase) **IPR001078; PF00198** (2-oxoacid dehydrogenases acyltransferase (catalytic domain)
			GO:0016746	M: acyltransferase activity				
			GO:0004149	M: dihydrolipoyllysine-residue succinyltransferase activity				
			GO:0045252	C: oxoglutarate dehydrogenase complex				
** **ydcV** Inner membrane transporter protein YdcV								
1,535,773	ins	CDS	GO:0055085	B: transmembrane transport	COG1177	Inorganic ion transport and metabolism; binding-protein-dependent transport systems inner membrane component	Missense at AA 264 (V to G) and truncated at AA 266/281	**IPR035906** (MetI-like superfamily) **PTHR43848** (Putrescine transport system permease protein)
			GO:0016020	C: membrane				
PROKKA_02100 hypothetical protein								
1,566,260	ins	CDS	-	-	COG5464	Transposase, function unknown	Two hypothetical genes merged	All conserved regions intact
* **wcaJ** UDP-glucose:undecaprenyl-phosphate glucose-1-phosphate transferase								
1,638,864	snp	CDS	-	-	COG2148	Cell wall / membrane / envelope biogenesis, sugar transferase	Missense at AA 374/464, G to A	**IPR017473; TIGR03023** (undecaprenyl-phosphate glucose phosphotransferase) **IPR017475; TIGR03025** (exopolysaccharide biosynthesis polyprenyl glycosylphosphotransferase) **IPR003362; PF02397** (bacterial sugar transferase) **PTHR30576** (colonic biosynthesis UDP-glucose lipid carrier transferase)
zapC_2 Cell division protein ZapC								
1,701,462	ins	pseudo	-	-	NOG01298	Cell division	Frameshift removed gene	**IPR009809; PF07126** (Cell-division protein ZapC)
*** **manC1** Mannose-1-phosphate guanylyltransferase 1								
1,867,898	ins	CDS	GO:0009058	B: biosynthetic process	COG0662	Carbohydrate transport and metabolism	Frameshift truncates at AA 129/471 and moves AAs 139-471 onto next frame with closest promoter >50bp	**IPR006375; TIGR01479** (mannose-1-phosphate guanylyltransferase/mannose-6-phosphate isomerase) **IPR005835; PF00483** (nucleotidyl transferase)
			GO:0000271	B: polysaccharide biosynthetic process	COG0836	Cell wall / membrane / envelope biogenesis		
			GO:0005976	B: polysaccharide metabolic process				

B4 position	Type	CDS/ pseudo	GO Terms		COGs		SNP effect	Disrupted conserved regions
			GO:0016779	M: nucleotidyltransferase activity			upstream from second half	**IPR029044** (nucleotide-diphospho-sugar transferases)
PROKKA_02481 hypothetical protein								
1,896,421	ins	CDS	-	-	-	-	Frameshift at AA 220/246, alters remaining AAs and extends to 250	-
PROKKA_02937 hypothetical protein								
2,205,206	ins	pseudo	-	-	-	-	Frameshift alters from AA 42/46, extends to 49	All conserved regions intact
shlA Hemolysin precursor **PROKKA_03153** hypothetical protein								
2,360,326	ins	CDS	-	-	COG3210	Intracellular trafficking, secretion and vesicular transport	Frameshift at AA 233/248 of shlA alters last 15 AAs and adds PROKKA_03153 unchanged onto end	All conserved regions intact
		CDS	-	-	-	-		All conserved regions intact
PROKKA_03155 hypothetical protein **PROKKA_03156** hypothetical protein								
2,361,515	ins	CDS	-	-	-	-	Frameshift merges PROKKA_03155 and PROKKA_03156 with 17 extra AAs between them	All conserved regions intact
		CDS	-	-	-	-		All conserved regions intact
* **pheS** Phenylalanine--tRNA ligase alpha subunit								
2,367,309	snp	CDS	GO:0043039	B: tRNA aminoacylation	COG0016	Translation, ribosomal structure and biogenesis	Missense at AA 59/327, V to G	**IPR022911; MF_00281** (Phenylalanine—tRNA ligase alpha subunit pheS) **IPR004529**; **x** (Phenylalanine—tRNA ligase alpha subunit) **IPR004188; PF02912** (Aminoacyl tRNA synthetase class II, N-terminal domain) **IPR010978; SSF46589** (tRNA-binding arm) **G3DSA:3.30.930.10** (Bira bifunctional protein; domain 2)
			GO:0006432	B: phenylalanyl-tRNA aminoacylation				
			GO:0005524	M: ATP binding				
			GO:0000049	M: tRNA binding				
			GO:0004812	M: aminoacyl-tRNA ligase activity				
			GO:0000166	M: nucleotide binding				
			GO:0004826	M: phenylalanine-tRNA ligase activity				
			GO:0005737	C: cytoplasm				
slyA_2 Transcriptional regulator SlyA								
2,407,904	ins	CDS	GO:0006355	B: regulation of transcription, DNA-templated	COG1846	Transcription regulator	Insertion 50bp in front of gene, unaltered AA sequence but moves	All conserved regions intact

B4 position	Type	CDS/ pseudo	GO Terms		COGs		SNP effect	Disrupted conserved regions
			GO:0003700	M: DNA-binding transcription factor activity			nearest start codon from 50bp away to 93bp away	
* **fnr** Fumarate and nitrate reduction regulatory protein								
2,467,583	snp	CDS	GO:0006355	B: regulation of transcription, DNA-templated	COG0664	Signal transduction mechanisms, transcriptional regulator, crp fnr family	Missense at AA 52/251, P to T	**IPR000595; cd00038** (effector domain of the CAP family of transcription factors), **PF00027** (cyclic nucleotide-binding domain), **PS50042** (cAMP/cGMP binding motif profile) **IPR014710** (RmlC-like jelly roll fold) **IPR018490** (cyclic nucleotide-binding-like) **PTHR24567** (CRP family transcriptional regulatory protein)
			GO:0003700	M: DNA-binding transcription factor activity				
			GO:0003677	M: DNA binding				
quiA_1 Quinate/shikimate dehydrogenase (qui-)								
2,562,282	ins	pseudo	-	-	COG4993	Carbohydrate transport and metabolism, dehydrogenase	Frameshift adds 22 AAs to start of gene	All conserved regions intact
PROKKA_03881 hypothetical protein								
2,873,885	ins	pseudo	-	-	NOG297256	Non supervised orthologous groups	Frameshift at 232/844 breaking into two proteins	-
chaA_2 Sodium/proton antiporter ChaA **chaA_3**								
3,207,262	ins	pseudo	-	-	COG0387	Inorganic ion transport and metabolism, calcium proton	Frameshift merges chaA_2 and chaA_3 with an extra 7 AAs between	All conserved regions intact
		pseudo	-	-				All conserved regions intact
PROKKA_04953 NMT1/THI5 like protein								
3,632,863	ins	CDS	-	-	-	-	Frameshift truncates at AA 269/291	All conserved regions intact
* **PROKKA_05204** hypothetical protein								
3,830,347	snp	CDS	GO:0005975	B: carbohydrate metabolic process	COG0726	Carbohydrate transport and metabolism, 4-amino-4-deoxy-alpha-L-arabinopyranosyl undecaprenyl phosphate biosynthetic process	Missense at AA 252/324, A to E	**G3DSA:3.20.20.370** (glycoside hydrolase/deacetylase) **cd10935** (putative catalytic domain of lipopolysaccharide biosynthesis protein WalW and its bacterial homologs)
			GO:0003824	M: catalytic activity				
yjiR_3 putative HTH-type transcriptional regulator YjiR								
3,962,584	ins	pseudo	GO:0003824	M: catalytic activity	COG1167	Transcriptional regulator, gntR family	Frameshift removes first 46/103 AAs	**IPR015422; G3DSA:3.90.1150.10** (aspartate aminotransferase, domain 1) **IPR015424; SSF53383** (PLP-dependent transferases)

B4 position	Type	CDS/ pseudo	GO Terms		COGs		SNP effect	Disrupted conserved regions
								PTHR42790:SF7 (transcriptional regulator-related)
** **PROKKA_05385** hypothetical protein								
3,967,547	ins	CDS	GO:0032506	B: cytokinetic process	COG3266	Function unknown, DamX-related protein	Frameshift removes first 38/288 AAs	**IPR032899; MF_02021** (Cell division protein DamX)
			GO:0042834	M: peptidoglycan binding				
			GO:0030428	C: cell septum				
— pSG4/3 —								
SGGMMB4_05911 hypothetical protein								
636	ins	CDS	-	-	-	-	Frameshift at AA 58/64, extends to 173	-
** **SGGMMB4_05912** hypothetical protein								
3,680	ins	CDS	GO:0008237	M: metallopeptidase activity	COG3291	PKD repeat	Frameshift truncates at AA 560/573	**IPR024079**; **G3DSA:3.40.390.10** (collagenase, catalytic domain) **SSF55486** (Metalloproteases ("zincins"), catalytic domain) **PF13688** (Metallo-peptidase family M12)
SGGMMB4_05921 hypothetical protein								
10,140	del	CDS	-	-	-	-	Frameshift truncates at AA 92/106	-
10,269	snp						Missense at AA 50, T to P	

3.6 Discussion

3.6.1 Deletions and polymorphisms are occurring in common functional groups

Similar gene groups were found within the 17 kb deletion and annotated by Snippy as containing small polymorphisms, indicating a common selection pressure for mutation and deletion. Within the 17 kb deletion most coding genes and pseudogenes were involved in transmembrane transport of organic and inorganic compounds, mainly of carbohydrates, or were membrane-bound transferases. Carbohydrate transmembrane transporters were also the most common group of genes containing small mutations, including those involved in transport of proline, mannose, and a GlcNAc repressor. Genes involved in protein modification, DNA transcription and regulation, and membrane production were also commonly mutated or deleted. Some gene groups appear to be in the first phase of accumulating small mutations before being deleted, including two phage lytic phase associated genes as well as genes involved in energy production and hemolysis.

3.6.2 Increasing co-dependency between *Sodalis* and *Wigglesworthia*

The Black Queen Hypothesis describes the co-dependencies between associated organisms in symbiotic relationships; these co-dependencies are formed by reductive genomic evolution driven by repeated population bottlenecks in a restricted environment [443]. An obligate mutualism of two bacteria within glassy sharpshooters has been described whereby two unrelated bacteria have evolved entirely complementary genomes to provide vitamins and amino acids respectively to their hosts and each other [503]. Our results suggest this could be the evolutionary

trajectory of *Sodalis* to become increasingly compatible with, and eventually entirely reliant on, both the tsetse host and *Wigglesworthia.*

In *Sg*GMMC1* multiple sulfur-associated transmembrane proteins were deleted; sulfur is used in the biosynthesis and modification of sulfur-containing amino acids, methionine and cysteine [504]. Methionine and cysteine are both produced by *Wigglesworthia* for the tsetse host [505], and the degradation of their production pathways in *Sodalis* have been predicted as it moves towards a symbiotic lifestyle [506]. Assuming the deleted sulfur and sulfite transmembrane proteins facilitated sulfur compound uptake, their deletion adds evidence to suggest degradation of methionine and cystem biosynthesis is occurring.

Thiamine is an essential vitamin for tsetse development that is not present in regular blood meals and is synthesised by the obligate endosymbiont *Wigglesworthia* [224, 485, 486]. Serially passaging *Sodalis* in thiamine-rich media *in vitro,* with the repeated population bottlenecks caused by passaging, resulted in the deletion of the gene for an essential cofactor for producing thiamine, *thiM*. This is a continuation of the degradation of the thiamine-production pathway which had already begun in the ancestral *Sg*GMMB4 strain [108] and is predicted to continue in an *in silico* experimental evolution study [486]. In tsetse *Wigglesworthia* and *Sodalis* are hypothesised to contribute complementary cofactors to produce an intact thiamine biosynthesis pathway [108, 486].

3.6.3 Adaptation is driven by environmental factors

While continuation of previously observed or predicted gene loss was seen in the *Sg*GMMC1* genome, impacts of the *in vitro* culture conditions on the genome were also observed. The loss of one *esaR* copy, which regulates quorum sensing gene transcription, could indicate that quorum sensing is less vital in a controlled, single-species and frequently passaged *in vitro* culture system than in a complex, multi-species host. The gene *hsrA*, which is putatively pseudogenised in *Sg*GMMC1* at the Tier 1 level, encodes a membrane transporter protein yieO which is putatively involved in toxin efflux during biofilm formation under stress [507]. The pseudogenisation of this gene may be due to the removed selection pressure of host immune system stresses. The deletion of multiple transmembrane transferase proteins indicates a change in the relation of *Sg*GMMC1* to its environment; in the tsetse host *Sodalis* lives a dual lifestyle as both an intracellular and extracellular symbiont, requiring the retention of virulence genes such as a TTSS to infect host cells [60]. This loss of quorum sensing and transmembrane genes may be a functional adaptation to the solely extracellular lifestyle of the *Sg*GMMC1* strain. Further work to confirm the activity and expression of these membrane proteins would be necessary to support this hypothesis.

3.6.4 Classification and implications of novel pseudogenes

It is possible for frameshifted pseudogenes to be functionally corrected by polymerase infidelity in homopolymeric tracts (where multiple bases of the same type occur in a row) during transcription to produce functional mRNA strands [508]. This phenomenon is common in endosymbionts, which contain large numbers of poly(A) tracts linked to their high A+T content, which has been well documented in the aphid

endosymbiont *Buchnera aphidicola* [508]. Because of this possibility, the threshold for classifying a gene as truly pseudogenised is less straightforward than identifying it as such in the genome [488]. One study of pseudogene expression in *Sodalis* observed a single insertion in one copy of the *prop_2* gene at position 869,745 which caused a frameshift; however, this frameshift was functionally corrected during transcription, therefore the gene was not annotated as a pseudogene [489]. In this study we observed another single-base insertion in the same copy of *prop_2* in the same homopolymeric position, which caused an additional frameshift. This supports the theory that mutations accumulate in genes that lack a selection pressure for their maintenance, eventually pseudogenising and inactivating the gene before its eventual deletion.

In this study a tiered system of classification for potentially pseudogenised genes has been used to describe the differing potential of mutations to pseudogenised a gene depending on the preservation of conserved regions, and changes to predicted tertiary structures were used to back up these assertations. Protein tertiary structure changes may affect the ability of the protein to bind substrates or catalyse reactions efficiently [509, 510]. The Tier 1-classified pseudogene *manC1* was severely truncated in *Sg*GMMC1*, causing large structural changes and increasing the likelihood of the produced protein being either unable to catalyse the production of GDP-mannose entirely, or at a reduced rate. The examined Tier 2 genes, however, had a largely unchanged tertiary structure with mostly intact alpha helices and beta strands, allowing the proteins to fold normally and therefore making the proteins more likely to perform normally. While this gives more information as to the potential impact of mutations and supports the tiered methodology of classifying potential

pseudogenes, more information on the protein effector sites and transcriptional modifications would be necessary to determine the full effect of each mutation.

3.6.5 Conclusions and further work

The pseudogenisation and deletion of genes described above support the hypothesis of symbiont genome size reduction through pseudogenisation and subsequent deletion of genes unnecessary to survival in a stable host environment [488]. In the case of *Sg*GMMC1*, 10 years of passaging in an extremely stable, nutrient-rich media with carefully controlled temperature and atmosphere has resulted in inactivation or deletion of genes unnecessary to this environment. This adaptation shows that *Sodalis* is in the process of switching lifestyles from free-living to obligately symbiotic, but currently retains the flexibility to continuously adapt while reducing its genome size. The serial passaging of *Sodalis* imitates the repeated population bottlenecks in natural host-symbiont systems, which are known to contribute to symbiont genome reduction. The *in vitro* passages were repeated every 10-14 days, whereas the lab-reared tsetse life cycle takes more than four weeks. The higher frequency of passaging *in vitro* than transmission of *Sodalis* between tsetse *in vivo* may have increased the bottlenecking effect on genome degradation.

Future work should address the effect of additional selection pressures more closely replicating those found in the tsetse host on the evolution of the *Sodalis* genome towards obligate symbiosis. The *Sg*GMMC1* strain should also be evaluated for its ability to infect tsetse hosts and its impact on tsetse refractoriness to trypanosome infection. The pseudogenes identified in this study should be validated using

transcriptomics to determine whether they are being functionally corrected during transcription.

3.7 Extended data

Raw Nanopore and MiSeq reads: ‘Sequencing an experimentally evolved strain of *Sodalis glossinidius*.’ **PRJEB23231**

Assembled and annotated genome: ‘Annotated genome of experimentally evolved *Sodalis glossinidius* strain C2.’ **DOI 10.17866/rd.salford.8052437**

Chapter 4 Elucidating the role of *Sodalis glossinidius* in the refractoriness of *Glossina morsitans morsitans* to *Trypanosoma congolense* using metatranscriptomics

4.1 Abstract

Glossina morsitans morsitans tsetse flies are the main vector species for *Trypanosoma congolense*, a causative agent of animal African trypanosomiasis. Tsetse flies are naturally resistant to trypanosome infection, but the presence of the secondary endosymbiont *Sodalis glossinidius* has been implicated in increasing tsetse susceptibility to trypanosome infection. The mechanism of this effect is unknown, but *Sodalis* is known to interact with the tsetse host, the trypanosome parasite, and other tsetse-associated bacteria.

In this chapter, transcriptional differences in *Serratia, Wolbachia*, *Sodalis* and *Wigglesworthia* in tsetse that are challenged with trypanosome infection are studied, to clarify the mechanism by which tsetse flies are made susceptible to trypanosome infection. Lab-reared tsetse flies were experimentally infected with *T. congolense* and sampled at early, mid and late infection stages. Total RNA was extracted from tsetse flight muscle and midguts, as well as from pure *Sodalis in vitro* cultures, for sequencing and analysis. *Wigglesworthia* showed transcriptional differences within infected flies over the course of the infection, whereas *Wolbachia* and *Serratia* both showed little or no differential expression in any treatment group. In *Sodalis* phage genes were strongly upregulated *in vitro* compared to *in vivo*, and differential expression between trypanosome-refractory and infected tsetse flies

indicate a role for *Sodalis* pathogenicity-related genes in increasing tsetse susceptibility to trypanosome infection. This indicates that examining *Sodalis* pathogenicity may inform novel control strategies for trypanosomiasis.

4.2 Introduction

4.2.1 Tsetse microbiota and trypanosome infection

Tsetse flies are naturally refractory (resistant) to infection with trypanosomes, with some variation in susceptibility attributed to host age, sex, reproductive status, the trypanosome strain and the presence or absence of other tsetse-associated microorganisms [209, 511]. Teneral – newly-emerged and unfed – tsetse flies are more susceptible to trypanosome infection (see Section 1.5.2). However, infection rates of wild tsetse in areas with an abundance of infected hosts are still as low as <1% [211, 512].

The microbiota of tsetse flies is of interest due to the presence of symbiotic bacteria at various stages of evolution from free-living lifestyles, as well as for its potential uses in vector control strategies. Until recently the bacterial microbiome of tsetse flies was believed to be restricted to three taxa: *Wigglesworthia glossinidia* [225], *Sodalis* [242] and *Wolbachia* [238, 239]; but recent studies have discovered a wider microbiome including *Serratia* (see Chapter 2 and Table 1-4). Insect symbionts range in their location within the host, known as their tissue tropism; more closely-associated primary symbionts are obligately intracellular, with some including *Wigglesworthia* existing in specialised host cells known as bacteriocytes [225]. *Sodalis*, which is a more recently-evolved symbiont of tsetse flies (see Chapter 3 for a description of the evolutionary pathway of *Sodalis* from free-living to symbiotic), has a broad tissue tropism and is found both intracellularly and extracellularly throughout the tsetse fly: concentrated in the gut, ovaries, salivary glands and milk glands, as well as muscle and fat body in older flies [104].

Sodalis is believed to increase tsetse susceptibility to trypanosome infection, although the mechanism for this is disputed – see Section 1.5.3.3. *Sodalis* is acquired by the larva through maternal milk and multiplies in the last milk meal retained in the pupal midgut, to produce an accumulation of factors such as GlcNAc in the teneral fly potentially explaining the teneral phenomenon [264]. However, the link between *Sodalis* and increased trypanosome infection has not been consistently observed in surveys of wild tsetse populations which demonstrate a complex relationship between the three species [102, 256, 396, 441]. *Sodalis* has been demonstrated to have potential as a paratransgenic control agent, whereby an anti-trypanosomal effector molecule is delivered to tsetse via a transgenic *Sodalis* strain [265, 266]. Without a thorough understanding of the mechanisms behind *Sodalis* increasing tsetse susceptibility to trypanosome infection, the use of transgenic *Sodalis* may not be sufficient to control trypanosome spread.

Serratia is a genus of Gram-negative bacteria with widespread parasitic and mutualistic associations in multiple animal hosts, including the *Buchnera aphidicola* endosymbiont *S. symbiotica* [61]. *S. marcescens* has been identified as a pathogen of tsetse flies [418], and *S. glossinae* was first discovered in the tsetse midgut microbiome [269]. Although the role of some *Serratia* species in the tsetse fly have not been described the genus is of particular interest due to the impact *S. marcescens* has in the *T. cruzi* vector *R. prolixus*, where it displays trypanolytic activity via production of the pigment prodigiosin [413].

4.2.2 Metatranscriptomics in tsetse

The simultaneous study of transcriptional signatures from all organisms within a host is known as metatranscriptomics, and allows the study of functional roles of microbiota within a host [513]. *Sodalis* transcriptional signatures in trypanosome-challenged flies have previously been examined in *G. palpalis gambiensis* infected with *T. brucei gambiense*, the causative agent of chronic HAT, using microarray technology. Soumana *et al*. (2014) found trypanosome challenge changed *Sodalis* gene expression regardless of the infection outcome and discovered potential involvement of *Sodalis* prophage elements in mediating tsetse susceptibility to trypanosomes [514]. However, sequencing all mRNA present in a sample via the more sensitive tool RNAseq can give more precise data on an entire metaorganism, allowing the study of interactions between individual species within a host [515, 516] This has been performed in teneral tsetse flies in a recent study; *Sodalis* genes involved in processing chitin were upregulated in *Gmm*, a highly competent trypanosome vector, compared to the low competence vector *G. brevipalis* [517]. These results provide evidence to suggest that the breakdown of chitin by *Sodalis* to produce the energy source GlcNAc, which inhibits trypanocidal lectins and facilitates trypanosome invasion of the peritrophic membrane, is responsible for increasing tsetse susceptibility to trypanosome infection [264].

4.3 Aims and objectives

4.3.1 Aims

In this study the transcriptome of the main AAT agent *T. congolense*, its most common vector *Gmm* [256, 518], and the associated bacteria *Wigglesworthia, Sodalis, Wolbachia* and *Serratia* are examined using RNA-seq. Transcriptional changes are observed over the course of a trypanosome infection, in exposed but refractory tsetse and successfully infected tsetse. In particular, this study aimed to understand the impact of *Sodalis* on tsetse permissiveness to trypanosome infection. The *T. congolense* lifecycle occurs in the midgut and proboscis of *Gmm* [260], rather than the salivary glands (see Figure 1-5). Samples were taken from midgut tissue, where direct contact occurs with trypanosomes, and compared to transcription in flight muscles, where no contact with trypanosomes occur, in order to examine the effect of direct trypanosome contact separately from impacts of the infected host. With this experimental design we aimed to identify transcriptional signatures of infection or refractoriness to trypanosomes in both the host and the commensal bacteria.

The transcriptomes of *Sodalis* populations in midgut, flight msucles and *in vitro* cultures of *Sodalis* were compared, to identify potential *Sodalis*-mediated causes of the increase in tsetse susceptibility to trypanosome infection and to better understand the metabolic profile of *Sodalis in vitro* and *in vivo*.

4.4 Methodology

4.4.1 Tsetse colony origin and maintenance

The *Gmm* colony at the Liverpool School of Tropical Medicine was established from the Bristol colony in 2002, which was itself established from wild Zimbabwean flies (personal communication, LH). The colony was maintained at 26°C and 80% relative humidity and fed three times a week on defibrinated horse blood through a silicon membrane. Tsetse were anaesthetised by chilling at 4°C or on ice before sorting by sex or dissecting. Males were identified for inclusion in the experiment by examining the genitals located at the tip of the abdomen (Figure 4-1). Males have a visible external armature named the hypopygium, which contains a folded clasper and phallosome. Females have genital plates and a longer, slightly more pointed abdomen.

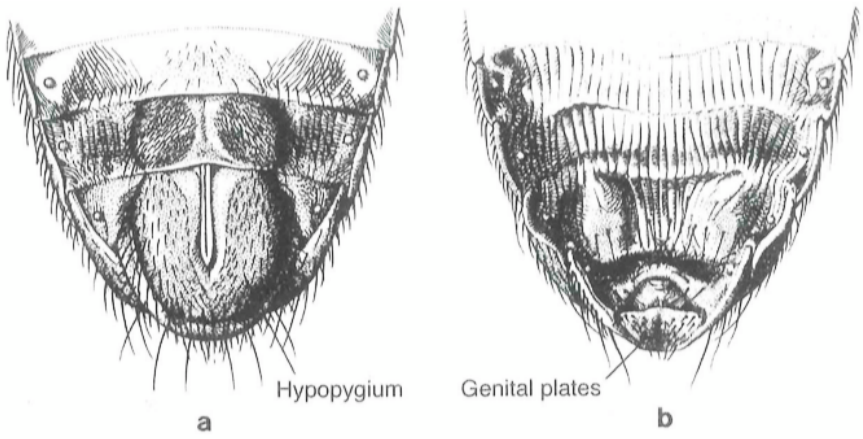


Figure 4-1 -Tsetse external genitalia illustrating the difference between a) males, which have a visible external armature called a hypopygium; and b) females, which have a pointed abdomen with external genital plates. From [1].

4.4.2 Experimental design

A visual representation of the experimental design and the testing groups can be found in Figure 4-2 and Table 4-1. Male tsetse flies which had eclosed within the same 24-hour period were separated from females and kept in groups of approximately 15 per cage, in normal insectary conditions without feeding, until 24-48 hours old. At this point (Day 0) 100 males were fed an uninfected blood meal while 300 males were fed an infected blood meal, as described below. Any unfed flies were discarded. Cages with experimental flies were kept in the insectary separately from the main colony and fed three times a week. Experimental flies were not fed in the 48 hours before dissection to avoid contamination of the samples with blood. Time points for dissection were decided based on the stages of *T. congolense* development in *Gmm*: early-stage infection (Day 3), mid-stage infection (Day 10) and late-stage infection (Day 30), based on previous studies using the same strain of *T.*

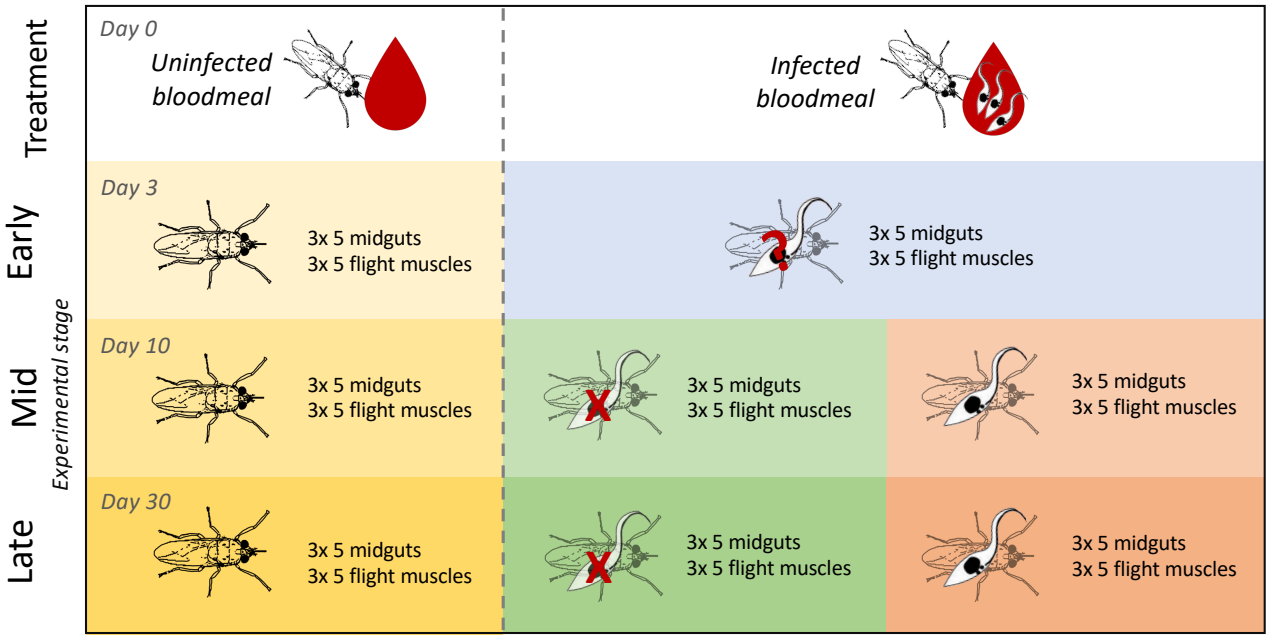


Figure 4-2 - Experimental design to collect tissues from uninfected (yellow), refractory (green) and infected (orange) tsetse flies at early, mid and late infection stages. Infection success cannot be determined in the early stage (blue). Three pools of each tissue were collected at each stage, with tissues from five individuals in each pool.

Table 4-1 – Nine experimental questions, and the relevant test group and control group comparisons needed to answer the questions for each species. C/I/R = Control/Infected/Refractory, trypanosome treatment group; 1/2/3 = early/mid/late experimental stage; MG/FM = midgut or flight muscle tissue; CULT = *in vitro Sodalis* culture

	Sodalis		***Serratia / Wolbachia***		***Glossina***		***Wigglesworthia***		***Trypanosoma***	
Question	**Test**	**Control**	**Test**	**Control**	**Test**	**Control**	**Test**	**Control**	**Test**	**Control**
1. Which genes are differentially expressed *in vitro* vs *in vivo*?	CULT vs C1MG	NA	NA	NA	NA	NA	NA	NA	NA	NA
2. What is the influence of early trypanosome infection on expression in the midgut?	C1MG vs I1MG	NA	C1MG vs I1MG	NA	C1MG vs I1MG	NA	C1MG vs I1MG	NA	NA	NA
3. What is the influence of early trypanosome infection on expression in the flight muscle?	C1FM vs I1FM	NA	C1FM vs I1FM	NA	C1FM vs I1FM	NA	NA	NA	NA	NA
4. Is there differential expression in the midgut of trypanosome-refractory flies compared to infected flies?	R2MG vs I2MG	C2MG vs R2MG	R2MG vs I2MG	C2MG vs R2MG	R2MG vs I2MG	C2MG vs R2MG	R2MG vs I2MG	C2MG vs R2MG	NA	NA
5. Is there differential expression in flight muscle of trypanosome-refractory flies compared to infected flies?	R2FM vs I2FM	C2FM vs R2FM	R2FM vs I2FM	C2FM vs R2FM	R2FM vs I2FM	C2FM vs R2FM	NA	NA	NA	NA
6. Do late-stage trypanosome infections alter expression in midguts?	I3MG vs C3MG	NA	I3MG vs C3MG	NA	I3MG vs I3FM	NA	I3MG vs C3MG	NA	NA	NA
7. Do late-stage trypanosome infections alter expression in flight muscles?	I3FM vs C3FM	NA	I3FM vs C3FM	NA	I3FM vs C3FM	NA	NA	NA	NA	NA
8. Are there changes in expression over the course of trypanosome infections in midguts?	I1MG vs I2MG vs I3MG	C1MG vs C2MG	I1MG vs I2MG vs I3MG	C1MG vs C2MG	I1MG vs I2MG vs I3MG	C1MG vs C2MG	I1MG vs I2MG vs I3MG	C1MG vs C2MG	I1MG vs I2MG vs I3MG	NA
9. Are there changes in expression over the course of trypanosome infection in flight muscle?	I1FM vs I2FM vs I3FM	C1FM vs C2FM	I1FM vs I2FM vs I3FM	C1FM vs C2FM	I1FM vs I2FM vs I3FM	C1FM vs C2FM	NA	NA	NA	NA

congolense and species of tsetse fly [260]. This allowed for examination of transcripts from each stage of tsetse infection, with an unchallenged control. The experimental flies were split into three groups at random, each to be dissected at a different time point. For each group and time point three pools of flight muscle and three of midguts from five individuals each were analysed. This experimental design is illustrated in Figure 4-2. Test comparisons and relevant controls are described in Table 4-1.

4.4.3 Infection with *T. congolense*

T. congolense Savannah strain 1/148 [519] rat blood stabilites were obtained from Dr Lee Haines (LH) at the Liverpool School of Tropical Medicine (LSTM), UK. Stabilites were assessed by microscopy for trypanosome number and activity, and 200 µl of stabilite was added to 5 L of defibrinated horse blood to create an infective blood meal. Teneral flies 48 hours post-eclosion were fed with either an infected or uninfected blood meal for 15-20 minutes. Based on an average blood meal size in this colony of 18-35 µl (LH, personal communication), the average intake per fly fed with an infective bloodmeal was 3-7 trypanosomes. A single trypanosome is considered sufficient to produce an infection [520].

4.4.4 Tsetse fly dissection and tissue collection

Tsetse flies for dissection were anaesthetised briefly by exposure to cold in a cold room and divided into Petri dishes with around 10 flies per dish. Dishes were sealed and labelled with treatment group (trypanosome unchallenged or challenged flies) before transportation to the dissection room. Flies in one Petri dish at a time were anaesthetised on ice for dissection; time between anaesthesia and dissection was

kept to less than 10 minutes, to minimise the impact of anaesthesia on transcription. Mouthparts, midgut and flight muscle tissues were dissected from each fly using fine forceps and a dissecting microscope and, where applicable, examined under a light microscope with phase contrast for trypanosome infection as described in section 4.4.4.1.

At the early infection stage, no samples were assessed for their infection status as trypanosomes from the initial blood meal would still be present regardless of whether the infection was successful or not. At the mid and late timepoints flies from the treatment group that received an infected bloodmeal were classified as ‘infected’ if trypanosomes were visible in mouthparts or midguts, and ‘refractory’ if not. Tissues were then placed in separate pre-labelled 1.5 ml Eppendorf tubes containing 300 µl of Zymo TRI Reagent (Cat # R2050-1-50), in pools of five per tube for each of the control, infected and refractory groups. Three pools of each tissue type were collected for each of the treatment groups in Figure 4-2..

4.4.4.1 Proboscis dissection for trypanosome infection identification

One fly at a time, using fine forceps cleaned with 70% ethanol, flies were moved to a sterile microscope slide on a dissecting microscope with three drops of sterile phosphate-buffer solution (PBS). The head was removed by grasping the thorax gently with one pair of forceps and pulling head off from the neck joint with the other. The head was squeezed gently to force the bulb of the proboscis to protrude, separating the mouthparts from the palps. The proboscis and bulb were then

removed from the head and moved to the first drop of PBS. The labellum and labrum were spread, and the hypopharynx was separated from the labellum by gently scraping the labellum with the tip of the forceps. The slide was moved to the light microscope with phase contrast: trypanosome infection was confirmed if single promastigotes were seen free swimming in the PBS (Figure 4-3), or if clumps of epimastigotes were seen attached to the inside wall of the hypopharynx or the labrum. One live trypanosome was enough to confirm infection.

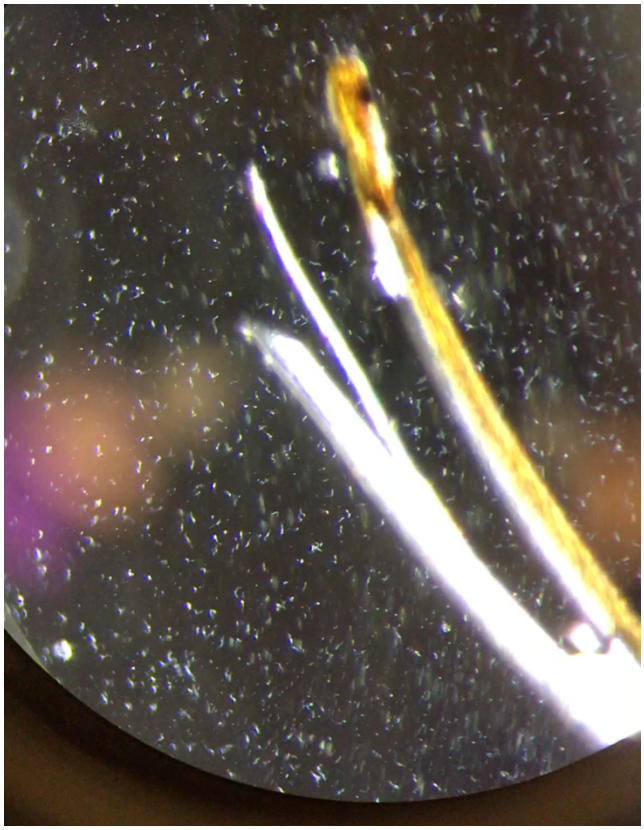


Figure 4-3 - Phase contrast microscopy of the tsetse mouthparts with a late-stage trypanosome infection; trypanosomes have spilled from the dissection into the surrounding liquid in large numbers.

4.4.4.2 Midgut dissection and trypanosome infection identification

With the tsetse fly on its back, the abdomen was grasped near the thoracic-abdominal joint with both forceps and a hole was torn in the abdomen with one swift movement to open the abdominal cavity and expose the organs. The thread of organs was pulled gently through the abdominal hole, allowing for the digestive tract to be removed while keeping the fat bodies and spermathecae inside the abdomen. The crop, if pulled out of the abdomen, was discarded.

If checking for trypanosome infection, the slide was moved to the light microscope with phase contrast: trypanosome infection was confirmed if promastigotes (see Figure 1-5) were seen free swimming in the PBS or within the gut lumen. One live trypanosome was enough to confirm infection. Midguts were always dissected after mouthparts to avoid the possibility of the messier gut dissection infecting the mouthpart dissection and giving a false positive late-stage infection. The entire gut was transferred to the appropriate pool in TRI reagent using the forceps, which were then wiped on a dry tissue before being resterilised with 70% ethanol.

4.4.4.3 Flight muscle dissection

Under the dissecting microscope the fly was moved adjacent to the final drop of PBS, on its back. The thorax was pierced between the segments with the tips of the forceps and opened completely with both sets of forceps to expose the flight muscle tissue, which was scraped off the inside of the thorax and any hard pieces of carapace were removed. The tissue was transferred to the appropriate pool in TRI reagent using the forceps, which were then wiped on a dry tissue before being resterilised with 70% ethanol. The fly and slide were then discarded.

4.4.4.4 Tissue pool storage

Pools of tissues were stored in TRI Reagent at -80°C until RNA extraction, which was no more than three weeks after dissection.

4.4.5 *Sodalis* isolation and culture from tsetse flies

This protocol for culturing *Sodalis* from live tsetse was developed by LH (unpublished). Under sterile conditions, *Sodalis* was isolated from tsetse tissue and cultured first in MMI before streaking onto Columbia agar plates supplemented with 10% defibrinated horse blood and incubated at 25°C under microaerophilic conditions for a week to isolate individual colonies. Single *Sodalis* colonies were streaked onto fresh plates, which were incubated in the same conditions for another seven days to produce pure, individual colonies. Several colonies from each plate were transferred into five sterile 15 ml conical Falcon tubes containing 5 ml of MMI, which were incubated at 25°C for two weeks. The five tubes from each plate were combined into 50 ml Falcon tubes and centrifuged at 4,000 x g for 10 minutes to form a bacterial pellet. Each pellet was dissolved in 300 µl of TRI Reagent and transferred to a 1.5 ml Eppendorf, before storing at -80°C until RNA extraction.

4.4.6 RNA extraction and purification

4.4.6.1 RNA extraction with a Zymo Direct-zol RNA Microprep kit

Before RNA extraction samples were thawed to room temperature (RT), and all work surfaces and tools were cleaned with RNase Zap (Thermo Scientific, Cat # AM9780). A glass pestle with a stainless-steel handle was cleaned with tissue dampened with RNase Zap and rinsed in MilliQ water.

Tsetse fly tissue samples were homogenised in their 1.5 ml Eppendorf tubes using the pestle for around 30 seconds. The pestle was then cleaned twice with clean blue roll dampened with RNase Zap and rinsed in MilliQ water, before use on the next sample. Pure *Sodalis* culture samples were simply thawed and vortexed for 30

seconds. All samples were then processed using a Zymo Direct-zol RNA Microprep kit (Cat # R2061) as described below.

First, 300 µl of absolute ethanol was added to each sample and mixed thoroughly by vortexing before centrifuging at 6,000 x g for one minute to remove debris. The supernatant was transferred into a Zymo-Spin IC Column in a collection tube and centrifuged at 15,000 x g for 30 seconds. Flow-through was discarded and 400 µl of RNA Wash Buffer was added to the column before centrifuging at 15,000 x g for 30 seconds. To remove DNA, 40 µl of DNA Digestion Mix (containing 5 µl DNAse 1 (6 U/ µl) and 35 µl DNA Digestion Buffer, freshly prepared) was added directly to the column matrix and incubated at RT for 15 minutes. Next, 400 µl of Pre-Wash was added and samples were centrifuged at 15,000 x g for 30 seconds; this step was repeated twice, with the flow-through discarded each time. To clean the purified RNA, 700 µl of Wash Buffer was added to the column and centrifuged at 15,000 x g for two minutes, before transferring the column to an RNase-free 1.5 ml Eppendorf tube. To elute RNA into the Eppendorf tube, 15 µl of MilliQ water was added directly to the column matrix and centrifuged at 15,000 x g for 30 seconds.

4.4.6.2 RNA Purification

To remove small RNAs and purify the RNA, extracted samples were cleaned using Agencourt RNAclean XP beads (Beckman Coulter Life Sciences, Cat # A63987). The XP beads were brought to RT before use, and RNA samples were vortexed briefly to collect any condensation. Ethanol was made to 80% fresh from absolute ethanol stocks, with MilliQ water.

Thoroughly vortexed beads were added to samples at a 1.8:1 ratio, meaning 27 µl of beads were added to each sample. Samples were mixed by pipetting up and down 10 times and incubated at RT for five minutes without shaking. Samples were placed on a magnetic stand for two minutes, or until the supernatant had cleared and a pellet of beads and attached RNA had formed against the wall of the tube. The supernatant was removed and discarded, and pellets were washed by adding and removing 80% ethanol from each tube twice. A fine pipette tip was used to remove any residual ethanol, and the pellet was allowed to dry slightly without cracking. Tubes were removed from the magnetic stand and beads were resuspended in 15.5 µl MilliQ water by pipetting up and down 10 times, eluting the RNA from the beads. Samples were incubated at RT for two minutes before being placed back on the magnetic stand for two minutes to pellet the beads out of solution, leaving RNA in solution in the water. With the sample still on the magnetic stand, the supernatant containing purified RNA was removed and transferred to a new 1.5 ml Eppendorf tube, without disturbing the bead pellet.

4.4.6.3 RNA quality control

RNA was analysed for quantity and quality using Qubit and NanoDrop assays. To assess purity and estimate quantity, 1 µl of each sample was analysed using a NanoDrop microvolume spectrophotometer. Absorption ratios for wavelengths 260/280 and 260/230 were recorded; samples with 260/280 ratios of ~2 and 260/230 ratios of 1.8-2-2 were deemed sufficiently pure.

To accurately quantify the amount of RNA in each sample a Qubit RNA High Sensitivity Assay kit (Invitrogen, Cat # Q32852) and a Qubit fluorometer were used. All reagents and samples were brought to RT before use. A master mix of 1 µl Qubit RNA HS Reagent and 199 µl of RNA HS Buffer per sample was made; 199 µl of master mix was transferred to one optically clear 0.5 ml tube per sample. 190 µl of master mix was transferred to two optically clear 0.5 ml tubes, which were labelled Standard 1 and Standard 2. 10 µl of each of Qubit RNA HS Standards #1 and #2 were pipetted into their respective tubes and vortexed thoroughly. 1 µl of each sample was pipetted into their respective tubes, labelled and vortexed thoroughly. Mixtures were incubated at RT for two minutes before calibration was performed with the standards, and measurements of the samples were taken in ng/µl. Purified and analysed samples were stored at -80°C before sequencing.

4.4.7 RNA sequencing

RNA samples were handed to the Centre for Genomic Research at the University of Liverpool for all subsequent lab work, performed by Margaret Hughes and Charlotte Nelson. At least 750 ng of total RNA was depleted of ribosomal RNA using an Epidemiology RiboZero Magnetic kit (Illumina); ribosomal RNA depletion was visualised using a Bioanalyser RNA pico chip (Agilent). RNA-Seq libraries were prepared from the enriched RNA using the NEBNext Ultra 11 Directional RNA Library Prep kit for Illumina (New England Biolabs), involving nine cycles of amplification and subsequent purification with AMPure XP beads (Beckman Coulter). Each library was quantified using the Qubit dsDNA High Sensitivity kit, and the size distribution was assessed using the Agilent Bioanalyser, before samples were pooled in equimolar amounts. The pools were assessed using the Agilent

Bioanalyser as well as by qPCR using the Illumina Quantification Kit (Kapa) as follows. Briefly, a 20 µl PCR reaction was performed in triplicate for each pool with 12 µl SYBR Green I Master Mix and 4 µl of diluted pooled cDNA. Thermal cycling was performed on a Roche Light Cycler LC480II, with conditions consisting of an initial denaturation at 95°C for 5 minutes, 35 cycles of 95°C for 30 seconds (denaturation) and 60°C for 45 seconds (annealing and extension), melt curve analysis to 95°C (continuous) and cooling at 37°C.

After each pool molarity was confirmed with qPCR data, template cDNA was diluted to 300 pM and denatured for eight minutes at room temperature with 0.2 N sodium hydroxide, terminated by the addition of 400 mM TrisCl. PhiX DNA was spiked in at 1% to improve sequencing quality control, and the libraries were sequenced on an Illumina NovaSeq 6000 on one lane of an S4 flow cell, generating 2 x 150 bp paired-end reads.

4.4.8 Initial read processing and quality control

Read processing and quality control was performed by the CGR using an in-house pipeline developed by Dr Richard Gregory, performed by Xuan Liu. Briefly, basecalling and de-multiplexing of indexed reads was performed using CASAVA version 1.8.2 (Illumina) to produce 39 samples in FASTQ format. Primer and adapter sequences were trimmed using Cutadapt version 1.2.1 [521] and reads were trimmed at bases with a minimum window quality score of 20 using Sickle version 1.200 [522]. At this point, any reads shorter than 20 bp were removed. Samples 10, 21, 22, 23 and 29 were all excluded from any differential expression analysis due to having low read counts.

4.4.9 Differential expression analysis

Bioinformatics analysis was performed by myself and Ian Goodhead. Reads were aligned to reference genomes for *Gmm*, *Wigglesworthia, Sodalis, Serratia, Wolbachia* and *T. congolense* from the sources listed in Table S4.1. Alignment was performed using bowtie2 version 2.3.4.1 [523], and differential expression analysis was performed using HTSeq [524].

4.4.10 Data handling and visualisation

All subsequent work was performed by myself. Differential expression (DE) tables were loaded into Degust [525] using the 'edgeR quasi-likelihood' method to create graphs of differentially expressed genes in parallel coordinates plots and to visualise the similarity of samples to each other using multi-dimensional scaling (MDS) plots. For all DE genes a false discovery rate (FDR) cut-off of <0.05 was applied, and tables were produced of counts per million (CPM) for each gene to standardise for library size. To determine possible protein functions, amino acid sequences of DE genes were searched in the STRING database [495] to identify COGs and the InterPro database to identify protein families and gene ontology terms [496]. The variance of mean reads mapped to *Sodalis* for the control, refractory and infected mid-stage midgut samples was examined using a Tukey ANOVA in R [435] and visualised using ggplot2 [526].

4.5 Results

4.5.1 Sequencing output

Unfortunately, for the late infection stage only tsetse with successful late-stage infections were able to be dissected. Figure 4-4 shows the raw read counts per sample, divided by infection stage and tissue type; five samples produced low read numbers and were removed from all subsequent analyses.

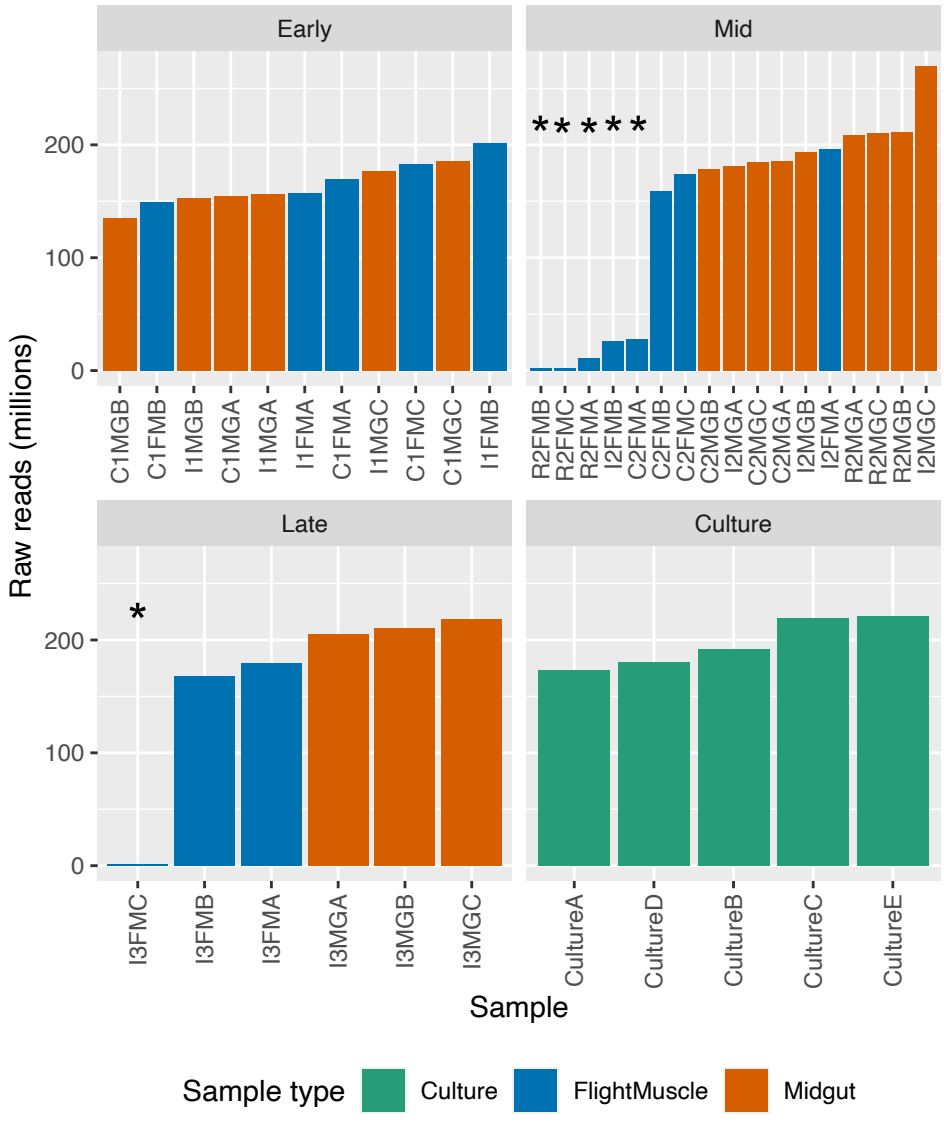


Figure 4-4 - Raw read counts per sample, divided by experimental stage and sample type. Sample names are formatted: Treatment (control, infected or refractory) + Stage (1: early, 2: mid, 3: late) + tissue type (midgut or flight muscle) + replicate (A, B or C). Samples labelled with an asterisk (*) contained too few reads and so were removed from the dataset.

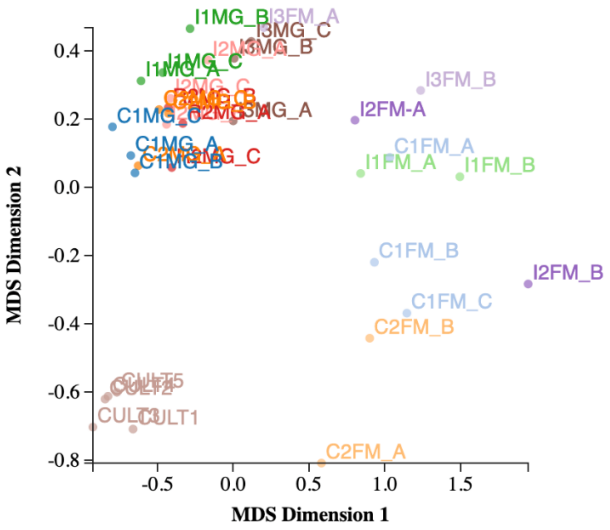
4.5.2 Differential expression of trypanosome and tsetse flies

Due to the complex nature of this dataset, the scope of the study was narrowed to look only at the bacterial expression in detail. Briefly, *T. congolense* transporter and membrane proteins were progressively upregulated from early to late-stage infections. There was no impact of early or mid-infection on tsetse midgut expression; however, immune genes including trypsins were upregulated in infected flight muscle tissue, and multiple immune response genes were upregulated in late infection stages of both tissue types, suggesting a systemic stress response to trypanosome infection [1, 527-531].

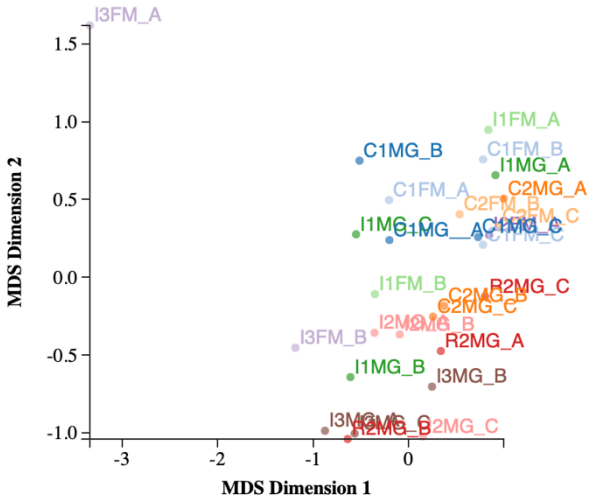
4.5.3 Differential expression between tissue types, treatments, and experimental times – MDS plots

In multidimensional scaling plots samples cluster based on the similarity of their expression. *Sodalis* expression was most similar within tissue types, with *in vitro* cultures clustering apart from *in vivo* (Figure 4-5a). Midgut samples clustered closely together, with flight muscles clustering more loosely. No effect of treatment on gene expression was immediately evident samples.

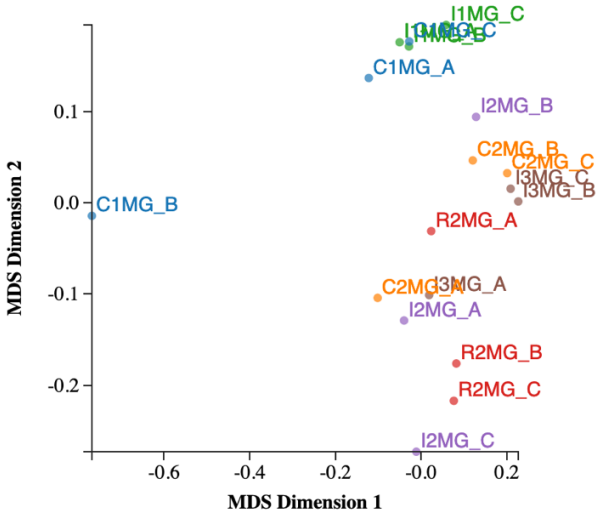
Wigglesworthia, Serratia and *Wolbachia* expression appeared largely similar regardless of treatment or time (Figure 4-5b, c & d).



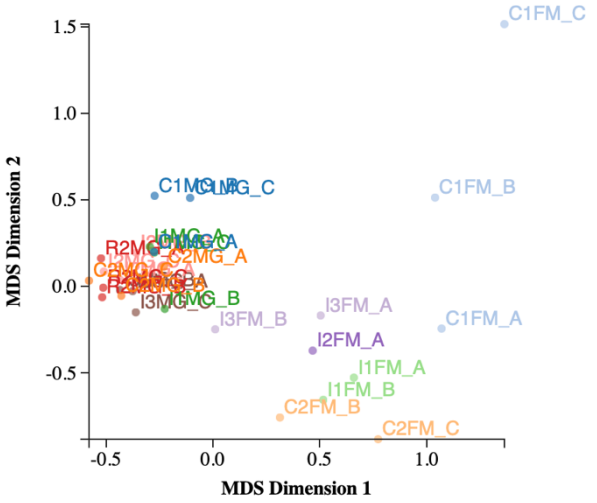
a) ***Sodalis***



b) ***Wolbachia***



c) ***Wigglesworthia***



d) ***Serratia***

Figure 4-5 - MDS plots of gene expression from pooled samples for each treatment, divided into expression from four bacterial species.
Sample key:
C/I/R = Control/Infected/Refractory
1/2/3 = Early/mid/late-stage experiment
MG/FM = Midgut/Flight muscle
A/B/C = replicate
CULT = *Sodalis in vitro* cultures, replicates 1-5

4.5.4 Q1: Differential expression of *Sodalis in vitro* compared to *in vivo*

A heatmap showing the major DE genes between *in vitro Sodalis* and early midgut *Sodalis* of control group tsetse can be seen in Table 4-2. In total 524 genes were DE between *Sodalis* from liquid *in vitro* culture and *Sodalis* from early-experiment tsetse fly midguts (Table 4-3). To simplify analysis DE genes were filtered to an absolute log fold change of >4 (a more than x16 difference), creating a subset of 52 DE genes. Most notably, 13 phage-associated genes were consistently upregulated *in vitro* compared to *in vivo*. Genes for antigen synthesis, ion transport and the flagellum were largely upregulated in the tsetse environment compared to the *in vitro* environment, with inconsistencies between replicates for some genes. Conversely, genes for protein synthesis and binding, stress response, structural genes and genes involved in transcription were downregulated in the tsetse compared to the *in vitro* cultures.

Table 4-2 – DE genes in *Sodalis* between early control midguts and *in vitro pure* cultures, with an FDR of <0.05 and an absolute log fold change (abs logFC) of >4. Numbers are counts per million (CPM), normalised for library size. Gene descriptions are summaries from the COG, GO terms and protein families for each gene. C1MG = early control midgut, replicates A-C; CULT = *in vitro* culture, replicates 1-5.

Gene name	C1MG_A	C1MG_B	C1MG_C	CULT1	CULT2	CULT3	CULT4	CULT5	Gene description
wecF	97.23	0.00	0.00	0.00	0.02	0.02	0.00	0.02	Antigen synthesis
wecB	24.31	29.74	0.00	0.00	0.00	0.02	0.00	0.00	
glpK	24.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Carbohydrate process
fliI_1	0.00	0.00	25.06	0.00	0.00	0.00	0.00	0.00	Flagellum
fliM_1	0.00	29.74	0.00	0.00	0.00	0.02	0.00	0.00	
fliG	0.00	29.74	0.00	0.00	0.00	0.02	0.00	0.00	
fliK_1	0.00	0.00	25.06	0.00	0.00	0.02	0.00	0.00	
flg_1	0.00	0.00	25.06	0.02	0.00	0.00	0.00	0.00	
fliH	24.31	0.00	0.00	0.00	0.02	0.02	0.00	0.00	
fliO_1	0.00	29.74	0.00	0.02	0.00	0.02	0.00	0.00	
flgC_1	0.00	29.74	25.06	0.12	0.08	0.14	0.018	0.05	
flgD_1	97.23	0.00	75.19	0.16	0.87	2.26	1.20	0.72	
fliJ	1652.85	1130.11	2506.39	9.15	33.17	71.93	55.14	25.79	
fliN_1	2746.65	862.45	1228.13	16.06	42.40	50.28	44.46	56.54	
alsT	2430.67	713.76	1328.39	15.19	39.47	45.15	43.38	51.87	Ion transport
SGP1_RS10710	121.53	89.22	225.58	1.01	4.83	5.60	5.63	3.01	Unknown function
SGP1_RS26425	680.59	594.80	1228.13	5.13	21.12	47.55	28.40	18.50	
SGP1_RS19990	170.15	29.74	350.90	1.20	4.38	10.55	9.25	4.19	
SGP1_RS31010	3913.37	2200.74	3634.27	39.18	92.62	106.08	109.20	148.72	
SG0763	24.31	29.74	75.19	0.35	0.95	2.75	1.97	0.88	
SG2226	364.60	684.02	1178.00	3.70	23.14	45.54	32.95	23.91	
SGP1_RS22280	121.53	208.18	150.38	2.86	5.88	12.62	5.72	5.83	
SG0319	170.15	237.92	75.19	1931.37	1785.48	1312.76	1543.82	1765.11	
SGP1_RS30580	97.23	29.74	75.19	1046.77	461.21	705.43	846.62	544.96	
SG0318	121.53	237.92	125.32	2671.69	557.61	1228.74	1564.28	2195.88	

Gene name	C1MG_A	C1MG_B	C1MG_C	CULT1	CULT2	CULT3	CULT4	CULT5	Gene description
SG0320	0.00	0.00	50.13	21.13	503.61	282.05	266.89	58.98	
SGP1_RS31015	97.23	29.74	25.06	919.50	397.72	602.02	717.93	421.46	Outer membrane
hmuU_2	170.15	0.00	25.06	1339.03	412.31	831.21	1095.65	522.07	Outer membrane (siderophore)
SGP1_RS07415	24.31	0.00	25.06	188.18	267.95	260.90	317.55	168.02	
SG0839	48.61	0.00	125.32	1280.35	439.21	702.23	1042.14	643.04	
SG0850	72.92	29.74	50.13	1168.96	532.34	1017.04	1031.30	335.20	
SG0836	243.07	0.00	25.06	2149.68	981.33	1646.28	2614.93	817.08	
SG0849	24.31	59.48	25.06	981.76	312.65	516.82	751.02	472.51	
SGP1_RS07255	48.61	59.48	0.00	1023.17	456.52	989.02	528.09	377.44	
SG0855	24.31	0.00	0.00	172.32	183.73	250.14	131.26	159.32	Phage
SG0820	24.31	0.00	0.00	310.58	132.60	150.13	286.35	180.15	
SG0835	48.61	0.00	0.00	529.58	338.33	453.15	731.69	399.53	
SG2348	24.31	148.70	50.13	19610.60	2116.75	4389.25	9178.69	11729.20	
SG2348	0.00	0.00	0.00	19.95	33.81	55.08	24.89	19.84	
SG2348	0.00	0.00	0.00	27.40	29.64	39.32	28.44	32.45	
SGP1_RS08705	0.00	0.00	0.00	15.38	25.39	52.35	32.88	40.05	
SGP1_RS30115	0.00	0.00	0.00	26.74	35.81	52.38	39.39	34.74	Protein binding
fucU	0.00	0.00	0.00	35.58	56.72	53.22	24.55	28.83	
esaI	0.00	0.00	0.00	37.67	46.84	53.22	34.78	27.24	Protein synthesis
aroF	0.00	0.00	0.00	57.12	33.60	36.71	49.67	28.44	
rsmG	0.00	0.00	0.00	41.75	53.40	64.50	41.54	44.63	rRNA processing
SG1493	0.00	0.00	0.00	63.91	42.27	49.75	56.39	34.24	Secreted protein
cat_1	0.00	0.00	0.00	77.04	53.05	67.08	81.27	50.19	Stress response
SG1847	0.00	0.00	0.00	62.36	85.27	82.40	65.19	70.40	Structural
SGP1_RS22145	0.00	0.00	0.00	98.04	61.545	75.79	82.20	49.88	
SGP1_RS18705	0.00	0.00	0.00	294.65	118.58	232.11	301.10	99.56	Transcription
SG2045	0.00	0.00	0.00	502.28	139.82	276.59	386.58	186.18	

4.5.5 Q2: The influence of early trypanosome infection on expression in the midgut

A table summarising the results from all comparison groups outlined in Table 4-1 can be found in Table 4-3. Two *Sodalis* genes were downregulated between early control and early trypanosome-infected tsetse fly midguts: the chitinase gene *chiA*; and *deoB* which is involved in metabolic intermediate biosynthesis. None of the four remaining bacteria showed any differential expression between control and infected groups in the early-stage midgut.

4.5.6 Q3: The influence of early trypanosome infection on expression in flight muscle

There were no DE genes between early control and trypanosome-infected tsetse fly flight muscle tissue in any of the bacteria studied.

4.5.7 Q4: Differential expression in the midgut of trypanosome-refractory tsetse compared to trypanosome-infected tsetse

Sodalis was the only bacterium to show differential expression between the control, refractory, and infected tsetse midgut groups. In the control group, which was a comparison of mid-stage control midgut tissue and mid-stage refractory midgut tissue, two *Sodalis* genes were DE. The flagellin gene *fliC* was downregulated in refractory tsetse compared to control (-3.06 log fold change (logFC)); and *sipB*, a TTSS needle formation protein, was upregulated in refractory tsetse flies (2.64 logFC).

Table 4-3 - Differential expression (DE) of genes from each of the four examined bacteria in comparison groups used to answer questions 2-4 and 8-9 in Table 4-1.
Key: dark green = >100 DE genes; medium green = 10-99 DE genes; light green = 1-9 DE genes; orange = no DE genes
*comparison groups only contain early and mid-stage samples, due to missing samples

	Sodalis		***Wigglesworthia***		***Serratia***		***Wolbachia***	
Question	**Test**	**Con**	**Test**	**Con**	**Test**	**Con**	**Test**	**Con**
1. Which genes are differentially expressed *in vitro* vs *in vivo*?	524	NA	NA	NA	NA	NA	NA	NA
2. What is the influence of early trypanosome infection on expression in the midgut?	2	NA	0	NA	0	NA	0	NA
3. What is the influence of early trypanosome infection on expression in the flight muscle?	0	NA	NA	NA	0	NA	0	NA
4. Is there differential expression in the midgut of trypanosome-refractory flies compared to infected flies?	7	2	0	0	0	0	0	0
8. Are there changes in expression over the course of trypanosome infections in midguts?	43	17*	165	74*	0	0*	5	-*
9. Are there changes in expression over the course of trypanosome infection in flight muscle?	0	0*	NA	NA	1	0*	0	0*

Table 4-4 - DE genes in *Sodalis* between mid-stage refractory and infected tsetse flies. with an FDR of <0.05. Numbers are CPM. Gene descriptions are summaries from the COG, GO terms and protein families for each gene. R2MG = mid-stage refractory midguts, replicates A-C; I2MG = mid-stage infected midguts, replicates A-C.

Gene name	R2MG_A	R2MG_B	R2MG_C	I2MG_A	I2MG_B	I2MG_C	Gene description
SG0618	0.00	43.68	19.44	259.62	484.99	177.13	Actin binding
SGP1_RS05470	11.53	14.56	29.16	302.89	250.86	200.74	Chitodextrinase
fliC	369.06	1776.28	281.86	4355.85	5953.68	6600.86	Flagellum
ompF	92.26	160.16	136.07	562.51	685.68	649.46	Outer membrane (porin)
SGP1_RS23530	23.07	29.12	77.75	331.74	250.86	366.06	Transcription
mhqR	196.06	247.51	87.47	807.71	668.95	862.01	
sicA	0.00	160.16	0.00	750.01	1153.94	614.03	TTSS

Table 4-5 - DE genes in *Sodalis* between early and mid-stage control midguts with an FDR of <0.05. Numbers are CPM. Gene descriptions are summaries from the COG, GO terms and protein families for each gene. C1MG = early-stage control midguts, replicates A-C; C2MG = mid-stage control midguts, replicates A-C.

Gene name	C1MG_A	C1MG_B	C1MG_C	C2MG_A	C2MG_B	C2MG_C	Gene description
purK	170.15	386.62	125.32	0.00	18.05	17.00	Biosynthetic processes
pflB	631.97	4193.31	576.47	238.11	397.09	271.95	Carbohydrate metabolism
chiA	704.89	3657.99	601.53	126.99	108.30	101.98	Chitinase
dppA_1	826.43	475.84	651.66	158.74	36.10	101.98	Membrane transport
SG1285	850.73	951.67	1754.47	222.24	198.55	169.97	Unknown function
yceM_1	388.91	178.44	225.58	47.62	0.00	17.00	Nucleotide binding
aceE	170.15	1724.91	150.38	79.37	36.10	118.98	Oxidoreductase
SGP2_0012	631.97	327.14	1052.68	2301.77	2057.65	1291.77	Plasmid
SG23SrRNA04	437.52	1070.63	476.21	206.37	18.05	17.00	
SG23SrRNA06	583.36	951.67	726.85	126.99	108.30	50.99	rRNA
SG23SrRNA07	656.28	1219.33	601.53	317.49	54.15	17.00	
sseB	2333.44	1397.77	2481.33	412.73	559.54	271.95	Secreted protein
sipB	1701.47	535.32	1453.71	190.49	198.55	101.98	
sipD	510.44	862.45	175.45	31.75	72.20	33.99	
SG1296	437.52	0.00	175.45	15.87	0.00	0.00	TTSS
SG1291	364.60	89.22	726.85	47.62	36.10	50.99	
ssaV	704.89	148.70	501.28	31.75	126.35	17.00	

In the test group (refractory vs infected tsetse fly midguts), seven genes were upregulated in infected flies compared to refractory flies (Table 4-4). This included two genes involved in transcription, an outer membrane porin gene *ompF*, the TTSS gene *sicA*, a chitodextrinase gene, a flagellum gene, and an actin binding gene. There was no statistically significant variation between the mean percentage of reads mapped to *Sodalis* in the control, refractory or infected mid-stage midguts (Table 4-6 and Figure 4-6)

Table 4-6 – Tukey's ANOVA of the percentage of reads in each sample that mapped to *Sodalis* from the mid-stage midguts, showing no significant differences between any of the comparisons.

Comparison	**P value**
Infected – Control	0.486
Refractory – Control	0.566
Refractory – Infected	0.134

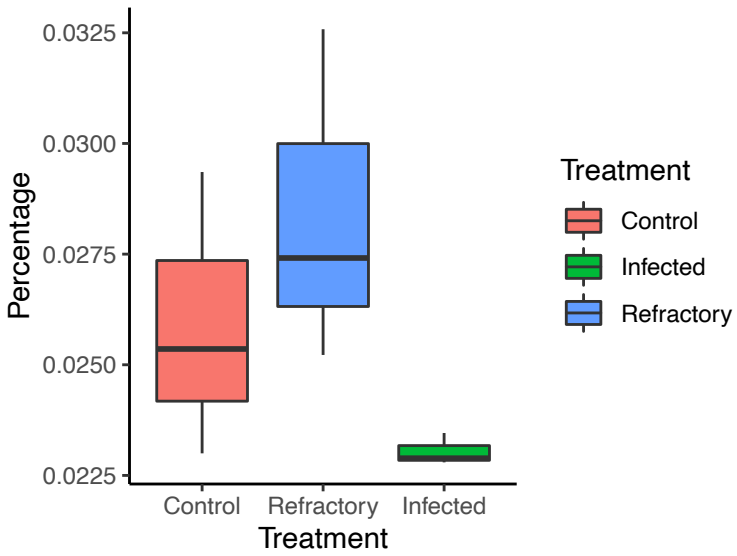


Figure 4-6 - A boxplot of the percentage of reads mapped to *Sodalis* in each sample from the mid-stage midguts, showing a non-significant difference between the three treatment groups.

4.5.8 Q5, Q6 and Q7 involving differential expression in flight muscle tissues at the mid- and late-experiment level, and midgut expression at the late-experimental level

As there were issues with the late-stage experimental tsetse flies, and the RNA extractions for several flight muscle samples was inadequate for sequencing (Figure 4-4), these three questions could not be addressed.

4.5.9 Q8: Expression changes in midgut tissue over the course of a trypanosome infection

Control comparisons for this question did not examine control flies from the late-experimental stage, as these samples were not successfully collected.

4.5.9.1 *Sodalis*

In the control comparison between early and mid-stage groups, 17 genes were differentially expressed (Table 4-5). One plasmid related gene was upregulated in the mid-stage midgut; the remaining 16 genes were downregulated at the mid-stage, including five TTSS-related genes and genes involved in carbohydrate metabolism, membrane transport, protein secretion and a chitinase.

Table 4-7 - DE genes in *Sodalis* from infected tsetse midguts across the three time points, with an FDR of <0.05. Numbers are CPM. Gene descriptions are summaries from the COG, GO terms and protein families for each gene. I1MG = early-stage infected midguts, replicates A-C; I2MG = mid-stage infected midguts, replicates A-C; I3MG = late-stage infected midguts, replicates A-C.

Gene name	I1MG_A	I1MG_B	I1MG_C	I2MG_A	I2MG_C	I2MG_B	I3MG_A	I3MG_B	I3MG_C	Gene description
grcA	7886.86	9933.94	10656.11	2293.31	3884.94	5418.51	1020.38	154.22	952.12	Catalytic reactions
SGP1_RS22485	104.60	289.34	21.75	129.81	247.98	250.86	671.30	154.22	1020.13	
ilvG	62.76	96.45	130.48	86.54	165.32	33.45	402.78	77.11	850.11	
yfbT_2	3786.53	3906.06	2740.14	360.58	366.06	568.61	295.37	411.26	544.07	
sseD	794.96	1928.92	652.42	100.96	200.74	100.34	0.00	0.00	102.01	Cell adhesion
SG0027	1004.16	1253.80	347.96	86.54	118.08	250.86	0.00	51.41	0.00	Flagellum
SG1297	376.56	385.78	195.73	57.69	23.62	16.72	0.00	0.00	0.00	
SG0028	1297.04	675.12	717.66	129.81	188.93	66.90	80.56	51.41	68.01	
SG2404	41.84	0.00	0.00	115.39	94.47	83.62	134.26	179.93	408.05	Flavodoxin
ftnA	2866.05	2121.81	2522.67	1197.14	1346.15	1505.14	349.08	591.18	374.05	Iron binding
SG1856	606.68	1109.13	369.70	57.69	129.89	16.72	161.11	282.74	34.00	Membrane protein
SGP1_RS22340	439.32	916.24	1022.12	389.43	673.08	484.99	1074.09	1876.37	1836.24	
yceM_1	209.20	192.89	152.23	0.00	11.81	0.00	0.00	0.00	0.00	Nucleotide binding
glpD	0.00	48.22	43.49	72.12	47.23	83.62	214.82	51.41	442.06	Oxidoreductase
amiD_4	899.56	1205.58	521.93	173.08	70.85	133.79	26.85	128.52	408.05	Peptidoglycan catabolic process
SGP1_RS30115	983.24	1060.91	543.68	187.50	366.06	518.44	617.60	1439.40	2176.28	Protein binding
nonCDS	2112.93	2266.48	2805.39	2134.66	3448.03	3261.14	3249.11	12029.30	14655.88	Non-CDS
nonCDS	2615.01	2989.83	2740.14	2552.93	3554.31	2792.88	3114.85	12209.23	13431.72	
nonCDS	2594.09	3327.39	2718.40	2596.20	3341.76	3194.25	3275.96	11849.38	13227.69	
sseB	2782.37	4581.18	1544.05	562.51	484.14	468.27	161.11	128.52	374.05	Secreted protein

Gene name	I1MG_A	I1MG_B	I1MG_C	I2MG_A	I2MG_C	I2MG_B	I3MG_A	I3MG_B	I3MG_C	Gene description
SGP1_RS11460	313.80	337.56	434.94	100.96	59.04	16.72	0.00	0.00	68.01	
dps	1485.33	964.46	1261.34	288.47	696.69	869.64	134.26	77.11	272.04	Stress response
SGP1_RS22345	22844.71	27342.43	26292.33	13774.30	23959.10	17108.45	41781.91	56393.78	96912.41	Transcription
fucP_1	1234.28	1109.13	739.40	115.39	224.36	284.31	214.82	257.04	204.03	Transmembrane transport
hmuV	0.00	0.00	0.00	0.00	0.00	0.00	80.56	77.11	68.01	Transport
SG1287	1422.56	2989.83	1130.85	360.58	106.28	234.13	0.00	77.11	204.03	TTSS
sipB	1380.72	3086.27	1152.60	288.47	307.02	200.69	26.85	128.52	136.02	
yscN	794.96	675.12	543.68	72.12	94.47	33.45	0.00	51.41	0.00	
LcrR	418.40	1350.24	543.68	144.23	70.85	66.90	80.56	25.70	34.00	
SsaM	606.68	771.57	282.71	14.42	106.28	50.17	53.70	0.00	34.00	
mxiD_2	690.36	1060.91	848.14	187.50	153.51	200.69	107.41	51.41	34.00	
yscJ	523.00	434.01	347.96	86.54	94.47	117.07	53.70	25.70	0.00	
prgK_2	167.36	337.56	152.23	14.42	59.04	66.90	0.00	0.00	0.00	
invA_2	439.32	482.23	239.22	43.27	94.47	50.17	26.85	51.41	68.01	
SG1291	376.56	530.45	369.70	201.93	118.08	66.90	0.00	51.41	34.00	
SG1285	1840.97	2121.81	1109.11	360.58	165.32	66.90	80.56	77.11	68.01	Unknown function
SG1155	2635.93	2459.37	1805.02	490.39	413.29	284.31	510.19	411.26	714.09	
SGP1_RS29400	648.52	1157.35	478.44	274.04	118.08	117.07	26.85	77.11	170.02	
slyX	523.00	626.90	282.71	72.12	153.51	150.51	80.56	0.00	34.00	
SGP1_RS10325	125.52	144.67	152.23	0.00	0.00	0.00	0.00	25.70	0.00	
SGP1_RS22515	16422.25	12248.64	10308.16	3014.48	3247.29	4766.29	4430.60	1259.48	3162.41	
SGP1_RS01805	41.84	48.22	86.99	173.08	177.13	301.03	375.93	359.85	272.04	
SG1631	251.04	289.34	130.48	375.01	141.70	301.03	0.00	25.70	68.01	

2808

In the treatment group, 43 genes were DE across the three time points (Table 4-7). Ten TTSS genes, three flagellar genes, genes for secreted proteins, stress responses and cell adhesion were progressively downregulated from the early to late infection stage. Genes involved in transport, transcription and protein binding were progressively upregulated from early to late infection stages.

4.5.9.2 *Wigglesworthia*

In the control comparison group 74 *Wigglesworthia* genes were DE between early and mid-stage midgut samples. These were filtered to 0.585 absolute logFC (>1.5x) for ease of analysis, leaving 33 DE genes (Table 4-8). Of these 33 genes 19 were identified as tRNA, all of which were downregulated in the mid-stage midguts. Also downregulated in the mid-stage were three transcription genes and genes for the flagellum, stress response, nucleotide synthesis and a pseudogene ancestrally involved in vitamin B1 production. Genes upregulated in the mid-stage midguts included three flagellum genes and a detoxification gene.

In the test comparison group of infected tsetse midguts 165 *Wigglesworthia* genes were DE; these were filtered to an absolute logFC of 1 (>2x), producing a subset of 20 DE genes (Table 4-9). Of these 20 genes, 12 were identified as tRNA, all of which were progressively downregulated from the early to late timepoints except for in I2MG_B, which showed similar expression to the early group. I2MG_B also showed similar expression to the early group samples in expression of biosynthesis mechanism genes and biotin synthesis genes, which were otherwise progressively downregulated from the early to late stages. Expression of other genes involved in the flagellum and heat shock response was less clearly defined by timepoint.

Table 4-8 – DE genes in *Wigglesworthia* from control tsetse midguts from the early and mid-stages, with an FDR of <0.05. Numbers are CPM. Gene descriptions are summaries from the COG, GO terms and protein families for each gene. C1MG = early-stage control midguts, replicates A-C; C2MG = mid-stage control midguts, replicates A-C. *pseudogene

Gene name	C1MG_A	C1MG_B	C1MG_C	C2MG_A	C2MG_B	C2MG_C	Gene description
bioC	3207.48	2673.28	3200.15	1852.98	2484.32	2601.12	Biotin biosynthesis
gloB	289.11	185.70	297.05	458.72	488.02	507.09	Detoxification
fliI	681.91	690.99	798.17	406.54	483.02	408.86	Flagellum
flgG	325.44	194.34	279.42	493.50	516.39	605.95	
flgF	249.76	183.55	252.10	397.12	410.44	412.05	
AFA40910.1	53.74	23.75	66.99	84.79	85.93	135.86	
AFA41244.1	301.98	215.94	354.35	169.57	200.21	199.65	Membrane protein
aspC	1654.45	1228.67	1763.81	823.95	1275.53	1466.40	Metabolism
pyrE	252.78	164.11	299.26	184.79	205.22	172.22	Nucleotide synthesis
bolA	174.07	187.86	237.56	105.80	151.83	166.48	Stress response
slyA	633.47	729.86	755.42	413.79	550.59	633.38	Transcription
rpoD	3823.55	3513.27	3773.55	2549.38	3185.07	3009.98	
nusG	3407.29	2690.55	3082.48	2136.32	2613.63	2639.39	
tRNA Glu	94.61	239.69	97.40	59.42	28.36	16.58	tRNA
tRNA Tyr	2770.79	2843.87	2387.00	1431.22	2043.02	1659.03	
tRNA Arg	1511.41	1969.33	1260.49	718.15	1095.34	889.15	
tRNA His	1597.67	2083.77	1338.50	771.05	1191.27	957.40	
tRNA Trp	93.09	190.02	138.83	37.68	49.22	54.22	
tRNA Asp	84.77	181.39	125.17	35.51	40.88	46.56	
tRNA Glu	97.63	235.37	104.89	74.64	26.70	14.033	
tRNA Met	367.07	561.43	363.60	235.52	262.78	282.56	
tRNA Thr	1942.80	1954.21	1695.06	1016.71	1492.43	1165.34	

Gene name	C1MG_A	C1MG_B	C1MG_C	C2MG_A	C2MG_B	C2MG_C	Gene description
tRNA Leu	365.55	574.39	363.16	236.97	273.63	278.74	
tRNA Met	217.21	202.98	204.94	132.61	153.50	137.14	
tRNA Gln	258.84	358.45	255.18	167.40	196.88	198.37	
tRNA Gly	2462.75	2442.23	2247.29	1607.32	1999.64	1816.58	
tRNA Gly	3926.48	4411.56	3932.21	2465.32	3497.07	2845.42	
tRNA Ala	430.64	602.46	449.11	263.78	416.28	386.53	
tRNA Leu	3933.29	4286.31	3933.09	2579.82	3586.33	2992.76	
tRNA Cys	4559.95	5206.20	4547.03	2885.63	4062.67	3387.58	
tRNA Ile	413.23	606.78	418.70	251.46	400.43	369.95	
tRNA Pro	387.50	395.16	320.85	215.23	330.35	300.42	
thiI*	84.77	17.28	96.08	42.76	49.22	44.65	Vitamin B1 production*

2839

Table 4-9 - DE genes in *Wigglesworthia* from infected tsetse midguts from all three timepoints, with an FDR of <0.05 and an abs logFC of 1. Numbers are CPM. Gene descriptions are summaries from the COG, GO terms and protein families for each gene. I1MG = early-stage infected midguts, replicates A-C; I2MG = mid-stage infected midguts, replicates A-C; I3MG = late-stage infected midguts, replicates A-C. *pseudogene

Gene name	I1MGA	I1MGB	I1MGC	I2MGA	I2MGB	I2MGC	I3MGA	I3MGB	I3MGC	Gene description
aspC	2873.00	2706.00	3892.00	1072.00	2648.00	892.00	821.00	831.00	986.00	Biosynthesis
bioF	15609.00	13413.00	21253.00	5258.00	14262.00	5831.00	5489.00	6070.00	6447.00	
bioD	2043.00	1951.00	3079.00	985.00	2591.00	1480.00	864.00	985.00	961.00	Biotin biosynthesis
bioC	5194.00	4534.00	7602.00	2014.00	5236.00	2467.00	1952.00	2253.00	2353.00	
ftsK*	7.00	5.00	4.00	0.00	3.00	2.00	4.00	25.00	7.00	Cell division*
flgD	162.00	157.00	270.00	171.00	332.00	318.00	315.00	365.00	292.00	Flagellum
hspQ	581.00	474.00	767.00	521.00	677.00	1241.00	372.00	372.00	261.00	Heat shock response
AFA41244.1	458.00	383.00	671.00	139.00	457.00	127.00	212.00	150.00	160.00	Membrane protein
tRNA Leu	189.00	160.00	255.00	34.00	131.00	65.00	111.00	77.00	66.00	tRNA
tRNA Tyr	3525.00	3196.00	4357.00	1059.00	3744.00	1319.00	1530.00	1427.00	954.00	
tRNA Leu	5517.00	5421.00	8635.00	2560.00	7296.00	2261.00	2663.00	2371.00	1643.00	
tRNA Gly	3620.00	3329.00	4480.00	1444.00	3895.00	1315.00	1606.00	1617.00	1359.00	
tRNA Thr	2338.00	2125.00	3036.00	752.00	2701.00	991.00	1141.00	1020.00	654.00	
tRNA Gly	5404.00	5294.00	8048.00	2339.00	7053.00	2273.00	2610.00	2336.00	1585.00	
tRNA Trp	183.00	147.00	162.00	33.00	116.00	28.00	19.00	64.00	23.00	
tRNA Cys	6444.00	6243.00	9643.00	2788.00	8309.00	2546.00	2960.00	2702.00	1820.00	
tRNA Asp	164.00	122.00	128.00	30.00	105.00	24.00	15.00	62.00	23.00	
tRNA His	1972.00	1868.00	2583.00	610.00	2199.00	775.00	1129.00	827.00	601.00	
tRNA Leu	499.00	491.00	566.00	156.00	544.00	224.00	202.00	279.00	160.00	
tRNA Arg	1856.00	1727.00	2405.00	562.00	2053.00	734.00	1072.00	778.00	534.00	

4.5.9.3 Serratia

No DE was observed in either the control or infected midguts across the time series.

4.5.9.4 Wolbachia

No DE genes were found in the control group, and five DE genes were described in the treatment group (Table 4-10). Three respiration-related genes were progressively upregulated from early to late-stage trypanosome-infected midguts, and genes for mRNA degradation and metabolic processing were downregulated from early to late-stage midguts.

4.5.10 Q9: Changes in expression over the course of trypanosome infection in flight muscle

4.5.10.1 *Sodalis, Wolbachia* and *Wigglesworthia*

Wigglesworthia, which is found exclusively in the midgut-associated bacteriome, was excluded from this analysis group. No *Sodalis* or *Wolbachia* genes were differentially expressed across time in tsetse flight muscles, either in the control group or the treatment group.

4.5.10.2 Serratia

No *Serratia* genes were DE between the early and mid-timepoints in control flight muscles. In the infected group one gene involved in amino acid activation was downregulated from early to mid-stage infections, then upregulated strongly in the late infection stage.

Table 4-10 - DE genes in *Wolbachia* from infected tsetse midguts from all three timepoints, with an FDR of <0.05. Numbers are CPM. Gene descriptions are summaries from the COG, GO terms and protein families for each gene. I1MG = early-stage infected midguts, replicates A-C; I2MG = mid-stage infected midguts, replicates A-C; I3MG = late-stage infected midguts, replicates A-C.

Gene name	I1MG_A	I1MG_B	I1MG_C	I2MG_A	I2MG_B	I2MG_C	I3MG_A	I3MG_B	I3MG_C	Gene description
WGMM_RS05355	1831.50	3099.81	8946.32	1977.59	1067.81	460.83	0.00	0.00	522.19	Metabolic process
rnpB	14346.76	0.00	6461.23	5932.76	4805.13	8294.93	0.00	324.15	522.19	mRNA degradation
nuoJ	610.50	0.00	0.00	659.20	1067.81	460.83	1473.84	11345.22	3655.35	Respiration
atpF	305.25	0.00	0.00	1318.39	1067.81	1382.49	3684.60	7131.28	4177.55	
nuoJ	610.50	619.96	1491.05	1977.59	2135.61	1843.32	3684.60	15235.01	2610.97	

4.6 Discussion

4.6.1 Study limitations

There were certain comparison groups that were unable to be studied in this work, either due to difficulties collecting the samples or extracting sufficient RNA for sequencing. There were no control or refractory flies in the late-stage group, which meant that effects of late-stage trypanosome infection could not be definitively attributed to the infection rather than the tsetse fly age. Six samples had to be removed due to low read outputs; all six were flight muscle tissue. Although these tissues contained more mass than the midgut tissues, RNA extractions were more likely to fail due to the difficulties of homogenising carapace present in the samples which blocked the extraction column filters. Subsequently all flight muscle samples in the mid-experimental refractory group were lost, preventing the examination of whether bacterial transcription in flight muscles may be related to refractoriness to trypanosome infection.

The amount of time available for analysis of this dataset was a limiting factor; the decision was made to focus on the four bacterial symbionts, without examining the trypanosome or tsetse transcription in detail. This may have resulted in important interactions between the bacteria, host, and parasite being overlooked. However, this approach allowed a deep study of the bacterial symbionts which was appropriate for this thesis, and the complete dataset will be available for further analysis later on.

4.6.2 *Sodalis* expression differences *in vitro* compared to *in vivo* show *Sodalis* has retained some metabolic flexibility

Sodalis is a secondary endosymbiont in the process of evolving a closer, more obligate symbiotic relationship with its tsetse host [60]. As such, it still retains many of the genes required to live outside of the host, allowing it to be cultured *in vitro*. *Sodalis* has previously been described as non-motile in tsetse midguts despite possessing a full set of genes capable of producing a flagellum [60, 242]. This is contradicted by the results in this study, which describe an increase in flagellar gene expression *in vivo* compared to *in vitro*. However, this could suggest an alternative role of *Sodalis* flagellar genes in the tsetse host; the flagellar apparatus is similar in structure to the TTSS, sharing a common ancestor [532], and flagellar proteins are known to facilitate host cell invasion in other proteobacteria including *Burkholderia* [533] and *Campylobacter jejuni* [534]. The upregulation of flagellar genes *in vivo* could indicate a similar role of flagellar proteins in *Sodalis* invasion of the tsetse host cells.

The main category of upregulated genes in *Sodalis in vitro* cultures compared to within the tsetse midgut were phage-associated genes, making up 13 of the 52 examined DE with an absolute log fold change of >4. Facultative symbionts like *Sodalis* often have multiple phage associations which can have an effect on the host; for example, the aphid symbiont *Hamiltonella defensa* phage APSE contain toxins that, when expressed, are responsible for terminating the development of parasitoid wasp larvae in the aphid [535]. The *Wolbachia* phage WO carries ANK genes which have been linked to the reproductive manipulation phenotype of *w*Mel, although the link between the WO phage and reproductive manipulation has not been confirmed

[536]. Phage elements make up 17.8% of the *Sodalis* genome CDS [108]; *Sodalis* lytic phage have been implicated in mediating the impact of *Sodalis* on tsetse fly refractoriness to trypanosome infection by affecting *Sodalis* density, with lower *Sodalis* densities associated with decreased susceptibility to trypanosomes [514, 537]. Most of the 13 DE phage genes are uncharacterised except for two tail-related proteins and a baseplate assembly protein, whose upregulation suggests that the lytic cycle may have been stimulated more in *in vitro* cultures compared to in tsetse [538]. Bacteriophage lytic phases can be triggered by increased bacterial density via quorum sensing [539]; the relatively high densities of *Sodalis* in the *in vitro* cultures may have triggered the lytic phase of *Sodalis* phage, explaining their substantial increase in expression compared to the *in vivo* population.

4.6.3 A potential role of *Sodalis* pathogenicity in increasing tsetse vector competence

No bacteria except *Sodalis* showed any differential expression in trypanosome infected flies compared to trypanosome refractory flies, whereas expression of seven *Sodalis* genes was significantly different. There was also no significant expression of tsetse host genes between infected and refractory flies (data not shown). This adds supporting evidence to the theory that maternally-inherited tsetse susceptibility to trypanosome infection can be mediated by *Sodalis* rather than other tsetse-associated bacteria or by the host itself [57].

All seven differentially expressed *Sodalis* genes between refractory and infected midguts were upregulated in infected flies, suggesting a potential response of *Sodalis* to trypanosome challenge that directly increases tsetse susceptibility to

trypanosome infection. Of these, two genes are associated with the TTSS, *sipB* and *sicA.* TTSS have a variable role in host-symbiont interactions: often used by parasitic bacteria for cellular invasion or delivery of effector molecules into host cells, they are widely classified as virulence factors [259]. However, TTSS have also been hypothesised to play an important role in the evolution of mutualistic endosymbiosis, facilitating colonisation of host cells [465]. *Sodalis* retains the ability to invade cells via two functional TTSS, with some genetic degradation indicating a diversification of function from classic virulence resulting in potentially novel and unknown roles of the *Sodalis* TTSS in host interactions [540]. The gene *sicA* codes for a chaperone protein involved in situating TTSS apparatus in the bacterial membrane; *sipB* is homologous to the *Salmonella enterica* TTSS gene *yspB*, an effector protein that facilitates modification of the host cytoskeleton during bacterial invasion [540-542]. *Sodalis* has two functional TTSS; SSR-1, which includes *sipB*, is required for invasion of host cells; and SSR-2, which includes *sicA*, facilitates bacterial proliferation after host cell entry [540]. The upregulation of genes from both TTSS in trypanosome-infected tsetse compared to refractory tsetse suggests that both host cell invasion and *Sodalis* proliferation within cells is linked to increasing tsetse susceptibility to trypanosomes.

Also upregulated in infected tsetse was *ompF*, an outer membrane porin which facilitates diffusion of small molecules across the bacterial membrane and acts as a membrane receptor [543]. Porins are protein structures which can allow the transport of molecules in and out of the bacterial cell, act as receptors for host effector molecules or for binding to host cells, and can facilitate the invasion of host cells [544]. The *Sodalis* porin gene *ompA* has an important role in modulating the tsetse

immune response and mediates *Sodalis* biofilm formation which is essential for successful colonisation of the tsetse midgut [421, 545]. Porin expression may have a role in increasing tsetse susceptibility to trypanosome infection, either through modifying *Sodalis* biofilm formation in the midgut or by mediating bacteria-host cell interactions.

As well as the TTSS and porin genes, which are classed as virulence factors due to their roles in cellular invasion [259, 546], the gene *fliC* was upregulated in infected tsetse midguts compared to refractory midguts. *FliC* forms the major flagellum subunit and is involved in cellular adhesion and invasion of host cells [547]. In the early, mid and late-stage midgut comparison groups, three flagellar genes and ten TTSS genes were strongly upregulated in the early infection group. While the trypanosome infection status of early-stage flies is unknown, the strong presence of flagellar and TTSS transcripts suggests that these systems are involved in the *Sodalis* response to trypanosome infection, and thus potentially the *Sodalis*-mediated susceptibility of tsetse flies to trypanosome infection.

A previous study by Soumana *et al.* [514] examined *Sodalis* expression in *G. palpalis gambiensis* infected with *T. brucei gambiense* and found 176 differentially expressed genes in refractory flies compared to infected flies, including upregulation of genes relating to cell death, bacteriolytic enzymes and defence mechanisms in trypanosome-refractory flies. Soumana *et al.* suggested based on their results that increased *Sodalis* density inducing the lytic phase of a bacteriophage may play a role in increasing tsetse refractoriness by triggering the tsetse fly innate immune response. Both TTSS and porin proteins can be linked to bacteriophage function as

receptors and were upregulated in infected flies in our study rather than refractory flies as would be suggested by the density-dependent hypothesis, but no DE of phage genes were noted. However, although the mean percentage of reads mapped to *Sodalis* in each treatment group (here used as a proxy for *Sodalis* density) were not significantly different, the infected tsetse appeared to have a lower density of *Sodalis* than the refractory tsetse (Figure 4-6).

Different tsetse species harbour genetically distinct populations of *Sodalis* with different expression profiles [517]; these differences have been linked to variation in vector competence between tsetse species [548], potentially explaining the discrepancy between this study and the Soumana *et al.* study [514]. Even within *Glossina* species and populations *Sodalis* demonstrates genetic diversity, with variation in the amount of diversity depending on the tsetse population size, and with some *Sodalis* gene flow between tsetse populations up to 150km apart [253]. Characterising the polymorphisms present in *Sodalis* populations and their effect on tsetse susceptibility to trypanosome infection may be essential for understanding the role of *Sodalis* in tsetse vector competence. If there are different *Sodalis* genotypes present in this lab-reared tsetse colony this may explain the variation in *Sodalis* responses to trypanosome challenge.

4.6.4 *Wigglesworthia* expression changes over time, with unclear links to ongoing trypanosome infection

Wigglesworthia gene expression was not significantly different between trypanosome-refractory and infected tsetse flies at the mid experimental stage, or between trypanosome-unchallenged and challenged tsetse flies at the early

experimental stage. This indicates that *Wigglesworthia* does not directly influence the early establishment of trypanosomes in the host midgut. However, differences in *Wigglesworthia* expression were observed over time in both control and trypanosome-infected flies, with double the number of DE genes in the test group. All expression differences were relatively weak (less than a 2.6-fold change for all DE genes in both the control and test groups, either up- or down-regulated) but statistically significant, suggesting the differences could be biologically meaningful. Biotin synthesis pathway genes were strongly upregulated in early infected midguts, with expression gradually reducing over the course of the infection; biotin is a B vitamin, which are essential components of the tsetse diet produced by *Wigglesworthia* [485]. The downregulation of biotin pathway genes seen in this dataset contrasts with the findings of another study of *Wigglesworthia* transcription, which found increasing expression of the biotin pathway associated with increased age in trypanosome-naïve wild *G. pallidipes* [549]. This indicates that *Wigglesworthia* may adjust its biotin production in response to the impacts of the worsening trypanosome infection on the host, such as a heightened tsetse immune response, tissue destruction or resource depletion.

In a similar study performed by Soumana *et al. Wigglesworthia* expression in *T. b. gambiense*-infected *G. p. gambiense* was studied using microarrays [550]. Direct comparison between the two datasets is difficult as the Soumana *et al.* study only compared time-matched infected and refractory midguts, and for these comparison groups our study found no DE genes. As with the differences between the present study and the *Sodalis* transcription study from the same group [551], whether the

disparities are due to differences between microarray and RNAseq data or between the tsetse and trypanosome species being studied is uncertain.

4.6.5 *Serratia* and *Wolbachia* have a limited response to trypanosome infection in tsetse

Very little DE was observed in both *Serratia* and *Wolbachia* in response to trypanosome challenge, or between the control groups. *Serratia* did show the only DE gene seen in flight muscle tissue; however, as only one gene was DE and many of the flight muscle samples were incompletely processed it is difficult to draw any conclusions from this. *Wolbachia* demonstrated a small response to increasingly advanced trypanosome infection with an expression of genes involved in respiration, possibly to compensate for the reduced resources being taken up by the trypanosomes. While conclusions are difficult to draw from such a small number of DE genes, it can be hypothesised that *Serratia* and *Wolbachia* do not have as important a role in determining the outcome of a trypanosome infection in the tsetse host as the symbionts *Sodalis* and *Wigglesworthia*.

4.6.6 Further study

The present study gives multiple novel avenues for exploration of the role of *Sodalis* in tsetse trypanosome susceptibility. *Sodalis,* as a secondary symbiont, exhibits broad tissue tropism within tsetse; the study of different *Sodalis* populations within trypanosome-refractory and infected tsetse tissues, such as intracellular populations in the bacteriome or extracellular populations in the haemolymph, may elucidate further the different roles *Sodalis* can play in tsetse vector competency. In addition, genomic or epigenetic differences between *Sodalis* populations from multiple tissues

within the tsetse may help explain the differences in trypanosome infection outcome and could be studied by the sequencing of the pangenome of *Sodalis* from a single lab-reared tsetse colony, or even a single individual. Finally, the density and biofilm formation of *Sodalis* within trypanosome-susceptible tsetse compared to trypanosome-resistant tsetse should be studied to determine whether these factors influence tsetse susceptibility to trypanosomes.

The full dataset produced in this study could not be examined in the depth warranted by its complexity, due to constraints on time and the scope of this thesis. Therefore, there have been questions raised by the data that have not been answered; for example, was the large amount of tRNA expression in the *Wigglesworthia* dataset – which was not seen in any of the other bacteria studied – due to host contamination, sequencing artefacts, or potentially a regulatory role of tRNA in *Wigglesworthia* gene expression [552]? There are also a substantial number of host and trypanosome reads that have only been superficially analysed, leaving a rich dataset for future investigation.

Whole metaorganism transcription datasets are large, complicated, and rely on the accuracy and completeness of gene databases. A large proportion of differentially expressed genes in this study were annotated as hypothetical proteins, with no known or proposed function based on homology to other genes. Until the genomes of the organisms involved in tsetse vector competence are more completely annotated, the use of genomics to elucidate their roles will be limited.

Chapter 5 Development of a 16S pipeline for microbiome analysis of low-biomass samples

5.1 Abstract

The microbiome of sand flies, sole vectors of leishmaniasis, has been shown to alter their vectorial capacity and the dissemination of *Leishmania* parasites in the mammalian host tissue. Sequencing the microbiome of small insects presents technical challenges arising from the low bacterial DNA yield and large proportion of contaminating taxa. The novelty of the field has resulted in multiple approaches to producing and analysing microbiome data with little consistency between studies, reducing the comparability of datasets and raising questions about what constitutes best practice in microbiome research. In this study a complete microbiome sequencing pipeline, from DNA extraction to data analysis, was produced and refined using a sample of 240 individual *Phlebotomus argentipes* sand flies collected from six districts in Bihar, India. Different approaches to sequencing the 16S gene were compared, including two different primer sets and novel contamination control strategies. The complete, filtered and accurate dataset of individual *P. argentipes* microbiota produced in this chapter will be investigated in Chapter 6 to determine the impact of location on the sand fly microbiome at fine scales. This work constitutes an important toolkit of approaches for sequencing microbiomes of individual small insects in an emerging field.

5.2 Introduction

5.2.1 The importance of studying individual sand fly microbiomes

Sand flies (genera *Phlebotomus* and *Lutzomyia*) are the sole vectors of the pathogenic kinetoplast *Leishmania*, which causes the neglected tropical disease leishmaniasis (Section 1.6). The microbiome of sand flies impacts *Leishmania* transmission dynamics, both by enabling infection of the sand fly with *Leishmania* parasites [89, 311] and facilitating parasite dissemination at the injection site of a mammalian host [553]. The nature of these interactions is currently unknown; so far only the impact of bacterial presence or absence on *Leishmania* transmission has been studied, rather than any particular bacterial species. The sand fly microbiome has not been described in as much detail as other vectors; for example, the tsetse fly microbiome has been well characterised and studied for over half a century [337]. This is due to the constraints of sequencing technology; sand flies are 1.5-3.5mm long [303] and therefore contain a low bacterial titre, with a very small amount of bacterial DNA available to sequence.

In contrast to tsetse flies, which have a primary symbiont and very similar microbiome structure throughout the genus, the life history of sand flies indicates that their microbiota may vary between individuals. Adult sand flies have a small home flight range of around 100m, with a varied diet of plant sugars; females only require a blood meal to produce egg batches, laying eggs in masonry or the ground which feed on leaf litter during larval stages [303]. Diet specialisation, such as the adenotrophic viviparity and obligate haematophagy of tsetse flies (see Section 1.5.2), often drives the evolution of symbiotic relationships between insects and

bacteria [100]. This pressure is absent in sand flies; sand fly life history suggests they will have a more fluid microbiome closer to that of their current environment, therefore strongly associated with fine geospatial scales and microhabitats. However, due to their low bacterial titre individual sand fly microbiomes are rarely studied using culture-independent methods (for a summary of microbiome study methods see Section 1.7). The sand fly species *Phlebotomus argentipes* is the sole vector of visceral leishmaniasis in its primary foci, the northern Indian district of Bihar (see Section 1.6.1.1). Despite the species' medical importance, and the potential for microbiome differences within a target population to impact future control strategies, the microbiome of *P. argentipes* individuals has never been described.

In this chapter we developed a complete pipeline of best practices for the accurate analysis of the microbiota of individual sand flies, from DNA extraction to bioinformatic analysis, in order to increase our understanding of *P. argentipes* microbiome dynamics across a fine geospatial scale in Bihar, India. The microbiome data produced using this pipeline, and the biological effects of geography on sand fly microbiota, will be discussed in Chapter 6.

5.2.2 Using sequencing in insect microbial community analysis

VBD transmission dynamics include the vector's resident microbiota as an integral factor that can affect development, survival and vector competence [554, 555]. Microbiome studies include measurements of microbial species richness (number of species) and evenness within a sample (alpha diversity), and the differences between samples in a population (beta diversity) [556]. Culture-independent methods have revolutionised how we understand the relationships between a host

and its resident microbiota. Previous reliance on culturing methods for microbiome studies meant that a large proportion of bacteria were poorly described or missed entirely (Section 1.7).

Improvements in sequencing technology have made culture-independent studies more affordable, practical and efficient, and increasingly large computing power is becoming available, allowing the generation and analysis of large sequencing datasets; for instance, a single sequencing run of the Illumina NovaSeq can produce up to 20 billion 250bp reads in 44 hours [557]. However, with these rapid increases in capability have come increases in the number of methodological variables and options; despite the large number of microbiome studies published in the last decade comparisons between them are difficult to draw, as there is very little standardisation of methods [360, 558]. Variation in methods at each step, from sample handling through to sequencing and bioinformatic analysis, has the potential to introduce variation comparable to true biological differences [559].

Microbiome composition can be determined using NGS in two ways: sequencing all genetic material present in a sample and aligning to whole known genomes, known as metagenomics [560, 561]; or amplifying and sequencing a region that is conserved between all species of the target population with small variations that can be used to identify the species of origin, known as amplicon sequencing or metabarcoding [562]. While metagenomics can provide a greater breadth of information without the restrictions of sequencing only a single region, amplicon-based microbiome studies can give more sequencing depth with the same amount of sequencing power and avoids sequencing host DNA instead of microbial DNA.

Amplicon sequencing also has reduced computational requirements compared to metagenomics, but is limited to the target organisms, excluding non-conforming targets or other interesting organisms in the analysed communities. Both of these techniques rely on accurate and accessible databases of known microbial sequences to compare sequencing reads to for phylogenetic classification.

The most commonly used sequence to determine the bacterial and archaea species in a community is the 16S gene, which codes part of the small ribosomal subunit of bacteria and archaea. Its use as a genetic marker was developed by Carl Woese to determine the phylogenetic origin of prokaryotes and led to the first description of the three domains of life [353, 354]. The 16S gene is present in all known bacteria and archaea with a largely conserved structure of nine variable regions (V1-V9) bounded by highly conserved regions [355, 356]. Variable regions in the gene can withstand some genetic drift over time without deleterious changes to the structure of the ribosome. These random mutations by genetic drift can happen at a predictable rate, allowing their use as a measurement of the time since divergence between two species [353]. Closely related prokaryote species have similar variable region sequences, allowing the use of 16S sequences to construct phylogenetic trees [356]. The 16S sequences of unknown prokaryotes can be identified by comparison with databases of known organisms' 16S sequences. Sequencing the entire ~1,500bp 16S gene gives the best taxonomic resolution [563], but necessitates the use of longer-read sequencing technologies (see Section 5.2.9). In order to increase sequencing depth (see Section 5.2.8) often only one or two variable regions are sequenced, with primers matching the conserved regions on either side of the targeted area.

5.2.3 Variations in 16S data collection methods

Microbiome methodologies are invariably complex, with many steps and multiple possible approaches to each step. Methodological variation can have a profound impact on not only the results of a microbiome study, but also how comparable those results are to other studies of the same subject matter. Under-reporting of methodology details can result in a lack of reproducibility and presents extra challenges when trying to compare datasets between different studies. There is an increasing drive to standardise methods for microbiome research in order to allow comparisons between studies and the combination of multiple datasets [360, 564-566]. Standardisation of methods requires selecting the 'best' approach for each study in a rapidly evolving field with no current consensus on what the best approaches are [360]. In this section a broad methodology for 16S-based insect microbiome studies will be outlined (Figure 5-1), with a brief discussion of the potential different approaches and the impact of those choices on the final dataset.

5.2.4 Insect sample collection

The type of sample taken for an insect microbiome study can vary sufficiently to affect the results, even within the same sample species. Single tissues may be sampled rather than the whole organism: different bacterial communities are present in different tissues, such as the midgut and cuticle [166]. Whole insects are often surface sterilised using ethanol, with differences in the duration and method of ethanol application. Some insect tissues contain PCR inhibitors, such as compound

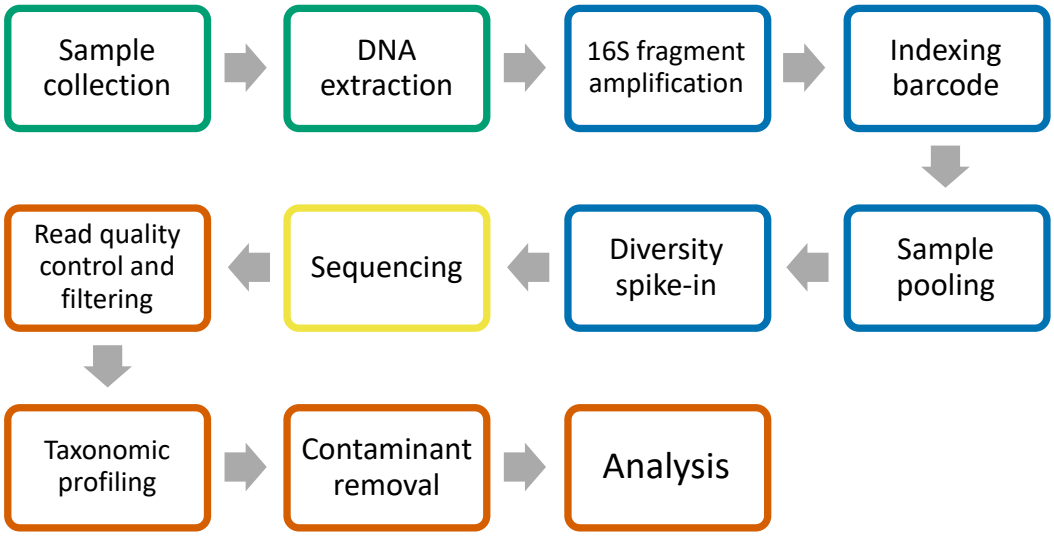


Figure 5-1 - A flowchart of the basic methodology of a 16S microbiome study with steps outlined in different colours. Green: initial sample preparation. Blue: sequencing library preparation. Yellow: sequencing. Orange: bioinformatics-based data analysis.

eyes [426], that can prevent the even amplification of microbial genes. Different sizes of insects or tissues will result in lower DNA yields, potentially reducing the ratio of sample reads to contaminant reads in the final dataset – the impact and removal of contaminating reads will be discussed later in this section. Sample storage before DNA extraction can also impact the DNA yield and microbial diversity observed: both the sample storage temperature, substrate and time can impact both alpha and beta diversity, although conflicting results have been reported [360, 555, 567]. Multiple different storage methods have been shown to preserve microbial diversity in samples; the method of storage used is not as important as ensuring that all samples in a study are treated and stored in exactly the same way to prevent confounding batch factors [360, 568].

The microbiome of insects is usually heavily influenced by their diet and environment [569], with the exception of symbionts which are obligately host-associated, such as

Wigglesworthia in tsetse flies [225]. As a rule, smaller organisms have a less diverse and more easily disrupted microbiome driven by their diet and habitat substrate [570]. Insects reared in laboratory settings are necessarily kept in highly controlled, almost sterile environments to reduce the risk of introducing pathogens to the colony. Lab-reared insects therefore come into contact with fewer microorganisms in their substrate and diet than wild populations and are exposed to fewer abiotic and biotic factor challenges. This is reflected in the structure of their native microbiota, which is much less diverse than that of wild-caught populations of the same species, meaning that microbiome studies of lab-reared insects will not always reflect the reality of their wild counterparts [571]. Lab-reared insects may also share bacteria with each other if housed together, adding confounding effects of housing groups to their microbial diversity.

5.2.5 Contamination control precautions

In low-biomass samples, where there is very little microbial DNA from the specimen, a high proportion of sequences may be primarily derived from contamination [568]. These contaminants are introduced to samples during processing from sources including reagents, dust in the laboratory air, the person processing the samples, or the tools and plastic consumables being used [572]. Some widespread contaminant taxa have been identified in blank control samples across multiple studies using an array of different methods, suggesting there are sources of contamination common to most laboratory environments or reagents (Table 5-1) [384, 572]. Cross-contamination between samples is often a result of human error and can occur at several stages: aerosolization or mistaken pipetting of extracted DNA or PCR product, or accidental transfer of barcode primers between wells, known as barcode

switching [573, 574]. Contamination control measures can be taken during the initial sample preparation and sequencing library preparation stages (Figure 5-1) including thorough cleaning of the lab surfaces and instruments to remove bacteria and environmental DNA, extra sterilisation steps of reagents using DNase treatments [575], and appropriate personal protective equipment including gloves and masks [572]. Steps to reduce the effect of sample batches include using the same working spaces, cleaning protocols and technicians for all samples, processing all samples at the same time and using the same DNA extraction kit as contaminants can vary between batches of the same extraction kit [384, 572, 576].

Table 5-1 - Commonly identified contaminating bacterial species from negative control samples. From Eisenhofer et al. (2019). *Genera previously identified in insect microbiome studies with sequenced negative controls.

Order	**Family**	**Genus**
Actinomycetales	Actinomycetaceae	*Actinomyces** [577]
	Micrococcaceae	*Arthrobacter** [578]*, Rothia*
Bacillales	Bacillaceae	*Bacillus** [579], *Geobacillus*
	Paenibacillaceae	*Brevibacillus** [580], *Paenibacillus** [581]
	Staphylococcaceae	*Staphylococcus** [582]
Bacteroidales	Porphyromonadaceae	*Porphyromonas*
	Prevotellaceae	*Prevotella*
Burkholderiales	Alcaligenaceae	*Achromobacter** [583]
	Burkholderiaceae	*Burkholderia** [584]*, Cupriavidus** [577]*, Ralstonia** [311]
	Comamonadaceae	*Acidovorax** [585], *Comamonas** [586], *Curvibacter*, *Leptothrix** [587], *Pelomonas** [306]
	Oxalobacteraceae	*Duganella** [587], *Herbaspirillum** [588]*, Janthinobacterium** [586]*, Massilia, Oxalobacter** [589]
Caulobacterales	Caulobacteraceae	*Brevundimonas*
Clostridiales	Clostridiaceae	*Clostridium*
	Lachnospiraceae	*Coprococcus** [590]
	Peptoniphilaceae	*Anaerococcus*
Coriobacteriales	Coriobacteriaceae	*Atopobium*
Corynebacteriales	Corynebacteriaceae	*Corynebacterium** [591]
Enterobacterales	Enterobacteriaceae	*Escherichia** [592]
Flavobacteriales	Flavobacteriaceae	*Capnocytophaga, Chryseobacterium** [312] ,
Fusobacteriales	Fusobacteriaceae	*Fusobacterium*

Order	Family	Genus
	Leptotrichiaceae	*Leptotrichia*
Lactobacillales	Aerococcaceae	*Abiotrophia*
	Carnobacteriaceae	*Granulicatella*
	Enterococcaceae	*Enterococcus** [579]
	Lactobacillaceae	*Lactobacillus** [593]
	Streptococcaceae	*Streptococcus** [583]
NA	NA	*UnclassifiedTM7*
Neisseriales	Neisseriaceae	*Kingella** [313], *Neisseria** [583]
Pasteurellales	Pasteurellaceae	*Haemophilus*
Propionibacteriales	Propionibacteriaceae	*Propionibacterium*
Pseudomonadales	Moraxellaceae	*Acinetobacter** [579]
	Pseudomonadaceae	*Pseudomonas** [579]
Rhizobiales	Bradyrhizobiaceae	*Afipia, Bradyrhizobium** [313]
	Hyphomicrobiaceae	*Devosia*
	Methylobacteriaceae	*Methylobacterium** [579]
	Phyllobacteriaceae	*Mesorhizobium, Phyllobacterium*
	Rhizobiaceae	*Rhizobium*
Rhodospirillales	Acetobacteraceae	*Roseomonas*
	NA	*Enhydrobacter** [312]
Selenomonadales	Veillonellaceae	*Dialister, Megasphaera, Veillonella*
Sphingobacteriales	Chitinophagaceae	*Sediminibacterium*
	Sphingobacteriaceae	*Pedobacter, Novosphingobium** [594], *Sphingobium** [594], *Sphingomonas** [594]
Xanthomonadales	Xanthomonadaceae	*Stenotrophomonas** [312], *Xanthomonas** [595]

In order to validate contamination reduction procedures and to perform post-sequencing contamination filtering, control samples should be sequenced alongside true samples. This includes negative and positive controls: negative controls contain no sample DNA and can include swabs from the sampling site or the lab space, reagent-only DNA extractions or reagent-only PCRs. Negative controls can be used in a variety of ways post-sequencing to reduce the effect of contaminating species in true samples – this will be discussed in Section 5.4.7.2. Positive controls are samples with a known quantity of known microbial species, such as a bacterial monoculture or a microbial community present in known ratios. Positive controls can be used to validate the accuracy of the sequencing and ensure that the extraction and sequencing methods used don’t under or over-represent certain taxa [568]; they

can also be used to identify contaminants, by examining any non-expected genera. Pre-prepared mock bacterial communities are available commercially and can either be whole cells or pre-extracted DNA. Both of these can be used to calculate whether certain taxa are under- or over-represented by the sequencing methods used; whole cell mock communities also assess the DNA extraction protocol for taxa biases. However, they are necessarily limited by being oversimplified compared to natural microbiomes and therefore can't be used to conclusively rule out misrepresentation of taxa in real samples that aren't present in the mock community [564]. Despite the utility of positive controls, mock communities are rarely included in standard microbiome sequencing runs. Including negative controls in sequencing runs is considered essential for accurate microbiome analysis [360, 384]; however, negative samples are rarely mentioned in microbiome studies, and when used are usually discarded before sequencing after PCR or qPCR analysis to confirm the low number of bacterial sequences.

5.2.6 DNA extraction and storage

DNA extraction methods can vary, with implications for the diversity and richness of the bacterial species identified; for example, in soil studies different extraction methods produce entirely disparate community profiles [596, 597]. DNA extraction methods must be chosen to best suit the sample type, in order to produce a high DNA yield and an even coverage of bacterial species, as different sample types present unique challenges. For example, whole insect samples need a thorough physical disruption step, such as bead-beating or freezing with liquid nitrogen before grinding, to break down the hard external chitin [598]. Enzymatic digestion steps, usually using proteinase K, are usually required in conjunction with mechanical

disruption to break down the thicker cell walls of gram-positive bacteria; without adequate disruption of gram-positive bacterial cell walls community analysis can be skewed towards gram-negative species [599, 600]. Finally, chemical disruption of both host and bacterial cell membranes is required to release DNA, usually performed using detergents such as sodium dodecyl sulphate or cetyltrimethylammonium bromide [349]. DNA is purified from the lytic solution either by isopropanol precipitation or by binding the DNA to a silica membrane before washing. Most commercial extraction kits use silica membranes in microcentrifuge columns which can be easily blocked by larger pieces of insect carapace, reducing the DNA yield; this can be avoided with adequate mechanical disruption and additional centrifugation steps to separate the supernatant from any large pieces. While commercial DNA extraction kits are quicker and easier to use than reagent-based methods such as phenol-chloroform precipitation [598], they can be an importance source of contamination despite careful contamination controls during manufacture, with variation between reagent batches [384].

5.2.7 Pooling samples

As the microbial DNA yield of some insects is very low, most studies of small insects will pool multiple individuals into a single sample before sequencing. Pooling can be performed before or after DNA extraction, and samples are pooled together based on shared characteristics, with the assumption that their microbial communities will be largely similar, if not identical. However, even highly similar individuals can have substantially varying microbiota [579], therefore using pooled samples risks erasing inter-individual variation [555]. Insect microbiota are often dynamically influenced by their immediate environment and therefore can react rapidly to changing conditions

or fine geographical scales and influence the adaptation of the host, making understanding the individual microbiota of vectors important for biocontrol strategies [601].

5.2.8 Sequencing library preparation

The methods used to prepare a microbiome DNA sample for sequencing can impact the diversity of the identified microbiota. Sequencing libraries are pools of DNA samples, fragmented to an appropriate size and diluted to an appropriate concentration for the sequencer, with attached adapters that allow the sequencer to capture and sequence each sample. These libraries either consist of all DNA in a sample (metagenomic sequencing) or targeted and amplified fragments (amplicon sequencing) (see Section 1.7.2). Metagenomic library production can also involve the universal amplification of all DNA in a low-biomass sample to enrich it sufficiently for sequencing [602, 603]. However, different amplification chemistries can amplify some microbial DNA either preferentially or ineffectively, skewing the observed microbial diversity and reducing comparability between studies [604]. Amplicon sequencing for microbiome studies usually targets one or more variable regions of the 16S gene; the choice of variable region and primer can significantly influence the microbiome by under- or over-amplifying certain taxa [362, 566, 605]. Amplification of genes by PCR can also introduce errors where two or more separate reads are merged together, producing a chimeric read; the chances of chimera production increase with the number of PCR cycles, producing a necessary trade-off between sufficient amplification of smaller samples and accuracy [606]. Current sequencing technologies require at least 1ng of template DNA for sequencing, creating the need for pooled samples (see Section 5.2.7), increased PCR cycles, or sample

enrichment processes that can introduce bias [607]. There is an urgent need for the development of unbiased amplification methods or improved library preparation strategies that require smaller DNA inputs [603].

In order to capture all taxa within a sample, and therefore produce data from which biological conclusions can reliably be drawn, each sample must be sequenced to an adequate depth. This is achieved by sequencing enough reads per sample that an increase in the number of reads sequenced would not produce an increase in the number of observed taxa. This is often measured using rarefaction curves, an ecological methodology whereby the number of observed taxa in multiple subsamples of the sample are plotted against the number of reads in the subset to produce a curve; if the curve flattens out with increasing reads then the sequencing depth is adequate for the sample [608]. Without sufficient sequencing depth rarer taxa can be missed, producing an artificially low microbial diversity [609]. Depth can be increased by using multiple sequencing runs to increase the overall number of reads sequenced; however, variation between runs can be a strong compounding factor, as will be discussed later. The number of samples in a sequencing library may also impact the sequencing depth. In order to pool multiple samples a unique barcoding index tag is attached to each sample before pooling, allowing DNA fragments to be demultiplexed by sample after sequencing. This theoretically allows a large number of samples to be sequenced at the same time; however, most sequencers are limited by the number of fragments they can sequence in a single run, therefore the more samples there are in the library the less depth the samples will be sequenced to. Additionally, although improvements have been made in modern sequencing technologies, pooled samples in a run can switch barcodes

either through human error during multiplexing or through sequencing errors, making accurate demultiplexing of samples after sequencing difficult [610].

There is naturally variation in the amount of DNA present in each sample; therefore in order to ensure an even sequencing depth of all samples, and prevent over-sequencing of more concentrated samples, they must be accurately quantified and pooled to an even concentration in the sequencing library. DNA quantification for next-generation sequencing can be performed using a variety of tools and methods, including spectrophotometry, fluorometry and qPCR, with no gold standard method currently agreed upon, reducing the comparability of studies using different methods [611-613]. The most accurate quantification methods – qPCR, Bioanalyzer and Qubit fluorometry – are also the most expensive, restricting the number of samples that can be analysed in a project. Inaccurate pooling will reduce the sequencing depth of the lowest samples and reduce the comparability of samples in the study.

5.2.9 Sequencing platform

The choice of sequencing platform used for microbiome studies can impact the observed species richness and diversity, whether through the library preparation method required by the platform [614] or the sequencing technology itself. Sequencing-by-synthesis (SBS) technologies, such as Illumina, replicate strands of DNA on a surface using fluorophores to distinguish each base as the strands are produced, which can be read to form a consensus of the read sequence [615]. Pacific Biosciences-developed single-molecule real time (SMRT) sequencing uses a similar fluorophore-based method, measuring the change in fluorescence as fluorophores detach from a single nucleic acid molecule in real time, allowing

continuous simultaneous detection of base pairs in single DNA strands [616]. The newest next-generation sequencing technology, produced by Oxford Nanopore Technologies, passes single nucleic acid strands through biological ‘pores’ on a membrane, measuring the changes in ionic current as they pass through to determine the base sequence [617] SBS produces high-quality shorter reads, whereas SMRT and Nanopore sequencing produce longer reads with higher error rates [618]. Different post-sequencing quality control is required depending on the sequencing platform; for instance, Illumina SBS reads deteriorate in quality towards the end, allowing the trimming of low quality reads, whereas SMRT and Nanopore sequencing have evenly distributed errors requiring different strategies to remove errors [618]. Read quality control will be described further in Section 5.4.7.

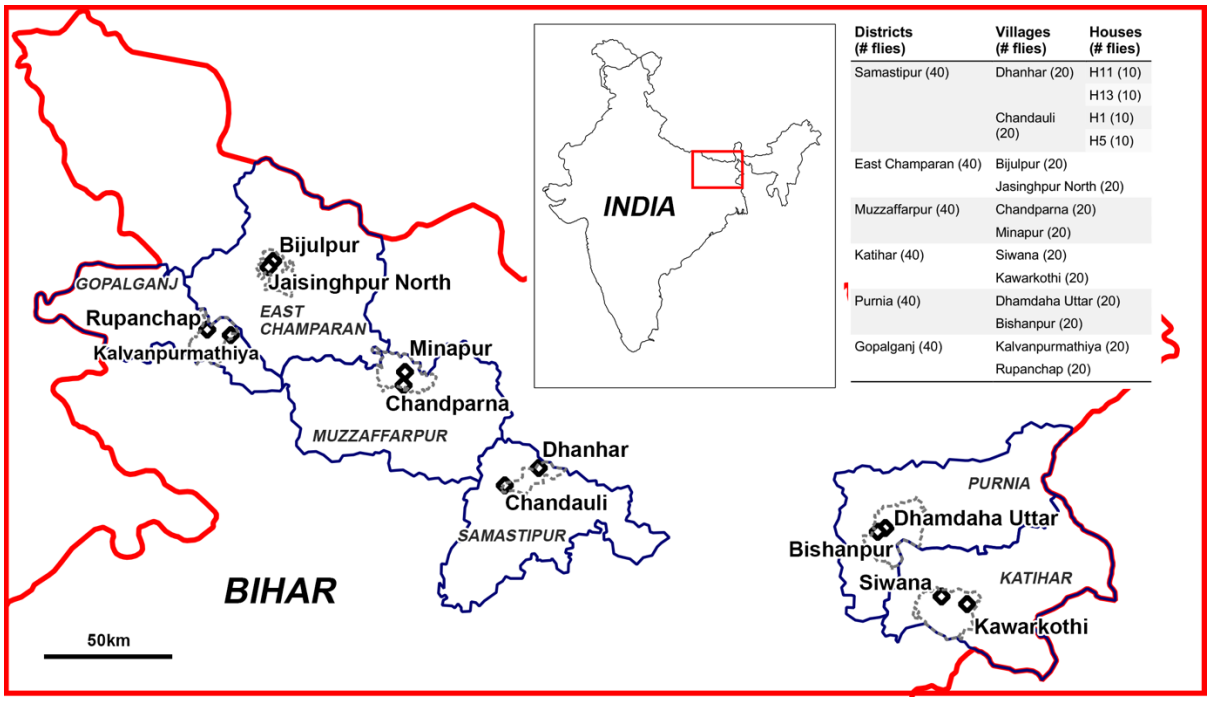
5.3 Aims and objectives

In this study, we aimed to produce a dataset of sand fly microbiome 16S reads using a well-designed sequencing methodology, comparing multiple strategies and incorporating the best practices for producing accurate results from low-biomass samples. With this dataset we aimed to create and assess a post-sequencing quality control and analysis pipeline suitable for low-biomass samples, including the removal of contaminating reads. The data generated from sand fly samples using this pipeline will be used to assess the core and peripheral microbiome of individual sand flies at fine geospatial scales in Chapter 6.

5.4 Methods

5.4.1 Sand fly collection and storage

All work in India was conducted by the Rajendra Memorial Research Institute and coordinated by Dr Mike Coleman and Dr Geraldine Foster at the Liverpool School of Tropical Medicine (LSTM). Sand flies were collected from sentinel sites in six VL-endemic districts in Bihar State, India: East Champaran, Gopalganj, Katihar, Muzaffarpur, Purnia, and Samastipur (Figure 5-2). Two villages within the same primary health centre block were selected per district. Sand fly collections were performed August-December 2016 in six randomly selected houses per village using mini CDC light traps installed overnight in participant household bedrooms. Collected sand flies were identified morphologically by sex and species and stored individually in 1.5 ml tubes with silica gel desiccant. Stored sand flies were transported to the LSTM for molecular characterisation. All work in the UK was performed by myself,



Districts (# flies)	Villages (# flies)	Houses (# flies)
Samastipur (40)	Dhanhar (20)	H11 (10)
		H13 (10)
	Chandauli (20)	H1 (10)
		H5 (10)
East Champaran (40)	Bijulpur (20)	
	Jasinghpur North (20)	
Muzzaffarpur (40)	Chandparna (20)	
	Minapur (20)	
Katihar (40)	Siwana (20)	
	Kawarkothi (20)	
Purnia (40)	Dhamdaha Uttar (20)	
	Bishanpur (20)	
Gopalganj (40)	Kalvanpurmathiya (20)	
	Rupanchap (20)	

Figure 5-2 - A map of Bihar, India showing the location of sand fly sampling sites within each district and primary health centre block, and the number of sand flies from each site in the final library preparation subset.

with assistance from Dr Foster and Gala Garrod from the LSTM during the DNA extraction process.

5.4.2 DNA extraction and 16S library preparation

DNA extraction was performed on 792 female sand flies less than 6 months after capture. Following surface sterilisation by swabbing with a Whatmann tissue soaked in 70% ethanol before air drying for two minutes, DNA was extracted from whole individuals using the Zymo Research Tissue and Insect DNA Kit, eluting into 25 µl of molecular biology-grade water. Sand flies were extracted in batches of 23 with a negative control in each batch containing 50 µl molecular biology-grade water. Four reagent-only extractions were also performed separately to the sample extraction runs, using the same equipment and methodology to give an accurate representation of the “kit-ome” without the risk of cross-contamination from sand fly samples. A mock community (ZymoBIOMICS Microbial Community Standard) containing whole cells of eight bacteria and two yeasts was also extracted with the same kits, for use as a positive control during sequencing. All DNA extraction kits were purchased at the same time to limit between-batch variation, and all extractions were performed at the same bench to reduce contamination differences between extraction runs. Extracted DNA from 792 individual flies, 38 water controls and four reagent-only controls was stored at -80°C before further use.

5.4.3 PCR-based confirmation of leishmania infection status and sand fly species

Details of all primers used can be found in Table 5-2. Each sand fly sample was tested for the presence of *Leishmania* parasites by ITS-1 PCR with OL1854/OL1853

primers [619, 620]. Briefly, 12.5 µl of Bioline MyTaq Red Mix (#BIO-25043), 8 µl of molecular biology-grade water, 1.25 µl of forward and reverse primers and 2 µl of extracted sand fly DNA were cycled through the following temperatures: 95°C for 1 minute, followed by 35 cycles of 95°C for 15 seconds, 55°C for 15 seconds and 72°C for 10 seconds, before a final extension step of 72°C for 2 minutes. PCR products were compared to extracted *L. donovani* DNA by gel electrophoresis. Positive amplifications were used in a restriction enzyme reaction to identify the species of *Leishmania* present: briefly, 12 µl of New England Biolabs (NEB) Hae III restriction enzyme (#R0108), 5 µl of NEB CutSmart buffer (#B7204G), 5 µl of NEB 1mg/ml bovine serum albumin (#B9000S), 33 µl of molecular-grade water and 5 µl of the previous PCR reaction were incubated at 37°C for 2 hours before examination by gel electrophoresis for fragment size.

Sand fly family was confirmed by PCR with SUN_PA primers (Table 5-2) specific to a region of the 18S gene unique to phlebotomine sand flies, with Hae III digestion to identify down to species level as previously described [620]. Conditions and reagents for the PCR and enzyme restriction were the same as used in detecting *Leishmania* parasites.

5.4.4 Geographic analysis design

Sand flies were identified by district, village, and – in Samastipur – house number (H1-H16). A sample subset was created from flies that had been positively identified as *P. argentipes* and as leishmania-free by PCR; this subset consisted of 40 flies per district in 6 districts, evenly split between two villages per district (Figure 5-2, inset).

Samastipur samples were further divided into two houses per village with 10 flies per house.

Table 5-2 - Primer details for sand fly species, leishmania presence and amplification of 16S gene for bacterial identification by MiSeq sequencing. Ambiguous nucleotides: Y = C or T; M = A or C; N = any; V = any but T; H = any but G.

Reaction	Primer name	Primer sequence	Target	Restriction enzyme digest	Target product size	Ref.
Sandfly species identification	SUN_PA_FW	5'-TCG AAT CTA TGG GTG GTG G-3'	Phlebotomine sandfly-specific 18S region	HAE III	241bp and 131bp fragments for *P. argentipes*	[620]
	SUN_PA_REV	5'-CAC AAT CCC AAC CAC GAA G-3'				
Leishmania infection status	OL1854	5'-TGA TAC CAC TTA TCG CAC TT-3'	ITS1 region specific to *Leishmania sp.*	HAE III	220bp and 130bp fragments for *L. major*	[619]
	OL1853	5'-CTG GAT CAT TTT CCG ATG-3'				
Earth Microbiome Project 16S (not including index sequences)	515F	5'-GTG YCA GCM GCC GCG GTA A-3'	V4 region of the 16S gene, all bacteria	None	~300-350bp	[621, 622]
	806R	5'-GGA CTA CNV GGG TWT CTA AT-3'				
Custom 16S primers (not including index, link or pad sequences)	V4F	5'-TAT GGT AAT TGT GTG CCA GCM GCC GCG GTA A-3'	V4 region of the 16S gene, all bacteria	None	~300bp	[382]
	V4R	5'-AGT CAG TCA GCC GGA CTA CHV GGG TWT CTA AT-3'				

5.4.5 16S library preparation

The subset of samples and their respective controls were divided blindly by sample number into three groups, with a negative PCR control and mock community sample in each. A separate library was prepared and sequenced of each group; groups two and three contained two samples also present in group one, in order to identify any inter-run bias. The approximately 300 bp-long v4 region of the 16S gene was amplified using primers recommended by the Earth Microbiome Project (Table 5-2) [565, 621].

To prepare the three libraries, the Illumina MiSeq V2 protocol for library preparation was followed. In a 25 µl reaction, 2.5 µl of each sample was amplified using 515F and 806R primers (Table 5-2) with overhang sequences on each primer to allow the ligation of barcode sequences (forward: 5'-TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG-3'; reverse: 5'-GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G-3'), with primers at a final concentration of 2 nM, and NEB Q5 High Fidelity 2X Master Mix (#M0492S). PCR cycles were as follows: 98°C for three minutes, followed by 35 cycles of 98°C for 30 seconds, 55°C for 30 seconds and 72°C for 30 seconds, with a final extension step of 72°C for five minutes. PCR products were cleaned using a homemade solid phase reversible immobilisation (SPRI) beads mixture, made as described by Rohland and Reich [623], at a ratio of 0.8:1 SPRI beads per sample, to remove any leftover primers. DNA captured on the magnetic SPRI beads was washed twice with freshly-prepared 80% ethanol, air dried to remove residual ethanol, and eluted into 50 µl of molecular biology-grade water. A second PCR was performed on the cleaned, amplified samples, to attach unique dual indices to each sample. The Illumina Nextera XT Index primers were used, with a unique combination of forward and reverse barcodes used for each sample. A 50 µl PCR was performed containing: 25 µl of NEB Q5 High Fidelity 2X taq, 5 µl each of the unique Nextera XT index primer combinations, 5 µl of cleaned DNA and 10 µl of molecular biology-grade water. PCR conditions were the same as for the initial 16S amplification, but with eight cycles instead of 35. Post-PCR products were cleaned with homemade SPRI beads as above, with a ratio of 1.12:1 beads per sample, and eluted into 25 µl.

In order to assess the quality of the samples, a random selection of 12 were run on the Agilent TapeStation with High Sensitivity D1000 reagents (#5067-5584, #5067-5603, #5067-5587), producing a graph of DNA fragment size for each sample. If a peak was observed at 40-100 bp due to latent primers or primer-dimers the SPRI bead clean step was repeated for all samples to select for the target DNA. All samples were then quantified using the Invitrogen Qubit 3.0 fluorometer with Invitrogen Qubit dsDNA HS Assay kit (#Q32851). In order to create pooled libraries with equal amounts of each sample, samples were diluted if necessary then pooled to the same concentration. The pools were then re-assessed on the TapeStation and re-cleaned with the SPRI bead mixture to remove pooled primers if necessary. The clean pools were re-assessed with the Qubit to determine the final pooled concentration.

Each pool was diluted to 4 nM before denaturing with 0.2 N NaOH in an equal volume and diluting further to produce 600 µl of 12.5 pM sequencing library. A sequencing library of PhiX (Illumina, #FC-110-3001), viral DNA with sequencing adapters, was denatured and diluted in the same manner. This PhiX library was 'spiked in' – mixed with the sample library – to make up 20% of the final library, increasing the diversity necessary for efficient SBS. All three libraries were sequenced in separate runs on the MiSeq at the University of Salford, using the MiSeq Reagent Kit V2 with 300 cycles (#MS-102-2002).

5.4.6 Custom primers library production

An additional run containing all samples and controls in a single library was performed, made possible by the use of custom dual-indexing primers developed by

Kozich *et al.* (Table 5-2) [382, 624]. The primers and workflow for library production were kindly provided by Dr Rachael Antwis at the University of Salford.

DNA samples and controls were separated randomly based on sample number into three PCR plates, with a mock community and negative PCR control in each set. Each plate was amplified using the custom 16S primers consisting of an Illumina adaptor, an 8 nucleotide (nt) index sequence, a 10 nt pad sequence, a 2 nt linker and the 16S V4 primer (Table 5-2) [382, 624]. PCR mixtures consisted of: 7.5 µl NEB Q5 Hot Start High Fidelity 2x Master Mix; 2.25 µl of each 2 µM dual-indexed primer pair; 1.5 µl of molecular-grade water and 1.5 µl of DNA sample. Each PCR reaction plate was run in duplicate simultaneously with the same master mix on different thermocyclers; duplicates were then pooled to reduce the effect of random amplification differences. Thermo-cycling conditions were 98°C for 30 seconds, followed by 35 cycles of 98°C for 10 seconds, 55°C for 10 seconds and 72°C for 30 seconds, and finished with a final extension step of 72°C for 2 minutes. Pooled duplicate plates were cleaned using the homemade SPRI bead mixture as described above with a ratio of 1.5:1 of beads to sample and eluted into 37 µl of molecular biology-grade water. Samples were then assessed by TapeStation and Qubit and pooled as above, with all three plates pooled together. The pooled library was denatured, diluted, spiked with 20% PhiX and sequenced as above.

5.4.7 Quality control and contamination removal

Demultiplexed samples from the three runs with Nextera primers were pooled together and analysed as a single dataset (hereafter referred to as the Nextera run); the sequencing run using custom primers (hereafter referred to as the Custom run)

was analysed separately. Both datasets were analysed using the same pipeline, produced with guidance from Dr Antwis. Reads were demultiplexed on the MiSeq and subsequent analysis steps were performed in R version 3.5.1 [435] using RStudio [436].

5.4.7.1 Read quality control with DADA2

Reads were trimmed and filtered using the DADA2 package (v 1.14.0) [625]. The default parameters of the `filterAndTrim` command were used to remove low-quality reads: forward reads were truncated at base 220 and reverse reads at base 200 (`truncLen=c(220,200)`), based on visual examination of quality plots. Any reads with unknown bases were discarded (`maxN=0`), and the threshold of maximum expected errors per read was set to two for both forward and reverse reads (`maxEE=c(2,2)`) corresponding to a 63% chance of a base being incorrect [626]. Any reads left were truncated at quality score two (`truncQ=2`), and any reads identified as PhiX were removed (`rm.phix=TRUE`).

The `learnErrors` function of DADA2 was used to estimate the dataset errors for forward and reverse reads separately, using a parametric model trained on the dataset to alternatingly estimate error rates and infer sample composition until they converged on a jointly consistent solution for each read [625]. This model was then applied to the filtered and trimmed sequence data for error correction, before a denoising step to return unique ASVs and their frequencies for each sample. Paired reads were merged, and any unpaired reads were discarded. Chimeric sequences were identified using `removeBimeraDenovo` with the “consensus” method,

whereby each ASV is independently checked; an ASV was removed if it could be exactly reconstructed by combining segments from two or more “parent” sequences.

Finally, taxonomy was assigned to ASVs using the `assignTaxonomy` function: this applies a naïve Bayesian classifier method to compare 8 nt segments of each ASV to a database of known sequences, or a “training set”, and scores ASVs based on their degree of likeness to known genera, assigning each ASV to the sequence with the highest score [627]. If no genus has a close match the taxonomic rank is reduced to family, and so on until a match is reached. Bootstrapping of sequences is performed, whereby subsets of the 8 nt segments are taken and used to calculate the joint probability that the assignment is correct; the default of 50 bootstrap trials was used. The training set used was version 132 of the SILVA ribosomal RNA gene database [628, 629].

5.4.7.2 Contamination control using phyloseq and decontam

The ASVs with assigned taxa were handed off to the phyloseq package (v1.30.0), which builds an object containing the sample metadata, ASV frequencies in each sample and the assigned taxon of each ASV [630]. Each sample was identified by the district, village and house (if applicable) of origin, as well as the DNA extraction run. The Nextera run metadata also included the sequencing run (1-3) of each sample, and whether the sample was included in more than one sequencing run for inter-run comparisons. An additional column was added in R to indicate whether each sample was a true sample, or a negative (DNA extraction or PCR) or positive control.

This phyloseq object was analysed for contaminating taxa from external sources (rather than cross-contamination between samples) using the prevalence-based strategy in the decontam package (v1.6.0) [631]. This strategy relies on the assumption that, as any identified taxa are only a sample of the total present in the sample, contaminating taxa are more likely to be present in negative samples than true samples. For each ASV a presence-absence table in negatives and true samples is constructed; the probability of the distribution being random is calculated using a modified Chi-squared analysis, and each ASV is given a score *P* ranging from 0 to 1. Small *P* scores indicate an ASV is likely to be a contaminant, but the final classification is based on a user-defined threshold (*Pi*), whereby if $P < Pi$ the ASV is labelled as a contaminant. For low-biomass samples where contaminants are likely to outnumber true ASVs, as in this study, the function `isNotContaminant` replaces the standard prevalence score *P* with *1-P* and positively identifies contaminants rather than true ASVs, in order to classify non-contaminants more conservatively. For this dataset both the `isContaminant` and `isNotContaminant` functions were used with thresholds of 0.5 to compare the two strategies; the final dataset was produced using `isNotContaminant` as is recommended for low-biomass datasets [631].

A new phyloseq object was created with the identified contaminants removed; reads identified as mitochondria or chloroplasts were also removed, as well as any ASVs with fewer than 10 reads, and finally all controls were removed to leave a final dataset of true sample taxa. A separate phyloseq object was created containing only the mock community samples, which were analysed by examining ASV counts and

comparing their relative proportions to the expected proportions of species declared in the ZymoBIOMICS literature.

5.4.8 Bioinformatic analyses

Samples were analysed using the packages microbiome (v1.4.2) [632] and vegan (v2.5.3) [633], an ecological community analysis toolkit; data was visualised using ggplot2 (v 3.1.0) [526]. Alpha diversity was calculated using the Shannon diversity index for each district, village and, in the case of the Nextera dataset, sequencing run. Non-metric multi-dimensional scaling (NMDS) plots with Bray-Curtis dissimilarity distances were plotted to visualise overall beta-diversity. Permutational multivariate analysis of variance (PERMANOVA) was calculated using the `adonis` function in the vegan package, to quantify the impacts and significance of origin, extraction run and sequencing run on the microbiome composition. An R^2 score is given between 0-1 as a measure of how much variation can be attributed to each term of the model. Scatter plots and line graphs to visualise the effects of the quality control pipeline were created in Excel and R. Linear regression analyses were conducted to assess the correlation between raw reads and the number of identified genera in each sample, with R^2 values adjusted for the degrees of freedom to assess correlation strength. Raw read data was not normally distributed and was therefore square-root transformed for linear regression analyses to produce a normally distributed dataset.

5.5 Results

5.5.1 Read quality and decontamination

More total reads were present in the Nextera dataset, which used Nextera XT adapters in three separate sequencing runs, than the dataset using Custom primers to sequence in a single run (13,509,999 and 9,504,258 reads respectively). The average number of raw reads for each sample was only slightly higher in the Nextera dataset (48,423) than the Custom dataset (36,138). A left-tail skewed distribution of reads was evident in the combined Nextera runs, with a more normal distribution in the Custom run (Figure 5-3). There was significant variation in the number of raw

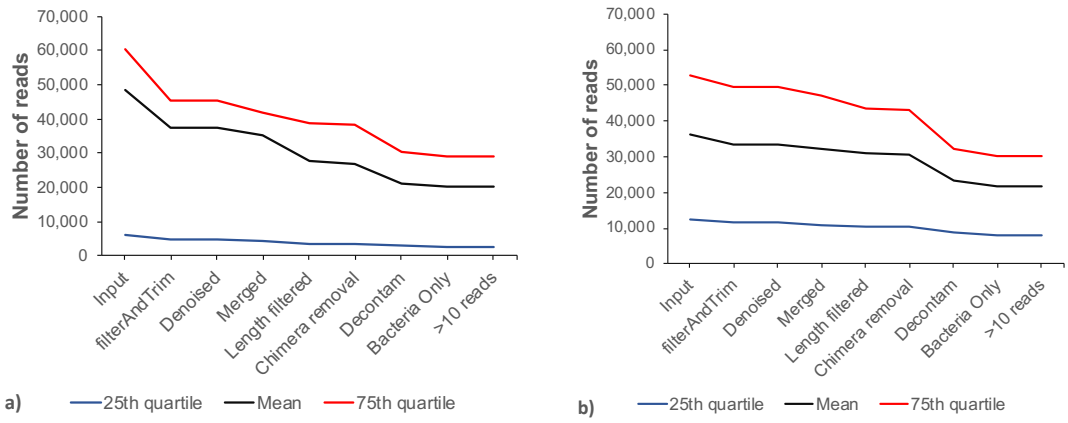


Figure 5-3 - The mean average and quartiles of reads per sample at each stage

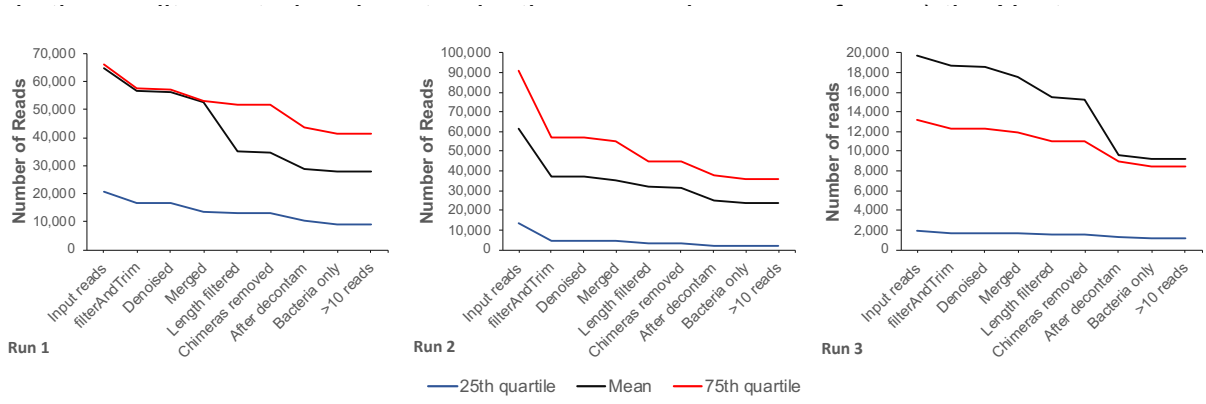


Figure 5-4 - The mean average and quartiles of reads per sample at each stage in the quality control and contamination removal process from each of the three sequencing runs with Nextera primers, showing the wide variation in read numbers and distributions between the runs. Note: the different y-axis values.

reads and the distribution of read numbers between samples in the three sequencing runs making up the Nextera dataset (Figure 5-4). The three sequencing runs had a mean average of 64,710, 61,222 and 19,708 raw reads per sample respectively; runs one and three had highly left-tailed skewed read variations, with a small number of very large samples having a considerable impact on the means as evidenced by the 75th quartile representing less than the mean in run three (Figure 5-4). The library sizes of negative controls were similar to the library sizes of most true samples in both the Nextera and Custom datasets (Figure 5-5).

The stage at which the most reads were lost varied by sequencing run. In the Custom sample set most reads were of sufficient quality to pass through the initial quality control steps, with the largest loss of reads at the decontamination stage (Figure 5-3b). Equal drops in read numbers were seen after the `filterAndTrim` function and the decontamination step in the Nextera dataset (Figure 5-3a). However, when broken down into the three separate sequencing runs only run two shows a large drop after the `filterAndTrim` step (Figure 5-4). Run one has a small drop after filtering and trimming, with the main reduction occurring when any reads outside of the expected size range were discarded. In run three, most of the read loss was due to the decontamination step.

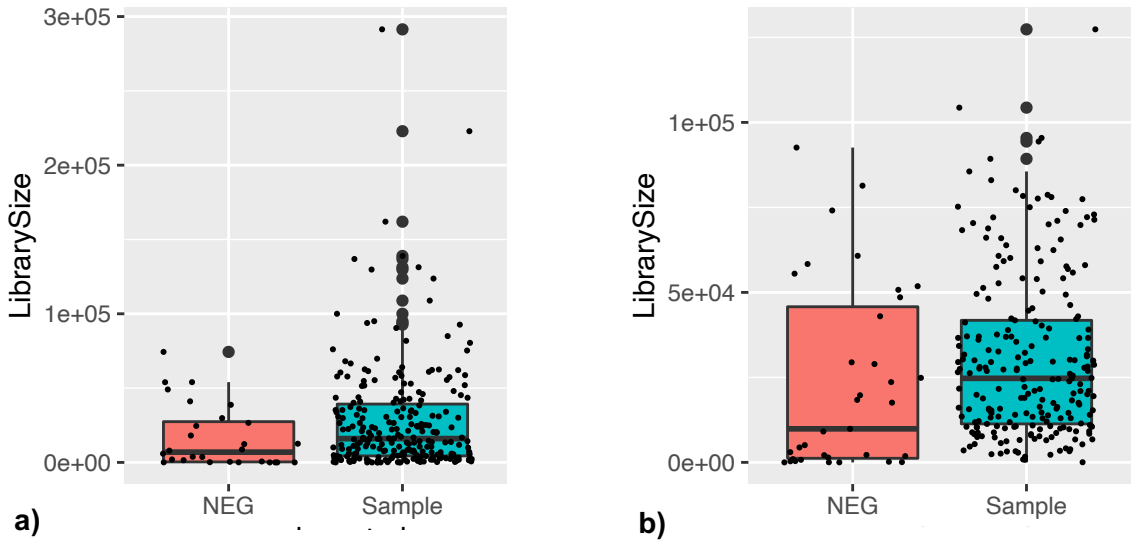


Figure 5-5 – Boxplots of the library sizes of negative controls and true samples, after quality control steps and before being run through the decontam program, for a) the Nextera sample set and b) the Custom sample set.

In both the Custom and Nextera datasets there was a weak but significant correlation between the number of raw reads and the number of genera identified in each sample (Custom $R^2 = 0.313$, Nextera $R^2 = 0.227$; $p < 0.001$) (Figure 5-6a and b). However, the correlation between raw reads and identified genera for the individual Nextera sequencing runs was variable (Figure 5-6c, d, e), although all three runs showed a weak but significant correlation (run one: $R^2 = 0.040$, $p = 0.04$; run two: $R^2 = 0.357$, $p < 0.001$; run three: $R^2 = 0.274$, $p < 0.001$). M449 was removed from run one for the linear regression analysis as an outlier, with 1,827,160 input reads making it over 10-fold more abundant than the next highest sample in the run.

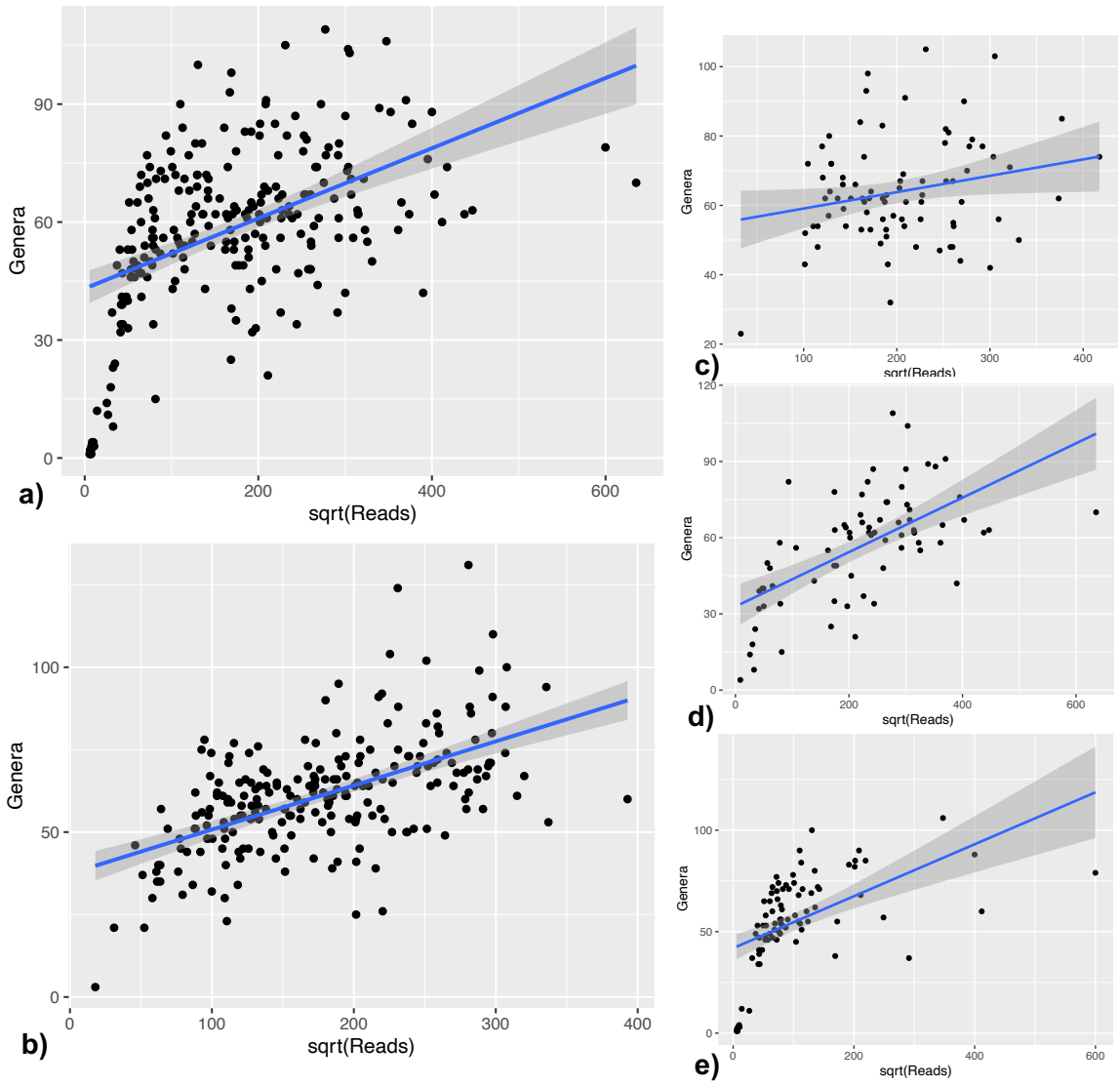


Figure 5-6 - Linear regression of transformed raw reads to identified genera in each sample of a) the Nextera dataset (R^2 = 0.227, p = <0.001) and b) the Custom dataset (R^2 = 0.313, p <0.001), suggesting a weak positive correlation between the input reads and the number of identified genera in a sample. The Nextera sample set was split by sequencing run: c) run one (R^2 = 0.040, p = 0.04); d) run two (R^2 = 0.357, p = <0.001); and e) run three (R^2 = 0.274, p = <0.001) showed a weak correlation between raw reads and identified genera. Sample M449 was removed from run one of the Nextera dataset as an outlier.

The two different decontam functions produced markedly different results in both datasets (Figure 5-7). The less stringent `isContaminant` function identified 6.5% of ASVs in the Custom dataset and 2.9% of ASVs in the Nextera dataset as contaminants. In contrast, the `isNotContaminant` function removed 80.7% of Nextera reads and 82.1% of Custom reads. With both functions, proportionally more reads were identified as contaminants in the Custom dataset than the Nextera dataset.

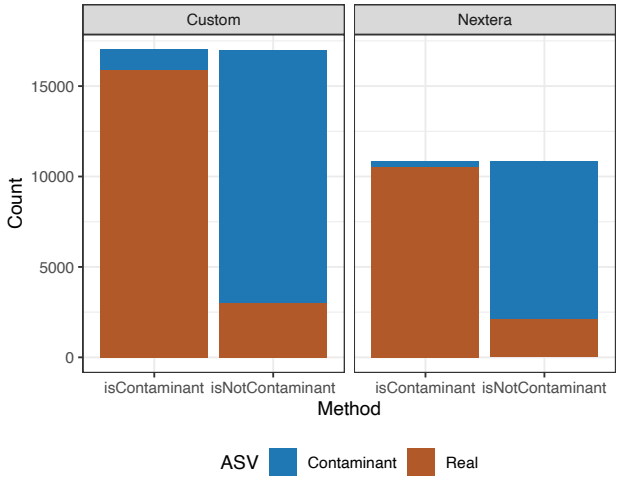


Figure 5-7 - ASVs identified as contaminants or real taxa present in sample by the decontam package, using the `isContaminant` and `isNotContaminant` functions, for the Custom and Nextera datasets.

5.5.2 Comparison of samples between sequencing runs

The mock communities included in each sequencing run showed variation both between the Custom and Nextera methods and between Nextera sequencing runs (Figure 5-8). Both the Custom and Nextera mock communities have under-represented *Listeria monocytogenes* and *Enterococcus faecalis*, with *Lactobacillus fermentum* over-represented in most samples. Run one of the Nextera dataset did not amplify any *Salmonella enterica* reads, despite this species being well represented in all other mock community samples. The three mock community replicates in the Custom dataset are largely similar, with very few reads identified outside of the expected eight species. The mock communities sequenced in separate runs in the Nextera dataset, however, are markedly different; in run one,

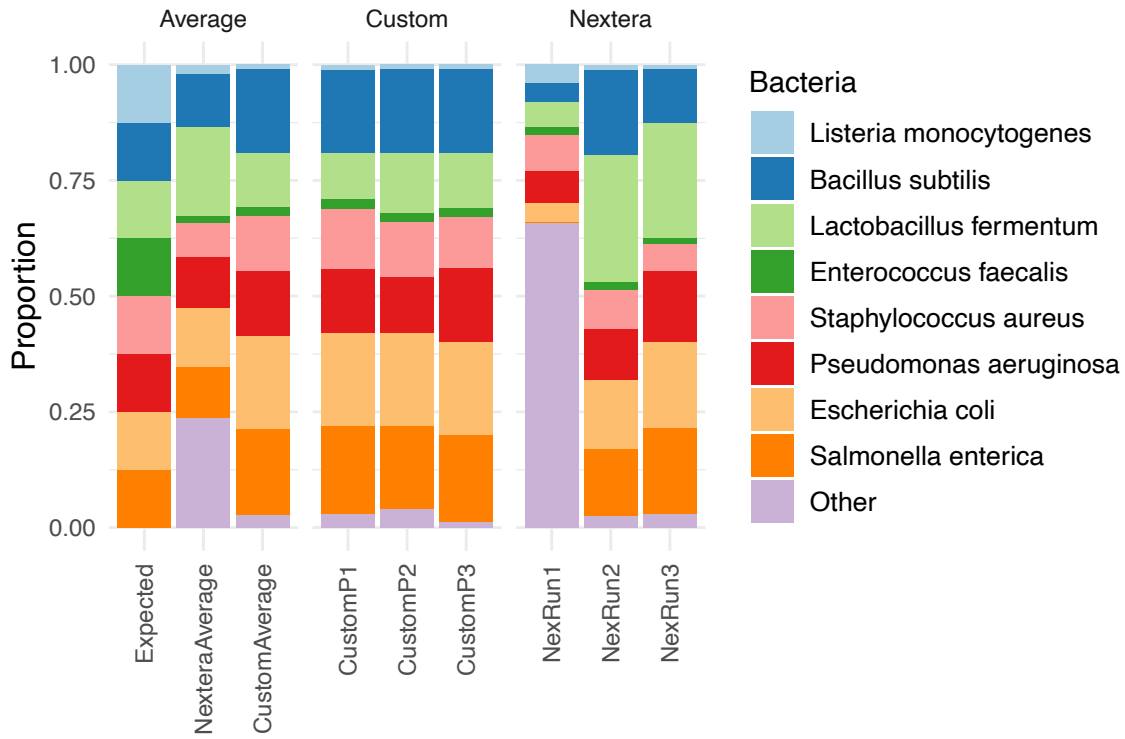


Figure 5-8 - Identified taxa from a Mock Community sequenced in triplicate in the Custom run (CustomP1 – CustomP3) and included in each sequencing run of the Nextera sample set. Expected proportions and species were from the ZymoBIOMICS supplementary information. Averages are the mean proportions of the mock communities in each run.

two thirds of reads came from genera not present in the mock community. Runs two and three were more similar, with fewer *Salmonella* and more *Bacillus* reads in run two, but low proportions of non-target genera in both.

Variation of observed microbiota in samples sequenced in multiple runs with the Nextera method ranged from minimal to considerable (Figure 5-9). Most samples sequenced in different runs clustered closely together, indicating that the observed communities in these samples were comparable between sequencing runs. However, three out of the seven samples sequenced in multiple runs showed a large difference in their observed communities between sequencing runs; the sample from

house five in Chandauli, Samastipur (SamChaH5) appeared to have a different community composition in each sequencing run. Notably, none of the samples which were sequenced in all three runs clustered together in all three.

When quantified, the effect of sequencing run has less of an effect on the Nextera samples community composition than the district or DNA extraction run of origin. PERMANOVA analysis of the Nextera samples found district (R^2 = 0.05), extraction run (R^2 = 0.20) and sequencing run (R^2 = 0.04) to all have significant impacts on the microbiome ($p < 0.001$). In the Custom sample set, district and extraction run of origin were also significant factors ($p < 0.001$), with R^2 values of 0.04 and 0.13 respectively.

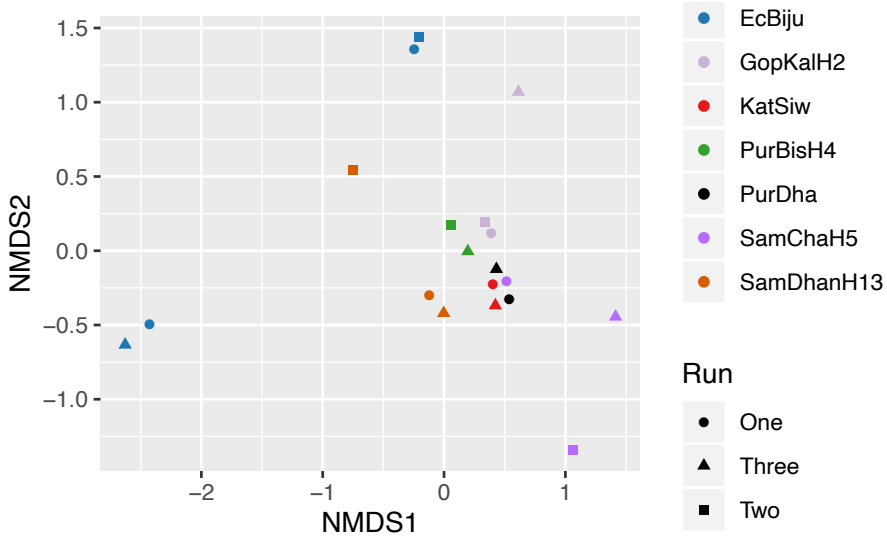


Figure 5-9 - NMDS plot of microbiome samples from sand flies sequenced with the Nextera method and included in more than one sequencing run, identified by their location. Each point represents a single sample; the colour of each point represents the sample origin in the format DistrictVillageHouse, the shape represents the sequencing run of the sample. The proximity of points to one another illustrates their similarity.

5.6 Discussion

5.6.1 Best practice methods still produce an imperfect sequencing library

A microbiome dataset is only as good as the methodology used to produce it. Despite careful consideration of the best practices for low-biomass microbiome studies there were still flaws evident in the methods used to produce the sand fly microbiome data described in Chapter 6. Two out of eight of the known bacterial species in the mock community included in both the Custom and Nextera runs were consistently under-represented in the datasets, representing a bias either in the DNA extraction strategy or the amplification of the 16S V4 region of these bacteria. *Enterococcus* and *Listeria* are both gram-positive commensal bacteria, and the former a known microbial partner of sand flies [327, 634]; their under-representation in the dataset suggests other similarly host-associated bacteria may also be under-identified in the sand fly samples. The reduced representation of important commensal bacteria in a dataset produced using the methods recommended for studying microbiomes illustrates the need for continued methodology refinement in this area.

5.6.2 Nextera primers vs Custom primers: key differences during quality control

Despite consisting of the same samples and the same region of their 16S gene, there was significant variation in outputs between the Nextera and the Custom primer pipelines. The Nextera sequencing pipeline with three separate runs should have had the capacity to produce three times as many raw reads as the Custom

pipeline, but only produced 1.4 times as many, either due to human error in the library dilution or an artifact of the primers used.

The proportion of reads lost at each stage of the quality control pipeline varied between the Custom and Nextera libraries (Figure 5-3). The Nextera library lost a greater proportion of reads than the Custom library during the initial `filterAndTrim` step, which removes reads based on expected errors and truncates reads based on quality. This seems to indicate that the Nextera raw reads were largely of worse quality than the Custom reads. A lower quality raw read library could be due to errors in the library production; for example, inadequate removal of primers or PCR reagents, non-optimal denaturation of dsDNA before loading on the sequencer, or overclustering the sequencing flowcell. The Nextera library production involved a second PCR step to ligate adapters to amplified 16S reads; as Qubit does not distinguish between DNA with or without adapters when calculating the concentration of a sample, the SPRI bead clean-up steps capture all DNA with or without adapters, and DNA fragments without adapters will not successfully attach to the MiSeq flowcell for sequencing, this step could potentially introduce a drop in read quality if all amplified reads do not have an adapter ligated.

5.6.3 The library size problem

The raw read outputs in both the Nextera and Custom datasets were left-tailed skewed, with a large number of samples in each set containing few reads and several samples with far more reads than the mean, producing a non-normal distribution of read numbers throughout the dataset that did not correlate with whether samples were negative controls or true samples (Figure 5-3 and Figure 5-5).

As the variation both the Custom and Nextera datasets was so high despite an equimolar sample pooling step, seemingly the methods used to quantify and pool samples in both pipelines were inadequate. Errors can be introduced during quantification or pooling easily by human error, through inaccurate pipetting of the minute quantities used; additionally, Qubit quantification is not the gold standard for sequencing low DNA concentrations, with the more expensive qPCR giving better results [611]. Whether the variation in sequencing depth between samples was due to human error or inaccurate quantification is difficult to ascertain from this data.

In theory, the more reads in a microbiome sample the more ASVs will be identified until the maximum diversity has been reached, representing the true diversity of the sample. Therefore, if microbiome samples are sequenced with enough depth and the true diversity of the sample has been discovered, the number of raw reads in a sample should not correlate with the number of identified genera in that sample. This is not the case with either the Custom or Nextera dataset: both show a very weak but positive correlation between raw reads and identified genera (Figure 5-6). While this suggests that the sequencing depth was not adequate, it could also be due to introduced artefacts or contaminants which were not adequately removed.

Large differences in library size between samples is a common issue in microbiome data, and several approaches have been suggested and subsequently refuted as being statistically inadmissible. For example, samples are commonly ‘rarefied’ by setting a threshold for the lowest library size and re-sampling all samples above the threshold to produce a dataset with equally-sized libraries [635]. This approach has been challenged as unnecessarily removing rare ASVs and not accurately reflecting

the microbial communities being subsampled [636]. The complexity of microbiome data, and the lack of statistical knowledge or consistency of application, makes comparing microbiome datasets without access to raw data perilous.

5.6.4 One dataset, multiple sequencing runs

At the time this experiment was designed, it was uncertain whether sequencing nearly 300 samples in one run would give sufficient sequencing depth for accurate microbiome analyses; therefore the Nextera XT dual-index system for 96 samples was chosen with samples split into three sequencing runs. During analysis it became clear that there were fundamental differences between the three runs that would make samples incomparable between them. Run three samples had an average number of reads three times lower than that of the other two runs, immediately countering the assumption that samples would be sequenced at comparable depths and therefore have comparable diversity. The three runs were filtered differently to one another: run one lost the largest proportion of reads when reads outside of the expected length were removed, indicating errors with the sequencing itself or with the library had a strong effect. Additionally, two-thirds of the mock community sequenced in run one consisted of genera not present in the sample; this suggests barcode switching, either during library preparation or sequencing, caused improper demultiplexing of samples. Run two lost most reads during the `filterAndTrim` step, suggesting low quality sequencing. Finally, run three lost most reads during decontamination. The differences in data between the three runs, illustrated best by comparing samples included in more than one sequencing run (Figure 5-9), were significant for some samples and minimal for others; the unpredictability of the differences between runs made correction for run effect impossible, leading to the

adopting of the Custom primer set in order to sequence all samples in a single run and risk reducing depth in order to eliminate inter-run variability.

5.6.5 Contamination and decontamination

The library sizes of negative controls in this dataset, and their equivalence to true sample library sizes, was further evidence of the necessity of an effective decontamination strategy in microbiome analysis (Figure 5-5). Many microbiome study designs do not include negative controls, either in the sequencing run or in the published dataset, or otherwise simply remove any ASVs present in negative samples from all true samples indiscriminately [568]. This strategy would remove true microbiota from our dataset in particular: many common contaminants of reagents are soil-associated microbes [384] therefore potentially overlapping with the sand fly microbiome, as sand flies spend part of their life cycle in soil [313].

Due to the large disparity of library sizes between samples it is not possible to ascertain whether negative samples with more raw reads truly contained more contaminating DNA, and therefore whether certain DNA extraction batches were more contaminated than others. The PERMANOVA analysis of both the Custom and Nextera datasets showed that DNA extraction batch had a greater effect on the microbiome of samples than their district of origin. This suggests that despite the large number of contaminating ASVs identified and removed by the decontam package, there are still contaminants present in the dataset. The decontam package is still under development and has recently added an option to treat samples by DNA extraction batch when at least one negative is present per batch. This would be the best strategy for a dataset of low-biomass samples like this study; however, as there

are uneven numbers of samples per extraction run included in the subset that was sequenced (ranging from 1 to 18), it was decided that batch treatment would not be appropriate. The decontam package had not been released when this study was designed so this issue was not foreseen, illustrating the speed at which tools in the microbiome field are evolving.

The dataset of sand fly microbiota produced by these methods, while as accurate as possible with current technology and software, is still not robust enough to rely entirely on quantitative and statistical analyses. It is worth noting that not one statistical analysis performed on this data gave any result except a low statistical score with a significant p value. Large datasets are prone to falsely attributing statistical significance to non-scientifically significant effects [637]; the field of microbiome analysis is still adapting to the increasing size of the datasets produced by current sequencing technology [638]. For example, the package DivNet in R was released in 2020 to give more accurate richness estimates with calculated errors based on the fact that microbiome data are a subsample, rather than a census, of the studied population [639]; however, the software is still being developed and is not ready to run large datasets. Even with contaminants removed bioinformatically, it is prudent to check the genera retained in the dataset against known or suspected contaminating genera by eye. Visualisation of the data by bar charts and box plots remains the safest way to draw conclusions from a microbiome dataset, with holistic integration of sequencing data and ecological, biological and microbiological knowledge.

Chapter 6 The core and peripheral microbiome of individual *Phlebotomus argentipes* at fine geospatial scales across Bihar, India

6.1 Abstract

Sand flies transmit *Leishmania* parasites across the world, causing the fatal disease VL and the highly disfiguring cutaneous leishmaniasis. India is one of 10 countries where 95% of VL cases occur, with 80% of Indian cases occurring in the northern Indian state Bihar. As the native microbiota of insect vectors are known to play a part in vectorial capacity, and may be exploited in vector control strategies, we sought to improve our knowledge of the microbiome of the sand fly *Phlebotomus argentipes*, the only VL vector in Bihar. The current knowledge of *P. argentipes* microbiota is restricted to the characterisation of pooled samples by aerobic culture. In this study we developed a pipeline to characterise the bacterial microbiota of independent *P. argentipes* samples for the first time, using 16S sequencing, whilst mitigating the challenges of low biomass and “kit-ome” contamination. Individual *P. argentipes* harbour a limited core microbiome of seven genera including *Staphylococcus*, *Acinetobacter* and *Corynebacterium*, and a highly variable peripheral microbiome. A common “dysbiotic” state was discovered in *P. argentipes* sand flies, where single bacterial species including *Orientia* dominate individual microbiomes, with implications for microbiota-based VL control strategies. The potential vector competency of sand flies for *Orientia*, which causes scrub typhus, is discussed. Targets for paratransgenesis-based vector control strategies are identified, including *Bacillus, Brevibacterium* and *Corynebacterium*, corroborating previous research into microbiota-mediated sand fly control strategies. These results form an important

baseline of individual sand fly microbiome study for future work and introduce the novel concept of sand flies harbouring an important pathogen for an emerging disease in the region, *Orientia.*

6.2 Introduction

6.2.1 Visceral leishmaniasis

VL is an NTD caused by the kinetoplast parasite *Leishmania,* which is transmitted by the bite of an infected female phlebotomine sand fly. VL infections occur in approximately 88 countries, with over 95% of cases in 10 countries concentrated within three geographic foci: East Africa (Ethiopia, Kenya, Somalia, Sudan and South Sudan), Brazil, and the Indian subcontinent (Bangladesh, India, Nepal) (Figure 6-1) [640, 641]. See Section 1.6 for an in-depth review of the leishmaniases and their vectors. In Bihar state, India, which accounts for 80% of Indian cases, the causative agent *Leishmania donovani* is transmitted by only one sand fly species, *Phlebotomus argentipes* [640, 642].

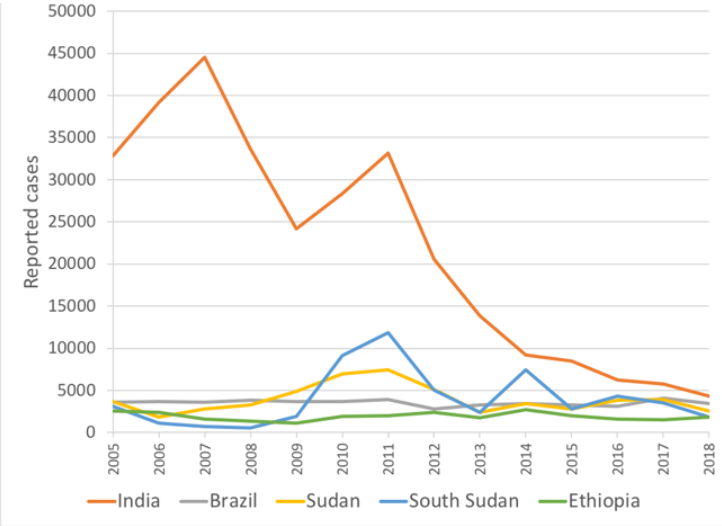


Figure 6-1 - Visceral leishmaniasis cases each year from 2005-2018, in the five countries with the highest infection rates. Data from the World Health Organization [4].

In 2005 the governments of India, Bangladesh and Nepal committed to a regional partnership to reduce the incidence of VL cases to one case per 10,000 people per primary health centre (PHC) region through the implementation of improved diagnosis and treatment, in addition to vector control [300, 643]. Vector control is delivered through indoor residual spraying, traditionally with DDT and more recently with alpha-cypermethrin [300]. The incidence of VL was reduced to 4,360 reported cases in 2018 [4], although actual VL incidence is thought

to be between 4-8x higher than official estimates [644, 645], giving a VL incidence of 0.03-0.25 per 10,000 people. However, several key districts in India still contain PHCs with more than 1 case per 10,000 people. In 2016 Bihar had 4,496 cases out of the 6,249 cases in India, giving it an incidence rate 10 times higher than the national average, and 15% of PHCs in Bihar still had an incidence rate above the 1 in 10,000 target [300]. Additionally, for every reported VL case in Bihar an estimated three cases co unreported [646].

6.2.2 Microbiota of vectors as VBD control agents

The development of novel methods of phlebotomine vector control has been identified as a necessary step to reach the VL elimination target. One such option is the use of microbial agents to successfully reduce the vectorial capacity of the host population [166] – see Section 1.4 for a summary of microbiota-mediated vector control strategies. Briefly, insect-associated microbes with natural or genetically modified traits to reduce vector capacity can be introduced to a target vector, or the target's naturally associated mutualistic symbionts can be disrupted. The use of microbiota to disrupt vector-borne pathogen transmission requires a deep understanding of the vector's microbiome, the diversity of species and their impact on the host.

Several insect vectors have associations with bacteria that impact their vector capacity: for example, *Sodalis glossinidius*, a commensal endosymbiont of the tsetse fly (genus *Glossina*), has been implicated in increasing tsetse susceptibility to the kinetoplast parasite *Trypanosoma* [57, 647]. *Sodalis* has also been successfully used as a paratransgenic target for reducing tsetse vector competence [266], and is

4004 an example of the multiple potential uses of symbiotic bacteria in vector disease
4005 control. The microbial communities present in vectors can also influence vector
4006 capacity less directly, through priming the immune response of the host to reduce
4007 infection and competing with the parasite for resources [648]. The gut microbiota of
4008 *L. longipalpis* is essential for the survival of *Leishmania infantum* [311].
4009
4010 Potential microbial agents for vector control must have either a measurable ability to
4011 reduce vectorial capacity or be amenable to manipulation to produce such a
4012 phenotypic effect. Logistically, such agents must be culturable, deliverable to wild
4013 hosts, be easily transmissible between individuals and be at least as fit as the natural
4014 host microbiota in order to outcompete the resident microflora [153]. It is clear,
4015 therefore, that further research describing the natural *P. argentipes* microbiota at
4016 both population and individual levels is essential for the identification of potential
4017 agents of such strategies. Most sand flies do not travel more than a few hundred
4018 metres during their lives [307], and VL incidence has fine spatial-temporal variation
4019 [649]. Thus, it can be hypothesised that given a limited range, the host microbiota
4020 will not fluctuate due to the shared environment and feeding targets. Furthermore,
4021 microbiota may vary more between sand flies isolated from wider geographical
4022 regions or if sand flies have a wider-than expected range. Elucidating spatial
4023 variation in microbiota between individual sand flies could reveal novel aspects of
4024 sand fly ecology that could be used to control vector behaviour.
4025
4026 There exist several studies investigating phlebotomine sandfly microbiota using
4027 culture-dependent studies: Maleki-Ravesan [650] identified *Enterobacteriaceae* as
4028 key components of the *P. papatasi* microbiota using biochemical and Sanger-

sequencing techniques. Hillesland *et al.* [322] identified two potential targets for paratransgenic intervention in *P. argentipes* (isolated from across Bihar): *Bacillus megaterium* and *Brevibacterium linens*, however this study used microbiological techniques linked to 16S rRNA sequencing to identify microbiota, limiting the study to aerobic bacteria, and only those that were amenable to culture-based methods.

Culture-independent (DNA sequencing based) studies are a powerful approach to understanding insect microbiota linked to vector competency as such studies mitigate such culture-related bias: Sant'Anna *et al* [651] used a pyrosequencing approach linked to 16S clone libraries to identify 38 bacterial phylotypes in Brazilian *L. longipalpis*, and Li [652] identified 29 phylotypes in *P. chinensis* - the key vector of VL in China. Kelly [311] sequenced laboratory-reared *L. longipalpis* microbiomes, identifying that blood feeding causes a decrease in midgut microbial diversity that was temporary in uninfected sand flies, but remained diminished in *L. infantum*-infected sand flies. It is important to note, however, that sequence-based microbiome analyses can be hindered through obfuscation of real microbial signatures due to contaminants introduced and amplified from the immediate, laboratory environment, or from kits used as part of the laboratory preparation of materials, particularly where the starting biomass in the sample is low [384]. In order to alleviate this, the majority of small insect microbiota studies use pools of individuals [651, 653]; this approach gives a higher starting biomass and better yield of sample DNA in proportion to contaminant DNA, but erases any variation in the microbiota between individual hosts which may be important to understand for developing microbiota-based control strategies. A methodological pipeline, from DNA extraction to bioinformatic analysis, was developed which allows the examination of individual sandfly microbiota while

addressing the risk of contamination – Chapter 5 describes the process of developing this pipeline.

P. argentipes were sampled from a range of locations at increasing spatial resolution – from district, village, and individual house levels – to provide a better understanding of the *P. argentipes* microbiome and *L. donovani* infection incidence. By sequencing the microbiota from individual sand flies and developing a pipeline for low-biomass microbiome samples, we describe a robust core microbiota of wild *P. argentipes*, simultaneously identifying spatial variation in peripheral microbiota between individuals and potential intervention targets.

6.3 Aims and objectives

The aim of this work was to describe the bacterial communities within individual *P. argentipes* collected at fine spatial scales across the important VL district Bihar, using culture-independent, NGS methods. By doing so, previous work to identify potential paratransgenesis targets can be validated and their prevalence within sand fly populations accurately calculated, as well as identifying novel targets. We aimed to examine the effects of *L. donovani* infection, association with buildings within the same village, villages within the same district, and districts across Bihar on the *P. argentipes* microbiome.

6.4 Methods

The methodology for sand fly collection, DNA extraction and library preparation can be found in Chapter 5; only reads from the single sequencing run using custom primers were used.

6.4.1 Analysis of sequencing data

Data were analysed in R using microbiome (v 1.4.2) [632] and vegan (v 2.5.3) [633], an ecological community analysis toolkit; data was visualised using ggplot2 (v 3.1.0) [526]. Alpha diversity for each village was calculated using the Shannon diversity index. NMDS plots with Bray-Curtis dissimilarity distances were plotted to visualise overall beta-diversity. Permutational multivariate analysis of variance was calculated using the `adonis` function in the vegan package, to quantify the impact of location on microbiome composition; an R^2 score is given between 0-1 as a measure of how much variation can be attributed to each term of the model. Krona plots were produced using psadd (v 0.1.2) [654] for visual examination of microbiota composition and percentage abundances. The core microbiome was calculated using the microbiome package *core* function; core genera were chosen at 95% prevalence and 1% relative abundance. A subset was created of all genera present at ≥1% abundance in the dataset for subsequent visualisation and analysis of biologically relevant microbiota.

6.5 Results

6.5.1 Leishmania prevalence

Out of 792 sand fly DNA samples two tested positive for *Leishmania sp.* (0.3%). As this would not be enough to statistically compare the microbiome of infected and uninfected sand flies, these samples were removed from the rest of the study.

6.5.2 Quality control

The small biomass of individual sand flies yielded low concentrations of DNA (<0.01 ng/µl, data not shown). As a result, 95% of ASVs were identified as contaminants and removed, leaving a final count of 226 genera in 229 fly samples. Figure 6-2 shows that the mock communities sequenced in triplicate were all similar in composition, confirming the lack of bias between amplicon plates. All expected genera from the mock communities were found; however, both *Listeria monocytogenes* and *Enterococcus faecalis* were under-represented.

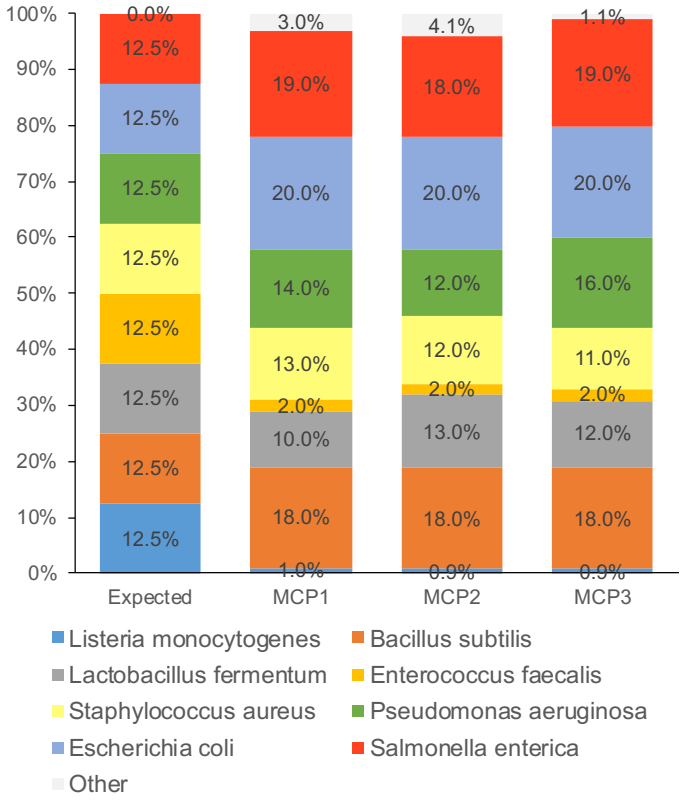


Figure 6-2 – The results of sequencing a Zymo Mock Community standard. A mock community was included in all three of the amplicon plates to ensure the accuracy of each amplification. The expected percentages are as described in the Zymo Mock Community technical documents. All expected species were seen with close to the expected proportions. MCP – Mock community plate.

6.5.3 The *P. argentipes* core microbiome

Sequencing 229 individual sand fly microbiomes and processing them with a stringent decontamination pipeline revealed a mean 61 (range: 3-131) bacterial genera per sample, and a total of 226 genera across all samples. A core microbiome of seven genera was identified, consisting of *Staphylococcus*, *Massilia*, *Noviherbaspirillum*, *Brevundimonas*, *Corynebacterium*, *Acinetobacter* and *Sphingomonas*.

At ≥1% relative abundance 23 genera were described (Figure 6-3). Despite making up a large proportion of the genera present at ≥1% relative abundance, *Orientia* is not identified as one of the core microbiota, indicating an imbalance of distribution. Table 6-1 further describes the core genera in the context of their association with VL vectors.

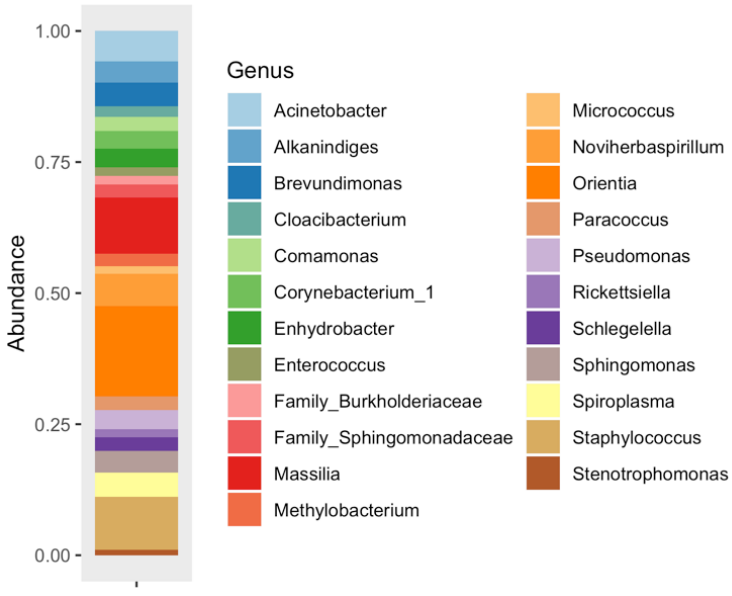


Figure 6-3 - All bacteria sequenced at ≥1% relative abundance in the *P. argentipes* microbiome (229 samples), identified to genera level.

Table 6-1 - The core and dominating genera of *P. argentipes* in this study, with previous observations in VL-transmitting sandfly species. Core genera are defined as any genera at ≥95% prevalence and ≥1% relative abundance in the sample set; dominating genera are those that make up ≥50% of any one sample.

Core genera			**Dominating genera**		
Description (type species)	Observations in VL-transmitting sandflies	# of flies	Description (type species)	Observations in VL-transmitting sandflies	# of flies (≥50%)
Massilia			*Orientia*		
Aerobe, gram-negative [655]	None	226	Obligate intracellular, gram-negative [656]	None	201 (19)
Sphingomonas			*Spiroplasma*		
Aerobe, gram-negative [657]	None	226	Facultative intracellular, wall-less bacteria [402, 658]	*P. chinensis* [652]	168 (3)
Staphylococcus			*Staphylococcus*		
Facultative anaerobe, gram-positive [659]	*P. chinensis* [652]; *P. argentipes* [322, 634, 660]; *L. longipalpis* [323, 661]; *L. cruzi* [651]; *L. evansi* [662]	225	Facultative anaerobe, gram-positive [659]	*P. chinensis* [652]; *P. argentipes* [322, 634, 660]; *L. longipalpis* [323, 661]; *L. cruzi* [651]; *L. evansi* [662]	225 (2)
Acinetobacter			*Rickettsiella*		
Aerobe, gram-negative [663]	None	224	Facultative intracellular, gram-negative [664, 665]	*P. chinensis* [652]; *L. longipalpis* [312]	64 (2)
Noviherbaspirillum			*Pseudomonas*		
Aerobe, gram-negative, [666]	None	222	Aerobe, gram-negative [667]	*P. chinensis* [652]; *P. argentipes* [322]; *L.longipalpis* [311, 323, 325]	214 (1)
Corynebacterium			*Wolbachia*		
Aerobe, gram-positive [668]	None	221	Obligate intracellular, gram-negative [669]	*P. perfiliewi transcaucasicus* [670]; *L. longipalpis* [671]	13 (1)
Brevundimonas			*Stenotrophomonas*		
Aerobe gram-negative, [672]	*P. perniciosus* [326]	218	Aerobe, gram-negative [673]	*P. chinensis* [652]; *P. argentipes* [322, 660]; *P. perniciosus* [326]; *L. longipalpis* [323, 325, 651]	123 (1)

6.5.4 Distribution of sand fly microbiota

There is high stability in the core microbiota across Bihar, with differences in peripheral microbiota between districts (Figure 6-4) (ADONIS R^2 = 0.039; p = <0.001).

Shannon diversity scores for individual sand flies did not vary significantly between districts, with average scores of 2.79-3.18 (standard deviation ± 0.53-0.94) at district level. At the village level average Shannon diversity scores were more varied, with averages of 2.71-3.36 (± 0.40-1.09); at house level in Samastipur, average Shannon scores were lower and standard deviation was higher, indicating a wider variety of scores (2.74-2.87 ± 0.49-1.12). The variation in Shannon scores between individuals within geographical areas was due to several low outlier scores per area, giving a left-tailed skew to the scores.

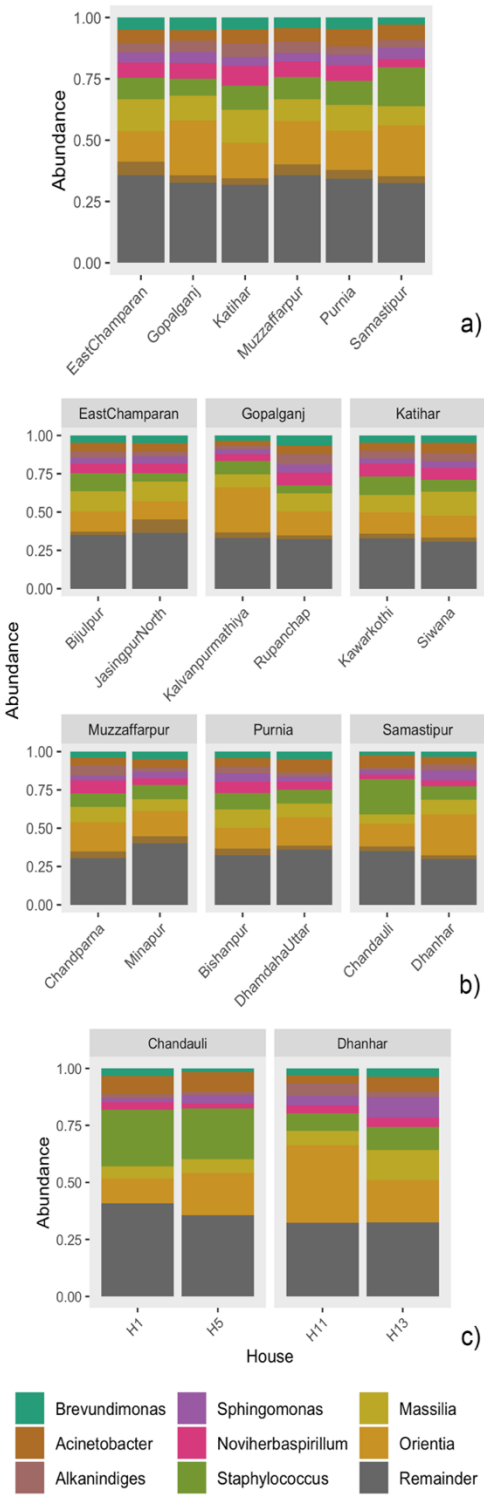


Figure 6-4 - Microbial genera present at ≥1% relative abundance in the P. argentipes microbiome dataset, organised by a) district; b) village; and c) house within Samastipur. Any genera that make up <1% of this subset have been labelled "Remainder".

6.5.5 Differences between individual sand flies

Microbial composition varied between individuals, including when collected from the same house (Figure 6-5). Samples with outlying low Shannon diversity scores were found to have highly uneven microbial communities, dominated by either a single genus or two genera. In 30/229 samples a single genus comprised ≥50% of the community, most often *Orientia* (19/30); details of these “dominating genera” can be found in Table 6-1.

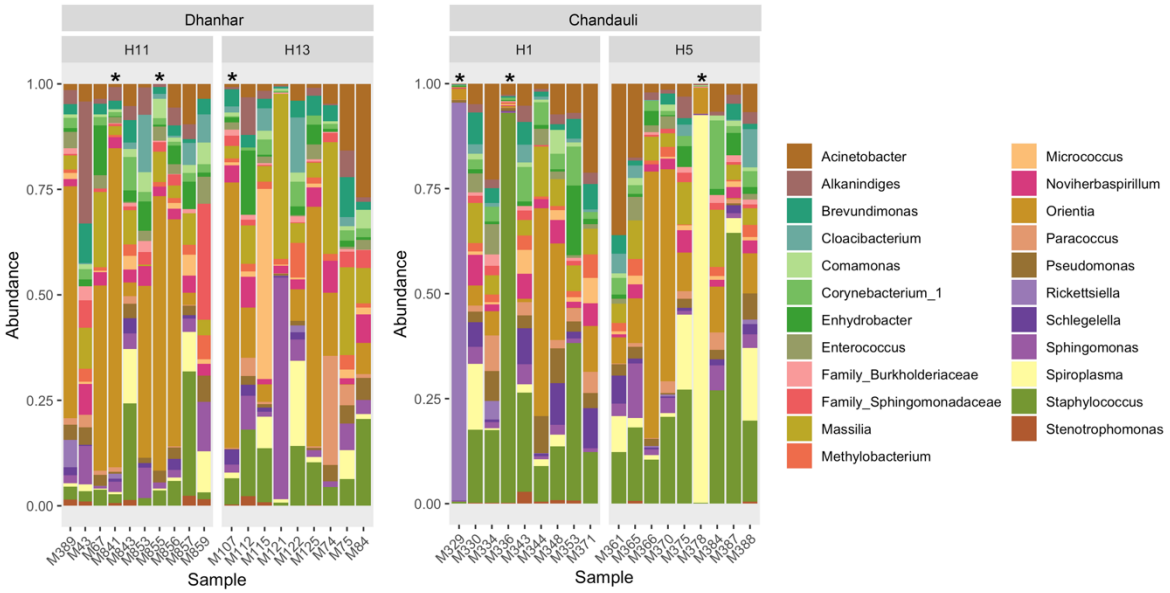


Figure 6-5 - The microbiota of individual P. argentipes grouped by house within two villages in Samastipur, Bihar. Genera shown are those found at ≥1% relative abundance in the data set. Samples where the total microbiota is made up of ≥50% one genera, referred to as “dysbiotic” samples, are labelled with an asterisk (*).

6.6 Discussion

Here, we developed a pipeline to assess the individual microbiomes of *P. argentipes* sand flies for the first time, circumventing previous difficulties surrounding using sequencing-only methods for microbiome characterisation in insects, including low-biomass related kit contaminants and run-to-run variation. Hitherto, phlebotomine sand fly microbiota studies have used either culture-dependent methods, or pools of samples to mitigate for low-biomass. Maleki-Ravasan *et al.* [650] identified *Enterobacteriaceae* as key components of the *P. papatasi* microbiota using biochemical and Sanger-sequencing techniques. Hillesland *et al.* [322] identified two potential targets for paratransgenic intervention in *P. argentipes* isolated from across Bihar: *Bacillus megaterium* and *Brevibacterium linens*. However, both of these studies used microbiological techniques linked to 16S rRNA sequencing to identify microbiota, limiting the study to aerobic bacteria, and only those that were amenable to culture-based methods [322]. Karakuş *et al.* [674] used culture-independent methods to describe the effect of geography on the microbiota of *P. papatasi* in Turkey, but only pools of 10 individual sand flies were examined, eliminating any differences between individuals. With these methodologies any “dysbiotic” individual sand flies would have been undetected. The methods presented in this study allow for the discovery and exploration of unusual microbial community phenotypes in individual sand flies while still describing the core microbiota of the population, both of which may have impacts on future microbiota-based VL control strategies.

6.6.1 *P. argentipes* has a core microbiome comprising seven genera and a variable peripheral microbiome

The development of this pipeline for examining individual sand flies with culture-independent methodology has allowed for the discovery of a core microbiota of seven genera, which were consistently present at high titres across Bihar, as well as a highly variable peripheral microbiome. The bacteria found in the core microbiome all possess key characteristics for use in vector control strategies: they are evidently capable of independently propagating and persisting within the *P. argentipes* population, allowing for their use in paratransgenic control of VL. The presence of a large and dynamic peripheral microbiome also suggests that *P. argentipes* are amenable to infection with a range of bacteria, making the use of bacteria in VL control strategies a promising alternative to traditional insecticide-based measures.

Investigating the transmission dynamics of core and peripheral sand fly microbiota will be important to inform future paratransgenesis-based *Leishmania* control strategies, as the method and efficiency of transmission will influence whether candidate bacteria species could be used. As continual dosage of wild populations with the chosen bacteria as described by Durvasula *et al.* [157] could be prohibitively expensive for long term interventions, candidate control bacteria should ideally be capable of consistent transmission between hosts. Although the core bacteria discovered in *P. argentipes* were found in over 95% of the samples in this study, these will not necessarily be vertically transmitted from parent to offspring. In the absence of close social contact bacteria can be transferred vertically by direct inoculation of eggs in the ovaries, egg-smearing or coprophagy [675]. However,

similarly consistent transmission of microbiota between non-social insects can occur when there is a close association with a common food source.

Several genera found in the *P. argentipes* core microbiome have previously been identified as having transmission routes between insect hosts that can be exploited in vector control strategies. *Acinetobacter* was present in 74% of a wild population of *Aedes albopictus* mosquitoes but was found exclusively in the midgut, indicating transmission from a food source [676]. Mosquitoes and sand flies share the essential features of their diet: males feed solely on plant sugars, and females feed on plant sugars with supplementary blood meals for egg production [677]. It can be hypothesised that the strong presence of the nectar microbe *Acinetobacter*, as well as the common fruit microbe *Massilia*, may be present in the core microbiome through ingestion rather than, or as well as, vertical transmission. Similarly, the human skin commensal bacterium *Corynebacterium* may be transferred to female *P. argentipes* passively or ingested during blood meals. The use of a constant food source for transferral of VL-controlling microbiota, as has been demonstrated in a proof-of-concept study for other vector-parasite systems [678], may circumvent the issue of small sand fly flight ranges, which would otherwise limit the horizontal transfer of bacteria used in control strategies.

6.6.2 Effects of *Leishmania* infection, and implications on the use of microbiota in sand fly control strategies

While it was not possible in this study to assess the impact of *Leishmania* infection on the *P. argentipes* microbiome due to a lack of *Leishmania*-positive samples, future work developing microbiota-based VL control strategies will need to explore

any *Leishmania*-mediated effects on *P. argentipes* microbiota. Kelly *et al.* [311] found that the microbiome of *L. longipalpis* was essential for the development of *L. infantum*, the causative agent of infantile VL in Latin America, into infective metacyclic promastigotes in the sand fly gut. An increase in *Acetobacteraceae* spp. and decrease in *Pseudomonadaceae* spp. was observed as *L. infantum* developed in the midgut, and removal of the midgut microbiome by antibiotics rendered *L. longipalpis* unable to support development of the parasites. Understanding whether similar host-microbiota dynamics occur in *P. argentipes* will be key to the development of microbiota-based VL control strategies.

6.6.3 Potential functions of core and dominating genera

The functions of the core microbiota in *P. argentipes* are uncertain, but ideas can be drawn from their effects in other host insects. For example, *Sphingomonas* is a known lignin-degrader, indicating a potential role in yeast digestion [679]. *Acinetobacter* has been implicated in aiding *Ae. albopictus* mosquitoes to survive in polluted, anthropised environments similar to the areas sampled in this study [676]. Associated bacteria can play important roles in insect fecundity: *Acinetobacter* has been demonstrated to be essential to the stable fly species *Stomoxys calcitrans* development, and *Brevundimonas* in horse manure has been observed to induce oviposition in *S. calcitrans* using the manure as a food source [680].

It should also be stated that diversity is not necessarily an inherent feature of healthy microbiomes; limited microbial diversity in the host microbiome may improve functionality [570]. If a bacterial species is conferring a benefit to the host that would increase host fitness proportional to the titre of the species, a microbiome dominated

by this species may not indicate “dysbiosis”, but rather adaptation. The identification of bacterial genera dominating individual sand flies in this study may be as important a discovery as identification of a core microbiome, in terms of enhancing the understanding of the role of sand fly microbiota and in terms of developing VL control strategies around key bacterial species.

6.6.4 A strong core microbiome and the confounding effect of “dysbiotic” individuals reduces the variation of the *P. argentipes* microbiome between locations

By looking at individual *P. argentipes* microbiomes at varying geographical scales, we discovered a stable core unaffected by location within Bihar, and a peripheral microbiome that varies widely in abundance at finer scales than have been previously examined. The effect of geographical location on the sand fly microbiome appears to be statistically minimal due to the presence of a strong core microbiome, but may be confounded by the presence of individuals with significantly uneven microbial compositions (“dysbiotic” individuals). This effect explains the decrease in diversity between groups as the geographical scale increases: the effect of confounding individuals is “diluted out”, as is the effect of a highly variable peripheral microbiome. The instability of the non-core microbiome, presumably by the influence of the environment on the sand fly microbiome, implies that bacteria used in control strategies may readily infect and establish in wild *P. argentipes* populations. Such variation in microbiota is common in small insects closely associated with soil and vegetation [570]. On the other hand, the vector control bacteria may be easily displaced in the wild *P. argentipes* host by other invading bacteria. The dynamics of

bacteria introduced into sand fly populations must be examined over time in order to confirm the viability of its use in long-term control strategies.

6.6.5 Targets for paratransgenesis-based *P. argentipes* control strategies

Several previously identified targets for paratransgenic VL control strategies in *P. argentipes* were discovered in this study: *Brevibacterium* and *Bacillus* were found in low titres, in 82/229 and 83/229 sand flies respectively. *Bacillus megaterium* is commonly used for industrial protein production [681], and *Brevibacterium linens* is used for enzyme production in the dairy industry [682], meaning both have been extensively studied and cultured in a laboratory setting, fulfilling one of the required traits for a paratransgenesis target. Hillesland *et al.* [322] identified *B. megaterium* and *B. linens* in *P. argentipes* from Bihar based on *in vitro* culture. While the taxa in this study were not identified down to species level so the presence of *B. megaterium* and *B. linens* cannot be confirmed, our results provide more evidence that *Bacillus* and *Brevibacterium* are capable of infecting and establishing in *P. argentipes*, making them important candidates for use in VL control strategies.

Other bacteria with the potential to be used in paratransgenesis-based VL control strategies were discovered in this study. *Pantoea,* here identified in 27/229 samples, has previously identified as a putative paratransgenic agent: *Pantoea agglomerans* was transformed to express anti-*Plasmodium* effector proteins to control malaria in *Anopheles* mosquitoes [683]. Identified as a core component of the *P. argentipes* microbiome (see Table 6-1), *Corynebacterium* has previously been transformed to produce an immunologically active single-chain antibody fragment, as a model for

4312 expression of similar antibodies that can control the related kinetoplastid
4313 *Trypanosoma cruzi* [340]. As its suitability for paratransgenesis has been
4314 demonstrated and we have shown its ubiquity in *P. argentipes*, *Corynebacterium*
4315 should be investigated for its use in VL control. The next steps in this process would
4316 involve identifying candidate genes to introduce into *Corynebacterium* and assessing
4317 the impact of introducing paratransgenic *Corynebacterium* into a *P. argentipes*
4318 population.
4319
4320 Previous investigations into the microbiome of *P. argentipes* have used culture-
4321 based methods to identify target bacteria for paratransgenesis-based VL control
4322 strategies, which may have under-represented non-culturable, difficult to culture or
4323 some anaerobic species where appropriate conditions were not achievable [322].
4324 Anaerobic culturing combined with a non-culture-based approach may complement
4325 current aerobic culture-based studies to more-fully describe the microbiome of *P.*
4326 *argentipes*, towards designing vector control strategies, although it should be noted
4327 that culturability is a key feature of such control strategy design. Indeed, the
4328 limitation of this high-throughput approach is the inability to identify bacteria further
4329 than genus level due to the lack of resolution in short amplicon sequencing. To
4330 further confirm the presence of specific bacteria of interest, in particular those with
4331 potential for paratransgenesis such as *Corynebacterium*, further studies utilising full
4332 metagenomic sequencing, or long-read length amplicon sequencing (e.g. Pacific
4333 Biosciences or Oxford Nanopore) would be necessary.
4334

6.6.6 Human pathogens

Accurate knowledge of infection routes of VBDs is essential for effective monitoring and targeting of control strategies; disease epidemiology can be monitored by detection of pathogens in insects [684, 685]. One genus present in all districts and frequently found in “dysbiotic” levels was *Orientia*, a genus comprising of two species, *O. tsutsugamushi* and *O. chuto*, both of which cause the serious neglected disease scrub typhus [656, 686, 687]. Currently the only confirmed vector of scrub typhus are chigger mites [688]; the transmission of *Orientia* is mechanical, indicating a similar transmission route may be possible in sand flies. *Orientia* has been implicated as a causative agent of acute encephalitis syndrome in Bihar [689], which killed over 100 children in June 2019 in Muzzaffarpur; various other potential causes of this acute encephalitis syndrome outbreak included enteroviruses and eating lychee fruit [690]. It has been suggested that unknown reservoirs and vectors of *Orientia* exist, evidenced by its high antigenic diversity and periodic re-emergence [691]. While the presence of *Orientia* does not confirm any vectorial capacity of *P. argentipes,* it does indicate that the ability of phlebotomine sand flies to transmit scrub typhus should be investigated. *P. argentipes* in northern India have been demonstrated to vary their blood meal host depending on the nearby host availability, with larger proportions of sand flies containing blood meals from humans caught in human dwellings (48.94%) compared to those caught in cattle sheds (1.84%) [692]. This is congruent with the small home ranges of sand flies, which are associated with single dwellings. However, individual *P. argentipes* have also been observed to take blood meals from multiple host species more frequently than *P. papatasi*; host meal combinations include human and bovine, human and caprine (goats), human and canine, and human and murine blood [692, 693]. This propensity

to feed on multiple hosts increases the potential competence of *P. argentipes* for spreading zoonotic diseases including *Orientia*, which has a wide host range that includes cows, goats, dogs and rodents [694].

The genus *Bartonella* was seen in East Champaran, Muzzaffarpur and Samastipur, and contains 11 human pathogenic species including the causative agents of Carrion's disease and cat-scratch disease [695]. *Bartonella* transmission is most often attributed to ticks but has also been historically associated with *Lutzomyia* sand flies, although this route has not been well characterised [696, 697]. The possibility of *Phlebotomus* sand flies being competent vectors of *Bartonella* should be investigated further.

6.6.7 Entomopathogens and reproductive manipulators

A proportion of individual *P. argentipes* sand flies had reduced microbiome diversity that included known insect pathogens, including *Spiroplasma* [698], *Rickettsiella* [699] and *Pseudomonas* [700]. *Spiroplasma* a reproductive manipulator of insects similar to *Wolbachia*, was present in 168/229 samples and in all districts [701]. *Spiroplasma* has been previously described in *Phlebotomus* sand flies [652, 702]; in one study by Karatepe *et al.* [702] only female *Phlebotomus* sand flies were infected, suggesting a potential effect of *Spiroplasma* on sex ratio. In contrast, *Wolbachia* was seen in 13/229 (5.7%) of our samples. A recent study of *Spiroplasma* and *Wolbachia* in Turkish *Phlebotomus* species contrasted with this, observing a 70% infection rate of *Wolbachia* and 13% infection rate of *Spiroplasma* [702]. While the presence of *Wolbachia* suggests that sand flies may be amenable to infection with another *Wolbachia* strain for control, as has been done in *Aedes aegypti* mosquitoes [140],

the existence of a natural *Wolbachia* infection is not a prerequisite for this and could conceivably hinder the introduction of a control strain by outcompeting it. Further study into the use of *Wolbachia* or *Spiroplasma* for controlling *P. argentipes* in a similar way to the successful *Wolbachia*-based strategies for mosquito control is warranted.

The insect pathogen *Rickettsiella* was present in 64/229 (28%) samples, in four out of six districts. *Rickettsiella* has been previously described in *P. chinensis* and *Lu. longipalpis* [312, 652], but as *Rickettsiella* is not currently culturable *in vitro* [703], it is likely a non-optimal choice for sand fly control strategies until this is resolved. *Orientia,* an obligately intracellular bacterium, was by far the most common dominating genus; however, its role in insect pathogenicity is currently poorly described.

In summary: there appears to be a core microbiome of *P. argentipes*, but the peripheral microbiome is heavily influenced by the environment and invading pathogenic bacteria, as is common for small insects. There appears to be no geographical effect at a fine scale except for this uniform environmental pressure. This instability in non-core microbiota suggests that native microbiota-based control strategies may be difficult to implement. However, the apparent susceptibility of *P. argentipes* to bacterial invasion could be important for the implementation of paratransgenic bacteria or heterologous association-based control strategies. This study also highlights some potential paratransgenesis targets, several of which have proof-of-concept studies already completed. The level and mechanisms of bacterial horizontal and vertical transmission need to be confirmed before bacterial control

4410 strategies could be implemented, particularly due to the small flight zones of this
4411 species.

Chapter 7 Discussion

7.1 Summary of research goals

Research into the microbiota of disease vectors has increased dramatically in the last 20 years due to the development of genomic technology, which is now widely available, cheaper, and more accessible for larger projects. While vector microbiota present important research avenues for novel vector control strategies, their application requires an in depth understanding of the vector-parasite-microbiome system. Many vector control strategies involving bacteria use genome editing to produce a paratransgenic bacterium capable of delivering effector molecules to the vector host; these strategies require a complete understanding of the potential for the genetically modified organism to exist outside of the target organism or to spread within a target population.

Two of neglected vector-borne diseases that have been identified as having potential for elimination but require novel control strategies to achieve this are trypanosomiasis, which is spread by the bite of an infected tsetse fly, and leishmaniasis, spread by the bite of an infected sand fly. In this project we aimed to use rapidly developing sequencing technologies to investigate the bacterial microbiota of these two vectors, to provide important information necessary for developing microbiota-based control strategies. Tsetse flies have a well characterised bacterial microbiome including the secondary symbiont *Sodalis glossinidius*, which is present to varying degrees in wild tsetse populations but becomes fixed in laboratory-reared tsetse populations. *Sodalis* has all of the required

properties of a target for vector control – it is culturable *in vitro*, horizontally transmissible, and non-pathogenic – however, it has been demonstrated to enhance tsetse vector competence by an unknown mechanism.

The sand fly microbiome is very different to the tsetse microbiome; sand flies have no known primary or secondary symbionts and are much more difficult to study due to their small size and therefore low bacterial load. While microbiome sequencing technologies have become more accessible in recent years the standardisation of techniques in microbiome studies is low, resulting in a lack of comparability between studies with little or no consensus on best practices.

In order to address the gaps in the knowledge base of these two vectors, we have applied several different sequencing and molecular biology-based techniques to answer questions about their native microbiota. Information was gathered on the distribution of *Sodalis, Wolbachia, Spiroplasma* and trypanosomes in wild Ugandan tsetse flies, a major focus of trypanosomiasis. An experimentally evolved strain of *Sodalis* was sequenced and compared with an ancestral genome to understand its evolutionary trajectory towards obligate endosymbiosis. The expression of the tsetse host and its symbiotic bacteria during exposure to trypanosomes was examined to determine which genes were involved in tsetse susceptibility to trypanosome infection. In sand flies, the microbiome of individual flies was sequenced to determine the differences between individuals and the effect of fine geospatial scales on microbiome structure. A novel pipeline using best practices for microbiome studies was developed to produce this dataset.

7.2 Discussion of findings compared to previous works

7.2.1 Comparison of microbiome data to previous studies of tsetse flies and sand flies

While the differences in methodology have a significant impact on the microbiomes described between studies, comparing incidence of bacterial genera between datasets is useful for understanding the natural variation of insect microbiomes. The microbiome of tsetse flies is dogmatically described as consisting of *Wigglesworthia*, *Sodalis* and *Wolbachia*; this is now understood to not be the case [228], spurring multiple recent studies of the wider tsetse microbiota with which the results from this thesis can be compared. The prevalence of *Wolbachia* (1.7%)*, Spiroplasma* (88%)*, Sodalis* (39%), and *Trypanosoma* (12%) in wild Ugandan tsetse flies (Chapter 2) contrasted with those of other studies despite sampling in the same country, often in similar regions and habitats. This demonstrates the impact of biotic and abiotic factors in microbiome studies, even in a group of insects previously believed to contain a restricted and relatively stable microbiome. The samples in Chapter 2 were collected in January, during the dry season; Schneider *et al.* observed lower prevalence of *Spiroplasma* in the dry and wet seasons in Uganda compared to the intermediate season [273], and lower prevalence overall compared to the study in this work, suggesting the variation cannot be explained only by seasonal differences.

Variation between studies can be due to a combination of methodological factors; for example, Schneider *et al.* used 16S sequencing and MLST to identify *Spiroplasma*, whereas Aksoy *et al.* used 16S sequencing and qPCR [272, 273], neither of which was employed in this study. Aksoy *et al.* only included flies with a wing fray analysis

score of ≥2 (estimated to be at least 3-4 weeks old) to ensure only flies with multiple blood meals were studied to maximise potential trypanosome exposure, something that was not considered in this study or the Schneider *et al.* study. Aksoy *et al.* also did not give information on in which seasons trapping was conducted. All of these factors, and many more besides, can introduce bias into a dataset that confounds comparisons with other datasets.

There were no datasets that could be directly compared with the one described in Chapters 5 and 6, due to the novelty of sequencing from individual sand flies; previous studies have either used culture-dependent methods or sequenced pools of samples. The advantages and disadvantages of these approaches have been described elsewhere, but regardless of whether one approach is superior to another or not the biases introduced by each makes comparisons between them difficult. In particular there is little in the way of comparable datasets of individual sand fly microbiomes, meaning that the 'dysbiotic' state seen in multiple individuals in Chapter 6 has been previously unnoticed. While culture-based methods have been used on individual sand flies, the difficulty of isolating live bacteria from samples means that many samples appear to have no bacteria at all, which is biologically unlikely due to their life cycle and habitats. This means that presence/absence data of individual bacterial species from culture-based studies is unreliable, and any 'dysbiosis' of single bacterial species dominating an individual's microbiome could be attributed to ineffective culture techniques.

7.2.2 Comparison of contamination removal in low-biomass datasets

A study of the microbiome of chigger mites, the vectors of scrub typhus (*Orientia sp.*), demonstrated a different approach to a similar issue of quality assurance in low-biomass microbiome data [704]. Trombiculoid mites have a larval stage which lives ectoparasitically on vertebrates and an adult stage which lives freely in the soil [705]. Chiggers are generally <250 µm long, making them difficult to use in molecular assays [706]. Chaisiri *et al.* sequenced both individual chigger mites and pools of 50 mites, as well as soil samples and negative controls [704]. Rather than removing sequences classified as contaminants as was performed in Chapter 5 and 6, Chaisiri *et al.* instead compared beta diversity of samples to controls using the 'ecodist' package in R, removing any samples that were >20% aligned to the soil and negative controls. Confirmation of the common contaminant *Geobacillus* as a true member of the chigger microbiome was conducted using Sanger sequencing and qPCR, proving that the dominant *Geobacillus* present in true samples was absent in background controls. This demonstrates that there is no single gold-standard approach for microbiome work; both Chapter 5 and the study by Chaisiri *et al.* produced highly filtered and robust microbiome datasets from samples with extremely low biomass. Both approaches, however, resulted in the loss of a significant proportion of reads; in Chapter 6, 82.1% of ASVs were identified as contaminants and removed from the dataset, but no samples were lost at this stage. The approach used by Chaisiri *et al.* resulted in the removal of 58.7% of samples, restricting the analysis that could be performed on several chigger species; however, the study confirmed results from individual chiggers by comparison with pools, a step that was not taken in our study.

The best way to compare these approaches would be to sequence serial dilutions of a known microbial community and perform both types of filtering. Karstens *et al.* performed a very similar study, sequencing 16S genes of a mock microbial community that had been 3-fold diluted 8 times and assessing contamination removal methods [707]. Four contamination removal methods were tried: removal of all ASVs present in the negative control; removal of ASVs present below a relative abundance threshold; use of the Decontam package with the frequency method; and the predictive modelling software SourceTracker which uses a Bayesian approach to predict the proportion of a sample that is made up of non-expected ASVs, such as common contaminants or ASVs from negative controls. Removal of all ASVs in negative controls and removal of ASVs below a certain threshold resulted in both removal of true ASVs and inclusion of contaminating ASVs. Decontam performed better than SourceTracker when the community being studied was not well defined, as is the case in most microbiome studies. While not a true comparison of the methodology used by Chaisiri *et al.* and the methodology used in Chapter 6, it does suggest that perhaps the filtering out of samples from the Chaisiri *et al.* dataset was overly stringent [704].

7.2.3 The evolving roles of *Sodalis*

Sodalis is most notable for its ability to live intracellularly within a host as well as *in vitro*; this is due to its recent evolution towards a symbiotic lifestyle and thus its relatively unspecialised genome [60]. Two free-living members of the *Sodalis* genus have been discovered in the last decade: *S. praecaptivus*, which was isolated from a human wound after the patient was impaled with the branch from a dead crab apple tree [251]; and *S. ligni*, isolated from dead wood [708]. These free-living species with

less genome degradation could represent close ancestors of *S. glossinidius* or share a common ancestor. Flux-balance analysis studies of *S. glossinidius* metabolism have described complete functionality of an ancestral metabolic network under minimal conditions *in silico*, with environmental conditions defining the trajectory of *in silico* evolution [486, 506]. The results in Chapter 4 align with these studies; they show the flexibility of the *Sodalis* transcription in response to different environments, with modified gene expression in *in vitro* culture compared to *in vivo* in tsetse flies and between infected and uninfected tsetse. The degradation of the thiamine biosynthesis pathway seen in *Sg*GMMC1* also reflects the prediction of Hall *et al.* in which *Sodalis* was predicted to evolve towards an intersymbiont dependence with *Wigglesworthia* for thiamine provision [486].

Working on the assumption that *Sodalis* is responsible for increasing tsetse susceptibility to trypanosome infection [264, 537], the question becomes: what factors promote the response of *Sodalis* to trypanosome infection that leads to tsetse susceptibility? In the experiment in Chapter 4 all tsetse flies were reared in the same environment, inoculated with trypanosomes from the same blood meal, and are all infected with *Sodalis*. One potential cause of differential *Sodalis* expression, and by extension increased tsetse susceptibility, may lie in *Sodalis* density. Quorum sensing, whereby bacteria communicate to modulate the gene expression of the population based on cues such as bacterial density [469, 470], has been shown to attenuate virulence in the free-living *Sodalis* relative *Sodalis praecaptivus* sufficiently to allow it to colonise tsetse and be transmitted vertically [472, 487]. Downregulation of virulence genes occurs when bacterial density increases to the point where the negative effect to the host reduces bacterial fitness [709]. In a similar study of

Sodalis transcription in a different tsetse-trypanosome species system, Soumana *et al.* suggested that increased *Sodalis* density may be responsible for increased refractoriness in tsetse due to stimulation of the tsetse innate immune system [514]. In this study we saw upregulation of multiple virulence-associated genes in infected flies compared to refractory flies, and while there was no significant difference in the percentage of reads mapped to *Sodalis* between the control, refractory and infected groups, there appeared to be a superficially lower number of *Sodalis* reads in infected flies (Figure 4-6). This suggests that, rather than high *Sodalis* density increasing tsetse refractoriness, low density of *Sodalis* and thus increased virulence gene expression may increase tsetse susceptibility. This may explain the fact that studies of trypanosome-*Sodalis* co-infection in wild tsetse populations find wide variation in the association of *Sodalis* with trypanosomes [102, 396, 441, 710]. A larger sample size may be needed to increase resolution of the differences, with a more accurate *Sodalis* density biomarker using a more sensitive methodology such as qPCR.

7.3 Future work

7.3.1 *Sodalis*: studies *in vitro*, *in vivo* and in the field

The strain *Sg*GMMC1* was generated by long-term *in vitro* culture with the simple aim of maintaining a *Sodalis* line. Long-term experimental evolution of *Sodalis* under varying selection pressures, such as increased oxidative stress or nutrient deprivation, may alter the evolutionary trajectory of *Sodalis* and elucidate the roles of varying gene systems. Similarly, long-term passaging of *Sodalis* through tsetse which are consistently exposed to trypanosomes may result in genetic, epigenetic or

transcription changes in the produced *Sodalis* strain, increasing our understanding of the tripartite interaction between symbiont, host, and parasite.

The lab strain *Sg*GMMC1* has diverged sufficiently from the tsetse-isolated ancestor strain that comparison studies of its ability to infect tsetse flies and its effect on tsetse vector competence should be performed. It is also important to ensure that work performed on *in vitro Sodalis* strains with potential real-world applications is representative of wild-type *Sodalis*; further work to improve field culture techniques from wild tsetse will be essential for this. A more in-depth study of *Sodalis* transcription in trypanosome-refractory and infected tsetse flies, with depletion of host and *Wigglesworthia* signals, may increase our insight into *Sodalis*' role in increasing tsetse trypanosome susceptibility. In particular, the expression of *Sodalis* pseudogenes could be studied in more detail by sequencing whole primary transcripts before any post-transcriptional modification. Finally, whole-genome sequencing of individual clones isolated from single tsetse colonies may reveal polymorphisms within the population that account for the differences in tsetse refractoriness.

7.3.2 Wild tsetse microbiota

With lab access restricted due to the COVID-19 pandemic, many tsetse fly samples collected in Uganda are yet to be analysed and no data has been collected on *Serratia* presence, trypanosome species or tsetse species. Completion of this work will increase the power of the study, allowing for a more robust analysis of correlations between the presence of tsetse microbiota in different tsetse species. Once the field culturing methodology has been developed it may also be possible to

culture wild isolates of *Spiroplasma* and *Serratia*, for further characterisation *in vitro* and for use in infection experiments using lab-reared tsetse. The widespread presence of *Spiroplasma* in the sampled tsetse indicates that it may be evolving towards secondary symbiosis, and the interactions with trypanosomes in tsetse observed in other studies requires further investigation.

7.3.3 Investigating the sand fly microbiota

The groundwork performed in this study to elucidate a strong core microbiome and highly variable peripheral microbiome of sand flies produces a large number of questions to be answered about the role of sand fly microbiota. Understanding the transmission routes of sand fly microbiota between individuals and from the environment will be essential for using bacteria in VL control strategies. As the leishmania infection rates were so low it was not possible to examine the impact of leishmania infection on bacterial microbiota; this could instead be studied using experimental infection of lab-reared sand fly populations. If the majority of the sand fly peripheral microbiome is influenced by the environment as this study suggests, any studies of the interactions between leishmania parasites and the bacterial microbiome should be conducted using semi-field conditions.

The origin and effect of the “dysbiotic” microbiome states requires further study, in particular to determine whether or not the reduction of microbiome diversity could be manipulated to spread a paratransgenic control bacterium. Additionally, as the scrub typhus agent *Orientia* is spread mechanically by the bite of chigger mites the potential for transmission by the bite of infected sand flies should be investigated urgently as a matter of public health.

7.3.4 Using and improving sequencing technology to study insect microbiota

Until recently the limiting factor of microbiome research has been the technology, which was not sufficiently advanced to be readily available for studying insect microbiota. As technology has improved, the limiting factor has become our ability to analyse and interpret the enormous amount of data generated by sequencing. Many raw sequencing datasets are made freely available, allowing for further study or use in other experiments, but there is a lack of suitable analysis methods and therefore the true potential of many sequencing datasets goes unexplored. Additionally, the reuse of datasets requires careful examination of the methodologies used, particularly if a meta-analysis or comparison between studies is performed.

A major hurdle to the full interpretation of sequencing data lies in database gaps. All of the sequencing performed in this thesis was entirely reliant on the accessibility, accuracy and completeness of gene databases: the silva database of known 16S gene sequences and their taxonomic identities for the sand fly experiment; the reference genomes and database of gene names and functions for the tsetse metatranscription experiment; the annotated ancestor strain *Sg*GMMB4 with labelled pseudogenes; and gene function annotations in wider gene databases for all *Sodalis* work. Although these databases represent an enormous amount of work, they are still far too incomplete to allow for full analyses of the data produced. In this work any putative protein genes with no annotated or related function, or 16S sequences with no taxonomic ID, result in a large proportion of potentially important information going unanalysed. Investing in database capacity building – using basic research to

fill in the gaps in genome, gene, and protein databases – will greatly improve the quality of microbial ecology studies across multiple fields.

7.4 Conclusions

Although the use of insect microbiota for vector control is in its infancy, the number of studies investigating vector microbiomes is increasing rapidly. The main hurdle to the use of bacteria in vector control strategies is the lack of understanding of the complex interactions between bacterial species and their hosts; trials using transgenic bacteria in wild or semi-wild host populations are rarely approved due to uncertainty surrounding the impact of these bacteria on non-intended targets. In order to release bacteria into wild vector populations the full implications of the action must be understood; this can only be achieved by detailed and accurate analysis of vector microbiota, their ecology and the mechanisms of their impact on the host.

Both trypanosomiasis and leishmaniasis are enormously detrimental to public health of vulnerable populations, as well as epidemiologically complex and recalcitrant to control by traditional methods. In this thesis valuable insights into the complex *Sodalis*-tsetse interaction have been made, hopefully pushing the field one step further towards a *Sodalis*-mediated control strategy for trypanosomiasis. Basic research is the cornerstone of effective microbiota-based control strategies; the work done in this thesis to build an improved sequencing pipeline for low biomass samples and using it to characterise the microbiota of individual sand flies is an early step in what will hopefully be a successful novel strategy for controlling leishmaniasis. The sequencing data produced over the course of this thesis could be

used to explore other aspects of the organisms in question – for example, the full expression of each of the organisms in the metatranscription experiment in Chapter 4 contain a wealth of information about their basic biology. The development of methodologies in this work could be extended to research into the control of other vector-borne diseases, particularly neglected diseases where medical intervention is not always possible due to the poverty or rurality of the affected populations. As the technology to understand the complex epidemiology of vector borne diseases improves, work towards their control and elimination must continue.

References

1. Leak, S. G. A. (1999). *Tsetse biology and ecology: their role in the epidemiology and control of trypanosomosis*. Wallingford, UK: CABI Publishing.

2. Gullan, P. J. and Cranston, P. S. (2014). *The Insects: An Outline of Entomology*. Fifth ed: Wiley & Sons.

3. World Health Organization (2019). Global Health Observatory data repository: Human African Trypanosomiasis. Global Health Observatory, Geneva. Available from: https://apps.who.int/gho/data/node.main.A1635?lang=en.

4. World Health Organization (2018). Number of cases of visceral leishmaniasis reported - Data by country. Global Health Observatory, Geneva. Available from: https://apps.who.int/gho/data/node.main.NTDLEISHVNUM?lang=en.

5. World Health Organization. (2019). Global Health Observatory Map Gallery. Available from: http://gamapserver.who.int/mapLibrary/app/searchResults.aspx.

6. Mitra, A. K. and Mawson, A. R. (2017). Neglected tropical diseases: epidemiology and global burden. *Tropical medicine and infectious disease*. **2** (36): 1-15.

7. World Health Organization. (2016). Recommendations for the adoption of additional diseases as neglected tropical diseases. Available from: https://www.who.int/neglected_diseases/diseases/Adoption_additional_NTDs.pdf.

8. World Health Organization. (2020). Neglected tropical diseases: a summary. Geneva. Available from: https://www.who.int/neglected_diseases/diseases/en/.

9. Roberts, L. S., Schmidt, G. D. and Janovy, J. (2004). *Foundations of Parasitology*: McGraw-Hill Higher Education.

10. World Health Organization. (2017). Vector-borne diseases. Available from: https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases.

11. Gubler, D. J. (2009). Vector-borne diseases. *Scientific and Technical Review of the Office International des Epizooties*. **28** (2): 583-8.

12. Finn, R. N., Chauvigné, F., Stavang, J. A., Belles, X. and Cerdà, J. (2015). Insect glycerol transporters evolved by functional co-option and gene replacement. *Nature Communications*. **6** (1): 1-7.

13. Paz, S. (2019). Effects of climate change on vector-borne diseases: an updated focus on West Nile virus in humans. *Emerging Topics in Life Sciences*. **3** (2): 143-52.

14. Vos, T., Lim, S. S., Abbafati, C., Abbas, K. M., Abbasi, M., Abbasifard, M.*, et al.* (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*. **396** (10258): 1204-22.

15. Puntasecca, C. J., King, C. H. and LaBeaud, A. D. (2021). Measuring the global burden of chikungunya and Zika viruses: A systematic review. *PLOS Neglected Tropical Diseases*. **15** (3): e0009055.

16. Mattioli, R. C., Feldmann, U., Hendrickx, G., Wint, W., Jannin, J. and Slingenbergh, J. (2004). Tsetse and trypanosomiasis intervention policies supporting sustainable animal-agricultural development. *Journal of Food Agriculture and Environment*. **2** (2): 310-4.

17. Centers for Disease Control and Prevention. (2021). Tickborne Disease Surveillance Data Summary. CDC, Available from: https://www.cdc.gov/ticks/data-summary/index.html.

18. Barker, C. M. and Reisen, W. K. (2019). Chapter 4 - Epidemiology of Vector-Borne Diseases. Ed: Mullen GR, Durden LA, In: *Medical and Veterinary Entomology*. vol.

19. Taylor, L. H., Latham, S. M. and Woolhouse, M. E. J. (2001). Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. **356** (1411): 983-9.

20. Centers for Disease Control and Prevention. (2018). Malaria Biology: Lifecycle. Atlanta, USA. cited 2020. Available from: https://www.cdc.gov/malaria/about/biology/index.html#.

21. Beier, J. C. (1998). Malaria Parasite Development in Mosquitoes. *Annual review of entomology*. **43** (1): 519-43.

22. Sharma, R., Gluenz, E., Peacock, L., Gibson, W., Gull, K. and Carrington, M. (2009). The heart of darkness: growth and form of *Trypanosoma brucei* in the tsetse fly. *Trends in Parasitology*. **25** (11): 517-24.

23. Baxter, R. H. G., Contet, A. and Krueger, K. (2017). Arthropod innate immune systems and vector-borne diseases. *Biochemistry*. **56** (7): 907-18.

24. Ukegbu, C. V., Giorgalli, M., Tapanelli, S., Rona, L. D. P., Jaye, A., Wyer, C.*, et al.* (2020). PIMMS43 is required for malaria parasite immune evasion and sporogonic development in the mosquito vector. *Proceedings of the National Academy of Sciences*. **117** (13): 7363-73.

25. Rogers, M. E., Chance, M. L. and Bates, P. A. (2002). The role of promastigote secretory gel in the origin and transmission of the infective stage of *Leishmania mexicana* by the sandfly *Lutzomyia longipalpis*. *Parasitology*. **124** (5): 495-507.

26. Pimenta, P. F. P., Modi, G. B., Pereira, S. T., Shahabuddin, M. and Sacks, D. L. (1997). A novel role for the peritrophic matrix in protecting *Leishmania* from the hydrolytic activities of the sand fly midgut. *Parasitology*. **115** (4): 359-69.

27. Anderson, R. M. and May, R. M. (1982). Coevolution of hosts and parasites. *Parasitology*. **85** (2): 411-26.

28. Zuzarte-Luís, V. and Mota, M. M. (2018). Parasite sensing of host nutrients and environmental cues. *Cell Host & Microbe*. **23** (6): 749-58.

29. Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W.*, et al.* (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*. **419** (6906): 498-511.

30. Olszewski, K. L., Morrisey, J. M., Wilinski, D., Burns, J. M., Vaidya, A. B., Rabinowitz, J. D.*, et al.* (2009). Host-parasite interactions revealed by *Plasmodium falciparum* metabolomics. *Cell Host & Microbe*. **5** (2): 191-9.

31. Olotu, A., Fegan, G., Wambua, J., Nyangweso, G., Leach, A., Lievens, M.*, et al.* (2016). Seven-year efficacy of RTS,S/AS01 malaria vaccine among young African children. *New England Journal of Medicine*. **374** (26): 2519-29.

32. Black, S. J. and Mansfield, J. M. (2016). Prospects for vaccination against pathogenic African trypanosomes. *Parasite Immunology*. **38** (12): 735-43.

33. Halstead, S. B. (2017). Dengvaxia sensitizes seronegatives to vaccine enhanced disease regardless of age. *Vaccine*. **35** (47): 6355-8.

34. Powers, A. M. (2018). Vaccine and therapeutic options to control chikungunya virus. *Clinical Microbiology Reviews*. **31** (1): e00104-16.

35. Hisaeda, H., Yasutomo, K. and Himeno, K. (2005). Malaria: immune evasion by parasites. *The International Journal of Biochemistry & Cell Biology*. **37** (4): 700-6.

36. MacGregor, P., Szöőr, B., Savill, N. J. and Matthews, K. R. (2012). Trypanosomal immune evasion, chronicity and transmission: an elegant balancing act. *Nature Reviews Microbiology*. **10** (6): 431-8.

37. Cupp, E., Sauerbrey, M., Cama, V., Eberhard, M., Lammie, P. J. and Unnasch, T. R. (2019). Elimination of onchocerciasis in Africa by 2025: the need for a broad perspective. *Infectious Diseases of Poverty*. **8** (1): 1-5.

38. Kamgno, J. and Djeunga, H. N. (2020). Progress towards global elimination of lymphatic filariasis. *The Lancet Global Health*. **8** (9): e1108-e9.

39. Büscher, P., Cecchi, G., Jamonneau, V. and Priotto, G. (2017). Human African trypanosomiasis. *The Lancet*. **390** (10110): 2397-409.

40. Lindner, A. K., Lejon, V., Chappuis, F., Seixas, J., Kazumba, L., Barrett, M. P.*, et al.* (2020). New WHO guidelines for treatment of *gambiense* human African trypanosomiasis including fexinidazole: substantial changes for clinical practice. *The Lancet Infectious Diseases*. **20** (2): e38-e46.

41. Centers for Disease Control and Prevention. (2015). History of Malaria: The Panama Canal. Atlanta, USA. Available from: https://www.cdc.gov/malaria/about/history/panama_canal.html.

42. De Raadt, P. (2005). The history of sleeping sickness: towards systematic control. World Health Organization, Available from: https://www.who.int/trypanosomiasis_african/country/history/en/index7.html.

43. Wilson, A. L., Courtenay, O., Kelly-Hope, L. A., Scott, T. W., Takken, W., Torr, S. J.*, et al.* (2020). The importance of vector control for the control and elimination of vector-borne diseases. *PLOS Neglected Tropical Diseases*. **14** (1): e0007831.

44. World Health Organization. (2018). Overview of the intervention classes and prototype/products under Vector Control Advisory Group (VCAG) review for assessment of public health value (As of 1 March 2020). Available from: https://apps.who.int/iris/handle/10665/274451.

45. Edwards, C. A. (1973). *Environmental Pollution by Pesticides*. First ed. London and New York: Plenum Press.

46. World Health Organization. (2004). Global strategic framework for integrated vector management. Available from: https://apps.who.int/iris/handle/10665/68624.

47. Brogdon, W. G. and McAllister, J. C. (1998). Insecticide resistance and vector control. *Emerging Infectious Diseases*. **4** (4): 605-13.

48. World Health Organization. (2017). Global vector control response 2017-2030. Available from: https://www.who.int/vector-control/publications/global-control-response/en/.

49. Lederberg, J. and McCray, A. T. (2001). 'Ome Sweet 'Omics - A Genealogical Treasury of Words. *The Scientist*. **15** (7): 8.

50. Moran, N. A., McCutcheon, J. P. and Nakabachi, A. (2008). Genomics and evolution of heritable bacterial symbionts. *Annual Review of Genetics*. **42**: 165-90.

51. Schnepf, E., Crickmore, N., Van Rie, J., Lereclus, D., Baum, J., Feitelson, J.*, et al.* (1998). *Bacillus thuringiensis* and its pesticidal crystal proteins. *Microbiology and Molecular Biology Reviews*. **62** (3): 775-806.

52. Cerstiaens, A., Verleyen, P., Van Rie, J., Van Kerkhove, E., Schwartz, J.-L., Laprade, R.*, et al.* (2001). Effect of *Bacillus thuringiensis* Cry1 Toxins in Insect Hemolymph and Their Neurotoxicity in Brain Cells of *Lymantria dispar*. *Applied Environmental Microbiology*. **67** (9): 3923-7.

53. Zwick, M. E., Joseph, S. J., Didelot, X., Chen, P. E., Bishop-Lilly, K. A., Stewart, A. C.*, et al.* (2012). Genomic characterization of the *Bacillus cereus* sensu lato species: backdrop to the evolution of *Bacillus anthracis*. *Genome Research*. **22** (8): 1512-24.

54. Cabral, C., Cherqui, A., Pereira, A. and Simoes, N. (2004). Purification and characterization of two distinct metalloproteases secreted by the entomopathogenic bacterium *Photorhabdus sp.* strain Az29. *Applied Environmental Microbiology*. **70** (7): 3831-8.

55. Maudlin, I. and Ellis, D. S. (1985). Association between intracellular rickettsial-like infections of midgut cells and susceptibility to trypanosome infection in *Glossina spp*. *Zeitschrift für Parasitenkunde*. **71** (5): 683-7.

56. Schaub, G. A. (1992). The Effects of Trypanosomatids on Insects. Ed: Baker JR, Muller R, In: *Advances in Parasitology*. vol. 31.

57. Welburn, S. C. and Maudlin, I. (1999). Tsetse-trypanosome interactions: rites of passage. *Parasitology Today*. **15** (10): 399-403.

58. King, E. G., Bell, J. V. and Martin, D. F. (1975). Control of the bacterium *Serratia marcescens* in an insect host-parasite rearing program. *Journal of Invertebrate Pathology*. **26** (1): 35-40.

59. Lamelas, A., Pérez-Brocal, V., Gómez-Valero, L., Gosalbes, M. J., Moya, A. and Latorre, A. (2008). Evolution of the secondary symbiont "*Candidatus serratia*

symbiotica" in aphid species of the subfamily lachninae. *Applied and Environmental Microbiology*. **74** (13): 4236-40.

60. Toh, H., Weiss, B. L., Perkin, S. A. H., Yamashita, A., Oshima, K., Hattori, M.*, et al.* (2006). Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Research*. **16** (2): 149-56.

61. Lamelas, A., Gosalbes, M. J., Manzano-Marín, A., Peretó, J., Moya, A. and Latorre, A. (2011). *Serratia symbiotica* from the aphid *Cinara cedri*: a missing link from facultative to obligate insect endosymbiont. *PLOS Genetics*. **7** (11): e1002357.

62. Lefoulon, E., Bain, O., Makepeace, B. L., d'Haese, C., Uni, S., Martin, C.*, et al.* (2016). Breakdown of coevolution between symbiotic bacteria *Wolbachia* and their filarial hosts. *PeerJ*. **4**: e1840.

63. Weinert, L. A., Araujo-Jnr, E. V., Ahmed, M. Z. and Welch, J. J. (2015). The incidence of bacterial endosymbionts in terrestrial arthropods. *Proceedings of the Royal Society B: Biological Sciences*. **282** (1807): 1-6.

64. Foster, J., Ganatra, M., Kamal, I., Ware, J., Makarova, K., Ivanova, N.*, et al.* (2005). The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLOS Biology*. **3** (4): e121.

65. Kaur, R., Shropshire, J. D., Cross, K. L., Leigh, B., Mansueto, A. J., Stewart, V.*, et al.* (2021). Living in the endosymbiotic world of *Wolbachia*: A centennial review. *Cell Host & Microbe*. **29** (6): 879-93.

66. Augustinos, A. A., Santos-Garcia, D., Dionyssopoulou, E., Moreira, M., Papapanagiotou, A., Scarvelakis, M.*, et al.* (2011). Detection and characterization of *Wolbachia* infections in natural populations of aphids: is the hidden diversity fully unraveled? *PLOS ONE*. **6** (12): e28695.

67. Baldo, L., Prendini, L., Corthals, A. and Werren, J. H. (2007). *Wolbachia* are present in Southern African scorpions and cluster with supergroup F. *Current Microbiology*. **55** (5): 367-73.

68. Bing, X. L., Xia, W. Q., Gui, J. D., Yan, G. H., Wang, X. W. and Liu, S. S. (2014). Diversity and evolution of the *Wolbachia* endosymbionts of *Bemisia* (Hemiptera: Aleyrodidae) whiteflies. *Ecology and Evolution*. **4** (13): 2714-37.

69. Bordenstein, S. and Rosengaus, R. B. (2005). Discovery of a novel *Wolbachia* supergroup in Isoptera. *Current Microbiology*. **51** (6): 393-8.

70. Glowska, E., Dragun-Damian, A., Dabert, M. and Gerth, M. (2015). New *Wolbachia* supergroups detected in quill mites (Acari: Syringophilidae). *Infection, Genetics and Evolution*. **30**: 140-6.

71. Haegeman, A., Vanholme, B., Jacob, J., Vandekerckhove, T., Claeys, M., Borgonie, G.*, et al.* (2009). An endosymbiotic bacterium in a plant-parasitic nematode: member of a new *Wolbachia* supergroup. *International Journal for Parasitology*. **39** (9): 1045-54.

72. Lo, N., Casiraghi, M., Salati, E., Bazzocchi, C. and Bandi, C. (2002). How many *Wolbachia* supergroups exist? *Molecular Biology and Evolution*. **19** (3): 341-6.

73. Ros, V. I., Fleming, V. M., Feil, E. J. and Breeuwer, J. A. (2009). How diverse is the genus *Wolbachia*? Multiple-gene sequencing reveals a putatively new *Wolbachia* supergroup recovered from spider mites (*Acari*: Tetranychidae). *Applied Environmental Microbiology*. **75** (4): 1036-43.

74. Baldo, L., Hotopp, J. C. D., Jolley, K. A., Bordenstein, S. R., Biber, S. A., Choudhury, R. R.*, et al.* (2006). Multilocus sequence typing system for the endosymbiont *Wolbachia pipientis*. *Applied and Environmental Microbiology*. **72** (11): 7098-110.

75. Yen, J. H. and Barr, A. R. (1971). New hypothesis of the cause of cytoplasmic incompatibility in *Culex pipiens L*. *Nature*. **232** (5313): 657-8.

76. Perrot-Minnot, M.-J., Guo, L. R. and Werren, J. H. (1996). Single and double infections with *Wolbachia* in the parasitic wasp *Nasonia vitripennis* effects on compatibility. *Genetics*. **143** (2): 961-72.

77. Shropshire, J. D., Leigh, B. and Bordenstein, S. R. (2020). Symbiont-mediated cytoplasmic incompatibility: what have we learned in 50 years? *eLife*. **9**: e61989.

78. Duron, O. and Weill, M. (2006). *Wolbachia* infection influences the development of *Culex pipiens* embryo in incompatible crosses. *Heredity*. **96** (6): 493-500.

79. Narita, S., Kageyama, D., Nomura, M. and Fukatsu, T. (2007). Unexpected mechanism of symbiont-induced reversal of insect sex: feminizing *Wolbachia* continuously acts on the butterfly *Eurema hecabe* during larval development. *Applied and Environmental Microbiology*. **73** (13): 4332-41.

80. Hurst, G. D., Jiggins, F. M., Hinrich Graf von der Schulenburg, J., Bertrand, D., West, S. A., Goriacheva, I. I.*, et al.* (1999). Male–killing *Wolbachia* in two species of insect. *Proceedings of the Royal Society of London Series B: Biological Sciences*. **266** (1420): 735-40.

81. Boivin, T., Henri, H., Vavre, F., Gidoin, C., Veber, P., Candau, J. N.*, et al.* (2014). Epidemiology of asexuality induced by the endosymbiotic *Wolbachia* across phytophagous wasp species: host plant specialization matters. *Molecular Ecology*. **23** (9): 2362-75.

82. Meyer, J. M. and Hoy, M. A. (2007). *Wolbachia*-associated thelytoky in *Diaphorencyrtus aligarhensis* (Hymenoptera: Encyrtidae), a parasitoid of the Asian citrus psyllid. *Florida entomologist*. **90** (4): 776-9.

83. Zchori-Fein, E., Roush, R. and Hunter, M. S. (1992). Male production induced by antibiotic treatment in *Encarsia formosa* (Hymenoptera: Aphelinidae), an asexual species. *Experientia*. **48** (1): 102-5.

84. Moreira, L. A., Iturbe-Ormaetxe, I., Jeffery, J. A., Lu, G., Pyke, A. T., Hedges, L. M.*, et al.* (2009). A *Wolbachia* symbiont in *Aedes aegypti* limits infection with dengue, chikungunya, and *Plasmodium*. *Cell*. **139** (7): 1268-78.

85. Himler, A. G., Adachi-Hagimori, T., Bergen, J. E., Kozuch, A., Kelly, S. E., Tabashnik, B. E.*, et al.* (2011). Rapid spread of a bacterial symbiont in an invasive whitefly is driven by fitness benefits and female bias. *Science (New York, NY)*. **332** (6026): 254-6.

86. Bockoven, A. A., Bondy, E. C., Flores, M. J., Kelly, S. E., Ravenscraft, A. M. and Hunter, M. S. (2020). What goes up might come down: the spectacular spread of an endosymbiont is followed by its decline a decade later. *Microbial Ecology*. **79** (2): 482-94.

87. Wei, G., Lai, Y., Wang, G., Chen, H., Li, F. and Wang, S. (2017). Insect pathogenic fungus interacts with the gut microbiota to accelerate mosquito mortality. *Proceedings of the National Academy of Sciences*. **114** (23): 5994-9.

88. Jefferies, D., Livesey, J. and Molyneux, D. (1986). Fluid mechanics of bloodmeal uptake by *Leishmania*-infected sandflies. *Acta Tropica*. **43** (1): 43-53.

89. Louradour, I., Monteiro, C. C., Inbar, E., Ghosh, K., Merkhofer, R., Lawyer, P.*, et al.* (2017). The midgut microbiota plays an essential role in sand fly vector competence for *Leishmania major*. *Cellular Microbiology*. **19** (10): e12755.

90. Hibbing, M. E., Fuqua, C., Parsek, M. R. and Peterson, S. B. (2010). Bacterial competition: surviving and thriving in the microbial jungle. *Nature Reviews Microbiology*. **8** (1): 15-25.

91. Sapp, J. (1994). *Evolution by Association: A History of Symbiosis*: Oxford University Press.

92. López-García, P., Eme, L. and Moreira, D. (2017). Symbiosis in eukaryotic evolution. *Journal of Theoretical Biology*. **434**: 20-33.

93. Douglas, A. E. (2011). Lessons from studying insect symbioses. *Cell Host & Microbe*. **10** (4): 359-67.

94. Zientz, E., Dandekar, T. and Gross, R. (2004). Metabolic interdependence of obligate intracellular bacteria and their insect hosts. *Microbiology and Molecular Biology Reviews*. **68** (4): 745-70.

95. Brecher, G. and Wigglesworth, V. B. (1944). The transmission of *Actinomyces rhodnii* Erikson in *Rhodnius prolixus* stål (hemiptera) and its influence on the growth of the host. *Parasitology*. **35** (4): 220-4.

96. Moran, N. A., Munson, M. A., Baumann, P. and Ishikawa, H. (1993). A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. *Proceedings of the Royal Society of London B: Biological Sciences*. **253** (1337): 167-71.

97. Hansen, A. K. and Moran, N. A. (2011). Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. *Proceedings of the National Academy of Sciences*. **108** (7): 2849-54.

98. Douglas, A. E. (1998). Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria *Buchnera*. *Annual review of entomology*. **43** (1): 17-37.

99. Douglas, A. E. (1989). Mycetocyte symbiosis in insects. *Biological Reviews*. **64** (4): 409-34.

100. Moran, N. A. and Telang, A. (1998). Bacteriocyte-associated symbionts of insects. *BioScience*. **48** (4): 295-304.

101. Tsuchida, T., Koga, R., Shibao, H., Matsumoto, T. and Fukatsu, T. (2002). Diversity and geographic distribution of secondary endosymbiotic bacteria in natural populations of the pea aphid, *Acyrthosiphon pisum*. *Molecular Ecology*. **11** (10): 2123-35.

102. Dennis, J. W., Durkin, S. M., Horsley Downie, J. E., Hamill, L. C., Anderson, N. E. and MacLeod, E. T. (2014). *Sodalis glossinidius* prevalence and trypanosome presence in tsetse from Luambe National Park, Zambia. *Parasites & Vectors*. **7** (1): 378-88.

103. Dobson, S. L., Bourtzis, K., Braig, H. R., Jones, B. F., Zhou, W., Rousset, F.*, et al.* (1999). *Wolbachia* infections are distributed throughout insect somatic and germ line tissues. *Insect Biochemistry and Molecular Biology*. **29** (2): 153-60.

104. Cheng, Q. and Aksoy, S. (1999). Tissue tropism, transmission and expression of foreign genes *in vivo* in midgut symbionts of tsetse flies. *Insect Molecular Biology*. **8** (1): 125-32.

105. Dale, C., Young, S. A., Haydon, D. T. and Welburn, S. C. (2001). The insect endosymbiont *Sodalis glossinidius* utilizes a type III secretion system for cell invasion. *Proceedings of the National Academy of Sciences*. **98** (4): 1883-8.

106. Brune, A. (2014). Symbiotic digestion of lignocellulose in termite guts. *Nature Reviews Microbiology*. **12** (3): 168-80.

107. Montllor, C. B., Maxmen, A. and Purcell, A. H. (2002). Facultative bacterial endosymbionts benefit pea aphids *Acyrthosiphon pisum* under heat stress. *Ecological Entomology*. **27** (2): 189-95.

108. Belda, E., Moya, A., Bentley, S. and Silva, F. J. (2010). Mobile genetic element proliferation and gene inactivation impact over the gemone structure and metabolic capabilities of *Sodalis glossinidius*, the secondary endosymbiont of tsetse flies. *BMC genomics*. **11** (449): 1-17.

109. McCutcheon, J. P. and Von Dohlen, C. D. (2011). An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Current Biology*. **21** (16): 1366-72.

110. Husnik, F., Nikoh, N., Koga, R., Ross, L., Duncan, R. P., Fujie, M.*, et al.* (2013). Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*. **153** (7): 1567-78.

111. López-Madrigal, S., Latorre, A., Porcar, M., Moya, A. and Gil, R. (2013). Mealybugs nested endosymbiosis: going into the 'matryoshka' system in *Planococcus citri* in depth. *BMC Microbiology*. **13** (1): 1-12.

112. Lo, W.-S., Huang, Y.-Y. and Kuo, C.-H. (2016). Winding paths to simplicity: genome evolution in facultative insect symbionts. *FEMS Microbiology Reviews*. **40** (6): 855-74.

113. Oliver, K. M., Russell, J. A., Moran, N. A. and Hunter, M. S. (2003). Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. *Proceedings of the National Academy of Sciences*. **100** (4): 1803-7.

114. Ferrari, J., Darby, A. C., Daniell, T. J., Godfray, H. C. J. and Douglas, A. E. (2004). Linking the bacterial community in pea aphids with host-plant use and natural enemy resistance. *Ecological Entomology*. **29** (1): 60-5.

115. Bensadia, F., Boudreault, S., Guay, J.-F., Michaud, D. and Cloutier, C. (2006). Aphid clonal resistance to a parasitoid fails under heat stress. *Journal of Insect Physiology*. **52** (2): 146-57.

116. van der Wilk, F., Dullemans, A. M., Verbeek, M. and van den Heuvel, J. F. (1999). Isolation and characterization of APSE-1, a bacteriophage infecting the secondary endosymbiont of *Acyrthosiphon pisum*. *Virology*. **262** (1): 104-13.

117. Degnan, P. H. and Moran, N. A. (2008). Evolutionary genetics of a defensive facultative symbiont of insects: exchange of toxin-encoding bacteriophage. *Molecular Ecology*. **17** (3): 916-29.

118. Moran, N. A., Degnan, P. H., Santos, S. R., Dunbar, H. E. and Ochman, H. (2005). The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. *Proceedings of the National Academy of Sciences USA*. **102** (47): 16919-26.

119. Su, Q., Oliver, K. M., Xie, W., Wu, Q., Wang, S. and Zhang, Y. (2015). The whitefly-associated facultative symbiont *Hamiltonella defensa* suppresses induced plant defences in tomato. *Functional Ecology*. **29** (8): 1007-18.

120. Sanchis, V. (2011). From microbial sprays to insect-resistant transgenic plants: history of the biospesticide *Bacillus thuringiensis*. A review. *Agronomy for Sustainable Development*. **31** (1): 217-31.

121. Boisvert, M. and Boisvert, J. (2000). Effects of *Bacillus thuringiensis* var. israelensis on Target and Nontarget Organisms: A Review of Laboratory and Field Experiments. *Biocontrol Science and Technology*. **10** (5): 517-61.

122. Arora, A. K. and Douglas, A. E. (2017). Hype or opportunity? Using microbial symbionts in novel strategies for insect pest control. *Journal of Insect Physiology*. **103** (2017): 10-7.

123. Hedges, L. M., Brownlie, J. C., O'Neill, S. L. and Johnson, K. N. (2008). *Wolbachia* and virus protection in insects. *Science (New York, NY)*. **322** (5902): 702.

124. Teixeira, L., Ferreira, Á. and Ashburner, M. (2008). The Bacterial Symbiont *Wolbachia* Induces Resistance to RNA Viral Infections in *Drosophila melanogaster*. *PLOS Biology*. **6** (12): e1000002.

125. Walker, T., Johnson, P. H., Moreira, L. A., Iturbe-Ormaetxe, I., Frentiu, F. D., McMeniman, C. J.*, et al.* (2011). The *w*Mel *Wolbachia* strain blocks dengue and invades caged *Aedes aegypti* populations. *Nature*. **476** (7361): 450-3.

126. Hughes, G. L., Koga, R., Xue, P., Fukatsu, T. and Rasgon, J. L. (2011). *Wolbachia* infections are virulent and inhibit the human malaria parasite *Plasmodium falciparum* in *Anopheles gambiae*. *PLOS Pathogens*. **7** (5): e1002043.

127. Kambris, Z., Blagborough, A. M., Pinto, S. B., Blagrove, M. S., Godfray, H. C. J., Sinden, R. E.*, et al.* (2010). *Wolbachia* stimulates immune gene expression and inhibits *Plasmodium* development in *Anopheles gambiae*. *PLOS Pathogens*. **6** (10): e1001143.

128. Caragata, E. P., Rancès, E., Hedges, L. M., Gofton, A. W., Johnson, K. N., O'Neill, S. L.*, et al.* (2013). Dietary cholesterol modulates pathogen blocking by *Wolbachia*. *PLOS Pathogens*. **9** (6): e1003459.

129. Rancès, E., Ye, Y. H., Woolfit, M., McGraw, E. A. and O'Neill, S. L. (2012). The relative importance of innate immune priming in *Wolbachia*-mediated dengue interference. *PLOS Pathogens*. **8** (2): e1002548.

130. Rancès, E., Johnson, T. K., Popovici, J., Iturbe-Ormaetxe, I., Zakir, T., Warr, C. G.*, et al.* (2013). The toll and Imd pathways are not required for *Wolbachia*-mediated dengue virus interference. *Journal of virology*. **87** (21): 11945-9.

131. Bian, G., Xu, Y., Lu, P., Xie, Y. and Xi, Z. (2010). The endosymbiotic bacterium *Wolbachia* induces resistance to dengue virus in *Aedes aegypti*. *PLOS Pathogens*. **6** (4): e1000833.

132. Rainey, S. M., Martinez, J., McFarlane, M., Juneja, P., Sarkies, P., Lulla, A.*, et al.* (2016). *Wolbachia* blocks viral genome replication early in infection without a transcriptional response by the endosymbiont or host small RNA pathways. *PLOS Pathogens*. **12** (4): e1005536.

133. Schultz, M. J., Connor, J. H. and Frydman, H. M. (2018). Group B *Wolbachia* Strain-Dependent Inhibition of Arboviruses. *DNA and Cell Biology*. **37** (1): 2-6.

134. Lu, P., Sun, Q., Fu, P., Li, K., Liang, X. and Xi, Z. (2020). *Wolbachia* inhibits binding of dengue and Zika viruses to mosquito cells. *Frontiers in Microbiology*. **11** (1750): 1-12.

135. O'Neill, S. L. (2018). The Use of *Wolbachia* by the World Mosquito Program to Interrupt Transmission of *Aedes aegypti* Transmitted Viruses. Ed: Hilgenfeld R, Vasudeva S, In: *Dengue and Zika: Control and Antiviral Treatment Strategies*. vol.

136. Kittayapong, P., Baisley, K. J., Baimai, V. and O'Neill, S. L. (2000). Distribution and diversity of *Wolbachia* infections in Southeast Asian mosquitoes (Diptera: Culicidae). *Journal of Medical Entomology*. **37** (3): 340-5.

137. Flores, H. A. and O'Neill, S. L. (2018). Controlling vector-borne diseases by releasing modified mosquitoes. *Nature Reviews Microbiology*. **16** (8): 508-18.

138. Joubert, D. A., Walker, T., Carrington, L. B., De Bruyne, J. T., Kien, D. H. T., Hoang, N. L. T.*, et al.* (2016). Establishment of a *Wolbachia* superinfection in *Aedes aegypti* mosquitoes as a potential approach for future resistance management. *PLOS Pathogens*. **12** (2): e1005434.

139. Nazni, W. A., Hoffmann, A. A., NoorAfizah, A., Cheong, Y. L., Mancini, M. V., Golding, N.*, et al.* (2019). Establishment of *Wolbachia* strain *w*AlbB in Malaysian populations of *Aedes aegypti* for dengue control. *Current Biology*. **29** (24): 4241-8.

140. Tantowijoyo, W., Andari, B., Arguni, E., Budiwati, N., Nurhayati, I., Fitriana, I.*, et al.* (2020). Stable establishment of *w*Mel *Wolbachia* in *Aedes aegypti* populations in Yogyakarta, Indonesia. *PLOS Neglected Tropical Diseases*. **14** (4): e0008157.

141. Xi, Z., Dean, J. L., Khoo, C. and Dobson, S. L. (2005). Generation of a novel *Wolbachia* infection in *Aedes albopictus* (Asian tiger mosquito) via embryonic microinjection. *Insect Biochemistry and Molecular Biology*. **35** (8): 903-10.

142. Schmidt, T. L., Barton, N. H., Rašić, G., Turley, A. P., Montgomery, B. L., Iturbe-Ormaetxe, I.*, et al.* (2017). Local introduction and heterogeneous spatial spread of dengue-suppressing *Wolbachia* through an urban population of *Aedes aegypti*. *PLOS Biology*. **15** (5): e2001894.

143. Caputo, B., Moretti, R., Manica, M., Serini, P., Lampazzi, E., Bonanni, M.*, et al.* (2020). A bacterium against the tiger: preliminary evidence of fertility reduction after release of *Aedes albopictus* males with manipulated *Wolbachia* infection in an Italian urban area. *Pest management science*. **76** (4): 1324-32.

144. Mains, J. W., Brelsfoard, C. L., Rose, R. I. and Dobson, S. L. (2016). Female adult *Aedes albopictus* suppression by *Wolbachia*-infected male mosquitoes. *Scientific Reports*. **6** (1): 1-7.

145. Mains, J. W., Kelly, P. H., Dobson, K. L., Petrie, W. D. and Dobson, S. L. (2019). Localized control of *Aedes aegypti* (Diptera: Culicidae) in Miami, FL, via inundative releases of *Wolbachia*-infected male mosquitoes. *Journal of Medical Entomology*. **56** (5): 1296-303.

146. Puggioli, A., Calvitti, M., Moretti, R. and Bellini, R. (2016). *w*Pip *Wolbachia* contribution to *Aedes albopictus* SIT performance: advantages under intensive rearing. *Acta Tropica*. **164**: 473-81.

147. Zheng, X., Zhang, D., Li, Y., Yang, C., Wu, Y., Liang, X.*, et al.* (2019). Incompatible and sterile insect techniques combined eliminate mosquitoes. *Nature*. **572** (7767): 56-61.

148. Ross, P. A., Wiwatanaratanabutr, I., Axford, J. K., White, V. L., Endersby-Harshman, N. M. and Hoffmann, A. A. (2017). *Wolbachia* infections in *Aedes aegypti* differ markedly in their response to cyclical heat stress. *PLOS Pathogens*. **13** (1): e1006006.

149. Amuzu, H. E., Tsyganov, K., Koh, C., Herbert, R. I., Powell, D. R. and McGraw, E. A. (2018). *Wolbachia* enhances insect-specific flavivirus infection in *Aedes aegypti* mosquitoes. *Ecology and Evolution*. **8** (11): 5441-54.

150. Martinez, J., Ok, S., Smith, S., Snoeck, K., Day, J. P. and Jiggins, F. M. (2015). Should symbionts be nice or selfish? Antiviral effects of *Wolbachia* are costly but reproductive parasitism is not. *PLOS Pathogens*. **11** (7): e1005021.

151. Kamtchum-Tatuene, J., Makepeace, B. L., Benjamin, L., Baylis, M. and Solomon, T. (2017). The potential role of *Wolbachia* in controlling the transmission of emerging human arboviral infections. *Current Opinion in Infectious Diseases*. **30** (1): 108-16.

152. Beard, C. B., Durvasula, R. V. and Richards, F. F. (1998). Bacterial symbiosis in arthropods and the control of disease transmission. *Emerging Infectious Diseases*. **4** (4): 581-91.

153. Beard, C. B., Cordon-Rosales, C. and Durvasula, R. V. (2002). Bacterial symbionts of the triatominae and their potential use in control of Chagas disease transmission. *Annual review of entomology*. **47** (1): 123-41.

154. Beard, C. B., Mason, P. W., Aksoy, S., Tesh, R. B. and Richards, F. F. (1992). Transformation of an insect symbiont and expression of a foreign gene in the Chagas' disease vector *Rhodnius prolixus*. *The American Journal of Tropical Medicine and Hygiene*. **46** (2): 195-200.

155. Dotson, E. M., Plikaytis, B., Shinnick, T. M., Durvasula, R. V. and Beard, C. B. (2003). Transformation of *Rhodococcus rhodnii*, a symbiont of the Chagas disease vector *Rhodnius prolixus*, with integrative elements of the L1 mycobacteriophage. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*. **3** (2): 103-9.

156. Durvasula, R. V., Gumbs, A., Panackal, A., Kruglov, O., Aksoy, S., Merrifield, R. B.*, et al.* (1997). Prevention of insect-borne disease: an approach using transgenic symbiotic bacteria. *Proceedings of the National Academy of Sciences USA*. **94** (7): 3274-8.

157. Durvasula, R. V., Kroger, A., Goodwin, M., Panackal, A., Kruglov, O., Taneja, J.*, et al.* (1999). Strategy for introduction of foreign genes into field populations of Chagas disease vectors. *Annals of the Entomological Society of America*. **92** (6): 937-43.

158. Bongio, N. J. and Lampe, D. J. (2015). Inhibition of *Plasmodium berghei* Development in Mosquitoes by Effector Proteins Secreted from *Asaia sp.* Bacteria Using a Novel Native Secretion Signal. *PLOS ONE*. **10** (12): e0143541.

159. Wang, S., Ghosh, A. K., Bongio, N., Stebbings, K. A., Lampe, D. J. and Jacobs-Lorena, M. (2012). Fighting malaria with engineered symbiotic bacteria from vector mosquitoes. *Proceedings of the National Academy of Sciences*. **109** (31): 12734-9.

160. Favia, G., Ricci, I., Damiani, C., Raddadi, N., Crotti, E., Marzorati, M.*, et al.* (2007). Bacteria of the genus *Asaia* stably associate with *Anopheles stephensi*, an Asian malarial mosquito vector. *Proceedings of the National Academy of Sciences*. **104** (21): 9047-51.

161. Whitten, M. M. A., Facey, P. D., Sol, R. D., Fernández-Martínez, L. T., Evans, M. C., Mitchell, J. J.*, et al.* (2016). Symbiont-mediated RNA interference in insects. *Proceedings of the Royal Society B-Biological Sciences*. **283** (1825): 1-9.

162. Zhao, R., Han, R., Qiu, X., Yan, X., Cao, L. and Liu, X. (2008). Cloning and heterologous expression of insecticidal-protein-encoding genes from *Photorhabdus luminescens* TT01 in *Enterobacter cloacae* for termite control. *Applied and Environmental Microbiology*. **74** (23): 7219-26.

163. Mandell, D. J., Lajoie, M. J., Mee, M. T., Tekeuchi, R., Kuznetsov, G., Norville, J. E.*, et al.* (2015). Biocontainment of genetically modified organisms by synthetic protein design. *Nature*. **518** (7537): 55-60.

164. Wilkinson, T. L. (1998). The elimination of intracellular microorganisms from insects: an analysis of antibiotic-treatment in the pea aphid (*Acyrthosiphon pisum*). *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*. **119** (4): 871-81.

165. Tancos, K. A., Villani, S., Kuehne, S., Borejsza-Wysocka, E., Breth, D., Carol, J.*, et al.* (2015). Prevalence of streptomycin-resistant *Erwinia amylovora* in New York apple orchards. *Plant Disease*. **100** (4): 802-9.

166. Douglas, A. E. (2015). Multiorganismal insects: diversity and function of resident microorganisms. *Annual review of entomology*. **60**: 17-34.

167. Login, F. H., Balmand, S., Vallier, A., Vincent-Monégat, C., Vigneron, A., Weiss-Gayet, M.*, et al.* (2011). Antimicrobial peptides keep insect endosymbionts under control. *Science (New York, NY)*. **334** (6054): 362-5.

168. Husseneder, C., Donaldson, J. R. and Foil, L. D. (2016). Genetically engineered yeast expressing a lytic peptide from bee venom (Melittin) kills symbiotic Protozoa in the gut of formosan subterranean termites. *PLOS ONE*. **11** (3): e0151675.

169. Turner, J. D., Sharma, R., Al Jayoussi, G., Tyrer, H. E., Gamble, J., Hayward, L.*, et al.* (2017). Albendazole and antibiotics synergize to deliver short-course anti-*Wolbachia* curative treatments in preclinical models of filariasis. *Proceedings of the National Academy of Sciences*. **114** (45): E9712.

170. van den Bossche, P. (2001). Some general aspects of the distribution and epidemiology of bovine trypanosomosis in southern Africa. *International Journal for Parasitology*. **31** (5): 592-8.

171. The Center for Food Security and Public Health. (2018). African Animal Trypanosomiasis. Iowa State University, Available from: http://www.cfsph.iastate.edu/Factsheets/pdfs/trypanosomiasis_african.pdf.

172. Franco, J. R., Simarro, P. P., Diarra, A. and Jannin, J. G. (2014). Epidemiology of human African trypanosomiasis. *Clinical Epidemiology*. **6**: 257-75.

173. Pérez-Molina, J. A. and Molina, I. (2018). Chagas disease. *Lancet (London, England)*. **391** (10115): 82-94.

174. World Health Organization. (2020). Trypanosomiasis, human African (sleeping sickness). Geneva. Available from: https://www.who.int/en/news-room/fact-sheets/detail/trypanosomiasis-human-african-(sleeping-sickness).

175. Jamonneau, V., Ilboudo, H., Kaboré, J., Kaba, D., Koffi, M., Solano, P.*, et al.* (2012). Untreated human infections by *Trypanosoma brucei gambiense* are not 100% fatal. *PLOS Neglected Tropical Diseases*. **6** (6): e1691-e.

176. Blum, J., Schmid, C. and Burri, C. (2006). Clinical aspects of 2541 patients with second stage human African trypanosomiasis. *Acta Tropica*. **97** (1): 55-64.

177. Checchi, F., Filipe, J. A., Haydon, D. T., Chandramohan, D. and Chappuis, F. (2008). Estimates of the duration of the early and late stage of gambiense sleeping sickness. *BMC Infectious Diseases*. **8** (1): 1-10.

178. Simo, G., Asonganyi, T., Nkinin, S. W., Njiokou, F. and Herder, S. (2006). High prevalence of *Trypanosoma brucei gambiense* group 1 in pigs from the Fontem sleeping sickness focus in Cameroon. *Veterinary Parasitology*. **139** (1-3): 57-66.

179. Njiokou, F., Laveissiere, C., Simo, G., Nkinin, S., Grebaut, P., Cuny, G.*, et al.* (2006). Wild fauna as a probable animal reservoir for *Trypanosoma brucei gambiense* in Cameroon. *Infection, Genetics & Evolution*. **6** (2): 147-53.

180. Balyeidhusa, A. S., Kironde, F. A. and Enyaru, J. C. (2012). Apparent lack of a domestic animal reservoir in Gambiense sleeping sickness in northwest Uganda. *Veterinary Parasitology*. **187** (1-2): 157-67.

181. Odiit, M., Kansiime, F. and Enyaru, J. (1997). Duration of symptoms and case fatality of sleeping sickness caused by *Trypanosoma brucei rhodesiense* in Tororo, Uganda. *East African medical journal*. **74** (12): 792-5.

182. Onyango, R., Van Hoeve, K. and De Raadt, P. (1966). The epidemiology of *Trypanosoma rhodesiense* sleeping sickness in alego location, Central Nyanza, Kenya I. Evidence that cattle may act as reservoir hosts of trypanosomes infective to man. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. **60** (2): 175-82.

183. Fèvre, E. M., Coleman, P. G., Odiit, M., Magona, J. W., Welburn, S. C. and Woolhouse, M. E. J. (2001). The origins of a new *Trypanosoma brucei rhodesiense* sleeping sickness outbreak in eastern Uganda. *The Lancet*. **358** (9282): 625-8.

184. Mugnier, M. R., Stebbins, C. E. and Papavasiliou, F. N. (2016). Masters of Disguise: Antigenic Variation and the VSG Coat in *Trypanosoma brucei*. *PLOS Pathogens*. **12** (9): e1005784.

185. Silva Pereira, S., de Almeida Castilho Neto, K. J. G., Duffy, C. W., Richards, P., Noyes, H., Ogugo, M.*, et al.* (2020). Variant antigen diversity in *Trypanosoma vivax* is not driven by recombination. *Nature Communications*. **11** (844): 1-13.

186. Autheman, D., Crosnier, C., Clare, S., Goulding, D. A., Brandt, C., Harcourt, K.*, et al.* (2021). An invariant *Trypanosoma vivax* vaccine antigen induces protective immunity. *Nature*. **595** (7865): 96-100.

187. Lambrecht, F. L. (1985). Trypanosomes and Hominid Evolution. *BioScience*. **35** (10): 640-6.

188. Steverding, D. (2008). The history of African trypanosomiasis. *Parasites & Vectors*. **1** (3): 1-8.

189. Lyons, M. (2002). *The Colonial Disease: A Social History of Sleeping Sickness in Northern Zaire, 1900-1940*. Cambridge: Cambridge University Press.

190. World Health Organization (2017). Trypanosomiasis, human African (sleeping sickness). *WHO*.

191. Yaro, M., Munyard, K. A., Stear, M. J. and Groth, D. M. (2016). Combatting African animal trypanosomiasis (AAT) in livestock: the potential role of trypanotolerance. *Veterinary Parasitology*. **225**: 43-52.

192. Smetko, A., Soudre, A., Silbermayr, K., Müller, S., Brem, G., Hanotte, O*., et al.* (2015). Trypanosomosis: potential driver of selection in African cattle. *Frontiers in Genetics*. **6** (137): 1-8.

193. Muhanguzi, D., Picozzi, K., Hattendorf, J., Thrusfield, M., Kabasa, J. D., Waiswa, C*., et al.* (2014). The burden and spatial distribution of bovine African trypanosomes in small holder crop-livestock production systems in Tororo District, south-eastern Uganda. *Parasites & Vectors*. **7** (603): 1-10.

194. Swallow, B. M. (2000). Impact of trypanosomiasis on African agriculture. International Livestock Research Institute, Nairobi, Kenya, Available from:

195. Wells, E. A. (1972). The importance of mechanical transmission in the epidemiology of nagana: a review. *Tropical Animal Health and Production*. **4** (2): 74-89.

196. Desquesnes, M. and Dia, M. L. (2003). *Trypanosoma vivax*: mechanical transmission in cattle by one of the most common African tabanids, *Atylotus agrestis*. *Experimental parasitology*. **103** (1): 35-43.

197. Coura, J. R. and Viñas, P. A. (2010). Chagas disease: a new worldwide challenge. *Nature*. **465** (7301): S6-S7.

198. World Health Organization. (2013). Control and surveillance of human African trypanosomiasis: report of a WHO expert committee. Available from: https://apps.who.int/iris/handle/10665/95732.

199. Wamwiri, F. N. and Changasi, R. E. (2016). Tsetse Flies (*Glossina*) as Vectors of Human African Trypanosomiasis: A Review. *BioMed Research International*. **2016** (6201350): 1-8.

200. Latif, A. A., Ntantiso, L. and De Beer, C. (2019). African animal trypanosomosis (nagana) in northern KwaZulu-Natal, South Africa: Strategic treatment of cattle on a farm in endemic area. *The Onderstepoort Journal of Veterinary Research*. **86** (1): e1-e6.

201. Baldry, D. A. T. (1980). Local distribution and ecology of *Glossina palpalis* and *G. tachinoides* in forest foci of West African human trypanosomiasis, with special reference to associations between peri-domestic tsetse and their hosts. *Insect Science and Its Application*. **1** (1): 85-93.

202. Garcia, H. A., Rodrigues, C. M. F., Rodrigues, A. C., Pereira, D. L., Pereira, C. L., Camargo, E. P.*, et al.* (2018). Remarkable richness of trypanosomes in tsetse flies (*Glossina morsitans morsitans* and *Glossina pallidipes*) from the Gorongosa National Park and Niassa National Reserve of Mozambique revealed by fluorescent fragment length barcoding (FFLB). *Infection, Genetics and Evolution*. **63**: 370-9.

203. Auty, H. K., Picozzi, K., Malele, I., Torr, S. J., Cleaveland, S. and Welburn, S. (2012). Using molecular data for epidemiological inference: assessing the prevalence of *Trypanosoma brucei rhodesiense* in tsetse in Serengeti, Tanzania. *PLOS Neglected Tropical Diseases*. **6** (1): e1501.

204. McNamara, J. J., Laveissière, C. and Masiga, D. K. (1995). Multiple trypanosome infections in wild tsetse in Côte d'Ivoire detected by PCR analysis and DNA probes. *Acta Tropica*. **59** (2): 85-92.

205. Solano, P., Salou, E., Rayaisse, J.-B., Ravel, S., Gimonneau, G., Traore, I.*, et al.* (2015). Do tsetse flies only feed on blood? *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*. **36**: 184-9.

206. Benoit, J. B., Attardo, G. M., Baumann, A. A., Michalkova, V. and Aksoy, S. (2015). Adenotrophic viviparity in tsetse flies: potential for population control and as an insect model for lactation. *Annual review of entomology*. **60** (1): 351-71.

207. Davis, S., Aksoy, S. and Galvani, A. (2011). A global sensitivity analysis for African sleeping sickness. *Parasitology*. **138** (4): 516-26.

208. Welburn, S. C., Fevre, E. M., Coleman, P. G., Odiit, M. and Maudlin, I. (2001). Sleeping sickness: a tale of two diseases. *Trends in Parasitology*. **17** (1): 19-24.

209. Aksoy, S., Gibson, W. C. and Lehane, M. J. (2003). Interactions between tsetse and trypanosomes with implications for the control of trypanosomiasis. *Advances in Parasitology*. **53** (1): 1-83.

210. Maudlin, I. and Welburn, S. C. (1987). Lectin mediated establishment of midgut infections of *Trypanosoma congolense* and *Trypanosoma brucei* in *Glossina morsitans*. *Tropical medicine and parasitology: official organ of Deutsche Tropenmedizinische Gesellschaft and of Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ)*. **38** (3): 167-70.

211. Haines, L. R. (2013). Examining the tsetse teneral phenomenon and permissiveness to trypanosome infection. *Frontiers in Cellular and Infection Microbiology*. **3** (84): 1-6.

212. Jackson, C. H. N. (1946). An artificially isolated generation of tsetse flies (Diptera). *Bulletin of Entomological Research*. **37** (2): 291-9.

213. Van Hoof, L., Henrard, C. and Peel, E. (1937). Influences modificatrices de la transmissibilité cyclique du *Trypanosoma gambiense* par *Glossina palpalis*. *Annales de la Societé Belge Médecine Tropicale*. **17**: 249-72.

214. Walshe, D. P., Lehane, M. J. and Haines, L. R. (2011). Post eclosion age predicts the prevalence of midgut trypanosome infections in *Glossina*. *PLOS ONE*. **6** (11): e26984.

215. Wigglesworth, V. B. (1929). Digestion in the tsetse-fly: a study of structure and function. *Parasitology*. **21** (3): 288-321.

216. Rose, C., Casas-Sánchez, A., Dyer, N. A., Solórzano, C., Beckett, A. J., Middlehurst, B.*, et al.* (2020). *Trypanosoma brucei* colonizes the tsetse gut via an immature peritrophic matrix in the proventriculus. *Nature Microbiology*. **5**: 909-16.

217. Chandra, M., Liniger, M., Tetley, L., Roditi, I. and Barry, J. D. (2004). TsetseEP, a gut protein from the tsetse *Glossina morsitans*, is related to a major surface glycoprotein of trypanosomes transmitted by the fly and to the products of a *Drosophila* gene family. *Insect Biochemistry and Molecular Biology*. **34** (11): 1163-73.

218. Haines, L. R., Lehane, S. M., Pearson, T. W. and Lehane, M. J. (2010). Tsetse EP protein protects the fly midgut from trypanosome establishment. *PLOS Pathogens*. **6** (3): e1000793.

219. Weiss, B. L., Jingwen, W., Maltz, M. A., Yineng, W. and Aksoy, S. (2013). Trypanosome infection establishment in the tsetse fly gut is influenced by microbiome-regulated host immune barriers. *PLOS Pathogens*. **9** (1): 1-12.

220. Douglas, A. E. (2009). The microbial dimension in insect nutritional ecology. *Functional Ecology*. **23** (1): 38-47.

221. Dillon, R. J. and Dillon, V. M. (2004). The gut bacteria of insects: nonpathogenic interactions. *Annual review of entomology*. **49** (1): 71-92.

222. Robinson, C. J., Schloss, P., Ramos, Y., Raffa, K. and Handelsman, J. (2010). Robustness of the bacterial community in the cabbage white butterfly larval midgut. *Microbial Ecology*. **59** (2): 199-211.

223. Wong, C. N. A., Ng, P. and Douglas, A. E. (2011). Low-diversity bacterial community in the gut of the fruitfly *Drosophila melanogaster*. *Environmental Microbiology*. **13** (7): 1889-900.

224. Wang, J., Weiss, B. L. and Aksoy, S. (2013). Tsetse fly microbiota: form and function. *Frontiers in Cellular and Infection Microbiology*. **3** (69): 1-6.

225. Aksoy, S. (1995). *Wigglesworthia* gen. nov. and *Wigglesworthia glossinidia* sp. nov., taxa consisting of the mycetocyte-associated, primary endosymbionts of tsetse flies. *International Journal of Systematic Bacteriology*. **45** (4): 848-51.

226. Chen, X., Li, S. and Aksoy, S. (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): 49-58.

227. Akman, L., Yamashita, A., Watanabe, H., Oshima, K., Shiba, T., Hattori, M.*, et al.* (2002). Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nature Genetics*. **32** (3): 402-7.

228. Doudoumis, V., Blow, F., Saridaki, A., Augustinos, A., Dyer, N. A., Goodhead, I. B.*, 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** (1): 1-13.

229. Attardo, G. M., Lohs, C., Heddi, A., Alam, U. H., Yildirim, S. and Aksoy, S. (2008). Analysis of milk gland structure and function in *Glossina morsitans*: milk protein production, symbiont populations and fecundity. *Journal of Insect Physiology*. **54** (8): 1236-42.

230. Wang, J., Wu, Y., Yang, G. and Aksoy, S. (2009). Interactions between mutualist *Wigglesworthia* and tsetse peptidoglycan recognition protein (PGRP-LB) influence trypanosome transmission. *Proceedings of the National Academy of Sciences*. **106** (29): 12133-8.

231. Weiss, B. L., Wang, J. and Aksoy, S. (2011). Tsetse immune system maturation requires the presence of obligate symbionts in larvae. *PLOS Biology*. **9** (5): e1000619.

232. Weiss, B. L., Maltz, M. and Aksoy, S. (2012). Obligate symbionts activate immune system development in the tsetse fly. *The Journal of Immunology*. **188** (7): 3395-403.

233. Schlein, Y. (1977). Lethal effect of tetracycline on tsetse flies following damage to bacterioid symbionts. *Experientia*. **33** (4): 450-1.

234. Nogge, G. (1978). Aposymbiotic tsetse flies, *Glossina morsitans morsitans* obtained by feeding on rabbits immunized specifically with symbionts. *Journal of Insect Physiology*. **24** (4): 299-304.

235. Nogge, G. and Gerresheim, A. (1982). Experiments on the elimination of symbionts from the tsetse fly, *Glossina morsitans morsitans* (Diptera: *Glossinidae*), by antibiotics and lysozyme. *Journal of Invertebrate Pathology*. **40** (2): 166-79.

236. Pais, R., Lohs, C., Wu, Y., Wang, J. and Aksoy, S. (2008). The obligate mutualist *Wigglesworthia glossinidia* influences reproduction, digestion, and immunity processes of its host, the tsetse fly. *Applied and Environmental Microbiology*. **74** (19): 5965-74.

237. Hertig, M. (1936). The rickettsia, *Wolbachia pipientis* (gen. et sp. n.) and associated inclusions of the mosquito, *Culex pipiens*. *Parasitology*. **28** (4): 453-86.

238. Alam, U., Medlock, J., Brelsfoard, C., Pais, R., Lohs, C., Balmand, S.*, et al.* (2011). *Wolbachia* symbiont infections induce strong cytoplasmic incompatibility in the tsetse fly *Glossina morsitans*. *PLOS Pathogens*. **7** (12): e1002415.

239. Doudoumis, V., Tsiamis, G., Wamwiri, F., Brelsfoard, C., Alam, U., Aksoy, E.*, et al.* (2012). Detection and characterization of *Wolbachia* infections in laboratory and natural populations of different species of tsetse flies (genus *Glossina*). *BMC Microbiology*. **12** (1): 1-13.

240. Doudoumis, V., Alam, U., Aksoy, E., Abd-Alla, A. M. M., Tsiamis, G., Brelsfoard, C.*, et al.* (2013). Tsetse-*Wolbachia* symbiosis: comes of age and has great potential for pest and disease control. *Journal of Invertebrate Pathology*. **112** (S1): S94-S103.

241. Lefoulon, E., Clark, T., Guerrero, R., Cañizales, I., Cardenas-Callirgos, J. M., Junker, K.*, et al.* (2020). Diminutive, degraded but dissimilar: *Wolbachia* genomes from filarial nematodes do not conform to a single paradigm. *Microbial Genomics*. **6** (12): mgen000487.

242. 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 and Evolutionary Microbiology*. **49** (1): 267-75.

243. Balmand, S., Lohs, C., Aksoy, S. and Heddi, A. (2013). Tissue distribution and transmission routes for the tsetse fly endosymbionts. *Journal of Invertebrate Pathology*. **112** (s1): S116-S22.

244. Beard, C. B., O'Neill, S. L., Mason, P., Mandelco, L., Woese, C. R., Tesh, R. B.*, et al.* (1993). Genetic transformation and phylogeny of bacterial symbionts from tsetse. *Insect Molecular Biology*. **1** (3): 123-31.

245. Welburn, S. C., Maudlin, I. and Ellis, D. S. (1987). *In vitro* cultivation of rickettsia-like-organisms from *Glossina spp*. *Annals of Tropical Medicine and Parasitology*. **81** (3): 331-5.

246. Amann, R. I., Ludwig, W. and Schleifer, K. H. (1995). Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiological Reviews*. **59** (1): 143-69.

247. Bergthorsson, U. and Ochman, H. (1995). Heterogeneity of genome sizes among natural isolates of *Escherichia coli*. *Journal of Bacteriology*. **177** (20): 5784-9.

248. Kaiwa, N., Hosokawa, T., Kikuchi, Y., Nikoh, N., Meng, X. Y., Kimura, N.*, et al.* (2010). Primary gut symbiont and secondary, *Sodalis*-allied symbiont of the Scutellerid stinkbug *Cantao ocellatus*. *Applied and Environmental Microbiology*. **76** (11): 3486-94.

249. Hosokawa, T., Kaiwa, N., Matsuura, Y., Kikuchi, Y. and Fukatsu, T. (2015). Infection prevalence of *Sodalis* symbionts among stinkbugs. *Zoological Letters*. **1** (1): 1-7.

250. Toju, H., Hosokawa, T., Koga, R., Nikoh, N., Meng, X. Y., Kimura, N.*, et al.* (2010). "*Candidatus* Curculioniphilus buchneri," a novel clade of bacterial endocellular symbionts from weevils of the genus *Curculio*. *Applied and Environmental Microbiology*. **76** (1): 275-82.

251. Chari, A., Oakeson, K. F., Enomoto, S., Jackson, D. G., Fisher, M. A. and Dale, C. (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** (5): 1400-5.

252. Clayton, A. L., Oakeson, K. F., Gutin, M., Pontes, A., Dunn, D. M., Niederhausern, A. C. v.*, 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.

253. Farikou, O., Thevenon, S., Njiokou, F., Allal, F., Cuny, G. and Geiger, A. (2011). Genetic diversity and population structure of the secondary symbiont of tsetse flies, *Sodalis glossinidius*, in sleeping sickness foci in Cameroon. *PLOS Neglected Tropical Diseases*. **5** (8): 1-9.

254. Alam, U., Hyseni, C., Symula, R. E., Brelsfoard, C., Wu, Y., Kruglov, O*., et al.* (2012). Implications of microfauna-host interactions for trypanosome transmission dynamics in *Glossina fuscipes fuscipes* in Uganda. *Applied and Environmental Microbiology*. **78** (13): 4627-37.

255. Lindh, J. M. and Lehane, M. J. (2011). The tsetse fly *Glossina fuscipes fuscipes* (Diptera: *Glossina*) harbours a surprising diversity of bacteria other than symbionts. *Antonie van Leeuwenhoek*. **99** (3): 711-20.

256. Geiger, A., Ravel, S., Frutos, R. and Cuny, G. (2005). *Sodalis glossinidius* (Enterobacteriaceae) and vectorial competence of *Glossina palpalis gambiensis* and *Glossina morsitans morsitans* for T*rypanosoma congolense* savannah type. *Current Microbiology*. **51** (1): 35-40.

257. de Vooght, L., Caljon, G., van Hees, J. and van den Abbeele, J. (2015). Paternal transmission of a secondary symbiont during mating in the viviparous tsetse fly. *Molecular Biology and Evolution*. **32** (8): 1977-80.

258. Trappeniers, K., Matetovici, I., van den Abbeele, J. and de Vooght, L. (2019). The tsetse fly displays an attenuated immune response to its secondary symbiont, *Sodalis glossinidius*. *Frontiers in Microbiology*. **10** (1650): 1-20.

259. Galán, J. E. and Collmer, A. (1999). Type III secretion machines: bacterial Devices for protein delivery into host cells. *Science (New York, NY)*. **284** (5418): 1322-8.

260. Peacock, L., Cook, S., Ferris, V., Bailey, M. and Gibson, W. (2012). The life cycle of *Trypanosoma (Nannomonas) congolense* in the tsetse fly. *Parasites & Vectors*. **5** (109): 1-13.

261. Ooi, C.-P., Schuster, S., Cren-Travaillé, C., Bertiaux, E., Cosson, A., Goyard, S*., et al.* (2016). The cyclical development of *Trypanosoma vivax* in the tsetse fly involves an asymmetric division. *Frontiers in Cellular and Infection Microbiology*. **6** (115): 1-16.

262. Dale, C. and Welburn, S. C. (2001). The endosymbionts of tsetse flies: manipulating host–parasite interactions. *International Journal for Parasitology*. **31** (5): 628-31.

263. Okolo, C. J., Molyneux, D. H., Wallbanks, K. R. and Maudlin, I. (1988). Fluorescein conjugated lectins identify different carbohydrate residues on *Glossina* peritrophic membranes. *Tropical medicine and parasitology : official organ of Deutsche Tropenmedizinische Gesellschaft and of Deutsche Gesellschaft fur Technische Zusammenarbeit (GTZ)*. **39** (3): 208-10.

264. Welburn, S. C., Arnold, K., Maudlin, I. and Gooday, G. W. (1993). *Rickettsia*-like organisms and chitinase production in relation to transmission of trypanosomes by tsetse flies. *Parasitology*. **107** (2): 141-5.

265. De Vooght, L., Caljon, G., Stijlemans, B., De Baetselier, P., Coosemans, M. and Van Den, A. (2012). Expression and extracellular release of a functional anti-trypanosome Nanobody® in *Sodalis glossinidius*, a bacterial symbiont of the tsetse fly. *Microbial Cell Factories*. **11** (23): 1-11.

266. De Vooght, L., Caljon, G., De Ridder, K. and Den Abbeele, J. V. (2014). Delivery of a functional anti-trypanosome Nanobody in different tsetse fly tissues via a bacterial symbiont, *Sodalis glossinidius*. *Microbial Cell Factories*. **13** (1): 1-23.

267. Geiger, A., Fardeau, M.-L., Grebaut, P., Vatunga, G., Josénando, T., Herder, S.*, et al.* (2009). First isolation of *Enterobacter*, *Enterococcus*, and *Acinetobacter spp.* as inhabitants of the tsetse fly (*Glossina palpalis palpalis*) midgut. *Infection, Genetics & Evolution*. **9** (6): 1364-70.

268. Geiger, A., Fardeau, M.-L., Njiokou, F., Joseph, M., Asonganyi, T., Ollivier, B.*, et al.* (2011). Bacterial diversity associated with populations of *Glossina spp.* from Cameroon and distribution within the Campo sleeping sickness focus. *Microbial Ecology*. **62** (3): 632-43.

269. Geiger, A., Fardeau, M. L., Falsen, E., Ollivier, B. and Cuny, G. (2010). *Serratia glossinae* sp nov., isolated from the midgut of the tsetse fly *Glossina palpalis gambiensis*. *International Journal of Systematic and Evolutionary Microbiology*. **60** (6): 1261-5.

270. Geiger, A., Fardeau, M. L., Njiokou, F. and Ollivier, B. (2013). *Glossina* spp. gut bacterial flora and their putative role in fly-hosted trypanosome development. *Frontiers in Cellular and Infection Microbiology*. **3** (34): 1-8.

271. Griffith, B. C., Weiss, B. L., Aksoy, E., Mireji, P. O., Auma, J. E., Wamwiri, F. N.*, et al.* (2018). Analysis of the gut-specific microbiome from field-captured tsetse flies, and its potential relevance to host trypanosome vector competence. *BMC Microbiology*. **18** (1): 55-67.

272. Aksoy, E., Telleria, E. L., Echodu, R., Wu, Y., Okedi, L. M., Weiss, B. L.*, et al.* (2014). Analysis of Multiple Tsetse Fly Populations in Uganda Reveals Limited Diversity and Species-Specific Gut Microbiota. *Applied and Environmental Microbiology*. **80** (14): 4301-12.

273. Schneider, D. I., Saarman, N., Onyango, M. G., Hyseni, C., Opiro, R., Echodu, R.*, et al.* (2019). Spatio-temporal distribution of *Spiroplasma* infections in the tsetse fly (*Glossina fuscipes fuscipes*) in northern Uganda. *PLOS Neglected Tropical Diseases*. **13** (8): e0007340.

274. Geiger, A., Fardeau, M.-L., Grebaut, P., Vatunga, G., Josénando, T., Herder, S.*, et al.* (2009). First isolation of *Enterobacter*, *Enterococcus*, and *Acinetobacter* spp. as inhabitants of the tsetse fly (*Glossina palpalis palpalis*) midgut. *Infection, Genetics and Evolution*. **9** (6): 1364-70.

275. Mediannikov, O., Audoly, G., Diatta, G., Trape, J. F. and Raoult, D. (2012). New *Rickettsia* sp. in tsetse flies from Senegal. *Comparative Immunology Microbiology and Infectious Diseases*. **35** (2): 145-50.

276. Burza, S., Croft, S. L. and Boelaert, M. (2018). Leishmaniasis. *The Lancet*. **392** (10151): 951-70.

277. World Health Organization. (2020). Leishmaniasis. Geneva. Available from: https://www.who.int/news-room/fact-sheets/detail/leishmaniasis.

278. Haouas, N. (2016). Estimations of cutaneous leishmaniasis burden: a constant challenge. *The Lancet Infectious Diseases*. **16** (5): 515-6.

279. de Vries, H. J. C., Reedijk, S. H. and Schallig, H. D. F. H. (2015). Cutaneous leishmaniasis: recent developments in diagnosis and management. *American journal of clinical dermatology*. **16** (2): 99-109.

280. Karimkhani, C., Wanga, V., Coffeng, L. E., Naghavi, P., Dellavalle, R. P. and Naghavi, M. (2016). Global burden of cutaneous leishmaniasis: a cross-sectional analysis from the Global Burden of Disease Study 2013. *The Lancet Infectious Diseases*. **16** (5): 584-91.

281. Bennis, I., Thys, S., Filali, H., De Brouwere, V., Sahibi, H. and Boelaert, M. (2017). Psychosocial impact of scars due to cutaneous leishmaniasis on high school students in Errachidia province, Morocco. *Infectious Diseases of Poverty*. **6** (46): 1-8.

282. Yanik, M., Gurel, M., Simsek, Z. and Kati, M. (2004). The psychological impact of cutaneous leishmaniasis. *Clinical and Experimental Dermatology: Clinical Dermatology*. **29** (5): 464-7.

283. Cincurá, C., de Lima, C. M. F., Machado, P. R., Oliveira-Filho, J., Glesby, M. J., Lessa, M. M.*, et al.* (2017). Mucosal leishmaniasis: a retrospective study of 327 cases from an endemic area of *Leishmania (Viannia) braziliensis*. *The American Journal of Tropical Medicine and Hygiene*. **97** (3): 761-6.

284. Elkhair, E. B. (2014). Elevated cortisol level due to visceral leishmaniasis and skin hyper-pigmentation are causally related. *International Journal Of Science Commerce and Humanities*. **2** (7): 86-92.

285. Moral, L., Rubio, E. M. and Moya, M. (2002). A leishmanin skin test survey in the human population of l'Alacanti region (Spain): implications for the epidemiology of *Leishmania infantum* infection in southern Europe. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. **96** (2): 129-32.

286. Ready, P. D. (2014). Epidemiology of visceral leishmaniasis. *Clinical Epidemiology*. **6**: 147-54.

287. Yimer, M., Abera, B., Mulu, W., Zenebe, Y. and Bezabih, B. (2014). Proportion of visceral leishmaniasis and human immune deficiency virus co-infection among clinically confirmed visceral leishmaniasis patients at the endemic foci of the Amhara National Regional State, north-west Ethiopia. *American Journal of Biomedical and Life Sciences*. **2** (1): 1-7.

288. Mock, D. J., Hollenbaugh, J. A., Daddacha, W., Overstreet, M. G., Lazarski, C. A., Fowell, D. J.*, et al.* (2012). *Leishmania* induces survival, proliferation and elevated cellular dNTP levels in human monocytes promoting acceleration of HIV co-infection. *PLOS Pathogens*. **8** (4): e1002635.

289. Burza, S., Mahajan, R., Sinha, P. K., van Griensven, J., Pandey, K., Lima, M. A.*, et al.* (2014). Visceral leishmaniasis and HIV co-infection in Bihar, India: long-term effectiveness and treatment outcomes with liposomal amphotericin B (AmBisome). *PLOS Neglected Tropical Diseases*. **8** (8): e3053.

290. Molina, R., Ghosh, D., Carrillo, E., Monnerat, S., Bern, C., Mondal, D.*, et al.* (2017). Infectivity of post-kala-azar dermal leishmaniasis patients to sand flies: revisiting a proof of concept in the context of the kala-azar elimination program in the Indian subcontinent. *Clinical Infectious Diseases*. **65** (1): 150-3.

291. Addy, M. and Nandy, A. (1992). Ten years of kala-azar in west Bengal, Part I. Did post-kala-azar dermal leishmaniasis initiate the outbreak in 24-Parganas? *Bulletin of the World Health Organization*. **70** (3): 341-6.

292. Singh, O. P., Hasker, E., Sacks, D., Boelaert, M. and Sundar, S. (2014). Asymptomatic *Leishmania* infection: a new challenge for *Leishmania* control. *Clinical Infectious Diseases*. **58** (10): 1424-9.

293. Petersen, C. (2020). Leishmaniosis in Dogs. MSD, Available from: https://www.msdvetmanual.com/generalized-conditions/leishmaniosis/leishmaniosis-in-dogs.

294. Singh, N., Mishra, J., Singh, R. and Singh, S. (2013). Animal reservoirs of visceral leishmaniasis in India. *Journal of Parasitology*. **99** (1): 64-7.

295. Alvar, J., Cañavate, C., Molina, R., Moreno, J. and Nieto, J. (2004). Canine Leishmaniasis. Ed: Baker JR, Muller R, Rollinson D, In: *Advances in Parasitology*. vol. 1.

296. Sundar, S. and Rai, M. (2002). Laboratory diagnosis of visceral leishmaniasis. *Clinical and Diagnostic Laboratory Immunology*. **9** (5): 951-8.

297. World Health Organization. (2010). Control of the leishmaniases: report of a meeting of the WHO Expert Committee on the Control of Leishmaniases. Available from: https://apps.who.int/iris/handle/10665/44412.

298. Fernández Cotrina, J., Iniesta, V., Monroy, I., Baz, V., Hugnet, C., Marañon, F.*, et al.* (2018). A large-scale field randomized trial demonstrates safety and efficacy of the vaccine LetiFend® against canine leishmaniosis. *Vaccine*. **36** (15): 1972-82.

299. World Health Organization. (2020). Ending the neglect to attain the Sustainable Development Goals: a road map for neglected tropical diseases 2021-2030. Available from: https://www.who.int/neglected_diseases/WHONTD-roadmap-2030/en/.

300. World Health Organization. (2017). Accelerated plan for kala-azar elimination. Available from: https://www.who.int/leishmaniasis/resources/Accelerated_plan_for_Kala-azar_Elimination_2017/en/.

301. Coleman, M., Foster, G. M., Deb, R., Pratap Singh, R., Ismail, H. M., Shivam, P.*, et al.* (2015). DDT-based indoor residual spraying suboptimal for visceral leishmaniasis elimination in India. *Proceedings of the National Academy of Sciences*. **112** (28): 8573-8.

302. Singh, O. P., Hasker, E., Boelaert, M. and Sundar, S. (2016). Elimination of visceral leishmaniasis on the Indian subcontinent. *The Lancet Infectious Diseases*. **16** (12): e304-e9.

303. Service, M. (2012). *Medical Entomology for Students*. 5th ed. New York: Cambridge University Press.

304. Bejarano, E. E., Uribe, S., Rojas, W. and Darío Vélez, I. (2002). Phlebotomine sand flies (Diptera: Psychodidae) associated with the appearance of urban leishmaniasis in the city of Sincelejo, Colombia. *Memórias do Instituto Oswaldo Cruz*. **97** (5): 645-7.

305. Faiman, R., Cuño, R. and Warburg, A. (2009). Control of phlebotomine sand flies with vertical fine-mesh nets. *Journal of Medical Entomology*. **46** (4): 820-31.

306. Telleria, E. L., Martins-da-Silva, A., Tempone, A. J. and Traub-Csekö, Y. M. (2018). *Leishmania*, microbiota and sand fly immunity. *Parasitology*. **145** (10): 1336-53.

307. Killick-Kendrick, R. (1999). The biology and control of phlebotomine sand flies. *Clinics in dermatology*. **17** (3): 279-89.

308. Feliciangeli, M. D. (2004). Natural breeding places of phlebotomine sandflies. *Medical and Veterinary Entomology*. **18** (1): 71-80.

309. Ready, P. D. (2013). Biology of Phlebotomine sand flies as vectors of disease agents. *Annual review of entomology*. **58**: 227-50.

310. Guernaoui, S., Garcia, D., Gazanion, E., Ouhdouch, Y., Boumezzough, A., Pesson, B.*, et al.* (2011). Bacterial flora as indicated by PCR-temperature gradient gel electrophoresis (TGGE) of 16S rDNA gene fragments from isolated guts of phlebotomine sand flies (Diptera: Psychodidae). *Journal of Vector Ecology*. **36**: S144-S7.

311. Kelly, P. H., Bahr, S. M., Serafim, T. D., Ajami, N. J., Petrosino, J. F., Meneses, C.*, et al.* (2017). The gut microbiome of the vector *Lutzomyia longipalpis* is essential for survival of *Leishmania infantum*. *mBio*. **8** (1): e01121-16.

312. Pires, A. C. A. M., Villegas, L. E. M., Campolina, T. B., Orfanó, A. S., Pimenta, P. F. P. and Secundino, N. F. C. (2017). Bacterial diversity of wild-caught *Lutzomyia longipalpis* (a vector of zoonotic visceral leishmaniasis in Brazil) under distinct physiological conditions by metagenomics analysis. *Parasites & Vectors*. **10** (627): 1-9.

313. Monteiro, C. C., Villegas, L. E. M., Campolina, T. B., Pires, A. C. M. A., Miranda, J. C., Pimenta, P. F. P.*, et al.* (2016). Bacterial diversity of the American sand fly *Lutzomyia intermedia* using high-throughput metagenomic sequencing. *Parasites & Vectors*. **9** (480): 1-6.

314. Volf, P., Kiewegová, A. and Nemec, A. (2002). Bacterial colonisation in the gut of *Phlebotomus duboseqi* (*Diptera*: Psychodidae): transtadial passage and the role of female diet. *Folia Parasitologica*. **49** (1): 73-7.

315. Manfredi, A. P., Perotti, N. I. and Martínez, M. A. (2015). Cellulose degrading bacteria isolated from industrial samples and the gut of native insects from Northwest of Argentina. *Journal of basic microbiology*. **55** (12): 1384-93.

316. Thapa, S., Ranjan, K., Ramakrishnan, B., Velmourougane, K. and Prasanna, R. (2018). Influence of fertilizers and rice cultivation methods on the abundance and diversity of phyllosphere microbiome. *Journal of basic microbiology*. **58** (2): 172-86.

317. Burke, G. R. and Moran, N. A. (2011). Massive genomic decay in *Serratia symbiotica*, a recently evolved symbiont of aphids. *Genome Biology and Evolution*. **3** (1): 195-208.

318. Diaz-Albiter, H., Sant'Anna, M. R. V., Genta, F. A. and Dillon, R. J. (2012). Reactive oxygen species-mediated immunity against *Leishmania mexicana* and *Serratia marcescens* in the phlebotomine sand fly *Lutzomyia longipalpis*. *Journal of Biological Chemistry*. **287** (28): 23995-4003.

319. van Rhijn, P. and Vanderleyden, J. (1995). The *Rhizobium*-plant symbiosis. *Microbiological Reviews*. **59** (1): 124-42.

320. Kaneko, T., Nakamura, Y., Sato, S., Minamisawa, K., Uchiumi, T., Sasamoto, S.*, et al.* (2002). Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. *DNA Research*. **9** (6): 189-97.

321. Denny, T. (2007). Plant pathogenic *Ralstonia* species. Ed: *Plant-associated bacteria*. vol. 1.

322. Hillesland, H., Read, A., Subhadra, B., Hurwitz, I., McKelvey, R., Ghosh, K.*, et al.* (2008). Identification of aerobic gut bacteria from the kala azar vector, *Phlebotomus argentipes*: a platform for potential paratransgenic manipulation of sand flies. *The American Journal of Tropical Medicine and Hygiene*. **79** (6): 881-6.

323. Oliveira, S. M. P. d., Moraes, B. A. d., Gonçalves, C. A., Giordano-Dias, C. M., d'Almeida, J. M., Asensi, M. D.*, et al.* (2000). Prevalence of the microbiota in the digestive tract of wild-caught females of *Lutzomyia longipalpis* (Lutz & Neiva) (Diptera: Psychodidae). *Revista da Sociedade Brasileira de Medicina Tropical*. **33** (3): 319-22.

324. Perira de Oliveira, S. M., de Morais, B. A., Gonçalves, C. A., Giordano-Dias, C. M., Vilela, M. I., Brazil, R. P.*, et al.* (2001). Digestive tract microbiota in female *Lutzomyia longipalpis* (Lutz & Neiva, 1912) (*Diptera*: Psychodidae) from colonies feeding on blood meal and sucrose plus blood meal. *Cadernos de saude publica*. **17** (1): 229-32.

325. Gouveia, C., Asensi, M. D., Zahner, V., Rangel, E. F. and Oliveira, S. M. P. d. (2008). Study on the bacterial midgut microbiota associated to different Brazilian populations of *Lutzomyia longipalpis* (Lutz & Neiva) (Diptera: Psychodidae). *Neotropical Entomology*. **37** (5): 597-601.

326. Fraihi, W., Fares, W., Perrin, P., Dorkeld, F., Sereno, D., Barhoumi, W.*, et al.* (2017). An integrated overview of the midgut bacterial flora composition of *Phlebotomus perniciosus*, a vector of zoonotic visceral leishmaniasis in the Western Mediterranean Basin. *PLOS Neglected Tropical Diseases*. **11** (3): e0005484.

327. Peterkova-Koci, K., Robles-Murguia, M., Ramalho-Ortigao, M. and Zurek, L. (2012). Significance of bacteria in oviposition and larval development of the sand fly *Lutzomyia longipalpis*. *Parasites & Vectors*. **5** (145): 1-8.

328. Moraes, C., Lucena, S., Moreira, B., Brazil, R., Gontijo, N. and Genta, F. (2012). Relationship between digestive enzymes and food habit of *Lutzomyia longipalpis* (*Diptera*: Psychodidae) larvae: characterization of carbohydrases and digestion of microorganisms. *Journal of Insect Physiology*. **58** (8): 1136-45.

329. Diaz-Albiter, H., Mitford, R., Genta, F. A., Sant'Anna, M. R. and Dillon, R. J. (2011). Reactive oxygen species scavenging by catalase is important for female *Lutzomyia longipalpis* fecundity and mortality. *PLOS ONE*. **6** (3): e17486.

330. Sant'Anna, M. R. V., Diaz-Albiter, H., Aguiar-Martins, K., Al Salem, W. S., Cavalcante, R. R., Dillon, V. M.*, et al.* (2014). Colonisation resistance in the sand fly gut: *Leishmania* protects *Lutzomyia longipalpis* from bacterial infection. *Parasites & Vectors*. **7** (329): 1-10.

331. Chakraborty, U., Chakraborty, B. and Basnet, M. (2006). Plant growth promotion and induction of resistance in *Camellia sinensis* by *Bacillus megaterium*. *Journal of basic microbiology*. **46** (3): 186-95.

332. Rattray, F. P. and Fox, P. F. (1999). Aspects of enzymology and biochemical properties of *Brevibacterium linens* relevant to cheese ripening: a review. *Journal of Dairy Science*. **82** (5): 891-909.

333. Braks, M., Anderson, R. and Knols, B. (1999). Infochemicals in mosquito host selection: human skin microflora and *Plasmodium* parasites. *Parasitology Today*. **15** (10): 409-13.

334. Meijerink, J. and van Loon, J. J. (1999). Sensitivities of antennal olfactory neurons of the malaria mosquito, *Anopheles gambiae*, to carboxylic acids. *Journal of Insect Physiology*. **45** (4): 365-73.

335. Hurwitz, I., Hillesland, H., Fieck, A., Das, P. and Durvasula, R. (2011). The paratransgenic sand fly: A platform for control of *Leishmania* transmission. *Parasites & Vectors*. **4** (82): 1-9.

336. Hurwitz, I., Forshaw, A., Yacisin, K., Ramalho-Ortigao, M., Satoskar, A. and Durvasula, R. (2014). Paratransgenic Control of Leishmaniasis: New Developments. Ed: Satoskar A, Durvasula R, In: *Pathogenesis of Leishmaniasis: New Developments in Research*. vol. 1.

337. Buchner, J. (1965). *Endosymbiosis of Animals with Plant Microorganisms*. New York: John Wiley & Sons, Inc.

338. Pontes, M. H. and Dale, C. (2006). Culture and manipulation of insect facultative symbionts. *Trends in Microbiology*. **14** (9): 406-12.

339. Dale, C., Beeton, M., Harbison, C., Jones, T. and Pontes, M. (2006). Isolation, pure culture, and characterization of *Candidatus* Arsenophonus arthropodicus, an intracellular secondary endosymbiont from the hippoboscid louse fly *Pseudolynchia canariensis*. *Applied and Environmental Microbiology*. **72** (4): 2997-3004.

340. Durvasula, R. V., Sundaram, R. K., Kirsch, P., Hurwitz, I., Crawford, C. V., Dotson, E.*, et al.* (2008). Genetic transformation of a Corynebacterial symbiont from the Chagas disease vector *Triatoma infestans*. *Experimental parasitology*. **119** (1): 94-8.

341. Kounatidis, I., Crotti, E., Sapountzis, P., Sacchi, L., Rizzi, A., Chouaia, B.*, et al.* (2009). *Acetobacter tropicalis* Is a major symbiont of the olive fruit fly (*Bactrocera oleae*). *Applied and Environmental Microbiology*. **75** (10): 3281-8.

342. Chouaia, B., Rossi, P., Montagna, M., Ricci, I., Crotti, E., Damiani, C.*, et al.* (2010). Molecular evidence for multiple infections as revealed by typing of *Asaia* bacterial symbionts of four mosquito species. *Applied and Environmental Microbiology*. **76** (22): 7444.

343. Darby, A. C., Choi, J. H., Wilkes, T., Hughes, M. A., Werren, J. H., Hurst, G. D. D.*, et al.* (2010). Characteristics of the genome of *Arsenophonus nasoniae*, son-killer bacterium of the wasp *Nasonia*. *Insect Molecular Biology*. **19** (s1): 75-89.

344. Chrudimský, T., Husník, F., Nováková, E. and Hypša, V. (2012). Candidatus *Sodalis melophagi sp. nov*.: phylogenetically independent comparative model to the tsetse fly symbiont *Sodalis glossinidius*. *PLOS ONE*. **7** (7): e40354.

345. Xu, Y., Buss, E. A. and Boucias, D. G. (2016). Culturing and characterization of gut symbiont *Burkholderia spp.* from the southern chinch bug, *Blissus insularis* (*Hemiptera*: Blissidae). *Applied and Environmental Microbiology*. **82** (11): 3319-30.

346. Brandt, J. W., Chevignon, G., Oliver, K. M. and Strand, M. R. (2017). Culture of an aphid heritable symbiont demonstrates its direct role in defence against parasitoids. *Proceedings of the Royal Society B-Biological Sciences*. **284** (1866): 1-9.

347. Chen, S., Blom, J. and Walker, E. D. (2017). Genomic, physiologic, and symbiotic characterization of *Serratia marcescens* strains isolated from the mosquito *Anopheles stephensi*. *Frontiers in Microbiology*. **8** (1483): 1-16.

348. Masson, F., Calderon Copete, S., Schüpfer, F., Garcia-Arraez, G. and Lemaitre, B. (2018). *In vitro* culture of the insect endosymbiont *Spiroplasma*

poulsonii highlights bacterial genes involved in host-symbiont interaction. *mBio*. **9** (2): e00024-18.

349. Prosdocimi, E. M., Mapelli, F., Gonella, E., Borin, S. and Crotti, E. (2015). Microbial ecology-based methods to characterize the bacterial communities of non-model insects. *Journal of microbiological methods*. **119**: 110-25.

350. Woo, P. C. Y., Lau, S. K. P., Teng, J. L. L., Tse, H. and Yuen, K. Y. (2008). Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clinical Microbiology and Infection*. **14** (10): 908-34.

351. Moran, N. A. and Mira, A. (2001). The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biology*. **2** (12): 1-12.

352. Bennett, G. M., Abbà, S., Kube, M. and Marzachì, C. (2016). Complete genome sequences of the obligate symbionts *Candidatus* Sulcia muelleri and *Nasuia deltocephalinicola* from the pestiferous leafhopper *Macrosteles quadripunctulatus* (Hemiptera: Cicadellidae). *Genome Announcements*. **4** (1): e01604.

353. Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews*. **51** (2): 221-71.

354. Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*. **74** (11): 5088-90.

355. Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F., Ludwig, W., Schleifer, K.*, et al.* (2014). Uniting the classification of cultured and uncultured Bacteria and Archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*. **12**: 635-45.

356. Gupta, R., Lanter, J. M. and Woese, C. R. (1983). Sequence of the 16S Ribosomal RNA from *Halobacterium volcanii*, an Archaebacterium. *Science (New York, NY)*. **221** (4611): 656-9.

357. Petti, C. A. (2007). Detection and identification of microorganisms by gene amplification and sequencing. *Clinical Infectious Diseases*. **44** (8): 1108-14.

358. Acinas, S. G., Klepac-Ceraj, V., Hunt, D. E., Pharino, C., Ceraj, I., Distel, D. L.*, et al.* (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*. **430** (6999): 551-4.

359. Callahan, B. J., McMurdie, P. J. and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*. **11** (12): 2639-43.

360. Pollock, J., Glendinning, L., Wisedchanwet, T. and Watson, M. (2018). The madness of microbiome: attempting to find consensus “best practice” for 16S microbiome studies. *Applied and Environmental Microbiology*. **84** (7): e02627-17.

361. McKain, N., Genc, B., Snelling, T. J. and Wallace, R. J. (2013). Differential recovery of bacterial and archaeal 16S rRNA genes from ruminal digesta in response to glycerol as cryoprotectant. *Journal of microbiological methods*. **95** (3): 381-3.

362. Fouhy, F., Clooney, A. G., Stanton, C., Claesson, M. J. and Cotter, P. D. (2016). 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiology*. **16** (123): 1-13.

363. Bahl, M. I., Bergström, A. and Licht, T. R. (2012). Freezing fecal samples prior to DNA extraction affects the Firmicutes to Bacteroidetes ratio determined by downstream quantitative PCR analysis. *Fems Microbiology Letters*. **329** (2): 193-7.

364. Kuske, C. R., Banton, K. L., Adorada, D. L., Stark, P. C., Hill, K. K. and Jackson, P. J. (1998). Small-scale DNA sample preparation method for field PCR detection of microbial cells and spores in soil. *Applied Environmental Microbiology*. **64** (7): 2463-72.

365. Schrader, C., Schielke, A., Ellerbroek, L. and Johne, R. (2012). PCR inhibitors–occurrence, properties and removal. *Journal of applied microbiology*. **113** (5): 1014-26.

366. Desneux, J. and Pourcher, A. M. (2014). Comparison of DNA extraction kits and modification of DNA elution procedure for the quantitation of subdominant bacteria from piggery effluents with real-time PCR. *Microbiology Open*. **3** (4): 437-45.

367. Mackenzie, B., Waite, D. W. and Taylor, M. W. (2015). Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. *Frontiers in Microbiology*. **6** (130): 130.

368. Gerasimidis, K., Bertz, M., Quince, C., Brunner, K., Bruce, A., Combet, E.*, et al.* (2016). The effect of DNA extraction methodology on gut microbiota research applications. *BMC research notes*. **9** (1): 1-10.

369. Guo, F. and Zhang, T. (2013). Biases during DNA extraction of activated sludge samples revealed by high throughput sequencing. *Applied Microbiology and Biotechnology*. **97** (10): 4607-16.

370. Hart, M. L., Meyer, A., Johnson, P. J. and Ericsson, A. C. (2015). Comparative evaluation of DNA extraction methods from feces of multiple host species for downstream next-generation sequencing. *PLOS ONE*. **10** (11): e0143334.

371. Kennedy, N. A., Walker, A. W., Berry, S. H., Duncan, S. H., Farquarson, F. M., Louis, P.*, et al.* (2014). The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLOS ONE*. **9** (2): e88982.

372. Henderson, G., Cox, F., Kittelmann, S., Miri, V. H., Zethof, M., Noel, S. J.*, et al.* (2013). Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. *PLOS ONE*. **8** (9): e74787.

373. Yang, B., Wang, Y. and Qian, P.-Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC bioinformatics*. **17** (135): 1-8.

374. Tremblay, J., Singh, K., Fern, A., Kirton, E. S., He, S., Woyke, T.*, et al.* (2015). Primer and platform effects on 16S rRNA tag sequencing. *Frontiers in Microbiology*. **6** (771): 1-15.

375. Cruaud, P., Vigneron, A., Lucchetti-Miganeh, C., Ciron, P. E., Godfroy, A. and Cambon-Bonavita, M.-A. (2014). Influence of DNA extraction method, 16S rRNA targeted hypervariable regions, and sample origin on microbial diversity detected by 454 pyrosequencing in marine chemosynthetic ecosystems. *Applied Environmental Microbiology*. **80** (15): 4626-39.

376. Ghyselinck, J., Pfeiffer, S., Heylen, K., Sessitsch, A. and De Vos, P. (2013). The effect of primer choice and short read sequences on the outcome of 16S rRNA gene based diversity studies. *PLOS ONE*. **8** (8): e71360.

377. Chakravorty, S., Helb, D., Burday, M., Connell, N. and Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods*. **69** (2): 330-9.

378. Bergmann, G. T., Bates, S. T., Eilers, K. G., Lauber, C. L., Caporaso, J. G., Walters, W. A.*, et al.* (2011). The under-recognized dominance of *Verrucomicrobia* in soil bacterial communities. *Soil Biology and Biochemistry*. **43** (7): 1450-5.

379. Kanagawa, T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of bioscience and bioengineering*. **96** (4): 317-23.

380. Ahn, J.-H., Kim, B.-Y., Song, J. and Weon, H.-Y. (2012). Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. *Journal of Microbiology*. **50** (6): 1071-4.

381. Wu, J.-Y., Jiang, X.-T., Jiang, Y.-X., Lu, S.-Y., Zou, F. and Zhou, H.-W. (2010). Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. *BMC Microbiology*. **10** (255): 1-7.

382. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. and Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and Environmental Microbiology*. **79** (17): 5112-20.

383. Fadrosh, D. W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R. M.*, et al.* (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*. **2** (1): 1-7.

384. Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F.*, et al.* (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology*. **12** (87): 1-12.

385. Picozzi, K., Fèvre, E. M., Odiit, M., Carrington, M., Eisler, M. C., Maudlin, I.*, et al.* (2005). Sleeping sickness in Uganda: a thin line between two fatal diseases. *BMJ (Clinical research ed)*. **331** (7527): 1238-41.

386. Welburn, S. and Odiit, M. (2002). Recent developments in human African trypanosomiasis. *Current Opinion in Infectious Diseases*. **15** (5): 477-84.

387. Ochan, B. (2004). Sleeping sickness re-emerges in Uganda. *BMJ : British Medical Journal*. **328** (70).

388. Cecchi, G., Wint, W., Shaw, A., Marletta, A., Mattioli, R. and Robinson, T. (2010). Geographic distribution and environmental characterization of livestock production systems in Eastern Africa. *Agriculture, Ecosystems & Environment*. **135** (1): 98-110.

389. Thornton, P. K., Kruska, R., Henninger, N., Kristjanson, P. M., Reid, R. S. and Robinson, T. P. (2003). Locating poor livestock keepers at the global level for research and development targeting. *Land Use Policy*. **20** (4): 311-22.

390. Muhanguzi, D., Mugenyi, A., Bigirwa, G., Kamusiime, M., Kitibwa, A., Akurut, G. G.*, et al.* (2017). African animal trypanosomiasis as a constraint to livestock health and production in Karamoja region: a detailed qualitative and quantitative assessment. *BMC Vet Res*. **13** (355): 1-13.

391. Okello, W. O., Muhanguzi, D., MacLeod, E. T., Welburn, S. C., Waiswa, C. and Shaw, A. P. (2015). Contribution of draft cattle to rural livelihoods in a district of southeastern Uganda endemic for bovine parasitic diseases: an economic evaluation. *Parasites & Vectors*. **8** (1): 1-9.

392. Muhanguzi, D., Okello, W. O., Kabasa, J. D., Waiswa, C., Welburn, S. C. and Shaw, A. P. (2015). Cost analysis of options for management of African Animal Trypanosomiasis using interventions targeted at cattle in Tororo District; south-eastern Uganda. *Parasites & Vectors*. **8** (387): 1-9.

393. Bengaly, Z., Vitouley, S. H., Somda, M. B., Zongo, A., Têko-Agbo, A., Cecchi, G.*, et al.* (2018). Drug quality analysis of isometamidium chloride hydrochloride and diminazene diaceturate used for the treatment of African animal trypanosomosis in West Africa. *BMC Veterinary Research*. **14** (361): 1-8.

394. Geerts, S., Holmes, P. H., Eisler, M. C. and Diall, O. (2001). African bovine trypanosomiasis: the problem of drug resistance. *Trends in Parasitology*. **17** (1): 25-8.

395. Shaw, A. P. M., Torr, S. J., Waiswa, C., Cecchi, G., Wint, G. R. W., Mattioli, R. C.*, et al.* (2013). Estimating the costs of tsetse control options: An example for Uganda. *Preventive Veterinary Medicine*. **110** (3): 290-303.

396. Farikou, O., Njiokou, F., Mbida Mbida, J. A., Njitchouang, G. R., Djeunga, H. N., Asonganyi, T.*, et al.* (2010). Tripartite interactions between tsetse flies, *Sodalis glossinidius* and trypanosomes—An epidemiological approach in two historical human African trypanosomiasis foci in Cameroon. *Infection, Genetics and Evolution*. **10** (1): 115-21.

397. Soumana, I. H., Berthier, D., Tchicaya, B., Thevenon, S., Njiokou, F., Cuny, G.*, et al.* (2013). Population dynamics of *Glossina palpalis gambiensis* symbionts, *Sodalis glossinidius,* and *Wigglesworthia glossinidia*, throughout host-fly development. *Infection, Genetics & Evolution*. **13**: 41-8.

398. Maudlin, I., Welburn, S. and Mehlitz, D. (1990). The relationship between rickettsia-like-organisms and trypanosome infections in natural populations of tsetse in Liberia. *Tropical medicine and parasitology: official organ of Deutsche Tropenmedizinische Gesellschaft and of Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ)*. **41** (3): 265-7.

399. Schneider, D. I., Garschall, K. I., Parker, A. G., Abd-Alla, A. M. M. and Miller, W. J. (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**: S104-S15.

400. Bastian, F. O., Elzer, P. H. and Wu, X. (2012). *Spiroplasma spp.* biofilm formation is instrumental for their role in the pathogenesis of plant, insect and animal diseases. *Experimental and molecular pathology*. **93** (1): 116-28.

401. Paredes, J. C., Herren, J. K., Schüpfer, F., Marin, R., Claverol, S., Kuo, C.-H.*, et al.* (2015). Genome sequence of the *Drosophila melanogaster* male-killing *Spiroplasma* strain MSRO endosymbiont. *mBio*. **6** (2): e02437-14.

402. Schwarz, R. S., Teixeira, É. W., Tauber, J. P., Birke, J. M., Martins, M. F., Fonseca, I.*, et al.* (2014). Honey bee colonies act as reservoirs for two *Spiroplasma* facultative symbionts and incur complex, multiyear infection dynamics. *MicrobiologyOpen*. **3** (3): 341-55.

403. Jaenike, J., Unckless, R., Cockburn, S. N., Boelio, L. M. and Perlman, S. J. (2010). Adaptation via Symbiosis: Recent Spread of a *Drosophila* Defensive Symbiont. *Science (New York, NY)*. **329** (5988): 212-5.

404. Łukasik, P., Guo, H., Van Asch, M., Ferrari, J. and Godfray, H. (2013). Protection against a fungal pathogen conferred by the aphid facultative endosymbionts *Rickettsia* and *Spiroplasma* is expressed in multiple host genotypes and species and is not influenced by co-infection with another symbiont. *Journal of Evolutionary Biology*. **26** (12): 2654-61.

405. Xie, J., Vilchez, I. and Mateos, M. (2010). *Spiroplasma* bacteria enhance survival of *Drosophila hydei* attacked by the parasitic wasp *Leptopilina heterotoma*. *PLOS ONE*. **5** (8): e12149.

406. Mouches, C., Bové, J. and Albisetti, J. (1984). Pathogenicity of *Spiroplasma apis* and other spiroplasmas for honey-bees in southwestern France. *Annales de l'Institut Pasteur/Microbiologie*. **135** (1): 151-5.

407. Clark, T., Whitcomb, R., Tully, J., Mouchès, C., Saillard, C., Bové, J.*, et al.* (1985). *Spiroplasma melliferum*, a new species from the honeybee (*Apis mellifera*). *International Journal of Systematic and Evolutionary Microbiology*. **35** (3): 296-308.

408. Humphery-Smith, I., Grulet, O., Le Goff, F. and Chastel, C. (1991). *Spiroplasma* (Mollicutes: Spiroplasmataceae) pathogenic for *Aedes aegypti* and *Anopheles stephensi* (Diptera: Culicidae). *Journal of Medical Entomology*. **28** (2): 219-22.

409. Anbutsu, H. and Fukatsu, T. (2003). Population dynamics of male-killing and non-male-killing spiroplasmas in *Drosophila melanogaster*. *Applied and Environmental Microbiology*. **69** (3): 1428-34.

410. Williamson, D. L., Sakaguchi, B., Hackett, K. J., Whitcomb, R. F., Tully, J. G., Carle, P.*, et al.* (1999). *Spiroplasma poulsonii* sp. nov., a new species associated

with male-lethality in *Drosophila willistoni,* a neotropical species of fruit fly. *International Journal of Systematic and Evolutionary Microbiology*. **49** (2): 611-8.

411. Tabata, J., Hattori, Y., Sakamoto, H., Yukuhiro, F., Fujii, T., Kugimiya, S.*, et al.* (2011). Male killing and incomplete inheritance of a novel *Spiroplasma* in the moth *Ostrinia zaguliaevi*. *Microbial Ecology*. **61** (2): 254-63.

412. Grimont, P. A. and Grimont, F. (1978). The genus serratia. *Annual Reviews in Microbiology*. **32** (1): 221-48.

413. Azambuja, P., Feder, D. and Garcia, E. (2004). Isolation of *Serratia marcescens* in the midgut of *Rhodnius prolixus*: impact on the establishment of the parasite *Trypanosoma cruzi* in the vector. *Experimental parasitology*. **107** (1-2): 89-96.

414. Azambuja, P., Garcia, E. S. and Ratcliffe, N. A. (2005). Gut microbiota and parasite transmission by insect vectors. *Trends in Parasitology*. **21** (12): 568-72.

415. Bando, H., Okado, K., Guelbeogo, W. M., Badolo, A., Aonuma, H., Nelson, B.*, et al.* (2013). Intra-specific diversity of *Serratia marcescens* in *Anopheles* mosquito midgut defines *Plasmodium* transmission capacity. *Scientific Reports*. **3** (1641): 1-9.

416. Gavini, F., Ferragut, C., Izard, D., Trinel, P., Leclerc, H., Lefebvre, B.*, et al.* (1979). *Serratia fonticola*, a new species from water. *International Journal of Systematic and Evolutionary Microbiology*. **29** (2): 92-101.

417. Kämpfer, P. and Glaeser, S. P. (2015). *Serratia glossinae* Geiger et al. 2010 is a later synonym of *Serratia fonticola* Gavini et al. 1979. *International Journal of Systematic and Evolutionary Microbiology*. **65** (5): 1406-8.

418. Poinar, G. O., Wassink, H. J., Leegwater-van der Linden, M. E. and van der Geest, L. P. (1979). *Serratia marcescens* as a pathogen of tsetse flies. *Acta Tropica*. **36** (3): 223-7.

419. Matthew, C. Z., Darby, A. C., Young, S. A., Hume, L. H. and Welburn, S. C. (2005). The rapid isolation and growth dynamics of the tsetse symbiont *Sodalis glossinidius*. *Fems Microbiology Letters*. **248** (1): 69-74.

420. Geiger, A., Ravel, S., Mateille, T., Janelle, J., Patrel, D., Cuny, G.*, et al.* (2007). Vector Competence of *Glossina palpalis gambiensis* for *Trypanosoma brucei* s.l. and Genetic Diversity of the Symbiont S*odalis glossinidius*. *Molecular Biology and Evolution*. **24** (1): 102-9.

421. Maltz, M. A., Weiss, B. L., O'Neill, M., Wu, Y. and Aksoy, S. (2012). OmpA-Mediated Biofilm Formation Is Essential for the Commensal Bacterium *Sodalis*

6047 *glossinidius* To Colonize the Tsetse Fly Gut. *Applied and Environmental*
6048 *Microbiology*. **78** (21): 7760-8.

6049 422. Ciosi, M., Masiga, D. K. and Turner, C. M. R. (2014). Laboratory Colonisation
6050 and Genetic Bottlenecks in the Tsetse Fly G*lossina pallidipes*. *PLOS Neglected*
6051 *Tropical Diseases*. **8** (2): e2697.

6052 423. Nei, M., Maruyama, T. and Chakraborty, R. (1975). The Bottleneck Effect and
6053 Genetic Variability in Populations. *Evolution*. **29** (1): 1-10.

6054 424. Symula, R. E., Alam, U., Brelsfoard, C., Wu, Y., Echodu, R., Okedi, L. M.*, et*
6055 *al.* (2013). *Wolbachia* association with the tsetse fly, *Glossina fuscipes fuscipes*,
6056 reveals high levels of genetic diversity and complex evolutionary dynamics. *BMC*
6057 *Evolutionary Biology*. **13** (1): 1-12.

6058 425. Haines, L. R. (2018). Method for culturing *Sodalis glossinidius* from tsetse
6059 flies. Personal communication to Pescod P [unpublished].

6060 426. Boncristiani, H., Li, J., Evans, J., Pettis, J. and Chen, Y. (2011). Scientific note
6061 on PCR inhibitors in the compound eyes of honey bees, *Apis mellifera*. *Apidologie*.
6062 **42** (4): 457-60.

6063 427. Matthew, C. Z. Biological and molecular aspects of *Sodalis glossinidius*.
6064 Edinburgh: University of Edinburgh; 2007.

6065 428. Zhou, W., Rousset, F. and O'Neill, S. (1998). Phylogeny and PCR–based
6066 classification of *Wolbachia* strains using *wsp* gene sequences. *Proceedings of the*
6067 *Royal Society of London Series B: Biological Sciences*. **265** (1395): 509-15.

6068 429. Mateos, M., Castrezana, S. J., Nankivell, B. J., Estes, A. M., Markow, T. A.
6069 and Moran, N. A. (2006). Heritable endosymbionts of *Drosophila*. *Genetics*. **174** (1):
6070 363-76.

6071 430. Fukatsu, T. and Nikoh, N. (2000). Endosymbiotic Microbiota of the Bamboo
6072 Pseudococcid A*ntonina crawii* (Insecta, Homoptera). *Applied and Environmental*
6073 *Microbiology*. **66** (2): 643-50.

6074 431. Zhu, H., Sun, S.-J. and Dang, H.-Y. (2008). PCR Detection of *Serratia* spp.
6075 Using Primers Targeting *pfs* and *luxS* Genes Involved in AI-2-Dependent Quorum
6076 Sensing. *Current Microbiology*. **57** (4): 326-30.

6077 432. Njiru, Z. K., Constantine, C. C., Guya, S., Crowther, J., Kiragu, J. M.,
6078 Thompson, R. C. A.*, et al.* (2005). The use of ITS1 rDNA PCR in detecting
6079 pathogenic African trypanosomes. *Parasitology Research*. **95** (3): 186-92.

433. FAO/IAEA. (2020). Standard operating procedures for identification of tsetse species from wild populations and laboratory colonies. Available from: https://www.iaea.org/sites/default/files/sop-for-tryp-id-2020.pdf.

434. Dyer, N. A., Lawton, S. P., Ravel, S., Choi, K. S., Lehane, M. J., Robinson, A. S.*, et al.* (2008). Molecular phylogenetics of tsetse flies (*Diptera: Glossinidae)* based on mitochondrial (COI, 16S, ND2) and nuclear ribosomal DNA sequences, with an emphasis on the palpalis group. *Mol Phylogenet Evol*. **49** (1): 227-39.

435. R Core Team. (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

436. R Studio Team. (2019). RStudio: Integrated Development for R. Boston, MA: RStudio, Inc.

437. Malele, I. I., Ouma, J. O., Nyingilili, H. S., Kitwika, W. A., Malulu, D. J., Magwisha, H. B.*, et al.* (2016). Comparative performance of traps in catching tsetse flies (Diptera: Glossinidae) in Tanzania. *The Onderstepoort Journal of Veterinary Research*. **83** (1): a1057-a.

438. Abila, P. P., Okello-Onen, J., Okoth, J. O., Matete, G. O., Wamwiri, F. and Politzar, H. (2007). Evaluation of the cost-effectiveness of pyramidal, modified pyramidal and monoscreen traps for the control of the tsetse fly, *Glossina fuscipes fuscipes*, in Uganda. *Journal of Insect Science*. **7** (47): 1-7.

439. Owaga, M. L. A. and Challier, A. (1985). Catch Composition of the Tsetse *Glossina pallidipes* Austen in Revolving and Stationary Traps With Respect to Age, Sex Ratio and Hunger Stage. *International Journal of Tropical Insect Science*. **6** (6): 711-8.

440. Williams, B., Dransfield, R. and Brightwell, R. (1992). The Control of Tsetse Flies in Relation to Fly Movement and Trapping Efficiency. *Journal of Applied Ecology*. **29** (1): 163-79.

441. Channumsin, M., Ciosi, M., Masiga, D., Turner, C. M. R. and Mable, B. K. (2018). *Sodalis glossinidius* presence in wild tsetse is only associated with presence of trypanosomes in complex interactions with other tsetse-specific factors. *BMC Microbiology*. **18** (1): 69-82.

442. Moran, N. A. and Wernegreen, J. J. (2000). Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends in Ecology & Evolution*. **15** (8): 321-6.

443. Morris, J. J., Lenski, R. E. and Zinser, E. R. (2012). The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. *mBio*. **3** (2): e00036-12.

444. Bourguignon, T., Kinjo, Y., Villa-Martín, P., Coleman, N. V., Tang, Q., Arab, D. A.*, et al.* (2020). Increased Mutation Rate Is Linked to Genome Reduction in Prokaryotes. *Current Biology*. **30** (19): 3848-55.

445. Lynch, M. and Gabriel, W. (1990). Mutation load and the survival of small populations. *Evolution*. **44** (7): 1725-37.

446. Bergstrom, C. T. and Pritchard, J. (1998). Germline Bottlenecks and the Evolutionary Maintenance of Mitochondrial Genomes. *Genetics*. **149** (4): 2135-46.

447. Peischl, S. and Kirkpatrick, M. (2012). Establishment of new mutations in changing environments. *Genetics*. **191** (3): 895-906.

448. Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*. **1** (1): 2-9.

449. Tutar, Y. (2012). Pseudogenes. *Comparative and Functional Genomics*. **2012** (424526): 1-4.

450. Sasidharan, R. and Gerstein, M. (2008). Protein fossils live on as RNA. *Nature*. **453** (7196): 729-31.

451. Mighell, A. J., Smith, N. R., Robinson, P. A. and Markham, A. F. (2000). Vertebrate pseudogenes. *FEBS Letters*. **468** (2): 109-14.

452. Lerat, E. and Ochman, H. (2005). Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Research*. **33** (10): 3125-32.

453. Cole, S., Eiglmeier, K., Parkhill, J., James, K., Thomson, N., Wheeler, P.*, et al.* (2001). Massive gene decay in the leprosy bacillus. *Nature*. **409** (6823): 1007-11.

454. Schneiker, S., Perlova, O., Kaiser, O., Gerth, K., Alici, A., Altmeyer, M. O.*, et al.* (2007). Complete genome sequence of the myxobacterium *Sorangium cellulosum*. *Nature Biotechnology*. **25** (11): 1281-9.

455. Bennett, G. M. and Moran, N. A. (2013). Small, Smaller, Smallest: The Origins and Evolution of Ancient Dual Symbioses in a Phloem-Feeding Insect. *Genome Biology and Evolution*. **5** (9): 1675-88.

456. Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H. E., Moran, N. A.*, et al.* (2006). The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science (New York, NY)*. **314** (5797): 267.

457. Wernegreen, J. J. (2002). Genome evolution in bacterial endosymbionts of insects. *Nature Reviews Genetics*. **3** (11): 850-61.

458. Clark, M. A., Moran, N. A. and Baumann, P. (1999). Sequence evolution in bacterial endosymbionts having extreme base compositions. *Molecular Biology and Evolution*. **16** (11): 1586-98.

459. Gillespie, J. J., Joardar, V., Williams Kelly, P., Driscoll, T., Hostetler Jessica, B., Nordberg, E.*, et al.* (2012). A *Rickettsia* Genome Overrun by Mobile Genetic Elements Provides Insight into the Acquisition of Genes Characteristic of an Obligate Intracellular Lifestyle. *Journal of Bacteriology*. **194** (2): 376-94.

460. Blanc, G., Ogata, H., Robert, C., Audic, S., Claverie, J.-M. and Raoult, D. (2007). Lateral gene transfer between obligate intracellular bacteria: evidence from the *Rickettsia massiliae* genome. *Genome Research*. **17** (11): 1657-64.

461. Kampfraath, A. A., Klasson, L., Anvar, S. Y., Vossen, R. H. A. M., Roelofs, D., Kraaijeveld, K.*, et al.* (2019). Genome expansion of an obligate parthenogenesis-associated *Wolbachia* poses an exception to the symbiont reduction model. *BMC genomics*. **20** (1): 1-14.

462. Ochman, H. and Groisman, E. A. (1996). Distribution of pathogenicity islands in *Salmonella* spp. *Infection and Immunity*. **64** (12): 5410-2.

463. Johansen, B. K., Wasteson, Y., Granum, P. E. and Brynestad, S. (2001). Mosaic structure of Shiga-toxin-2-encoding phages isolated from Escherichia coli O157: H7 indicates frequent gene exchange between lambdoid phage genomes. *Microbiology*. **147** (7): 1929-36.

464. Wagner, P. L. and Waldor, M. K. (2002). Bacteriophage control of bacterial virulence. *Infection and Immunity*. **70** (8): 3985-93.

465. Dale, C., Plague, G. R., Wang, B., Ochman, H. and Moran, N. A. (2002). Type III secretion systems and the evolution of mutualistic endosymbiosis. *Proceedings of the National Academy of Sciences*. **99** (19): 12397-402.

466. Foultier, B., Troisfontaines, P., Müller, S., Opperdoes, F. R. and Cornelis, G. R. (2002). Characterization of the ysa pathogenicity locus in the chromosome of *Yersinia enterocolitica* and phylogeny analysis of type III secretion systems. *Journal of Molecular Evolution*. **55** (1): 37-51.

467. Hensel, M. (2000). *Salmonella* pathogenicity island 2. *Molecular Microbiology*. **36** (5): 1015-23.

468. Oliver, K. M., Degnan, P. H., Hunter, M. S. and Moran, N. A. (2009). Bacteriophages Encode Factors Required for Protection in a Symbiotic Mutualism. *Science (New York, NY)*. **325** (5943): 992-4.

469. Fuqua, W. C., Winans, S. C. and Greenberg, E. P. (1994). Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. *Journal of Bacteriology*. **176** (2): 269-75.

470. Nealson, K. H., Platt, T. and Hastings, J. W. (1970). Cellular control of the synthesis and activity of the bacterial luminescent system. *Journal of Bacteriology*. **104** (1): 313-22.

471. Papenfort, K. and Bassler, B. L. (2016). Quorum sensing signal–response systems in Gram-negative bacteria. *Nature Reviews Microbiology*. **14** (9): 576-88.

472. Enomoto, S., Chari, A., Clayton, A. L. and Dale, C. (2017). Quorum sensing attenuates virulence in *Sodalis praecaptivus*. *Cell Host & Microbe*. **21** (5): 629-36.

473. Moran, N. A. (2002). Microbial Minimalism: Genome Reduction in Bacterial Pathogens. *Cell*. **108** (5): 583-6.

474. Lenski, R. E., Rose, M. R., Simpson, S. C. and Tadler, S. C. (1991). Long-term experimental evolution in *Escherichia coli.* I. Adaptation and divergence during 2,000 generations. *The American Naturalist*. **138** (6): 1315-41.

475. Elena, S. F. and Lenski, R. E. (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Reviews Genetics*. **4** (6): 457-69.

476. Chaston, J. and Douglas, A. E. (2012). Making the Most of “Omics” for Symbiosis Research. *The Biological Bulletin*. **223** (1): 21-9.

477. Medina, M. and Sachs, J. L. (2010). Symbiont genomics, our new tangled bank. *Genomics*. **95** (3): 129-37.

478. Chen, X., Hitchings, M. D., Mendoza, J. E., Balanza, V., Facey, P. D., Dyson, P. J.*, et al.* (2017). Comparative Genomics of Facultative Bacterial Symbionts Isolated from European *Orius* Species Reveals an Ancestral Symbiotic Association. *Frontiers in Microbiology*. **8** (1969): 1-14.

479. Smits, T. H. M. (2019). The importance of genome sequence quality to microbial comparative genomics. *BMC genomics*. **20** (662): 1-4.

480. Miyamoto, M., Motooka, D., Gotoh, K., Imai, T., Yoshitake, K., Goto, N.*, et al.* (2014). Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC genomics*. **15** (1): 1-9.

481. Loman, N. J., Quick, J. and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*. **12** (8): 733-5.

482. Tyson, J. R., O'Neil, N. J., Jain, M., Olsen, H. E., Hieter, P. and Snutch, T. P. (2018). MinION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. *Genome Research*. **28** (2): 266-74.

483. Wick, R. R., Judd, L. M., Gorrie, C. L. and Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*. **13** (6): e1005595.

484. Bashir, A., Klammer, A., Robins, W. P., Chin, C.-S., Webster, D., Paxinos, E.*, et al.* (2012). A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology*. **30** (7): 701-7.

485. Snyder, A. K. and Rio, R. V. M. (2015). *Wigglesworthia morsitans* Folate (Vitamin B) Biosynthesis Contributes to Tsetse Host Fitness. *Applied and Environmental Microbiology*. **81** (16): 5375-86.

486. Hall, R. J., Flanagan, L. A., Bottery, M. J., Springthorpe, V., Thorpe, S., Darby, A. C.*, 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.

487. Medina Munoz, M., Spencer, N., Enomoto, S., Dale, C. and Rio, R. V. M. (2020). Quorum sensing sets the stage for the establishment and vertical transmission of *Sodalis praecaptivus* in tsetse flies. *PLOS Genetics*. **16** (8): e1008992.

488. Goodhead, I. and Darby, A. C. (2015). Taking the pseudo out of pseudogenes. *Current Opinion in Microbiology*. **23**: 102-9.

489. Goodhead, I., Blow, F., Brownridge, P., Hughes, M., Kenny, J., Krishna, R.*, et al.* (2017). Large scale and significant expression from pseudogenes in *Sodalis glossinidius* - a facultative bacterial endosymbiont. *Microbial Genomics*. **6** (1): e000285.

490. Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. 0.11.4 ed: Babraham Bioinformatics.

491. Chen, S., Zhou, Y., Chen, Y. and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. **34** (17): i884-i90.

492. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. **30** (14): 2068-9.

493. Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G. and Parkhill, J. (2005). ACT: the Artemis Comparison Tool. *Bioinformatics*. **21** (16): 3422-3.

494. Seemann, T. (2017). Snippy: Rapid haploid variant calling and core genome alignment. 3.2 ed.

495. Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S.*, et al.* (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*. **49** (D1): D605-D12.

496. Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A.*, et al.* (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*. **49** (D1): D344-D54.

497. Wick, R. R., Schultz, M. B., Zobel, J. and Holt, K. E. (2015). Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics*. **31** (20): 3350-2.

498. Sullivan, M. J., Petty, N. K. and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics*. **27** (7): 1009-10.

499. Carver, T., Thomson, N., Bleasby, A., Berriman, M. and Parkhill, J. (2009). DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics*. **25** (1): 119-20.

500. Addgene. (2021). Addgene: Analyze sequence. Available from: https://www.addgene.org/analyze-sequence/.

501. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. and Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*. **10** (6): 845-58.

502. Zhang, Z., Schwartz, S., Wagner, L. and Webb, M. (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol*. **7** (1-2): 203-14.

503. Wu, D., Daugherty, S. C., Van Aken, S. E., Pai, G. H., Watkins, K. L., Khouri, H.*, et al.* (2006). Metabolic Complementarity and Genomics of the Dual Bacterial Symbiosis of Sharpshooters. *PLOS Biology*. **4** (6): e188.

504. Mbaye, M. N., Hou, Q., Basu, S., Teheux, F., Pucci, F. and Rooman, M. (2019). A comprehensive computational study of amino acid interactions in membrane proteins. *Scientific Reports*. **9** (1): 1-14.

505. Bing, X., Attardo, G. M., Vigneron, A., Aksoy, E., Scolari, F., Malacrida, A.*, et al.* (2017). Unravelling the relationship between the tsetse fly and its obligate symbiont *Wigglesworthia*: transcriptomic and metabolomic landscapes reveal highly integrated physiological networks. *Proc R Soc B*. **284** (1857): 1-10.

506. Belda, E., Silva, F. J., Peretó, J. and Moya, A. (2012). Metabolic Networks of *Sodalis glossinidius*: A Systems Biology Approach to Reductive Evolution. *PLOS ONE*. **7** (1): e30652.

507. Kvist, M., Hancock, V. and Klemm, P. (2008). Inactivation of Efflux Pumps Abolishes Bacterial Biofilm Formation. *Applied and Environmental Microbiology*. **74** (23): 7376-82.

508. Tamas, I., Wernegreen, J. J., Nystedt, B., Kauppinen, S. N., Darby, A. C., Gomez-Valero, L.*, et al.* (2008). Endosymbiont gene functions impaired and rescued by polymerase infidelity at poly(A) tracts. *Proc Natl Acad Sci U S A*. **105** (39): 14934-9.

509. Goodsell, D. S. (2009). *The machinery of life*: Springer Science & Business Media.

510. Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T.*, et al.* (2021). Structure-based protein function prediction using graph convolutional networks. *Nature Communications*. **12** (1): 1-14.

511. Peacock, L., Ferris, V., Bailey, M. and Gibson, W. (2012). The influence of sex and fly species on the development of trypanosomes in tsetse flies. *PLOS Negl Trop Dis*. **6** (2): e1515.

512. Okoth, J. 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): 459-61.

513. Bashiardes, S., Zilberman-Schapira, G. and Elinav, E. (2016). Use of Metatranscriptomics in Microbiome Research. *Bioinformatics and Biology Insights*. **10** (1): 19-25.

514. Soumana, I. H., Loriod, B., Ravel, S., Tchicaya, B., Simo, G., Rihet, P.*, et al.* (2014). The transcriptional signatures of *Sodalis glossinidius* in the *Glossina palpalis gambiensis* flies negative for *Trypanosoma brucei gambiense* contrast with those of this symbiont in tsetse flies positive for the parasite: Possible involvement of a *Sodalis*-hosted prophage in fly *Trypanosoma* refractoriness? *Infection, Genetics & Evolution*. **24**: 41-56.

515. Croucher, N. J. and Thomson, N. R. (2010). Studying bacterial transcriptomes using RNA-seq. *Current Opinion in Microbiology*. **13** (5): 619-24.

516. Mantione, K. J., Kream, R. M., Kuzelova, H., Ptacek, R., Raboch, J., Samuel, J. M.*, et al.* (2014). Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. *Medical Science Monitor Basic Research*. **20**: 138-41.

517. Medina Munoz, M., Brenner, C., Richmond, D., Spencer, N. and Rio, R. V. M. (2021). The holobiont transcriptome of teneral tsetse fly species of varying vector competence. *BMC genomics*. **22** (1): 400-.

518. Solano, P., Michel, J.-F., Lefrançois, T., de La Rocque, S., Sidibe, I., Zoungrana, A.*, et al.* (1999). Polymerase chain reaction as a diagnosis tool for detecting trypanosomes in naturally infected cattle in Burkina Faso. *Veterinary Parasitology*. **86** (2): 95-103.

519. Young, C. J. and Godfrey, D. G. (1983). Enzyme polymorphism and the distribution of *Trypanosoma congolense* isolates. *Annals of Tropical Medicine & Parasitology*. **77** (5): 467-81.

520. Maudlin, I. and Welburn, S. C. (1989). A single trypanosome is sufficient to infect a tsetse fly. *Annals of Tropical Medicine & Parasitology*. **83** (4): 431-3.

521. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet*. **17** (1): 10-2.

522. Joshi, N. A. and Fass, J. N. (2011). Sickle: a sliding window, adaptive, quality-based trimming tool for FastQ files. 1.200 ed. https://github.com/najoshi/sickle.

523. Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*. **9** (4): 357-9.

524. Anders, S., Pyl, P. T. and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*. **31** (2): 166-9.

525. David, P., Michael, M., Andrew, P. and Kim, S. (2019). Degust: interactive RNA-seq analysis. 4.1.4 ed: Zenodo.

526. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

527. Liniger, M., Acosta-Serrano, A., Abbeele, J. V. D., Renggli, C. K., Brun, R., Englund, P. T.*, et al.* (2003). Cleavage of trypanosome surface glycoproteins by alkaline trypsin-like enzyme(s) in the midgut of *Glossina morsitans*. *International Journal for Parasitology*. **33** (12): 1319-28.

528. Hu, C. and Aksoy, S. (2006). Innate immune responses regulate trypanosome parasite infection of the tsetse fly *Glossina morsitans morsitans*. *Molecular Microbiology*. **60** (5): 1194-204.

529. Bateta, R., Wang, J., Wu, Y., Weiss, B. L., Warren, W. C., Murilla, G. A.*, et al.* (2017). Tsetse fly (*Glossina pallidipes*) midgut responses to *Trypanosoma brucei* challenge. *Parasites & Vectors*. **10** (614): 1-12.

530. Lehane, M. J., Aksoy, S., Gibson, W., Kerhornou, A., Berriman, M., Hamilton, J.*, et al.* (2003). Adult midgut expressed sequence tags from the tsetse fly *Glossina morsitans morsitans* and expression analysis of putative immune response genes. *Genome Biology*. **4** (10): 1-10.

531. Roditi, I. and Lehane, M. J. (2008). Interactions between trypanosomes and tsetse flies. *Current Opinion in Microbiology*. **11** (4): 345-51.

532. Saier Jr, M. H. (2004). Evolution of bacterial type III protein secretion systems. *Trends in Microbiology*. **12** (3): 113-5.

533. Tomich, M., Herfst, C. A., Golden, J. W. and Mohr, C. D. (2002). Role of flagella in host cell invasion by *Burkholderia cepacia*. *Infect Immun*. **70** (4): 1799-806.

534. Song, Y. C., Jin, S., Louie, H., Ng, D., Lau, R., Zhang, Y.*, et al.* (2004). FlaC, a protein of *Campylobacter jejuni* TGH9011 (ATCC43431) secreted through the flagellar apparatus, binds epithelial cells and influences cell invasion. *Molecular Microbiology*. **53** (2): 541-53.

535. Oliver, K. M. and Higashi, C. H. V. (2019). Variations on a protective theme: *Hamiltonella defensa* infections in aphids variably impact parasitoid success. *Current Opinion in Insect Science*. **32**: 1-7.

536. Kent, B. N. and Bordenstein, S. R. (2010). Phage WO of *Wolbachia*: lambda of the endosymbiont world. *Trends in Microbiology*. **18** (4): 173-81.

537. Welburn, S. C. and Maudlin, I. (1991). *Rickettsia*-like organisms, puparial temperature and susceptibility to trypanosome infection in *Glossina morsitans*. *Parasitology*. **102** (2): 201-6.

538. Ziermann, R., Bartlett, B., Calendar, R. and Christie, G. E. (1994). Functions involved in bacteriophage P2-induced host cell lysis and identification of a new tail gene. *Journal of Bacteriology*. **176** (16): 4974-84.

539. Silpe, J. E. and Bassler, B. L. (2019). A Host-Produced Quorum-Sensing Autoinducer Controls a Phage Lysis-Lysogeny Decision. *Cell*. **176** (1): 268-80.

540. Dale, C., Jones, T. and Pontes, M. (2005). Degenerative evolution and functional diversification of type-III secretion systems in the insect endosymbiont *Sodalis glossinidius*. *Molecular Biology and Evolution*. **22** (3): 758-66.

541. Tucker, S. C. and Galán, J. E. (2000). Complex function for SicA*,* a *Salmonella enterica* serovar *typhimurium* type III secretion-associated chaperone. *Journal of Bacteriology*. **182** (8): 2262-8.

542. Kaniga, K., Tucker, S., Trollinger, D. and Galán, J. E. (1995). Homologs of the *Shigella* IpaB and IpaC invasins are required for *Salmonella typhimurium* entry into cultured epithelial cells. *Journal of Bacteriology*. **177** (14): 3965-71.

543. Shearer, J., Jefferies, D. and Khalid, S. (2019). Outer Membrane Proteins OmpA, FhuA, OmpF, EstA, BtuB, and OmpX Have Unique Lipopolysaccharide Fingerprints. *Journal of Chemical Theory and Computation*. **15** (4): 2608-19.

544. Achouak, W., Heulin, T. and Pagès, J.-M. (2001). Multiple facets of bacterial porins. *Fems Microbiology Letters*. **199** (1): 1-7.

545. Weiss, B. L., Wu, Y., Schwank, J. J., Tolwinski, N. S. and Aksoy, S. (2008). An insect symbiosis is influenced by bacterium-specific polymorphisms in outer-membrane protein A. *Proceedings of the National Academy of Sciences*. **105** (39): 15088-93.

546. Hejair, H. M. A., Zhu, Y., Ma, J., Zhang, Y., Pan, Z., Zhang, W.*, et al.* (2017). Functional role of ompF and ompC porins in pathogenesis of avian pathogenic *Escherichia coli*. *Microb Pathog*. **107**: 29-37.

547. Haiko, J. and Westerlund-Wikström, B. (2013). The role of the bacterial flagellum in adhesion and virulence. *Biology*. **2** (4): 1242-67.

548. Geiger, A., Cuny, G. and Frutos, R. (2005). Two tsetse fly species, *Glossina palpalis gambiensis* and G*lossina morsitans morsitans*, carry genetically distinct

populations of the secondary symbiont *Sodalis glossinidius*. *Applied and Environmental Microbiology*. **71** (12): 8941-3.

549. Medina Munoz, M., Pollio, A. R., White, H. L. and Rio, R. V. M. (2017). Into the Wild: Parallel Transcriptomics of the Tsetse-*Wigglesworthia* Mutualism within Kenyan Populations. *Genome Biology and Evolution*. **9** (9): 2276-91.

550. Soumana, I. H., Tchicaya, B., Simo, G. and Geiger, A. (2014). Comparative gene expression of *Wigglesworthia* inhabiting non-infected and *Trypanosoma brucei gambiense*-infected *Glossina palpalis gambiensis* flies. *Frontiers in Microbiology*. **5** (620): 1-16.

551. Soumana, I. H., Tchicaya, B., Loriod, B., Rihet, P. and Geiger, A. (2014). Identification of overexpressed genes in *Sodalis glossinidius* inhabiting trypanosome-infected self-cured tsetse flies. *Frontiers in Microbiology*. **5** (255): 1-8.

552. Ren, B., Wang, X., Duan, J. and Ma, J. (2019). Rhizobial tRNA-derived small RNAs are signal molecules regulating plant nodulation. *Science (New York, NY)*. **365** (6456): 919-22.

553. Dey, R., Joshi, A. B., Oliveira, F., Pereira, L., Guimarães-Costa, A. B., Serafim, T. D.*, et al.* (2018). Gut Microbes Egested during Bites of Infected Sand Flies Augment Severity of Leishmaniasis via Inflammasome-Derived IL-1β. *Cell Host & Microbe*. **23** (1): 134-43.

554. Saldaña, M. A., Hegde, S., Hughes, G. L., Saldaña, M. A., Hegde, S. and Hughes, G. L. (2017). Microbial control of arthropod-borne disease. *Memórias do Instituto Oswaldo Cruz*. **112** (2): 81-93.

555. Rodríguez-Ruano, S. M., Juhaňáková, E., Vávra, J. and Nováková, E. (2020). Methodological Insight Into Mosquito Microbiome Studies. *Frontiers in Cellular and Infection Microbiology*. **10** (86): 1-10.

556. Whittaker, R. H. (1972). Evolution and Measurement of Species Diversity. *Taxon*. **21** (2/3): 213-51.

557. Illumina. (2021). Illumina sequencing platforms. illumina.com. Available from: https://www.illumina.com/systems/sequencing-platforms.html.

558. Clooney, A. G., Fouhy, F., Sleator, R. D., O' Driscoll, A., Stanton, C., Cotter, P. D.*, et al.* (2016). Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. *PLOS ONE*. **11** (2): e0148028.

559. Sinha, R., Abu-Ali, G., Vogtmann, E., Fodor, A. A., Ren, B., Amir, A.*, et al.* (2017). Assessment of variation in microbial community amplicon sequencing by the

6441 Microbiome Quality Control (MBQC) project consortium. *Nature Biotechnology*. **35**
6442 (11): 1077-86.

6443 560. Riesenfeld, C. S., Schloss, P. D. and Handelsman, J. (2004). Metagenomics:
6444 Genomic Analysis of Microbial Communities. *Annual Review of Genetics*. **38** (1):
6445 525-52.

6446 561. Schloss, P. D. and Handelsman, J. (2003). Biotechnological prospects from
6447 metagenomics. *Current Opinion in Biotechnology*. **14** (3): 303-10.

6448 562. Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D. and Knight, R. (2007).
6449 Short pyrosequencing reads suffice for accurate microbial community analysis.
6450 *Nucleic Acids Research*. **35** (18): e120-e.

6451 563. Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz,
6452 P., Chen, L.*, et al.* (2019). Evaluation of 16S rRNA gene sequencing for species and
6453 strain-level microbiome analysis. *Nature Communications*. **10** (1): 1-11.

6454 564. Bokulich, N. A., Ziemski, M., Robeson, M. and Kaehler, B. (2020). Measuring
6455 the microbiome: best practices for developing and benchmarking microbiomics
6456 methods. *Computational and Structural Biotechnology Journal*. **18** (2020): 4048-62.

6457 565. Earth Microbiome Project. (2016). 16S Illumina Amplicon Protocol: Earth
6458 Microbiome Project. Available from: https://earthmicrobiome.org/protocols-and-
6459 standards/its/.

6460 566. Gohl, D. M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A.*, et al.*
6461 (2016). Systematic improvement of amplicon marker gene methods for increased
6462 accuracy in microbiome studies. *Nature Biotechnology*. **34** (9): 942-9.

6463 567. Menke, S., Gillingham, M. A. F., Wilhelm, K. and Sommer, S. (2017). Home-
6464 Made Cost Effective Preservation Buffer Is a Better Alternative to Commercial
6465 Preservation Methods for Microbiome Research. *Frontiers in Microbiology*. **8** (102):
6466 1-12.

6467 568. Kim, D., Hofstaedter, C. E., Zhao, C., Mattei, L., Tanes, C., Clarke, E.*, et al.*
6468 (2017). Optimizing methods and dodging pitfalls in microbiome research.
6469 *Microbiome*. **5** (52): 1-14.

6470 569. Yun, J.-H., Roh, S. W., Whon, T. W., Jung, M.-J., Kim, M.-S., Park, D.-S.*, et*
6471 *al.* (2014). Insect Gut Bacterial Diversity Determined by Environmental Habitat, Diet,
6472 Developmental Stage, and Phylogeny of Host. *Applied and Environmental*
6473 *Microbiology*. **80** (17): 5254-64.

570. Reese, A. T. and Dunn, R. R. (2018). Drivers of Microbiome Biodiversity: A Review of General Rules, Feces, and Ignorance. *mBio*. **9** (4): e01294.

571. Ibarra-Juarez, L. A., Desgarennes, D., Vázquez-Rosas-Landa, M., Villafan, E., Alonso-Sánchez, A., Ferrera-Rodríguez, O.*, et al.* (2018). Impact of Rearing Conditions on the Ambrosia Beetle's Microbiome. *Life (Basel, Switzerland)*. **8** (4): 1-22.

572. Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R. and Weyrich, L. S. (2019). Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology*. **27** (2): 105-17.

573. Olomu, I. N., Pena-Cortes, L. C., Long, R. A., Vyas, A., Krichevskiy, O., Luellwitz, R.*, et al.* (2020). Elimination of “kitome” and “splashome” contamination results in lack of detection of a unique placental microbiome. *BMC Microbiology*. **20** (1): 1-19.

574. Hornung, B. V. H., Zwittink, R. D. and Kuijper, E. J. (2019). Issues and current standards of controls in microbiome research. *FEMS microbiology ecology*. **95** (5): fiz045.

575. Stinson, L. F., Keelan, J. A. and Payne, M. S. (2019). Identification and removal of contaminating microbial DNA from PCR reagents: impact on low-biomass microbiome analyses. *Letters in applied microbiology*. **68** (1): 2-8.

576. Glassing, A., Dowd, S. E., Galandiuk, S., Davis, B. and Chiodini, R. J. (2016). Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut pathogens*. **8** (1): 1-2.

577. Waltmann, A., Willcox, A. C., Balasubramanian, S., Borrini Mayori, K., Mendoza Guerrero, S., Salazar Sanchez, R. S.*, et al.* (2019). Hindgut microbiota in laboratory-reared and wild *Triatoma infestans*. *PLOS Neglected Tropical Diseases*. **13** (5): e0007383.

578. González-Dominici, L. I., Saati-Santamaría, Z. and García-Fraile, P. (2021). Genome Analysis and Genomic Comparison of the Novel Species *Arthrobacter ipsi R*eveal Its Potential Protective Role in Its Bark Beetle Host. *Microbial Ecology*. **81** (2): 471-82.

579. Novakova, E., Woodhams, D. C. W., Rodriguez-Ruano, S. M., Brucker, R. M., Leff, J. W., Maharaj, A.*, et al.* (2017). Mosquito Microbiome Dynamics, a Background for Prevalence and Seasonality of West Nile Virus. *Frontiers in Microbiology*. **8** (526): 1-17.

580. Marche, M. G., Mura, M. E. and Ruiu, L. (2016). *Brevibacillus laterosporus* inside the insect body: Beneficial resident or pathogenic outsider? *Journal of Invertebrate Pathology*. **137**: 58-61.

581. Keller, A., Brandel, A., Becker, M. C., Balles, R., Abdelmohsen, U. R., Ankenbrand, M. J.*, et al.* (2018). Wild bees and their nests host *Paenibacillus* bacteria with functional potential of avail. *Microbiome*. **6** (1): 1-10.

582. Mann, A. E., Mitchell, E. A., Zhang, Y., Curtis-Robles, R., Thapa, S., Hamer, S. A.*, et al.* (2020). Comparison of the bacterial gut microbiome of North American *Triatoma* spp. with and without *Trypanosoma cruzi*. *Frontiers in Microbiology*. **11** (364): 1-10.

583. Segata, N., Baldini, F., Pompon, J., Garrett, W. S., Truong, D. T., Dabiré, R. K.*, et al.* (2016). The reproductive tracts of two malaria vectors are populated by a core microbiome and by gender- and swarm-enriched microbial biomarkers. *Scientific Reports*. **6** (1): 1-10.

584. Kaltenpoth, M. and Flórez, L. V. (2020). Versatile and Dynamic Symbioses Between Insects and *Burkholderia* Bacteria. *Annual review of entomology*. **65** (1): 145-70.

585. Hegde, S., Khanipov, K., Albayrak, L., Golovko, G., Pimenova, M., Saldaña, M. A.*, et al.* (2018). Microbiome Interaction Networks and Community Structure From Laboratory-Reared and Field-Collected *Aedes aegypti, Aedes albopictus,* and *Culex quinquefasciatus* Mosquito Vectors. *Frontiers in Microbiology*. **9** (2160): 1-16.

586. Straub, T. J., Shaw, W. R., Marcenac, P., Sawadogo, S. P., Dabiré, R. K., Diabaté, A.*, et al.* (2020). The *Anopheles coluzzii* microbiome and its interaction with the intracellular parasite *Wolbachia*. *Scientific Reports*. **10** (13847): 1-11.

587. Duguma, D., Rugman-Jones, P., Kaufman, M. G., Hall, M. W., Neufeld, J. D., Stouthamer, R.*, et al.* (2013). Bacterial Communities Associated with *Culex* Mosquito Larvae and Two Emergent Aquatic Plants of Bioremediation Importance. *PLOS ONE*. **8** (8): e72522.

588. Frankel-Bricker, J. (2020). Shifts in the microbiota associated with male mosquitoes (*Aedes aegypti*) exposed to an obligate gut fungal symbiont (*Zancudomyces culisetae*). *Scientific Reports*. **10** (1): 1-11.

589. Ivens, A. B. F., Gadau, A., Kiers, E. T. and Kronauer, D. J. C. (2018). Can social partnerships influence the microbiome? Insights from ant farmers and their trophobiont mutualists. *Molecular Ecology*. **27** (8): 1898-914.

590. Arévalo-Cortés, A., Mejia-Jaramillo, A. M., Granada, Y., Coatsworth, H., Lowenberger, C. and Triana-Chavez, O. (2020). The midgut microbiota of Colombian

6545 *Aedes aegypti* populations with different levels of resistance to the insecticide
6546 lambda-cyhalothrin. *Insects*. **11** (9): 1-19.

6547 591. Dumonteil, E., Ramirez-Sierra, M.-J., Pérez-Carrillo, S., Teh-Poot, C.,
6548 Herrera, C., Gourbière, S.*, et al.* (2018). Detailed ecological associations of
6549 triatomines revealed by metabarcoding and next-generation sequencing: implications
6550 for triatomine behavior and *Trypanosoma cruzi* transmission cycles. *Scientific*
6551 *Reports*. **8** (1): 1-13.

6552 592. Ferguson, L. V., Dhakal, P., Lebenzon, J. E., Heinrichs, D. E., Bucking, C.
6553 and Sinclair, B. J. (2018). Seasonal shifts in the insect gut microbiome are
6554 concurrent with changes in cold tolerance and immunity. *Functional Ecology*. **32**
6555 (10): 2357-68.

6556 593. Killer, J., Dubná, S., Sedláček, I. and Švec, P. (2014). *Lactobacillus apis* sp.
6557 nov., from the stomach of honeybees (Apis mellifera), having an in vitro inhibitory
6558 effect on the causative agents of American and European foulbrood. *Int J Syst Evol*
6559 *Microbiol*. **64** (Pt 1): 152-7.

6560 594. Minard, G., TRAN, F.-H., tran-van, v., Goubert, C., Bellet, C., Lambert, G.*, et*
6561 *al.* (2015). French invasive Asian tiger mosquito populations harbor reduced
6562 bacterial microbiota and genetic diversity compared to Vietnamese autochthonous
6563 relatives. *Frontiers in Microbiology*. **6** (970): 1-15.

6564 595. Sun, Z., Liu, Z., Zhou, W., Jin, H., Liu, H., Zhou, A.*, et al.* (2016). Temporal
6565 interactions of plant-insect-predator after infection of bacterial pathogen on rice
6566 plants. *Scientific Reports*. **6** (1): 1-12.

6567 596. de Lipthay, J. R., Enzinger, C., Johnsen, K., Aamand, J. and Sørensen, S. J.
6568 (2004). Impact of DNA extraction method on bacterial community composition
6569 measured by denaturing gradient gel electrophoresis. *Soil Biology and Biochemistry*.
6570 **36** (10): 1607-14.

6571 597. Feinstein, L. M., Sul, W. J. and Blackwood, C. B. (2009). Assessment of bias
6572 associated with incomplete extraction of microbial DNA from soil. *Applied and*
6573 *Environmental Microbiology*. **75** (16): 5428-33.

6574 598. Rubin, B. E. R., Sanders, J. G., Hampton-Marcell, J., Owens, S. M., Gilbert, J.
6575 A. and Moreau, C. S. (2014). DNA extraction protocols cause differences in 16S
6576 rRNA amplicon sequencing efficiency but not in community profile composition or
6577 structure. *MicrobiologyOpen*. **3** (6): 910-21.

6578 599. Salonen, A., Nikkilä, J., Jalanka-Tuovinen, J., Immonen, O., Rajilić-Stojanović,
6579 M., Kekkonen, R. A.*, et al.* (2010). Comparative analysis of fecal DNA extraction
6580 methods with phylogenetic microarray: effective recovery of bacterial and archaeal
6581 DNA using mechanical cell lysis. *Journal of microbiological methods*. **81** (2): 127-34.

600. Scupham, A., Jones, J. and Wesley, I. (2007). Comparison of DNA extraction methods for analysis of turkey cecal microbiota. *Journal of applied microbiology*. **102** (2): 401-9.

601. Guégan, M., Zouache, K., Démichel, C., Minard, G., Potier, P., Mavingui, P.*, et al.* (2018). The mosquito holobiont: fresh insight into mosquito-microbiota interactions. *Microbiome*. **6** (1): 1-17.

602. Yokouchi, H., Fukuoka, Y., Mukoyama, D., Calugay, R., Takeyama, H. and Matsunaga, T. (2006). Whole-metagenome amplification of a microbial community associated with scleractinian coral by multiple displacement amplification using ϕ29 polymerase. *Environmental Microbiology*. **8** (7): 1155-63.

603. Bowers, R. M., Clum, A., Tice, H., Lim, J., Singh, K., Ciobanu, D.*, et al.* (2015). Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC genomics*. **16** (856): 1-12.

604. Yilmaz, S., Allgaier, M. and Hugenholtz, P. (2010). Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nature Methods*. **7** (12): 943-4.

605. Walker, A. W., Martin, J. C., Scott, P., Parkhill, J., Flint, H. J. and Scott, K. P. (2015). 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome*. **3** (1): 1-11.

606. Schloss, P. D., Gevers, D. and Westcott, S. L. (2011). Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLOS ONE*. **6** (12): e27310.

607. Pettengill, J. B., McAvoy, E., White, J. R., Allard, M., Brown, E. and Ottesen, A. (2012). Using metagenomic analyses to estimate the consequences of enrichment bias for pathogen detection. *BMC research notes*. **5** (378): 1-7.

608. Hughes, J. B. and Hellmann, J. J. (2005). The Application of Rarefaction Techniques to Molecular Inventories of Microbial Diversity. Ed: *Methods in Enzymology*. vol. 397.

609. Zaheer, R., Noyes, N., Ortega Polo, R., Cook, S. R., Marinier, E., Van Domselaar, G.*, et al.* (2018). Impact of sequencing depth on the characterization of the microbiome and resistome. *Scientific Reports*. **8** (5890): 1-11.

610. Carlsen, T., Aas, A. B., Lindner, D., Vrålstad, T., Schumacher, T. and Kauserud, H. (2012). Don't make a mista (g) ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecology*. **5** (6): 747-9.

611. Robin, J. D., Ludlow, A. T., LaRanger, R., Wright, W. E. and Shay, J. W. (2016). Comparison of DNA Quantification Methods for Next Generation Sequencing. *Scientific Reports*. **6** (24067): 1-10.

612. Hussing, C., Kampmann, M. L., Mogensen, H. S., Børsting, C. and Morling, N. (2015). Comparison of techniques for quantification of next-generation sequencing libraries. *Forensic Science International: Genetics Supplement Series*. **5**: e276-e8.

613. Nakayama, Y., Yamaguchi, H., Einaga, N. and Esumi, M. (2016). Pitfalls of DNA Quantification Using DNA-Binding Fluorescent Dyes and Suggested Solutions. *PLOS ONE*. **11** (3): e0150528.

614. Jones, M. B., Highlander, S. K., Anderson, E. L., Li, W., Dayrit, M., Klitgord, N.*, et al.* (2015). Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proceedings of the National Academy of Sciences*. **112** (45): 14024-9.

615. Illumina. (2021). Sequencing by Synthesis (SBS) Technology. Available from: https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html.

616. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G.*, et al.* (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science (New York, NY)*. **323** (5910): 133-8.

617. Oxford Nanopore Technologies. (2020). How does nanopore DNA/RNA sequencing work? nanoporetech.com. Available from: https://nanoporetech.com/how-it-works.

618. Schloss, P. D., Jenior, M. L., Koumpouras, C. C., Westcott, S. L. and Highlander, S. K. (2016). Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. *PeerJ*. **4**: e1869.

619. Al-Nahhas, S. A. and Kaldas, R. M. (2013). Characterization of *Leishmania* Species Isolated from Cutaneous Human Samples from Central Region of Syria by RFLP Analysis. *International Scholarly Research Notices*. **2013** (1): e308726.

620. Tiwary, P., Kumar, D., Singh, R. P., Rai, M. and Sundar, S. (2012). Prevalence of Sand Flies and *Leishmania donovani* Infection in a Natural Population of Female *Phlebotomus argentipes* in Bihar State, India. *Vector Borne and Zoonotic Diseases*. **12** (6): 467-72.

621. Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J.*, et al.* (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*. **108** (S1): 4516-22.

622. Parada, A. E., Needham, D. M. and Fuhrman, J. A. (2016). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology*. **18** (5): 1403-14.

623. Rohland, N. and Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*. **22** (5): 939-46.

624. Kozich, J. J., Schloss, P., Baxter, N. T., Jenior, M. L., Koumpouras, C. C. and Bishop, L. MiSeq Wet Lab SOP [Internet]. Bishop L, editor. github.com2013 8-August-2019. [cited 2021]. Available from: https://github.com/SchlossLab/MiSeq_WetLab_SOP/blob/master/MiSeq_WetLab_SOP.md.

625. Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A. and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*. **13** (7): 581-3.

626. Edgar, R. C. and Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*. **31** (21): 3476-82.

627. Wang, Q., Garrity, G. M., Tiedje, J. M. and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*. **73** (16): 5261-7.

628. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P.*, et al.* (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*. **41** (D1): D590-D6.

629. Callahan, B. J. (2018). Silva taxonomic training data formatted for DADA2 (Silva version 132). Zenodo. Available from: https://zenodo.org/record/1172783.

630. McMurdie, P. J. and Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE*. **8** (4): e61217.

631. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. and Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*. **6** (226): 1-14.

632. Lahti, L., Shetty, S., Blake, T. and Salojarvi, J. (2012). microbiome R package.

633. Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D.*, et al.* (2019). vegan: Community Ecology Package.

634. Karimian, F., Vatandoost, H., Rassi, Y., Maleki-Ravasan, N., Mohebali, M., Shirazi, M. H.*, et al.* (2019). Aerobic midgut microbiota of sand fly vectors of zoonotic visceral leishmaniasis from northern Iran, a step toward finding potential paratransgenic candidates. *Parasites & Vectors*. **12** (1): 1-12.

635. Navas-Molina, J. A., Peralta-Sánchez, J. M., González, A., McMurdie, P. J., Vázquez-Baeza, Y., Xu, Z.*, et al.* (2013). Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol*. **531**: 371-444.

636. McMurdie, P. J. and Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Computational Biology*. **10** (4): e1003531.

637. Van Calster, B., Steyerberg, E. W., Collins, G. S. and Smits, T. (2018). Consequences of relying on statistical significance: Some illustrations. *European Journal of Clinical Investigation*. **48** (5): e12912.

638. Willis, A. D. (2019). Rigorous Statistical Methods for Rigorous Microbiome Science. *mSystems*. **4** (3): e00117-19.

639. Willis, A. D. and Martin, B. D. (2020). Estimating diversity in networked ecological communities. *Biostatistics*. **00** (00): 1-17.

640. WHO. (2019). Leishmaniasis. WHO, Available from: https://www.who.int/news-room/fact-sheets/detail/leishmaniasis.

641. Pigott, D. M., Bhatt, S., Golding, N., Duda, K. A., Battle, K. E., Brady, O. J.*, et al.* (2014). Global distribution maps of the leishmaniases. *eLife*. **3**: e02851.

642. National Vector Borne Disease Control Programme (2019). Kala-Azar situation in India. Directorate General of Health Services MoHaFW, Government of India, nvbdcp.gov.in. Available from: https://nvbdcp.gov.in/index4.php?lang=1&level=0&linkid=467&lid=3750.

643. Narain, J. P., Dash, A. P., Parnell, B., Bhattacharya, S. K., Barua, S., Bhatia, R.*, et al.* (2010). Elimination of neglected tropical diseases in the South-East Asia Region of the World Health Organization. *Bulletin of the World Health Organization*. **88** (3): 206-10.

644. Singh, S. P., Picado, A., Boelaert, M., Gidwani, K., Andersen, E. W., Ostyn, B.*, et al.* (2010). The epidemiology of *Leishmania donovani* infection in high transmission foci in India. *Tropical Medicine & International Health*. **15** (2): 12-20.

645. Singh, S. P., Reddy, D. C. S., Rai, M. and Sundar, S. (2006). Serious underreporting of visceral leishmaniasis through passive case reporting in Bihar, India. *Tropical Medicine & International Health*. **11** (6): 899-905.

646. Singh, V. P., Ranjan, A., Topno, R. K., Verma, R. B., Siddique, N. A., Ravidas, V. N.*, et al.* (2010). Estimation of under-reporting of visceral leishmaniasis cases in Bihar, India. *The American Journal of Tropical Medicine and Hygiene*. **82** (1): 9-11.

647. Wamwiri, F. N., Ndungu, K., Thande, P. C., Thungu, D. K., Auma, J. E. and Ngure, R. M. (2014). Infection With the Secondary Tsetse-Endosymbiont *Sodalis glossinidius* (Enterobacteriales: Enterobacteriaceae) Influences Parasitism in *Glossina pallidipes* (Diptera: Glossinidae). *Journal of Insect Science*. **14** (272): 1-4.

648. Dennison, N. J., Jupatanakul, N. and Dimopoulos, G. (2014). The mosquito microbiota influences vector competence for human pathogens. *Current Opinion in Insect Science*. **3** (1): 6-13.

649. Bhunia, G. S., Kesari, S., Chatterjee, N., Kumar, V. and Das, P. (2013). Spatial and temporal variation and hotspot detection of kala-azar disease in Vaishali district (Bihar), India. *BMC Infectious Diseases*. **13** (64): 1-12.

650. Maleki-Ravasan, N., Ali Oshaghi, M., Hajikhani, S., Saeidi, Z., Ahmad Akhavan, A., Gerami-Shoar, M.*, et al.* (2014). Aerobic Microbial Community of Insectary Population of *Phlebotomus papatasi*. *Journal of Arthropod-Borne Diseases*. **8** (1): 69-81.

651. Sant'Anna, M. R. V., Darby, A. C., Brazil, R. P., Montoya-Lerma, J., Dillon, V. M., Bates, P. A.*, et al.* (2012). Investigation of the Bacterial Communities Associated with Females of *Lutzomyia* Sand Fly Species from South America. *PLOS ONE*. **7** (8): e42531.

652. Li, K., Chen, H., Jiang, J., Li, X., Xu, J. and Ma, Y. (2016). Diversity of bacteriome associated with *Phlebotomus chinensis* (Diptera: Psychodidae) sand flies in two wild populations from China. *Scientific Reports*. **6** (1): 1-8.

653. McCarthy, C. B., Diambra, L. A. and Pomar, R. V. R. (2011). Metagenomic Analysis of Taxa Associated with *Lutzomyia longipalpis*, Vector of Visceral Leishmaniasis, Using an Unbiased High-Throughput Approach. *PLOS Neglected Tropical Diseases*. **5** (9): e1304.

654. Ondov, B. D., Bergman, N. H. and Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC bioinformatics*. **12** (1): 1-10.

655. La Scola, B., Birtles, R. J., Mallet, M.-N. and Raoult, D. (1998). *Massilia timonae gen. nov., sp. nov.*, Isolated from Blood of an Immunocompromised Patient with Cerebellar Lesions. *Journal of Clinical Microbiology*. **36** (10): 2847-52.

656. Tamura, A., Ohashi, N., Urakami, H. and Miyamura, S. (1995). Classification of *Rickettsia tsutsugamushi* in a new genus, *Orientia gen. nov.*, as *Orientia*

6757 *tsutsugamushi comb. nov*. *International Journal of Systematic Bacteriology*. **45** (3):
6758 589-91.

6759 657. Yabuuchi, E., Yano, I., Oyaizu, H., Hashimoto, Y., Ezaki, T. and Yamamoto,
6760 H. (1990). Proposals of *Sphingomonas paucimobilis gen. nov*. and *comb. nov*.,
6761 *Sphingomonas parapaucimobilis sp. nov., Sphingomonas yanoikuyae sp. nov.,*
6762 *Sphingomonas adhaesiva sp. nov., Sphingomonas capsulata comb, nov*., and Two
6763 Genospecies of the Genus *Sphingomonas*. *Microbiology and Immunology*. **34** (2):
6764 99-119.

6765 658. Saglio, P., Lhospital, M., Laflèche, D., Dupont, G., Bové, J. M., Tully, J. G*., et*
6766 *al.* (1973). *Spiroplasma citri gen*. and *sp. n*.: A *Mycoplasma*-Like Organism
6767 Associated with “Stubborn” Disease of Citrus. *International Journal of Systematic*
6768 *and Evolutionary Microbiology*. **23** (3): 191-204.

6769 659. Rosenbach, R. J. (1884). *Mikro-organismen bei den Wund-Infections-*
6770 *Krankheiten des Menschen*. Wiesbaden: Wiesbaden J. F. Bergmann.

6771 660. Gunathilaka, P., Wijerathna, A., Perera, H. and Rodrigo, W. (2019). Bacterial
6772 Diversity in the Lumen of Wild Caught Leishmaniasis Vector *Phlebotomus argentipes*
6773 (Psychodidae: Phlebotominae). Proceedings of 12th International Research
6774 Conference 2019, KDU; Sri Lanka: IRKDU. 1174-9.

6775 661. Braga, G. d. S., Alves, L. and da Silva, L. (2011). Fungal and bacterial agents
6776 in a laboratory colony of *Lutzomyia longipalpis*, Imperatriz population. *PUBVET*. **5**
6777 (4): 1-12.

6778 662. Vivero, R. J., Jaramillo, N. G., Cadavid-Restrepo, G., Soto, S. I. U. and
6779 Herrera, C. X. M. (2016). Structural differences in gut bacteria communities in
6780 developmental stages of natural populations of *Lutzomyia evansi* from Colombia's
6781 Caribbean coast. *Parasites & Vectors*. **9** (1): 1-20.

6782 663. Brisou, J. and Prévot, A. R. (1954). Études de systématique bactérienne. X.
6783 Révision des espèces réunies dans le genre *Achromobacter*. *Annales de l’Institut*
6784 *Pasteur (Paris)*. **86** (6): 722-8.

6785 664. Philip, C. B. (1956). Comments on the classification of the Order Rickettsiales.
6786 *Canadian Journal of Microbiology*. **2** (3): 261-70.

6787 665. Bouchon, D., Cordaux, R. and Grève, P. (2011). *Rickettsiella*, Intracellular
6788 Pathogens of Arthropods. Ed: Zchori-Fein E, Bourtzis K, In: *Manipulative Tenants:*
6789 *Bacteria Associated with Arthropods*. vol.

6790 666. Lin, S.-Y., Hameed, A., Arun, A. B., Liu, Y.-C., Hsu, Y.-H., Lai, W.-A*., et al.*
6791 (2013). Description of *Noviherbaspirillum malthae gen. nov., sp. nov*., isolated from
6792 an oil-contaminated soil, and proposal to reclassify *Herbaspirillum soli,*

Herbaspirillum aurantiacum, Herbaspirillum canariense and *Herbaspirillum psychrotolerans* as *Noviherbaspirillum soli comb. nov.*, *Noviherbaspirillum aurantiacum comb. nov., Noviherbaspirillum canariense comb. nov.* and *Noviherbaspirillum psychrotolerans comb. nov.* based on polyphasic analysis. *International Journal of Systematic and Evolutionary Microbiology*. **63** (11): 4100-7.

667. Migula, W. (1894). Über ein neues System der Bakterien. *Arbeiten aus dem Bakteriologischen Institut der Technischen Hochschule zu Karlsruhe*. **1**: 235-8.

668. Lehmann, K. B. and Neumann, R. O. (1896). *Atlas und Grundriss der Bakteriologie und Lehrbuch der speciellen bakteriologischen Diagnostik*. 1st ed. München: Lehmann.

669. Hertig, M. and Wolbach, S. B. (1924). Studies on *Rickettsia*-Like Micro-Organisms in Insects. *The Journal of medical research*. **44** (3): 329-74.

670. Parvizi, P., Fardid, F. and Soleimani, S. (2013). Detection of a new strain of *Wolbachia pipientis* in *Phlebotomus perfiliewi transcaucasicus*, a potential vector of visceral leishmaniasis in North West of Iran, by targeting the major surface protein gene. *Journal of Arthropod-Borne Diseases*. **7** (1): 46-55.

671. Da Rocha, N. O., Lambert, S. M., Dias-Lima, A. G., Julião, F. S. and Souza, B. M. P. S. (2018). Molecular detection of *Wolbachia pipientis* in natural populations of sandfly vectors of *Leishmania infantum* in endemic areas: first detection in *Lutzomyia longipalpis*. *Medical and Veterinary Entomology*. **32** (1): 111-4.

672. Segers, P., Vancanneyt, M., Pot, B., Torck, U., Hoste, B., Dewettinck, D.*, et al.* (1994). Classification of *Pseudomonas diminuta* Leifson and Hugh 1954 and *Pseudomonas vesicularis* Büsing, Döll, and Freytag 1953 in *Brevundimonas gen. nov.* as *Brevundimonas diminuta comb. nov.* and *Brevundimonas vesicularis comb. nov.*, Respectively. *International Journal of Systematic and Evolutionary Microbiology*. **44** (3): 499-510.

673. Palleroni, N. J. and Bradbury, J. F. (1993). *Stenotrophomonas*, a New Bacterial Genus for *Xanthomonas maltophilia* (Hugh 1980) Swings *et al.* 1983. *International Journal of Systematic and Evolutionary Microbiology*. **43** (3): 606-9.

674. Karakuş, M., Karabey, B., Kalkan, Ş. O., Özdemir, G., Oğuz, G., Kasap, Ö. E.*, et al.* (2017). Midgut Bacterial Diversity of Wild Populations of *Phlebotomus (P.) papatasi*, the Vector of Zoonotic Cutaneous Leishmaniasis (ZCL) in Turkey. *Scientific Reports*. **7** (1): 1-7.

675. Engel, P. and Moran, N. A. (2013). The gut microbiota of insects – diversity in structure and function. *FEMS Microbiology Reviews*. **37** (5): 699-735.

676. Minard, G., Mavingui, P. and Moro, C. V. (2013). Diversity and function of bacterial microbiota in the mosquito holobiont. *Parasites & Vectors*. **6** (146): 1-12.

677. Lehane, M. J. (2005). *The Biology of Blood-Sucking Insects*. 2nd ed. New York: Cambridge University Press.

678. Arora, A. K., Forshaw, A., Miller, T. A. and Durvasula, R. (2015). A delivery system for field application of paratransgenic control. *BMC Biotechnology*. **15** (59): 1-11.

679. Masai, E., Katayama, Y. and Fukuda, M. (2007). Genetic and biochemical investigations on bacterial catabolic pathways for lignin-derived aromatic compounds. *Bioscience, biotechnology, and biochemistry*. **71** (1): 1-15.

680. Albuquerque, T. A. and Zurek, L. (2014). Temporal changes in the bacterial community of animal feces and their correlation with stable fly oviposition, larval development, and adult fitness. *Frontiers in Microbiology*. **5** (590): 1-9.

681. Vary, P. S., Biedendieck, R., Fuerch, T., Meinhardt, F., Rohde, M., Deckwer, W.-D*., et al.* (2007). *Bacillus megaterium* - from simple soil bacterium to industrial protein production host. *Applied Microbiology and Biotechnology*. **76** (5): 957-67.

682. Bikash, C., Ghosh, T., Sienkiewicz, T. and Krenkel, K. (2000). *Brevibacterium linens* - a useful enzyme producer for cheese: a review. *Milchwissenschaft*. **55** (11): 628-32.

683. Schallmey, M., Singh, A. and Ward, O. P. (2004). Developments in the use of *Bacillus* species for industrial production. *Canadian Journal of Microbiology*. **50** (1): 1-17.

684. Cruaud, A., Gonzalez, A.-A., Godefroid, M., Nidelet, S., Streito, J.-C., Thuillier, J.-M*., et al.* (2018). Using insects to detect, monitor and predict the distribution of *Xylella fastidiosa*: a case study in Corsica. *Scientific Reports*. **8** (1): 1-13.

685. Braks, M., van der Giessen, J., Kretzschmar, M., van Pelt, W., Scholte, E.-J., Reusken, C*., et al.* (2011). Towards an integrated approach in surveillance of vector-borne diseases in Europe. *Parasites & Vectors*. **4** (1): 1-11.

686. Izzard, L., Fuller, A., Blacksell, S. D., Paris, D. H., Richards, A. L., Aukkanit, N*., et al.* (2010). Isolation of a Novel *Orientia* Species (*O. chuto* sp. nov.) from a Patient Infected in Dubai. *Journal of Clinical Microbiology*. **48** (12): 4404-9.

687. Xu, G., Walker, D. H., Jupiter, D., Melby, P. C. and Arcari, C. M. (2017). A review of the global epidemiology of scrub typhus. *PLOS Neglected Tropical Diseases*. **11** (11): e0006062.

688. Pham, X. D., Otsuka, Y., Suzuki, H. and Takaoka, H. (2001). Detection of *Orientia tsutsugamushi* (Rickettsiales: Rickettsiaceae) in Unengorged Chiggers (Acari: Trombiculidae) from Oita Prefecture, Japan, by Nested Polymerase Chain Reaction. *Journal of Medical Entomology*. **38** (2): 308-11.

689. Jain, P., Prakash, S., Tripathi, P. K., Chauhan, A., Gupta, S., Sharma, U.*, et al.* (2018). Emergence of *Orientia tsutsugamushi* as an important cause of Acute Encephalitis Syndrome in India. *PLOS Neglected Tropical Diseases*. **12** (3): e0006346.

690. BBC. (2019). India encephalitis death toll rises to 103.

691. Paris, D. H., Shelite, T. R., Day, N. P. and Walker, D. H. (2013). Unresolved Problems Related to Scrub Typhus: A Seriously Neglected Life-Threatening Disease. *The American Journal of Tropical Medicine and Hygiene*. **89** (2): 301-7.

692. Palit, A., Bhattacharya, S. K. and Kundu, S. N. (2005). Host preference of *Phlebotomus argentipes* and *Phlebotomus papatasi* in different biotopes of West Bengal, India. *International journal of environmental health research*. **15** (6): 449-54.

693. Garlapati, R. B., Abbasi, I., Warburg, A., Poché, D. and Poché, R. (2012). Identification of Bloodmeals in Wild Caught Blood Fed P*hlebotomus argentipes* (Diptera: Psychodidae) Using Cytochrome b PCR and Reverse Line Blotting in Bihar, India. *Journal of Medical Entomology*. **49** (3): 515-21.

694. Elliott, I., Pearson, I., Dahal, P., Thomas, N. V., Roberts, T. and Newton, P. N. (2019). Scrub typhus ecology: a systematic review of *Orientia* in vectors and hosts. *Parasites & Vectors*. **12** (1): 1-36.

695. Edouard, S. and Raoult, D. (2010). *Bartonella henselae*, un agent d'infections ubiquitaires. *Médecine et Maladies Infectieuses*. **40** (6): 319-30.

696. Billeter, S. A., Levy, M. G., Chomel, B. B. and Breitschwerdt, E. B. (2008). Vector transmission of *Bartonella* species with emphasis on the potential for tick transmission. *Medical and Veterinary Entomology*. **22** (1): 1-15.

697. Townsend, C. H. T. (1913). The transmission of Verruga by *Phlebotomus*. *Journal of the American Medical Association*. **61** (19): 1717-8.

698. Fukatsu, T., Tsuchida, T., Nikoh, N. and Koga, R. (2001). *Spiroplasma* Symbiont of the Pea Aphid, *Acyrthosiphon pisum* (Insecta: Homoptera). *Applied and Environmental Microbiology*. **67** (3): 1284-91.

699. Weiss, E., Dasch, G. A. and Chang, K. P. (1984). *Genus VIII. Rickettsiella Philip 1956*. Baltimore, MD: The Williams & Wilkins Co.

700. Lacey, L. A. (1997). *Manual of Techniques in Insect Pathology*: Academic Press.

701. Duron, O., Bouchon, D., Boutin, S., Bellamy, L., Zhou, L., Engelstädter, J.*, et al.* (2008). The diversity of reproductive parasites among arthropods: *Wolbachia* do not walk alone. *BMC biology*. **6** (1): 1-12.

702. Karatepe, B., Aksoy, S. and Karatepe, M. (2018). Investigation of *Wolbachia spp.* and *Spiroplasma spp.* in *Phlebotomus* species by molecular methods. *Scientific Reports*. **8** (1): 1-5.

703. Fournier, P.-E. and Raoult, D. (2015). *Rickettsiella*. Ed: Whitman WB, Rainey F, Kämpfer P, Trujillo M, Chun J, DeVos P, et al., In: *Bergey's Manual of Systematics of Archaea and Bacteria*. vol.

704. Chaisiri, K., Gill, A. C., Stekolnikov, A. A., Hinjoy, S., McGarry, J. W., Darby, A. C.*, et al.* (2019). Ecological and microbiological diversity of chigger mites, including vectors of scrub typhus, on small mammals across stratified habitats in Thailand. *Animal Microbiome*. **1** (1): 1-17.

705. Shatrov, A. B. and Kudryashova, N. I. (2006). Taxonomy, life cycles and the origin of parasitism in trombiculid mites. Ed: *Micromammals and macroparasites*. vol.

706. Kumlert, R., Chaisiri, K., Anantatat, T., Stekolnikov, A. A., Morand, S., Prasartvit, A.*, et al.* (2018). Autofluorescence microscopy for paired-matched morphological and molecular identification of individual chigger mites (Acari: Trombiculidae), the vectors of scrub typhus. *PLOS ONE*. **13** (3): e0193163.

707. Karstens, L., Asquith, M., Davin, S., Fair, D., Gregory, W. T., Wolfe Alan, J.*, et al.* (2019). Controlling for Contaminants in Low-Biomass 16S rRNA Gene Sequencing Experiments. *mSystems*. **4** (4): e00290-19.

708. Tláskal, V., Pylro, V. S., Žifčáková, L. and Baldrian, P. (2021). Ecological Divergence Within the Enterobacterial Genus *Sodalis*: From Insect Symbionts to Inhabitants of Decomposing Deadwood. *Frontiers in Microbiology*. **12** (1499): 1-13.

709. Rutherford, S. T. and Bassler, B. L. (2012). Bacterial quorum sensing: its role in virulence and possibilities for its control. *Cold Spring Harbor perspectives in medicine*. **2** (11): a012427.

710. Wamwiri, F. N., Alam, U., Thande, P. C., Aksoy, E., Ngure, R. M., Aksoy, S.*, et al.* (2013). *Wolbachia*, *Sodalis* and trypanosome co-infections in natural populations of *Glossina austeni* and *Glossina pallidipes*. *Parasites & Vectors*. **6** (1): 1-11.

Appendix

Table S4.1 – The annotated genome sources used to analyse the origin of RNA reads in the tsetse metatranscription project.

Organism	Database	Genome URL
Glossina morsitans morsitans	VectorBase	https://www.vectorbase.org/downloads?field_organism_taxonomy_tid%5B%5D=GmorY1&field_status_value=Current
Serratia marcescens	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCF_000513215.1/
Sodalis glossinidius	NCBI	https://www.ncbi.nlm.nih.gov/genome/?term=*Sodalis*+glossinidius+str.+%27morsitans%27
Trypanosoma congolense	TriTrypDB	https://tritrypdb.org/common/downloads/release-47/TcongolenseIL3000/
Wigglesworthia glossindia	NCBI	https://www.ncbi.nlm.nih.gov/genome/?term=wigglesworthia+glossinidia
Wolbachia	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCF_000689175.1/