

**WestminsterResearch**

<http://www.westminster.ac.uk/westminsterresearch>

**The evolution of immune genes in tsetse flies (*Glossina*) and  
insights into tsetse-symbiont-trypanosome interactions**

**Bruce, Robert**

This is a PhD thesis awarded by the University of Westminster.

© Mr Robert Bruce, 2022.

<https://doi.org/10.34737/w060x>

The WestminsterResearch online digital archive at the University of Westminster aims to make the research output of the University available to a wider audience. Copyright and Moral Rights remain with the authors and/or copyright owners.

The evolution of immune genes in  
tsetse flies (*Glossina*) and insights into  
tsetse-symbiont-trypanosome  
interactions

Robert Calam Bruce

A thesis submitted in partial fulfilment of the requirements  
of the University of Westminster for the degree of Doctor  
of Philosophy

31<sup>st</sup> May 2022

# Abstract

Tsetse flies (genera *Glossina*) are the sole biological vectors of African *Trypanosoma* species, the infectious agents of African Trypanosomiasis. Vector control is a key inhibitor of disease transmission; however, long-term control measures are economically and ecologically unsustainable and therefore, alternatives must be explored. In this thesis we aim to explore the evolution of three important immune genes: attacin-A (*AttA*), Defensin (*Def*) and Toll-like receptor 2 (*TLR2*), in relation to symbionts and parasitic interactions. This could in turn lay the foundations for genetic control methods

The successful identification of novel attacin orthologues confirmed the previous descriptions of attacin clusters within the *Glossina* genome, while a single novel defensin orthologue was identified in each of the six *Glossina* genomes. A total of six TLRs were confirmed within the *Glossina* genome, and three additional TLRs were potentially identified, though these are unconfirmed. The evolutionary history of the attacin cluster remains undetermined, however concerted evolution likely impacts the evolution of *AttA*, while *Def* and *TLRs* are governed by strict Darwinian selection.

A wild population sample of *Glossina morsitans morsitans* illustrated differing levels of nucleotide variation in each gene, *Def* being the least polymorphic (n = 8) and *TLR2* being the most (n = 22). All genes indicated a recent population expansion event and deviations from neutrality, indicative of population expansion and balancing selection. Genetic variation in both *AttA* and *TLR2* was found to be maintained via purifying selection, while *Def* exhibited signs of the Red Queen arms race and balancing selection. Trypanosome infection rates were unexpectedly high (69.35%), consisting of mixed species infections. Advantageous *Def* variants were observed to reduce infection rates within samples, while an observable relationship between *TLR2* and symbiont variation, and infection rate requires further research.

The results within described the impacts of evolution and population change on immune genes and how the interactions with symbiont populations can influence trypanosome infection rates. This thesis indicates that an understanding of the evolution and interactions of the tsetse-symbiont-trypanosome triplet could be used to inform novel genetic control methods.

# Table of Contents

Abstract .....	i
Table of Contents .....	ii
List of Figures .....	viii
List of Tables .....	xv
List of Appendices .....	xvii
Acknowledgements .....	xviii
Author's declaration.....	xix
List of Abbreviations.....	xx
<b>1: Introduction .....</b>	<b>1</b>
1.1: African Trypanosomiasis and Tsetse flies.....	1
1.2: The tsetse fly: Hippoboscoidea superfamily .....	4
1.2.1: Glossinidae.....	5
1.3: Parasite - Vector Interactions.....	10
1.3.1: Dipteran immunity.....	11
1.3.2: <i>Trypanosoma brucei</i> spp. and tsetse interactions.....	16
1.3.3: The role of endosymbionts in trypanosome transmission.....	19
1.4: Gaps in knowledge .....	20
1.5: Aims and objectives.....	21
<b>2: The identification and characterisation of the attacin clusters and defensin genes within <i>Glossina</i> species genome assemblies.....</b>	<b>24</b>
2.1: Introduction.....	24
2.1.1: Aims and Objectives .....	29
2.2: Methodology .....	31
2.2.1: Identification of <i>G. morsitans</i> attacin and defensin genes.....	31
2.2.2: Identification of transcripts within <i>Glossina</i> genomes.....	31

2.2.3: Mapping of predicted genes and the identification of missing attacin genes ...	32
2.2.4: CLUSTALW gene alignments .....	32
2.2.5: Phylogenetic analysis .....	33
2.2.6: Inter-species nucleotide variation analysis .....	33
2.2.7: Pairwise distance Principle Component Analysis .....	34
2.2.8: Three-dimensional protein modelling .....	34
2.3: Results .....	35
2.4: Attacins .....	35
2.4.1: Attacin gene cluster identification and variation .....	35
2.4.2: Interspecies analysis .....	63
2.5: Defensin .....	74
2.5.1: <i>Glossina</i> defensin identification .....	74
2.5.2: Interspecies variation .....	83
2.6: Discussion .....	90
2.7: Conclusion .....	94
<b>3: Molecular variation of Attacin-A and Defensin, in relation to trypanosome infection and symbionts within a wild <i>Glossina morsitans morsitans</i> population. ....</b>	<b>95</b>
3.1: Introduction .....	95
3.1.1: Aims and Objectives .....	99
3.2: Materials and Methods .....	101
3.2.1: Tsetse samples collection .....	101
3.2.2: gDNA extraction of <i>Glossina</i> specimens .....	102
3.2.3: Primer design and Polymerase Chain Reaction .....	102
3.2.4: Gel extraction of trypanosome ITS bands .....	106
3.2.5: Sanger sequencing and sequence analysis .....	106
3.2.6: Intra-species phylogenetic analysis .....	107
3.2.7: Haplotype analysis .....	107

3.2.8: Intra-species nucleotide variation analysis .....	107
3.2.9: dN/dS: Synonymous vs non-synonymous variation .....	108
3.2.10: Gene flow analysis .....	108
3.2.11: Demographic change and test for neutrality: Pairwise mismatch, Tajima's D, Fu's Fs and Coalescent Simulation.....	109
3.2.12: Recombination analysis .....	109
3.2.13: <i>Wigglesworthia</i> haplotype variation and population genetics .....	110
3.2.14: Association of AMP and symbiont nucleotide variation .....	110
3.3: Results .....	111
3.3.1: Intra-species genetic variation and population genetics of wild <i>Glossina morsitans morsitans</i> .....	111
3.3.2: Tests for neutrality and demographic change.....	125
3.3.3: <i>Wigglesworthia</i> endosymbiosis and genetic variation.....	131
3.3.4: <i>Trypanosoma</i> infection and genetic variation.....	138
3.3.5: Comparison of the genetic variation .....	142
3.4: Discussion .....	144
3.5: Conclusion .....	148
<b>4: Structural and functional analysis of attacin-A and defensin as a result of single nucleotide polymorphism: Impacts on infection and evolution .....</b>	<b>149</b>
4.1: Introduction.....	149
4.1.1: Aims and objectives .....	151
4.2: Methodology .....	153
4.2.1: gDNA extraction, Polymerase Chain Reaction, and bioinformatics analysis:...	153
4.2.2: Comparison of protein variation and infection .....	153
4.2.3: Indication of Selection: Z-tests .....	153
4.2.4: Indication of Selection: HyPHY based analysis .....	154
4.2.5: TreeSAAP.....	155

4.2.6: Three-dimensional protein modelling .....	156
4.2.7: Active site prediction .....	156
4.3: Results .....	158
4.3.1: Protein variation and <i>Trypanosoma</i> infection.....	160
4.3.2: Indicators of natural selection .....	163
4.3.3: Characterisation of significant codons. ....	168
4.3.4: Structural and functional analysis .....	171
4.3.5: Attacin-A .....	171
4.3.6: Defensin .....	185
4.4: Discussion .....	208
4.5: Conclusion .....	210
<b>5: The identification and characterisation of Toll-like receptor protein families within the <i>Glossina</i> genome assemblies .....</b>	<b>212</b>
5.1: Introduction.....	212
5.1.1: Aims and Objectives .....	216
5.2: Methods .....	218
5.2.1: Identification of TLR genes within <i>Glossina</i> and other dipteran genera.....	218
5.2.2: Transcript and domain structure.....	219
5.2.3: Phylogenetic analysis.....	219
5.2.4: Interspecies variation .....	219
5.2.5: Three-dimensional protein prediction .....	220
5.3: Results .....	221
5.3.1: Gene structure and characterisation.....	221
5.3.2: Phylogenetic analysis.....	240
5.3.3: Interspecies variation .....	244
5.3.4: 3-dimensional protein structure analysis.....	250
5.4: Discussion .....	257

5.5: Conclusion .....	261
<b>6: Intraspecies variation of the Toll-like Receptor 2 gene within wild <i>Glossina morsitans morsitans</i>, and the association with endosymbiont and trypanosome infection.....</b>	<b>262</b>
6.1: Introduction.....	262
6.1.1: Aims and Objectives .....	265
6.2: Materials and Methods .....	267
6.2.1: Tsetse samples collection and gDNA extraction.....	267
6.2.2: Primer design and Polymerase Chain Reaction .....	267
6.2.3: Sanger sequencing and sequence analysis.....	269
6.2.4: Intra-species phylogenetic analysis .....	269
6.2.5: Haplotype analysis .....	269
6.2.6: Intra-species nucleotide variation analysis .....	269
6.2.7: dN/dS: Synonymous vs non-synonymous variation .....	270
6.2.8: Gene flow analysis .....	270
6.2.9: Demographic change and test for neutrality: Pairwise mismatch, Tajima's D, Fu's Fs and Coalescent Simulation .....	270
6.2.10: Recombination analysis .....	271
6.2.11: Indication of Selection: Z-tests and HyPhy based analysis .....	271
6.2.12: Three-dimensional protein modelling and function impact .....	271
6.2.13: <i>Wigglesworthia</i> haplotype variation and population genetics .....	271
6.2.14: Association of AMP and symbiont nucleotide variation .....	272
6.3: Results .....	273
6.3.1: Intra-species genetic variation and population genetics of wild <i>Glossina morsitans morsitans TLR2 gene</i> .....	273
6.3.2: The impacts of genetic variation on structure, functionality, and selection ...	280
6.3.3: <i>Wigglesworthia</i> endosymbiosis and <i>Trypanosoma</i> infection .....	284
6.4: Discussion .....	288



6.5: Conclusion .....	290
<b>7: General discussion and conclusions .....</b>	<b>292</b>
7.1: Limitations of this study .....	295
7.2: Impacts, implications, and future research.....	296
7.3: Conclusion .....	301
<b>Appendix.....</b>	<b>303</b>
<b>List of references.....</b>	<b>320</b>

## List of Figures

Figure 1.1: The life cycle of <i>T. brucei</i> within both the mammalian and tsetse hosts. ....	3
Figure 1.2: The evolutionary history of the Hippoboscoidea superfamily. ....	5
Figure 1.3: Illustrations of distinguishing features of <i>Glossina</i> spp. ....	6
Figure 1.4: The life cycle of <i>Glossina</i> spp. (Leak, 1999). ....	8
Figure 1.5: A simplified model of the Toll/Dif/Dorsal and IMD-Rel signaling cascades in <i>Drosophila melanogaster</i> . ....	12
Figure 1.6: The structure of TLR proteins illustrating all three subdomains. ....	13
Figure 2.1: Predicted structures of <i>D. melanogaster</i> AttA and AttB produced using AlphaFold V2 (Jumper <i>et al.</i> , 2021). ....	26
Figure 2.2: A linear representation of the <i>Glossina</i> attacin clusters. ....	27
Figure 2.3: The structure of the C-terminal mature <i>Def</i> region of insect defensin-A. ....	28
Figure 2.4: A linear representation of the attacin cluster in <i>Glossina morsitans morsitans</i> . ....	35
Figure 2.5: The comparison of the CDS sequences of two <i>G. m. morsitans</i> AttA genes, GMOY010521 and GMOY013348. ....	36
Figure 2.6: A) The comparison of two <i>G. m. morsitans</i> AttA sequences, GMOY01021 and GMOY010522. B) The comparison of <i>G. m. morsitans</i> AttA (GMOY01021) and AttB (GMOY010523). ....	38
Figure 2.7: The comparison of CDS sequences <i>G. m. morsitans</i> AttA (GMOY01021) and AttD (GMOY010524). ....	39
Figure 2.8: A linear representation of the predicted Attacin cluster within <i>Glossina austeni</i> . ....	40
Figure 2.9: The comparison of the CDS sequences of two <i>G. austeni</i> AttA genes GAUT047992 and GAUT047990. ....	41
Figure 2.10: A) The comparison of two <i>G. austeni</i> AttA sequences, GAUT047992 and GAUT048001. B) The comparison of <i>G. austeni</i> AttA (GAUT047992) and AttB (GAUT048006). ....	42
Figure 2.11: The comparison of the CDS sequence of <i>G. austeni</i> AttA (GAUT047992) and AttD (GAUT047991). ....	43

Figure 2.12: A linear representation of the predicted Attacin cluster within <i>Glossina pallidipes</i> . .....	44
Figure 2.13: The comparison of the CDS sequences of <i>G. pallidipes AttA</i> (GPAI040759) and <i>AttB</i> (GPAI040754) genes.....	45
Figure 2.14: A) The comparison of two <i>G. pallidipes AttA</i> sequenes, GPAI040759 and GPAI040769. B) The identifeciation of a novel <i>AttA</i> C-terminal sequences on <i>G. pallidipes</i> contig. JMRR01001001. ....	46
Figure 2.15: The comparison of CDS sequences <i>G. pallidipes AttA</i> (GPAI040759) and <i>AttD</i> (GPAI040752) genes.....	47
Figure 2.16: A linear representation of the predicted Attacin cluster within <i>Glossina fuscipes fuscipes</i> . .....	48
Figure 2.17: A) The comparsion of the CDS sequences of <i>G. f. fuscipes AttB</i> (GFUI014658) and two <i>AttA</i> genes: A) GFUI014668, B) GFUI014661. ....	49
Figure 2.18: The alignment of <i>G. f. fuscipes AttB</i> (GFUI0104658) and contigs. JFJR01000138 and JFJR01000139. ....	50
Figure 2.19: The alignment of <i>G. f. fuscipes AttB</i> (GFUI0104658) and the reverse strand of contig. JFJR01000137. ....	51
Figure 2.20: The comparison of CDS sequences from <i>G. m. morsitans AttD</i> (GMOY010524) and <i>G. f. fuscipes AttD</i> (GFUI014660). ....	52
Figure 2.21: A linear representation of the predicted Attacin cluster within <i>Glossina palpalis gambiensis</i> . .....	53
Figure 2.22: A) The comparison of <i>G. m. morsitans AttA</i> sequenes (GMOY010521) and <i>G. p. gambiensis AttA</i> (GPPI020332). B) The alignment of <i>G. m. morsitans AttA</i> sequenes (GMOY010521) and the untranslated region of <i>G. p. gambiensis gene</i> GPPI020339.....	54
Figure 2.23: The comparison of CDS sequences from <i>G. m. morsitans AttD</i> (GMOY010524) and <i>G. p. gambiensis AttD</i> (GPPI020339). ....	55
Figure 2.24: The alignment of <i>G. m. morsitans AttA</i> (GMOY010521) and <i>G. p. gambiensis</i> contigs. JXJN01009050 and JXJN01009051. ....	56
Figure 2.25: A linear representation of the predicted Attacin cluster within <i>G. brevipalpis</i> . .....	57

Figure 2.26: A) The comparison of <i>G. m. morsitans AttA</i> sequenes (GMOY010521) and <i>G. brevipalpis AttA</i> (GBRI004567). B) The comparison of <i>G. m. morsitans AttA</i> sequenes (GMOY010521) and <i>G. brevipalpis</i> contig. JFJS01007046. ....	59
Figure 2.27: A) The comparison of <i>G. m. morsitans AttA</i> sequenes (GMOY010521) and <i>G. brevipalpis AttB</i> (GBRI004559). B) The comparison of <i>G. m. morsitans AttA</i> sequenes (GMOY010521) and <i>G. brevipalpis</i> contig. JFJS01007041. ....	60
Figure 2.28: The comparison of CDS sequences from <i>G. m. morsitans AttD</i> (GMOY010524) and <i>G. brevipalpis AttD</i> (GBRI004558). ....	62
Figure 2.29: Phylogenetic analyses of all predicted attacin genes within the <i>Glossina</i> genus conducgted using both the Maximum Likelihood method (A) and the Neighbour-Joining method (B). ....	65
Figure 2.30: Sliding window analysis of the predicted Attacin genes CDS within the <i>Glossina</i> spp. A) <i>AttA</i> , B) <i>AttB</i> and C) <i>AttD</i> .. ....	67
Figure 2.31: Principle component analysis (PCA) plot of all predicted attacin genes within the <i>Glossina</i> genus. ....	69
Figure 2.32: Structural alignment of complete predicted <i>Glossina</i> Attacin protein families A) <i>AttA</i> ; B) <i>AttB</i> and C) <i>AttD</i> . ....	72
Figure 2.33: Comparison of predicted attacin structures between the <i>Glossina</i> spp. ....	73
Figure 2.34: The identification of <i>Def</i> within the <i>Glossina morsitans morsitans</i> genome. ....	75
Figure 2.35: The identification of <i>Def</i> within the <i>Glossina austeni</i> genome. ....	77
Figure 2.36: The identification of <i>Def</i> within the <i>Glossina pallidipes</i> genome ....	79
Figure 2.37: The identification of <i>Def</i> within the <i>Glossina fuscipes fuscipes</i> genome. ....	80
Figure 2.38: The identification of <i>Def</i> within the <i>Glossina palpalis gambiensis</i> genome....	81
Figure 2.39: The identification of <i>Def</i> within the <i>Glossina brevipalpis</i> genome.....	82
Figure 2.40: Phylogenetic analyses of <i>Glossina Def</i> using both the Maximum Likelihood method (A) the Neighbour-Joining method (B). ....	84
Figure 2.41: Interspecies variation of <i>Glossina Def</i> genes. ....	85
Figure 2.42: Principle component analysis (PCA) plot of all predicted defensin genes within the <i>Glossina</i> genus. ....	86
Figure 2.43: Predictions of <i>Glossina Def</i> proteins structures. ....	88
Figure 2.44: Comparision of predicted Def structures between the <i>Glossina</i> spp.....	89

Figure 3.1: A map of Zimbabwe detailing the collection sites for wild <i>Glossina morsitans morsitans</i> samples used in this study. ....	101
Figure 3.2: Phylogenetic analyses of the <i>COI</i> sequence fragments. ....	113
Figure 3.3: Phylogenetic analyses of the <i>AttA</i> sequence fragments. ....	115
Figure 3.4: Phylogenetic analyses of the <i>Def</i> sequence fragments. ....	117
Figure 3.5: The frequency and distribution of genetic haplotypes within each geographical location.....	119
Figure 3.6: TCS haplotype networks of each target gene, A) <i>COI</i> ; B) <i>AttA</i> and C) <i>Def</i> , within the sample population. ....	121
Figure 3.7: Sliding window graphs of the PCR amplification of <i>G. morsitans morsitans</i> immune genes; A) <i>AttA</i> and B) <i>Def</i> . ....	124
Figure 3.8: Pairwise mismatch analysis of the <i>COI</i> , <i>AttA</i> , and <i>Def</i> genes showing the observed and expected frequencies of nucleotide variation. ....	129
Figure 3.9: TCS haplotype network of the <i>Wigglesworthia 16S</i> gene within the <i>G. m. morsitans</i> sample populations. ....	132
Figure 3.10: Pairwise mismatch analysis of the <i>W. glossinidia 16S</i> gene showing the observed and expected frequencies of nucleotide variation. ....	134
Figure 3.11: TCS haplotype networks for the <i>G. m. morsitans</i> immune genes <i>COI</i> (A), <i>AttA</i> (B) and <i>Def</i> (C) showing the frequency of <i>Wigglesworthia 16S</i> haplotypes within the exhibited haplotypes.....	137
Figure 3.12: TCS haplotype network of both <i>G. m. morsitans AttA</i> (A) and <i>Def</i> (B) genes showing the frequency of infected and uninfected samples within each haplotype. ....	140
Figure 3.13: TCS haplotype network of the <i>W. glossinidia 16S</i> gene showing the frequency of infected and uninfected samples within each haplotype. ....	141
Figure 3.14: The association of <i>P</i> -distance between <i>AttA</i> (A) and <i>Def</i> (B) and successful <i>W. g. morsitans</i> amplification. ....	143
Figure 4.1: Non-synonymous mutations within the <i>AttA</i> (A) and <i>Def</i> (B) genes.....	159
Figure 4.2: A comparison of the number of samples exhibiting each protein variant and the number of infected samples. A) <i>AttA</i> variants, B) <i>Def</i> variants. ....	160
Figure 4.3: A comparison of the number of samples exhibiting each protein variant, and the frequency of infection within each variant. A) Illustrates the relationship between the	

number of samples expressing each AttA variant and infection rates. B) Shows the relationship between the number of samples expressing each Def variant and infection frequency. ....	162
Figure 4.4: dN – dS values at each codon of a protein sequence fragment: A) AttA; B) Def. ....	167
Figure 4.5: PROVEAN scores of each codon along the AttA (A) and Def (B) protein fragments. ....	170
Figure 4.6: Principle component analysis (PCA) plot of wild AttA variant.....	175
Figure 4.8: The alignment of all wild <i>G. m. morsitans</i> AttA protein variants showing the predicted active site.....	177
Figure 4.9: Prediction of the active site within the wild <i>G. m. morsitans</i> AttA variant.....	180
Figure 4.10: Prediction of the active site within the AttA-N35 variant. ....	181
Figure 4.11: Prediction of the active site within the AttA-T43 variant.....	182
Figure 4.12: Prediction of the active site within the AttA-D49 variant. ....	183
Figure 4.13: Prediction of the active site within the AttA-A86 variant.....	184
Figure 4.14: Principle component analysis (PCA) plot of wild defensin protein variants. ....	189
Figure 4.15: Comparison of the Def active site within <i>Allomyrina dichotoma</i> , <i>Oryctes rhinoceros</i> and <i>Glossina morsitans morsitans</i> . ....	190
Figure 4.16: Prediction of the active site within the <i>G. m. morsitans</i> Def using the amino acid sequence identified in Chapter 3. ....	192
Figure 4.17: Prediction of the active site within the natural Def variant. ....	197
Figure 4.18: Prediction of the active site within the Def-F45 variant.....	198
Figure 4.19: Prediction of the active site within the Def-S18 variant.....	199
Figure 4.20: Prediction of the active site within the Def-I18 variant. ....	200
Figure 4.21: Prediction of the active site within the Def-I82 variant.....	201
Figure 4.22: Prediction of the active site within the Def-S18/E42 variant.....	202
Figure 4.23: Prediction of active the site within the Def-S18/F45/I82 variant.....	203
Figure 4.24: Prediction of active the site within the Def-E42 variant.....	204
Figure 4.25: Prediction of active the site within the Def-I18/F45 variant. ....	205
Figure 4.26: Prediction of active the site within the Def-S18/F45 variant. ....	206
Figure 4.27: Prediction of active the site within the Def-S18/I82 variant. ....	207

Figure 5.1: The phylogeny of TLR gene within arthropod taxa.....	213
Figure 5.2: The overall structure of TLR proteins illustrating all three subdomains. ....	214
Figure 5.3: A model of the Spz-TLR signaling cascade in <i>D. melanogaster</i> .....	216
Figure 5.4: A full gene alignment of predicted TLR1 genes within the <i>Glossina</i> genome assemblies with reference genes of <i>D. melanogaster</i> , <i>S. calcitrans</i> and <i>M. domestica</i> ....	223
Figure 5.5: A full gene alignment of predicted TLR2 genes within the <i>Glossina</i> genome assemblies with reference genes of <i>D. melanogaster</i> , <i>S. calcitrans</i> and <i>M. domestica</i> ....	225
Figure 5.6: The gene alignment of predicted TLR3 genes within the <i>Glossina</i> genome assemblies with reference genes of <i>D. melanogaster</i> , <i>S. calcitrans</i> and <i>M. domestica</i> ....	227
Figure 5.7: The gene alignment of predicted TLR5 genes within the <i>Glossina</i> genome assemblies with reference genes of <i>D. melanogaster</i> , <i>S. calcitrans</i> and <i>M. domestica</i> ....	229
Figure 5.8: The gene alignment of predicted TLR6 genes within the <i>Glossina</i> genome assemblies with reference genes of <i>D. melanogaster</i> , <i>S. calcitrans</i> and <i>M. domestica</i> ....	231
Figure 5.9: The gene alignment of predicted TLR7 genes within the <i>Glossina</i> genome assemblies with reference genes of <i>D. melanogaster</i> , <i>S. calcitrans</i> and <i>M. domestica</i> ....	233
Figure 5.10: The gene alignment of predicted TLR8 genes within the <i>Glossina</i> genome assemblies with reference genes of <i>D. melanogaster</i> , <i>S. calcitrans</i> and <i>M. domestica</i> ....	235
Figure 5.11: The gene alignment of predicted TLR9 genes within the <i>Glossina</i> genome assemblies with reference genes of <i>D. melanogaster</i> , <i>S. calcitrans</i> and <i>M. domestica</i> ....	237
Figure 5.12: The gene alignment of predicted TLR13 genes within the <i>Glossina</i> genome assemblies with reference genes of <i>D. melanogaster</i> , <i>S. calcitrans</i> and <i>M. domestica</i> ....	239
Figure 5.13: Phylogenetic analyses of all predicted <i>Glossina</i> TLR proteins. A) Maximum Likelihood method, B) Neighbour-Joining method.....	243
Figure 5.14: Sliding window analysis showing dN/dS variation across the predicted TLR genes within the <i>Glossina</i> spp.. ....	247
Figure 5.15: Principle component analysis (PCA) plot of all predicted TLR gene within the <i>Glossina</i> genus.....	249
Figure 5.16: Structural alignment of complete predicted <i>Glossina</i> TLR proteins.....	252
Figure 5.17: Structural alignment of partial predicted <i>Glossina</i> TLR proteins.. ....	253
Figure 5.18: A heatmap comparing the conservation of TLR protein structures across all identified <i>Glossina</i> TLR genes. ....	255

Figure 5.19: Principle component analysis (PCA) plot of all TLR protein structures within the <i>Glossina</i> genus.....	256
Figure 6.1: Phylogenetic analysis of the <i>TLR2</i> gene fragments. A) A Maximum Likelihood tree, B) Neighbour-Joining method. ....	274
Figure 6.2: TCS haplotype networks of the <i>TLR2</i> gene fragment within the sample population. ....	276
Figure 6.3: Sliding window graphs of the <i>G. moristans morsitans TLR2</i> gene fragment; A) illustrates the distribution of polymorphic sites throughout the gene fragment. B) illustrates the characteristic (Synonymous and Non-synonymous) of each polymorphic site.. ....	277
Figure 6.4: Pairwise mismatch analysis of the <i>TLR2</i> gene fragment showing the observed and expected frequencies of nucleotide variation. ....	279
Figure 6.5: dN – dS values at each codon of the <i>TLR2</i> sequence fragment.....	283
Figure 6.6: TCS haplotype networks for the <i>G. m. morsitans TLR2</i> gene, showing the frequency of <i>Wigglesworthia 16S</i> haplotypes within the exhibited haplotypes.....	285
Figure 6.7: TCS haplotype network of the <i>TLR2</i> gene fragment showing the frequency of infected and uninfected samples within each haplotype.....	286
Figure 6.8: The association of genetic <i>P</i> -distance between <i>TLR2</i> genes and successful <i>W. g. morsitans</i> amplification. ....	287



## List of Tables

Table 1.1: Ligands/PAMPs responsible for trypanosome identification in triatomine spp. (Uematsu and Akira, 2008; Kumar <i>et al.</i> , 2009).....	18
Table 2.1: Attacin genes identified by Trappeniers <i>et al.</i> (2019).....	31
Table 3.1: Primers used in the amplification of <i>AttA</i> , <i>Def</i> , <i>Wigglesworthia glossinidia 16S</i> and Trypanosome <i>ITS</i> genes. ....	104
Table 3.2: The PCR cycling conditions used to amplify gene fragments of <i>AttA</i> , <i>Def</i> , <i>Wigglesworthia glossinidia 16S</i> and Trypanosome <i>ITS</i> genes. ....	105
Table 3.3: The gene flow results for the <i>AttA</i> , <i>Def</i> and <i>COI</i> genes across all collections localities. ....	125
Table 3.4: Mantel test results showing the correlation between geographical distance and Fst values of <i>AttA</i> , <i>Def</i> and <i>COI</i> . ....	126
Table 3.5: Tajima's D and Fu's Fs statistics of <i>COI</i> , <i>AttA</i> and <i>Def</i> .....	127
Table 3.6: The sites of recombination within the <i>AttA</i> and <i>Def</i> fragments.....	130
Table 3.7: The gene flow results for the <i>W. g. morsitans 16S</i> and <i>G. m. morsitans</i> between all collections localities.....	133
Table 3.8: Mantel test results showing the correlation between geographical distance and Fst values of <i>Wigglesworthia glossinidia morsitans 16S</i> and <i>G. m. morsitans COI</i> . ....	134
Table 3.9: Mantel test results showing the correlation between <i>G. m. morsitans</i> AMPs and <i>W. glossinidia 16S</i> . ....	142
Table 4.1: A matrix of Z-scores showing indicated selection between <i>Def</i> haplotypes. ....	164
Table 4.2: All radical amino acid property changes in <i>AttA</i> and <i>Def</i> identified by TreeSAAP. . .....	169
Table 4.3: Predicted 3-Dimensional structures of each of the <i>G. m. morsitans AttA</i> variants. .....	172
Table 4.4: Predicted 3-Dimensional structures of each of the <i>G. m. morsitans Def</i> variants. .....	186

Table 5.1: Reference genes for each TLR identified within related dipteran species..	.....218
Table 5.2: Genetic variation across the <i>Glossina</i> within each TLR gene.....	245
Table 6.1: TLR2 prime information. ....	268
Table 6.2: The gene flow results for <i>TLR2</i> gene fragment across all collections localities. .....	278
Table 6.3: Indicated sites of recombination within the <i>TLR2</i> gene fragments.....	279
Table 6.4: Predicted 3-Dimensional structures of each of the <i>G. m. morsitans</i> TLR2 variants. .....	281

## List of Appendices

<b>Appendix 1: Tables of identified attacin and defensin genes.....</b>	<b>303</b>
<b>Appendix 2: Statistical equations .....</b>	<b>305</b>
<b>Appendix 3: <i>Glossina</i> genomes .....</b>	<b>309</b>
<b>Appendix 4: Gel images .....</b>	<b>310</b>
<b>Appendix 5: Protein surface structure.....</b>	<b>317</b>
<b>Appendix 6: Topology.....</b>	<b>319</b>

## Acknowledgements

I would like to extend my deepest thanks to my supervisory team: Firstly, to Dr Polly Hayes for her invaluable guidance and support throughout this PhD, and to Dr John Murphy and Prof. Wendy Gibson for their advice and assistance during my study. Secondly, I would like to thank the University of Westminster and the School of Liberal Arts and Sciences, for the essential use of their facilities and the Studentship, which made this PhD possible.

I would further like to extend my thanks to Prof. Stephen Torr of the Liverpool School of Tropical Medicine (United Kingdom), who provided the samples used in this study.

Finally, I would like to thank my family and friends for their continued support, motivation, and assistance. Specifically, I would like to thank my parents John and Jennifer Bruce for the emotional and financial support over the years. I would also like to thank my partner Sophie for keeping me sane and motivated throughout this journey. And lastly, I would like to thank Joe Avis, Daniel Brookes and Bruce Wight for their assistance and expertise in GIS and mathematics.

## Author's declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

## List of Abbreviations

<b>Abbreviation</b>	<b>Meaning</b>
+G	With Discrete Gamma distribution
+I	With Invariant sites
16S	16S Ribosomal RNA
3D	Three-dimensional
aa	Amino Acid
AAT	Animal African Trypanosomiasis
Ala/A	Alanine
AMP	Antimicrobial Peptides/Proteins
AP-1	Activator protein 1
Arg/R	Arginine
Asn/N	Asparagine
Asp/D	Aspartic acid
AT	African Trypanosomiasis
AttA	Attacin-A
AttB	Attacin-B
AttC	Attacin-C
AttD	Attacin-D
BLAST	Basic Local Alignment Search Tool
bp	Base Pair
CDS	Coding Sequence
COI	Cytochrome Oxygenase 1
cont.	Continued
CPP	Cell-penetrating peptides
Cys/C	Cysteine
Def	Defensin
DEM	Digital elevation model
DNA	Deoxyribonucleic Acid
DnaSP	DNA Sequence Polymorphism
E.g.	<i>exempli gratia</i> ("for example")
<i>et al.</i>	<i>et alia</i> ("and others")
FEL	Fixed Effects Likelihood
Fig.	Figure
Fst	Fixation index
FUBAR	Fast, Unconstrained Bayesian Approximation
GARD	Genetic Algorithm for Recombination Detection
gDNA	Genomic Deoxyribonucleic Acid
GIS	Geographic Information System
Gln/Q	Glutamine
Glu/E	Glutamic acid
Gly/G	Glycine
GNBPs	Gram-negative Binding Proteins
HAT	Human African Trypanosomiasis
Haps	Haplotypes

His/H	Histidine
Hs	Haplotype statistic
i.e.	<i>in est</i> ("that is")
Ile/I	Isoleucine
IMD	Immunodeficiency
I-TASSER	Iterative Threading Assembly Refinement
JC	Juke Cantor
kDa	Kilodalton
Km	Kilometres
Ks	Nucleotide statistic
Leu/L	Leucine
LRR	Leucine-rich Repeats
LRR-CT	Leucine-rich Repeats C-terminal flanking region
LRR-NT	Leucine-rich Repeats N-terminal flanking region
Lys/K	Lysine
MEGA	Molecular Evolutionary Genetics Analysis
MEME	Mixed Effects Model of Evolution
Met/M	Methionine
mm	Millimetres
mtDNA	Mitochondrial Deoxyribonucleic Acid
MUSCLE	Multiple Sequence Comparison by Log-Expectation
MyD88	Myeloid differentiation primary response 88
N/A	Not Applicable
NCBI	National Centre for Biotechnology Information
NF- $\kappa$ B	Nuclear factor kappa-light-chain-enhancer of activated B cells
NTD	Neglected Tropical Disease
PAMP	Pathogen-Associated Molecular Patterns
PCA	Principle Component Analysis
PCR	Polymerase Chain Reaction
PDB	Protein database
P-distance	Pairwise distance
PDP	Pathogen Detector Proteins
PGRPLB	Peptidoglycan Recognition Protein-LB
Phe/F	Phenylalanine
Pro/P	Proline
QGIS	Quantum Geographic Information System
Ser/S	Serine
SLAC	Single-Likelihood Ancestor Counting
spp	Species
Spz	Spätzle
SRTM	The Shuttle Radar Topography Mission
TCS	Templeton, Crandall, and Sing
TEP3	Thioester-containing protein 3
Thr/T	Threonine
TIR	Toll/Interleukin Receptor
TLR	Toll-like Receptor
TM-Score	Template Modelling Score

Tyr/Y	Tyrosine
US\$	United State Dollars
USGS	United States Geological Survey
UT	Untranslated region
Val/V	Valine
WAG	Whelan And Goldman Model
WHO	World Health Organisation



# 1: Introduction

## 1.1: African Trypanosomiasis and Tsetse flies

Parasitic infections are widely recognised as one of the greatest inhibitors of human economic and social development worldwide (Sachs and Malaney, 2002; Hotez and Kamath, 2009; Brooker, 2010). Yet, despite the threat that these diseases continue to present, many remain both under-funded and under-researched. Such infections are known as Neglected Tropical Diseases (NTDs) (Feasey *et al.*, 2010). Hotez and Kamath (2009) observed that NTDs affect an estimated 500 million people in Sub-Saharan Africa alone, which equates to approximately one-half of the economic and social burden resultant from malarial infections in the same area.

Human African trypanosomiasis (HAT), commonly referred to as African sleeping sickness, is one such NTD (Hotez and Kamath, 2009; Brun *et al.*, 2010). Although recorded cases of HAT have decreased drastically over the last two decades: from nearly 50,000 annual cases in the 1990's (Hide, 1999), to fewer than 977 in 2019 (Gao *et al.*, 2020); it is thought that many asymptomatic cases remain unrecognised (Capewell *et al.*, 2016). Significantly, Capewell *et al.* (2016) highlighted the importance of mammalian extracellular infection as potential reservoirs for HAT, with asymptomatic infections facilitate future outbreaks. Therefore, in addition to documented cases it is estimated that at least 70 million further individuals are at risk of HAT (Simarro *et al.*, 2012). Human African trypanosomiasis is caused by two sub-species of the protozoan parasite *Trypanosoma brucei*, namely *T. b. gambiense* and *T. b. rhodesiense* (Brun *et al.*, 2010; Stephens *et al.*, 2012).

The continued fall of recorded HAT cases in recent years, has resulted in calls to shift focus to Animal African Trypanosomiasis (AAT) (Morrison *et al.*, 2016). Animal African Trypanosomiasis, also known as nagana, is primarily associated with *Trypanosoma vivax* and *Trypanosoma congolense* infections, while *T. brucei* is considered to be a secondary pathogen (Losos and Ikede, 1972; Courtin *et al.*, 2008; Kasozi *et al.*, 2021). This change of focus aims to reduce the mounting economic impact of AAT in the central African countries, with estimates in 2013 of an annual economic loss of \$1-4 billion (USD) across the continent (Chanie *et al.*, 2013). This loss is generally attributed to the reduced production of milk and

meat, the mortality of working livestock and efforts to treat and prevent the disease (Chanie *et al.*, 2013). Animal African Trypanosomiasis targets a large number of both domestic and wild animals, with ruminants (including bovines, sheep and goats), horses, donkeys, cats, dogs and monkeys all at risk of infection (Losos and Ikede, 1972; Kasozi *et al.*, 2021)

Trypanosomiasis is a vector borne disease and requires a vector/intermediate host to facilitate biological development of trypanosome parasites and the infection of new mammalian hosts (Wamwiri and Changasi, 2016; Smith *et al.*, 2017). Tsetse flies (genus *Glossina*) are the sole biological vector of African trypanosome species, while all *Glossina* spp. are capable of transmitting the human infectious agent, the vectoral capacity of *T. b. gambiense* and *T. b. rhodesiense* varies within the *Glossina* genus (Wamwiri and Changasi, 2016).

The life cycle of African trypanosomes is characterised by two distinct stages: mammalian and tsetse. The mammalian stage of the life cycle (Stages 1-2 in Fig. 1.1) starts with the injection of infective metacyclic trypomastigotes from the salivary glands of the tsetse fly into the dermis of the host. The initial metacyclic trypomastigotes are short and stumpy and are preadapted for existence within the vector. Upon migration to the cardiovascular system, the metacyclic trypomastigotes undergo morphological transformation, stimulated by enzymatic and climatic triggers, adopting a long slender form for continued existence within the mammalian host. These slender trypomastigotes undergo rapid proliferation by binary fission, the resulting daughter cells adopt the stumpy morphology for ingestion by a feeding tsetse fly where they are able transformation into procyclic trypomastigotes (El-Sayed *et al.*, 2000; Matthews *et al.*, 2004). As the slender trypomastigotes mature they are capable of penetrating the blood vessel endothelium, migrating into extracellular tissues of the lymph system, the central nervous system and across the blood-brain barrier (Masocha and Kristensson, 2012).

The tsetse aspect of African trypanosomes life cycle starts with the ingestion of metacyclic trypomastigotes during a blood meal (Fig. 1.1, stages 3-7). Stimulated by the presence of protease enzymes and a change in temperature, stumpy trypomastigotes undergo a morphological transformation into procyclic trypomastigotes that rapidly proliferate within the tsetse midgut. Maturation to proventricular trypomastigotes, enables migration from

the midgut to the ectoperitrophic space, where proventricular trypomastigotes transform to the epimastigote stage enabling a final migration to the salivary glands. Upon entering the salivary glands epimastigotes attachment to the epithelium and undergo further proliferation by binary fission, before detaching from the salivary gland epithelium and becoming pathogenic free swimming metacyclic trypomastigotes (Caljon *et al.*, 2014).

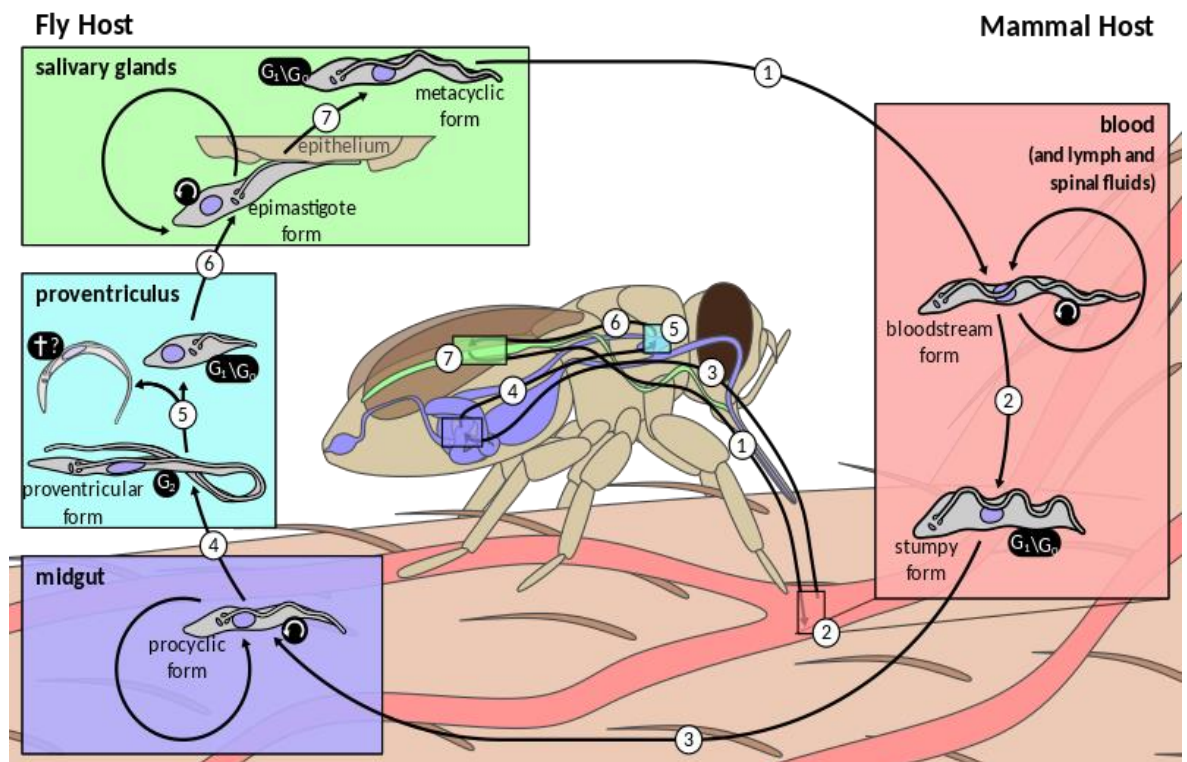


Figure 1.1: The life cycle of *T. brucei* within both the mammalian and dipteran hosts (with permission; Richard Wheeler, 2016). The focus of this review and all subsequent work lies between stages four and seven.

The mammalian adaptive immune system initiates a strong response to metacyclic trypomastigotes infection and eliminates the majority of trypanosomes throughout the course of infection. The successful infection of a mammalian host is achieved primarily through antigenic variation whereby trypanosomes alter their surface antigens, helping to mask their presence from mammalian antibodies (Borst and Cross, 1982; Horn, 2014).

The incubation period of *T. b. gambiense* and *T. b. rhodesiense* varies, acute sleeping sickness caused by *T. b. rhodesiense* often presents 1-3 weeks after infection, while chronic sleeping sickness, resulting from *T. b. gambiense* infection, has a longer but undefined incubation period. Progression of HAT can be characterised by two distinct phases: hemolympathic and neurological (Lundkvist *et al.*, 2004). During hemolympathic phase infected patients often present non-specific symptoms such as: severe headaches,

intermittent fever, joint pain and inflammation of the lymph nodes (Winterbottom's sign) (Lundkvist *et al.*, 2004; Brun *et al.*, 2010).

The latter neurological phase occurs following migration of trypanosomes to the central nervous system by crossing of the blood brain barrier. The neurological phase first presents signs 21-60 days after infection in acute HAT, and between 500-600 days in chronic infections. This stage of infection can be characterised by changes to a patient's personality and extreme lethargy resulting from the 24 hour cycle of interrupted sleep-wake patterns, ultimately resulting in falling into coma, organ failure and death (Lundkvist *et al.*, 2004).

Although this study focuses on *Glossina* spp. as the exclusive biological vector of AT, other species of haematophagic flies, specifically members of the *Tabanidae* and *Stomoxys* families, have been recorded as mechanical vectors (Desquesnes and Dia, 2003a, 2003b; Gao *et al.*, 2020). These mechanical vectors are neither required for nor capable of, facilitating parasite development and maturation. Mechanical transmission occurs following an interrupted meal, where it is common for the metacyclic trypomastigotes to remain in mouth parts of a vector without being ingested. Experimental data showed that *T. brucei* has the highest successful mechanical transmission rate (11.5%) of all *Trypanosoma* species (Mihok *et al.*, 1995).

While vectorial transmission is responsible for the vast majority of infections, there have been sporadic and occasional reports of both vertical and horizontal transmission of HAT (Lindner and Priotto, 2010; Biteau *et al.*, 2016; Gaillot *et al.*, 2017). In 2017, Gaillot *et al.* recorded a case of vertical transmission, from mother to daughter through the placenta, stating, "Whilst cases of vertical transmission are rare, they are also most likely highly underestimated". Additionally, rare, and isolated reports of horizontal transmission via either sexual intercourse or blood transfusion have also been recorded (Rocha *et al.*, 2004; Biteau *et al.*, 2016).

## 1.2: The tsetse fly: Hippoboscoidea superfamily

The superfamily Hippoboscoidea comprises of four families; *Glossinidae* (tsetse flies), *Hippoboscidae* (ked or louse flies), *Nycteribiidae* and *Streblidae* (both referenced as bat flies) (Fig. 1.2) (Petersen *et al.*, 2007). Characterised by their adenotrophic viviparous life

cycles, the members of this superfamily are all obligate haematophages and vectors of several notable pathogens.

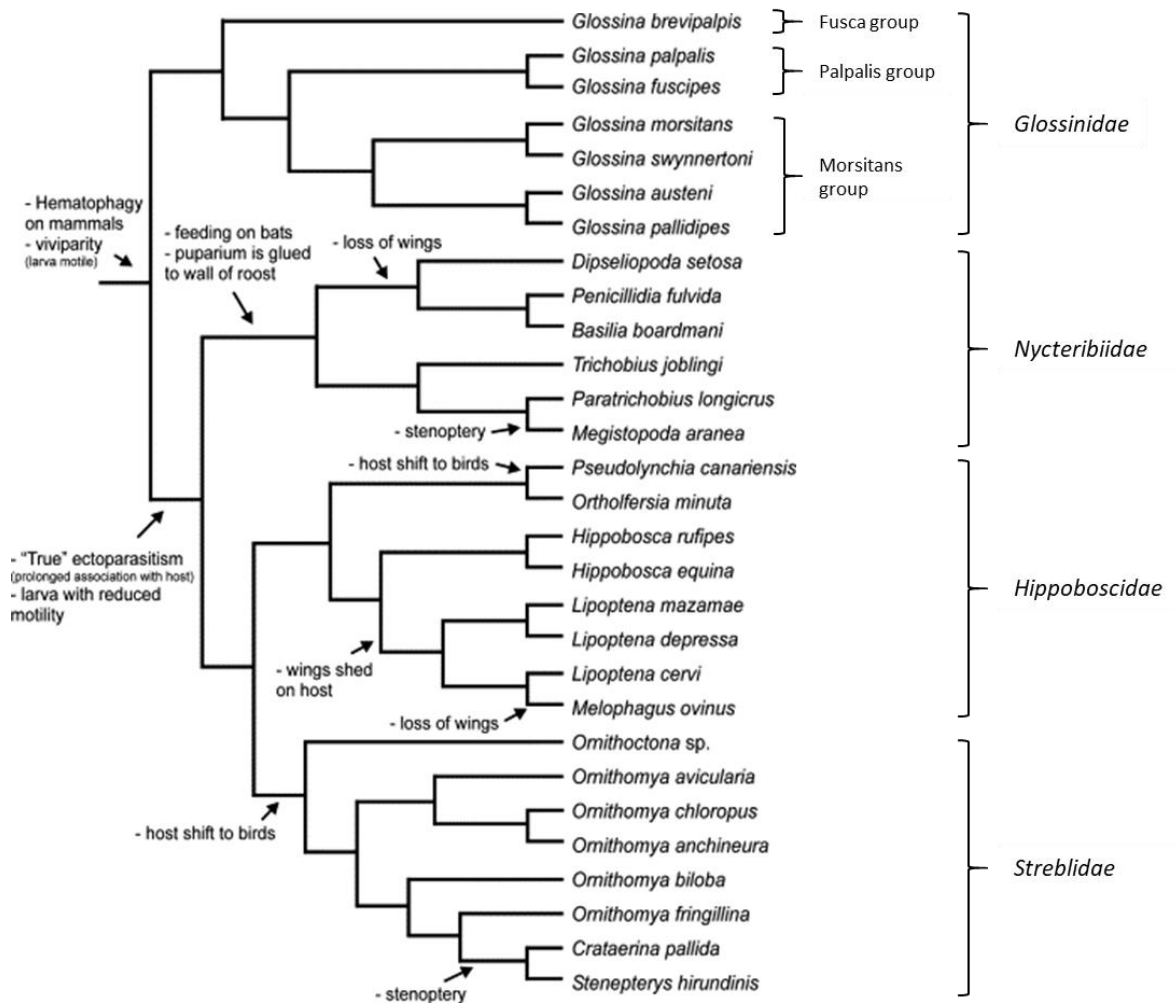


Figure 1.2: The evolutionary history of the Hippoboscoidea superfamily indicating the key evolutionary stages as published by Petersen *et al.* (2007).

### 1.2.1: Glossinidae

Tsetse flies are obligate parasites that feed on avian, reptilian and mammalian species including members of the Primates, Bovidae and Suidae families (Weitz, 1963). Tsetse are native to the African continent and are confined within definitive boundaries, between 10° North and 20° South of the equator. The *Glossinidae* family is separated into three subgenera characterised by habitat, behavioural and morphological traits (Fig. 1.2) (Cecchi *et al.*, 2008). The *Morsitans* (or, savannah group), including several prominent species like *G. morsitans* subsp, *G. pallidipes* and *G. swynnertoni*, favour open brush areas and are the primary vectors of *T. b. rhodesiense* in East Africa. The Palpalis (riverine) group species, including *G. palpalis* subsp., *G. fuscipes* subsp and *G. tachinoides*, prefer forested or

previously forested river banks. This group contains the primary vectors of *T. b. gambiense* in Western and central Africa. The Fusca (or, forest group) are primarily vectors of AAT, including important vectors such as *G. fusca* subsp., *G. tabaniformis* and *G. brevipalpis*.

The *Glossina* genus has four unique identifying morphological characteristics. Firstly, the shape and size of proboscis, which is long, thick, and attaching to the mouth beneath the head, quite distinctively from other dipteran species. Secondly, a tsetse will rest its wings completely, folding one over the other. Thirdly, the central wing cell has a unique and individually distinct “hatchet” shape. Finally, the antennae are covered in branching arista hairs. Differentiation of specific species can be undertaken using the colour of the tarsi, clarity of thorax markings and abdominal band size and colour. Additionally, species specific traits can be used for example, increased eye size and prominence in *G. pallidipes* (Fig. 1.4) (Austen, 1911; Newstead, 1924).

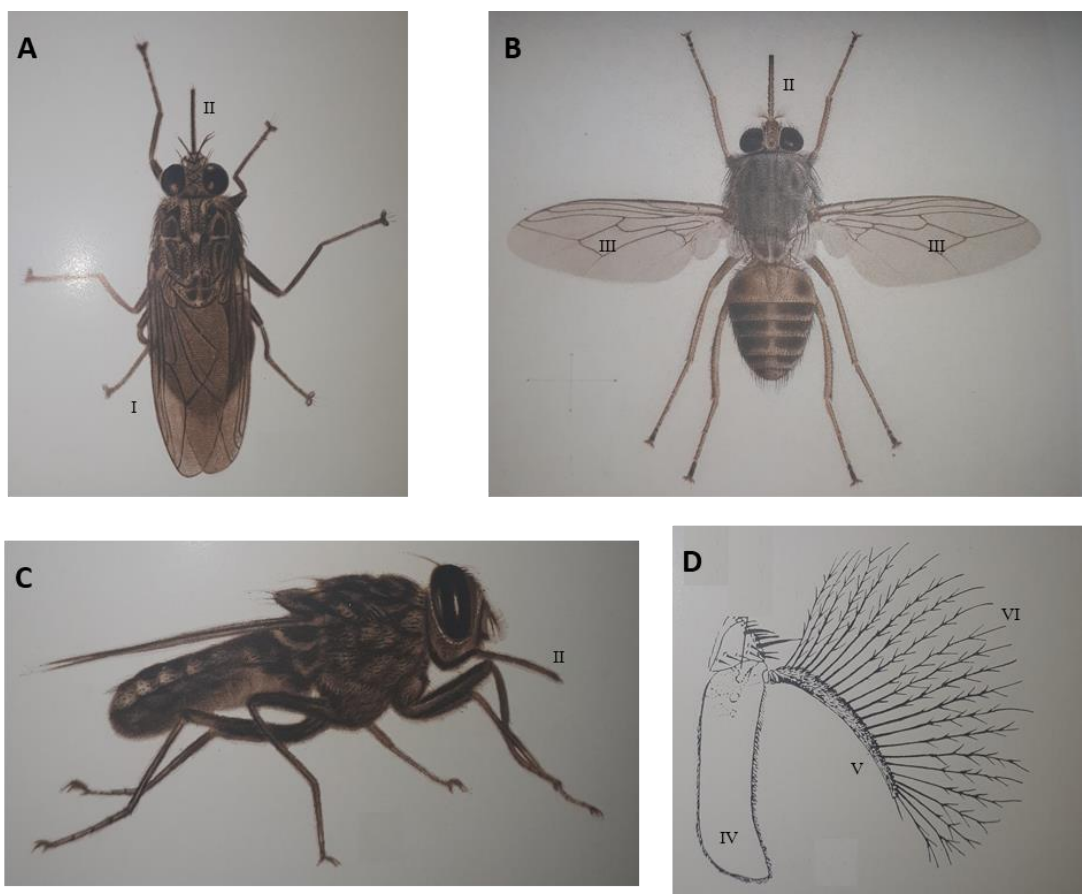


Figure 1.3: Illustrations of distinguishing features of *Glossina* spp. A) A dorsal view of a female *G. morsitans* at rest, the wings are folded along her back (I), also note the proboscis extending laterally from the head (II) (x4.5 magnification). B) A dorsal view of a female *G. morsitans* (x6 magnification), III marks the position of the hatchet cell. C) A profile view of a female *G. morsitans* at rest, showing the proboscis extending from the underside of the head (II) and the antennae (IV). D) A magnified image of the antennae (IV) of *G. morsitans* clearly showing the arista (V) and the branching hairs (VI) (Austen, 1911; Newstead 1924).

Determining the sex of tsetse flies can be done visually, with the average length of females being larger than that of males (female average: 10.0 mm; male average: 9.3 mm). Alternatively, dissection of the sexual organs can be undertaken.

Over the last decade climate change and human interactions appear to have influenced the nature of these species and their habitat preference. For example, Courtin *et al.*, (2010) observed that the “tsetse belt” in Burkina Faso had shifted significantly southwards whilst the central population diminished. The authors commented further that these changes were attributable to an increase in human population, causative of the destruction of breeding habitat, and climatic change, resulting in decreased rainfall and severe droughts. The shift in distribution is indicative of the migration recorded in several other vectors species, including members of the *Culicidae* (Mosquitoes) (Elbers *et al.*, 2015). Migration is a primary concern of health organisations, with vector species moving to previously uninfected areas and preventing the containment of not only trypanosomiasis but also other vector borne diseases such as, schistosomiasis and malaria (Tabachnick, 2010).

Although climate change is likely to have an impact on tsetse populations, Thornton *et al.*, (2006) observed that the specific habitat preferences and behaviour of tsetse mean that climate change will be less damaging to tsetse populations than other vector species. However, the authors continued to state that the destruction of habitats during human expansion is likely to have a far greater impact than the effects of natural climate change alone.

#### 1.2.1i: Glossinidae life cycle

Female tsetse mate only once, fertilise one egg at a time and, during the first three larval stages, juveniles remain within the uterus (Krafsur, 2009). While in the uterus, females feed the juvenile on a milky substance secreted from their milk glands. Approximately 7-9 days after fertilization, upon development to the third larval stage, the independent larva leave the female and burrow into the ground developing a hard puparial case for protection during maturation into an adult fly. This developmental stage can take up to 30 days to complete. As the juvenile does not feed in this time all nutrients must be obtained from the mother during the initial nine day development. In that time the blood meals ingested by the female must be capable of supporting her own needs, the needs of the developing juvenile and supplying an excess to store for the pupal development stages. Once the

juvenile adult fly emerges from the pupa, it must inflate its wings and feed, before reaching sexual maturity approximately 14 days after emergence (Mellanby, 1937). Figure 1.3 below illustrates the life of the tsetse fly. This slow reproduction cycle is a critical aspect of tsetse population control, as estimates state that a daily mortality rate of 4% within the female population will result in the rapid extinction of the population (Hargrove, 1988).

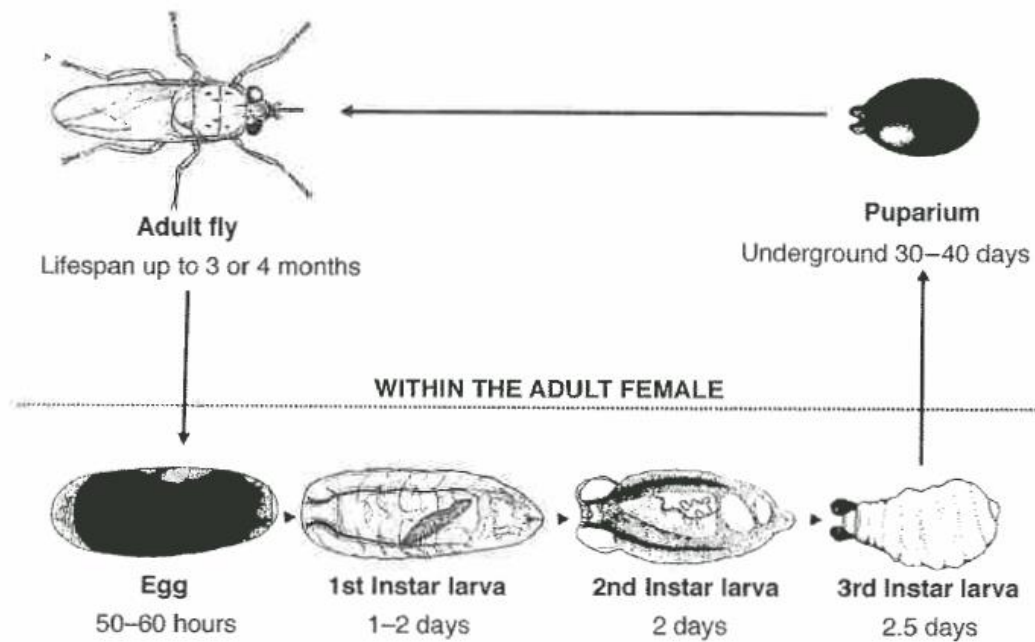


Figure 1.4: The life cycle of *Glossina* spp. (Leak, 1999).

### 1.2.1ii: Tsetse population control

In the absence of an effective vaccine and inexpensive treatments for HAT, along with mounting resistance to current trypanocidal prophylactics, vector control is now considered the most effective and sustainable way of managing African Trypanosomiasis (Tirados *et al.*, 2015; Meyer *et al.*, 2018; Percoma *et al.*, 2018). In an attempt to eradicate HAT and reduce the economic impact of AAT throughout the African continent, four primary methods of tsetse control have been adopted (Esterhuizen, Njiru *et al.*, 2011; Esterhuizen, Rayaisse *et al.*, 2011; Abd-Alla *et al.*, 2013; Vale, Hargrove, Lehane *et al.*, 2015):

- Insecticide treated, odour-baited target traps.
  - These targets have three components; a visual target, typically either a black or blue sheet of cloth; an odour attractant consisting of 1-octen-3-ol (octenol) (0.5 mg/h) and either acetone (100 mg/h) or butanone (15 mg/h);



and an insecticide treatment applied to the visual target (Vale *et al.*, 1988; Langley *et al.*, 1993).

- Pyrethroid-treated cattle.
  - Pyrethroids are highly effective as insecticidal synthetic compounds with a low level of mammalian toxicity. Treatment of cattle by aerosol spray is considered the cheapest tsetse control method, however the ecological impact on other beneficial fauna is highly contentious (Hargrove *et al.*, 2000; Ndeledje *et al.*, 2013; Vale *et al.*, 2015).
- Ground and aerial insecticide spraying techniques.
  - This method aims to interrupt the life cycle of the tsetse by targeting adult females and larval stages following emergence from the pupa. As with all indiscriminate insecticide treatments, the ecological impacts of this method are high (De Deken and Bouyer, 2018).
- Sterile insect methods.
  - Sterilisation methods within tsetse utilise symbiont interaction to regulate fertility (Abd-Alla *et al.*, 2013). Although this method was effective at eradicating an isolated population of *Glossina austeni* in Zanzibar (Vreysen *et al.*, 2000), the cost of maintaining sterile populations and the logistical issues associated with the method in non-isolated populations may not justify this control method (Abd-Alla *et al.*, 2013).

The success of vector control on disease management is evident from the continued reduction in HAT cases over the last 25 years (Hide, 1999; Gao *et al.*, 2020). However, the total cost of sustained vector control suggests that alternative solutions are required. The cost of continuous vector control over a 20-year period with no discounts is estimated to range between US\$894/Km<sup>2</sup> for insecticide treated cattle to US\$11,666/Km<sup>2</sup> for 10 traps/Km<sup>2</sup> (Shaw *et al.*, 2013). While the prolonged employment of these control measures can dramatically reduce tsetse populations, the eradication of a population is considerably more difficult and often requires an integrated approach, combining several control methods (Percoma *et al.*, 2018).

While the efficiency of vector control cannot be denied, there is a necessity to maintain control measures almost indefinitely. Following the termination of tsetse control in the

Lambwe Valley of western Kenya, populations returned to and stabilized at pre-treatment levels within 12 months (Turner and Brightwell, 1986). Modelling of tsetse infestation of cleared areas suggested that a cleared area of 100Km<sup>2</sup> could be lost within a year, while 10,000Km<sup>2</sup> could be lost in just two years if there was no bar to migration and breeding (Hargrove, 2000). It is critical therefore, that more economically stable and enduring methods of trypanosome control are investigated. One such potential avenue of pathogen control is to utilise the natural interactions between vectors and parasites to break transmission, however, a detailed understanding of these mechanisms is required before this can be explored in detail.

### 1.3: Parasite - Vector Interactions

The interactions between parasites and vectors are critical to the survival and development of the parasite. Parasites must be able to conceal their presence from the immune system of both the host and vector to avoid triggering an immune response. Dipteran and other arthropod vectors do not possess an adaptive immune system like mammalian hosts, but rather rely on the enzymes and anti-microbial peptides (AMPs) comprising their innate immune system. This immune response is genetically predetermined and could therefore potentially be manipulated if fully understood, however gaining a comprehensive understanding of the complexity of the interactions between pathogen and host requires extensive research.

The interactions between vector and parasite species result in an evolutionary arms race under the concept of coevolution (Anderson and May, 1982). This theory states that the interactions between the two organisms is one of the primary evolutionary drivers within individuals (Anderson and May, 1982; Feeney *et al.*, 2012). The concept of coevolution is strongly linked to Red Queen hypothesis (Van Valen, 1973) which comprised two contrasting theories (the Red Queen arms race and Red Queen dynamics) that influence the evolution of the interacting species. Equally, the traditional understanding of a parasite-vector relationship states that the parasite relies upon one single vector to survive, and the extinction of the vector will result in the extinction of the parasite. However, recent observations of seemingly random “jumps” to new vectors are common and observable in real time; this process is described as the “Parasite paradox” (Agosta *et al.*, 2010) and illustrates the complexity of these organisms and the need for further research.

### 1.3.1: Dipteran immunity

The most prominent interaction between pathogens and hosts is the immune response to infection. As mentioned above, dipteran species do not possess an adaptive immune system, relying instead solely upon their innate immune system to defend against infection. This innate response is, however, not indiscriminate but initiated by the interactions of pathogen detector proteins (PDPs) and pathogen-associated molecular patterns (PAMPs) to mount a limited adaptive response. Capable identifying and initiating differing predetermined responses to gram-negative and gram-positive bacteria, viruses, fungi and parasites (Wachinger *et al.*, 1998; Imler and Hoffmann, 2000; Hao *et al.*, 2001; Ageitos *et al.*, 2017).

The innate immune system of both vertebrates and invertebrates is comprised of two separate pathways (Akira *et al.*, 2006). Firstly, the cellular pathway incorporating macrophage-like cells to phagocytose pathogens and micro-organisms. Secondly, the molecular pathway, which utilises PAMPs to identify pathogens and AMPs to combat infection. This study focuses solely on the molecular pathway, as the cellular pathway functions exclusively in the hemocoel and, thus, is not triggered by trypomastigote infection.

The molecular innate response follows a series of signalling pathways to stimulate AMP expression in response to pathogen detection, the Toll-like (TLR), immune deficiency (IMD) and Jak/Stat pathways are vital for the identification of pathogenic infections (Rolff and Reynolds, 2009; Caljon *et al.*, 2014). These pathways result in the expression of specific AMPs to combat the detected infection (Figure 1.5).

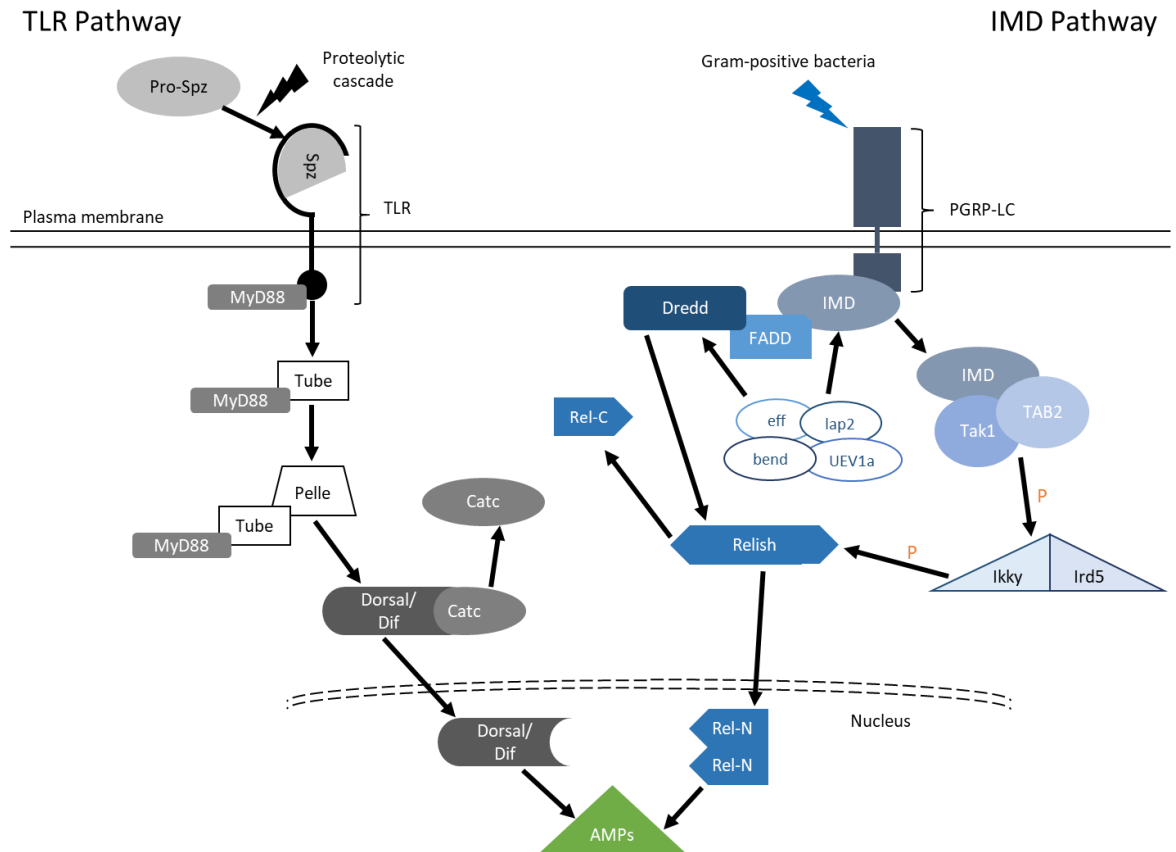


Figure 1.5: A simplified model of the Toll/Dif/Dorsal and IMD-Rel signaling cascades in *Drosophila melanogaster*. The TLR pathway was adapted from the model published by Lemaitre *et al.*, (1996), while the IMD pathway was adapted from the model published by Myllymäki *et al.*, (2014).

### 1.3.1i: The TLR pathway

The TLR pathway is present in both the vertebrate and invertebrate immune system, TLR proteins are highly conserved, transmembrane proteins found across the cell plasma membrane. First identified in 1985 within *Drosophila*, a total of nine TLR gene (TLR1-9) have now been identified within the *D. melanogaster* genome (Anderson *et al.*, 1985; Valanne *et al.*, 2011; Levin and Malik, 2017). In addition to the immunological function, TLRs have been shown to also play a vital role in, development, providing vital cues for dorsal/ventral differentiation within the early stages of growth (Anderson *et al.*, 1985; Hashimoto *et al.*, 1988; Jang *et al.*, 2006).

Structurally TLR proteins are comprised of three “subdomains”. The largest being the N-terminal ectodomain, typically consisting of 16-23 leucine-rich repeats (LLRs) (in mammalian proteins) this forms the basis for PAMP recognition (Leulier and Lemaitre, 2008; Kumar *et al.*, 2009). LLRs form parallel  $\beta$ -sheets on the concave surface while helices form the convex outer surface, generating their distinctive horseshoe shape (Fig. 1.5A) (Bell

*et al.*, 2005; Jin *et al.*, 2007). Following the ectodomain is a single helical transmembrane region, this traverses the phospholipid bilayer connecting the ectodomain to the *Toll/IL-1* receptor (TIR) domain (Bell *et al.*, 2005). Structurally the TIR domain consists of five central parallel  $\beta$ -sheets, surrounded by  $\alpha$ -helices and a single “BB loop” (Fig. 1.5B) (Xu *et al.*, 2000; Khan *et al.*, 2004). Whilst it is the LLRs that give TLRs their horseshoe shape, the TIR is responsible for the dimers that form between them, resulting in the “M” shape most often associated with TLRs (Fig. 1.5C). The formation of both homo and heterodimers between TLR proteins enables them to detect and respond to different pathogens (Khan *et al.*, 2004; Jin *et al.*, 2007). TLR-4 predominately forms homodimers while TLR2 will form dimers with TLR1 and TLR6 (Zhang and Ghosh, 2002).

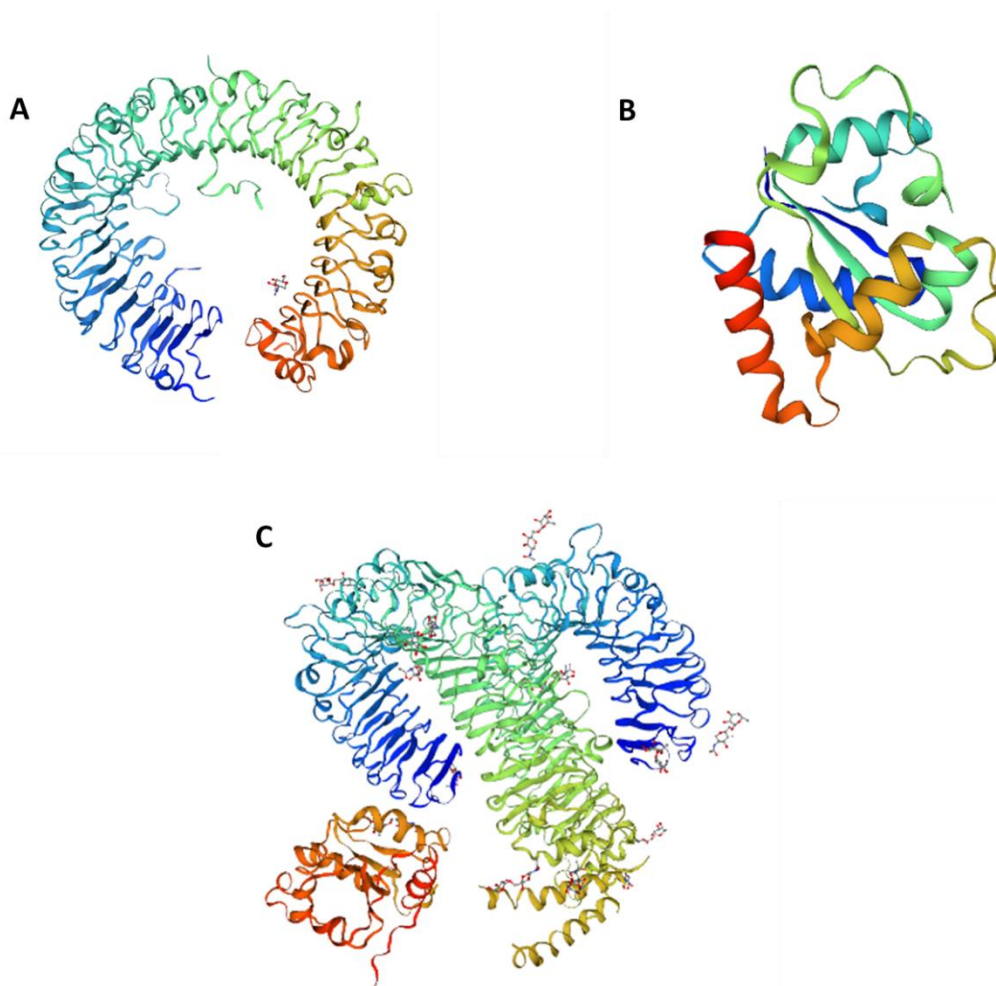


Figure 1.6: The overall structure of TLR proteins illustrating all three subdomains. A) The ectodomain of TLR 9 as a monomer (Ohto *et al.*, 2015). B) The structure of the TIR domain, comprising the central  $\beta$ -sheets surrounded by  $\alpha$ -helices as recorded by Xu *et al.* (2000). C) The homodimer structure of TLR5 exhibiting all three subdomains. The ectodomain can be seen in green and blue; the transmembrane region in yellow and the TIR in red (Zhou *et al.*, 2012). All images produced using SWISS-MODEL (Guex *et al.*, 2009; Waterhouse *et al.*, 2018).

While the structure and function of TLRs is conserved among all classes, the mode of action varies between vertebrates and invertebrates. The vertebrate TLR pathway relies on the direct interaction of the ectodomain and pathogen-associated molecule proteins (PAMPs), while the arthropod TLR pathway relies on the binding of endogenous ligand proteins to stimulate the TLR pathway and immune response (Stein and Nüsslein-Volhard, 1992). The primary endogenous ligand protein involved in dipteran immune responses, and embryo development, is Spätzle (Spz) (Weber *et al.*, 2003). Spätzle is an evolutionary conserved, dimeric protein that is synthesized as a pro-protein requiring maturation via a series of conformational changes induced by a serine protease in order to bind to TLR proteins (Weber *et al.*, 2003; Buchon *et al.*, 2009).

The arthropod TLR signalling cascade (Fig. 1.4) is initiated by the activation of Spz. Following the identification of a pathogen, by pathogen receptor proteins, such as gram-negative binding proteins (GNBPs), the proteolytic cascade is triggered. This results in cleaving of Spz and the release of the pro-domain, exposing the C-terminal, thereby enabling Spz to bind with the TLR extracellular domain (Lemaitre *et al.*, 1996; Weber *et al.*, 2003; Tanji *et al.*, 2007; Arnot *et al.*, 2010; Valanne *et al.*, 2011). Upon Spz-TLR binding, the TIR domain binds to Myeloid differentiation primary response 88 (MyD88), which subsequently binds to Tube and Pelle, forming the MyD88-Tube-Pelle heterotrimeric complex (Horng and Medzhitov, 2001; Sun *et al.*, 2002; Tauszig-Delamasure *et al.*, 2002). This complex is vital for the degradation of the Dorsal/Dif-Cactus (Cact) complex, enabling the nuclear translocation of Dorsal/Dif, which in turn results in the synthesis of AMPs, such as Attacin, Defensin and Dipterecin (Wu and Anderson, 1998; Akira *et al.*, 2006; Valanne *et al.*, 2011).

### 1.3.1ii: Anti-microbial peptides and proteins

AMPs are a diverse protein superfamily and are frequently characterised by their size, structure, and function. Smaller AMPs (peptides) are typically defined as between 12 and 50 amino acids, while the larger AMPs (proteins) typically exceed 100 amino acids in length (Ganz, 2003). The structure and function of AMPs is highly diverse helping to maintain the efficiency of the innate immune system against different pathogens. Antimicrobial proteins can be characterised into four broad structural groups (Reddy *et al.*, 2004; Brogden, 2005; Dhople *et al.*, 2006):

- $\alpha$ -helical dominant structures (e.g.: Cecropin and andropin);

- $\beta$ -sheet dominant structures (e.g.: tachyplesins and Lactoferricin B);
- Sequences rich in cysteine residues (e.g.: Defensins and cryptidins);
- Sequences rich in other specific amino acid residues: namely proline, glycine, histidine and tryptophan (e.g.: Attacins and drosocin).

And exhibited one of two primary modes of action (Yeaman and Yount, 2003; Brogden, 2005; Otvos, 2005; Torrent, et al., 2012):

- Cell membrane disruption resulting in cell lysis and death;
- Direct interference of essential intracellular mechanisms such as protein synthesis and folding.

The amino acid sequence and structure of AMPs dictates the mode of action, with characteristic structure and amino acids resulting in specific functions. For example, cecropins are small (31-37 aa),  $\alpha$ -helical AMPs (Steiner, et al., 1981; Hoskin and Ramamoorthy, 2008) known to form ion channels or pores in the cell membrane using the “toroidal pore” model (Huang, 2000; Reddy, et al., 2004). This membrane disruption is only possible due to the hydrophobic and cationic surface of the protein that results from the disulphide dependent, twin helical structure (Reddy, et al., 2004; Brogden, 2005). Other membrane disruption methods such as the “barrel-stave” and “carpet” models rely on different structural and charge variations. The “barrel-stave” model relies on the small, amphipathic helical AMPs that directly insert into the cell membrane. As the hydrophobic surface binds to the centre of the membrane, the arguments of hydrophilic surface produces a pore which is increased in size as more monomers are recruited (Yang, et al., 2001; Reddy, et al., 2004; Brogden, 2005). The “carpet” model of pore formation results from a high concentration of AMPs aligning parallel to the cell membrane. Membrane disruption is achieved as the hydrophilic surfaces reorientate towards the core of the membrane causing the disintegration of the membrane (Reddy, et al., 2004; Brogden, 2005).

In order for AMPs to interfere directly within intracellular pathways, they must first cross the cell membrane without causing terminal damage. The mechanism for this has been hypothesised to be similar to that of Cell-penetrating peptides (CPPs) (Nicolas, 2009; Torrent, et al., 2012). One method of intracellular action is to disrupt ATP production, as exhibited by the human AMP histatin 5 (Luque-Ortega, et al., 2008). Following exposure to

Hst5 the trypanosomatid parasite *Leishmania* shows signs of morphological changes to the mitochondria and reduced respiration rate, followed by collapse of the parasitic cell (Luque-Ortega, et al., 2008; Torrent, et al., 2012). As the role of AMPs and the innate immune system cannot be understated in combatting pathogenic infection, it presents a unique opportunity as genetic control target to break parasitic transmission, however, the extent of the interactions between the parasite and vector must be understood before such control methods can be contemplated.

### 1.3.2: *Trypanosoma brucei* spp. and tsetse interactions

The establishment of an effective trypanosome colony within the midgut or salivary gland of a tsetse fly relies upon several factors permitting trypomastigotes to overcome the refractoriness of the tsetse fly (Akoda, et al., 2009). Research into the factors influencing trypanosome infection in tsetse flies has been continuing since the early 20th century. Lloyd stated in 1930 that the abundance of 'factors' influencing the infection rate of trypanosomes in tsetse made it impossible to accurately measure the transmission rates of any single strain (Lloyd, 1930) and so it was not until the late 20th century that the full extent of the trypanosome-tsetse relationship became clear. While external factors influencing infection have been identified, complex parasite-host interactions are emerging as a vital area for research, especially the innate immune response of the tsetse fly against trypanosome infection through AMPs and enzymes.

#### 1.3.2i: Physical and internal factors

In addition to the immune response within a vector, several other factors directly influence the success rate and intensity of trypanosome infection. Temperature, sex, age and food abundance, are documented as affecting midgut infection rates (Dipeolu and Adam, 1974; Otieno, et al., 1983; Akoda, et al., 2009). The effect of temperature was recorded in the early 20th century, with both Lloyd (1930) and Taylor (1932) observing that increasingly substantial trypanosome infections were indicative of a higher temperature of approximately 37 °C, whilst temperatures between 20 - 30 °C resulted in a decreased transmission rate. Sex and age of the tsetse fly also plays a critical role, with juvenile males (up to 32 hours after emergence) being the most susceptible to infection (Distelmans, et al., 1982; Otieno, et al., 1983). It has been observed further that nutritional stress has a



direct effect on the concentration of AMPs present in the tsetse midgut both prior to and following infection (Akoda *et al.*, 2009). Akoda *et al.* (2009) also noted that tsetse presenting a lower concentration the AMP Attacin were more susceptible to infection, reinforcing Attacin's contribution as a parasiticide agent (Akoda, et al., 2009; Beschin, et al., 2014).

Physical factors causing bottlenecks to the effective transmission of trypanosomes can be found throughout their life cycle. Caljon *et al.* (2014) observed that the first three days following an infected blood meal are vital to the establishment of a successful trypanosome colony within the tsetse fly. Within this time frame, trypanosomes must penetrate the peritrophic matrix lining the epithelial wall of the tsetse midgut, enter the ectoperitrophic space and continue to the salivary glands. This migration presents the largest bottleneck of their life cycle. Less than 5 % of the trypomastigotes penetrate the salivary glands successfully and complete their life cycle (Oberle, et al., 2010).

#### 1.3.2ii: Tsetse innate immune system

Current literature regarding the tsetse immune response to trypanosomal infection has focused almost solely on the expression of AMPs. Therefore, information on the interactions between trypanosomes and Toll-like proteins within *Glossina* species is currently limited. However, four TLRs (TLR2, TLR4, TLR6 and TLR9) have been reported to recognise ligands from the *Trypanosoma* genus, with the majority of research having been conducted using the causative agent of Chagas Disease, *Trypanosoma cruzi*, and its triatomine vectors (Bafica *et al.*, 2006; Uematsu and Akira, 2008; Kumar *et al.*, 2009). There are several trypanosomal ligands/PAMPs recognised by TLRs as shown in Table 1.1.

Table 1.1: TLRs and their corresponding ligands/PAMPs responsible for trypanosome identification (Uematsu and Akira, 2008; Kumar *et al.*, 2009).

TLR	Ligand/PAMP
TLR2	Glycoinositolphospholipids Glycylphosphatidylinositol anchors Unsaturated alkylacylglycerol Lipophosphoglycan
TLR4	Glycoinositolphospholipids Glycylphosphatidylinositol anchors
TLR6	Glycylphosphatidylinositol anchors
TLR9	Genomic DNA

It should be noted however, that the identification of Glycylphosphatidylinositol anchors by TLR6 is reliant upon the formation of a heterodimer with TLR2 (Uematsu and Akira, 2008). Consequently, the binding of these ligands/PAMPs stimulates the Toll pathway which, through interactions with secondary signalling molecules, results in the expression of AMPs including attacin and defensin.

Boulanger *et al.*, (2002) identified four primary AMPs, namely attacin, defensin, cecropin and dipterin, expressed during the tsetse immune responses to *T. brucei brucei*. Hu and Aksoy (2006) concluded similarly that AMPs are a major factor in the tsetse innate immune response to trypanosome infection following research concerning the IMD pathway. Notably, cecropin, attacin and dipterin were identified in the tsetse midgut (Hu and Aksoy, 2006; Roditi and Lehane, 2008) following both bacterial and parasitic infection (Hao *et al.*, 2001; Boulanger *et al.*, 2002). Hao *et al.*, (2001) concurred with this observation, commenting further the tsetse innate immune system indicated pathogen specific responses, recognising differences between bacterial, fungal, and parasitic infections as well as life cycle stages of trypanosomes present within the midgut.

Boulanger *et al.*, (2002) observed further that the expression of three of the AMPs, defensin, attacin and cecropin, as a result of trypanosome infection was seen only in the first week following an infectious blood meal. Dipterin, however was, and is, constantly expressed throughout the tsetse adult life and is unregulated during trypanosome infections (Boulanger *et al.*, 2002, 2006).

### 1.3.3: The role of endosymbionts in trypanosome transmission

Symbiotic relationships between dipteran species and bacteria are common, often supplying the host crucial nutrients that their restrictive diets (mammalian blood in the case of tsetse) cannot provide (Bing *et al.*, 2017). However, these relationships are also emerging as a significant factor in the successful transmission of African trypanosomiasis.

*Glossina* are known to harbour three genera of endosymbiotic bacteria, *Wolbachia*, *Wigglesworthia*, and *Sodalis* (Kikuchi *et al.*, 2009; Balmand *et al.*, 2013; Sasser *et al.*, 2013). The primary endosymbiont found in tsetse flies is the gram-negative bacteria *Wigglesworthia glossinidia* subsp. (Pais *et al.*, 2008; Sasser *et al.*, 2013). These bacteria reside both extra and intercellularly within the milk gland and a bacteriome organ (Wang *et al.*, 2009; Caljon *et al.*, 2014). The association between *W. glossinidia* and tsetse has become obligatory, meaning that tsetse flies are dependent upon this endosymbiotic relationship which influences their life cycle in two ways, the sexual maturation of female tsetse flies and the development of the innate immune system within juveniles. Pais *et al.* (2008) documented that female flies bred without *W. glossinidia* were sterile and unable to reproduce whilst the males were unaffected by this variation. Secondly, it was observed that whilst juvenile flies without *W. glossinidia* had a similar trypanosome infection rate to natural juvenile flies, this changed in adulthood. Adult flies without *W. glossinidia* were considerably more susceptible to trypanosome infection than the natural flies (Kikuchi, 2009; Sasser *et al.*, 2013).

*Sodalis glossinidius* is another endosymbiont characteristically associated with the tsetse - trypanosome tripartite. The involvement and full interactions of *S. glossinidius* in trypanosome infections remains unclear. However, unlike *W. glossinidia* its presence within the tsetse host appears to be purely mutualistic rather than obligatory (Toh *et al.*, 2006). Despite this however, the presence of *S. glossinidius* has been shown to significantly promote trypanosome infection of tsetse (Dale and Maudlin, 1999; Toh *et al.*, 2006).

The gram-negative bacterium *Wolbachia* has been observed in many parasite-host relationships, being present in the Culicidae vectors of nematodes and *Plasmodium* spp. (Kramer *et al.*, 2008; Kambris *et al.*, 2010). Whilst the function of *Wolbachia* sp. remains unclear in the interactions of the nematode *Dirofilaria immitis* and its mosquito vectors, it has been theorised that the bacterium plays a role in parasite sexual maturation whilst

simultaneously masking the presence of the nematode once in the definitive host (Holley, 2011). However, the role of *Wolbachia* in protozoan-vector interactions appears to differ to that observed between nematode-vector interactions. Interestingly, the presence of *Wolbachia* sp. within a tsetse exhibited two results: firstly, when present in the germ line tissue of female tsetse flies *Wolbachia* appears to inhibit fertilization of eggs resulting in the termination of the larval tsetse (Cheng *et al.*, 2000). Secondly, it was noted that *Wolbachia* presented a strong anti-parasitic reaction within the vector resulting in the clearance of trypanosome infection (Kambris *et al.*, 2010; Alam *et al.*, 2011; Sasser *et al.*, 2013).

Interestingly, all three endosymbionts are gram-negative bacteria. This could account for the developmental relationship between endosymbionts, such as *W. glossinidia*, and the tsetse immune system. With both attacins and defensins targeting primarily gram-negative bacteria the presence of these endosymbiotic bacteria could result in a continuous low-level expression of AMPs. This would, in turn, increase the resistance to trypanosome infection minimising the time required to mount a full immune response.

#### 1.4: Gaps in knowledge

Despite increasing research into both the innate immune system and endosymbiont interactions on trypanosome transmission, several fundamental questions remain. Firstly, how could the tsetse-symbiont-trypanosome tripartite be used to break the transmission cycle of African Trypanosomiasis? As stated previously, current control measure of AT focus almost exclusively on vector control. While this has been proven successful in reducing cases of HAT, the successful elimination of a vector population incurs large costs and, if not maintained following termination of the control program, vector populations are likely to return to pre-treatment levels (Turner and Brightwell, 1986; Hargrove, 2000; Shaw *et al.*, 2013). Furthermore, there is an increasing awareness of the ecological impacts of untargeted insecticide spraying. While this method only accounts for a one aspect of tsetse control, the global decline in insect populations requires an innovative solution (Forister *et al.*, 2019; Simmons *et al.*, 2019). Therefore, there is an ongoing demand for novel, economically and ecologically sustainable control methods.

Genetic intervention such as the introduction of sterile male tsetse, similar to those documented by Benedict and Robinson (2003) to combat malaria, would ultimately result in a short-term increase of potential vectors as, unlike mosquitoes, both male and female are obligate haematophages and therefore the release of sterile individuals would only increase the risk of trypanosome transmission (Caljon *et al.*, 2014). It is, therefore, vital that an alternative solution be identified. One such solution presented by Caljon *et al.* (2014) was the manipulation of the *S. glossinidius*-tsetse relationship. This relationship presents an opportunity to examine the complex interactions of endosymbionts and hosts to introduce genetic interventions to trypanosomes before transmission to mammalian hosts can occur. The manipulation of the endosymbiotic relationship between *W. glossinidia* and *Glossina* would offer a long-term, economically, and ecologically sustainable control method. However, for this method to be considered, a greater understanding of the extent of *S. glossinidius* and *W. glossinidia* interactions with the tsetse fly and the influence on *Trypanosoma* infection must be gained.

Further, to the interactions of the tripartite above, there is a severe lack of literature concerning the inter- and intra-species variation of the tsetse innate immune system. Understanding the fundamental aspects and population dynamics of tsetse immune responses is essential to establishing potential future genetic control measures. Given the natural refractory nature of tsetse to trypanosomal infection, it could be possible to promote resistance to trypanosome infection and thus stop transmission within the vector. However, as with the manipulation of endosymbionts, this requires a far greater understanding of the fundamental aspects of tsetse immunity than is currently published.

## 1.5: Aims and objectives

The aim of this thesis is to undertake an evolutionary study of three important immune genes within the *Glossina* genome, and to provide insights into the relationship between tsetse evolution, bacterial symbiosis, and trypanosome infection. Developing our understanding of these interactions is central to the development of future genetic control measures for African Trypanosomiasis.

In this study we will focus on two AMP families, attacin and defensin, and one receptor gene family, the TLRs. Given the paucity of published literature concerning these genes

within the *Glossina* genus we further aim to identify each of these gene families within the available *Glossina* genomes. There this achieve our aim five specific objectives were investigated:

- Objective 1: To identify, characterise and map the attacin and defensin gene families, and to assess interspecies variation within the available *Glossina* genomes.
  - Previously published nucleotide sequences will be used to identify members of the attacin and defensin gene families within the six available *Glossina* genome assemblies. Identification of the genomic loci and structure of each gene will be achieved using tBLASTn to identify related gene sequences. Furthermore, genetic variation and protein structure will be assessed across the *Glossina* genera.
- Objective 2: To evaluate the intraspecies variation of attacin-A and defensin at population level in relation to endosymbionts and trypanosome infection.
  - This will be achieved using newly extracted gDNA and PCR amplification to produce nucleotide sequences for both *AttA* and *Def*. These will then be submitted to standard evolutionary and population genetic analysis to examine nucleotide variation in relation to *Wigglesworthia* and trypanosome infection both within a wild population of *G. m. morsitans*.
- Objective 3: To assess the impacts of selection upon both the structural and functional aspects of these immune genes.
  - This will be done using evolutionary genetic analysis to determine the location and nature (positive, negative, or neutral) of selective pressures being exerted on amino acid mutations. Furthermore, three-dimensional modeling of protein variants will be used to visualise any alteration in the protein structure, which could be indicative of a functional change.
- Objective 4: To identify, characterise and map the TLR genes within the available *Glossina* genomes. And assess the variation within TLR genes across the *Glossina* genus.

- This will be achieved using cross genomic analysis to provide a template of TLR nucleotide sequences from previously identified genes in *D. melanogaster*. These templates can then be used to identify TLR genes within the *Glossina* genome assemblies available on VectorBase.
- Objective 5: To evaluate the inter- and intra-species variation of Toll-Like Receptors in relation to symbiont and trypanosome infection.
  - This object will be achieved using the same methods and approaches utilised to address objectives 2 and 3.

## 2: The identification and characterisation of the attacin clusters and defensin genes within *Glossina* species genome assemblies

### 2.1: Introduction

The innate immune system is an integral part of the tsetse response to trypanosome infection. Following stimulation of the Toll-like (TLR) and immune deficiency (IMD) pathways, antimicrobial peptides (AMPs) are synthesised in order to combat the invading pathogens (Fig. 1.5). Antimicrobial peptides are a diverse superfamily and are frequently characterised by their size, structure and function, capable of combatting a range of predetermined pathogens including gram-negative and gram-positive bacteria, viruses, parasites and fungi (Wachinger *et al.*, 1998; Ageitos *et al.*, 2017).

The broad spectrum of properties exhibited by AMPs helps to maintain the efficiency of the innate immune system. There are two primary modes of action associated with AMPs: i) cell membrane disruption resulting in cell lysis and death, and ii) direct interference of essential intracellular mechanisms such as protein synthesis and folding (Yeaman and Yount, 2003; Brogden, 2005; Otvos, 2005; Torrent *et al.*, 2012). These mechanisms are outlined in greater detail in Chapter 1.

Antimicrobial proteins can be further characterised into four broad structural groups: i)  $\alpha$ -helical dominant structures, ii)  $\beta$ -sheet dominant structures, iii) sequences rich cysteine residues and, iv) sequences rich in other specific amino acid residues (namely proline, glycine, histidine and tryptophan) (Reddy *et al.*, 2004; Dhople *et al.*, 2006). While these groups broadly explain the diversity within the AMP superfamily, individual AMP families tend to show high levels of both inter and interspecies conservation.

One important AMP family found within the Arthropoda is the attacin protein family. First identified in *Hyalophora cecropia*, attacins are a glycine-rich AMP family, weighing approximately 20 kDa, and primarily associated with the immune response to gram-negative bacteria (Hultmark *et al.*, 1983; Imler and Bulet, 2005). Their expression is controlled by a combination of the IMD/Relish, Toll/Dif and Toll/Dorsal pathways following



the activation of the signalling protein Spätzle by pathogen specific binding proteins (Fig. 2.1) (Imler and Hoffmann, 2000; Hao *et al.*, 2001). Their mode of action is considered to disrupt the outer membrane of pathogens, resulting in increased permeability, changes to cell shape and disrupted cell cycles (Engström *et al.*, 1984; Bulet *et al.*, 1999; Ravi *et al.*, 2011). The exact method by which this is achieved is yet to be determined, however, experimental data into the mode of action of gloverin (another glycine rich AMP) indicated that cell lysis was achieved through inhibiting the synthesis of important membrane proteins, increasing permeability. Engström *et al.* (1984) also observed that the presence of attacin had a direct effect on the AMP cecropin B, with cells becoming far more susceptible to cecropin B following exposure to attacin. To date, four attacin paralogues (attacin-A (*AttA*), attacin-B (*AttB*), attacin-C (*AttC*) and attacin-D (*AttD*)) have been identified within the *D. melanogaster* genome (Hedengren *et al.*, 2000; Lazzaro and Clark, 2001).

Literature regarding the secondary and tertiary structure of attacin proteins is scarce; a preliminary study Gunne *et al.* (1990) concluded that the secondary structure likely featured several random-coil structures. However, recent developments in protein modelling, using AlphaFold 2.0 software, have predicted the structures for *D. melanogaster AttA* and *AttB* (available on the Pfam database) (El-Gebali *et al.*, 2019; Jumper *et al.*, 2021). These models suggest that the structures of both *AttA* and *AttB* proteins include a partial enclosed channel, consisting of 12 consecutive anti-parallel  $\beta$ -sheets (Fig. 2.1); the primary difference between the two proteins is the presence of a single  $\alpha$ -helix at the 5' terminal of *AttA* (Fig. 2.1).

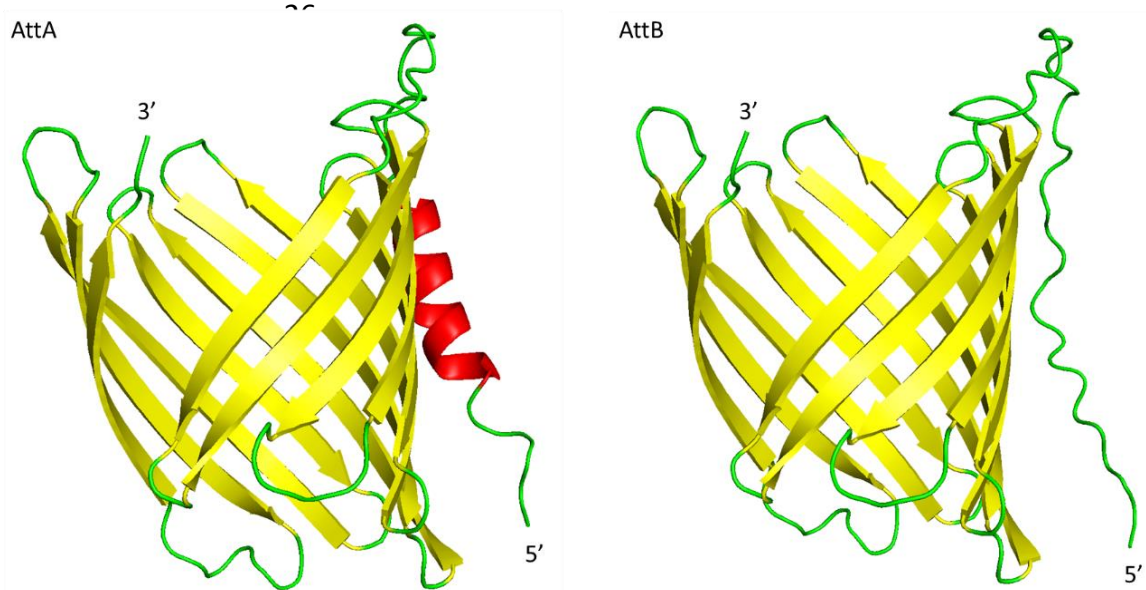


Figure 2.1: Predicted structures of *D. melanogaster* AttA and AttB produced using AlphaFold V2 (Jumper *et al.*, 2021), using UniProt entries P45885 (AttA) and Q9V751 (AttB). Secondary structures are represented by colour, coils are shown in green, helices in red and  $\beta$ -sheet in yellow.

Three attacin genes, namely *AttA*, *AttB* and *AttD*, are encoded within the *G. m. morsitans* genome (Wang *et al.*, 2008; Trappeniers *et al.*, 2019). Comparative analysis of these genes with *Drosophila* orthologues revealed two defining features. Firstly, both *Glossina AttA* and *AttB* are missing the propeptide (activation) domain present in the *D. melanogaster* orthologues. Although, both *Glossina* orthologues do exhibit similar binding sites for both NF- $\kappa$ B and AP-1 within the promoter regions (Senger *et al.*, 2004; Wang *et al.*, 2008). Secondly, *G. morsitans AttD* appears to be missing both the pro- and pre-peptide domains characteristic of other insect attacins, with no obvious promoter binding sites (Hao *et al.*, 2001; Wang *et al.*, 2008).

Wang *et al.* (2008) grouped the attacin genes within *G. m. morsitans* into three clusters of individual paralogues, though there is currently no literature available concerning their origins or functional distinction:

- **Cluster 1:** contains two paralogues of *AttA* in a head-to-head orientation.
- **Cluster 2:** contains a third *AttA* paralogue and an *AttB* gene again in a head-to-head orientation.
- **Cluster 3:** contains a single *AttD* gene (Fig. 2.3).

Wang *et al.* (2008) demonstrated that *AttA* and *AttB* were structurally identical, except for two amino acid substitutions: His187Asn and Gln195Arg (*AttA/AttB*). The authors

further stated that nucleotides encoding *AttD* varied more significantly from the other attacin genes.

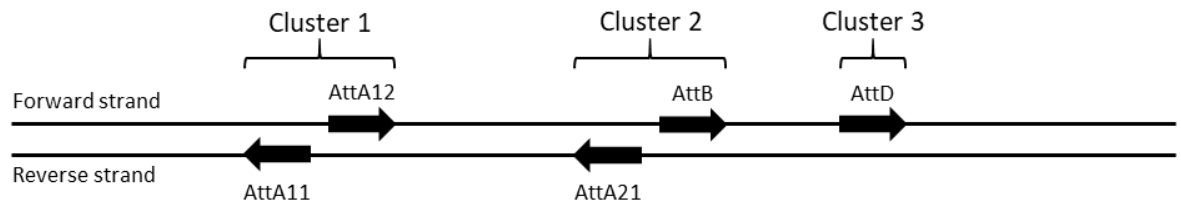


Figure 2.2: A linear representation of the attacin clusters within the *G. m. morsitans* genome as proposed by Wang *et al.* (2008) (not to scale).

Attacins are crucial in combating trypanosome infection (Hao *et al.*, 2001; Boulanger *et al.*, 2002; Hu and Aksoy, 2006; Roditi and Lehane, 2008). *AttA* and *AttB* are expressed in response to both gram-negative bacteria and trypanosomes, while *AttD* expression is only induced upon trypanosome infection (Wang *et al.*, 2008). There is limited literature concerning the expression of attacins in relation to other parasite-vector interactions. Christophides *et al.*, (2004) identified just one family of attacin proteins present in the malarial vector *Anopheles gambiae*. However, attacin does not appear to be synthesized in response to *Plasmodium* spp. infection (Lehane *et al.*, 2004).

Another family of AMPs critical to the tsetse immune response are the insect defensins. These are small, cysteine-rich AMPs characterised by the presence of six conserved cysteine regions that form stabilising disulphide bonds (Varkey *et al.*, 2006) (Figure 2.4). Both the IMD/Relish and Toll/Dif pathways regulate the expression of defensin within the tsetse immune response (Imler and Hoffmann, 2000). The conserved C-terminal structure of insect defensin has been thoroughly documented and relies heavily on the aforementioned disulphide bonds; an N-terminal loop, leads to a  $\alpha$ -helix and an antiparallel  $\beta$ -sheet to form a cysteine-stabilized alpha beta structure (Figure 2.4) (Bonmatin *et al.*, 1992a; Bonmatin *et al.*, 1992b; Cornet *et al.*, 1995; Yi *et al.*, 2014). The mode of action of defensin has been documented to follow either the “toroidal pore” or “barrel-stave” model to form channels within the cell membrane, resulting in the loss of cytoplasmic potassium, depolarisation of the membrane and inhibition of respiration (Cociancichs *et al.*, 1993; Aksoy, 1995; Yang *et al.*, 2001; Reddy *et al.*, 2004; Yi *et al.*, 2014).

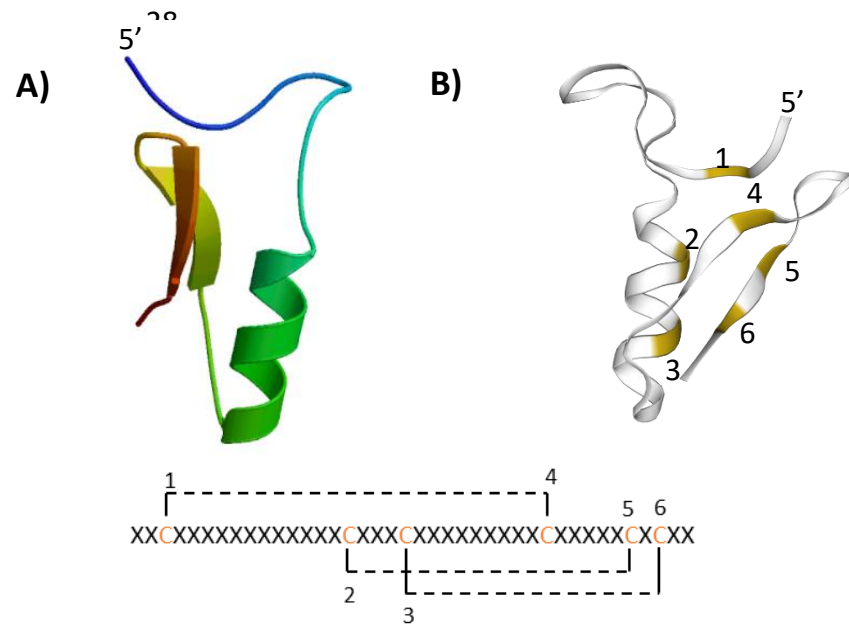


Figure 2.3: The structure of the C-terminal mature Def region of insect defensin-A as documented by Cornet et al., (1995). A) The overall C-terminal structure of the arthropod defensin-A protein. B) The structure of the arthropod defensin-A protein highlighting the position of cytosine residues and their interactions forming disulphide bonds in the conserved arthropod defensin domain as defined by Pfam (El-Gebali et al., 2019). Structures produced by SWISS-MODEL, template SMTL = 1ica.1 (Guex et al., 2009; Waterhouse et al., 2018).

Literature regarding *Glossina* defensins is minimal; however, Hao *et al.* (2001) briefly described the defensin sequences amplified from *G. m. morsitans*. They noted that tsetse defensin was 457 nucleotides in length and coded for an 87 amino acid prepropeptide. This protein consisted of a 49 nucleotide noncoding region, preceding a 19 amino acid hydrophobic signalling region finishing at Ala19 (Hao *et al.*, 2001). Furthermore, cleaving of the protein at Lys34 was found to produce the mature defensin peptide. This mature peptide contained the six conserved cysteine residues and was found to have an isoelectric point of  $\sim 8.3$ , suggesting that the protein exhibited the cationic properties previously documented in insect defensins (Bulet *et al.*, 1999; Hao *et al.*, 2001).

Under neutral conditions (i.e. when no pathogen is present), defensin expression is minimal within the tsetse midgut; however, following infection and transformation of metacyclic to procyclic trypomastigotes expression increases dramatically (Hao *et al.*, 2001; Hu and Aksoy, 2006). Furthermore, variation in expression between infected and non-infected flies indicates that continued defensin expression is reliant upon the continued binding of trypanosome related PAMPs (Hao *et al.*, 2001). The expression of defensins has been reported in *An. gambiae* in response to infection of both gram-negative bacteria and

*Plasmodium* spp. (Dimopoulos *et al.*, 1997; Richman *et al.*, 1997; Christophides *et al.*, 2002). Crucially, defensin was reported within the mosquitoes' midgut prior to establishment of an infection and externally following penetration of the midgut epithelium, indicating the expression of defensin in response to parasitic infection and migration. It should be noted, that Dimopoulos *et al.* (1997) state that defensin primarily targets the latter stages of infection rather than those that penetrate the midgut epithelium. This supports the observations of Hao *et al.* (2001) and Hu and Aksoy (2006), who both observed that defensin expression was elevated following the maturation of procyclic trypomastigotes within the midgut, rather than targeting the proventricular forms penetrating the midgut epithelium.

### 2.1.1: Aims and Objectives

An analysis of attacins in *Glossina* genome by Wang *et al.* (2008) illustrated the fundamental characteristics of attacin genes within the *G. m. morsitans* genome. Wang *et al.* (2008) documented the presence of the three attacin orthologues, *AttA*, *AttB* and *AttD*, arranged into the three-cluster organisation within the genome, and determined the variation of characteristic pro- and pre-peptide domains between *G. m. morsitans* and other dipteran species. While Trappeniers *et al.* (2019) recently identified and annotated the four attacin genes within the *G. m. morsitans* genome (Table 2.1), comprehensive studies across the *Glossina* genus are lacking, as is any published structural analysis of dipteran attacin protein. Furthermore, there appears to be no published literature detailing the defensins within any of the *Glossina* spp. genomes.

The aim of this chapter is to address this paucity by identifying and characterising the attacin clusters and the defensin genes within the available *Glossina* spp. genomes (Objective 1, see section 1.6). The primary aim of this chapter is to identify the attacin clusters and defensin genes within the available *Glossina* genome assemblies. Identification of *Glossina* attacin will be achieved using the attacin genes annotated by Trappeniers *et al.* (2019) (Table 2.1). Amino acid sequences for each of the attacin paralogues will be used as a template to identify orthologues within *Glossina* spp. using a tBLASTn method as detailed below. While no defensin genes have been annotated within any *Glossina* genomes, a *Glossina* defensin sequences is available in the NCBI data base and can be used to identify

defensin sequences within the *Glossina* genome assemblies using simple tBLASTn mining methods.

The in depth study of immune genes within *Drosophila* spp. showed that AMPs are often encoded by gene families, with multiple genes encoding members of the same gene family (Khush and Lemaitre, 2000). This trend was also observed within the *Glossina* attacin gene family, and supported by the characterisation of the attacin cluster (Wang *et al.*, 2008). This attacin gene cluster (Wang *et al.*, 2008), presents a clear foundation for future genomic research to establish the loci of attacin genes within the *Glossina* genomes, thereby enabling the assessment of genomic variation between attacin genes within the *Glossina* genus. In turn, genomic variation could be indicative of the evolutionary divergence between *Glossina* spp. and must be considered alongside nucleotide and amino acid variation.

Since no defensin genes have been previously annotated within any of the *Glossina* genomes, the identification of these genes within the genomes is necessary prior to any extensive genomic analysis. The characterisation of tsetse defensin by Hao *et al.* (2001) provides a solid foundation for further analysis, however, the defensin gene within other *Glossina* spp. remains undocumented and full identification is required in order to assess interspecies genomic and protein variation.

Our secondary aim is to assess interspecies genomic and protein variation within the attacin and defensin gene families. While interspecies nucleotide and amino acid variation within AMPs have been examined previously, data on AMP interspecies structural variation within a genus is lacking. Structural analysis could offer an alternative insight into the evolutionary history of these critical proteins.

## 2.2: Methodology

### 2.2.1: Identification of *G. morsitans* attacin and defensin genes

Trappeniers *et al.* (2019) identified four attacin genes within the available *G. m. morsitans* genomes in the VectorBase database (available at: [www.vectorbase.org](http://www.vectorbase.org)) (Giraldo-Calderón *et al.*, 2015), see Table 2.1, which were used as reference sequences while searching for attacin genes within the other *Glossina* spp. for this chapter.

Table 2.1: Attacin genes identified by Trappeniers *et al.* (2019), giving the genetic accession number and the attacin gene they code for.

Gene accession number	attacin gene
GMOY010521	<i>AttA</i>
GMOY010522	Partial <i>AttA</i>
GMOY010523	<i>AttB</i>
GMOY010524	<i>AttD</i>

While defensin has not been annotated within the *G. m. morsitans* genome, a sequence is available on NCBI (available at: [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)) (Accession number: AAL34112.1) as published by Hao *et al.* (2001) this was used to identify other defensin orthologous within the available *Glossina* genomes.

### 2.2.2: Identification of transcripts within *Glossina* genomes

Potential novel attacin and defensin transcripts were identified using the tBLASTn search method in VectorBase. The protein sequences from GMOY010521 (*AttA*), GMOY010523 (*AttB*), GMOY010524 (*AttD*) and AF368907.1 (Def) were submitted to searches across all *Glossina* spp. genomes (see Appendix 3 for information on genomes utilised). Results for potential transcripts were analysed using Pfam (available at: <https://pfam.xfam.org/>) (El-Gebali *et al.*, 2018) to assess for the presence of attacin terminal domains within the identified transcripts. Transcripts coding for either one of or both of the attacin N and C-terminals were then aligned to known attacin sequences, using the online MUSCLE (Multiple Sequence Comparison by Log-Expectation) sequence alignment software

(available at: [www.ebi.ac.uk/Tools/msa/muscle/](http://www.ebi.ac.uk/Tools/msa/muscle/)) (Madeira *et al.*, 2019), to predict the encoded attacin protein.

Analysis of identified defensin sequences was conducted in the same way, using Pfam to identify the C-terminal defensin domain (El-Gebali *et al.*, 2019).

### 2.2.3: Mapping of predicted genes and the identification of missing attacin genes

In order to establish a comprehensive understanding of attacin genes within the *Glossina* genus, the fundamental variation in the attacin clusters assembly and gene structure must be considered alongside nucleotide and amino acid variation. As such, linear maps illustrating the attacin cluster were constructed for each of the available *Glossina* genomes. By comparing the predictions to the structure documented by Wang *et al.* (2008), any variation between *Glossina* species or groups can be observed. Furthermore, if attacin paralogues are not identified, these maps can provide an outline of the region's most likely to containing missing attacin genes.

Missing attacin genes were identified using the same tBLASTn method as described above (section 2.2.2). Partial sequences identified within the predicted cluster region were aligned to known attacin sequences using MUSCLE software (Madeira *et al.*, 2019). Where strong alignment was observed, the amino acid sequences were submitted to a Pfam search for the characteristic attacin domains (El-Gebali *et al.*, 2018), and the location was added to the genome map produced above.

Defensin transcripts were mapped to the Scaffold in the same way. Maps were adapted from figures given by VectorBase and produced in Microsoft PowerPoint.

### 2.2.4: CLUSTALW gene alignments

In order to observe nucleotide and amino acid conservation within species and gene families, alignments of the nucleotide and amino acid sequences were conducted. An initial Pearson/FASTA alignment was conducted using MUSCLE (Multiple Sequence Comparison by Log-Expectation) (Madeira *et al.*, 2019), before the result was realigned using ExPASy Boxshade (available at: [http://embnet.vital-it.ch/software/BOX\\_form.html](http://embnet.vital-it.ch/software/BOX_form.html) (now unavailable)). This produced a CLUSTALW alignment using RTF\_new shading. For alignments containing both nucleotide and amino acid sequences; nucleotide alignment



was conducted as described above, while amino acid alignments were produced using MUSCLE to produce a CLUSTALW alignment which was aligned manually (on a codon-by-codon basis) to the Boxshade results.

#### 2.2.5: Phylogenetic analysis

Phylogenetic analysis was conducted to assess the relationship between the identified genes and the attacin gene families. The evolutionary history of all predicted attacin and defensin genes was investigated using MEGAX (Kumar *et al.*, 2018). Pearson/FASTA alignments of the amino acid sequences were created using MUSCLE (Madeira *et al.*, 2019). Neighbour-Joining trees (Saitou and Nei, 1987) were constructed using the Poisson correction method (Zuckermandl and Pauling, 1965), while Maximum-likelihood trees were constructed using the most appropriate model for each data set, as predicted using the MEGAX model comparison software (Kumar *et al.*, 2018). In this case, the Whelan And Goldman (WAG) model (Whelan and Goldman, 2001) was used for the attacin sequences while the Dayhoff model (Schwarz and Dayhoff, 1979) was used for defensin. Both protein alignments also utilised Discrete Gamma distribution (+G) in the construction of the Maximum Likelihood phylogenetic trees.

All trees were run using 1,000 bootstrap replicates (Felsenstein, 1985), while missing data was set to partial deletion with a cut off of 50% for attacin and 95% for defensin. Amino acid sequences from *D. melanogaster* (*AttA*, *AttB*, *AttD* and *Def*), *S. calcitrans* (*AttA* and *Def*) and *M. domestica* (*AttA* and *Def*) were used as outgroups.

#### 2.2.6: Inter-species nucleotide variation analysis

Nucleotide diversity and conservation within each gene family was illustrated using sliding window analysis. This produced a visual representation of the regions of highest diversity and conservation within the CDS. Nucleotide variation ( $\pi$ ) analysis was calculated using the 'DNA polymorphism' function in DnaSP (version 6) (Rozas *et al.*, 2017), this analysis was conducted separately over the full alignments of *AttA*, *AttB*, *AttD* and *Def*. Sliding window analysis was also conducted using the 'DNA polymorphism' function in DnaSP (version 6) (Rozas, *et al.*, 2017) to illustrate the presence of nucleotide variation throughout the CDS. Nucleotide variation was calculated using  $\pi$  on a codon specific scale (window size = 3; step size = 3). The raw data was then extracted from DnaSP and exported to excel to produce a

linear representation of nucleotide variation within the genes. It should be noted that only *Glossina* spp. were used and defensin sequences were edited so that only the defensin coding regions were included within the analysis. Nucleotide variation ( $\pi$ ) was estimated using equation defined by Nei (1987) and Nei and Miller (1990) (equation 1, appendix 2).

#### 2.2.7: Pairwise distance Principle Component Analysis

Pairwise distance (*P*-distance) was calculated to assess the proportion of amino acid differences between two protein sequences. This provided an inside into the degree of variation between the predicted proteins and protein families. The pairwise distance between sequences was calculated in MEGAX (Kumar *et al.*, 2018). All sites with less than 50% coverage were eliminated from the attacin analysis, while all sites with less than 95% coverage were removed from the *Def* analysis. A matrix was then constructed and a Principle Component Analysis (PCA) run in PAST3 (Hammer *et al.*, 2001). *P*-distance was calculated using equation 2 (Appendix 2).

#### 2.2.8: Three-dimensional protein modelling

Prediction of the protein tertiary structure was undertaken to assess for variation within the protein families. Amino acid sequences were submitted to the I-TASSER online software (available at: <https://zhanggroup.org/I-TASSER/>) (Yang and Zhang, 2015; Yang, *et al.*, 2015), using standard conditions to predict the secondary and tertiary structure of each of the identified genes. The C-score and TM value of each model indicated the reliability of each predicted structure, structures presenting a higher reliability (highest C-scores and a TM value > 0.5) were selected for analysis. Only full predicted genes were used in this analysis as partial sequences could illustrate false variation due to missing domains. All models were visualised and aligned using PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre).

While visual comparison of the structures was conducted using I-TASSER, more detailed analysis of the three-dimensional (3D) protein structures was conducted using the DALI online server (available at: <http://ekhidna2.biocenter.helsinki.fi/dali/>) (Holm, 2020). Protein database (PDB) files were uploaded to the server and an All-vs-All analysis undertaken. This produced a heatmap and distance matrices used to produce a second PCA in the same manner as described above (see section 2.2.7).

## 2.3: Results

## 2.4: Attacins

A total of 28 predicted attacin genes were identified within the six available *Glossina* genomes (including those previously identified by Trappeniers *et al.* (2019)) (see Supplementary Table 3, Appendix 3) Five predicted attacin genes were identified in all species except *G. palpalis gambiensis* and *G. brevipalpis*, where four predicted genes were identified. Of these 28 genes, 13 were found to be partial or incomplete transcripts which exhibited fundamental characteristics of attacin genes though did not code for a complete attacin protein.

### 2.4.1: Attacin gene cluster identification and variation

#### 2.4.1i: *Glossina morsitans morsitans*

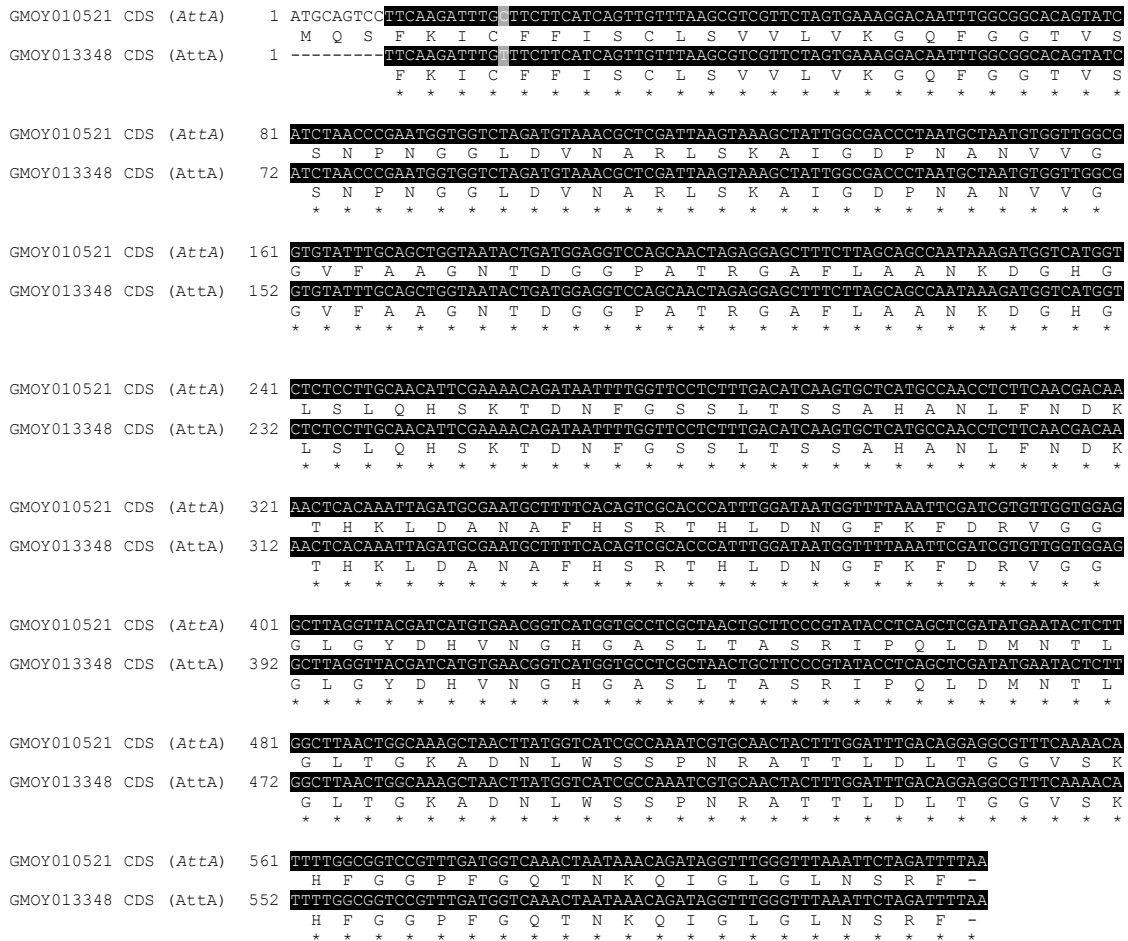
A model of the *G. m. morsitans* attacin cluster has been produced previously by both Wang *et al.* (2008) and Trappeniers *et al.* (2019) (Fig. 2.2). Figure 2.4 (below) shows the location of each previously identified attacin genes within SuperContig scf7180000652149 and the predicted fifth attacin gene (GMOY013348).



Figure 2.4: A linear representation of previously identified attacin genes in the scf7180000652149 SuperContig, drawn in Microsoft PowerPoint and adapted from the VectorBase models. Sequenced contigs. are represented by the blue, with gaps in the SuperContig are shown by the white spaces. Each gene is shown by the brown area; coding regions are denoted by the filled areas with non-coding regions shown in white. Newly predicted genes are shown in green.

BLAST results identified the missing *AttA* gene, *AttA12*, which is coded for by GMOY013348, located at the start of contig. CCAG010006534. This was further supported by an alignment of the CDS sequences of GMOY010521 and GMOY013348 (Fig. 2.5). This has an almost identical nucleotide alignment with a single substitution being observed at C21 to T21. However, this substitution is synonymous with both resultant codons, 'TGC' and 'TGT', coding for cysteine. As such, despite the slight nucleotide variation, the amino acid

sequence is identical in both GMOY010521 and GMOY013348. This observation is supported further by the results of a Pfam protein domain search which illustrated that both sequences coded for the same protein domains, the attacin N- and C-terminal domains, with similar significant E-values (Fig. 2.5).



Sequence	Pfam domains	e-value	Domain structure
GMOY010521	attacin N-Terminal	5.8e <sup>-21</sup>	
	attacin C-Terminal	4.8e <sup>-44</sup>	
GMOY013348	attacin N-Terminal	6.0e <sup>-21</sup>	
	attacin C-Terminal	5.0e <sup>-44</sup>	

Figure 2.5: Shows a ClustalW alignment of the CDS sequences of GMOY010521 and GMOY013348, showing both the DNA alignment and codon translation constructed using ExPASy Boxshade. The degree of similarity of the aligned nucleotides is indicated by the degree of shading, black = direct match, grey = synonymous mutation and white = non-synonymous mutations or missing data. Protein similarity is denoted using standard alignment methodology, \* = complete conservation. The intron is highlighted in yellow. A table is so present showing the proteins domains present in both GMOY010521 and GMOY013348. The E-value for each identified domain is also given, as is a linear diagram of the domain structure. The domain images were generated by Pfam (El-Gebali et al., 2018), the attacin N-terminal is represented by the red oval while the C-terminal is represented by the green oval.

Cluster 2 encompasses GMOY010522, a partial *AttA*, and GMOY010523, *AttB* (Trappeniers *et al.*, 2019), however, neither appears to be fully characterised. GMOY010522 codes for a partial *AttA* gene which can be confirmed by aligning GMOY010522 to GMOY01021 (Fig. 2.6A). This alignment shows a complete match with the C-terminal of both genes while the N-terminal is missing. This is supported, again, by Pfam results that illustrate a partial attacin C-terminal domain in GMOY010522 (Fig. 2.6A). It is likely, however, that the majority of the N-terminal and 5' sequence of the C-terminal domains are located within the 302 base pair gap between contigs CCAG010006534 and CCAG010006535. This is supported strongly by the reverse sequence of contig CCAG010006534 prior to the start of GMOY010522 exon 2 which, when added to the 5' of exon 2, extends the alignment by a further 39 residues (Fig. 2.6A).

GMOY010523 contains a much larger non-coding region than expected whilst coding only for a partial attacin C-terminal domain. An alignment of the protein sequences of GMOY010521 (*AttA*) and the translated GMOY010523 (*AttB*) cDNA sequence shows a much closer alignment than the current VectorBase CDS sequence (Figure 2.6B). This translation also explains the large non-coding region observed on the VectorBase gene map, as there is no start codon prior to Met156 (Figure 2.6B). Interestingly, when translated using reading frame 1 rather than reading frame 2, the "MQSFKIC" motif can be identified. However, this is the only aspect of the attacin gene present in that reading frame. This suggests further that there has been a mistake during annotation or sequencing of contig. CCAG010006535, as previously mentioned.

A)


```

GMOY010521 1 MQSFKICFFISCLSVVLVKQFGGTVSSNPNGGLDVNARLSKAIGDPNANVVGGVFAAGNTDGGPATRGAFLAANKDGHGSLQHSKTDN
GMOY010522 1 -----MASRKRVTIRIIVRTLELV-----

GMOY010521 91 FGSSLTSSAHANLFNDKTHKLDANAFHSRTHLDNGFKFDRVGGGLGYDHNHGHGASLTASRIPQLDMNTLGLTGKANLWSSPNRATTLDL
GMOY010522 21 -----SRTHLDNGFKFDRVGGGLGYDHNHGHGASLTASRIPQLDMNTLGLTGKANLWSSPNRATTLDL

GMOY010521 181 TGGVSKHFGGPFDDGQTNKQIGLGLNSRF
GMOY010522 48 TGGVSKHFGGPFDDGQTNKQIGLGLNSRF

```

Sequence	Pfam domains	E-value	Domain structure
GMOY010522	attacin C-Terminal	6.6e <sup>-16</sup>	

B)

```

GMOY010521 1 -----MQSFKICFFISCLSVVLVKQFGGTVSSNPNGGLDVNARLSKAIGDPNANVVGGVFAAGNTDGGPATR
GMOY010523 1 NRRTFQNSIGNCLKK*QKIYQHAVLQDLFFISCLSVVLVKQFGGTVSSNPNGGLDVNARLSKAIGDPNANVVGGVFAAGNTDGGPATR

GMOY010521 69 GAFLAANKDGHGSLQHSKTDNFGSSLTSSAHANLFNDKTHKLDANAFHSRTHLDNGFKFDRVGGGLGYDHNHGHGASLTASRIPQLDMN
GMOY010523 90 GAFLAANKDGHGSLQHSKTDNFGSSLTSSAHANLFNDKTHKLDANAFHSRTHLDNGFKFDRVGGGLGYDHNHGHGASLTASRIPQLDMN

GMOY010521 159 TLGLTGKANLWSSPNRATTLDLTGRVSKHFGGPFDDGQTNKQIGLGLNSRF-----
GMOY010523 180 TLGLTGKANLWSSPNRATTLDLTGRVSKHFGGPFDDGQTNKQIGLGLNSRF*ALYFELNVFRLEWKIKRKGK

```



Sequence	Pfam domains	E-value	Domain structure
GMOY010523 (Vectorbase)	attacin C-Terminal	2.3e <sup>-13</sup>	
GMOY010523 (translation)	attacin N-Terminal	5.9e <sup>-21</sup>	 <b>Attacin_C</b>
	attacin C-Terminal	1.2e <sup>-42</sup>	

Figure 2.6: A) ClustalW protein alignment of GMOY01021 (*AttA*) and GMOY010522 (partial *AttA*), with the additional 39 residues added from CCAG010006534 (highlighted in gold). It is clear the second exon of GMOY010522 has an identical match with the C-terminal of GMOY010521, while the first exon shows little alignment to the other *AttA* gene. A table is also given showing the results of a Pfam search, partial attacin C-terminal domains was detected, cover the aligned C-terminal within the alignment. B) The protein alignment of GMOY01021 (*AttA*) and GMOY010523 (*AttB*) following translation of the full cDNA sequence, produced by ExPASy Boxshade. This shows a 95% identify match between the two genes, while clearly illustrating that GMOY010523 codes for *AttB* with the N187 and R195 visible. A table is also given comparing the results of a Pfam search of GMOY010523 protein sequence from VectorBase and the translated cDNA sequence. The attacin N-terminal is represented by the red oval while the C-terminal is represented by the green oval. Conservation within the aligned amino acids is indicated by the degree of shading, black = complete conservation, dark grey = residues with a Gonnet PAM 250 score > 0.5, light grey = residues with residues with a Gonnet PAM 250 score < 0.5 and white = no similarity. \* indicates a stop codon.

Cluster 3 contains a single *AttD* gene identified as GMOY010524 by Trappeniers *et al.* (2019). Given the greater degree of variation between *AttD* and the other attacin genes within the *Glossina* genomes observed Wang *et al.* (2008), it is likely that Figure 2.7 illustrates the correct identification of GMOY010524 as *AttD*.

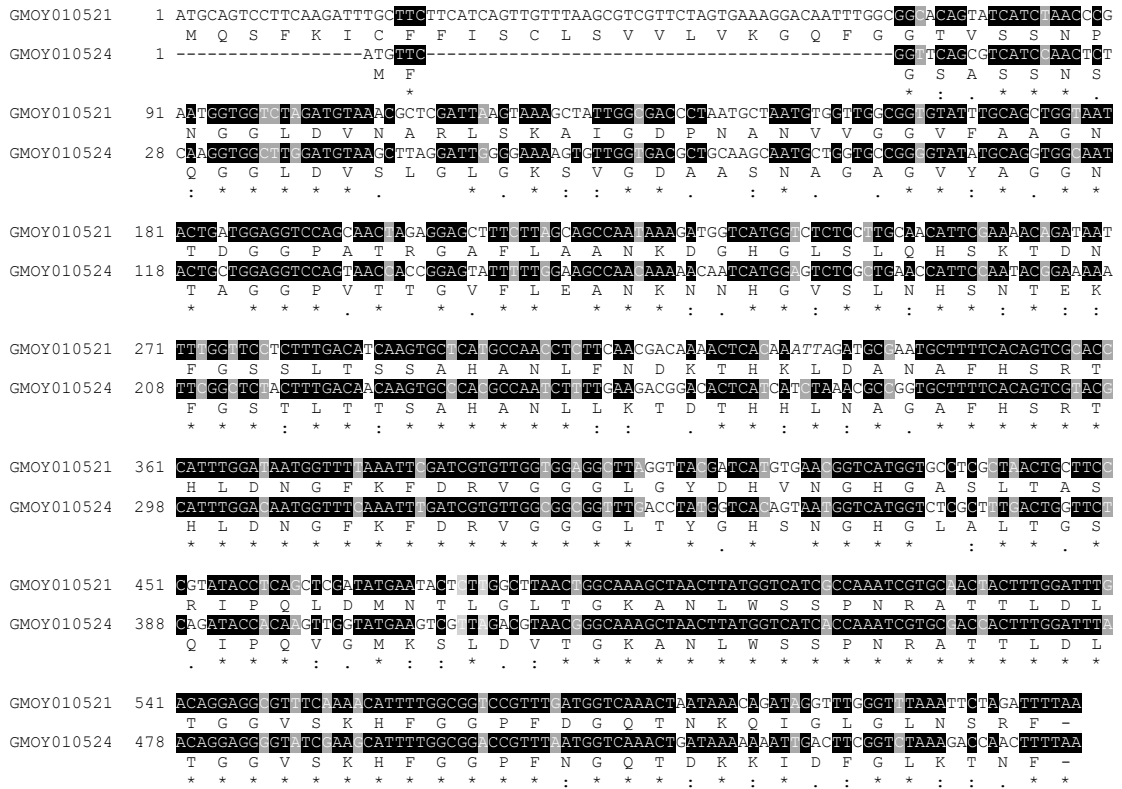


Figure 2.7: The nucleotide and protein alignment of GMOY010521 (*AttA*) and GMOY010524 (*AttD*) produced by ExPASy Boxshade Conservation within the nucleotide sequence is indicated by the degree of shading, Black = complete conservation, Grey = synonymous mutation and White = non-synonymous mutations or missing data. Protein similarity is denoted using standard alignment methodology, \* = complete conservation, : = residues with a Gonnet PAM 250 score > 0.5, . = residues with residues with Gonnet PAM 250 score < 0.5 and a gap = no similarity. A hyphen (-) in the amino acid sequence indicates a stop codon or missing data.

### 2.4.1ii: *Glossina austeni*

Of the five Attacin genes within the *G. austeni* genome, one (GAUT047992) had been identified previously as *AttA*. GAUT047990 was predicted to code for the second *AttA* gene in cluster 1, while GAUT047991 was predicted to encode *AttD*. The *AttA* and *AttB* genes of cluster 2, were predicted to be coded by GAUT048001 and GAUT048006, respectively. A model of the Attacin cluster within the *G. austeni* genome was constructed in Figure 2.8. This illustrates the presence of each Attacin gene cluster within the *G. austeni* genome though in reverse order to that observed in *G. m. morsitans*.

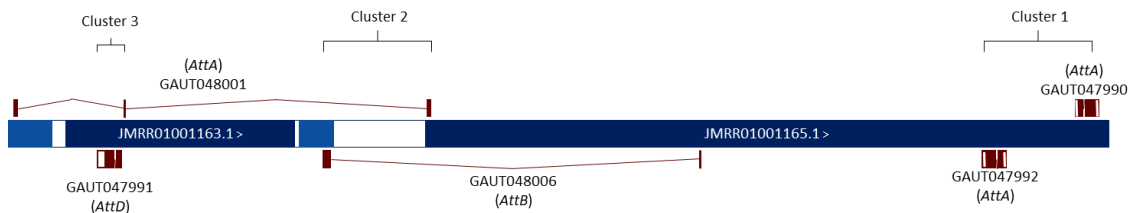


Figure 2.8: A linear representation of the predicted Attacin cluster within *G. austeni* Scaffold7, drawn in Microsoft PowerPoint and adapted from the VectorBase models. Sequenced contigs. are represented by the blue, with gaps in the scaffold are shown by the white spaces. Each gene is shown by the brown area, coding regions are denoted by the filled areas with non-coding regions shown in white.

GAUT047990 was predicted to code for *AttA* following nucleotide and amino acid alignments. This illustrated a 98.56 % nucleotide alignment and a 99.04 % amino acid alignment. Just one amino acid substitution occurred within the Attacin N-terminal domain identified by Pfam (Fig. 2.9). Both amino acid substitutions (V16L and A43T) showed Gonnet PAM 250 score > 0.5 indicating similar properties between the two amino acids.

GAUT047992	1	ATGCAGTCCTCAAGATTGCTTCTTCATCAGTTGTTAAGCGTCG	TTCTCGTAAAAGGACAATTGGCGGCACAGTATCATCTAACCCG		
GAUT047990	1	ATGCAGTCCTCAAGATTGCTTCTTCATCAGTTGTTAAGCGTCG	TTCTCGTAAAAGGACAATTGGCGGCACAGTATCATCTAACCCG		
		M Q S F K I C F F I S C L S V V L V K G Q F G G T V S S N P	M Q S F K I C F F I S C L S V L L V K G Q F G G T V S S N P		
		* *	* *		
GAUT047992	91	AATGGTGGTCTAGATGT	AATGCTCGATTAAGTAAA	CTATTGGCGA	CCTAATGCTAATGTGGTGGCGGTGATTTGCAGCTGGTAA
GAUT047990	91	AATGGTGGTCTAGATGT	AATGCTCGATTAAGTAAA	CTATTGGCGA	CCTAATGCTAATGTGGTGGCGGTGATTTGCAGCTGGTAA
		N G G L D V N A R L S K T I G D P N A N V V G G V F A A G N	N G G L D V N A R L S K T I G D P N A N V V G G V F A A G N		
		* *	* *		
GAUT047992	181	ACTGCTGG	GGTCCAGCAACTAGAGGAGCTTCTTAGCAGCCAATAAAGATGGTCATGGTCTCTCCTTGCA	CATTCGAAAACAGATAAT	
GAUT047990	181	ACTGCTGG	GGTCCAGCAACTAGAGGAGCTTCTTAGCAGCCAATAAAGATGGTCATGGTCTCTCCTTGCA	CATTCGAAAACAGATAAT	
		T A G G P A T R G A F L A A N K D G H G L S L Q H S K T D N	T A G G P A T R G A F L A A N K D G H G L S L Q H S K T D N		
		* *	* *		
GAUT047992	271	TTTGG	TCCTCTTTGACATCAAGTGCATGCCAACCTCTTCAACGACAAAACCTCACAAATTAGATGCGAATGCTTTTCCACTGCGACC		
GAUT047990	271	TTTGG	TCCTCTTTGACATCAAGTGCATGCCAACCTCTTCAACGACAAAACCTCACAAATTAGATGCGAATGCTTTTCCACTGCGACC		
		F G S S L T S S A H A N L F N D K T H K L D A N A F H T R T	F G S S L T S S A H A N L F N D K T H K L D A N A F H T R T		
		* *	* *		
GAUT047992	361	CATTGGATAATGGTTTAAATTCGATCGTGTGGTGGAGG	TTAGGTTACGATCATGC	AGCGGTCATGGTGCCTCGCTAACTGCTTCC	
GAUT047990	361	CATTGGATAATGGTTTAAATTCGATCGTGTGGTGGAGG	TTAGGTTACGATCATGC	AGCGGTCATGGTGCCTCGCTAACTGCTTCC	
		H L D N G F K F D R V G G G L G Y D H A S G H G A S L T A S	H L D N G F K F D R V G G G L G Y D H A S G H G A S L T A S		
		* *	* *		



```

* * * * *
GAUT047992 451 CGTATACCTCAGCTCGATATGAACACTCTGGCTTAACTGGAAAAGCTAACTTATGGTCATCGCCAAATCGTGCCACCACCTTTGGATTG
R I P Q L D M N T L G L T G K A N L W S S P N R A T T L D L
GAUT047990 451 CGTATACCTCAGCTCGATATGAACACTCTGGCTTAACTGGAAAAGCTAACTTATGGTCATCGCCAAATCGTGCCACCACCTTTGGATTG
R I P Q L D M N T L G L T G K A N L W S S P N R A T T L D L
* * * * *
GAUT047992 541 ACAGGAGGAGTTTCAAAACATTTGGCGGTCCGTTTGATGGTCAAACTAATAAACAGATTGGCTTGGGTTTAAATTCAGATTCTAA
T G G V S K H F G G P F D G Q T N K Q I G L G L N S R F -
GAUT047990 541 ACAGGAGGAGTTTCAAAACATTTGGCGGTCCGTTTGATGGTCAAACTAATAAACAGATTGGCTTGGGTTTAAATTCAGATTCTAA
T G G V S K H F G G P F D G Q T N K Q I G L G L N S R F -
* * * * *

```


Sequence	Pfam domains	E-value	Domain structure
GAUT047990	Attacin N-Terminal	1.3e <sup>-20</sup>	
	Attacin C-Terminal	9.1e <sup>-44</sup>	

Figure 2.9: Shows a ClustalW alignment of the GAUT047992 and GAUT047990 CDS showing both the DNA alignment and codon translation constructed using ExPASy Boxshade. The degree of similarity of the aligned nucleotides is indicated by the degree of shading, Black = direct match, grey = synonymous mutation and white = non-synonymous mutations or missing data. Protein similarity is denoted using standard alignment methodology, \* = complete conservation, : = residues with a Gonnet PAM 250 score > 0.5. A table is so present showing the proteins domains present in GAUT047990. The E-value for each identified domain is also given, as is a linear diagram of the domain structure. The domain images were generated by Pfam (El-Gebali *et al.*, 2018), the Attacin N-terminal is represented by the red oval while the C-terminal is represented by the green oval.

Cluster 2 was observed to contain two partial Attacin genes GAUT048001 and GAUT048006. As seen within the *G. m. morsitans* these genes contain a single intron that crosses a gap in the contig. and sequence for an Attacin C-terminal domain. This shows that the C-terminal domain of GAUT048001 shows an almost complete match to that of GAUT04992, with just two amino acid substitutions in the C-terminal sequence (Fig. 2.10A). The first being the start codon, Met67, and the second at Ser86. This conservation indicates that GAUT048001 encodes an *AttA* gene. The GAUT048006 alignment shows a longer C-terminal alignment; however, it contains five amino acid substations (Fig. 2.10B). Notably, one these substitutions include a Gln195Arg substitution observed between *G. m. morsitans AttA* and *AttB*, which suggests that GAUT048006 codes for the *AttB* gene within Cluster 2. Interestingly, both cluster 2 genes exhibit a N86S substitution despite encoding different Attacin genes, this could be a specific mutation within the Cluster 2 though any impact on functionality and structure is yet to be determined.

The absence of N-terminal domains in both of these genes is likely explained by a gap between contigs JMRR01001164 and JMRR01001165. This gap in the scaffold is approximately 3,500 nucleotides in length, indicating both missing N-terminals could be encoded within the missing sequences.

A)

```

GAUT047992 1 MQSFKICFFISCLSVVLVKQFGGTVSSNPNGLDVARLSKATGDPANVVGVAAGNTAGGEPATRGAFLAANKDGHGSLQHSKTDN
GAUT048001 1 -----MRSRANAWNLDDRRQLLTSTIEVNYNANLWTVVTRTV-----

GAUT047992 91 FGSSLTSSAHANLFNDRTHKLDANAFHTRTHLDNGFKFDRVGGGLGYDHASGHGASLTASRIPOLDMNTLGLTGKANLWSSPNRATTLDD
GAUT048001 43 ----LSTICRNTFVSOVVHDIPLKALOS-----M LDMNTLGLTGKANLWSSPSRATTLDD

GAUT047992 180 LTGGVSKHFGGPPFDGQTNKQIGLGLNSRF
GAUT048001 94 LTGGVSKHFGGPPFDGQTNKQIGLGLNSRF
  
```

B)

```

GAUT047992 1 MQSFKICFFISCLSVVLVKQFGGTVSSNPNGLDVARLSKATGDPANVVGVAAGNTAGGEPATRGAFLAANKDGHGSLQHSKTDN
GAUT048006 1 MRSR-----GRKTFVLRNFVGH-----

GAUT047992 91 FGSSLTSSAHANLFNDRTHKLDANAFHTRTHLDNGFKFDRVGGGLGYDHASGHGASLTASRIPOLDMNTLGLTGKANLWSSPNRATTLDD
GAUT048006 20 ----HANAFHTRTHLDNGFKFDRVGGGLGYDHASGHGASLTASRIPOLDMNTLGLTGKANLWSSPSRATTLDD

GAUT047992 181 LTGGVSKHFGGPPFDGQTNKQIGLGLNSRF
GAUT048006 89 LTGGVSKHFGGPPDGRINKQIGLGLNSRF
  
```



Sequence	Pfam domains	E-value	Domain structure
GAUT048001	Attacin C-Terminal	4.3e <sup>-15</sup>	
GAUT048006	Attacin C-Terminal	7.1e <sup>-35</sup>	

Figure 2.10: A) ClustalW protein alignment of GAUT047992 (*AttA*) and GAUT048001 (partial *AttA*), produced by ExpASY Boxshade. It is clear the second exon of GAUT048001 illustrates a high conservation between the C-terminal of GAUT047992, while the first exon shows little alignment to the other *AttA* gene. B) The protein alignment of GAUT047992 (*AttA*) and GAUT048006 (partial *AttB*), produced by ExpASY Boxshade. This shows increased variation between the two C-terminal sequences, while clearly indicating that GAUT048006 codes for *AttB* with the R195 substitution being exhibited. Conservation within the aligned amino acids is indicated by the degree of shading, Black = complete conservation, dark grey = residues with a Gonnet PAM 250 score > 0.5, light grey = residues with residues with Gonnet PAM 250 score < 0.5 and white = no similarity. \* indicates a stop codon. A table is also given illustrating the results of a Pfam search of GAUT048001 and GAUT048006 protein sequence from VectorBase. Partial Attacin C-terminal domains were detected in both genes, the E-value of each result is given and a linear diagram illustrating Attacin C-terminal. The green oval shows the presence of an Attacin C-terminal.

The identification of GAUT047991 as *AttD* was undertaken by aligning GAUT047991 (*AttA*) with GAUT047992. This alignment illustrated a far greater degree of variation between GAUT047992 and other predicted Attacin genes, this variation can be observed in Figure 2.11 below. This indicated that GAUT047991 coded for *AttD*, this was supported by the phylogeny seen in Figure. 2.29, where GAUT047991 can be observed to in the *AttD* clade.

```

GAUT047992 1 ATGCAGTCCTTCAAGATTGGTTTCATCAGTTGTTTAAAGCGTCGTTCTCGTAAAGGACAATTGGCGGACAGTATCATCATAAACC
GAUT047991 1 -----ATGTC-----GGTCACCGTCATCATAATCG
          M Q S F K I C F F I S C L S V V L V K G Q F G G T V S S N P
          * : . * * * .

GAUT047992 91 AATCGTGGCTTGCATGTAAAGCTCCAGTAAAGCTAATGGGAGCCTAATGCTAATCTGCTGCGGCTATTTGCAGCTGGTAAT
GAUT047991 28 AAAGGTGGCTTGCATGTAAATTAGGATTGCGAAAGCTTTGGGAGCTGCAAGCAATCTGCTGCGGCTATTTGCAGCTGGTAAT
          K G G L D V N L G L G K S V G D A A S N A G A G V Y A A G N
          : * * * * * * : * : * * : * * *

GAUT047992 181 ACTGCTGGGGCCAGCAACAGAGGAGCTTTTTCAGCCAAATAAAGTTGCTCATGGCTCTCCTGCAACATTCATAAAGCATTAAT
GAUT047991 118 ACTGCTGGGGCCAGCAACGCCGAGTATTTTCAGCCAAATAAAGTTGCTCATGGCTCTCCTGCAACTCATTCATAAAGCATTAAT
          T A G G P A T R G A F L A A N K D G H G L S L Q H S K T D N
          T A G G P A T A G V F L E A N K N N H G V S L T H S N T E K
  
```



### 2.4.1iii: *Glossina pallidipes*

As with *G. austeni*, one predicted Attacin gene, GPAI040754, had been identified previously as *AttA*. However, there is evidence to suggest that GPAI040754 codes for *AttB* (see below). tBLASTn results identified three further predicted Attacin genes, GPAI040769, GPAI040759 and GPAI040752, in addition to a partial gene fragment. The locality of these genes suggests that the GPAI040754 and GPAI040759 comprise Cluster 2, while GPAI040752 encode the *AttD* gene of Cluster 3. GPAI040769 and the partial gene fragment were predicted to make up Cluster 1 (Figure. 2.12).



Figure 2.12: A linear representation of the predicted Attacin cluster within *G. pallidipes* Scaffold62, drawn in Microsoft PowerPoint and adapted from the VectorBase models. Sequenced contigs. are represented by the blue, with gaps in the scaffold are shown by the white spaces. Each gene is shown by the brown area, coding regions are denoted by the filled areas with non-coding regions shown in white. Newly predicted genes are shown in green.

While GPAI040754 has been annotated as *AttA* within the *G. pallidipes* genome on VectorBase, an alignment to GPAI040759 and its location in the Attacin cluster suggest it encodes *AttB*. The location of GPAI040754 matches that of GMOY010523 (predicted to encode *G. m. morsitans AttB*), while an alignment to GPAI040759 indicated the presence of the Q195R substitution indicative of *AttB* (Figure. 2.13). Three other amino acid substitutions were also identified between the two genes: R121H, N173S and L202F.

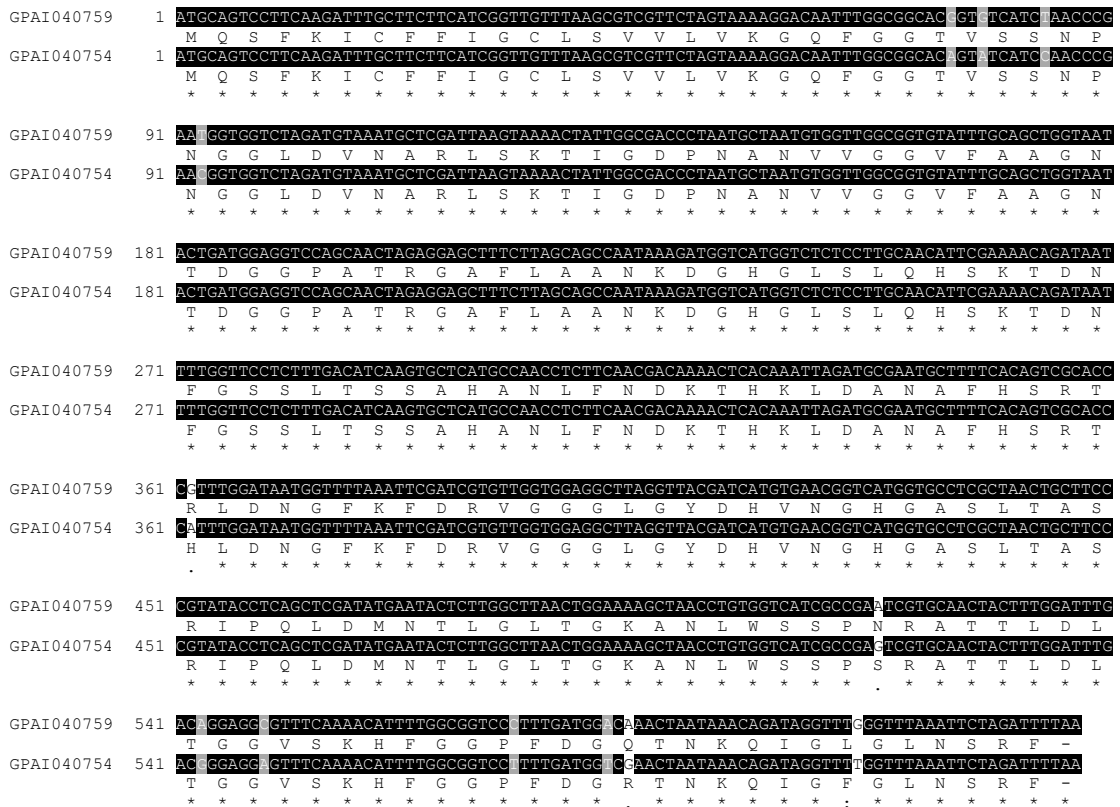


Figure 2.13: The alignment of GPAI040759 (*AttA*) and GPAI040754 (*AttB*) produced by ExPASy Boxshade. Conservation within the nucleotide sequence is indicated by the degree of shading, Black = complete conservation, Grey = synonymous mutation and White = non-synonymous mutations. Protein similarity is denoted using standard alignment methodology, \* = complete conservation, . = residues with a Gonnet PAM 250 score > 0.5, - = residues with residues with Gonnet PAM 250 score < 0.5 and a gap = no similarity. A hyphen (-) in the amino acid sequence indicates a stop codon or missing data.

Cluster 1 can be seen to span the gap between contigs. JMRR01001000 and JMRR01001001. GPAI040769 was identified to encode an Attacin C-terminal domain following a Pfam search, while an alignment to GPAI040759 indicates a high conservation between the C-terminal of the two genes (Fig. 2.14A). As with several genes predicted in other species, GPAI040769 contains an intron crossing the gap in the Scaffold sequence. It is likely therefore, that the N-terminal of GPAI040769 is encoded in this missing sequence. A second Attacin N-terminal was also identified at the start of contig. JMRR01001001. Both the nucleotide and amino acid sequences of this protein domain show complete conservation between the new identified Attacin sequences and GPAI040759 (FIG. 2.14B), indicating that an *AttA* gene is present. However, like GPAI040769, the N-terminal is found within the missing sequence data.



phylogenetic analysis (Fig. 2.29) and an alignment between GPAI040759 (*AttA*) and GPAI040752. This alignment illustrated a large degree of variation between the two genes, as previously observed between *G. m. morsitans* and *G. austeni*, *AttA* and *AttD* genes (Fig. 2.15). This prediction is supported further by the shorter CDS length of GPAI040752, which has also been observed in other *AttD* genes.

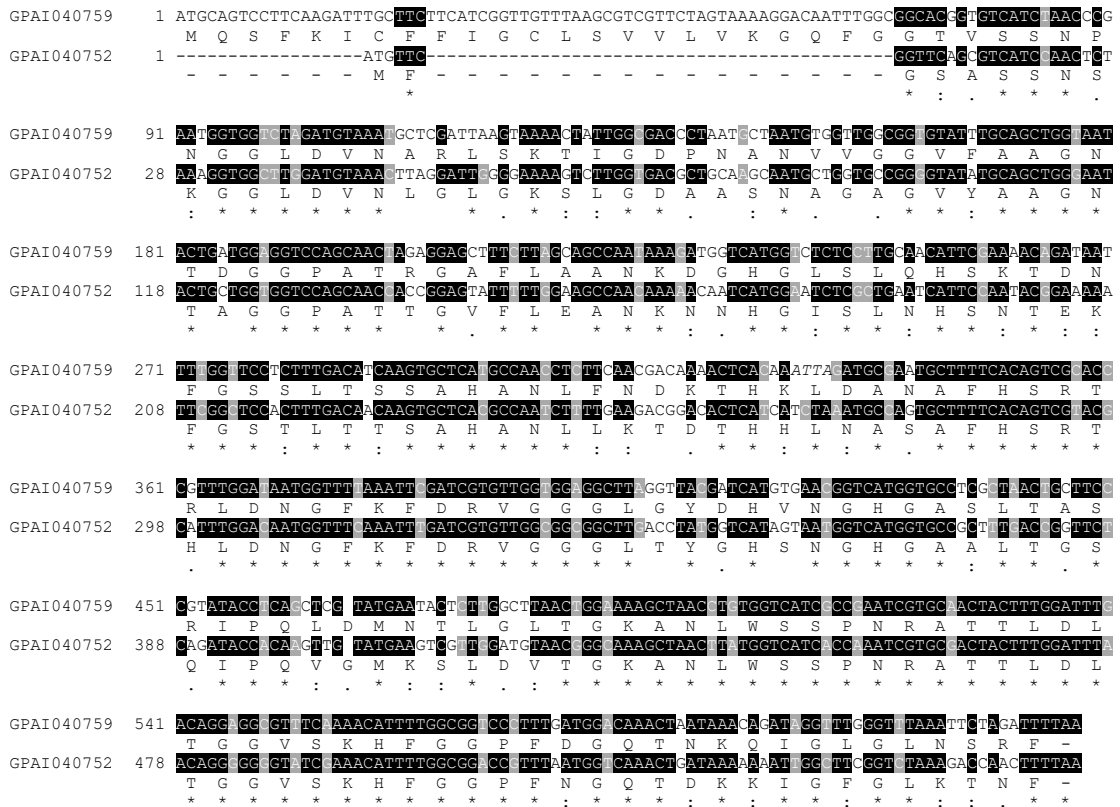


Figure 2.15: The nucleotide and amino acid alignment of GPAI040759 (*AttA*) and GPAI040752 (*AttD*) produced by ExPASy Boxshade. Conservation within the nucleotide sequence is indicated by the degree of shading, Black = complete conservation, Grey = synonymous mutation and White = non-synonymous mutations. Protein similarity is denoted using standard alignment methodology, \* = complete conservation, : = residues with a Gonnet PAM 250 score > 0.5, . = residues with residues with Gonnet PAM 250 score < 0.5 and a gap = no similarity. A hyphen (-) in the amino acid sequence indicates a stop codon or missing data.

#### 2.4.1iv: *Glossina fuscipes fuscipes*

The Attacin cluster within *G. f. fuscipes* was less complete than those observed in other *Glossina* species. GFUI014658 has been previously annotated as *AttA* on VectorBase; an additional three genes, GFUI014660, GFUI014661 and GFUI014668, were identified following a tBLASTn search of the genome. Two partial sequences were identified coding for a section of an Attacin gene, the first on the reverse strand of JFJR01000137, and the second on the forward strand crossing the gap between contigs. JFJR01000138 and JFJR01000139 (Fig. 2.16). The location of the three identified genes suggests that GFUI014661 and GFUI014668 form a cluster, though once again this cluster spans a gap in the Scaffold sequence (Fig. 2.16). While GFUI014660 forms an individual cluster, suggesting the presence of *AttD* (Fig. 2.16). One sequence fragment is found in a head-to-head orientation with GFUI014658, suggesting the presence of the third Attacin cluster (Fig. 2.16).



Figure 2.16: A linear representation of the predicted Attacin cluster within *G. f. fuscipes* Scaffold1, drawn in Microsoft PowerPoint and adapted from the VectorBase models. Sequenced contigs are represented by the blue, with gaps in the scaffold are shown by the white spaces. Each gene is shown by the brown area, coding regions are denoted by the filled areas with non-coding regions shown in white. Introns are shown by thin lines; dotted introns indicated the following exons are absent from the Figure. Newly predicted genes are shown in green.

As noted above GFUI014661 and GFUI014668 form a cluster across the gap between contigs. JFJR01000136 and JFJR01000137, with both genes coding for an Attacin C-terminal domain. An alignment of each gene with the previously identified GFUI014658 illustrates two things. Firstly, that both GFUI014661 and GFUI014668 code for *AttA*, and secondly, that GFUI014658 likely codes for *AttB* rather than *AttA* as annotated (Fig. 2.17). These observations were made apparent as the amino acid alignment of both genes contain the amino acid substitutions within the C-terminal domains. This indicates that both GFUI014661 and GFUI014668 contain the same *AttA* C-terminal sequences within their second exon, while varying from that of GFUI014658. As such it is likely that this cluster



forms Attacin Cluster 1 (containing two *AttA* genes) despite being located in the middle of the Attacin cluster.

### A)

```

GFUI014658 1 MQSFKICFFTSCLSVVVLVKGQFGGTVSSNPNGLDVTARLSKTIGDPNANLVGGVFASGNTDGGPATRGAFLCANFDGHGLSLQHSKTDN
GFUI014668 1 -----MFRMAG-----GKATE-----LGRDC-----

GFUI014658 91 FGSSLTSSAHANLFNDKTHKLDANAFHSRTHLDNGFKFDRVGGGIGYHVNHGASLTASRI PQLDMNTLGLTGKANLWSSPSRATTLDL
GFUI014668 17 -----IILNRTHLDNGFKFDRVGGGIVYHVNHGASLTASRI PQLDMNTLGLTGKANLWSSPSRATTLDL

GFUI014658 181 TGGVSKHFGGPPDGGQTNKQIGLGLNSRF
GFUI014668 83 TGGVSKHFGGPPDGGQTNKQIGLGLNSRF

```

### B)

```

GFUI014658 1 MQSFKICFFTSCLSVVVLVKGQFGGTVSSNPNGLDVTARLSKTIGDPNANLVGGVFASGNTDGGPATRGAFLCANFDGHGLSLQHSKTDN
GFUI014661 1 -----MFRMAG-----GKATE-----LGRDC-----

GFUI014658 91 FGSSLTSSAHANLFNDKTHKLDANAFHSRTHLDNGFKFDRVGGGIGYHVNHGASLTASRI PQLDMNTLGLTGKANLWSSPSRATTLDL
GFUI014661 1 -----MFRMAG-----GKATE-----LGRDC-----

GFUI014658 181 TGGVSKHFGGPPDGGQTNKQIGLGLNSRF
GFUI014661 69 TGGVSKHFGGPPDGGQTNKQIGLGLNSRF

```

Sequence	Pfam domains	E-value	Domain structure
GFUI014668	Attacin C-Terminal	6.9e <sup>-29</sup>	
GFUI014661	Attacin C-Terminal	5.5e <sup>-29</sup>	

Figure 2.17: A) ClustalW protein alignment of GFUI014658 (*AttB*) and GFUI014668 (partial *AttA*), produced by ExPASy Boxshade. The second exon of GAUT048001, commencing from R21, illustrates high conservation between the C-terminal of GFUI014658 with just four sites of variation, while the first exon shows little alignment to the other Attacin gene. B) The protein alignment of GFUI014658 (*AttB*) and GFUI014661 (partial *AttA*), produced by ExPASy Boxshade. This an almost identical alignment to that of GFUI014668. With high conservation in the second exon, with the same four sites of variation. Conservation within the aligned amino acids is indicated by the degree of shading, black = complete conservation, dark grey = residues with a Gonnet PAM 250 score > 0.5, light grey = residues with residues with a Gonnet PAM 250 score < 0.5 and white = no similarity. \* indicates a stop codon. A table is also given illustrating the results of a Pfam search of GFUI014668 and GFUI014661 protein sequence from VectorBase. Partial Attacin C-terminal domains were detected in both genes, the E-value of each result is given and a linear diagram illustrating Attacin C-terminal. The green oval shows the presence of an Attacin C-terminal.

The position of Cluster 2 within the Scaffold suggests that GFIO014658 codes for *AttB* (Fig. 2.18 and 2.19). However, there is no clear indication of the third *AttA* gene found in this cluster. One possible location for this gene is across the contig. gap as highlighted in Figure. 2.16. An alignment of GFIO014658 and the sequence fragments identified in this location (Figure. 2.18) shows a clear alignment of the N and C-terminals of the Attacin sequence. However, there is no way of identifying the sequence as *AttA* or *AttB*. Furthermore, the gap between the contigs. is 543 nucleotides in length, while 508 nucleotides are missing from the CDS of the identified Attacin gene, inferring an intron length of only 35 nucleotides.



```

GFUI014658 1 ATGCAATCCTTCAAGATTTGTTTCTTCATCAGTTGCTTAAGCGTCGTTCTAGTAAAAGGACAATTTGGCGGCACAGTATCATCCAACCCG
Scaffold1 1 -----
GFUI014658 91 AACGGTGGTCTAGATGTAAGTCTCGATTAAAGTAAAACATTGGCGACCCTAATGCCAATCTGGTTGGCGGTGATTTGCATCTGGTAAT
Scaffold 1 -----
GFUI014658 181 ACTGATGGAGGTCCAGCAACTAGAGGAGCTTCTTGGGAGCCAATAAAGATGGTCATGGTCTCTCCTTGCAACATTCGAAAACAGATAAT
Scaffold1 1 -----
GFUI014658 271 TTTGGTTCCTCTTTGACATCAAGTGCATGCAACCTCTTCAACGACAAAACACAAATAGATGCGAATGCTTTTTCACAGTCGCACC
Scaffold1 1 -----
GFUI014658 361 CATTGGATAATGGTTTTAAATTCGATCGTGTGGTGGAGGCTTAGGTTACGATCATGTGAACGGTCATGGCGCCTCGCTAACTGCTTCC
Scaffold1 1 -----
GFUI014658 451 CGTATACCTCAGCTCGATATGAATACTCTTGGCTTAACTGGTAAAGCTAACTTATGCTCATCGCCGAGTCGGCCCAACCACCTTTGGATTG
JFJR01000137 1 -----GTCATCGCCGAGTCGGCCCAACCACCTTTGGATTG
S S P S R A T T L D L
S S P S R P T T L D L
* * * * * . * * * * *
GFUI014658 541 AC GGA AGC GTTTC AAAA CATT TCG CCG C CCGTTTGATGGTCAAAC TAATAAACAGATTGGGTTGGGTTTAAATTC TAGATTTTAA
T G G V S K H F G G P F D G Q T N K Q I G L G L N S R F -
JFJR01000137 35 AC GGA AAT TGG C AAAA ATG TATG C C CCGTTTGATGGTCAAAC TAATAAACAGATTGGGTTGGGTTTAAATTC TAGATTTTAA
T G I L R K M Y V S P F D G Q T N K Q I G L G L N S R F -
* * : * : . * * * * * * * * * * * * * * * * * * * * * *

```

Figure 2.19: The nucleotide and amino acid alignment of GFUI0104658 (*AttB*) and the reverse strand of contig. JFJR01000137 produced by ExpASY Boxshade. Conservation within the nucleotide sequence is indicated by the degree of shading, Black = complete conservation, Grey = synonymous mutation and White = non-synonymous mutations. Protein similarity is denoted using standard alignment methodology, \* = complete conservation, : = residues with a Gonnet PAM 250 score > 0.5, . = residues with residues with a Gonnet PAM 250 score < 0.5 and a gap = no similarity. A hyphen (-) in the amino acid sequence indicates a stop codon or missing data.

The identification of *AttD* within the *G. f. fuscipes* Attacin cluster was undertaken by aligning GMOY010524, *G. m. morsitans AttD*, to the sequence of GFUI01460 (Fig. 2.20). This illustrated a much greater degree of conservation between the two genes than was observed between GFUI01458 and GFUI01460. Given the conservation between *AttD* genes, it can be concluded that GFUI01460 codes for *AttD* rather than the missing *AttA* gene. Unusually, however, GFUI01460 has been annotated as Sorbitol Dehydrogenase within the *G. f. fuscipes* genome, with Attacin N- and C-terminals being coded alongside a short chain dehydrogenase domain.

```

GMOY010524 1 MFGSASSNSGGLDVSLCLGKSVGDASNAAGAVVAGGNTAGGPMVTGFLEANNNHGVSLNHSNTEKGGSTLTSSAHANLLKTDTHHL
GFUI014660 1 MFGSASSNSGGLDVNLCCGKSVGDASNWCAGITAAAGNTLGGPATTGFLEANNNHGVSLNHSNTEKGGSTLTSSAHANLLKTDTHHL

GMOY010524 91 NACAFHSRTHLDNGFKFDRVGGGLTYGHSNGHCLALTGSQIPOGMRSLDVTGKANLWSSPNRATTLDTGGVSKHFGGPFNG---QIDR
GFUI014660 91 NAFHSRTHLDNGFKFDRVGGGLTCSHNGHGAALTGSQIPOGMRSLDVTGKANLWSSPNRATTLDTGGVSKHFGGPFNGPFYRRTN

GMOY010524 178 KID-----
GFUI014660 181 RIDGKVVIVTGCNTGIGKETALELARRGARLYMACRDAARCEAARLEIERTQNPVFNRTLDLASLSSVRQFAERFLAEEEDRLDILINN

GMOY010524 181 -----FG-----
GFUI014660 271 AGVMATPRKLTVDGFPEQQLGINHLGHFLLTNLLLDRLKKSAPSRIVVSSAAYMFRINKNDLNSKRYWPFEGAYAQSKLANILFTRKL

GMOY010524 183 -----LKTTF-----
GFUI014660 361 AELLKDTSVTANCLHPGIVRTELMRYNCLSPKTFHLITRSPKAGAQTTLYLALDPKFDLTSGGYYEYMLRLLPLLPWARKETANLW

GMOY010524 -----
GFUI014660 451 EESEKMVGLKHNDDESKLTNITTVQNQKTYGADMNDRPVHFPEESEQIKAEELN

```

Sequence	Pfam domains	E-value	Domain structure
GFUI014660	Attacin N-Terminal	6.7e <sup>-12</sup>	
	Attacin C-Terminal	4.9e <sup>-35</sup>	
	Short Chain	7.2e <sup>-34</sup>	
	Dehydrogenase		

Figure 2.20: A ClustalW protein alignment of *G. m. morsitans* GMOY010524 (*AttD*) and GFUI014660 (*AttD*), produced by ExPASy Boxshade. This shows a high level of conservation between the two *AttD* genes spanning the first two exons of GFUI014660. Conservation within the aligned amino acids is indicated by the degree of shading, Black = complete conservation, dark grey = residues with a Gonnet PAM 250 score > 0.5, light grey = residues with residues with a Gonnet PAM 250 score < 0.5 and white = no similarity. \* indicates a stop codon. A table is also given illustrating the results of a Pfam search of GFUI014660 protein sequence from VectorBase. An Attacin N and C-terminal domain was detected in the genes (Purple and green ovals respectively), while a short chain dehydrogenase domain was also identified (seen in red). The E-value of each result is given and a linear diagram illustrating the Attacin C-terminal.

### 2.4.1v: *Glossina palpalis gambiensis*

Unlike for the previous species no Attacin genes had been identified previously within the *G. palpalis gambiensis* genome on VectorBase. A tBLASTn search identified two annotated genes and two sequences within the genome as predicted Attacin genes (Fig. 2.21). GPPI020339 is the only full Attacin gene identified, while GPPI020332 codes for a partial Attacin C-terminal domain. An almost complete Attacin gene was identified within the non-coding region of GPPU020339, while two sequences coding for the N and C-terminals of an Attacin gene were identified on the reverse strand downstream of the other sequences.

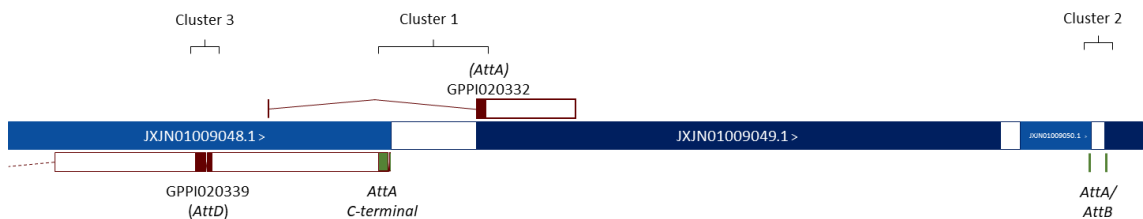


Figure 2.21: A linear representation of the predicted Attacin cluster within *G. palpalis gambiensis* Scaffold114, drawn in Microsoft PowerPoint and adapted from the VectorBase models. Sequenced contigs. are represented by the blue, with gaps in the scaffold are shown by the white spaces. Each gene is shown by the brown area, coding regions are denoted by the filled areas with non-coding regions shown in white. Introns are shown by thin lines, dotted introns indicated the following exons are absent from the Figure. Newly predicted genes are shown in Green.

As with the *G. f. fuscipes* Attacin cluster, there is no clear indication of the three clusters within the *G. palpalis gambiensis* genome. An alignment of GPPI020332 and GMOY010521 (*G. m. morsitans* AttA), shows a high level of conservation between the two genes, with just three amino acid substitution (Fig. 2.22A). This suggests that GPPI020332 codes for an AttA protein though the N-terminal is, once again, missing from the sequences and likely found in the gap between contigs. JXJN01009048 and JXJN01009049. A second AttA sequence was identified in a head-to head orientation with GPPI020332, in the non-coding region of GPPI020339. When aligned to GMOY010521 this showed a high conservation between the genes, with just two amino acid substitutions (Fig. 2.22B). Notably, both GPPI020332 and the newly identified sequence contain the same N118S substitution.

**A)**

```

GMOY010521 1 MQSFKICFFISCLSVVLVKQFGGTVSSNPNGLDVNARLSKAIGDPNANVVGVFAGNTDGGFATRGAFLANRFDGHGLSLOHSKTDN
GPPI020332 1 -----M-----R-----FDGHGLSLOHSKTDN

GMOY010521 91 FGSSLTSSAHANLFNDKTHKLDANAFHSRTHLDNGFKFDRVGGGLGYDHNHGHASLTASRIPLDMNTLGLTGKANLWSSFNRRATTLDL
GPPI020332 18 FGSSLTSSAHANLFNDKTHKLDANAFHSRTHLDNGFKFDRVGGGLGYDHNHGHASLTASRIPLDMNTLGLTGKANLWSSFNRRATTLDL

GMOY010521 181 TGGVSKHFGGPF DGQTNKQIGLGLNSRF
GPPI020332 108 TGGVSKHFGGPF DGQTNKQIGLGLNSRF
  
```

**B)**

```

GMOY010521 1 ATGCAGTCCTTCAAGATTGCTTCTTCATCAGTTGTTTAAAGCGTCGTTCTAGTGAAAGGACAATTTGGCGGCACAGTATCATCTAACCCG
GPPI020339-UT 1 M Q S F K I C F F I S C L S V V L V K G Q F G G T V S S N P

GMOY010521 91 AATGGTGGTCTAGATGTAACCGCTCGATTAAGTAAAGCTATTGGCGACCTAATGCTAATGGTTGGCGGTGTAFTTGCAGCTGGTAAT
GPPI020339-UT 1 N G G L D V N A R L S K A I G D P N A N V V G G V F A A G N
                                                                TTTGCAGCTGGTAAT
                                                                F A A G N
                                                                * * * * *

GMOY010521 181 ACTGAGGAGTCCAGCAACTAGAGGAGCTTCTTCCAGCCAATAAAGATGGTCATGGTCTCTCCTTGCACATTCGAAAACAGATAAT
GPPI020339-UT16 T D G G P A T R G A F L A A N K D G H G L S L Q H S K T D N
ACTGAGGAGTCCAGCAACTAGAGGAGCTTCTTCCAGCCAATAAAGATGGTCATGGTCTCTCCTTGCACATTCGAAAACAGATAAT
* * * * * P * * * * * . * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

GMOY010521 271 TTTGGTTCCTCTTTGACATCAAGTGCATGCCAACCTCTTCAACGACAAAACACAAAATAGATGCGAATGCTTTTCACAGTCGCACC
GPPI020339-UT106 G S S L T S S A H A N L F N D K T H K L D A N A F H S R T
TTTGGTTCCTCTTTGACATCAAGTGCATGCCAACCTCTTCAACGACAAAACACAAAATAGATGCGAATGCTTTTCACAGTCGCACC
F G S S L T S S A H A N L F N D K T H K L D A N A F H S R T
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

GMOY010521 361 CATTTGGATAATGGTTTAAATTCGATCGTGTGGTGGAGGCTTAGGTTACGATCATGTGAACGGTCATGGGCCTCGCTAACTGCTTCC
GPPI020339-UT196 H L D N G F K F D R V G G G L G Y D H V N G H G A S L T A S
CATTTGGATAATGGTTTAAATTCGATCGTGTGGTGGAGGCTTAGGTTACGATCATGTGAACGGTCATGGGCCTCGCTAACTGCTTCC
H L D N G F K F D R V G G G L G Y D H V N G H G A S L T A S
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

GMOY010521 451 CGTATACCTCAGCTCGATATGAATACTCTTGGCTTAACTGGAAAGCTAACCTATGGTCATCGCCATTCGGCAACACTTTGGATTTC
GPPI020339-UT286 R I P Q L D M N T L G L T G K A N L W S S P N R A T T L D L
CGTATACCTCAGCTCGATATGAATACTCTTGGCTTAACTGGAAAGCTAACCTATGGTCATCGCCATTCGGCAACACTTTGGATTTC
R I P Q L D M N T L G L T G K A N L W S S P S R A T T L D L
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

GMOY010521 541 ACGGAGCGTTTCAAACATTTGGCGGTCGTTGATGGTCAAACCTAATAACAGATGGTTGGGTTAAATCTAGATTTTAA
GPPI020339-UT376 T G G V S K H F G G P F D G Q T N K Q I G L G L N S R F -
ACGGAGCGTTTCAAACATTTGGCGGTCGTTGATGGTCAAACCTAATAACAGATGGTTGGGTTAAATCTAGATTTTAA
T G G I S K H F G G P F D G Q T N K Q I G L G L N S R F -
* * * : * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
  
```

Sequence	Pfam domains	E-value	Domain structure
GPPI020332	Attacin C-Terminal	7.0e <sup>-45</sup>	
GPPI020339-UT	Attacin N-Terminal Attacin C-Terminal	6.5e <sup>-10</sup> 1.0e <sup>-44</sup>	

Figure 2.22: A) ClustalW protein alignment of GMOY010521 (*AttA*) and GPPI020332 (partial *AttA*), produced by ExPASy Boxshade. This illustrates the conservation between the two genes indicating that GPPI020332 codes for *AttA*. B) The protein and nucleotide alignment of GMOY010521 (*AttA*) and part of the GPPI020339 untranslated region (UT), produced by ExPASy Boxshade. Illustrates the high conservation between the identified *AttA* gene and the newly identified sequence. Conservation within the aligned amino acids is indicated by the degree of shading, Black = complete conservation, dark grey = residues with a Gonnet PAM 250 score > 0.5, light grey = residues with residues with a Gonnet PAM 250 score < 0.5 and white = no similarity. \* indicates a stop codon. A table is also given illustrating the results of a Pfam search of GPPI020332 and the untranslated region of GPPI020339 from VectorBase. Attacin C-terminal domains were detected in both genes, with an N-terminal also being identified in GPPI020339. The E-value of each result is given and a linear diagram illustrating the Attacin C-terminal. The green oval shows the presence of an Attacin C-terminal.

The only complete predicted Attacin gene, GPPI020339, appears to code for *AttD*. This was a result of the alignment of GPPI020339 and GMOY010524 (*G. m. morsitans AttD*), which showed a much greater degree on conservation between the two genes (Fig. 2.23), and a similar level of conservation to that observed between *G. f. fuscipes* and *G. m. morsitans AttD* genes. Furthermore, the location of *AttD* (GPPI020339) within the *G. palpalis gambiensis* Attacin cluster mirrors that observed in the *G. f. fuscipes* Attacin cluster supporting the conservation of genes within the Palpalis phylogenetic group (Fig. 2.16 and 2.21).

```

GMOY010524 1 MFCSSSSNSGGLDVSLCLGKSVGDAASNAAGAGVACGNTAGGPVITGPLEANKNNHGVSLSHSNTEKGSLLTTSAHANLLKTDTHHL
GPPI020339 1 MFCSSSSNSGGLDVSLCLGKSVGDAASNAAGAGVACGNTAGGPVITGPLEANKNNHGVSLSHSNTEKGSLLTTSAHANLLKTDTHHL

GMOY010524 91 NAAAFHSRTHLDNGFKFDRVGGGLTYGHSNGHGLALITGSQIPQGMKSLDTGKANLWSSPNRATLDTGGVSKHFGGPFNGQTDKKID
GPPI020339 91 NAAAFHSRTHLDNGFKFDRVGGGLTCSHNGHGLALITGSQIPQGMKSLDTGKANLWSSPNRATLDTGGVSKHFGGPFNGQTDKKIC

GMOY010524 181 FGLTTF*
GPPI020339 181 FGLTTF*

```

Figure 2.23: A ClustalW protein alignment of *G. m. morsitans* GMOY010524 (*AttD*) and GPPI020339 (*AttD*), produced by ExPASy Boxshade. This shows a high level of conservation between the two *AttD* genes. Conservation within the aligned amino acids is indicated by the degree of shading, Black = complete conservation, dark grey = residues with a Gonnet PAM 250 score > 0.5, light grey = residues with residues with a Gonnet PAM 250 score < 0.5 and white = no similarity. \* indicates a stop codon

There was little evidence of Cluster 2 within the genome. A single Attacin gene was identified spanning the gap between contigs. JXJN0109050 and JXJN0109051, when aligned to GMOY010521 both the N and C-terminals show a high conservation with just two amino acid substitutions in the C-terminal (Fig. 2.24). However, there is insufficient data to identify this gene as *AttA* or *AttB*. Furthermore, there is no evidence to indicate the presence of a second Attacin gene in this cluster, or even a fifth gene within the cluster.

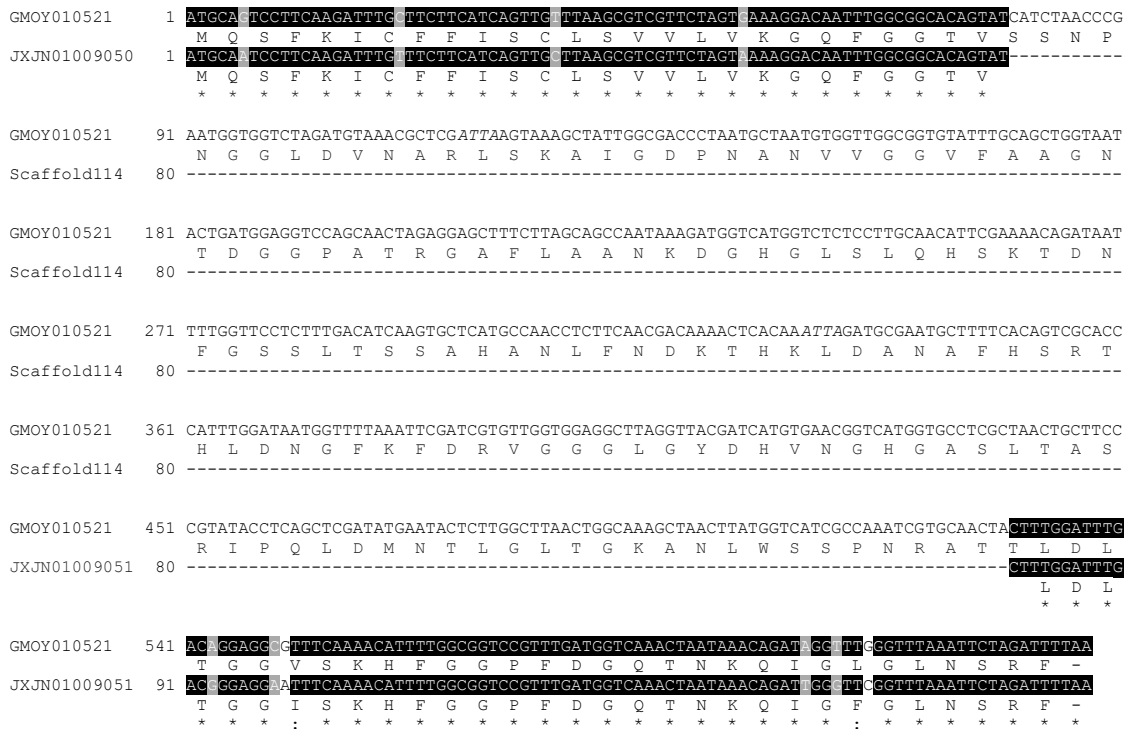


Figure 2.24: The nucleotide and amino acid alignment of GMOY010521 (*AttA*) and contigs. JXJN01009050 and JXJN01009051 produced by ExPASy Boxshade. Conservation within the nucleotide sequence is indicated by the degree of shading, Black = complete conservation, Grey = synonymous mutation and White = non-synonymous mutations. Protein similarity is denoted using standard alignment methodology, \* = complete conservation, : = residues with a Gonnet PAM 250 score > 0.5, . = residues with a Gonnet PAM 250 score < 0.5 and a gap (-) in the amino acid sequence indicates a stop codon or missing data.



### 2.4.1vi: *Glossina brevipalpis*

The identification of Attacin genes within the *G. brevipalpis* genome was undertaken using a tBLASTn search. This revealed three predicted genes and two partial gene sequences, GBRI004567, GBRI00559 and GBRI004558 appear to code of Attacin genes while a partial sequence on the reverse strand of contig. JFJS01007041 and another on the forward strand of JFJS01007046, also appear to code for Attacin genes (Figure. 2.25).



Figure 2.25: A linear representation of the predicted Attacin cluster within *G. brevipalpis* Scaffold114, drawn in Microsoft PowerPoint and adapted from the VectorBase models. Sequenced contigs. are represented by the blue, with gaps in the scaffold are shown by the white spaces. Each gene is shown by the brown area, coding regions are denoted by the filled areas with non-coding regions shown in white. Introns are shown by thin lines. Newly predicted genes are shown in green.

Of the predicted genes, GBRI004567 codes for an Attacin C-terminal with seven amino acid substitutions between itself and GMOY010521 (Fig. 2.26A). While the N-terminal is missing this is likely found in the contig. gap downstream of the GBRI004567 CDS. There is no evidence of a second Attacin gene in the head-to-head orientation, while it is possible that this gene is located in the large sequence gap between contigs. JFJS01007048 and JFJS01007049 there is currently no data to confirm this. An alternative location is presented by the Attacin sequence identified in contig. JFJS01007046. An alignment of this sequence to GMOY010521 shows a greater degree of variation though still sufficiently similar to code for an Attacin C-terminal domain (Fig. 2.26B). However, unlike previously identified partial genes there is no evidence of N-terminal in the contig. sequence prior to the C-terminal (Fig. 2.25).

**A)**

GMOY010521 1 ATGCAGTCCTTCAAGATTGGCTTCTTCATCAGTTGTTAAAGCGTCGTTCTAGTCAAAGGACAATTTGGCGGCACAGTATCATCTAAACCCG  
 M Q S F K I C F F I S C L S V V L V K G Q F G G T V S S N P  
 GBRI004567 1 -----

GMOY010521 91 AATGGTGGTCTAGATGTAACGCTCGATTAAAGTAAAGCTATTGGCGACCCTAATGCTAATGTGGTGGCGGTGATTTGCAGCTGGTAAT  
 N G G L D V N A R L S K A I G D P N A N V V G G V F A A G N  
 GBRI004567 1 -----

GMOY010521 181 ACTGATGGAGGTCCAGCAACTAGAGGAGCTTCTTAGCAGCCAATAAAGATGGTCATGGTCTCTCCTTGCAACATTGCAAAACAGATAAT  
 T D G G P A T R G A F L A A N K D G H G L S L Q H S K T D N  
 GBRI004567 1 -----

GMOY010521 271 TTTGGTCTCTTTGACATCAAGTGCTCATGCCAACTCTCTTAAACGAACTCAACAATAGATGCGAATGCTTTTCACAGTCCGACC  
 F G S S L T S S A H A N L F N D K T H K L D A N A F H S R T  
 GBRI004567 1 -----GCCAACTCTCTTAAACGAACTCAACAATAGATGCGAATGCTTTTCACAGTCCGACC  
 A N L F N D Q T H K I D A N A F H S R T  
 \* \* \* \* \* : \* \* \* \* \* : \* \* \* \* \* \* \* \* \* \*

GMOY010521 361 CATTGGATAATGGTTTTAAATTCGATCGTTGGTGGAGGTTAGGTTACGTCATCCTCAACCGGTCAAGGTGCCTGTAACCTCTCC  
 H L D N G F K F D R V G G G L G Y D H V N G H G A S L T A S  
 GBRI004567 61 CATTGGATAATGGTTTTAAATTCGATCGTTGGTGGAGGTTAGGTTACGTCATCCTCAACCGGTCAAGGTGCCTGTAACCTCTCC  
 H L D N G F K F D R V G G G L G Y E H A R G H G A S L T G S  
 \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* : \* . \* \* \* \* \* \* \* \* \* \*

GMOY010521 451 CGTATCCTCACTGATGAATACCTGGCTTAACGGAAAGCTAATTATGGTCTCCAAAACCGTGCACACTTTGGATTTC  
 R I P Q L D M N T L G L T G K A N L W S S P N R A T T L D L  
 GBRI004567 151 CGTATCCTCACTGATGAATACCTGGCTTAACGGAAAGCTAATTATGGTCTCCAAAACCGTGCACACTTTGGATTTC  
 R I P Q L D M N T L G L T G K A N L W S S P N R A T T L D L  
 \*

GMOY010521 541 ACAGGAGCGTTTCAAACATTTGGGGTCCCTTATGGTCAAACTAATAAACTCATGGTTGGTTAAATTCAGATTTTAA  
 T G G V S K H F G G P F D G Q T N K Q I G L G L N S R F -  
 GBRI004567 241 ACAGGAGCGTTTCAAACATTTGGGGTCCCTTATGGTCAAACTAATAAACTCATGGTTGGTTAAATTCAGATTTTAA  
 T G G V S K H F G G P F N G Q T N K N I G L G L N S R F -  
 \* : \* \* \* \* \* \* \* \* \* \*

**B)**

GMOY010521 1 ATGCAGTCCTTCAAGATTGGCTTCTTCATCAGTTGTTAAAGCGTCGTTCTAGTCAAAGGACAATTTGGCGGCACAGTATCATCTAAACCCG  
 M Q S F K I C F F I S C L S V V L V K G Q F G G T V S S N P  
 JFJS01007046 1 -----

GMOY010521 91 AATGGTGGTCTAGATGTAACGCTCGATTAAAGTAAAGCTATTGGCGACCCTAATGCTAATGTGGTGGCGGTGATTTGCAGCTGGTAAT  
 N G G L D V N A R L S K A I G D P N A N V V G G V F A A G N  
 JFJS01007046 1 -----

GMOY010521 181 ACTGATGGAGGTCCAGCAACTAGAGGAGCTTCTTAGCAGCCAATAAAGATGGTCATGGTCTCTCCTTGCAACATTGCAAAACAGATAAT  
 T D G G P A T R G A F L A A N K D G H G L S L Q H S K T D N  
 JFJS01007046 1 -----

GMOY010521 271 TTTGGTCTCTTTGACATCAAGTGCTCATGCCAACCTCTTCAACGACAAAACCTCAACAATAGATGCGAATGCTTTTCACAGTCCGACC  
 F G S S L T S S A H A N L F N D K T H K L D A N A F H S R T  
 JFJS01007046 1 -----

GMOY010521 361 CATTGGATAATGGTTTTAAATTCGATCGTTGGTGGAGGCTTAGGTTACGATCATGTAACCGGTCAAGGTGCCTGCTACTGCTCTCC  
 H L D N G F K F D R V G G G L G Y D H V N G H G A S L T A S  
 JFJS01007046 1 -----ACTGCTCTCC  
 T G S  
 \* . \*

GMOY010521 451 CGTATACCTCACTCGATGAATACCTGGCTTAACGGAAAGCTAATTATGGTCTCCAAAACCGTGCACACTTTGGATTTC  
 R I P Q L D M N T L G L T G K A N L W S S P N R A T T L D L  
 JFJS01007046 10 CGTATACCTCACTCGATGAATACCTGGCTTAACGGAAAGCTAATTATGGTCTCCAAAACCGTGCACACTTTGGATTTC  
 R I P Q L G M N T F D L G K A N L W S S P N R A T T L D L  
 \* \* \* \* \* . \* \* \* : . \*

GMOY010521 541 ACAGGAGCGTTTCAAACATTTGGGGTCCCTTATGGTCAAACTAATAAACTCATGGTTGGTTAAATTCAGATTTTAA  
 T G G V S K H F G G P F D G Q T N K Q I G L G L N S R F -  
 JFJS01007046 100 ACAGGAGCGTTTCAAACATTTGGGGTCCCTTATGGTCAAACTAATAAACTCATGGTTGGTTAAATTCAGATTTTAA  
 T G G V S K H F G G P F D G Q T N K H I G L G L N S R F -  
 \* : \* \* \* \* \* \* \* \* \* \*



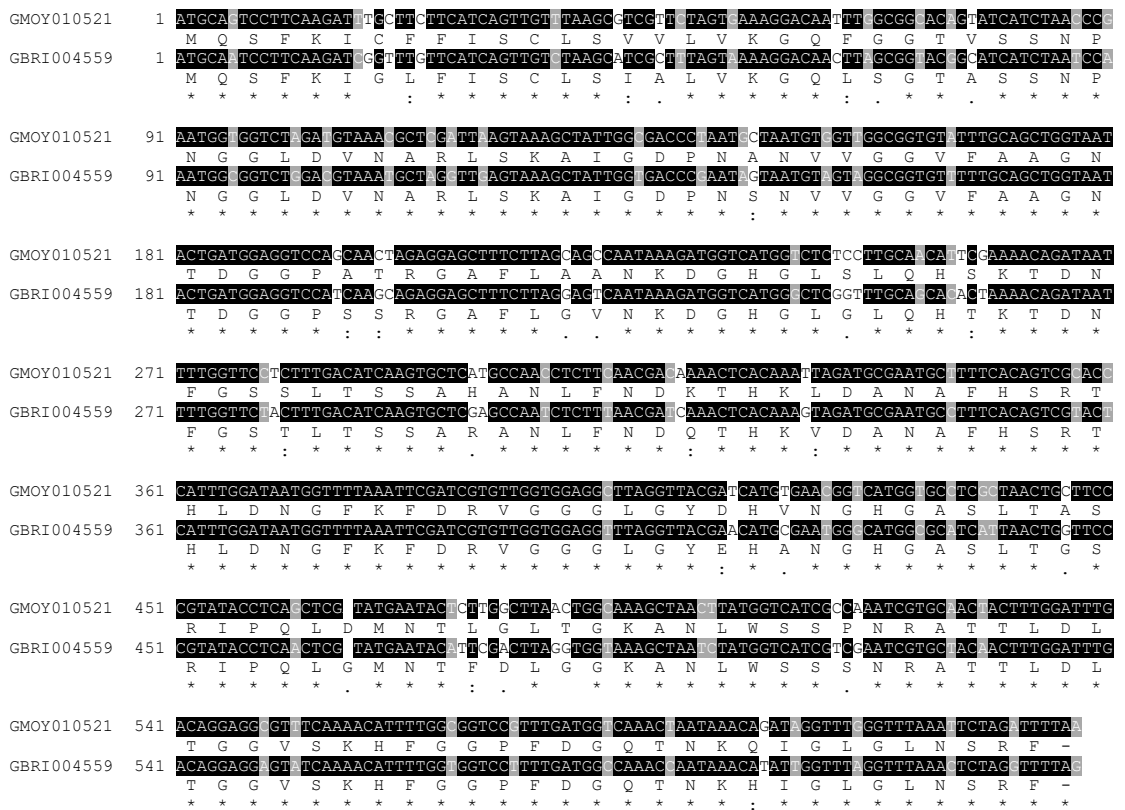
Sequence	Pfam domains	E-value	Domain structure
GBRI004567	Attacin C-Terminal	1.4e <sup>-40</sup>	
JFJS01007046	Attacin C-Terminal	2.5e <sup>-18</sup>	

Figure 2.26: A) ClustalW protein alignment of GMOY010521 (*AttA*) and GBRI004567 (partial *AttA*), produced by ExPASy Boxshade. This illustrates the conservation between the two genes C-terminals indicating that GBRI004567 codes for *AttA*. B) The protein and nucleotide alignment of GMOY010521 (*AttA*) and JFJS01007046, produced by ExPASy Boxshade. Illustrates the alignment between the identified Attacin C-terminal and the newly identified sequence. Conservation within the aligned amino acids is indicated by the degree of shading, Black = complete conservation, dark grey = residues with a Gonnet PAM 250 score > 0.5, light grey = residues with residues with a Gonnet PAM 250 score < 0.5 and white = no similarity. \* indicates a stop codon. A table is also given illustrating the results of a Pfam search of GBRI004567 and the sequence of JFJS01007046 from VectorBase. Attacin C-terminal domains were detected in both genes, with an N-terminal also being identified in GPPI020339. The E-value of each result is given and a linear diagram illustrating the Attacin C-terminal. The green oval shows the presence of an Attacin C-terminal.

A single full Attacin gene was identified within the reverse strand of contig. JFJS01007044, with a partial Attacin C-terminal sequence being identified on the reverse sequence of contig. JFJS01007041. The full gene, GBRI004559, has no previous annotation and, when aligned to GMOY010521, it shows an increased level of nucleotide and amino acid variation (Fig. 2.27A). The identified Attacin C-terminal sequence shows a higher level of conservation to GMOY010521, with six amino acid substitutions between the two sequences (Fig. 2.27B). Given the difference in the sequence conservation, GBRI004559 appears to code for *AttB* with a higher level of variation, while the high conservation between GMOY010521 and the identified Attacin C-terminal indicated the presence of *AttA*. Furthermore, it is likely that the missing N-terminal is located within the gap between contigs. JFJS01007041 and JFJS01007042 (Fig. 2.25).

**A)**



**B)**

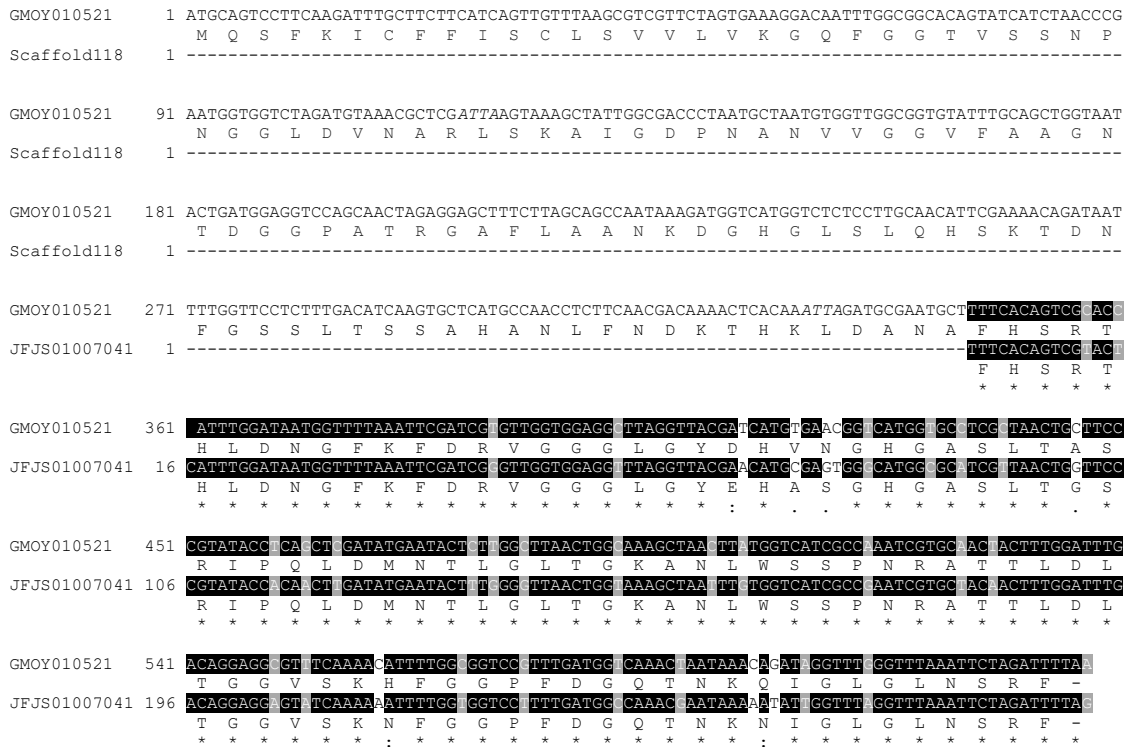


Figure 2.27: A) ClustalW nucleotide and protein alignment of GMOY010521 (*AttA*) and GBRI004559 (*AttB*), produced by ExPASy Boxshade. This illustrates the variation between the two genes indicating that GBRI004567 codes for *AttB*. B) The protein and nucleotide alignment of GMOY010521 (*AttA*) and JFJS01007041, produced by ExPASy Boxshade. Illustrates the alignment between the Attacin C-terminal and the newly identified sequence, the greater degree of conservation indicates that the partial gene sequence likely codes for an *AttA* gene. Conservation

within the aligned amino acids is indicated by the degree of shading, Black = complete conservation, dark grey = residues with a Gonnet PAM 250 score > 0.5, light grey = residues with residues with a Gonnet PAM 250 score < 0.5 and white = no similarity. \*indicates a stop codon.

The final gene identified by the tBLASTn search is GBRI004558. This gene shows a larger number of exons, three, compared to other Attacin genes and a longer CDS. The location of GBRI004558 within the attacin cluster suggests that it codes for *AttD* and this is supported by an alignment with GMOY010524 (Fig. 2.28). This alignment illustrated an increased level of conservation between the *G. brevipalpis* and *G. m. morsitans AttD* sequences than was observed between GBRI004558 and other attacin gene sequences further reinforcing the prediction that GBRI004558 codes for *AttD*.

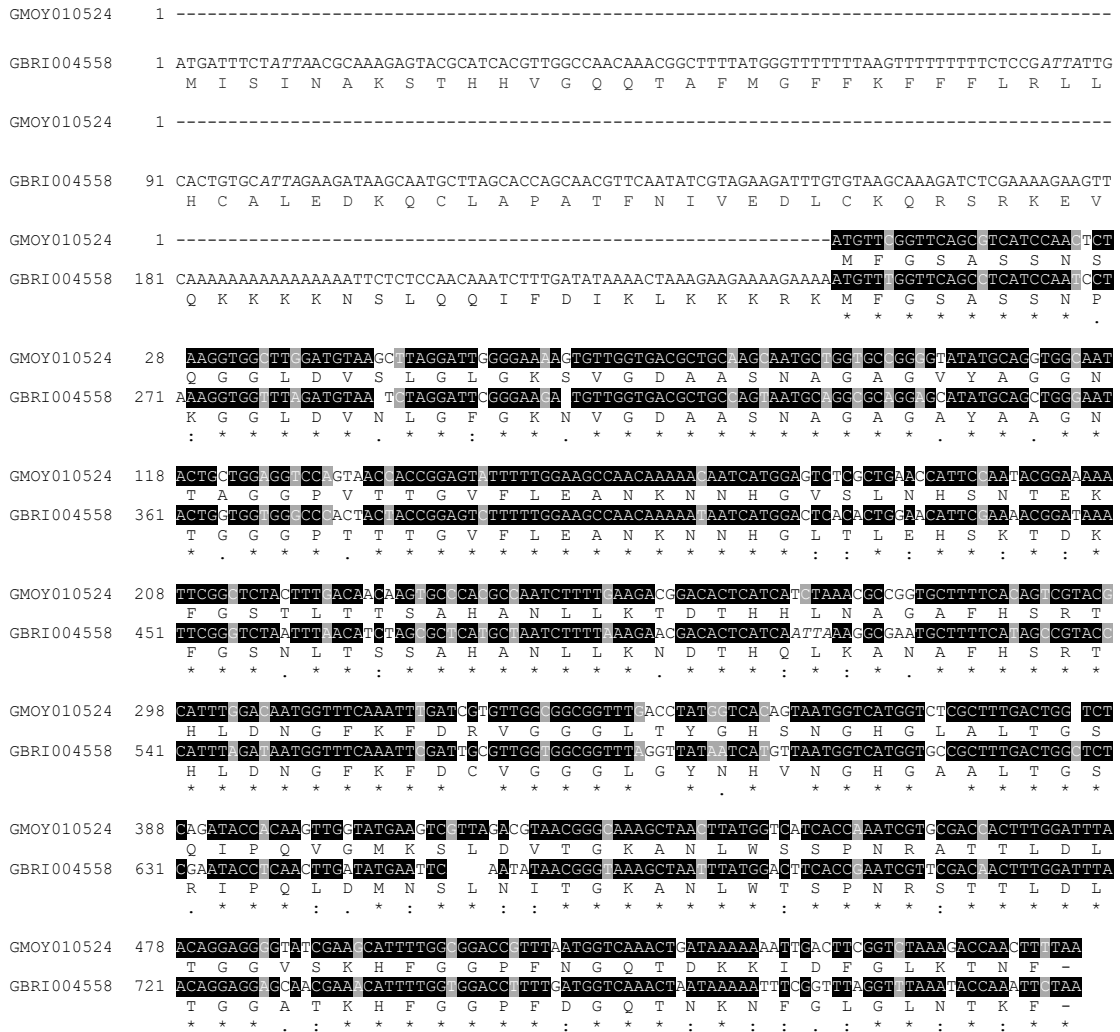


Figure 2.28: The nucleotide and amino acid alignment of GMOY010521 (*AttD*) and GBRI004558 (*AttD*) produced by ExPASy Boxshade. Conservation within the nucleotide sequence is indicated by the degree of shading, Black = complete conservation, Grey = synonymous mutation and White = non-synonymous mutations. Protein similarity is denoted using standard alignment methodology, \* = complete conservation, : = residues with a Gonnet PAM 250 score > 0.5, . = residues with residues with a Gonnet PAM 250 score < 0.5 and a gap = no similarity. A hyphen (-) in the amino acid sequence indicates a stop codon or missing data.

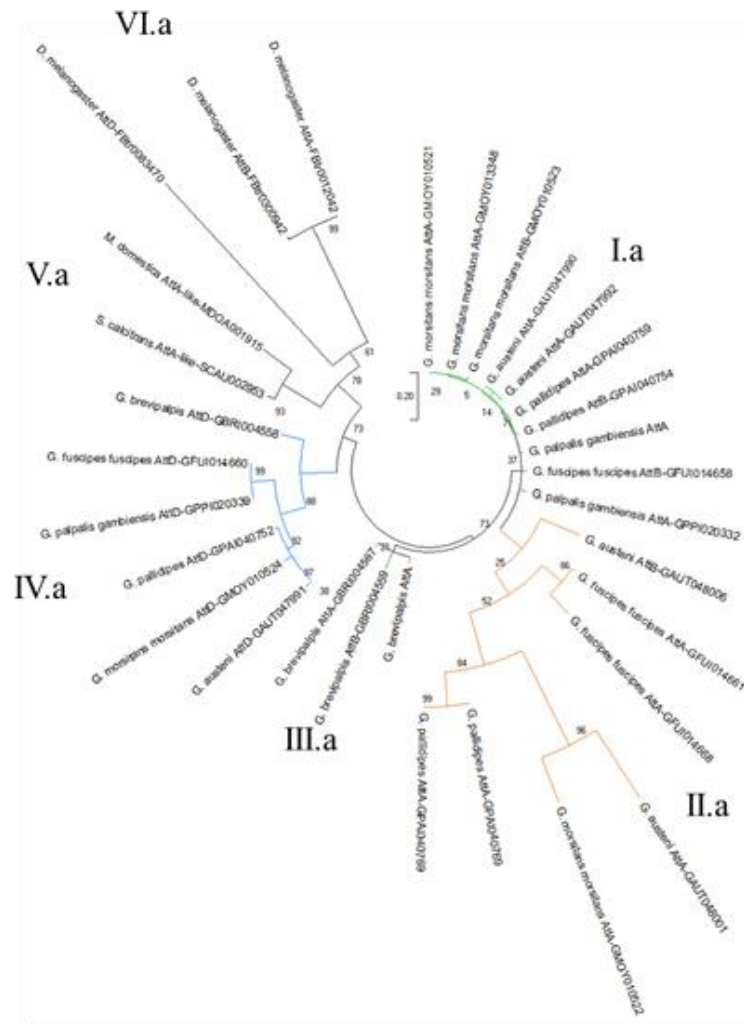
### 2.4.2: Interspecies analysis

The evolutionary history of these genes was assessed using both Maximum Likelihood and Neighbour-Joining methods and illustrated two differing topologies (Fig. 2.29). The Maximum Likelihood methodology demonstrated the presence of four *Glossina* clades and two outgroups containing Attacin sequences from other Dipteran species (Fig. 2.29A).

Two *Glossina* clades indicated a clear separation between complete and partial AttA/AttB proteins within the Morsitans and Palpalis groups (Clades I. and II.a). Interestingly the AttA/AttB proteins identified within the Fusca group species *G. brevipalpis* form a distinct sister clade (Clade III.a) to the other AttA/AttB proteins. The final *Glossina* clade illustrates the divergence between AttD and the other attacin proteins (Clade IV.a). This suggests that the evolution of attacin proteins follows the same evolutionary pathway as the *Glossina* species.

The topology revealed by the Neighbour-Joining method (Fig. 2.29B) was in variance to the Maximum-Likelihood method. Three distinct *Glossina* clades were observed, highlighting the separation of complete and partial AttA/AttB proteins (Clades I.b and II.b), while AttD formed an identical sister clade (Clade III.b) to that exhibited by the Maximum-Likelihood method (Fig. 2.29). Interestingly, the *G. brevipalpis* sister clade exhibited by the Maximum-Likelihood tree (Clade III.a, Fig. 2.29A) forms a subclade of the primary complete AttA/AttB protein clade (Clade I.b), although a single *G. brevipalpis* AttA protein does show divergence from all other *Glossina* proteins (Fig. 2.29B).

**A)**



**B)**

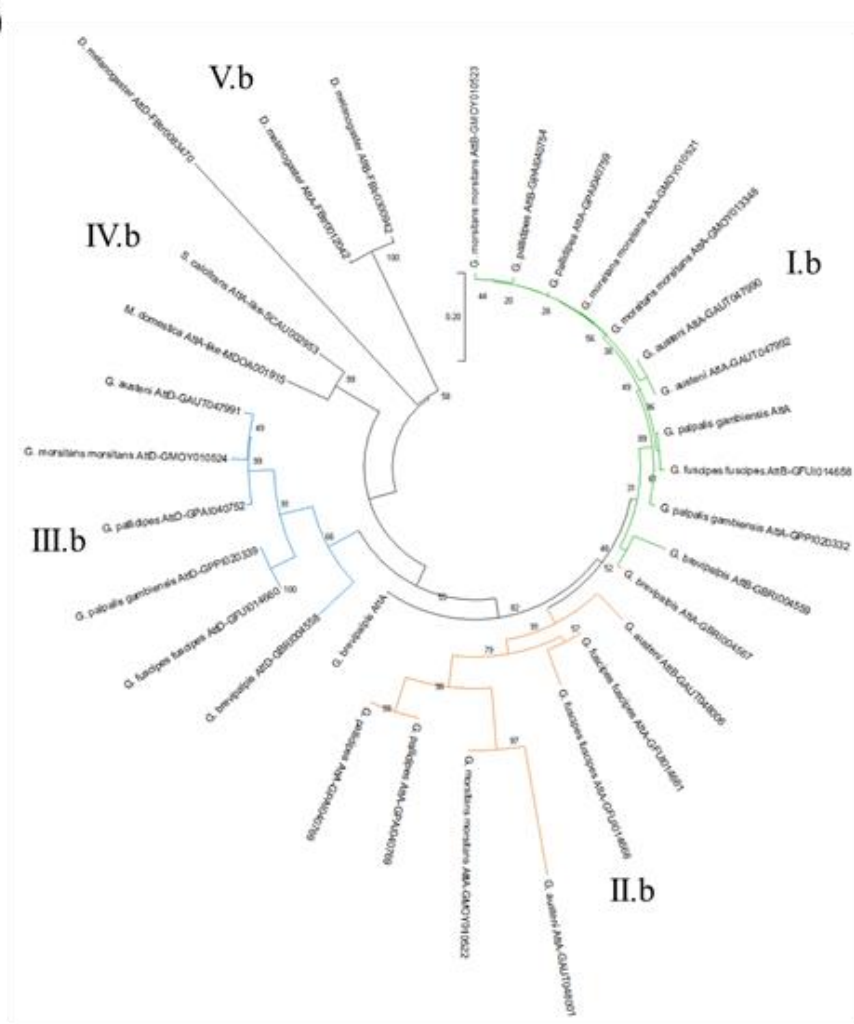




Figure 2.29: Evolutionary analyses were conducted in MEGAX (Kumar et al., 2018). A) The evolutionary history of the attacin gene family inferred by the Maximum Likelihood method, using the WAG model (Whelan and Goldman, 2001), a discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories ([+G], parameter = 2.6171)). B) The evolutionary history of the attacin gene family as inferred by the Neighbour-Joining method (Saitou and Nei, 1987). The optimal tree with the sum of branch length = 3.15867194 is shown. The evolutionary distances were computed using the Poisson correction method (Zuckerandl and Pauling, 1965). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). Both trees are drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions with less than 50% site coverage were eliminated. *D. melanogaster*, *M. domestica* and *S. calcitrans* were added as an outgroup. Branches coloured in green show complete AttA and AttB proteins, those in orange show partial AttA and AttB proteins, while AttD is shown in blue.

#### 2.4.2i: Attacin Family Nucleotide Variation

There is a clear difference between each Attacin gene family based on the level of nucleotide variation. Attacin-A shows a high level of conservation in the N-terminal while the C-terminal shows considerably more variation (Fig. 2.30A). This is likely due to the greater number of genes exhibiting the C-terminal as all but five of the predicted *AttA* genes were missing the N-terminal. Variation within *AttB* was consistent throughout the gene, whilst also indicating a greater degree of variation than *AttA* (Fig. 2.30B). This was supported by the higher values of nucleotide variation observed in *AttB* in addition to the greater number of mutations. In comparison, *AttD* contained fewer points of variation than *AttB*, though nucleotide variation ( $\pi$ ) was considerably higher throughout *AttD* than that observed in *AttA* or *AttB* (Fig. 2.30C).

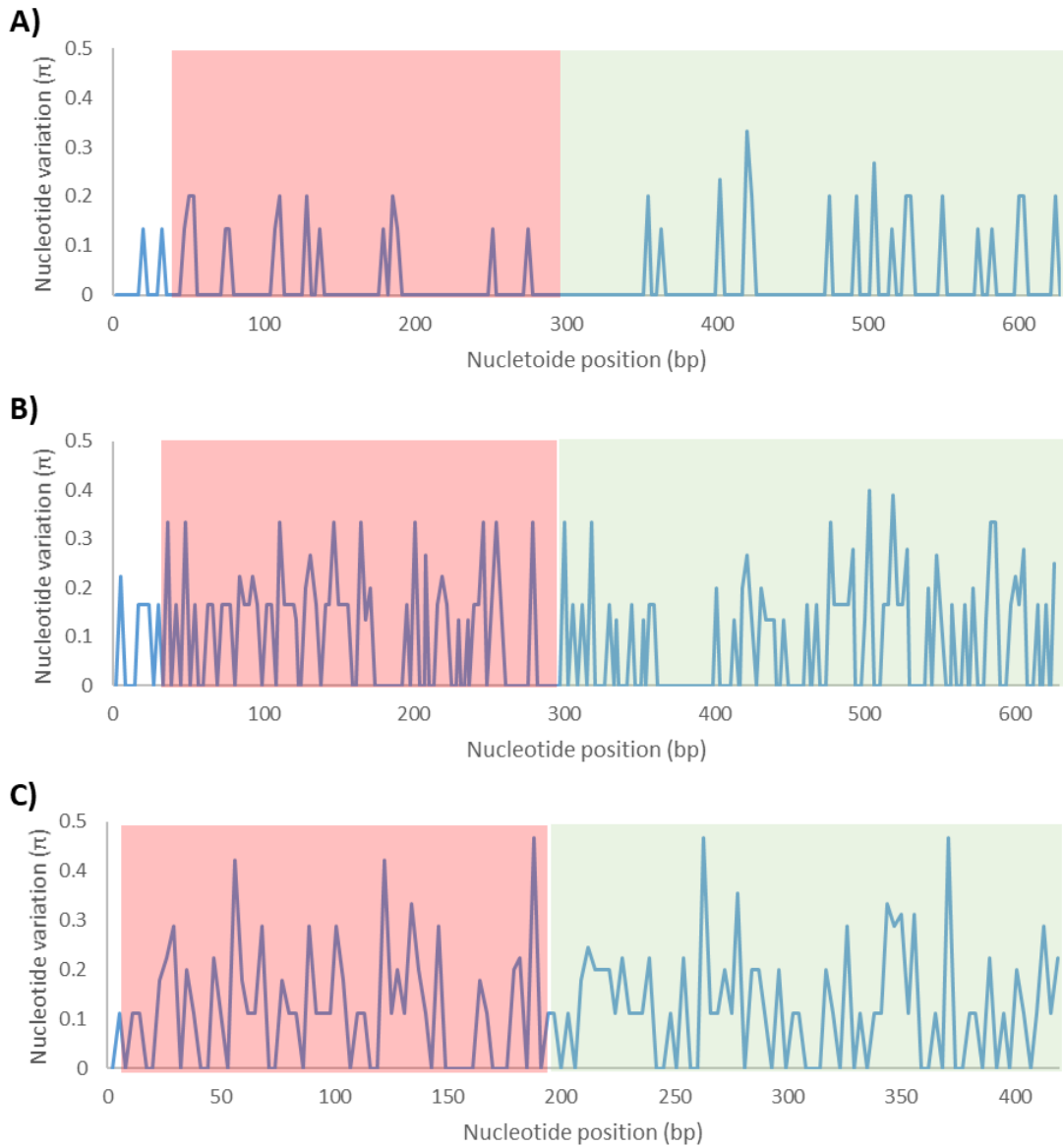


Figure 2.30: Sliding window analysis of the predicted Attacin genes CDS within the *Glossina* spp. A) Predicted *AttA* genes, contain both full and partial sequences. B) Predicted *AttB*, contain all five predicted genes. C) All predicted *AttD* genes. Sliding window analysis was run in DnaSP (version 6), window size = 3 and step size = 3.  $\pi$  was calculated as defined in equation 1. The red shade area indicates the location of the *AttA* N-terminal domain, while the green area shows the C-terminal domain.

#### 2.4.2ii: Pairwise Distance Principle Component Analysis

Principle component analysis was conducted on the pairwise distance ( $P$ ) of all predicted *Glossina* attacin genes, this illustrated the divergence between attacin genes by comparing the proportion of amino acid difference between sequences. Principle components (PC) 1 and 2 were used in the analysis with Eigenvalues and percentage of variance equalling 0.185 and 79.169% (PC1) and 0.026 and 11.057% (PC2) (Fig. 2.31). This plot clearly supports the observations made previously during the phylogenetic analysis (Fig. 2.29), whilst also illustrating further the relationship between the predicted attacin genes. As with the phylogenetic analysis the PCA plot indicates a clear separation of the predicted *AttD* genes from the other attacin genes. As expected from the phylogenetic analysis and gene alignments, *AttA* and *AttB* show a close conservation between the two genes, with partial genes clustering slightly apart from the fully predicted genes. Additionally, the diversification of the Fusca group species *G. brevipalpis* was illustrated further.

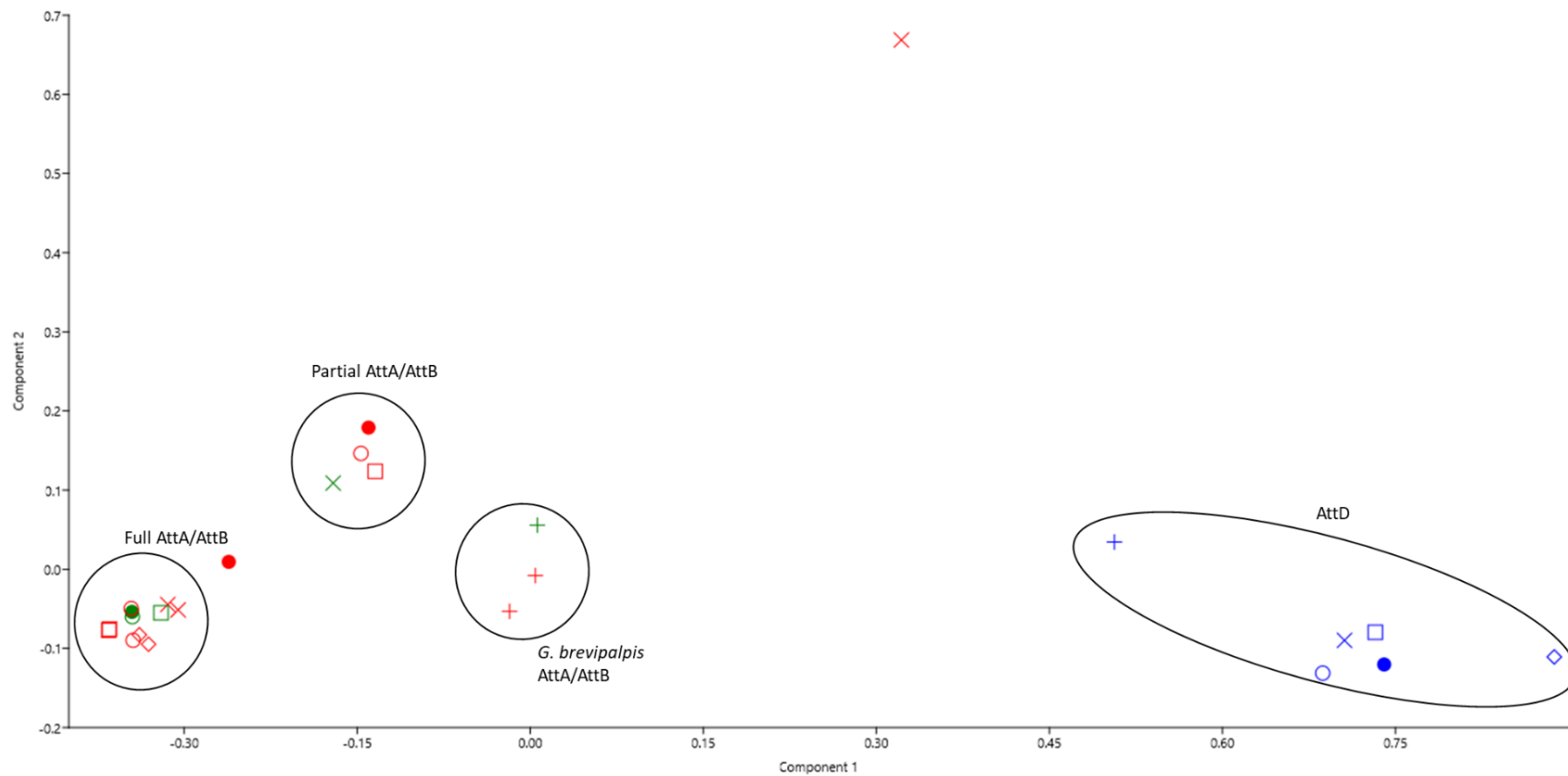


Figure 2.31: Principle component analysis (PCA) plot of all predicted Attacin genes within the *Glossina* genus using the first and second principle components (Eigenvalues: PC1 = 0.185 (79.169 % variance); PC2 = 0.026 (11.057 % variance)). Pairwise distance estimations were conducted in MEGAX (Kumar *et al.*, 2018), using pairwise distance. All sites with less than 50 % coverage were eliminated. A distance matrix was produced in Microsoft Excel and PCA analysis was conducted in PAST3 (Hammer *et al.*, 2001). Individual predicted Attacin genes are shown by plots, species is denoted by shape: *G. austeni* = X; *G. brevipalpis* = +; *G. f. fuscipes* = ●; *G. m. morsitans* = □; *G. pallidipes* = ○; *G. palpalis gambiensis* = ◇. Gene families are denoted by colour. *AttA* = Red; *AttB* = Green; *AttD* = Blue.

#### 2.4.2iii: Three-dimensional structural analysis

The 3-D structure of each predicted full Attacin protein was predicted using the I-Tasser online server. This revealed that all three predicted Attacin genes illustrated a highly conserved concave tertiary structure, consisting primarily of coiled N- and C-terminals with a series of anti-parallel  $\beta$ -sheets forming the majority of the structure (Fig. 2.32A). Of the three predicted *AttA* structures, all followed the same overall structure with between six and 11 anti-parallel  $\beta$ -sheets following a coiled N-terminal. Further to the differing number of  $\beta$ -sheets, the only other point of variation was observed in the predicted *G. pallidipes AttA* structure, which featured a single helix in the N-terminal between S14 and L17 (Fig. 2.32A).

The predicted structures of the four full *AttB* genes had a higher percentage of random coil structures, making direct alignment more difficult (Fig. 2.32B); nevertheless, they maintain a similar overall structure to that observed within *AttA* and *AttD*. The N-terminal appears to consist of random coils with a single helix (similar to that observed in *G. pallidipes AttB*), leading to either nine or 10 anti-parallel  $\beta$ -sheets comprising the majority of the concave protein structure preceding a coiled C-terminal. Notably, the predicted *G. brevipalpis AttB* structure does not exhibit the N-terminal helix, instead consisting purely of coiled structures (Fig. 2.23B).

Despite the nucleotide and amino acid variation observed across the identified *AttD* genes, their predicted structure exhibits a more rigid and conserved structure across all *Glossina* species (Fig. 2.32C). A short, coiled N-terminal precedes a series of between nine and 11 anti-parallel  $\beta$ -sheets and another short, coiled C-terminal. The only major variation to this structure was observed in *G. pallidipes* which exhibits a single helix between P132 and V134 (Fig. 2.32C).

Both DALI Z-scores and PCA analysis support the observations made during the structural alignments. Two main clusters of structural conservation were highlighted within the samples (Fig. 2.33). Firstly, five of the six predicted *AttD* structures show a higher degree of conservation with Z-scores ranging between 16.2 and 22. Secondly, a cluster consisting of *G. m. morsitans AttA*, *G. f. fuscipes AttB*, *G. pallidipes AttA and AttB*, and *G. austeni AttA* show higher levels of conservation between them ( $17.8 \leq Z \leq 19.2$ ) (Fig. 2.33A). Interestingly, three structures appear to show little to no similarity to the other attacin

samples. While *G. m. morsitans AttB*, shows little similarity with any of the *AttD* cluster samples ( $Z \leq 12.5$ ), it also illustrates a surprising lack of similarity to other predicted *AttA* or *AttB* structures ( $Z \leq 17.1$ ). Similarly, *G. f. fuscipes AttD* indicated a similarity degree of similarity to all other predicted structures, though this was relatively low ( $12.4 \leq Z \leq 16.1$ ). Finally, *G. brevipalpis AttB* showed almost no similarity to any other predicted structure (Fig. 2.33A).

Principle component analysis illustrated a comparable distribution of samples (Fig. 2.33B). Both the previously observed *AttA/AttB* cluster and *AttD* cluster are clearly visible. As expected from the heatmap results (Fig. 2.33), *G. m. morsitans AttB* shows higher similarity to the *AttA/AttB* cluster, while *G. f. fuscipes AttD* falls roughly between both of the clusters. The PCA plot of *G. brevipalpis AttB* further indicates the lack of similarity between itself and the other predicted Attacin structures (Fig. 2.33B).

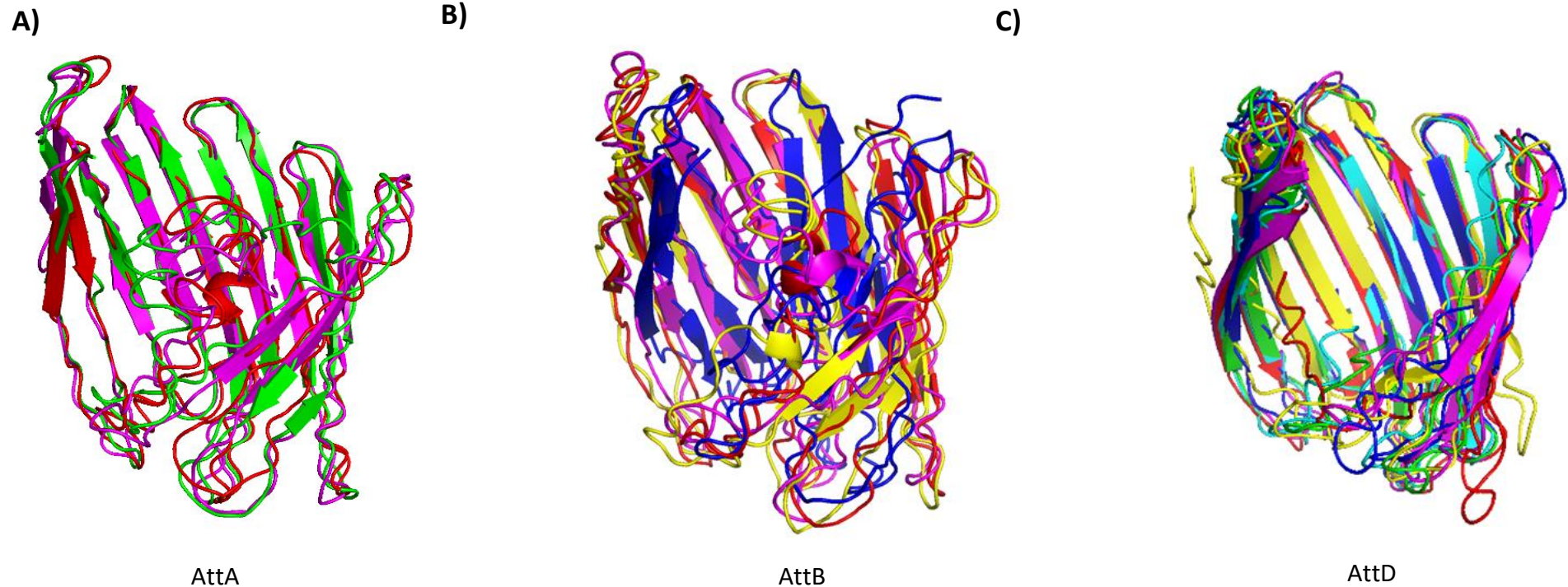


Figure 2.32: Structural alignment of complete predicted *Glossina* Attacin protein families A) *AttA*; B) *AttB* and C) *AttD*. PDB files were produced using I-TASSER server (Yang and Zhang, 2015; Yang *et al.*, 2015) and models visualized and aligned in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre). A) *AttA* protein structures illustrate little alignment at the N-terminal due to the high concentration of random coils. However, the conserved concave structure is observable forming around the anti-parallel  $\beta$ -sheets. B) The alignment of *AttB* structures shows dramatically less conservation, again due to this high concentration of coiled structures. However, the series of anti-parallel  $\beta$ -sheets is clearly visible. C) The level of conservation within *AttD* structures was unexpected, however the rigid structure formed around the anti-parallel  $\beta$ -sheets explains this high degree of conservation. All models exhibited low C-Scores ( $C < -1.8$ ), model colour denotes the species of the predicted protein structure: Red = *G. pallidipes*; Green = *G. austeni*; Purple = *G. m. morsitans*; Yellow = *G. f. fuscipes*; Cyan = *G. p. gambiensis*; Blue = *G. brevipalpis*.



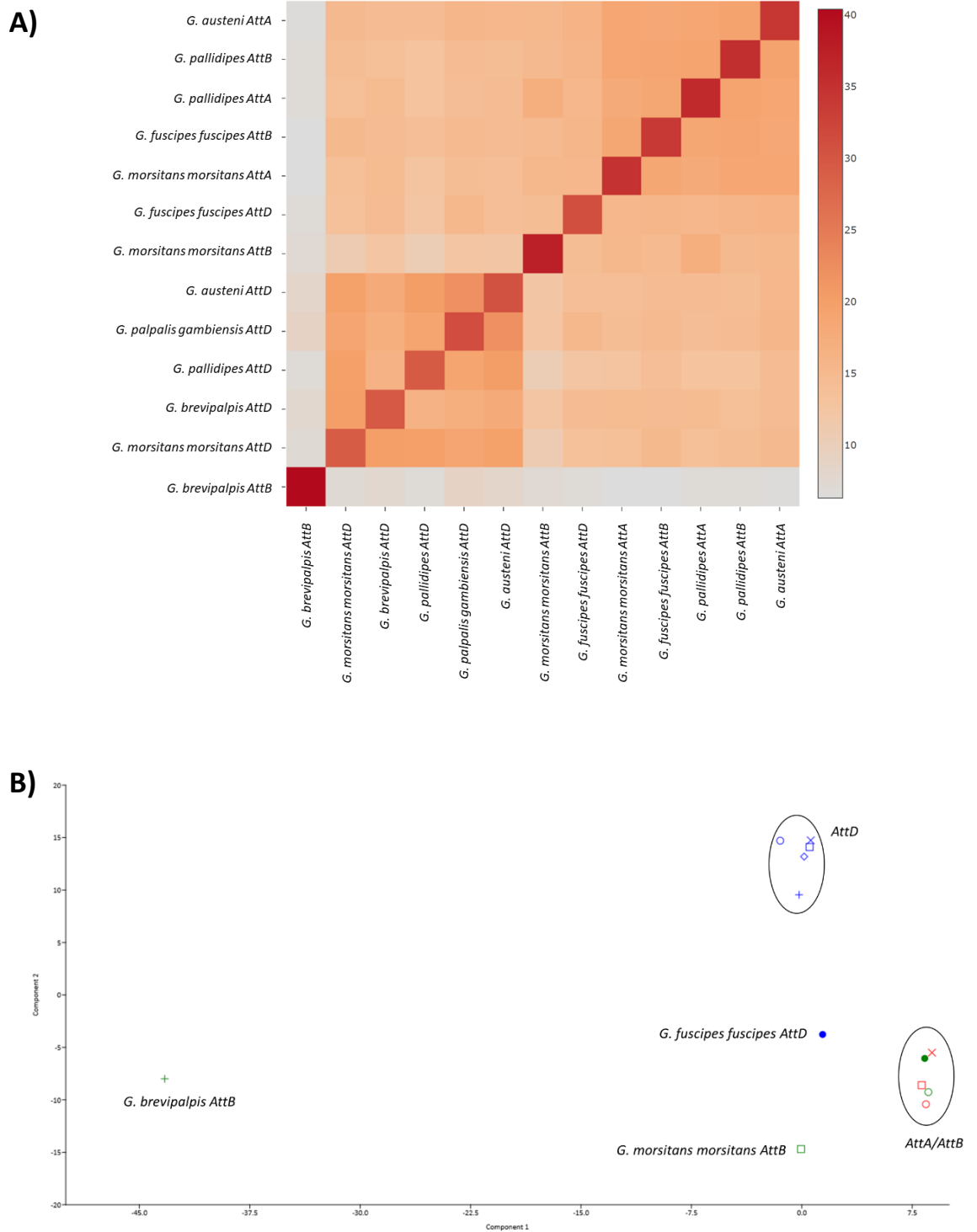


Figure 2.33: A) A heatmap comparing the conservation of Attacin protein structures across all complete predicted *Glossina* Attacin genes. The heatmap was constructed using DALI server (Holm, 2020). Colour represents the Z-value as estimated by DALI, higher Z-score and conservation are shown by the deep red colouration, while low Z-scores and shown by the grey. B) Principle component analysis (PCA) plot of predicted Attacin proteins within the *Glossina* genus using the first and second principle components (Eigenvalues: PC1 = 186.36 (36.29 % variance); PC2 = 127.268 (24.783 % variance)). Z-values were calculated, and a matrix produced using DALI (Holm, 2020). PCA analysis was conducted in PAST3 (Hammer *et al.*, 2001). Individual predicted Attacin genes are shown by plots, species is denoted by shape: *G. austeni* = X; *G. brevipalpis* = +; *G. f. fuscipes* = ●; *G. m. morsitans* = □; *G. pallidipes* = ○; *G. p. gambiensis* = ◇. Gene families are denoted by colour. AttA = Red; AttB = Green and AttD = Blue.

## 2.5: Defensin

Previously, no defensin (*Def*) genes had been annotated within any *Glossina* genomes. Identification of *Def* genes within the *Glossina* genomes was achieved using a tBLASTn search of the GenBank sequence AAL34112.1, published previously by Hao *et al.* (2001). This identified a single predicted *Def* within each of the available *Glossina* genomes. These predicted genes showed variation in gene structure will differing CDS sizes, and the number of exons within each gene.

### 2.5.1: *Glossina* defensin identification

#### 2.5.1i: *Glossina morsitans morsitans*

Surprisingly, defensin has not been annotated previously within the *G. m. morsitans* genome. A tBLASTn search yielded a 100% amino acid match, with a 98.86% nucleotide alignment, with a sequence within contig. CCAG010013027 (SuperContig. Scf7180000644371) (Fig. 2.34). The result of this search places the defensin gene on the forward strand of SuperContig scf7180000644371 between nucleotides 18,960 and 19,280. The alignment suggests the gene contains a single exon with non-coding regions at the N and C terminals, a single arthropod defensin domain was identified at the 5' C-terminal by Pfam (Fig. 2.34).

A)



B)

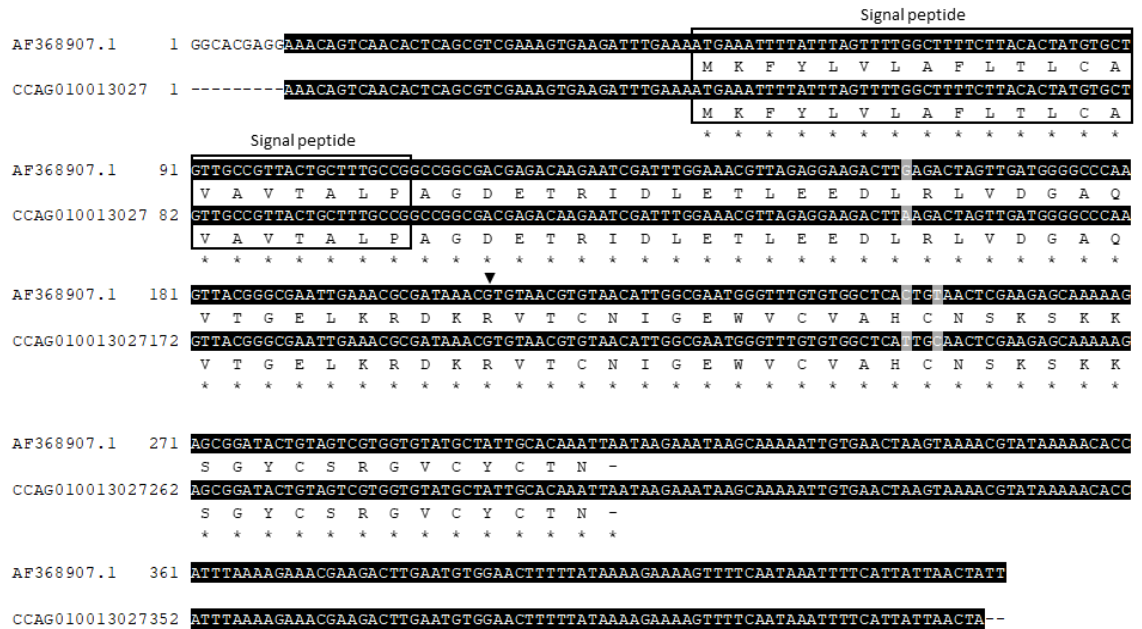


Figure 2.34: A) The linear map of CCAG010013027352 is also given, the location of the defensin gene is shown by the green area. Coding regions are represented by the filled areas, while non-coding regions are shown by the outlined sections. The position of the arthropod defensin domain identified by Pfam is shown by the light green box. B) An alignment of the *G. m. morsitans* defensin sequences produced by Hao *et al.* (2001) and the corresponding section of Contig. CCAG010013027352, the amino acid translation of each sequence is given below the nucleotide sequence. This shows a strong alignment between the two sequences with three synonymous nucleotide mutations being observed between them. Nucleotide conservation is denoted by the degree of shading where black = complete conservation, grey = synonymous mutations and white = non-synonymous mutations. Amino acid conservation is shown using standard alignment denotation (\* = a conserved residue, : = residues with a Gonnet PAM 250 score > 0.5 and . = residues with residues with a Gonnet PAM 250 score < 0.5). □ indicates the start of the mature defensin peptide as described by Hoa *et al.* (2001).

### 2.5.1ii: *Glossina austeni*

The predicted Def gene, GAUT030101, showed a considerably longer CDS than that observed within *G. m. morsitans*, with a CDS length of 942 base pairs over 4 exons. When aligned to AF368907.1 a greater degree of variation was observed with 12 amino acid substitutions within the identified arthropod defensin domain, an amino acid alignment of 86.2 % (Fig. 2.35). This variation was observed within in the phylogenetic analysis (Fig. 2.41), where the GAUT030101 diverged from the other members of the Morsitans group. An arthropod defensin domain was identified by Pfam in the C-terminal of gene with an e-value of  $1.3e-05$  (Fig. 2.35), no other protein domains were identified within the CDS.

A)



B)

AF368907.1	1	-----	
GAUT030101	1	ATGACTGCCACTGAAAGACGAGGGGATGTTAATGTCCATAACCAAAGTTCACCGTTAGACAAAATATGAAAAATGAAAACGGCCACTATG	M T A T E R R G D V N V H N Q S S P L D K Y E N E N W R T M
AF368907.1	44	-----	
GAUT030101	91	TTCATGCTTGTGTTTACAGATATGATACTAACGGATTTTCATGTTCCACACGATATGTTTCTACATTTCTAACGTTTGGCATATTGCAATAC	F M L V F T D M I L T D F H V H T I C F Y I L T F G I L Q Y
AF368907.1	44	-----	
GAUT030101	181	CGTTCTACAGCGTATGTTGGTGTACCGTCTCTTCCATGCGCTTATACCAATTATAGTCGTGATCGTCTAATAGTCCACACCTGCTGC	R F Y S V C W C T V L F H A L I P I I V V I V L I V H T C C
AF368907.1	44	-----	
GAUT030101	271	GATGTCATTACCAACTTCCAGAACCTGAGTATAATAATCCCAATAATAGCACAAATTTAGCGCCATTCGCTAGTTGTATGTCGTTGG	D V I T N F Q N L S I I I P I I A Q I V A A I P L V V C R W
AF368907.1	44	-----	
GAUT030101	361	AATCGTTCGTGTGACGATCAAGCCTTGATGACTGGTATACATGCTCATTTCGGTCTTATAGTGGTTATAATTTATGTTGACCCCTAGTC	N R S C D D Q A L M T G I H A H F G L I V V I I Y W W T L V
AF368907.1	44	-----	
GAUT030101	451	TTCTTAATAAGGCGTCCCTTTGAAGCTTGCATGGTTACCATAGTCTATATGCTGTGTTTCTGGCAGTGTACTTACGCCACAGTTTTAT	F L I R R P F E A C M V T I V Y M L C F L A V L L T P Q F Y
AF368907.1	44	-----	
GAUT030101	541	GGTGTGCGCAAATCTTTCACCTGAAAGCAAAACATTAACCTACGATATCATTGAAAGAGTTTCGCGAACAAATTCATCAGACCCGATT	G V A K S F T E K K N I K P T I S L K E F R E Q I Q S R P I
AF368907.1	55	-----	
GAUT030101	631	ACCATACAGTCGCGATCGCAGTTGCAACAAATACAATCAACTGCATCAAAGCGCTCAAGTCCAAACCTTTGGCTTTTCTTCACTTTGTCGCT	T I Q S R S Q L Q Q I Q S T A S K A
AF368907.1	91	-----	
GAUT030101	721	GTTGCCGTTACGCTTTCGCCGCGCGGACGAGACAAAGAAATCGATTGGAAACCTTAGACCAAGACCTGAGACTAGTTGATGGCCCA	V A V T A L P A G D E T R I D L E T L E E D L R L V D G A Q
AF368907.1	181	-----	
GAUT030101	811	GCTACGGGGAATTGAAACGGAAACCGTGTAACTGGCGAATGGGCTTTGTTGGCTCAATGAACTCGAAGAGCAAAAAG	A T G E L K R D K R V T C N I G E W A C V A H C N S K S K K
AF368907.1	271	-----	
GAUT030101	901	AGCGGTACTCGAGTCGTGGTGTATGCTATTGCACAAATTAATAAGAAATAAGCAAAAATTTGTGAACCTAAGTAAAAAGTATAAAAAACAC	S G Y C S R G V C Y C T N -
AF368907.1	361	-----	
GAUT030101			

Figure 2.35: A) A linear map of JMRR01003512 is also given, the location of GAUT030101 is shown but the brown area. Coding regions are represented by the filled areas, while introns are shown in the linear sections. The position of the arthropod defending domain identified by Pfam is shown by the light green box. B) An alignment of AF368907.1 CDS produced by Hao *et al.* (2001) and GAUT030101, the amino acid translation of each sequence is given below the nucleotide sequence. This shows a greater degree of variation between the predicted *G. austeni* Def sequence and the identified *G. m. morsitans* sequence. Nucleotide conservation is denoted by the degree of shading where black = complete conservation, grey = synonymous mutations and white = non-synonymous mutations. Amino acid conservation is shown using standard alignment denotation (\* = a conserved residue, : = residues with a Gonnet PAM 250 score > 0.5 and . = residues with residues with a Gonnet PAM 250 score < 0.5). ▼ indicates the start of the mature defensin peptide as described by Hoa *et al.* (2001).

### 2.5.1iii: *Glossina pallidipes*

A tBLASTn search against the *G. pallidipes* genome indicated that the defensin protein was coded for by GPAI019770 on the forward strand of Scaffold235. GPAI019770 contains a large CDS of 420 nucleotides over 4 exons, with the highest alignment to the arthropod defensin domain on the third exon (Fig. 2.36A). However, a Pfam search of the protein sequence of GPAI019770 yielded an insignificant match with arthropod defensin domain (E = 0.0011). Closer evaluation of the Scaffold235 sequence around the third exon yielded a complete arthropod defensin domain extending into the third intron, between nucleotides 149,713-150,033 (Fig. 2.36C). This domain was found to have a significant match (E = 2.8e<sup>-05</sup>). The lack of protein domains in the other three exons indicates a mistake within the annotation of Scaffold235.

A)



B)

GPAI019770	1	ATGCATTGCGGCGTTGAACTGCTGTGGACATCTATCTGGAAGCTTCGGCGATATCACAA	ATGCAATGCTTTT	TCTTACGAAAATAAA
		M H C G V E L L W T S I W K L S A I S Q	K Q F F L S Y E N K	
AF368907.1	1	-----	ATGCAATGCTTTT	-----
		- - - - -	M K F Y L	- - - - -
			: * : *	
			Signal peptide	
GPAI019770	91	AGCAAACAGCAAATGATGTTCCAACTAACATCAGATTTAA	GCTTTGGCTTTTCTTACACTAT	TGGCTGTGCCGTTACTGCTTTGCC
		S K Q Q M M F P T N I R F N	A L A F L T L F A V A V T A	L P
AF368907.1	16	-----	GCTTTGGCTTTTCTTACACTAT	TGGCTGTGCCGTTACTGCTTTGCC
		- - - - -	V L A F L T L C A V A V T A	L P
			* * * * *	* * * * *
GPAI019770	181	GCCGGCAGCAGACAGAATCGTTTGGAAAC	TTAGAGGAAGACTT	AGACTAGT
		A G D E T R I G L E T L E E D L R L V D V D Q V T G E L K R	GATGCGC	CCAAGT
AF368907.1	64	GCCGGCAGCAGACAGAATCGTTTGGAAAC	TTAGAGGAAGACTT	AGACTAGT
		A G D E T R I D L E T L E E D L R L V D G A Q V T G E L K R	GATGCGC	CCAAGT
		* * * * *	* * * * *	* * * * *
GPAI019770	271	GATAAACGTGTAACGTAAACATTGGCGAATGGGTTT	GTGGCTCA	TGTAAC
		D K R V T C N I G E W V C V A H C N S K S K K S G Y C Q R G	TGTAAC	CGAAGAGCAA
AF368907.1	154	GATAAACGTGTAACGTAAACATTGGCGAATGGGTTT	GTGGCTCA	TGTAAC
		D K R V T C N I G E W V C V A H C N S K S K K S G Y C S R G	TGTAAC	CGAAGAGCAA
		* * * * *	* * * * *	* * * * *
GPAI019770	361	TCAATGGAAAATGAGGAAAATGGCAACGAGCG	ATGAGCCGGCAA	TATTGATTAATTG
		S M E N E E N W Q R A N E P A N I D N		
AF368907.1	244	-----	GTATGCTATTGC	ATAA
		- - - - -	V C Y C T N	- -
			*	

c)

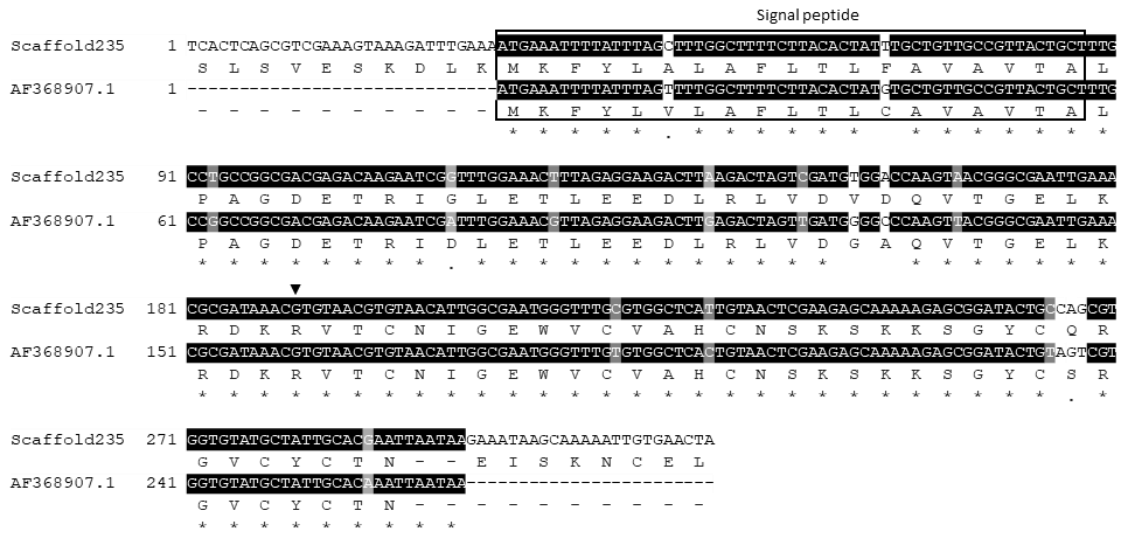


Figure 2.36: A) A linear map of JMRR01003512 is also given, the location of GAUT030101 is shown but the brown area. Coding regions are represented by the filled areas, while introns are shown in the linear sections. The position of the arthropod defending domain identified by Pfam is shown by the light green box. B) An alignment of AF368907.1 CDS produced by Hao *et al.* (2001) and GPAI019770, the amino acid translation of each sequence is given below the nucleotide sequence. This shows a higher level of alignment between the third exon and identified *G. m. morsitans* sequence. However, the final six residues of the Arthropod defending domain are no present within the GPAI019770 sequence. C) An alignment of AF368907.1 CDS produced by Hao *et al.* (2001) and Scaffold235, the amino acid translation of each sequence is given below the nucleotide sequence. This shows an almost identical alignment for plate A, though the last six residues are present. Nucleotide conservation is denoted by the degree of shading where black = complete conservation, grey = synonymous mutations and white = non-synonymous mutations. Amino acid conservation is shown using standard alignment denotation (\* = a conserved residue, : = residues with a Gonnet PAM 250 score > 0.5 and . = residues with Gonnet PAM 250 score < 0.5). ▼ indicates the start of the mature defensin peptide as described by Hoa *et al.* (2001).

2.5.1iv: *Glossina fuscipes. fuscipes*

The predicted *Def* gene within the *G. f. fuscipes* genome, GFUI031425, is located on the reverse strand of Scaffold40, containing a CDS of 534 nucleotides over 4 exons. Pfam identified an Arthropod defensin domain within the first exon of GFUI031425 ( $E = 7.5e^{-06}$ ), though no other protein domains were identified in the sequence. This is supported by the alignment of GFUI031425 to AF368907.1 which illustrates a conserved alignment throughout the first exon, with just the stop codes aligning to the fourth exon (Fig. 2.37).

A)



B)

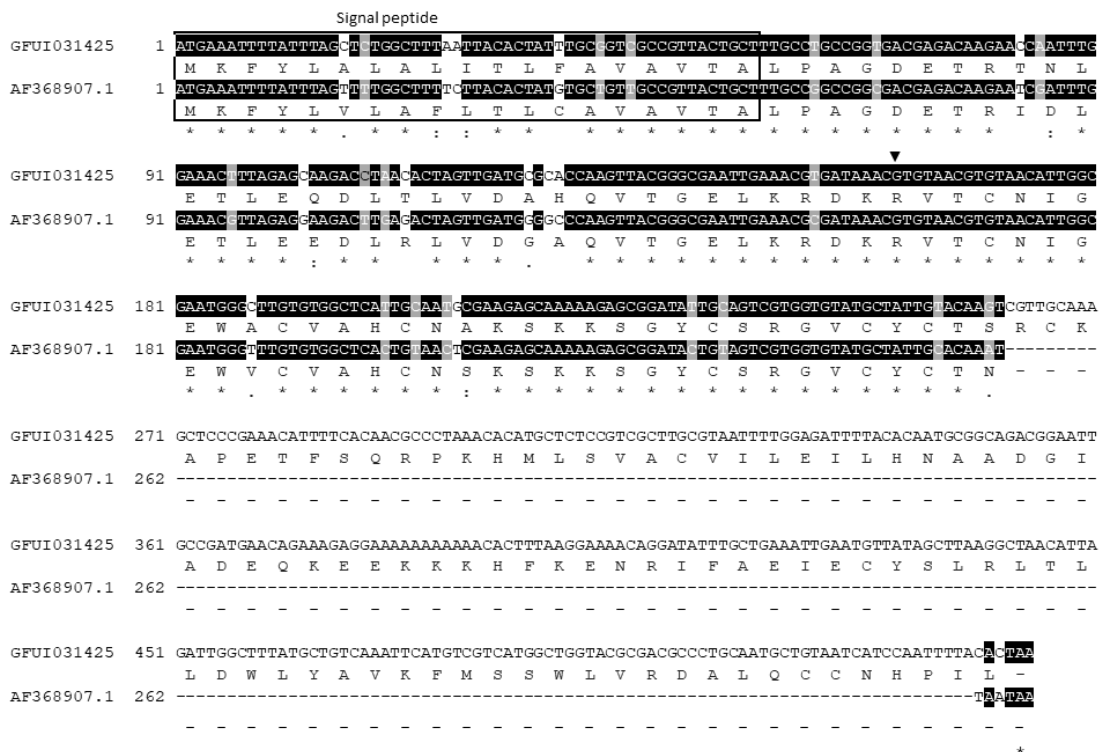


Figure 2.37: A) A linear map of JFJR01001831 is also given, the location of GFUI031425 is shown but the brown area. Coding regions are represented by the filled areas, while introns are shown in the linear sections. The position of the arthropod defensin domain, identified by Pfam, is shown by the light green box. B) An alignment of AF368907.1 CDS produced by Hao *et al.* (2001) and GFUI031425, the amino acid translation of each sequence is given below the nucleotide sequence. This illustrates the conservation of the Arthropod defensin within the first exon of GFUI031425. Despite showing a relatively high level of nucleotide variation to the identified *G. m. morsitans* sequence, the amino acid alignment shows a high level of conservation. Nucleotide conservation is denoted by the degree of shading where black = complete conservation, grey = synonymous mutations and white = non-synonymous mutations. Amino acid conservation is shown using standard alignment denotation (\* = a conserved residue, : = residues with a Gonnet PAM 250 score > 0.5 and . = residues with residues with Gonnet PAM 250 score < 0.5). ▼ indicates the start of the mature defensin peptide as described by Hoa *et al.* (2001).



### 2.5.1v: *Glossina palpalis gambiensis*

Of the predicted defensin genes, GPPI029745, shows the highest structural similarity to that observed within the *G. m. morsitans* genome. A single exon with CDS length of 264 nucleotides, codes for an arthropod defensin domain ( $E = 2.1e^{-06}$ ), found on the reverse strand of Scaffold228 between nucleotides 261,990 and 262,253. When aligned to AF368907.1, GPPI029745 shows a lower level of amino acid conservation than observed within the other *Glossina Def* genes (82.76%), though this is likely due to the evolutionary separation of *G. palpalis gambiensis* and *G. m. morsitans* (Fig. 2.38).

A)



B)

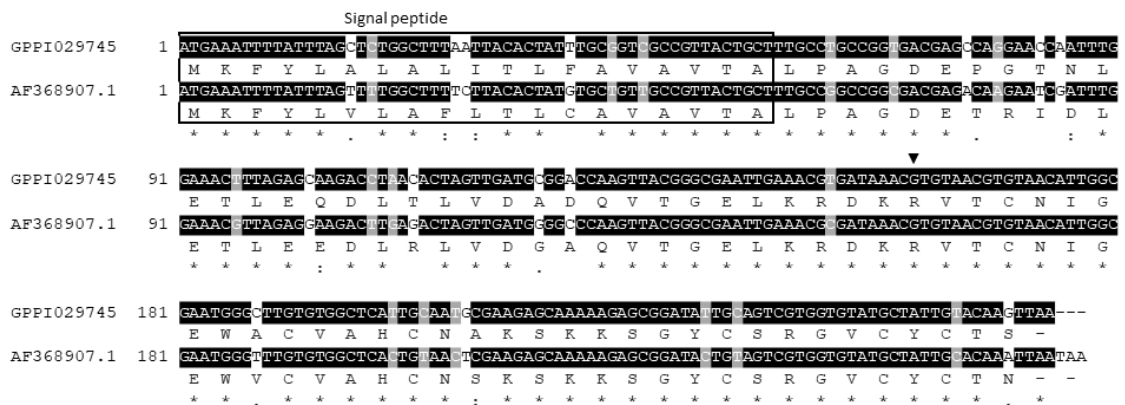
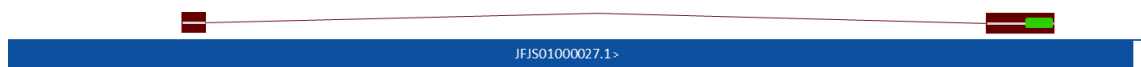


Figure 2.38: A) A linear map of JXJN01014089 is also given, the location of GPPI029745 is shown but the brown area. Coding regions are represented by the filled areas, while introns are shown in the linear sections. The position of the arthropod defensin domain, identified by Pfm, is shown by the light green box. B) An alignment of AF368907.1 CDS produced by Hao *et al.* (2001) and GPPI029745, the amino acid translation of each sequence is given below the nucleotide sequence. This illustrates the variation between the two defensin genes while maintaining the Arthropod defensin domain. Nucleotide conservation is denoted by the degree of shading where black = complete conservation, grey = synonymous mutations and white = non-synonymous mutations. Amino acid conservation is shown using standard alignment denotation (\* = a conserved residue, : = residues with a Gonnet PAM 250 score > 0.5 and . = residues with residues with Gonnet PAM 250 score < 0.5). ▼ indicates the start of the mature defensin peptide as described by Hoa *et al.* (2001).

### 2.5.1vi: *Glossina brevipalpis*

The predicted *Def* gene in *G. brevipalpis*, GBRI000865, is located on the forward strand of Scaffold0 between nucleotides 2,773,250 and 2,776,202 and consists of 2 exons. GBRI000865 demonstrated the largest degree of variation to AF368907.1 with a 64.37% amino acid match, though the C-terminal shows a greater degree of conservation (Fig. 2.39). Pfam identified an Arthropod defensin domain at the C-terminal of GBRI000865, with a significant E value of  $4.1e^{-05}$  (Fig. 2.39).

A)



B)

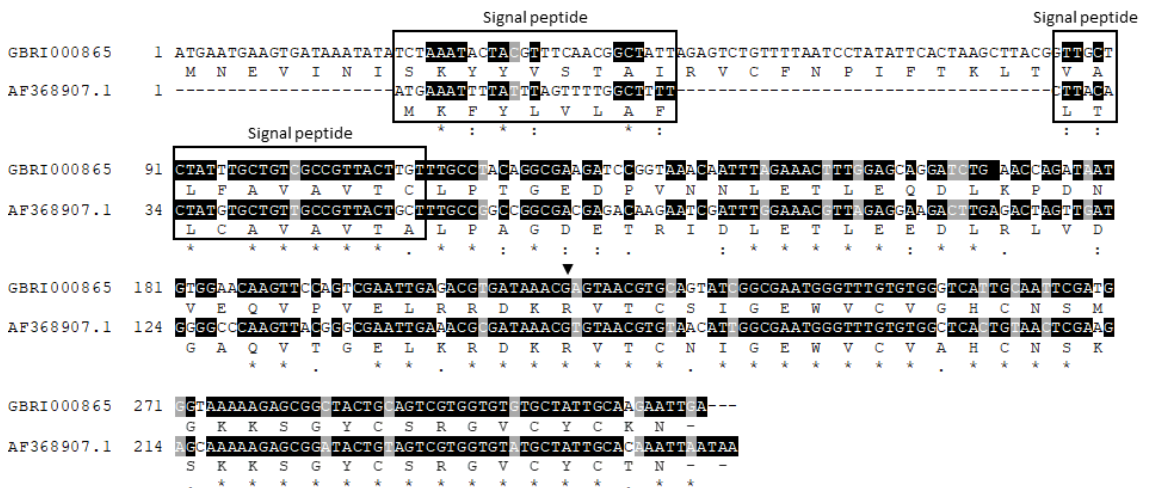


Figure 2.39: A) A linear map of JFJS0100027.1 is also given, the location of GBRI000865 is shown but the brown area. Coding regions are represented by the filled areas, while introns are shown in the linear sections. The position of the arthropod defensin domain, identified by Pfam, is shown by the light green box. B) An alignment of AF368907.1 CDS produced by Hao *et al.* (2001) and GBRI000865, the amino acid translation of each sequence is given below the nucleotide sequence. The variation between the N-terminal is apparent while conservation within C-terminal indicates the presence of an Arthropod defensin domain. Nucleotide conservation is denoted by the degree of shading where black = complete conservation, grey = synonymous mutations and white = non-synonymous mutations. Amino acid conservation is shown using standard alignment denotation (\* = a conserved residue, : = residues with a Gonnet PAM 250 score > 0.5 and . = residues with residues with a Gonnet PAM 250 score < 0.5). ▼ indicates the start of the mature defensin peptide as described by Hoa *et al.* (2001).

### 2.5.2: Interspecies variation

Phylogenetic analysis of these genes illustrated that the predicted Def proteins broadly follow the same evolutionary pathway to the *Glossina* genus as established by Dyer et al. (2008) (Fig. 2.40). The Maximum-Likelihood phylogeny indicated that, while the Morsitans and Fusca groups formed two clear clades (Clades I.a and IV.a), the Palpalis group formed two distinct clades (Clades II.a and III.a) showing a clear divergence between *G. f. fuscipes* and *G. palpalis gambiensis* (Fig. 2.40A). However, this was not strongly supported by bootstrap values, with only two nodes being strongly supported (<75 %) (Fig. 2.40A). The Neighbour-Joining method did not support the separation of Palpalis group species, illustrating that the Morsitans (Clade I.b), Palpalis (Clade II.b) and Fusca (Clade III.b) groups all formed specific clades (Fig. 2.40B). Interestingly this was supported by strong bootstrap values (>80 %), though taxa within the Morsitans group exhibited low bootstrap support.

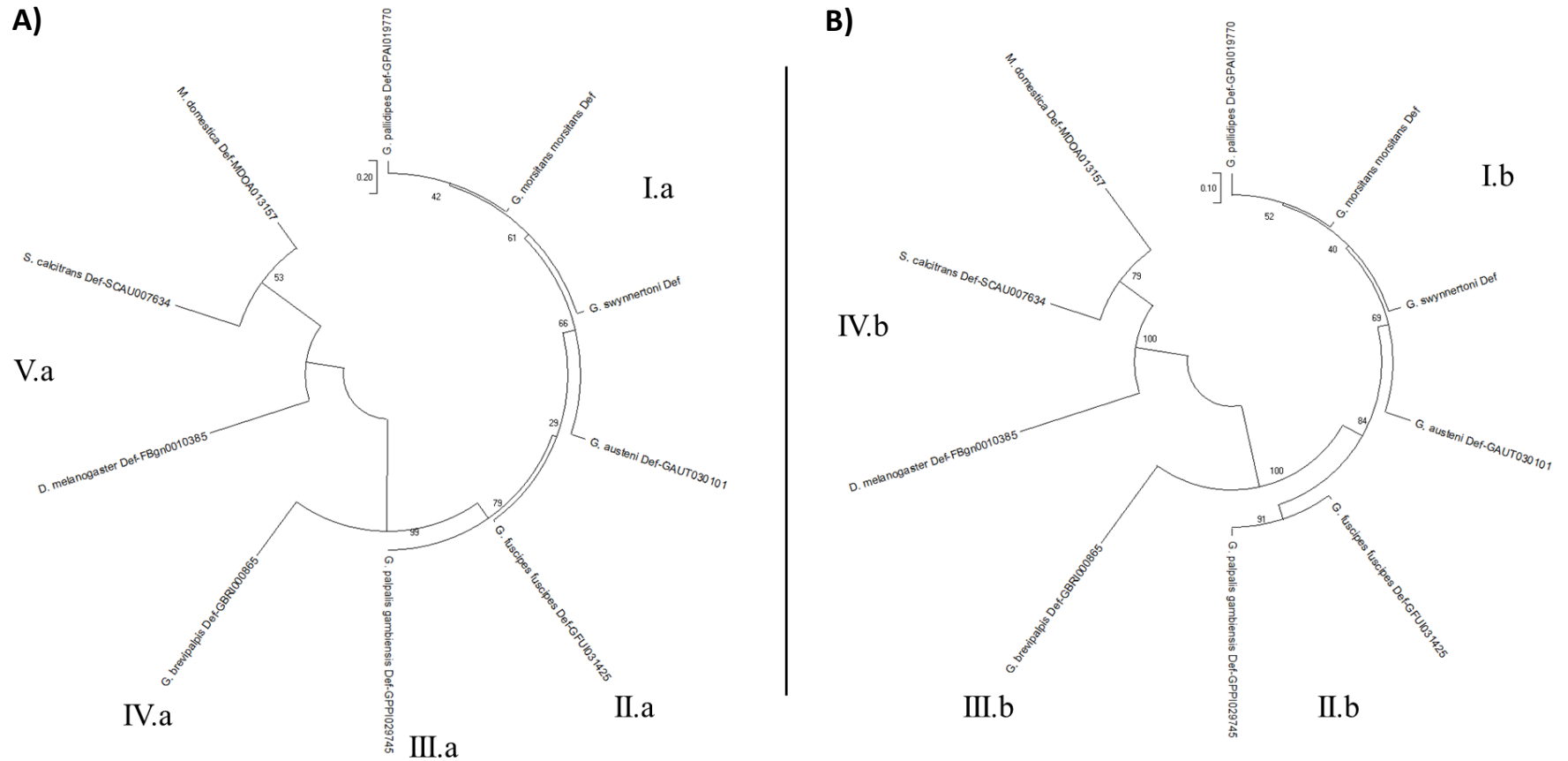
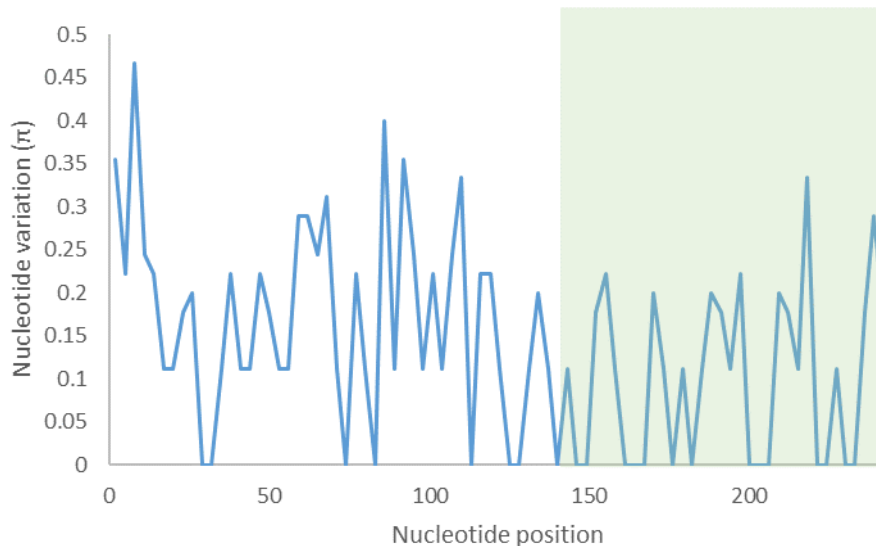


Figure 2.40: Evolutionary analyses were conducted in MEGAX (Kumar *et al.*, 2018). A) The evolutionary history of defensin inferred using the Maximum Likelihood method using the Dayhoff model (Kimura, 1980) and Discrete Gamma distribution (5 categories (+G, parameter = 2.3590)). The tree with the highest log likelihood (-937.73) is shown. B) The evolutionary history of defensin inferred using the Neighbour-Joining method (Saitou and Nei, 1987), the optimal tree is shown, the evolutionary distances were computed using the Poisson correction method (Zuckerkanndl and Pauling, 1965). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). Both trees are drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions with less than 95% site coverage were eliminated. Sequences for *Glossina* species were taken from the VectorBase genomes, *G. swynertoni* defensin sequence was produced by PCR independently. Sequences from *D. melanogaster*, *M. domestica* and *S. calcitrans* were added as an outgroup.

## 2.5.2i: Defensin nucleotide variation

The conserved nature of insect *Def* is clearly illustrated within the predicted *Glossina* defensin genes (Fig. 2.41). Nucleotide variation within the *Glossina* genus was found to be higher throughout the signal domain and pre-peptide regions preceding the mature *Def* region (Fig. 2.41A). Amino acid variation between the Morsitans, Palpalis and Fusca groups is apparent (Fig. 2.41B), the Fusca group species, *G. brevipalpis*, demonstrated the greatest degree of variation, exhibiting five species specific variants within the mature *Def* region (N58S, A66G, K71M, S72G, and T86K). The Palpalis group exhibited two unique variations at amino acids S70A and N87S, though interestingly both of the Palpalis group species and the Morsitans groups species *G. austeni* exhibited the same V63A variation (Fig. 2.41B).

**A)**



**B)**

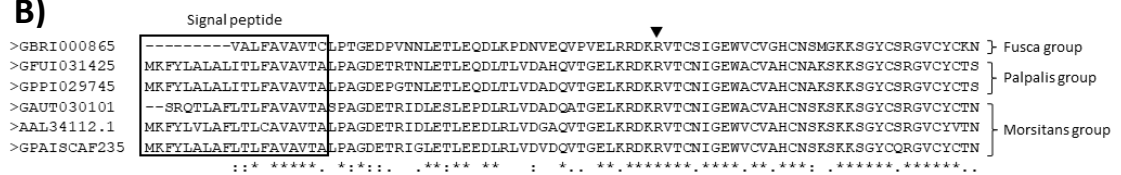


Figure 2.41: A) Sliding window analysis illustrating nucleotide variation within the predicted *Glossina Def* genes. Run in DnaSP (version 6), window size = 3 and step size = 3,  $\pi$  was calculated as defined in equation 1. The mature defensin domain is highlighted demonstrating the lower level of nucleotide variation with this region. B) An amino acid alignment of the predicted *Glossina Def* genes. A ClustalW alignment of the translated *Def* sequences was produced in MUSCLE (Madeira *et al.*, 2019). The signaling peptide is highlighted, while ▼ indicates the start of the mature *Def* peptide. \* = conservation of an amino acid across the alignment. : = conservation of amino acids with similar properties (Gonnet PAM 250 matrix > 0.5), . = conservation of amino acids with a Gonnet PAM 250 matrix < 0.5, while a space shows no conservation between residues.

## 2.5.2ii: Pairwise Distance Principle Component Analysis

As with the predicted Attacin genes, *P*-distance PCA was conducted between all predicted *Def* genes. This illustrated further the variation between the three *Glossina* groups with the *Morsitans* group species clustering around (-0.08,0.075) apart from the Palpalis group at (-0.07,-0.105) and the Fusca group species, *G. brevipalpis*, showing a large degree of variation and separation at (0.37,0.002) (Fig. 2.42). This supports the structure of the observed phylogeny in Figure 2.34. The first and second PCs were used in the analysis with PC1 exhibiting an Eigenvalue of 0.0327 and 75.345 % of the variance, PC2 had an Eigenvalue of 0.00813 and a percentage variance of 18.72 %.

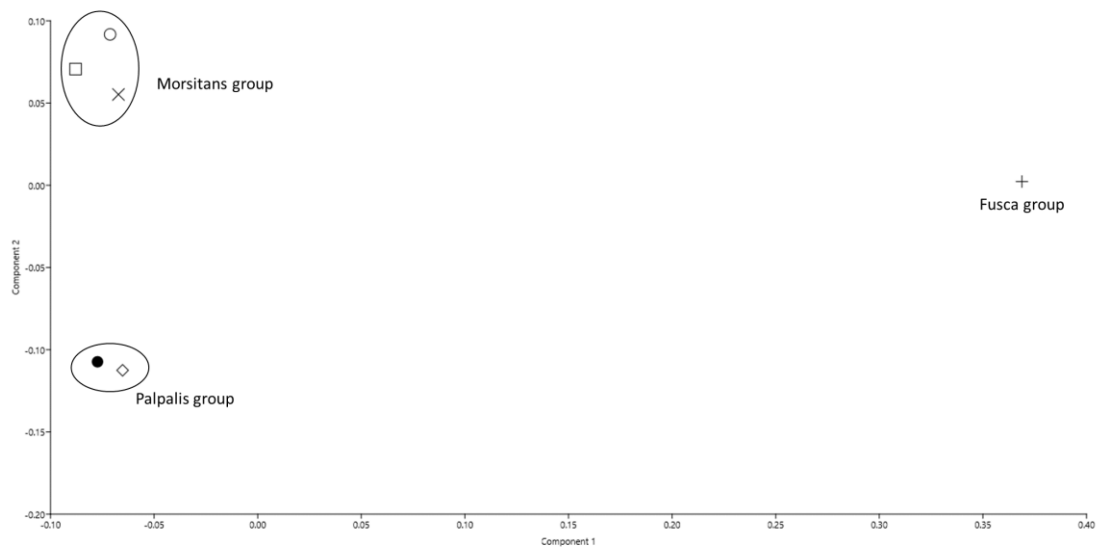
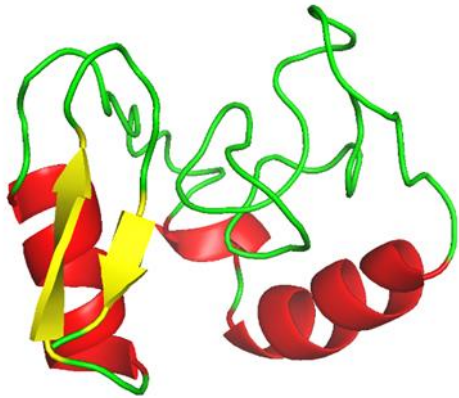


Figure 2.42: Principle component analysis (PCA) plot of all predicted defensin genes within the *Glossina* genus using the first and second principle components (Eigenvalues: PC1 = 0.0327 (75.345 % variance); PC2 = 0.00813 (18.72 % variance)). Pairwise distance estimations were conducted in MEGA7 (Kumar *et al.*, 2016), using pairwise distance calculated by equation 2. All sites with less than 50 % coverage were eliminated. A distance matrix was produced in Microsoft Excel and PCA analysis was conducted in PAST3 (Hammer *et al.*, 2001). Individual predicted Attacin genes are shown by plots, species is denoted by shape: *G. austeni* = X; *G. brevipalpis* = +; *G. f. fuscipes* = ●; *G. m. morsitans* = □; *G. pallidipes* = ○; *G. palpalis gambiensis* = ◇.

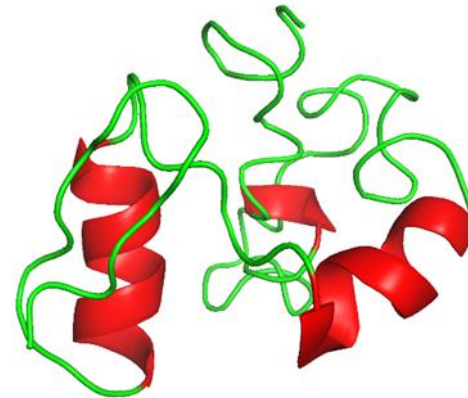
### 2.5.2iii: Defensin Three-dimensional Structural Analysis

Prediction of the defensin 3-D protein structures indicated an unexpected structural variation. All structures exhibited different N-terminal regions, each indicating either a combination of coils and  $\alpha$ -helix or purely coiled structures (Fig. 2.43). The C-terminals exhibited a similar level of variation with just two predicted structures (*G. m. morsitans* and *G. palpalis gambiensis*) showing the expected  $\alpha$ -helix and an antiparallel  $\beta$ -sheet complex that is characteristic of insect defensins. All other species had a helical structure at the C-terminal, though none indicated a  $\beta$ -sheet, instead displaying two anti-parallel coiled structures (Fig. 2.43). These helices differed considerably in size with *G. pallidipes* exhibiting a short five residue helix between codons V65 and S70, and *G. austeni* illustrating a ten-residue helix (E61 – K71) (Fig. 2.43).

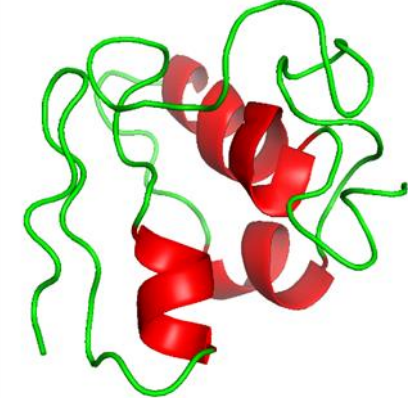
Statistical similarity analysis, using DALI and PCA, gave conflicting results. A heatmap, produced by DALI, indicated a degree of similarity between *G. m. morsitans* and *G. f. fuscipes*, as well as *G. austeni* and *G. brevipalpis*. However, this was not strongly supported with Z-scores of 2.8 and 2.1 respectively (Fig. 2.44A). There was no indication of similarity between any other structures ( $Z = 0.1$ ). Interestingly, while PCA analysis did indicate similarities between *G. m. morsitans* and *G. f. fuscipes* (Fig. 2.44B;  $X = -10$ ,  $Y = 11$ ), it indicated no similarity between *G. austeni* and *G. brevipalpis*. Rather, similarity was detected between *G. pallidipes* and *G. palpalis gambiensis* (Fig. 2.44C;  $X = -4$ ,  $Y = -12.5$ ).



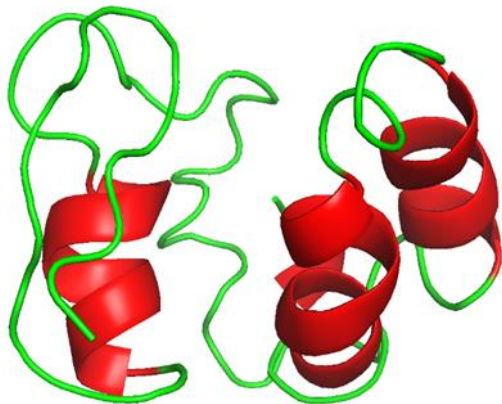
*G. morsitans morsitans*



*G. austeni*



*G. pallidipes*



*G. fuscipes fuscipes*



*G. palpalis gambiensis*



*G. brevipalpis*

Figure 2.43: Structural prediction of *Def* proteins structures produced using I-TASSER server (Yang and Zhang, 2015; Yang *et al.*, 2015), the produced PDB files were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre). Secondary structures are represented by colour, coils are shown in green, helices in red and  $\beta$ -sheet in yellow, all models exhibited low C-Scores ( $C < -2.4$ ), indicating that the models show little similarity to the templates.



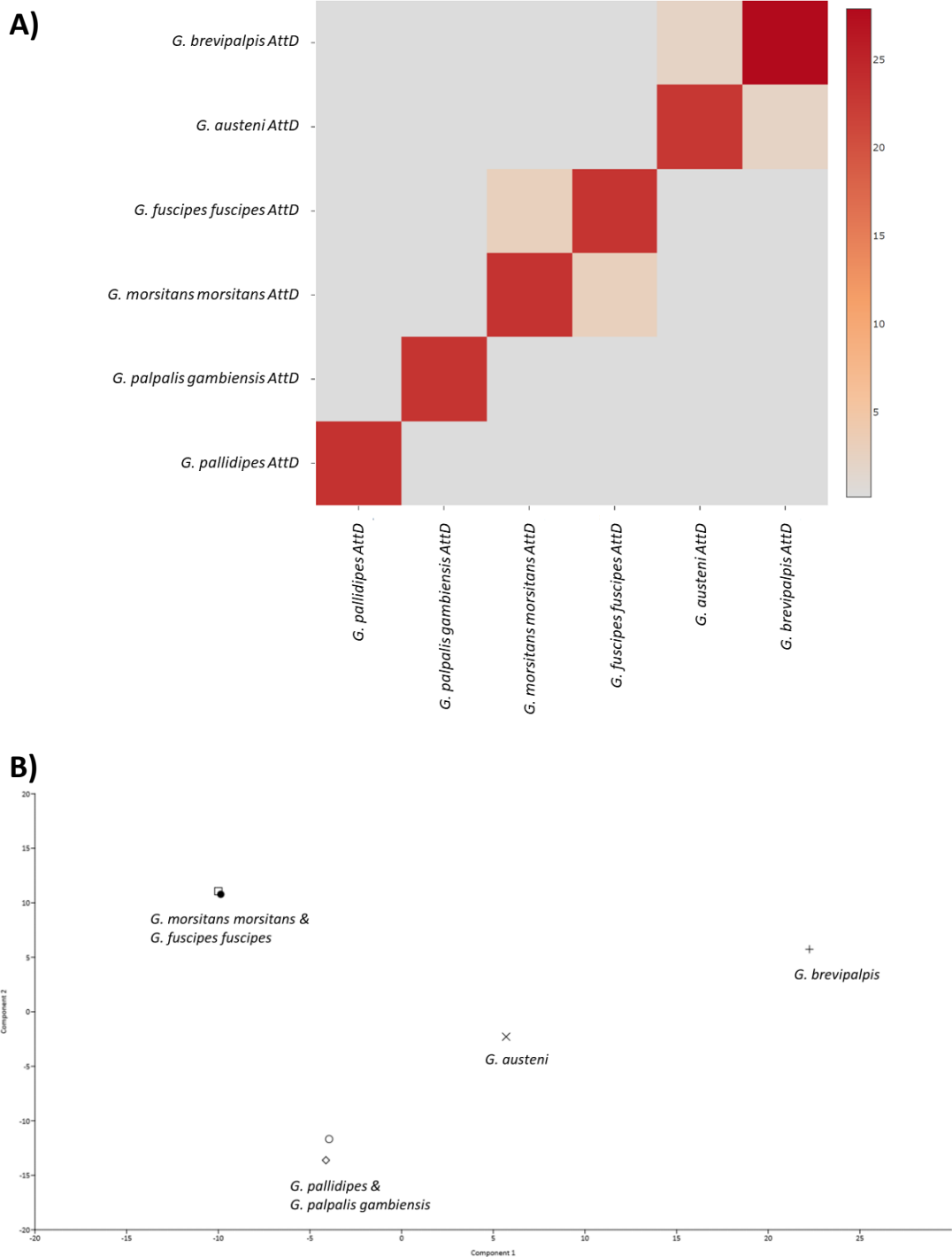


Figure 2.44: A) A heatmap comparing the conservation of Def protein structures across all complete predicted *Glossina* Def genes. The heatmap was constructed using DALI server (Holm, 2020). Colour represents the Z-value as estimated by DALI, higher Z-score and conservation are shown by the deep red colouration, while low Z-scores are shown by the grey. B) Principle component analysis (PCA) plot of predicted Def proteins within the *Glossina* genus using the first and second principle components (Eigenvalues: PC1 = 151.469 (27.022 % variance); PC2 = 119.737 (21.361 % variance)). Z-values were calculated, and a matrix produced using DALI (Holm, 2020). PCA analysis was conducted in PAST3 (Hammer *et al.*, 2001). Individual predicted Attacin genes are shown by plots, species is denoted by shape: *G. austeni* = X; *G. brevipalpis* = +; *G. f. fuscipes* = ●; *G. m. morsitans* = □; *G. pallidipes* = ○; *G. palpalis gambiensis* = ◇.

## 2.6: Discussion

An extensive understanding of immune genes within vector genera are vital to establishing a comprehensive understanding of parasite-host interactions. The successful identification of 24 novel attacin orthologues, in addition to the four identified previously by Trappeniers *et al.* (2008), supports the previously published results of Wang *et al.* (2008) and Trappeniers *et al.* (2019). Each of these orthologues was identified within a recognisable attacin cluster as described previously (Wang *et al.*, 2008). Furthermore, a single novel defensin gene was identified within each of the six *Glossina* species examined.

Phylogenetic analysis of the attacin gene family inferred that the overall evolutionary history of both the attacin and defensin genes followed the species diversification of the *Glossina* genus as specified by Dyer *et al.*, (2008). Furthermore, the observed evolutionary divergence between *AttA/B* and *AttD* supports the theory that the expression of *AttD* is regulated by alternative stimuli and that the protein may be functionally different to *AttA* and *AttB* (Hedengren *et al.*, 2000; Wang *et al.*, 2008). Wang *et al.* (2008) commented further that there was a considerable difference in transcription levels between *AttA/B* and *AttD*, with *AttA* and *AttB* being considerably more abundant. This imbalance in transcription may be explained by the presence of three *AttA* homologues within the genome enabling rapid expression of the gene. However, this is not the case for *AttB* and therefore this variation is likely a consequence of differing signalling pathways and stimuli as implied by the absence of pre- and pro-peptide domains (Hedengren *et al.*, 2000; Wang *et al.*, 2008; Trappeniers *et al.*, 2019).

The presence of a clade containing partial genes is to be expected given the differences in sequence length and coding domains between complete and partial attacin genes (Yue *et al.*, 2009). This was almost solely the result of gaps in the Scaffold sequence, between sequenced contigs, which made identification considerably more difficult. However, two factors indicate these genes are expressed. Firstly, each partial gene exhibited a clear, partial attacin C-terminal domain ( $E \leq 4.3e^{-15}$ ), there is no evidence within the literature to suggest that any of the attacin paralogues are expressed as single C-terminal domains. Secondly, the high level of amino acid conservation between genes suggests that these genes are still expressed within in response to immune stimuli

(Dushay *et al.*, 2000). Therefore, it must be concluded that the N-terminal domain is located within the missing sequence data.

Unexpectedly, an additional partial attacin sequences was identified within both *G. f. fuscipes* and *G. brevipalpis*. The location of each of these partial genes suggests that they are not part of the primary cluster structure; while the increased nucleotide variation suggests that these represent pseudogenes. While the presence of pseudo-immune genes is not well documented, a cecropin pseudogene has been identified within the *Drosophila* genome (Imler and Bulet, 2005). The origin of this pseudogene remains undetermined, but may be due to evolutionary silencing of Attacin-C. Despite *AttC* being observed within the *Drosophila* genome (Hedengren *et al.*, 2000), it is notably absent from the *Glossina* genome. The sequence identified within the *G. brevipalpis* contig JFJS01007046 shows no evidence of either a start codon or a N-terminal - possibly a result of recombination or a frame shift in the region (Harrison *et al.*, 2003). Unfortunately, the sequence identified on *G. f. fuscipes* contig. JFJR01000137 is located downstream of a contig. break, with no way to confirm either the presence or absence of the N-terminal. Nevertheless, the high degree of genetic variation within the available sequence data suggests that this sequence is not expressed as an attacin within the *G. f. fuscipes* genome.

As expected, both nucleotide and amino acid variation between *AttA* and *AttB* homologues were considerably lower than the *AttD* homologues (Lazzaro and Clark, 2001; Wang *et al.*, 2008). Though the degree of conservation within the *AttA* genes was unexpected given the high variation exhibited within other insect genera (Gunne *et al.*, 1990; Bulet *et al.*, 1999). Variation within attacin gene isoforms differs, with *AttA* and *AttB* demonstrating a small degree of variation, while *AttD* shows a far greater degree of variation.

This interspecies conservation between *AttA* paralogous is likely to result from the duplication of a single attacin cluster, although the process behind this duplication remains undetermined. Wang *et al.*, (2008) hypothesised that cluster 2 was the result of mariner transposition, having identified a mariner transposase gene ~200 bp down stream of cluster 2. Alternative this conservation could be the result concerted evolution within the *Glossina AttA* gene. Concerted evolution results in a lower level of

intraspecies variation within a gene family, compared to interspecies variation of the same gene family (Liao, 1999). Therefore, the reduced interspecies variation within the *AttA* gene sequences could result from the large number of highly conserved sequence within each species. Furthermore, this high level of conservation could result from the process of coevolution, and the ongoing arms race between *Glossina* sp. and *Trypanosoma* sp. Given the refractory nature of *Glossina* to trypanosomal infection it is likely that AMPs involved in the elimination of the parasite are subject to high levels of purifying selection to maintain effectivity against pathogens. This in turn, supports the previous hypothesis that the increased variation within *AttD* is a consequence of differing stimuli and targets requiring a greater degree of genetic variation (Anderson and May, 1982; Hedengren *et al.*, 2000; Wang *et al.*, 2008; Feeney *et al.*, 2012).

Attacin paralogues all illustrated the same concave structure to that produced using AlphaFold V2 (Jumper *et al.*, 2021). However, though a relatively high degree of conservation was observed within each Attacin family, none of them was supported by significant C-scores or TM values (Zhang and Skolnick, 2005; Yang and Zhang, 2015). All previous literature has described Attacin as forming random coils, allowing for an increase in nucleotide diversity without compromising the functionality (Gunne *et al.*, 1990; Bulet *et al.*, 1999). Furthermore, studies into related Glycine-rich immune gene domains, such as that observed in dipterin, indicate a similar random structure (Cudic *et al.*, 1999); while a previous study into the structural-functional relationship of immune genes indicated that  $\beta$ -sheets were generally stabilised by one or two disulphide bonds, which were absent from the predicted attacin structure (Hwang and Vogel, 1998).

As the exact mode of action of attacin is yet to be determined, this structure could offer an insight. Bulet *et al.* (1999), suggested that attacins utilise a similar mode of action to gloverin, increasing permeability by inhibiting protein synthesis. Under this presumption, attacins must be able to cross the cell membrane without causing terminal disruption, and therefore may exhibited similar structural characteristics to Cell-Penetrating Peptides (CPPs) (Nicolas, 2009; Torrent *et al.*, 2012). Interestingly, the structural flexibility of the CPP Penetratin has been well documented, varying between an unfolded coil to  $\beta$ -sheet in the presence of negatively charged phospholipids (Magzoub *et al.*, 2001, 2002; Su *et al.*, 2008; Eiríkisdóttir *et al.*, 2010). Consequently, this

could explain the contrasting descriptions of attacin protein structure as random coil (Gunne *et al.*, 1990) and  $\beta$ -sheet.

The observation that predicted *Def* genes followed a similar evolution pathway to speciation events of the *Glossina* is supported strongly by the PCA results, illustrating the separation of the Morsitans, Palpalis and Fusca groups. However, the low range of PC values (PC1 range = 0.46; PC2 range = 0.2) indicates a high level of conservation within the genes. A high level of conservation within the defensin gene family has been described previously in several Insecta species (Varkey *et al.*, 2006; Altincicek and Vilcinskas, 2007; Wiesner and Vilcinskas, 2010).

While the overall structure of the predicted genes varied between species, with the exception of *G. brevipalpis*, all predicted defensin genes illustrated a single exon sequence containing the signalling peptide domain and mature protein region. The structure of dipteran defensin genes is poorly described in the literature, however, similarities between the *Drosophila* defensin and the predicted *G. m. morsitans* and *G. p. gambiensis* orthologues suggest that a single exon gene, encoding a 87 - 92 residue protein is the most likely structure of the defensin genes (Dimarcq *et al.*, 1994). The presence of a single gene, rather than a gene family of tandem repeats, suggests that the defensin gene is maintained through high Darwinian evolution, minimising deleterious polymorphism and promoting advantageous variations through natural selection (Sackton *et al.*, 2007). As with all AMPs this selective pressure could be driven by the ongoing coevolution of pathogens and hosts (Anderson and May, 1982).

Given the widespread conservation of the Def protein across arthropod families, a high level of conservation was expected within the *Glossina* genus (Bonmatin *et al.*, 1992a; Bonmatin *et al.*, 1992b; Cornet *et al.*, 1995; Yi *et al.*, 2014). The Def structure can be characterised by the presence of an  $\alpha$ -helix and anti-parallel  $\beta$ -sheet C-terminal, stabilised by three disulphide bonds (Hwang and Vogel, 1998). While structure was observed in both *G. m. morsitans* and *G. p. gambiensis* predicted Def protein structures, it was absent in all other species, the C-terminal  $\alpha$ -helix was exhibited by all species though the subsequent anti-parallel  $\beta$ -sheets were absent. Rather intriguingly, the initial predicted structure of *G. austeni* Def exhibited these anti-parallel  $\beta$ -sheets, those visualisation within PyMOL suggested that they were absent. This is likely a result of

borderline values in the processing of the PDB file, though confidence in these predictions is not high, as all exhibited low C-scores and TM-values (Zhang and Skolnick, 2005; Yang and Zhang, 2015). The absence of this critical structure was surprising given the observed amino acid conservation at within the *Def* transcripts. Given the thoroughly documented conservation of Def protein structure (Bonmatin *et al.*, 1992a; Bonmatin *et al.*, 1992b; Cornet *et al.*, 1995; Yi *et al.*, 2014), it is unlikely that this is a true representation of the Def protein structure within *G. pallidipes*, *G. austeni*, *G. f. fuscipes* and *G. brevipalpis*. While experimental data, such as protein crystallisation, would have yielded results that are more accurate this was not feasible in the period of this study.

## 2.7: Conclusion

While considerable further research is required to understand fully the evolution and relationship of AMPs within the *Glossina* genus, this chapter lays the essential foundations for future genomic and evolutionary analysis.

The aim of this chapter was to identify and characterise the attacin clusters and the defensin genes within the available *Glossina* spp, and to investigate the evolutionary history and interspecies variation of these two important AMPs. From the observations above it is possible to confirm the presence of the *Glossina* attacin cluster in all six species examined. A single defensin gene was identified within each of the *Glossina* species, showing a strong affinity to the established speciation of the *Glossina* genus, with high levels of conservation between gene.

The origins of the attacin cluster remains uncertain and requires further research, however, intraspecies variation and population genetic analysis may offer a further indication into the origins of the gene cluster. Unlike, the observed attacin gene cluster, the presence of a single defensin gene suggests that the gene is under high levels of selective pressure driven by the ongoing arms race between parasite and vector. Both of these concepts are explored in further detail in chapters 3 and 4.

### 3: Molecular variation of Attacin-A and Defensin, in relation to trypanosome infection and symbionts within a wild *Glossina morsitans morsitans* population.

#### 3.1: Introduction

Tsetse flies (*Glossina*) have an innate refractory nature to African trypanosomes. While this phenomenon is still not fully understood, it likely results from a combination of genotypic and phenotypic characteristics, as well as the relationship with symbiotic bacteria within the tsetse fly (Akoda *et al.*, 2009) (see Chapter 1 for more detail). Arguably, the most crucial aspect of this refractory nature is the stimulation of the TLR and IMD pathways and the resultant expression of antimicrobial peptides (AMPs) (Caljon *et al.*, 2014). This immune response is a vital aspect of the tsetse haematophagic lifestyle, and has likely evolved to counter the increased microbial populations ingested during a blood meal (Mosser and Edelson, 1984; Ooi *et al.*, 2015). In addition to controlling ingested microorganisms, the IMD and TLR pathways also maintain endosymbiont populations within the flies (Wang *et al.*, 2009). This endosymbiosis plays a fundamental role in the life cycle and development of the tsetse fly and helps to establish the refractory response to trypanosome infection (Symula *et al.*, 2011; Weiss *et al.*, 2012; Sassera *et al.*, 2013). However, studies into the population and evolutionary history of AMPs and endosymbionts within the *Glossina* genus are severely lacking, with previous studies of the population dynamics and evolution of the broader *Glossina* genus focusing primarily on the mitochondrial genome (Leak, 1999).

When considering the evolutionary history of the *Glossina* immune system, the concept of pathogen/parasite-host coevolution must be considered (Anderson and May, 1982). This concept suggests that the close association between parasites and their host is a major driving factor in the evolution of both organisms (Anderson and May, 1982; Feeney *et al.*, 2012). Under this concept, the immune genes responsible for the suppression of pathogen populations would be subject to high levels of selective pressures to maintain the advantage when combating infection (Niaré *et al.*, 2002;

Jiggins and Hurst, 2003; Lehmann *et al.*, 2009). However, the realities of host immune evolution are considerably more complex.

The evolution of dipteran immune genes has been thoroughly investigated within the *Drosophila* genera and, to a lesser extent, the Culicidae family (mosquitoes). Within the *Drosophila* genera, several studies have investigated the presence and impact of selective pressure on immune genes (Date *et al.*, 1998; Lazzaro and Clark, 2001; Jiggins and Hurst, 2003; Lazzaro and Clark, 2003; Chapman *et al.*, 2019 and Hill *et al.*, 2019). Four AMP families: attacins, defensins, cecropins and dipterocins, have been extensively studied with relatively high levels of synonymous variation having been detected in all (Date *et al.*, 1998; Lazzaro and Clark, 2001; Jiggins and Hurst, 2003; Lazzaro and Clark, 2003). Interestingly, two of these studies formed similar conclusions, specifically that it is unlikely the *Drosophila* immune system is driven solely by the basis of pathogen-host coevolution (Jiggins and Hurst, 2003; Lazzaro and Clark, 2003). This was supported further by the later findings of Hill *et al.* (2019), who noted that evolution of immune genes within dipteran species is likely strongly influenced by the specific pathogens faced and the make up of the species immune system.

Studies into the evolution of Culicidae immunity focus primarily on the malarial mosquito *Anopheles gambiae* and the yellow fever mosquito *Aedes aegypti* (Niaré *et al.*, 2002; Little and Cobbe, 2005; Lehmann *et al.*, 2009). Early studies into the coevolution of *An. gambiae* and the malarial parasite *Plasmodium falciparum*, found a high level of neutral alleles responsible for resistance within the *An. gambiae* genome (Niaré *et al.*, 2002). While a later study of both *An. gambiae* and *Ae. aegypti* by Little and Cobbe (2005), indicated the presence of purifying selection among the receptor gene Peptidoglycan Recognition Protein-LB (*PGRPLB*), while the high levels polymorphisms observed within the Thioester-containing protein 3 (*TEP3*) gene suggested effector genes are likely submitted to repeated periods of positive selection (Little and Cobbe, 2005). However, while a higher degree of selective pressure was observed within the Culicidae, it does not appear that coevolution, driven solely by parasite interactions, is entirely responsible for this selective pressure (Lehmann *et al.*, 2009).

The direct arms-race between pathogens and hosts is often the primary focus of evolutionary studies, however, an equally important but often overlooked aspect is the



interaction between pathogens and hosts at a population level. The importance of understanding population level interactions between hosts and pathogens was briefly alluded to by Hill *et al.* (2019), who stated that the rapid evolution of immune genes may vary between species depending on the pathogens encountered.

Interspecies molecular population genetics studies of insect immune genes have been undertaken previously, a comprehensive study of the population genetics of *Drosophila* viral resistance locus *ref(2)P*, illustrated that the site was highly polymorphic, thus deviating from neutrality and under selective pressure (Wayne *et al.*, 1996). Interestingly, a higher evolutionary rate was detected within the rhabdovirus sigma virus susceptible *D. melanogaster* lineage compared to the non-susceptible *D. simulans* lineage (Wayne *et al.*, 1996). While the study conducted by Wayne *et al.* (1996) utilised lab raised fly lineages, wild populations have also exhibited high rates of polymorphisms and selective sweeps at the *ref(2)P* locus suggesting ongoing co-evolutionary interactions between the host and pathogens (Juneja and Lazzaro, 2009). However, whether intraspecies evolution differs as a result of specific pathogen interactions remains to be seen.

Population level variation within the dipteran immune system was further observed in the AMPs cecropin, andropin and dipterin (Clark and Wang, 1997). It was observed that while the AMPs predominantly rejected neutrality in favour of balancing selection, their interpopulation and interspecies variation differed considerably. This was clearly illustrated in the difference between cecropin-B (*cecB*) and andropin; *cecB* exhibited significant differences between populations, however a low degree of interspecies divergence was detected despite a high level of polymorphisms (Clark and Wang, 1997). Andropin on the other hand, showed no significant heterogeneity between populations despite indicating a clear interspecies divergence (Clark and Wang, 1997). This supports the observation that the evolution of AMPs is closely associated with, but not exclusively driven by, the direct interaction of specific pathogens within a subpopulation (Gandon, 2002; Hill *et al.*, 2019).

While the importance of a population level approach to understanding dipteran immune evolution has been stated on several occasions (Travis, 1993; Clark and Wang, 1997), there is little detailed literature regarding parasite-host interactions at a molecular

population level. Studies of spatial patterns of subpopulations and coevolution suggest that sympatric subpopulations are likely to be more compatible than allopatric populations (Gandon, 2002; Woolhouse *et al.*, 2002). However, in reality local adaptation and maladaptation have been observed in the interaction of *An. gambiae* and *P. falciparum*, where two resistant markers were observed to be more effective against either sympatric or allopatric infections (Niaré *et al.*, 2002; Juneja and Lazzaro, 2009).

Symbiont diversity, at both a molecular and population level, has been studied in several arthropod genera, most extensively in aphids and their symbionts *Buchnera* and *Acyrtosiphon* (Funk *et al.*, 2001; Abbot and Moran, 2002; Tsuchida *et al.*, 2002; Swanevelder *et al.*, 2010). Nonetheless, several studies have also been conducted examining the symbiosis between *Glossina* and the bacteria genera *Wigglesworthia* and *Sodalis* (Aksoy, 1995; Aksoy *et al.*, 1997; Geiger *et al.*, 2006; Symula *et al.*, 2011; Rio *et al.*, 2012). The evolution of symbiotic bacteria is hypothesised to follow co-speciation with the host, and indeed this has been observed between *Glossina* and *Wigglesworthia* (Aksoy *et al.*, 1997; Symula *et al.*, 2011), though the recent evolutionary history of symbionts remains unclear.

Genetic diversity within the mitochondrial genome of both American aphid populations and their primary symbiont, *Buchnera*, was generally found to be low, suggesting a high level of gene flow between aphid populations (Abbot and Moran, 2002; Swanevelder *et al.*, 2010). It is interesting to note however, that the ant symbiont *Blochmannia floridanus* illustrated elevated levels of genetic variation within non-coding regions indicating a deletion bias within the genome to reduce the genome size of the symbiont (Gómez-Valero *et al.*, 2008). Nucleotide diversity within the *Wigglesworthia glossinidia fuscipes* mitochondrial genome was observed to be approximately three times lower than that of the *G. f. fuscipes*, suggesting that the symbiont is evolving slower than the host (Symula *et al.*, 2011). At a population level, Symula *et al.* (2011) observed further that *W. g. fuscipes* population clearly diverged into two gene lineages corresponding with northern and southern populations. To date however, this phenomenon has not been observed in the *W. glossinidia* strains of any other *Glossina* spp.

### 3.1.1: Aims and Objectives

As detailed above, several studies offer an insight into the evolution of dipteran immune genes, however, there remains a complete absence of literature regarding the evolution of immune genes within the *Glossina* genus. This chapter aims to address this by undertaking an indepth analysis of two important AMPs, Attacin-A (*AttA*) and Defensin (*Def*). Specifically this chapter aims to address objective 3 (see chapter 1, section 1.6): To evaluate the intra-species variation of *AttA* and *Def* in association to symbiont and trypanosome infection within a wild tsetse population.

Despite the low level of variation detected in Attacin-A (see chapter 2) it was selected for this study for three reasons: firstly, the highly elevated expression of *AttA* in both the midgut and fatty tissues of the tsetse during trypanosome infection indicates that the protein is vital to the immune response to parasite infection. Secondly, the presence of three *AttA* homologues within the *Glossina* genome (see chapter 2) suggests the importance of *AttA* and facilitates the high level of expression. Thirdly, while a low level of interspecies variation was observed between *Glossina AttA* orthologues, genetic variation, and the influence of selective pressure at population level remain uninvestigated.

The genetic variation of immune genes within wild tsetse populations was assessed using gDNA extracted from three *G. m. morsitans* subpopulations from Northern Zimbabwe. The evolutionary relationship of these sequences was measured using standard phylogenetic practices, and allele diversity was assessed using haplotypic analysis. Having established the extent of allele diversity, each set of sequences was screened for synonymous and nonsynonymous sites to illustrate the intraspecies nucleotide variation. Recent evolutionary history was estimated using population genetics, while gene flow was measured between the subpopulations and test of neutrality and demographic change were also conducted.

In order to produce a comprehensive study into the evolutionary history and population dynamics of a wild *G. m. morsitans* population of both the mitochondrial and nuclear genomes was assessed. Cytochrome Oxygenase 1 (*COI*) was employed in this study as a mitochondrial comparator for the nucleotide genes, *AttA* and *Def*. Furthermore, the

primarily neutral evolution of *COI* would highlight any divergence from neutrality by the nuclear genes.

Genetic variation within the *W. g. morsitans* endosymbiont population and the association of specific tsetse and *W. glossinidia* haplotypes was assessed. Haplotype analysis illustrated the allele diversity within the *W. g. morsitans* population, while gene flow and tests for neutrality illustrated the relationship between symbiont and host. The association of *G. m. morsitans* and *W. g. morsitans* genetic variation was assessed by comparing the exhibited haplotypes of *W. g. morsitans 16S* rRNA gene and specific *AttA* and *Def* haplotypes.

Finally, the relationship between genetic variation and trypanosome infection was also assessed. The presence of *Trypanosoma* spp. within the *G. m. morsitans* samples was established using PCR, the association to genetic variation was assessed by simply observing the percentage of infected and uninfected samples within each immune gene and *W. g. morsitans 16S* haplotype. Finally, the relationship between immune gene variation, *W. g. morsitans 16S* variation and infection was assessed to establish any association between the three aspects of the triplet.

## 3.2: Materials and Methods

### 3.2.1: Tsetse samples collection

The 63 *G. m. morsitans* specimens used in this study were provided by Prof. Stephen Torr from the Liverpool School of Tropical Medicine, United Kingdom. These had been collected from three trapping sites in North-Eastern Zimbabwe during May 2015. These sites are the Rekomitjie Research Station on the banks of the Ruckomeshi River ( $16^{\circ}08'19''$  S,  $29^{\circ}24'03''$  E) (n=20), the Nykasanga area of the Hurungwe Safari Area ( $16^{\circ}09'05''$  S,  $29^{\circ}06'48''$  E) (n=14) and the village of Makuti ( $16^{\circ}18'51''$  S,  $29^{\circ}15'03''$  E) (n=29).



Figure 3.1: A map of Zimbabwe. The red shaded area shows the location of the collection sites in Northern Zimbabwe. The insert shows a detailed mapped of the collection sites showing location and distance between each site (Km). N = Nykasanga, R = Rekomitjie, M = Makuti. Both maps were generated using Google maps online software.

Collection had been undertaken using three types of traps: Epsilon (an insecticide treated, odour trap), vehicle electric trap (VET) and Fly-round (both traps use a moving target to attract flies, VET traps use an electrical current to stun flies, while the fly round trap utilise a net to capture following flies). VET traps were employed to capture the

majority of samples (44 of the 49) collected in Rekomitjie and Makuti, with the remaining five samples collected using Epsilon traps with acetone, 1-octen-3-ol, 4-methylphenol and 3-n-propylphenol (AOP) attractants (Hall *et al.*, 1990). All samples from the Nykasanga area were collected using Fly-round traps. Tsetse flies are considered a pest species and disease vectors and are not endangered therefore no ethical clearance was required for this study (Ethics application ID: ETH1718-0190).

### 3.2.2: gDNA extraction of *Glossina* specimens

Whole genomic DNA (gDNA) extraction of individual *Glossina* specimens was undertaken employing a QIAGEN DNeasy Blood and Tissue kit (QIAGEN, UK) following the manufacturer's protocol with two minor modifications. Firstly, incubation at 56 °C was conducted for 18 hours rather than the recommended four hours to ensure full lysis of cells. Secondly, 100 µl, rather than 200 µl, of elution buffer (Buffer AE) were added to the spin column membrane prior to centrifugation. This was repeated, for a total yield of 200 µl of eluted gDNA. The concentration and purity of extracted gDNA was assessed using a Nanodrop 1000 Spectrophotometer (Thermo Scientific). Concentration was measured in ng/ml to ensure a sufficient quantity of DNA was present for PCR amplification. The purity was assessed using 260/280 and 260/230 values, where a 260/280 value > 1.8 and a 260/230 value between 2 and 2.2 was indicative of pure DNA.

### 3.2.3: Primer design and Polymerase Chain Reaction

Polymerase Chain Reaction (PCR) was conducted to amplify the genes of interest prior to sequencing. Where no previous publications described suitable primers for this study, primers were designed using NCBI Primer BLAST (Ye *et al.*, 2012). *Glossina* genome data on VectorBase (Giraldo-Calderón *et al.*, 2015) and NCBI (Sayers *et al.*, 2009) were mined for relevant sequences to provide a template for primer design. Table 3.1 contains information on all the primers used in this study, primer working stocks were made from a 1:10 dilution of master stocks at 100 pMol/ml, to reach a working concentration of 10 µM/ml.

Amplification of target gene fragments by PCR was conducted using 12.5 µl of DreamTaq™ PCR master mix (2X DreamTaq buffer, 0.4 mM of each dNTP, 4mM MgCl<sub>2</sub>)

(Thermo Scientific, UK), 1 µl each of gene-specific forward and reverse primers and 1-5 µl of gDNA, with a final reaction volume of 25 µl made up with PCR grade water.

In order to establish the infection status of each sample trypanosome identification was conducted using Nested PCR as detailed by Adams *et al.* (2006). The first amplification was conducted using 12.5 µl of DreamTaq™ PCR master mix (Thermo Scientific, UK), 1 µl of primers TYP3 and TYP4, and 2 µl of gDNA. The reaction was made up to 25 µl using PCR grade water. The second amplification used primers TYP1 and TYP2 and 5 µl of the previous PCR product as replacements for primers TYP3 and TYP4 and gDNA. In the case of negative controls gDNA was absent and the appropriate volume of PCR grade water was added. PCRs were run on either a Prime (Techne) thermal cycler or MJ mini personal thermal cycler (Bio-Rad) and cycling conditions are shown in Table 3.2.

Following completion of the PCR program, gel electrophoresis was used to determine the success of amplification. 5 µl of all PCR products was mixed with 6x Gel loading dye (Thermo Scientific, UK) and GelRed™ Stain (Cambridge Bioscience, UK) before being run on an agarose gel. Gel percentages and DNA ladders differed depending on the fragment size; a 1% agarose gel using a 1Kb Hyperladder (Bioline, UK) was used for *G. m. morsitans AttA* and *COI*, and *W. g. morsitans 16S* samples. For *Def*, a 1.5% gel using a 100bp Hyperladder (Bioline, UK) was used, and a 2% gel with 100bp Hyperladder (Bioline, UK) was utilised for *Trypanosoma ITS* samples. All gels, excluding *Trypanosoma ITS*, were run at 90 V for 45 minutes prior to visualisation on an UV transilluminator. While *Trypanosoma ITS* samples were run for 50 minutes at 130 V before visualisation (see Supplementary Figures 1-5 in Appendix 4 for gel images).

It must be noted that 15 samples utilised in this study were initially examined by Akuzike Kalizang'oma in fulfilment of his MSc dissertation. The data he generated, using the methods described within this chapter, produced 15 *AttA* sequences and screen these same samples for *Wigglesworthia* and *Trypanosoma* infections. The gels produced by Akuzike Kalizang'oma can be seen in Supplementary Figures 1, 4 and 5 (see Appendix 4).

Table 3.1: The target gene, given primer name, nucleotide sequence, length, melting temperature (Tm), G-C percentage and expected fragment size of each primer used in this study. \* = Trypanosome identification by ITS amplification was undertaken using nested PCR, as such multiple bands are produced each specific to a species.

Target gene	Name	Sequence (5'-3')	Length (bp)	Tm (°C)	GC%	Fragment size	Designed by
<b>Attacin-A</b>	AttA-F2	TGTTTAAGCGTCGTTCAAGT	20	N/A	40.00	674bp	This work
<b>Attacin-A</b>	AttA-R2	CTTAATCCGAAATACAAGGCT	21	N/A	38.10		This work
<b>Defensin</b>	NDEF-F3	ACACTCAGCGTCGAAAGTG	19	58.11	52.63	380bp	This work
<b>Defensin</b>	NDEF-R3	TAAAAAGTTCCACATTCAAGTCTTC	25	56.24	32.00		This work
<b>Wigglesworthia 16S</b>	170F	ATAAAGCCTTGCGTTT	16	49.1	37.50	~800bp	Chen <i>et al.</i> , 1999
<b>Wigglesworthia 16S</b>	1227R	CCATTGTAGCACGTGT	16	49.2	50.00		Chen <i>et al.</i> , 1999
<b>Trypanosoma ITS</b>	TYP1	AAGCCAAGTCATCCATCG	18	N/A	N/A	N/A	Adams <i>et al.</i> , 2006
<b>Trypanosoma ITS</b>	TYP2	TAGAGGAGGCAAAAAG	15	N/A	N/A		Adams <i>et al.</i> , 2006
<b>Trypanosoma ITS</b>	TYP3	TGCAATTGGTCGCGC	18	N/A	N/A	Multiple*	Adams <i>et al.</i> , 2006
<b>Trypanosoma ITS</b>	TYP4	CTTTGCTGCGTTCTT	15	N/A	N/A		Adams <i>et al.</i> , 2006
<b>Cytochrome oxygenase 1</b>	CI-J-2195	TTGATTTTTTGGTCATCCAGAAGT	24	N/A	N/A	~600bp	Simon <i>et al.</i> , 1994
<b>Cytochrome oxygenase 1</b>	CULR	TGAAGCTTAAATTCATTGCACTAATC	26	N/A	N/A		Dyer <i>et al.</i> , 2008



Table 3.2: The cycling conditions for all PCR amplifications giving the time and temperature through each stage. <sup>1/2</sup> = both reactions of the trypanosome identification protocol utilised these cycling conditions. \* = Unsuccessful amplifications of the *Wigglesworthia 16S* were repeated using increased gDNA volumes and these cycling conditions.

Target gene	Initial Denaturation (Time/temperature)	Denaturation (Time/temperature)	Annealing (Time/temperature)	Extension (Time/temperature)	Final extension (Time/temperature)
			<b>X35</b>		
<b>Attacin-A</b>	5 minutes / 94 °C	30 seconds / 94 °C	30 seconds / 53 °C	45 seconds / 72 °C	10 minutes / 72 °C
<b>Defensin</b>	5 minutes / 94 °C	30 seconds / 94 °C	30 seconds / 52 °C	45 seconds / 72 °C	10 minutes / 72 °C
<b><i>Wigglesworthia</i></b>	5 minutes / 94 °C	1 minute / 94 °C	1 minutes / 56 °C	1 minutes / 72 °C	10 minutes / 72 °C
<b><i>Trypanosoma ITS</i><sup>1/2</sup></b>	1 minute / 94 °C	1 minute / 94 °C	1 minute / 54 °C	30 seconds / 72 °C	5 minutes / 72 °C
<b>Cytochrome oxygenase 1</b>	5 minutes / 95 °C	1 minutes / 93 °C	1 minutes / 55 °C	2 minutes / 72 °C	7 minutes / 72 °C
			<b>X45</b>		
<b><i>Wigglesworthia</i>*</b>	5 minutes / 94 °C	1 minute / 94 °C	2 minutes / 56 °C	2 minutes / 72 °C	10 minutes / 72 °C

### 3.2.4: Gel extraction of trypanosome ITS bands

As detailed above (section 3.2.3), the amplification of *Trypanosoma* species ITS regions was undertaken using primers published previously by Adams *et al.* (2006). This nested PCR approach produced an ITS amplicon specific to each *Trypanosoma* spp. that could be differentiated by size (Adams *et al.*, 2006). To confirm the successful amplification of trypanosomal ITS regions, nine bands of varying sizes were cut from the 2% agarose gels following visualization. Gel purification was used to extract the DNA amplicon from the gel using a QIAquick® Gel Extraction Kit (QIAGEN, UK) and PCR Purification combo Kit (Thermo Fisher, UK), in accordance with the manufacturer's protocol. This enabled each amplicon to be sequenced (see section 3.2.5) to confirm the trypanosome species indicated by gel electrophoresis where possible.

### 3.2.5: Sanger sequencing and sequence analysis

Nucleotide sequences of the PCR amplicons were produced using Sanger Sequencing (Sanger *et al.*, 1977). The gel extraction products were also submitted for sequencing to verify the presence of trypanosome infection. This was conducted by the DNA Sequencing Facility at The Natural History Museum, London. *Glossina AttA*, *Def*, *COI*, *Wigglesworthia 16S* and trypanosome ITS PCR primers were utilised, with Fluorescent Dye Terminator Sequencing Kits (Applied Biosystems™), and then run on an Applied Biosystems™ 3730XL automated sequencer.

Analysis of resultant nucleotide sequences was undertaken using SnapGene software (from GSL Biotech; available at [snapgene.com](http://snapgene.com)) to visualise forward and reverse chromatographs. These were then aligned using the MUSCLE (Multiple Sequence Comparison by Log-Expectation) sequence alignment tool (Madeira *et al.*, 2019) and a contiguous sequence was then constructed and edited, removing intron sequences from *AttA* sequences, and removing non-coding sequences following the stop coding in *Def* (Okonechnikov *et al.*, 2012). Each sequence was then submitted to a blast search against the VectorBase *Glossina* data base and NCBI blast search to verify the gene fragments produced from the PCR reactions.

### 3.2.6: Intra-species phylogenetic analysis

Phylogenetic analysis was conducted to estimate the evolutionary relationship of each gene within the sample population. Multiple sequence alignments were performed using MUSCLE (Madeira *et al.*, 2019) and these were then edited in UGENE (Okonechnikov *et al.*, 2012) to ensure equal sequence length (resultant sequences lengths; *COI* = 782bp; *AttA* = 480bp and *Def* = 246bp). Phylogenetic analysis was conducted using MEGAX (Kumar *et al.*, 2018), Neighbour-joining trees were constructed using the Jukes-Cantor model with 1000 bootstrap replicates. Maximum-likelihood trees were constructed using the model of best fit, estimated using the MEGAX 'Find Best NDA/Protein Models' function. This indicated that, Tamura 3-parameter model best suited *COI*, Jukes-Cantor (JC) was the best model for use with the *AttA* alignment and Juke Cantor + Gamma distribution with Invariant sites (JC+G+I) for *Def*, and 1000 bootstrap replicates were used once again. This provided two comparable insights into the nature of the relationship between *G. m. morsitans* samples within the subpopulations, while additionally indicating variation between mitochondrial and nuclear genes.

### 3.2.7: Haplotype analysis

Further to phylogenetic analysis, the haplotype variation of each gene was analysed to provide a further understanding of genetic and allele variation. Haplotype data files were generated in DnaSP (version 6) (Rozas *et al.*, 2017), this identified the number of haplotypes (h), the variation between those identified, and which samples exhibited each haplotype within the sample population. The frequencies of the identified haplotypes were then plotted by geographical loci in Microsoft Excel to give an indication of haplotype frequency within each collection site. PopART (Leigh and Bryant, 2015) was used to construct TCS haplotype networks and perform simple AMOVA analyses, illustrating the relationship between identified haplotypes within the geographical distribution.

### 3.2.8: Intra-species nucleotide variation analysis

Having predicted the evolutionary and haplotype variation within each gene, the sites of DNA polymorphisms and nucleotide variation at those points was also investigated.

This gave an insight into the location and degree of variation within each of the genes of interest. Nucleotide variation ( $\pi$ ) analysis was calculated using the 'DNA polymorphism' function in DnaSP (version 6) (Rozas *et al.*, 2017). The region of analysis was set between 103 and 582 base pairs for *AttA* and between 16 and 261 base pairs for DEF, codon specific sliding window analysis was conducted (window size = 3; step size = 3). Nucleotide variation ( $\pi$ ) is estimated using the following equation 1 in Appendix 2 (Nei, 1987, equation 10.5 or 10.6; Nei and Miller, 1990, equation 1).

### 3.2.9: dN/dS: Synonymous vs non-synonymous variation

Having established the presence of DNA polymorphisms within the gene fragments, the nature of these mutations was also established in order to predict the potential functional and structural impacts on the protein synthesis. The presence of synonymous (dS) and non-synonymous (dN) mutations was accessed in DnaSP (version 6) (Rozas *et al.*, 2017) using the 'Polymorphism and Divergence' function. The region of analysis was set to the fragment size using the same parameters as described in 3.2.8. Inter-specific population analysis was conducted between the geographical collection loci and either 'synonymous only' or 'non-synonymous only' changes were considered. Pi ( $\pi$ ) values were calculated using equation 1 (Nei, 1987, equation 10.5 or 10.6; Nei and Miller, 1990, equation 1). dN/dS ratios (also referred to as Ka/Ks) were calculated using the 'Pi(a)/Pi(s) and Ka/Ks ratios' function, with all parameters remaining the same as above.

### 3.2.10: Gene flow analysis

Given the severe lack of literature concerning wild tsetse populations, population genetic analysis was conducted to provide a novel insight into the *G. m. morsitans* population. The 'Gene Flow and Genetic Differentiation' analysis conducted in DnaSP (version 6) (Rozas *et al.*, 2017) measures the extent of DNA divergence among populations in order to estimate the average gene flow. The coding region was set, and a permutation test was conducted with 1000 replicates (Pseudorandom Number Seed was randomly generated by DnaSP). Values for the haplotype and nucleotide statistics (Hs and Ks, respectively), the fixation index (Fst), the average number of nucleotide substitutions (Dxy) and the net nucleotide substitution per site between populations (DA) were produced as described in equations 3-6 (Appendix 2).

A Mantel-test was conducted using PAST3 (Hammer *et al.*, 2001), geographical distance between collection locations (section 3.2.1) was compared to  $F_{st}$  between sites.

### 3.2.11: Demographic change and test for neutrality: Pairwise mismatch, Tajima's D, Fu's $F_s$ and Coalescent Simulation

The impacts of population genetics on genetic variation were considered by examining the demographic change and neutrality of the genes. Tests of neutrality, Tajima's D (Tajima, 1989) and Fu's  $F_s$  (Fu, 1997), were performed to indicate the presence of selection and indicate population growth. While pairwise mismatch (Watterson 1975; Slatkin and Hudson 1991; Rogers and Harpending 1992) and raggedness ( $r$ ) (Harpending, 1994) analysis were undertaken to estimate any recent or historical population expansion events. This was conducted in DnaSP (version 6) (Rozas *et al.*, 2017), Pairwise mismatch was conducted using the 'Population size change' function using a constant population size model. As before the region to be analysed was set to fragment parameters described in section 3.2.6. The equations for the above statistics can be found in Appendix 2, equations 7 – 10. Raggedness ( $r$ ) was also calculated using 'The Coalescent Simulations (DnaSP V5)' with the assumption of free recombination using DnaSP V6 (Rozas *et al.*, 2017). Theta per gene was calculated using 'DNA polymorphism' as described in 3.2.8, with 95% confidence intervals.

### 3.2.12: Recombination analysis

As pairwise mismatch analysis can only be applicable in a population with no recombination between genetic sites, both *AttA* and *Def* were screened for recombination. Genetic recombination was assessed to evaluate population growth by observing new allele combinations within populations and to further illustrate any inferred selection. This was estimated in DnaSP (V6) (Rozas *et al.*, 2017) using the equation 11 (Appendix 2).

The 'Recombination' function within DnaSP (V6) calculates the number of recombinant offspring ( $R$ ) (as in equation 11) and the minimum number of recombination events ( $R_m$ ), as described by Hudson and Kaplan (1985) (Appendix 2). Values were generated using the 'Recombination' option in the analysis menu, and the coding regions were set as described in 3.2.13.

Recombination was assessed further using the Genetic Algorithm for Recombination Detection (GARD) (Kosakovsky *et al.*, 2006) tool on the Datamonkey webserver (Sergei *et al.*, 2005; Weaver *et al.*, 2018). GARD screens multiple sequences alignments for the presence of putative recombination breaking points and analysing them over several phylogenetic trees. Furthermore, GARD can be used to screen recombinant sequences for positive selection (Kosakovsky *et al.*, 2006).

### 3.2.13: *Wigglesworthia* haplotype variation and population genetics

Given the obligatory nature of symbiosis between *W. glossinidia* and *Glossina* spp. the *Wigglesworthia 16S* gene was submitted to the same population genetic analysis. PCR amplification and sequencing of *W. g. morsitans 16S* rRNA gene was conducted as described in sections 3.2.3 to 3.2.5. Amplification was successful in 34 of the 63 (53.97%) *G. m. morsitans* samples. Eleven samples were successfully amplified from both the Nykasanga and Rekomitjie collection sites, and 12 sequences were obtained in samples from Makuti. This equates to 78.57% of Nykasanga samples, 55% from Rekomitjie and only 42.86% of the Makuti samples. A multiple sequence alignment was conducted as described in section 3.2.5 with a final sequence length of 1180bp. In order to assess the relationship between *G. m. morsitans* and *W. g. morsitans* the same analytical methods were used, haplotype analysis was conducted as described in section 3.2.7, while gene flow, Tajima's D, Fu's Fs and demographic change were all assessed as described above in section 3.2.10 and 3.2.11.

### 3.2.14: Association of AMP and symbiont nucleotide variation

The association between *AttA* and *Def* nucleotide variation and that observed within *W. g. morsitans 16S* was assessed using a standard Pairwise-distance analysis in MEGAX (Kumar *et al.*, 2018). The *P*-distance between tsetse AMP sample was calculated and plotted against the corresponding *W. g. morsitans 16S* *P*-distance to illustrate the relationship between the two, a mantel-test was also conducted as described in section 3.2.10. This was undertaken in order to provide an insight into the influence of endosymbiosis on genetic variation.

### 3.3: Results

#### 3.3.1: Intra-species genetic variation and population genetics of wild *Glossina morsitans morsitans*

Intra-species analysis was conducted on all successfully sequenced samples of wild *G. m. morsitans* (*COI* n = 63; *Def* n = 62; *AttA* n = 51) to assess genetic variation within and between the target genes in relation to geographic distribution.

##### 3.3.1i: Phylogenetic analysis

While no clear relationship was observed between gene variation and geographical location, both nuclear genes exhibited an elevated level of nucleotide variation compared to the mitochondrial *COI* gene. Both the Neighbour-Joining and Maximum-likelihood methods illustrated an identical topology within the *COI* gene tree, consisting of two primary clades (I and II) divided into two subclades (a and b) (Fig. 3.2). Both main clades I and II contain a similar number of samples, with Clade II being slightly larger containing 33 of the 63 successfully sequenced samples (52.38 %). However, the bootstrap values do not strongly support the illustrated topology with values ranging between 61 - 64 % in the Maximum Likelihood tree (Fig. 3.4A), and 62 - 66 % for the primary nodes within the Neighbour-Joining tree (Fig. 3.4B), meaning all nodes fall below the 70% robustness threshold.

The two immune genes illustrate differing topologies from each other and *COI*. The two phylogenies of *AttA* (Fig. 3.3) produced similar overall topologies with a smaller monophyletic clade (Clade I), containing just 14 (27. 45%) samples, diverging from the majority of the remaining samples. However, while the Maximum Likelihood method produced one other primary clade (clade II), containing three subclades (a, b, and c) (Fig. 3.3A), the Neighbour-Joining method produced two further clades (II and III) (Fig. 3.3B). In contrast to both *COI* and *AttA*, the observed phylogenies of *Def* show a diverse evolutionary history (Fig. 3.4). The Maximum Likelihood method produced three primary clades (I, II and III), each containing three or four subclades (a - d) (Fig. 3.4A). While the Neighbour-Joining method produced four primary clades (I - IV) (Fig. 3.4B). As observed in *AttA*, this increase in variation is a result of a predicted ancestor

separating Maximum Likelihood clade III.c from the rest of the clade III. Furthermore, there is a clear increase in the number of observed subclades within the Neighbour-Joining phylogeny. Clade I increases from three to five subclades with the addition of an extra sister clade (clade I.a) and a common ancestor separating the samples of Maximum Likelihood clade I.c to form Neighbour-Joining subclades I.d and I.e (Fig. 3.4B). Yet, as with the observed COI phylogenies, the bootstrap values remain low in both Maximum Likelihood and Neighbour-Joining trees, ranging between 19 - 64 %, and 2 - 64 % respectively in the *AttA* phylogenies and 5 - 47 %, and 4 - 63 % respectively in the *Def* phylogenies.



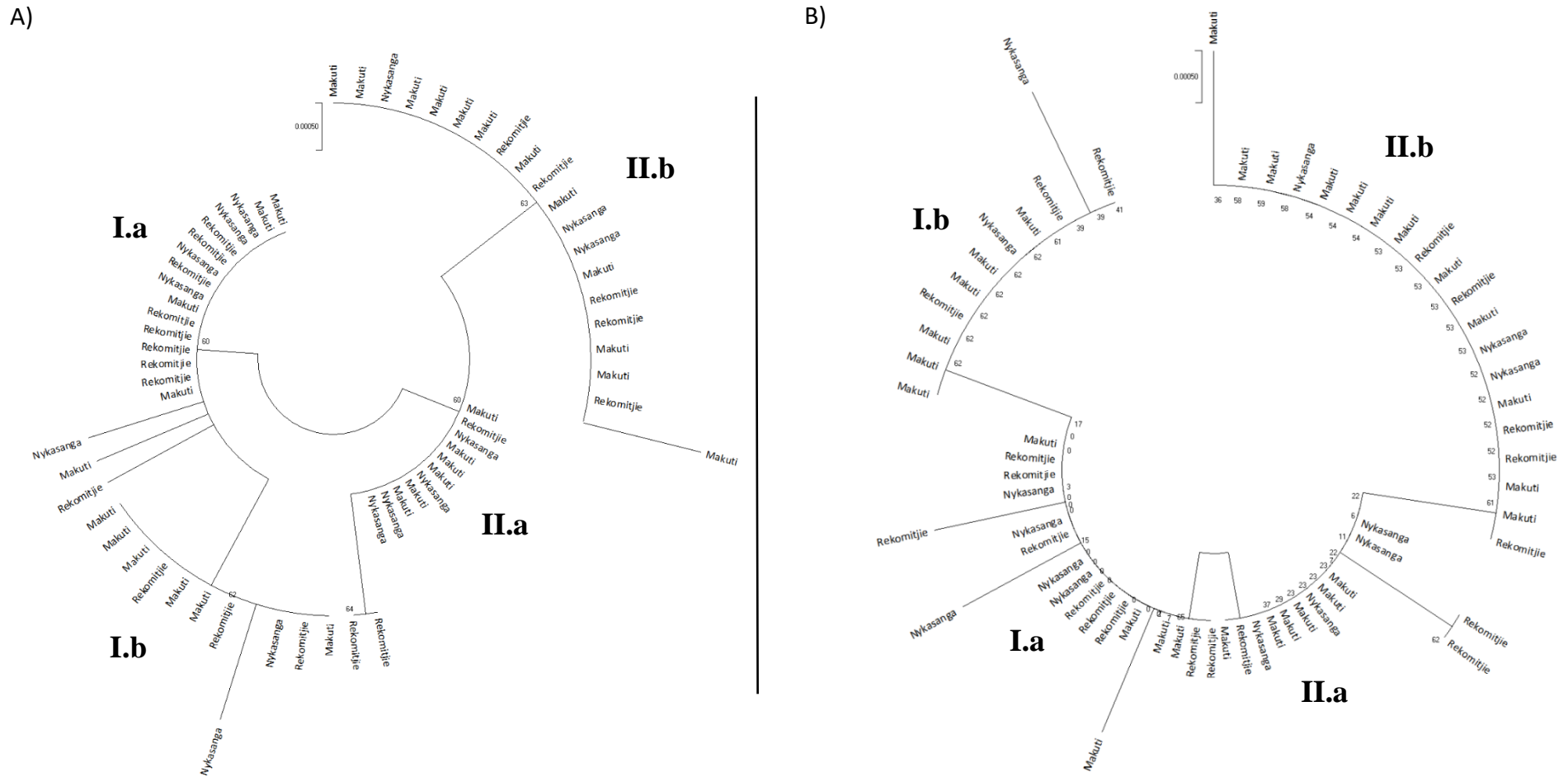


Figure 3.2: Evolutionary analyses of the COI sequence fragments was conducted in MEGA X (Kumar *et al.*, 2018). A) The Maximum Likelihood method, based on the Tamura 3-parameter model (Tamura, 1992) was used. The tree with the highest log likelihood (-1078.27) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbour-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. B) Neighbour-Joining method (Saitou and Nei, 1987), the evolutionary distances were

computed using the Jukes-Cantor method (Jukes and Cantor, 1969) and are in the units of the number of base substitutions per site. The optimal tree is shown with the sum of branch length = 0.01153156. Both trees are drawn to scale, with branch lengths measured in the number of substitutions per site. Codon positions included were 1st+2nd+3rd+Noncoding, while all positions with less than 95% site coverage were eliminated. 1000 bootstrap replicates were used.

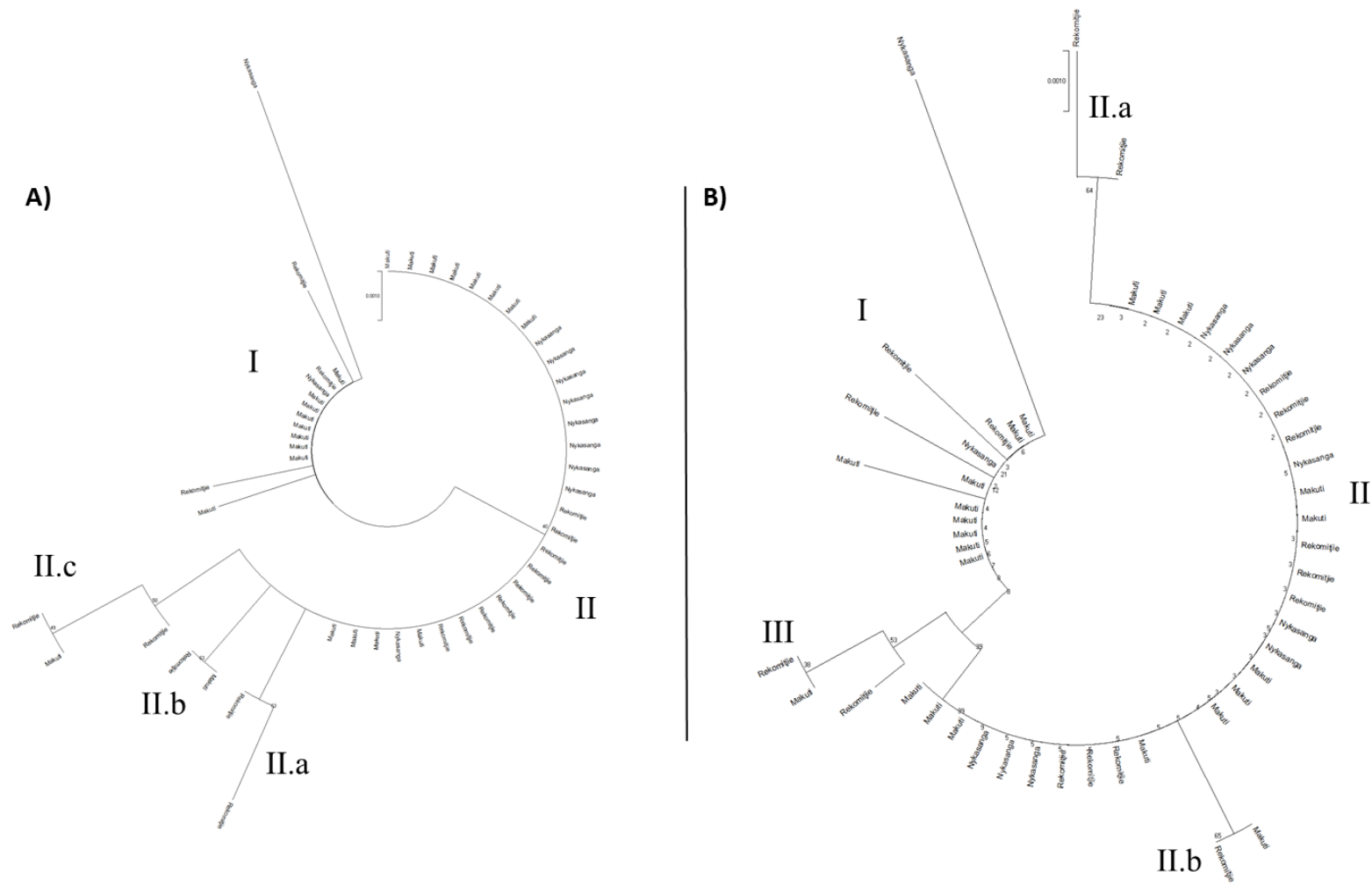


Figure 3.3: Evolutionary analyses of the *AttA* sequence fragments was conducted in MEGA X (Kumar *et al.*, 2018). A) The Maximum Likelihood method based on the Jukes-Cantor model (Jukes and Cantor, 1969). The tree with the highest log is shown (-761.38). B) Neighbour-Joining method (Saitou and Nei, 1987), the evolutionary

distances were computed using the Jukes-Cantor method (Jukes and Cantor, 1969) and are in the units of the number of base substitutions per site. The optimal tree is shown with the sum of branch length = 0.02404575. This analysis involved 51 nucleotide sequences; codon positions included were 1st+2nd+3rd+Noncoding. Both trees are drawn to scale, all positions with less than 95% site coverage were eliminated. 1000 bootstrap replicates were used.

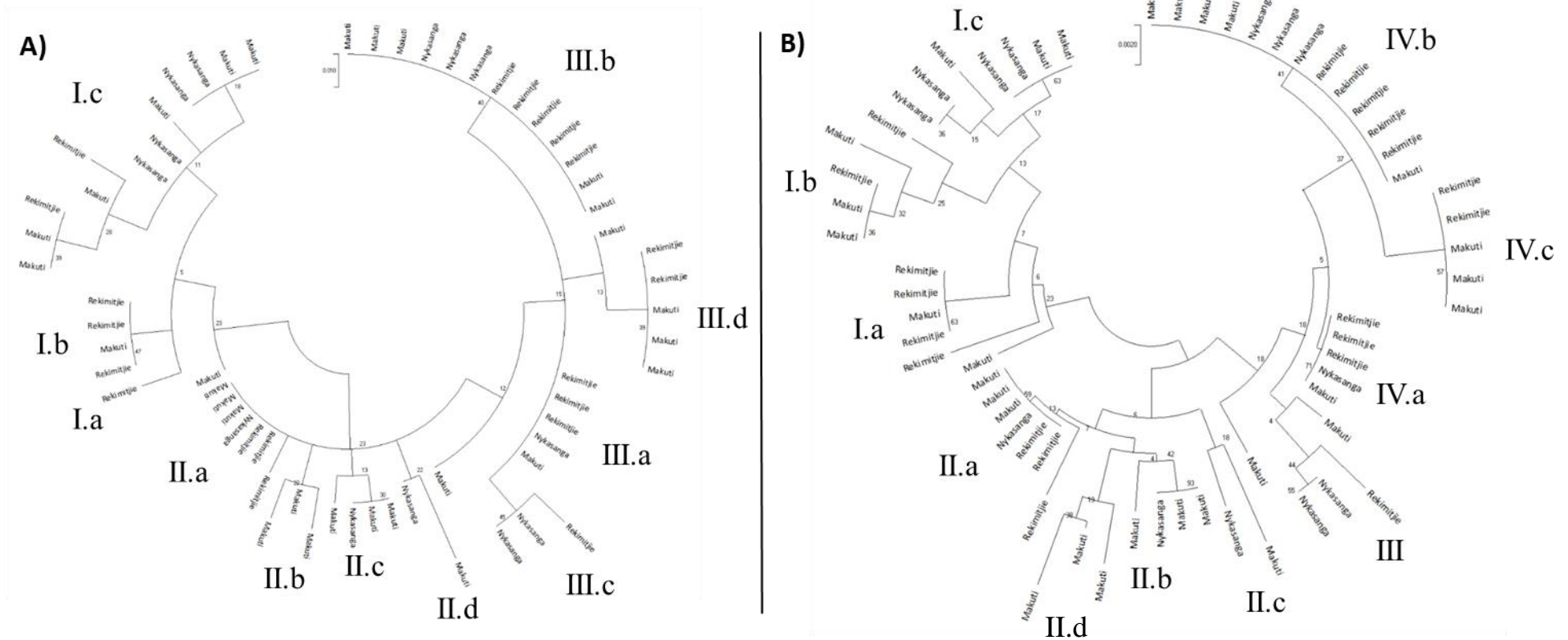


Figure 3.4: Evolutionary analyses of the Def sequence fragments were conducted in MEGAX (Kumar *et al.*, 2018). A) Maximum Likelihood method based on the Jukes-Cantor model (Jukes and Cantor, 1969). The tree with the highest log is shown (-494.39). A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.0500)) and the rate variation model allowed for some sites to be evolutionarily invariable ([+I], 48.37% sites). B) Neighbour-Joining method (Saitou and Nei, 1987), the evolutionary distances were computed using the Jukes-Cantor method (Jukes and Cantor, 1969) and are in the units of the number of base substitutions per site. The optimal tree is shown with the sum of branch length = 0.0942. Both trees are drawn to scale, while all positions with less than 95% site coverage were eliminated. 1000 bootstrap replicates were used.

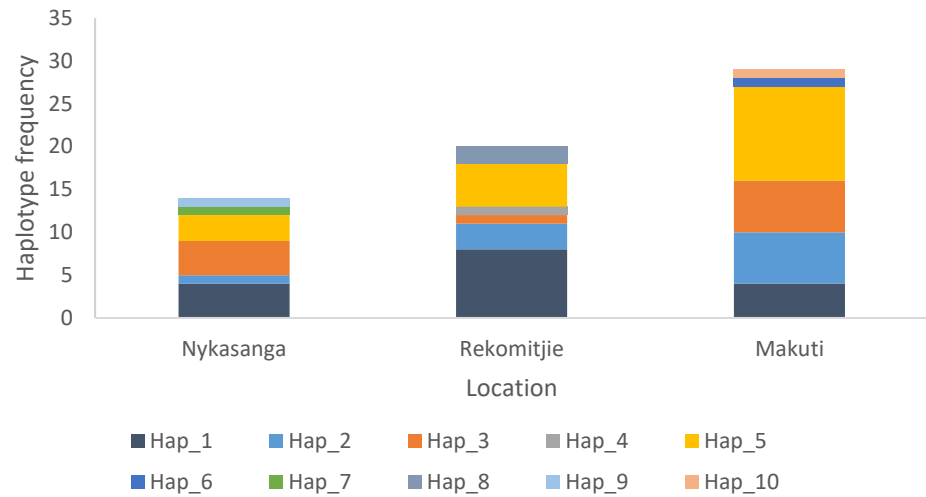
### 3.3.1ii: Haplotype analysis

Haplotype analysis of the *Glossina* genes shows differing degrees of genetic variation within each gene. The COI samples exhibited a total of ten haplotypes with the sample population ( $H = 10$ ;  $Hd = 0.799$ ), of which four (Haps 1-3 and 5) were observed in all three collection localities. These four haplotypes were exhibited by 88.89% of the sample population (56/63 samples), while the remaining seven samples were observed to exhibit six geographic specific haplotypes (Figs 3.5A and 3.6A). There is an even distribution of haplotypes throughout the sample population, with each locality exhibiting six haplotypes (the four common haplotypes and two specific variations) (Fig. 3.5A).

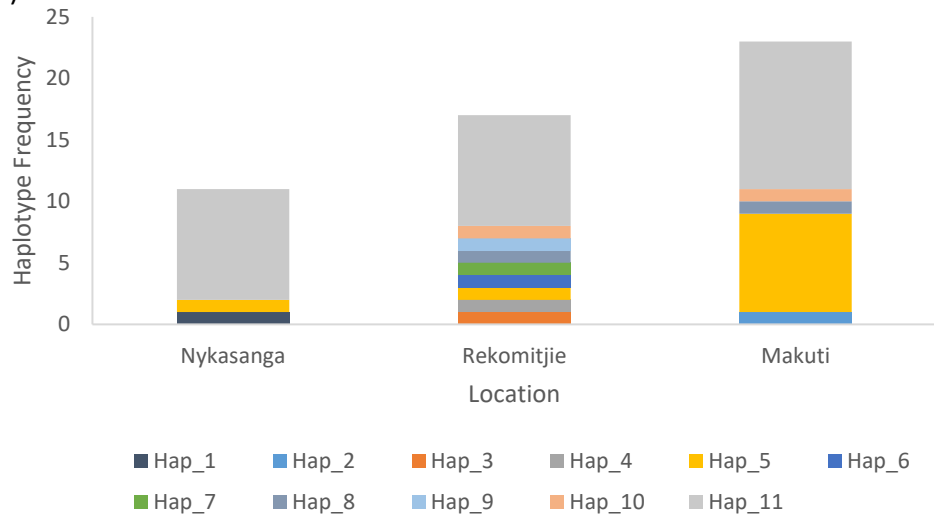
While more haplotypes were observed within the *AttA* samples ( $H = 11$ ), haplotype diversity was observed to be lower than that seen in COI (*AttA*  $Hd = 0.622$ ) (Fig. 3.5B and 3.6B). Two of these (Haps 5 and 11) were observed in all geographical locations, these haplotypes were observed in 78.43% of the overall population across the 3 localities (40/51 samples), of which haplotype 11 accounted for 58.82% of the overall sample population alone. Two haplotypes (Haps 8 and 10) were observed in both Rekomitjie and Makuti, though they were absent from Nykasanga. The remaining seven haplotypes were only observed in single samples, one each from Makuti and Nykasanga, and five from Rekomitjie (Fig. 3.5B and 3.6B). Rekomitjie exhibits the largest haplotype variation, exhibiting nine haplotypes in 17 samples, while Nykasanga and Makuti exhibited three and five haplotypes respectively (Fig. 3.5B).

Finally, *Def* illustrated the greatest level of haplotype variation, exhibiting a total of 25 haplotypes ( $H = 25$ ;  $Hd = 0.93$ ) (Fig. 3.5C and 3.6C). Of which just three (Haps 5, 6 and 17) were observed in all collection subpopulations being exhibited by 38.71% of the sample population (24/62 samples). A further five haplotypes (Haps 7, 16, 18, 20, and 24) were exhibited by sample from two collection points. Haplotypes 7, 18 and 24 were observed in samples from Rekomitjie and Makuti, while haplotypes 16 and 20 were observed in Nykasanga and Makuti, the remaining 19 samples all exhibited location specific haplotypes (Fig. 3.5C and 3.6C). Nine haplotypes were observed within the 14 Nykasanga samples, while Makuti samples exhibit 17 and Rekomitjie ten (Figure 3.5C).

COI)



AttA)



Def)

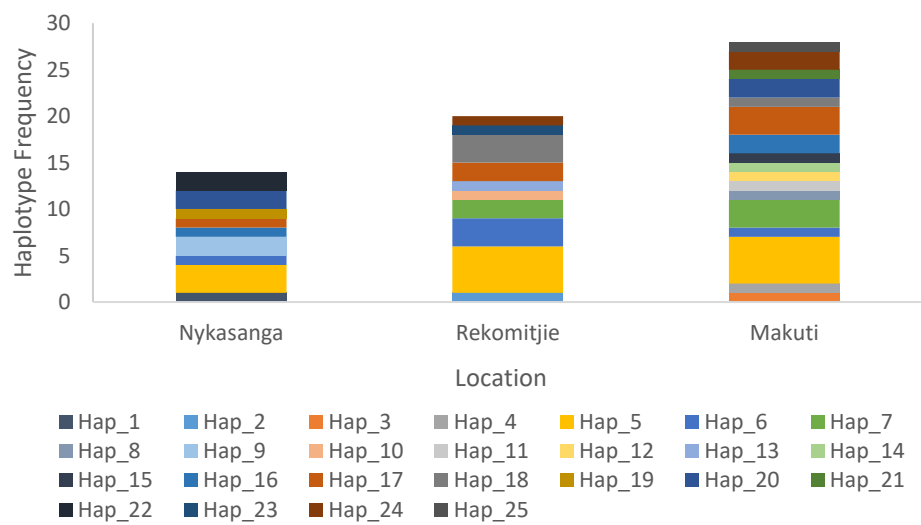


Figure 3.5: The frequency and distribution of genetic haplotypes within each geographical location.

The evolutionary relationship of each haplotype variation can be observed by the construction of TCS networks (Fig. 3.6). Figure 3.6A illustrates the relationship between the COI haplotypes, four common haplotypes (Hap 1-3 and 5) are all separated by a single polymorphism and form the primary lineage. The six location specific haplotypes (Hap 4, 6-10) also diverge from the primary lineage following a single polymorphism. Interestingly half of these variations diverge from haplotype 1, while a single location specific variant diverges from the haplotypes 2, 3 and 5, possibly indicating a level of geographical specific variation. Nevertheless, this low level of haplotype diversity within the COI gene suggests further that the sample population is likely evolving neutrally.

In contrast to this, both *AttA* and *Def* show considerably more extensive haplotype networks (Fig. 3.6B and C). Haplotypes 5 and 11 appear to indicate the primary *AttA* haplotype lineage, being separated by a single mutation, with all other haplotypes branching from these primary alleles (Fig. 3.6B). The majority of the variations result from a single polymorphism, however, separation between haplotypes 1 and 5 results from three polymorphisms. The extent of *Def* haplotype network (Fig. 3.6C) reinforces the previous suggestion of a recent selective sweep within the immune gene. No primary haplotype lineage can be observed within the network and all haplotypes are separated by a single mutation. The presence of three predicted haplotypes suggesting that the haplotype diversity within *Def* is potentially greater than seen within our sample populations.

An AMOVA test was conducted in PopART to assess the genetic variation between the populations. This indicated no relationship between genetic variation and geographical location and suggests that the populations are freely interbreeding (COI  $\phi_{st} = -0.00562$  ( $P = 0.465$ ), *AttA*  $\phi_{st} = -0.00432$  ( $P = 0.344$ ), and *Def*  $\phi_{st} = -0.00457$  ( $P = 0.483$ )).



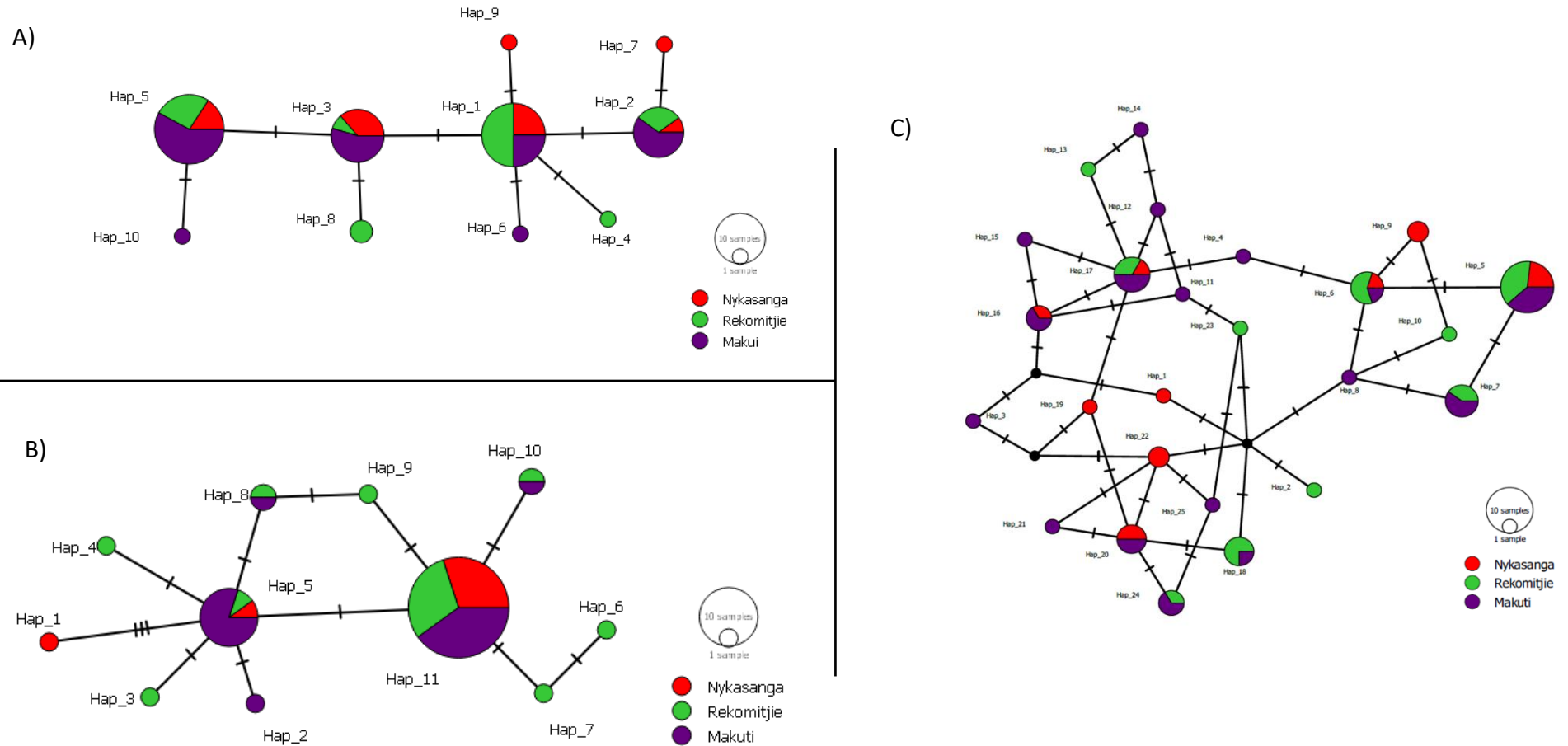


Figure 3.6: TCS haplotype networks of each target gene, A) *COI*; B) *AttA* and C) *Def*, within the sample population. The circle size represents of the number of samples within a haplotype, while the colours represent the geographical location of each sample. Black lines crossing a branch indicate the number of nucleotide mutations between haplotypes and solid black circles signify inferred or missing haplotype. All networks were produced in PopArt (Leigh and Bryant, 2015).

### 3.3.1iii: Intra-species nucleotide variation

Genetic variations across the *AttA* and *DEF* genes were assessed using sliding window analysis. Figure 3.7 illustrates polymorphic sites within the *AttA* and *Def* amplicons. Values of  $\pi$  were generated using DnaSP (Version 6) (Rozas *et al.*, 2017), using equation 1, in order to assess variation. For example, the two peaks at base pairs 522 and 559 in Figure 3.7A indicate these DNA polymorphisms, within the respective codons, show much greater variation compared to the other observed points of mutation. Nucleotide variation across all samples with the sample population was found to be low (*COI*  $\pi$  = 0.00187; *AttA*  $\pi$  = 0.00256 and *Def*  $\pi$  = 0.0127), though at a subpopulation level this was found to increase.

Eleven sites of mutation were observed within the *AttA* gene fragment; variation around nucleotide 522 is observed in all three loci, while variation at nucleotides 103 and 146 was not observed in Nykasanga samples. The remaining sites of variation (nucleotides 127, 198, 207, 246, 256, 399, 504 and 558) were all found to be location specific (Fig. 3.7A). The presence of geographical specific variation suggests that there is a degree of genetic diversification within the subpopulations. The largest point of variation was detected at nucleotide 522 ( $0.12821 \leq \pi \leq 0.17077$ ), while Rekomitjie exhibited the most polymorphic sites of all the subpopulations ( $n = 7$ ) (Fig. 3.7A)

Eight points of nucleotide variation were identified within the *Def* samples (Fig. 3.7B). Seven of these, at nucleotides 53, 111, 125, 133, 201, 204 and 244, were observed in all three subpopulations. A single polymorphic site at nucleotide 234 was observed in Rekomitjie and Makuti samples but was absent from the Nykasanga population. While the number of mutation sites is lower across the *Def* fragment,  $\pi$  is consistently higher than that observed in *AttA* (*Def*  $\bar{\pi}$  = 0.135178; *AttA*  $\bar{\pi}$  = 0.068868) (Fig. 3.7A/B). Having identified the distribution of nucleotide variation within the *AttA* and *Def* fragments, the nature of these mutations (whether synonymous or non-synonymous) was examined.

Of the eleven polymorphic sites identified in the *AttA* gene fragment (Fig. 3.7A), seven (nucleotides 198, 207, 246, 399, 504, 522 and 558) were found to be synonymous ( $S = 7$ ), while four (nucleotides 103, 127, 146 and 256) were found to be non-synonymous ( $N = 5$ ) (Fig. 3.7C). Six of the eight location specific polymorphic sites were found to be synonymous: three (nucleotides 198, 207 and 504) from Rekomitjie, two (nucleotides

246 and 399) from Nykasanga and one from Makuti (nucleotide 558). The remaining two (Rekomitjie 127 and Nykasanga 256) were found to be non-synonymous. The three-shared sites exhibited two non-synonymous and one synonymous site.

Within the Def samples, three synonymous (nucleotides 111, 201 and 234) and three non-synonymous (nucleotides 53, 133 and 244) mutations were identified ( $S = 3$ ,  $N = 3$ ) (Fig. 3.7D). Interestingly the two non-synonymous polymorphic sites at nucleotides 125 and 204 are absent for the results. The values of  $\pi$  remained consistently higher in the Def gene than that observed in AttA, supported this previous observation that genetic variation within the Def gene is considerably higher than in AttA.

The effect of these mutations on the genes was also considered by calculating dN/dS ratios between locations. This yielded similar results for both *AttA* and *Def* with *AttA* dN/dS = 0.117 and *Def* dN/dS = 0.269, suggesting signs of purifying selection in both genes.

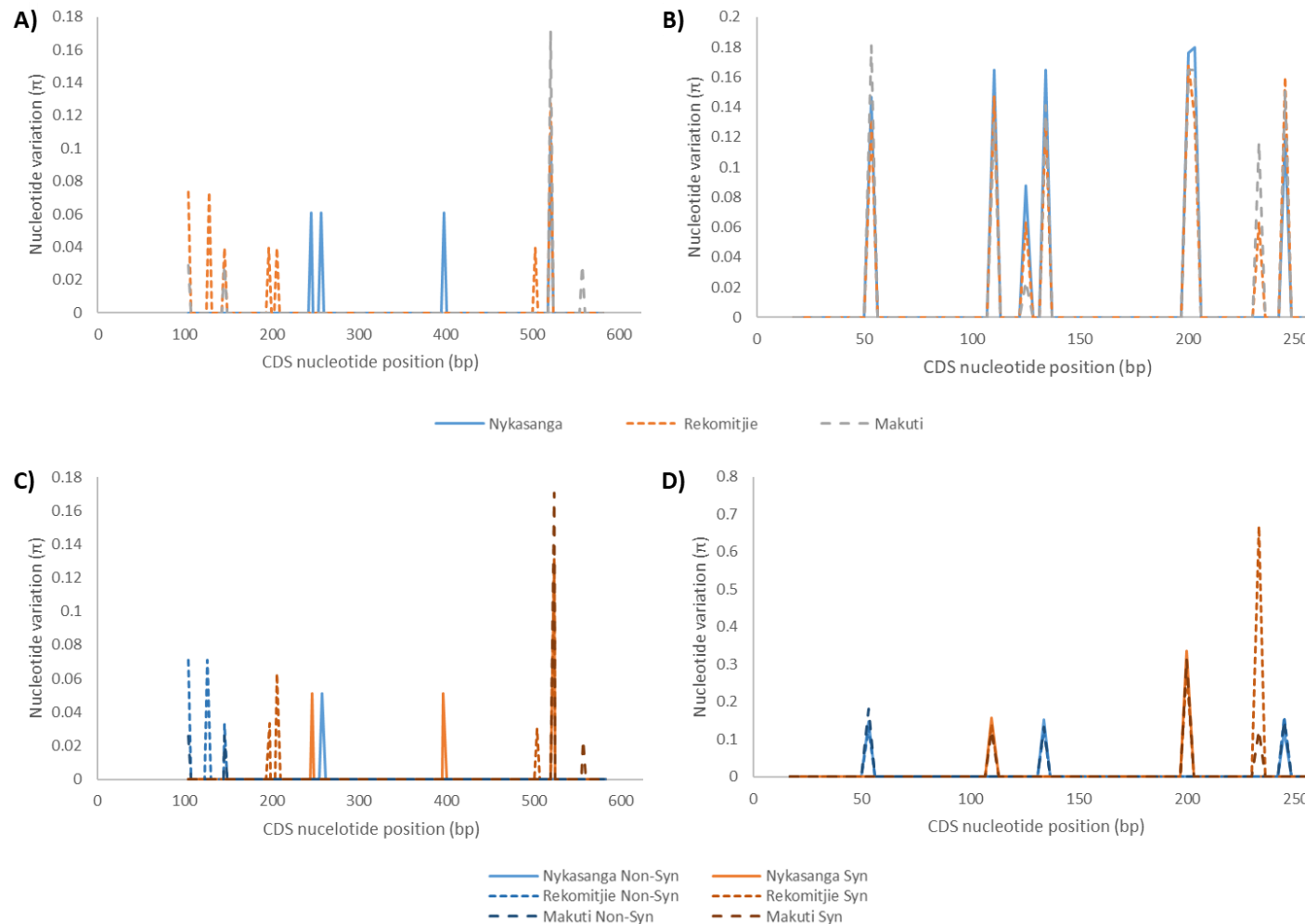


Figure 3.7: Sliding window graphs of the PCR amplification of *G. m. morsitans* immune genes; A) *AttA* and B) *Def*, illustrate the distribution of polymorphic sites throughout each gene fragment. Each subpopulation is represented by the colour of the line. C) *AttA* and D) *Def*, illustrating the characteristic (Synonymous and Non-synonymous) of each polymorphism. Subpopulations are denoted by the style of line, while the nature of mutation is denoted by colour. Produced in DnaSP V6 (Rozas *et al.*, 2017),  $\pi$  ( $\pi$ ) represents nucleotide diversity within the fragment against the nucleotide position of the mutation. The coding region in *AttA* samples was set between 103 - 582bp and between 64 - 309bp in *Def*. Window size = 3, step size = 3.

### 3.3.2: Tests for neutrality and demographic change

#### 3.3.2i: Gene flow and populations genetics.

Given the genetic and haplotype diversity observed in section 3.3.1, a test for selective neutrality and population analysis were conducted to assess for potential contributing factors. Gene flow analysis was conducted to measure the degree of genetic divergence between populations. Whilst the *COI* gene was used as a comparator between the nuclear and mitochondrial genomes.

Table 3.3: The gene flow results for the *AttA*, *Def* and *COI* genes across all collections localities. Values are rounded to three d.p where possible. M = Makuti; N = Nykasanga and R = Rekomitjie. Hs: Haplotype statistic. Ks: Nucleotide statistic.  $F_{st}$ : Fixation index. Dxy: The average number of nucleotide substitutions. Da: The net nucleotide substitution per site between populations. For equations used see section Appendix 2.

Pop. 1	Pop. 2	Hs	Ks	$F_{st}$	Dxy	Da
<b><i>COI</i></b>						
N	R	0.801	1.45	-0.039	0.002	-0.00007
N	M	0.794	1.46	-0.004	0.002	-0.00001
R	M	0.778	1.46	0.021	0.002	0.00004
<b><i>AttA</i></b>						
N	R	0.589	1.129	-0.008	0.002	-0.00002
N	M	0.544	0.806	0.044	0.002	0.00008
R	M	0.673	0.995	0.016	0.002	0.00003
<b><i>Def</i></b>						
N	R	0.917	3.044	-0.016	0.012	-0.0002
N	M	0.945	3.185	-0.020	0.013	-0.00025
R	M	0.932	3.122	0.024	0.013	0.00031

The mitochondrial gene, *COI*, indicted a high level of gene flow between subpopulations. Both the haplotype diversity (Hs) and nucleotide diversity (Ks) results indicated a relatively high level of diversity between subpopulations. However, the number of substitutions between sites was negligible (Dxy = 0.002). The fixation index ( $F_{st}$ ) values suggest that the subpopulations of *G. m. morsitans* are freely interbreeding and part of a more extensive panmictic population ( $F_{st} \approx 0$ ). This is supported by the earlier AMOVA test that also suggested free interbreeding between subpopulations (Section 3.3.1ii).

The observed levels of Hs and Ks within *AttA* supports the observation of gene flow between subpopulations, Hs and Ks are considerably lower in *AttA* than the other genes. Interestingly, the number of substitutions between sites is identical to that observed within *COI* ( $D_{xy} = 0.002$ ). Conversely, while Hs and Ks are relatively low in *AttA*, both values are far greater in *Def*. Not only does this support the observation made previously in section 3.3.1, but it supplies further evidence to support the proposed gene flow between subpopulations. Interestingly, despite the increased Hs and Ks values  $D_{xy}$  remains low ( $D_{xy} = 0.012$  or  $D_{xy} = 0.013$ ) indicating that the number of substitutions between sites is still very low. The  $F_{st}$  values of *AttA* and *Def* subpopulations further reinforce the suggestion that all populations are freely interbreeding and part of a larger panmictic population ( $F_{st} \approx 0$ ).

A Mantel test comparing geographical distance and  $F_{st}$  between subpopulations illustrated that the correlation between distance and interbreeding varied considerably between genes (Table 3.4). *AttA* and *COI* showed a positive correlation between distance and interbreeding, while *Def* illustrate a strong negative correlation. However, none of these were found to be statistically significant ( $P > 0.05$ )

Table 3.4: Mantel test results showing the correlation between geographical distance and  $F_{st}$  values. The statistical significance of each is given that all exceed the  $P < 0.05$  significance threshold.

Gene	Correlation (R)	P-Value
<i>COI</i>	0.4049	0.5104
<i>AttA</i>	0.5989	0.3405
<i>Def</i>	-0.7112	0.8342

### 3.3.2ii: Test of neutrality and demographic change

Tests of neutrality (Tajima's D and Fu's  $F_s$ ) were conducted to determine deviation from neutrality within the sample population. Both tests indicated that *COI* is evolving neutrally as well as presenting evidence of a recent population expansion event, possibly following the indicated recent genetic bottleneck ( $F_s < 0$ ). However, both values were statistically insignificant ( $P > 0.05$ ). These observations were mirrored within *AttA* with

negative values for both tests, indicating a population expansion following a genetic bottleneck. However, once again neither of these results were statistically significant. Interestingly, the Tajima's D result for *Def* indicated signs of balancing selection across the gene fragment. While the Fu's  $F_s$  statistic indicates the same genetic bottleneck, followed by population expansion event observed in *COI* and *AttA*. Whilst Tajima's D remain statistically insignificant, Fu's  $F_s$  was found to be significant ( $P < 0.05$ ).

Table 3.5: Shows the Tajima's D and Fu's  $F_s$  statistics of *COI*, *AttA* and *Def*. \* indicates statistically significant results ( $P < 0.05$ ).

	<i>COI</i>	<i>AttA</i>	<i>Def</i>
<b>Tajima's D</b>	-0.63082	-1.73805	1.69765
<b>Fu's <math>F_s</math></b>	-2.908	-6.899	-15.032*

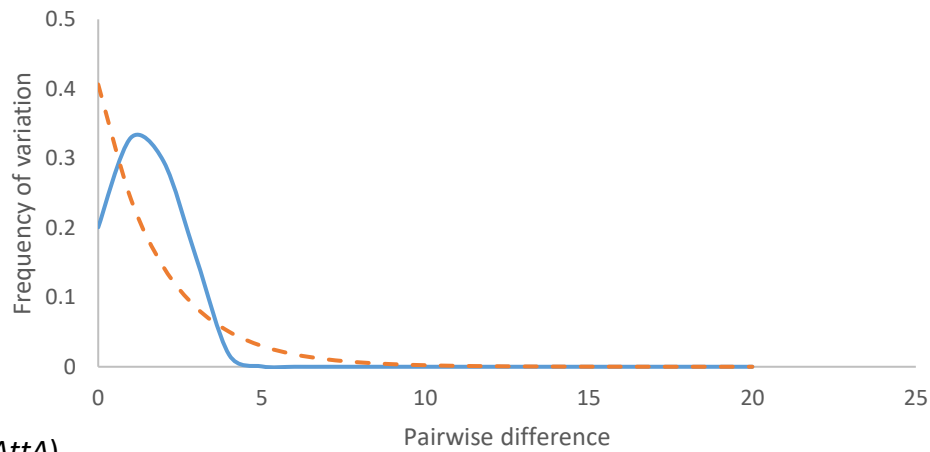
Pairwise mismatch and Raggedness ( $r$ ) were used to assessed demographic change by examining the expected and observed frequencies of mutation (Harpending, 1994). Raggedness ( $r$ ) values illustrate recent and ancient population expansions where  $r \approx 0$  indicates a recent population expansion event under the presumptions of a constant population size and no recombination between sites. Pairwise mismatch within the *COI* gene indicated a recent population expansion, with a high frequency of genotypes with a low pairwise difference. (Fig. 3.9A). This observation is supported by the Raggedness value ( $r = 0.0568$ ) which further indicated a recent population event. However, this was not found to be statistically significant,  $P > 0.05$ . These predictions could also be drawn from the results generated by the *AttA* and *Def* genes. Both illustrated a high frequency of genotype variants with a low frequency of pairwise differences (Fig. 3.9B and C), while  $r \approx 0$  further supported the hypothesis of a recent expansion event (*AttA*  $r = 0.0589$ ,  $P > 0.05$ ; *Def*  $r = 0.0242$ ,  $P > 0.05$ ).

When Raggedness was calculated using the Coalescent theory under the presumptions of a constant population size and free recombination between sites, there is very little change in the resultant  $r$  values. Under these assumptions, the average  $r$  value of *COI* increases to 0.11214, with a 95% Confidence interval between 0.04422 and 0.26597. Likewise, the average  $r$  value of *AttA* also increased to 0.17112, with a larger 95%

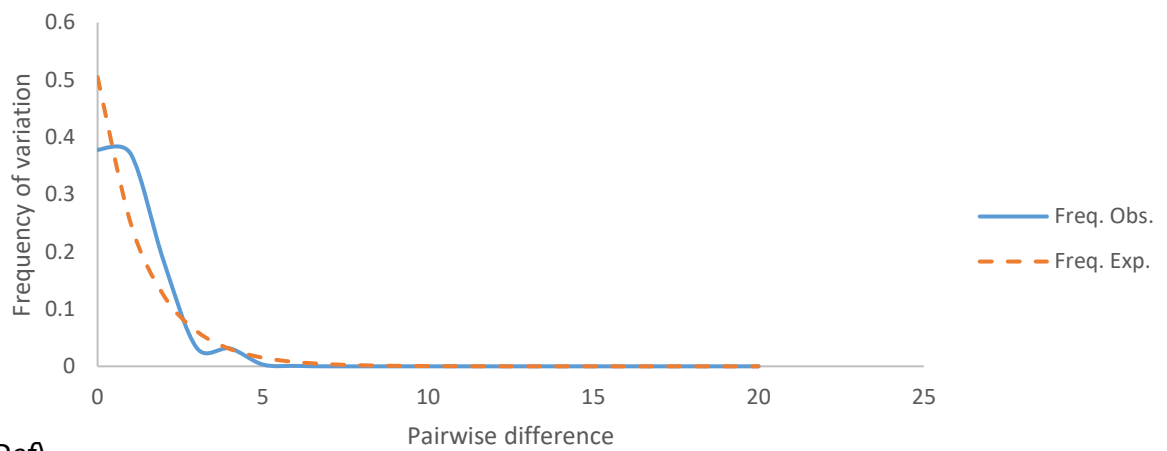
confidence interval range between 0.0562 and 0.72064. The raggedness value for *Def* also increased though by a far lesser degree ( $\bar{r} = 0.04922$ ), while the 95% confidence interval covers a much smaller range between 0.02159 and 0.10435.



*COI*)



*AttA*)



*Def*)

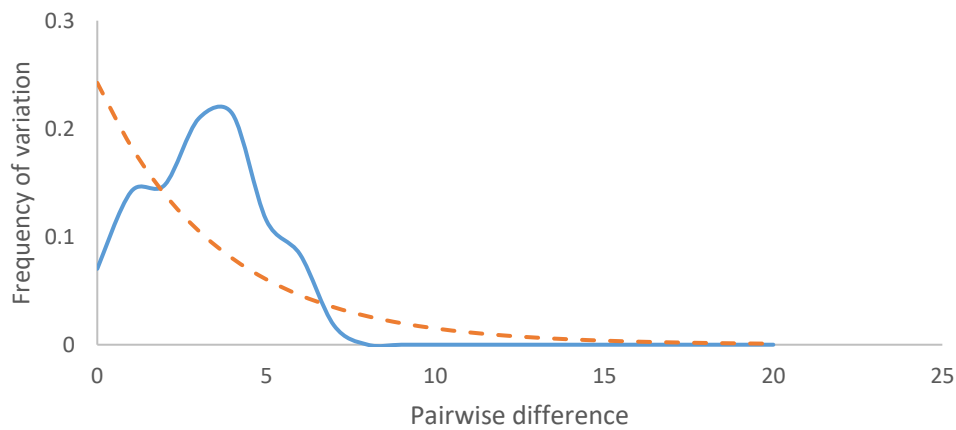


Figure 3.8: Pairwise mismatch analysis of the *COI*, *AttA* and *Def* genes showing the observed and expected frequencies of nucleotide variation. Produced in DnaSP (V6) using equations 7a and 7b (Appendix 2) to calculate the expected frequency. *AttA* illustrates the presence of a recent population expansion within the tsetse population. *Def* also demonstrates a population expansion, though the presence of a genetic bottleneck can also be seen in the bimodal observed frequency curve.

### 3.3.2iii: Recombination analysis

Genetic recombination was evaluated to estimate mutation rates in both nucleotide genes. DnaSP (V6) found one sites of recombination within the *AttA* gene ( $R_m = 1$ ) (Table 3.7), however, the recombination parameter was estimated to be much higher ( $R = 60.1$ ). Within *Def*,  $R_m$  was higher than seen in *AttA* ( $R_m = 5$ ) while  $R$  remained high ( $R = 53.0$ ) (Table 3.7). No evidence of recombination was identified within the *COI* gene fragment ( $R_m = 0$ ).

Table 3.6: The sites of recombination within the *AttA* and *Def* fragments. The nucleotide position substitution and nature of mutation are given, any detected recombination between sites is also shown.

Nucleotide position	Nucleotide substitution	Synonymous or Non-Synonymous	Recombination	Between sites
<b><i>AttA</i></b>				
103	G → A	Non-synonymous	No	N/A
127	G → A	Non-synonymous	No	N/A
146	C → A	Non-synonymous	No	N/A
198	A → C	Synonymous	No	N/A
207	A → G	Synonymous	Yes	207 – 522
246	C → G	Synonymous	No	N/A
256	T → G	Non-synonymous	No	N/A
399	A → G	Synonymous	No	N/A
504	A → G	Synonymous	No	N/A
522	T → C	Synonymous	No	N/A
558	A → G	Synonymous	No	N/A
<b><i>Def</i></b>				
53	C → G	Non-synonymous	No	N/A
	C → T	Non-synonymous	No	N/A
111	A → G	Synonymous	Yes	159 - 181
125	G → A	Synonymous	No	N/A
133	G → T	Non-synonymous	Yes	181 - 249
200	T → C	Synonymous	Yes	249 - 252
204	T → C	Synonymous	Yes	525 - 292
234	C → T	Synonymous	No	N/A
244	G → A	Non-synonymous	No	N/A

Interestingly, when the *AttA* and *Def* alignments were submitted to GARD (Kosakovsky *et al.*, 2006) via the Datamonkey online software (Sergei *et al.*, 2005; Weaver *et al.*, 2018), no evidence of recombination was found in either of the genes. However, breaking points corresponding to the points of variation were identified.

### 3.3.3: *Wigglesworthia* endosymbiosis and genetic variation

#### 3.3.3i: *Wigglesworthia* genetic and population analysis

From the 34 *Wigglesworthia* sequences, a total of 15 haplotypes were identified ( $H = 15$ ;  $Hd = 0.6934$ ;  $\pi = 0.00292$ ). A single common haplotype (Haplotype 12) was exhibited by 19 samples (seven from Nykasanga, eight from Rekomitjie and four from Makuti), while haplotype 9 contained one sample from Nykasanga and Makuti. The remaining 13 haplotypes (Hap 1-8, 10-11, 13-15) were location specific, three were identified in Nykasanga (Hap 4, 5 and 10) and Rekomitjie (Hap 6, 7 and 11) while seven were exhibited by sample from Makuti (Hap 1-3, 8, 13-15). Despite an almost equal distribution of successfully amplified samples across the three sites, Makuti exhibits a considerably higher haplotype diversity than the other collection sites ( $H = 9$ ). The TCS haplotype network (Fig. 3.11) illustrates this diversity within the *Wigglesworthia 16S* samples. Unlike the haplotype networks observed with *COI*, *AttA* and *Def* (Fig. 3.6), the number of polymorphic sites between haplotypes varied considerably in the *Wigglesworthia 16S* gene, ranging between one and four polymorphisms between haplotypes. Furthermore, the presence of three inferred haplotypes suggests that the full extent of the haplotype diversity is not represented within this network (Fig. 3.10). A simple AMOVA test showed there was no significant relationship between the genetic variation and geographic location ( $\phi_{st} = 0.04130$ ;  $P = 0.083$ ), whilst not statistically significant this does infer that the subpopulations are interbreeding.

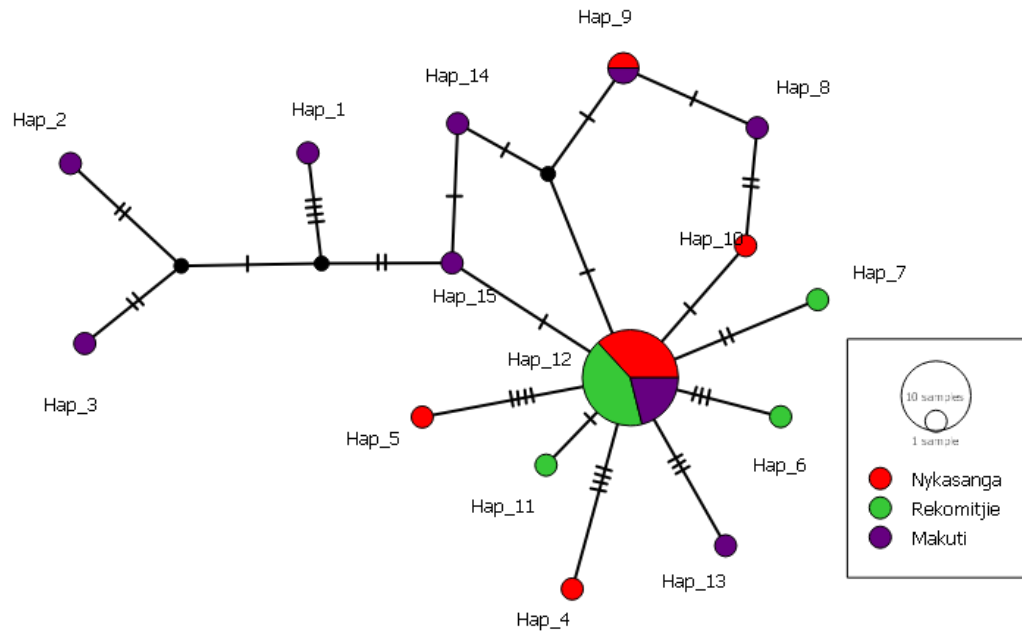


Figure 3.9: TCS haplotype network of the *Wigglesworthia 16S* gene within the *G. m. morsitans* sample populations, with reference to geographical location. Produced in PopART, the circle size represents of the number of samples within a haplotype, while the colours represent the collection location of samples within the haplotype. Black lines crossing a branch indicate the number of nucleotide mutations between haplotypes. Solid black circles signify inferred or missing haplotype.

Gene flow within the three *Wigglesworthia* subpopulations indicated similar results those observed previously in the *Glossina COI*, *AttA* and *Def* (section 3.3.2i). Table 3.8 shows that haplotype diversity ( $H_s$ ) was relatively high, while nucleotide diversity ( $K_s$ ) was higher than observed within the *G. m. morsitans* genes. However, given the increase in polymorphic sites between haplotypes (Fig. 3.10) this is to be expected. Interestingly, the number of mutations per mutation site ( $D_{xy}$ ) remains low despite this increase in nucleotide variation be haplotypes. Finally, the  $F_{st}$  values to continue to support the presence of a recent population expansion event ( $F_{st} \approx 0$ ). It should be noted however, that the  $F_{st}$  value representing interbreeding between Rekomitjie and Makuti is higher than expected ( $F_{st} \approx 0.1$ ). While not high enough to be considered a sign of no interbreeding, it may indicate that individuals from these sites are not interbreeding as freely as other sites.

Table 3.7: The gene flow results for the *W. g. morsitans 16S* and *G. m. morsitans* between all collections localities. Values are rounded to three d.p where possible. M = Makuti; N = Nykasanga and R = Rekomitjie. Hs: Haplotype statistic. Ks: Nucleotide statistic.  $F_{st}$ : Fixation index. Dxy: The average number of nucleotide substitutions. Da: The net nucleotide substitution per site between populations. For equations used, see Appendix 2.

Pop. 1	Pop. 2	Hs	Ks	$F_{st}$	Dxy	Da
<b><i>W. g. morsitans 16S</i></b>						
N	R	0.556	1.56	0.000	0.002	0.00000
N	M	0.739	2.97	0.054	0.004	0.0002
R	M	0.679	2.54	0.109	0.003	0.00036
<b><i>G. m. morsitans COI</i></b>						
N	R	0.801	1.45	-0.039	0.002	-0.00007
N	M	0.794	1.46	-0.004	0.002	-0.00001
R	M	0.778	1.46	0.021	0.002	0.00004

The test of neutrality (Tajima's D and Fu's  $F_s$ ) supported the observations made in section 3.3.2ii indicating a recent genetic bottleneck and recovery event ( $D = -2.19142$ ,  $P < 0.01$ ;  $F_s = -6.073$ ,  $P < 0.02$ ). However, unlike the results in section 3.3.2ii, both of these values were found to be statistically significant ( $P < 0.05$ ). This adds considerable weight to the hypothesis that all three subpopulations are currently expanding. However, pairwise mismatch analysis shows considerable difference to the results observed within the *G. m. morsitans* genes. This indicates that the *Wigglesworthia 16S* gene shows low frequency of genotype with a higher level of pairwise mismatches (Fig. 3.11). Despite this,  $r$  remains low ( $r = 0.0697$ ,  $P > 0.05$ ) once again reinforcing the observed population expansion. As with observed previously, when  $r$  was calculated under the assumption of free recombination using Coalescent theory, the mean  $r$  value rose to 0.072 with a 95% confidence interval between 0.02905 and 0.14759.

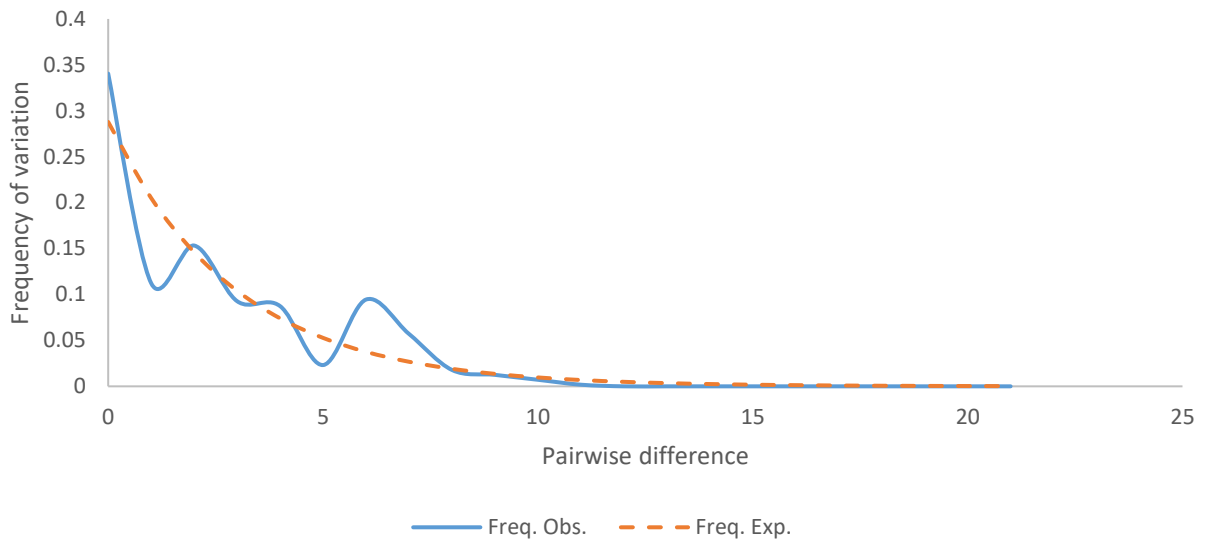


Figure 3.10: Pairwise mismatch analysis of the *W. glossinidia 16S* gene showing the observed and expected frequencies of nucleotide variation. Produced in DnaSP (V6) using equations 7a and 7b to calculate the expected frequency. This illustrates the low frequency of genotype with a higher level of pairwise mismatches.

A mantel test illustrated that *W. g. morsitans 16S* and *G. m. morsitans COI* exhibited an identical moderate positive correlation (Table 3.8). However, this was found to be statistically significant ( $P > 0.05$ ).

Table 3.8: Mantel test results showing the correlation between geographical distance and  $F_{st}$  values. The statistical significance of each is given that all exceed the  $P < 0.05$  significance threshold.

Gene	Correlation (R)	P-Value
<i>W. g. morsitans 16S</i>	0.4049	0.6712
<i>G. m. morsitans COI</i>	0.4049	0.5104

Recombination was detected between three sites ( $R_m = 3$ ) within the *W. g. morsitans 16S* gene fragment, between nucleotides 252/287; 296/657; and 657/954. Interestingly, the recombination parameter per gene (R) was found to be considerably lower than that observed within *G. m. morsitans* immune genes ( $R = 0.2$ ). Interestingly, recombination was also identified by GARD suggesting that recombination may play a role in the genetic variation of the *W. g. morsitans 16S* gene.

### 3.3.3ii: Association of *W. g. morsitans* 16S and *G. m. morsitans* COI, AttA and Def

The association of specific haplotypes between the *W. g. morsitans* 16S gene and the *G. m. morsitans* mitochondrial and immune genes could offer novel insights into the relationship between the tsetse host and the symbiont. Comparison of the *G. m. morsitans* gene haplotype networks (Fig. 3.6) and *W. g. morsitans* 16S haplotypes illustrated the association of genetic variation between the two organisms (Fig. 3.13). Where no success *W. g. morsitans* 16S amplification was achieved the sample and corresponding *G. m. morsitans* haplotype was removed from the analysis.

*Glossina m. morsitans* COI exhibited ten haplotypes (Fig. 3.6A), the four common haplotypes (Haps 1, 2, 3 and 5) each exhibited multiple *W. g. morsitans* 16S haplotypes, though the number of samples and haplotypes present within each of these varied (Fig. 3.13A). All four exhibited samples from *Wigglesworthia* 16S haplotype 12, while COI haplotype 3, exhibited just one other *W. g. morsitans* 16S haplotype (Hap 14), COI haplotype 2 exhibited a total of three *W. g. morsitans* 16S haplotypes (Haps 5, 8 and 12). Interestingly, both the samples of *W. g. morsitans* 16S haplotype 9 were exhibited by COI haplotype 1, as were haplotypes 10, 11 and 12, while COI haplotype 5 exhibited eight *W. g. morsitans* 16S haplotypes (Haps 1, 3, 4, 6, 7, 12, 13 and 15) (Fig. 3.13A). Furthermore, of the six location specific haplotypes exhibited by COI only two, Haps 4 and 10, exhibited *W. g. morsitans* 16S haplotypes; COI Haplotype 4 exhibited *W. g. morsitans* 16S haplotype 12, while COI haplotype 10 exhibited *W. g. morsitans* 16S haplotype 2 (Fig. 3.13A). Interestingly, *W. g. morsitans* 16S haplotype 2 is also location specific and observed within a single a single sample (Fig. 3.11).

Of the 11 haplotypes observed in AttA (Fig. 3.6B), three (Haps 3, 4 and 6) contained no identified *W. g. morsitans* 16S haplotype (Fig. 3.13B). A further four (Haps 5 and 11) contained several missing samples. *Wigglesworthia glossinidia morsitans* 16S haplotype 12 was exhibited in five of the eight AttA haplotypes assessed, Haplotypes 1, 5, 8, 10 and 11 (Fig. 3.13B). *W. g. morsitans* 16S haplotype 12 was exhibited by one sample in both AttA haplotypes 8 and 10, while *W. g. morsitans* 16S haplotypes 14 and 15 were also observed within these AttA haplotypes. Three other *W. g. morsitans* 16S haplotypes were observed within location specific haplotypes, 16S haplotypes 7, 9 and 11, were exhibited by AttA haplotypes 2, 7 and 9 respectively *glossinidia* (Fig. 3.13B).

Eleven *Def* haplotypes contained no identified *W. g. morsitans 16S* haplotypes, while a further five were missing identified *W. glossinidia 16S* haplotypes (Fig. 3.13C). Six *Def* haplotypes exhibited a single *W. g. morsitans 16S* haplotype (Haps 8, 11, 13, 15, 19 and 23) of which three (Haps 11, 13 and 15) exhibited *W. g. morsitans 16S* haplotype 12. Defensin haplotypes 8, 19 and 23 exhibited *W. g. morsitans 16S* haplotypes 15, 1 and 14, respectively. Both samples exhibiting *W. g. morsitans 16S* haplotype 9, were found to also exhibited *G. m. morsitans Def* haplotype 17 (Fig. 3.13C).

Despite this observed variation a simple AMOVA test showed no significant relationship between the immune genes and *W. glossinidia 16S* haplotype variation (*COI*  $\phi_{st}$  = 0.15882, P = 0.108; *AttA*  $\phi_{st}$  = -0.31313, P = 0.571; *Def*  $\phi_{st}$  = 0.11752; P = 0.125).



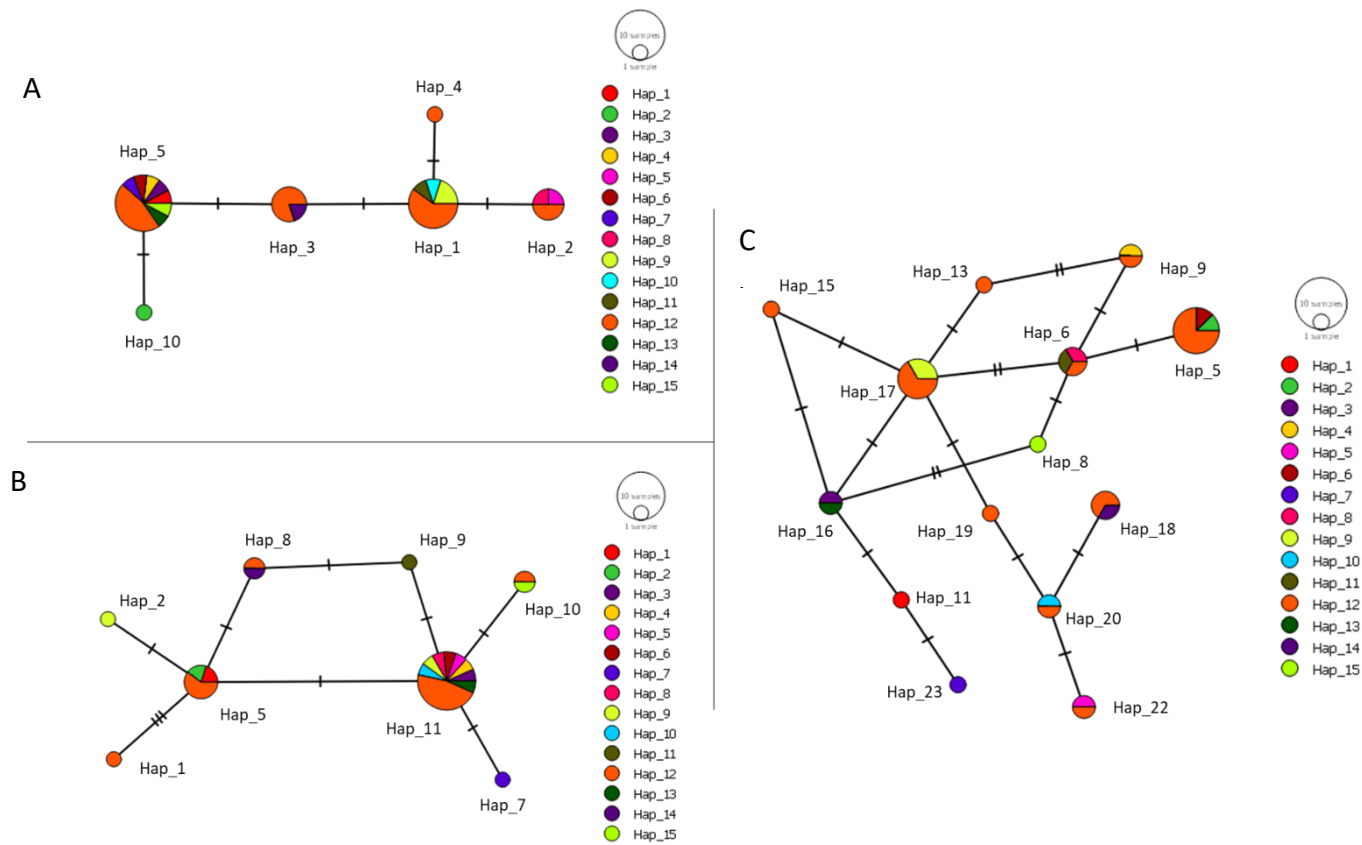


Figure 3.11: TCS haplotype networks for the *G. m. morsitans* immune genes *COI* (A), *AttA* (B) and *Def* (C) showing the frequency of *Wigglesworthia 16S* haplotypes within the exhibited haplotypes. Produced in PopART, the circle size represents the number of samples exhibiting a haplotype, while the colour represent the presence of *Wigglesworthia* haplotypes within the samples. White filled sections represent the samples that failed to amplify during PCR, while black circles represent missing or inferred haplotypes. Black lines crossing a branch indicate the number of nucleotide mutations between haplotypes.

### 3.3.4: *Trypanosoma* infection and genetic variation

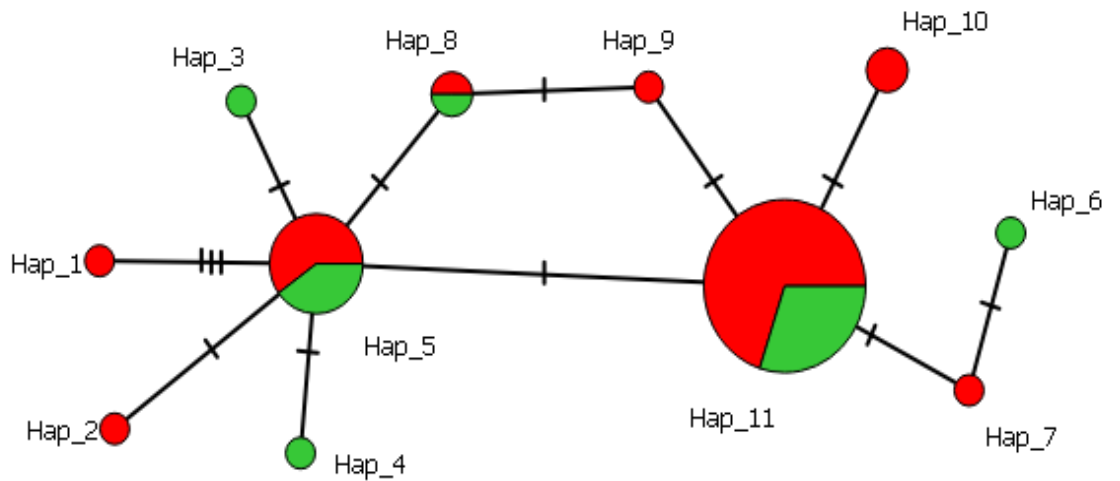
A surprisingly high infection rate was detected within tsetse sample populations, 43/62 (69.35%) of the sample population tested positive for *Trypanosoma* infection. The majority of these were mixed infections comprising two or three *Trypanosoma* spp., while the remaining 19 were negative. Of the nine samples purified using gel extraction six yielded successful sequencing results, confirming the presence of three African Trypanosome species, *Trypanosoma evansi*, *Trypanosoma congolense* and *Trypanosoma vivax* within the sample population. A fourth unconfirmed species was detected though the poor quality of sequencing result made full confirmation impossible. Gel electrophoresis results indicated that a further three *Trypanosoma* spp., namely *Trypanosoma grayi*, *Trypanosoma godfreyi* and *Trypanosoma brucei*, could be present within the sample population though these species were not confirmed by sequencing.

The association between *G. m. morsitans* immune gene variation and trypanosome infection was assessed by reconstructing the haplotype networks (Fig 3.6B and C), illustrating the infection status of each sample within the haplotype. Five of the 11 AttA haplotypes (Haps 1, 2, 7, 9 and 10) showed a 100% infection rate, three (Haps 5, 8 and 11) showed a mixture of infected and uninfected samples, while just three (Haps 3, 4 and 6) showed no sign of infection (Fig 3.14A). Of the infected haplotypes four were location specific containing a single sample (One from each of Makuti and Nykasanga and two from Rekomitjie), while the sixth contained two samples from different localities (Makuti and Rekomitjie), while all three of uninfected haplotypes were location specific haplotypes from Rekomitjie. None of the three mixed infection haplotypes were location specific, though haplotype 8 contained two samples one from Rekomitjie (infected) and one from Makuti (uninfected). The remaining two haplotypes (Haps 5 and 11) contained samples from all 3 locations, haplotype 5 contained six infected samples and four uninfected samples, while the most common AttA haplotype (Hap 11) contained 21 infected and nine uninfected samples. Despite some observed association between haplotypes and infection a simple AMOVA test showed there was no significant relationship between them ( $\phi_{st} = -0.00199$ ;  $P = 0.42$ ).

A total of 15 *Def* haplotypes showed a 100% infection rate, 13 of these (Haps 1-2, 4, 8, 10-14, 19, 21, 23 and 25) contained a single sample, while haplotypes 9 and 22 contained two

samples both from Nykasanga (Fig. 3.14B). Two haplotypes (3 and 15) were the only localised haplotypes that showed no sign of infection. The remaining eight haplotypes (Haplotypes 5-7, 16-18, 20 and 24) contained a mixture of infected and uninfected samples. Haplotypes 5-6 and 17 contained samples from all three locations, the remaining five haplotypes contained samples from two collection sites. Haplotype 5 contains a total of 13 samples, eight of which were infected, while haplotype 17 contains 5 infected and a single uninfected sample. Unlike the previous two haplotypes, haplotype 6 had more uninfected (3) than infected (2) samples. The remaining five haplotypes were all observed in two localities, haplotypes 7, 16 and 24 all contained more infected than uninfected samples. Haplotype 20 contained an equal number of both infected and uninfected while haplotype 18 contained more uninfected samples (Fig. 3.14B). As previously with *AttA*, a simple AMOVA test showed no significant relationship between haplotype variation within the sample population and trypanosome infection ( $\phi_{st} = -0.00230$ ;  $P = 0.422$ ).

A)



B)

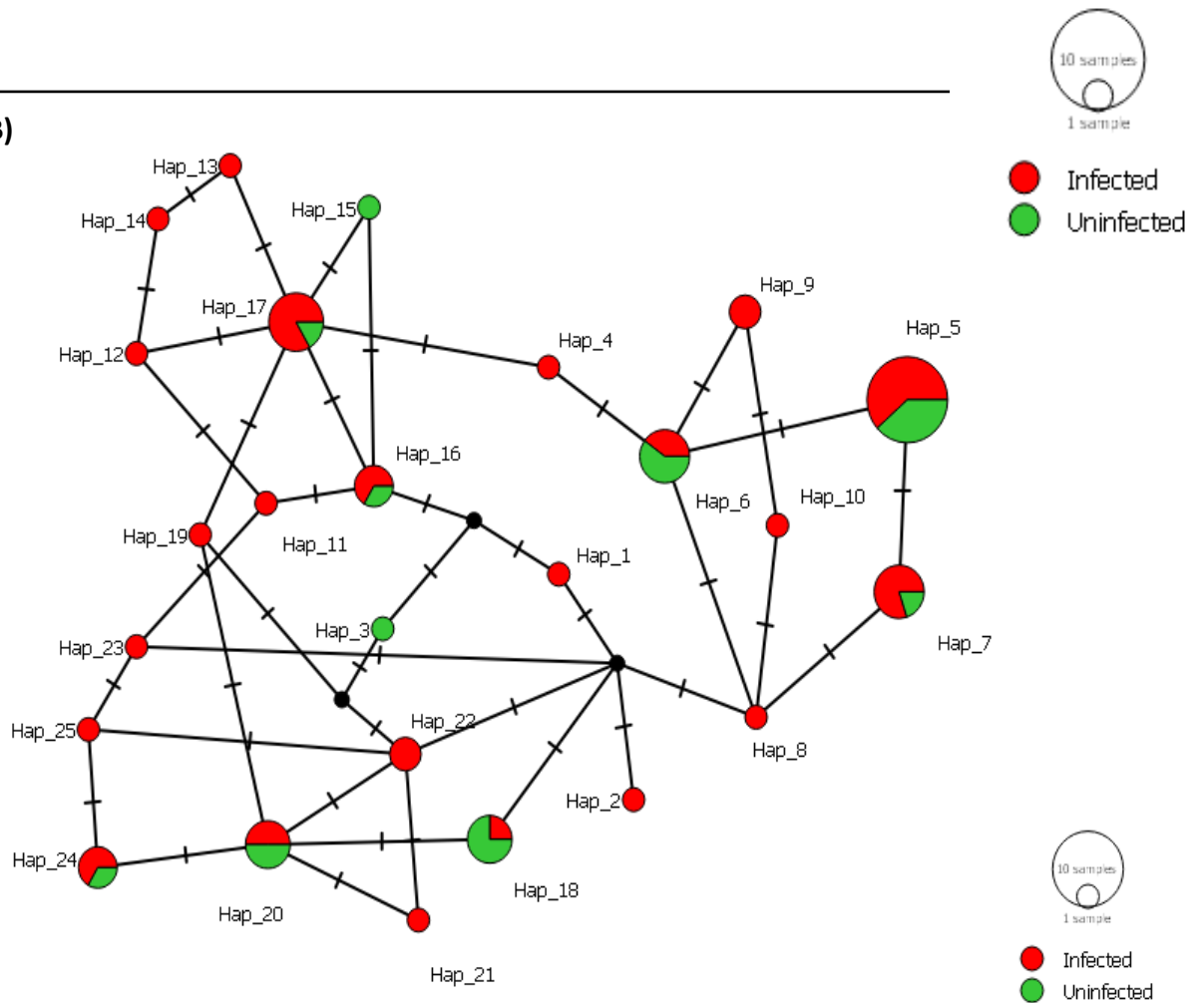


Figure 3.12: TCS haplotype network of both *G. m. morsitans AttA* (A) and *Def* (B) genes showing the frequency of infected and uninfected samples within each haplotype. Produced in PopART, the circle size represents the number of samples exhibiting a haplotype, while the colours represent the infection status of samples within the haplotype. Black circles represent inferred or missing haplotypes, while black lines crossing a branch indicate the number of nucleotide mutations between haplotypes.

### 3.3.4i: *Wigglesworthia* 16S variation and trypanosome infection

The infection rate of 70.59% (24/34 successful amplification), within the *Wigglesworthia* samples was comparable to the overall trypanosome infection rate of 69.35%. Of the 15 16S haplotypes identified in section 3.3.3i, ten (Haps 1, 4, 5, 6, 7, 8, 10, 11, 14 and 15) were found to have a 100% infection rate, haplotypes 9 and 12 contained both infected and uninfected samples, and just three haplotypes showed no signs of infection (Haps 2, 3 and 14) (Fig. 3.15) . All the haplotypes exhibited a 100 % infection rate contain a single sample and are specific to a single geographical location, while the most commonly exhibited haplotype (Hap 12) showed a 70 % infection rate. Perhaps of most interest are haplotypes 2 and 3, both diverge from the primary haplotype cluster through a series of unidentified haplotypes, and both show no signs of infection. However, an AMOVA test suggests there is no statistical relationship between haplotypes and infection ( $\phi_{st} = 0.01217$ ;  $P = 0.247$ ).

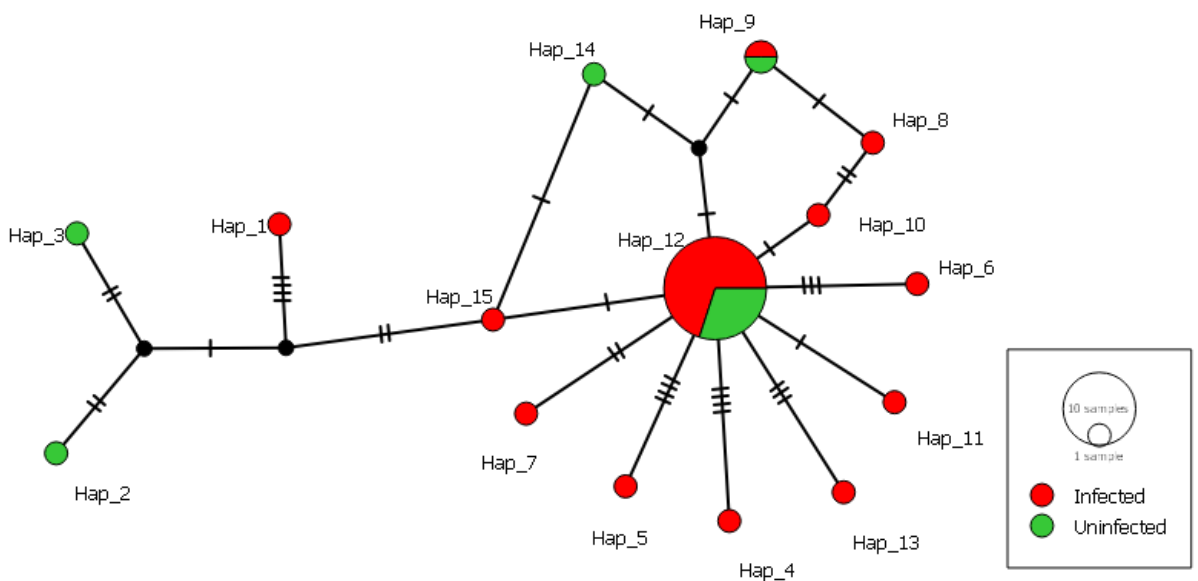


Figure 3.13: TCS haplotype network of the *W. glossinidia* 16S gene showing the frequency of infected and uninfected samples within each haplotype. Produced in PopART, the circle size represents the number of samples exhibiting a haplotype, while the colours represent the infection status of samples within the haplotype. Black circles represent inferred or missing haplotypes, while black lines crossing a branch indicate the number of nucleotide mutations between haplotypes.

### 3.3.5: Comparison of the genetic variation

The association between AMP nucleotide and symbiont variation, indicated two different results (Fig. 3.16). Firstly, the association between *AttA* and *W. g. morsitans 16S* showed a clear negative correlation, meaning as *AttA* variation increased *W. g. morsitans 16S* decreased (Fig. 3.16A). This was supported by the results of the Mantel test which also indicated a negative correlation between the two, though this is statistically insignificant. On the other hand, a positive correlation was observed between *Def* and *W. g. morsitans 16S*. This was again supported by the Mantel test results, though once again these were found to be statistically insignificant.

Table 3.9: Mantel test results showing the correlation between *G. m. morsitans* AMPs and *W. glossinidia 16S*.

<b>Genes</b>	<b>Correlation (R)</b>	<b>P-Value</b>
<i>AttA vs 16S</i>	-0.09845	0.8061
<i>Def vs 16S</i>	0.03068	0.2488

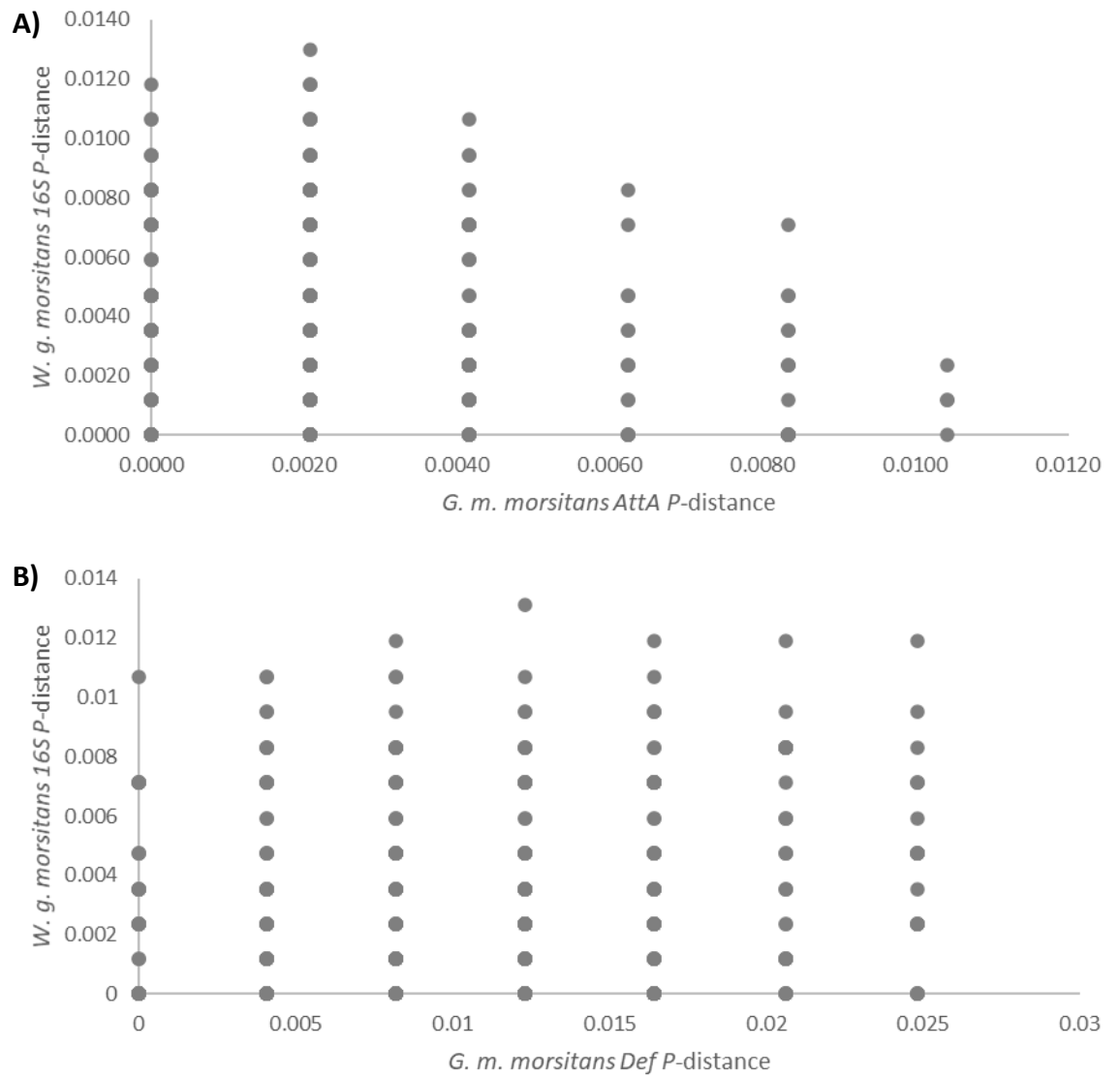


Figure 3.14: The association of *P*-distance between *AttA* (A) and *Def* (B) and successful *W. g. morsitans* amplification. *P*-distance was calculated in MEGAX (Kumar *et al.*, 2018), scatter plots were plotted Microsoft excel.

### 3.4: Discussion

This chapter offers a novel insight into the evolutionary genetic history of a wild *G. m. morsitans* population, as well as exploring the genetic relationship between wild *G. m. morsitans*, the endosymbiont bacteria *W. g. morsitans* and *Trypanosoma* spp.

Initial phylogenetic analysis of the *G. m. morsitans* mitochondrial and nuclear genomes showed no relationship between genetic variation and geographic location. The phylogenetic analysis of *COI* and *AttA* indicated a relatively high level of relatedness between samples, whereas *Def* indicated a more diverse evolutionary history. Overall the phylogenetic topology, haplotype and nucleotide diversity of the *G. m. morsitans* mitochondrial gene *COI* was found to be comparable to other dipteran species (de Jong *et al.*, 2011; Qin *et al.*, 2016), suggesting that the wild *G. m. morsitans* mitochondrial genome is stable and evolving neutrally (Powell *et al.*, 1986; Caccone *et al.*, 1988).

The low level of intraspecies variation within *AttA* could be a consequence of the evolutionary history of *Glossina* tsetse flies. The unexpectedly high level of interspecies variation of *AttA* (Chapter 3, Section 2.3A.7i), combined with the low level of intraspecies nucleotide variation strongly alludes to concerted evolution driving nucleotide variation within *Glossina* *AttA* (Liao, 1999). An alternate possibility, is the maintenance of genetic variation via balancing selection, given the high number of rare *AttA* haplotypes (at frequency < 5% within a population (Datta *et al.*, 2018) observed within sample population. This process has been observed previously within *D. melanogaster* *AttA* (Lazzaro and Clark, 2001), however, this premise was not supported by the test of neutrality conducted within this study which showed no indication of balancing selection.

The elevated levels of nucleotide and haplotype variation observed within *Def* are comparable to that observed in other immune genes previously (Unckless and Lazzaro, 2016a; Chapman *et al.*, 2019). Furthermore, unlike *AttA* which forms part of a gene family, *Def* was identified as an individual gene within the *Glossina* genome (Chapter 3), and as such, is likely subject to higher levels of selective pressure (Jiggins and Hurst, 2003; Lehmann *et al.*, 2009). Therefore, the indication of balancing selection observed within the *Def* sample is not only expected (Clark and Wang, 1997; Chapman *et al.*, 2019), but also



offers an explanation for the increased levels of nucleotide and haplotype variation to maintain multiple functional alleles within the population (Siewert and Voight, 2017).

Interestingly, both AMPs and *COI* displayed high levels of gene flow, and results for Fu's  $F_s$ , Tajima's  $D$ , Pairwise mismatch and raggedness tests all of which indicated a recent population expansion event, possibly representing the recovery from a population bottleneck (Tajima, 1989; Rogers and Harpending, 1992; Harpending, 1994; Fu, 1997). Additionally, the high number of rare haplotypes but low variation between haplotypes is characteristic of a population expansion event (Slatkin and Hudson, 1991; Venkatesan *et al.*, 2007; Allcock and Strugnell, 2012). Curiously, *Def* appeared to illustrate a haplotype topology less indicative of a population expansion (Allcock and Strugnell, 2012), however, balancing selection within the population to maintain genetic diversity could be responsible for this variation (Clark and Wang, 1997; Chapman *et al.*, 2019). Furthermore, Allcock and Strugnell (2012) hypothesised that a similar pattern may be observed if isolated populations survived a bottleneck event and then began recolonization, maintaining non-geographically specific genetic variation.

Interestingly, a recent population bottleneck could be a consequence of tsetse control measures within the collection area. Tsetse control in the collection area of this study terminated in 2001 (Shereni *et al.*, 2016), and while estimates suggest that clear areas can be recolonized within one to two years (Turner and Brightwell, 1986; Hargrove, 2000) this is still a comparatively recent event. While it could be argued similar results would be expected from a founding effect following recolonization (Raupach *et al.*, 2010; Allcock and Strugnell, 2012), there is no evidence to suggest the original tsetse population was eliminated, furthermore it is uncommon for founding effects to be accompanied by evidence of a rapid population expansion (Allcock and Strugnell, 2012).

The high gene flow between subpopulations is likely an outcome of the indicated free interbreeding between the subpopulations and implied panmictic population. However, two important physical factors must be considered before the presence of a panmictic population is accepted; the physical geography surrounding the collection sites and, the distance between collection sites. The physical geography of the collection sites illustrates no physical barrier preventing interbreeding between the three populations (Supplementary Figure 10, Appendix 6). Although, tsetse favour lower altitudes, below

1000m the presence of a tsetse population in Makuti and further up the escarpment indicates that migration to higher altitudes is possible (Shereni *et al.*, 2016). Interestingly, the observed negative correlation between  $F_{st}$  and geographical distance infers a greater degree of interbreeding between the two furthest geographical sites (Nykasanga and Rekomitjie).

The average distance between collection sites is 26.30 Km, whilst this distance could be considered relatively close in terms of population genetic analysis, it does present a long migration for tsetse flies. The average daily movement of approximately 1 Km (Hargrove, 2000), and the average lifespan of 14-21 days for males and 30-120 days for females (Encyclopædia Britannica, 2017) of tsetse suggests that while these loci are theoretically within the limits of tsetse migration, the subpopulations are more likely breeding with surrounding populations between the collection sites, rather than directly interbreeding.

The genetic variation within *W. g. morsitans 16S* was found to be highly comparable to previous studies into genetic variation of *W. glossinidia* subsp. (Symula *et al.*, 2011) and other primary endosymbionts (Funk *et al.*, 2001; Abbot and Moran, 2002). As *W. glossinidia* is maternally inherited by off spring, the population dynamics of *W. glossinidia* and the *Glossina* mtDNA should be similar (Symula *et al.*, 2011). The population expansion event outlined above was mirrored within the symbiont population. Low nucleotide variation, elevated haplotype diversity and a “star-like” haplotype network clearly indicate a population expansion (Slatkin and Hudson, 1991; Venkatesan *et al.*, 2007; Allcock and Strugnell, 2012). Furthermore, similar values for Fu’s  $F_s$  and Tajima’s  $D$ , and an identical Mantel test result show the clear relationship between the symbiont and *Glossina* mtDNA as well as reinforcing the observed population expansion (Tajima, 1989; Fu, 1997). It is important to consider that endosymbiont populations undergo frequent bottlenecks as an inherent aspect of their lifestyle, and therefore the evidence of a population expansion event within the symbiont population must be interpreted cautiously (Funk *et al.*, 2001).

Genetically distant tsetse haplotypes were seen to contain the same symbiont haplotype, a phenomenon observed in other symbiotic relationships (Chong and Moran, 2016; Tseng *et al.*, 2019). Interestingly, it was found that an increase in symbiont diversity was negatively correlated with *AttA* diversity, but positively correlated to *Def* diversity. While this does not

signify causality between the two events, it does imply that the maintenance of AMP nucleotide variation is independent of symbiont variation.

An alternative explanation for the presence of one predominant common haplotype within the *W. g. morsitans 16S* sample is possibly indicative of selfishness within symbiont populations, whereby a genetic variant, which is potentially detrimental to the host, is selected for within the symbiont population (Bennett and Moran, 2015; Rispe and Moran, 2015; Chong and Moran, 2016). However, there is no empirical evidence for this within the sample population.

The high trypanosome infection rate with the tsetse sample population could be a result of a skewed sample dynamic. All samples used in the study were tsetse males, and as juvenile males are the most susceptible to trypanosome infection, it is possible that the 69.35% infection is a drastic overestimation of the true wild infection rate (Distelmans *et al.*, 1982; Otieno *et al.*, 1983). Indeed, a study a much larger (n = 2092), mixed sex sample population, collected at a similar time and location to the flies used in this study, indicated an infection rate of just 6.31% (Shereni *et al.*, 2016).

Although comparison of AMP genetic variation and infection revealed no significant relationship, it is likely that nucleotide variation has little direct influence on infection. The relatively high number of synonymous mutations observed within *AttA* suggests that the potential for protein variation is limited, and therefore the high infection rate is not likely associated to nucleotide polymorphisms. The variation within *Def* offers a slightly different insight, the higher number of non-synonymous sites within the gene indicate that protein variation could be greater, thus having a greater effect on trypanosomal infection.

Interestingly, the association of *W. g. morsitans 16S* variation and infection suggests that direct divergence away from the common *16S* haplotype results in infection. As the presence of *W. glossinidia* directly influences the susceptibility of tsetse flies to trypanosome infection (Kikuchi, 2009; Sasser *et al.*, 2013), it can be hypothesised that genetic variation within the symbiont population may also affect susceptibility. As the mechanics of this symbiosis are not fully understood and literature regarding the impacts of symbiont variation on infection is lacking, the current author presents two possible theories. Firstly, as symbiont populations play a critical role in the development of the immune system in juvenile flies, variation could result in an underdeveloped immune

system increasing susceptibility to infection (Kikuchi, 2009; Symula et al., 2011; Weiss et al., 2012; Sasser et al., 2013). Secondly, although the relationship between *W. glossinidia* and tsetse flies is symbiotic, the AMPs are utilised to regulate symbiont population. If it is possible that symbiont variation could improve the bacterial evasion of the host immune system, reducing expression of AMPs and facilitating trypanosome infection. However, both propositions require substantial further research.

On the other hand, there is evidence to suggest a possible genetic relationship influencing resistance variants within the sample population. Under the premise presented above, genetic variation should result in increased susceptibility to trypanosome infection. However, the increased genetic distance between these two uninfected samples and the common 16S haplotype alludes to a strong resistant sub-strain of *W. g. morsitans* within the population. However, given the small sample size and lack of evolutionary data this remains undetermined.

### 3.5: Conclusion

The aim of this chapter was to evaluate the intraspecific variation of *AttA* and *Def* within a wild tsetse population, and examine the relationship between tsetse nucleotide variation, symbiont genetic variation and trypanosome infection.

Genetic variation within *AttA* and *Def* was found to differ, while population genetics revealed that the wild tsetse population has undergone a recent expansion event, following a bottleneck period. This observation was supported by the relationship with the *W. glossinidia* symbiont, which illustrated almost identical results. However, the exact relationship between tsetse and symbiont genetic variation warrants further research to understand fully. The association of tsetse genetic variation and trypanosome infection yielded inconclusive results, however it is possible that protein variation will offer greater insight into this relationship, as well as the role of natural selection on AMP diversity, a concept that is investigated further in Chapter 4.

## 4: Structural and functional analysis of attacin-A and defensin as a result of single nucleotide polymorphism: Impacts on infection and evolution

### 4.1: Introduction

Protein structure is determined by the folding of amino acid transcripts to form protein secondary structures ( $\beta$ -sheets,  $\alpha$ -helices, and coils), while functionality depends upon the biochemical properties of amino acids. Alteration in the amino acid sequence by non-synonymous mutations can have drastic impacts on the stability, structure and function of proteins (Anfinsen, 1973; Lorch *et al.*, 1999, 2000; Tiede *et al.*, 2006; Ung *et al.*, 2006).

It has long been recognised that genetic mutation is a primary driver of selection and adaptation within organisms (Lynch, 2010; Watari *et al.*, 2010). Non-synonymous mutations resulting in the variation of amino acid sequences are subject to high levels of both purifying and positive selection, while nonsense mutations resulting in the early termination of transcripts are subject to strong purifying selection (Haddrill *et al.*, 2010; Booker *et al.*, 2017; Campos *et al.*, 2017; Chu and Wei, 2019). The impact of synonymous mutations was thought to be minimal and thus only subject to weak selection, however, while these mutations are often referred to as 'silent', evidence suggests they may impact protein function (Supek *et al.*, 2014; Kristofich *et al.*, 2018).

Evolution is heavily influenced by the interactions of an organism with the environment and other organisms (coevolution). The Red Queen hypothesis (Van Valen, 1973), is often associated with predator-prey and parasite-host coevolution (Ebner, 2006; Soares and Yilmaz, 2016). At its most elementary, this hypothesis states that a more advantageous variation will survive and become prominent within a population, while deleterious mutations will be removed. It can be further divided into two contrasting processes: the Red Queen arms race, resulting in the fixation of advantageous alleles within a population; and Red Queen dynamics which promotes genetic variation and the maintenance of multiple alleles within a population via balancing selection (Woolhouse *et al.*, 2002).

Frequency-dependent selection, like most selective process, fall under two types: positive and negative. Positive frequency-dependent selection is the process whereby the fitness of a phenotype or genotype increases with frequency, while negative frequency-dependent selection can be described in two ways, either as an increase of fitness as frequency decreases, or a decrease in fitness as frequency increases (Ayala and Campbell, 1974). In systems where genetic diversity is required, such as parasite-host interactions and coevolution, negative frequency dependent selection is more common and is considered the primary driver of coevolution and the Red Queen dynamic, as it does not enable the fixation of a specific genotype within a population (Burdon *et al.*, 2013; Unckless and Lazzaro, 2016a). Therefore, promoting genetic variation through balancing selection and the Red Queen dynamic.

There are a multitude of factors that affect the evolutionary rate of proteins, including: expression level (Drummond *et al.*, 2005), structural stability (Bloom *et al.*, 2006), function (Cherry, 2010) and site specific variation (McCandlish and Stoltzfus, 2014; Echave *et al.*, 2016; Echave and Wilke, 2017; Jimenez *et al.*, 2018; Marcos and Echave, 2020). It has been observed that majority of non-synonymous mutations have minimal impact on functionality (Zuckerandl and Pauling, 1965; Bloom and Arnold, 2009). However, there are numerous accounts of mutations altering the structure and functional properties of proteins (Mahalingam *et al.*, 2001; Sawai *et al.*, 2002; Bolintineanu *et al.*, 2007; Portelli *et al.*, 2018; Vedithi *et al.*, 2018).

Under the concept of coevolution the antimicrobial proteins responsible for the suppression of pathogen infection would be subject to intense level of selection (Anderson and May, 1982; Niaré *et al.*, 2002; Jiggins and Hurst, 2003; Lehmann *et al.*, 2009). Interestingly, the idea that AMPs are unlikely to undergo positive selection due to their nonspecific nature has been presented on several occasions (Sackton *et al.*, 2007; Simard *et al.*, 2007). However, the adoption of advantageous alleles is not unheard of and significant aspect of the interspecies arms race (Woolhouse *et al.*, 2002; Tennessen, 2005).

Balancing selection has been well documented within AMPS and has become a fundamental aspect of coevolution in most systems (Clark and Wang, 1997; Woolhouse *et al.*, 2002; Chapman *et al.*, 2019). This process of maintaining multiple functional alleles within a population helps to increase genetic diversity within populations while keeping

fitness constant (Pasvol *et al.*, 1978; Woolhouse *et al.*, 2002; Charlesworth, 2006; Key *et al.*, 2014a). This process has been particularly well documented in immune genes across several vertebrate and invertebrate species (Lazzaro and Clark, 2001; Lehmann *et al.*, 2009; Key *et al.*, 2014a; Unckless and Lazzaro, 2016b; Unckless *et al.*, 2016).

In the previous two chapters, both attacin-A (*AttA*) and defensin (*Def*) showed differing inter and intraspecies variation. The interspecies conservation of *Def* supports the idea that AMPs are conserved and maintained through high levels of selective pressure (Chapter 2) (Jiggins and Hurst, 2003; Lehmann *et al.*, 2009). Furthermore, the high level of intraspecific variation and indication of balancing selection observed in Chapter 2, would also suggest that genetic diversity within *G. m. morsitans* *Def* is being maintained by the Red Queen dynamic (Lazzaro and Clark, 2001; Woolhouse *et al.*, 2002; Key *et al.*, 2014b; Chapman *et al.*, 2019). However, the presence of five non-synonymous mutations within the gene fragments could induce a heavy selective on protein variation if functionality is altered (Haddrill *et al.*, 2010; Booker *et al.*, 2017). Interestingly, *AttA* was found to exhibit a higher interspecies variation than intraspecies variation (Chapters 2 and 3), given that *AttA* is part of a larger gene family this could be indicative of concerted evolution within the attacin gene family (Liao, 1999). Although, *AttA* did exhibit some variation it is likely that this is a result of the maintenance of genetic variation through balancing selection though this was not evident from the test of neutrality performed (chapter 3).

#### 4.1.1: Aims and objectives

Four and five non-synonymous mutations were identified in both *AttA* and *Def* respectively (Chapter 3). While this does not guarantee functional variation, it may offer further insight into the evolution of each gene. In this chapter, the aim is to assess the structural and functional consequences of these mutations in relation to infection and protein evolution. Comparisons of nucleotide variation and trypanosome infection indicated little correlation (chapter 3); however, greater insight could be gained from the direct comparison of protein variation to infection. Furthermore, the difference in interspecies nucleotide variation between *AttA* and *Def* suggests that these genes are under different selective pressures (as discussed in chapter 3). Sweeps for natural selection within the transcripts of the two AMP genes could help to quantify this observation.

Additionally, understanding the changes in amino acid properties resulting from non-synonymous mutations can indicate potential consequences on both structure and functionality of the proteins. Detection of any radical changes in biochemical properties can directly lead to alteration in stability, bind affinity and secondary structure of protein variants. As such, the three-dimensional structures of each identified protein variant will be predicted to identify any drastic alterations in secondary structure.

Finally, the aim is also to predict the active site of both AttA and Def within the *G. m. morsitans* protein variants and assess the implications of functional variation. As functional variation is not necessarily predetermined, this would provide a link between protein structural and functional variation and infection rates.



## 4.2: Methodology

### 4.2.1: gDNA extraction, Polymerase Chain Reaction, and bioinformatics analysis:

This chapter builds on the data generated in the previous chapter. For information on the extraction of gDNA, amplification of gene fragments and sequencing, as well as previous analysis including phylogenetic, evolutionary and population genetics analysis please see Chapter 3.

The identification of synonymous and non-synonymous substitutions, undertaken in Chapter 3, showed the location of nucleotide variation within amplified gene fragments. These sequences were translated and aligned using UGENE (Okonechnikov *et al.*, 2012) to illustrate amino acid variation and indicate any potential property changes between the wild AMP variants.

### 4.2.2: Comparison of protein variation and infection

While no relationship was observed between nucleotide variation and infection (see Chapter 3), the relationship between protein variants and infection may offer a deeper insight into the interactions between tsetse and trypanosomes and the ability of an individual to combat infection. This was achieved in two ways: firstly, by directly comparing the total number of samples to the number of infected samples within a protein variant. Secondly, infection frequency was compared to the total number of samples in each protein variant. This provided two observations: firstly, an overall comparison to establish the general trend of infection within the subpopulation (i.e.: does the number of samples directly correlate to infection?) and secondly, to illustrate whether specific variants broke this trend when frequency was assessed (i.e.: do specific variants show reduced infection frequencies despite a higher number of samples?).

### 4.2.3: Indication of Selection: Z-tests

If there is relationship between infection and specific amino acid variation, whether increased susceptibility or resistance, this could be further indicated by the presence of natural selection to remove or promote variation within the population. An initial test for natural selection was conducted using a Z-test in MEGAX (Kumar *et al.*, 2018). This used a

standard Z-test to assess for deviations from the population mean neutral state along a full sequence, this in turn can imply the presence of selective pressure upon a specific sequence. Using the equation 8 (Appendix 2), this enabled the testing of the null hypothesis of neutrality:  $H_0(dN = dS)$ , versus either positive selection:  $H_1(dN > dS)$ , or purifying selection:  $H_2(dN < dS)$ .

#### 4.2.4: Indication of Selection: HyPHy based analysis

Z-tests offer an indication of the selective pressures currently influencing the evolution of a gene by testing the null hypothesis of neutrality against positive and purifying selection (see above). However, in order to determine which codons variants are under selection HyPHy (Hypothesis testing using Phylogenies) was used (Pond and Muse, 2005). Methodologies run using the premise presented by HyPHy: such as SLAC (Single-Likelihood Ancestor Counting) (Kosakovsky Pond and Frost, 2005), FEL (Fixed Effects Likelihood) (Kosakovsky, Pond and Frost, 2005), MEME (Mixed Effects Model of Evolution) (Murrell *et al.*, 2012) and FUBAR (Fast, Unconstrained Bayesian Approximation) (Murrell *et al.*, 2013), can be used to provide a more comprehensive assessment of the selective nature of specific genes and codons. Each method measures selection in a different way, either by assessing the presence of selection across the full sequence or on a codon-by-codon basis. This site based analysis was selected as the focus of this chapter is to assess the evolution of the protein variants, rather than specific branches of the sample population (Weaver *et al.*, 2018).

HyPHy was run using MEGA7 (Kumar *et al.*, 2016) to estimate the number of synonymous (s) and non-synonymous (n) mutations as well as the number of synonymous (S) and non-synonymous (N) sites within each codon using the methods published previously by Felsenstein (1981) and, Muse and Gaut (1994). As in the Z-test, dS and dN represent the number of synonymous substitutions per synonymous site (s/S) and dN the number of non-synonymous substitutions per non-synonymous site (n/N), while dN – dS can be used to detect codons under positive selection. P-values give the probability of rejecting the null hypothesis of neutrality:  $H_0(dN = dS)$  in favour of the presence of positive selection:  $H_1(dN > dS)$  (Suzuki and Gojobori, 1999; Pond and Muse, 2005).

SLAC, FEL, MEME and FUBAR were all run using the Datamonkey webserver (available at: [www.datamonkey.org/](http://www.datamonkey.org/)) (Pond and Frost, 2005; Weaver *et al.*, 2018), all use the premise

detailed above in their methodology however, certain aspects vary within each analytical tool. Both SLAC and FEL predict pervasive selection (Kosakovsky, Pond and Frost, 2005) using a maximum-likelihood approach to determine dS and dN, while SLAC also incorporates a counting method. FUBAR, utilises a Bayesian approach to estimate pervasive selection, and thus a posterior probability value is used to infer significance (Murrell *et al.*, 2013). All three of these approaches use the assumption that selective pressure is constant throughout the phylogeny.

MEME uses a mixed affect likelihood to predicted sites under episodic selection. This approach varies considerable from the others, generating a single value for dS (denoted as  $\alpha$ ) and two values for dN (denoted as  $\beta^-$  and  $\beta^+$ ). Positive selection of a codon is indicated when  $\beta^+ > \alpha$  and is shown to be statistically significant (Murrell *et al.*, 2012).

#### 4.2.5: TreeSAAP

As each amino acid can exhibited fundamentally different properties, which can influence the structural and functional properties of proteins, TreeSAAP was employed to categorise these variations.

TreeSAAP was employed to fully assess the different properties of each amino acid substitution as well as the level variance between the two amino acids. This was done following the user's manual, a NEXUS alignment file was generated in DnaSP (V6) ((Rozas *et al.*, 2017), while a separate phylogenetic tree file was generated in MEGAX (Kumar *et al.*, 2018). By utilising a sliding window approach TreeSAAP compared the total number of amino acid replacements relative to those evolving through neutrality across the Atta and Def phylogeny respectively. The analyses detect the occurrence of 20 physiochemical properties over eight categories, where 1-3 are considered to have little effect on the resultant protein, 4-5 having medium effect and 6-8 have the most substantial impact on the biochemistry and essentially the function of the protein. The significance of each individual amino acid was assessed using a z score, where scores  $> 3.09$  had  $P < 0.001$ .

PROVEAN (Protein Variation Effect Analyzer) online software (available at: <http://provean.jcvi.org/index.php>) (Choi, 2012; Choi *et al.*, 2012) was utilised to predict the impacts of each of the amino acid substitutions on the protein functionality. A PROVEAN

score < -2.5 indicates functional alterations resulting from a specific mutation, while a score above this value indicates a neutral mutation.

#### 4.2.6: Three-dimensional protein modelling

Variation of amino acid properties within a transcript could have direct impact on the structure of a protein. The secondary and tertiary structures of each protein variation was assessed using I-TASSER online software (Zhang, 2008; Roy *et al.*, 2010; Yang *et al.*, 2015). Structural predictions were made as described previously in Chapter 2 (section 2.2.8). All secondary structure models were visualised, and tertiary/surface structures were generated using PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre).

Statistical comparison of these structural predictions was conducted using the DALI online server (Holm, 2020) (see chapter 3; section 2.2.8), and PCA analysis was conducted in PAST3 (Hammer *et al.*, 2001) as detailed in see chapter 3; section 2.2.7, to illustrate the relationship between protein variants.

#### 4.2.7: Active site prediction

The impact of amino acid substitutions and structural variation on protein function could directly influence the function of a protein. Visualisation of binding pockets and residues involved in the active site of protein could help to associate protein structure and infection variance. Prediction of the binding sites within *Glossina* AttA and Def enabled the comparison of the observed wild protein variants to these reference proteins (identified in Chapter 2). Prediction of the active sites was conducted using two online servers: PrankWeb (available at: <https://prankweb.cz>) (Jendele *et al.*, 2019) and FTSite (available at: <https://FTSite.bu.edu>) (Ngan *et al.*, 2012; Kozakov *et al.*, 2015).

As there is no published literature detailing the AttA active site, predicted AttA structures from *G. m. morsitans*, *G. austeni* and *G. pallidipes* (generated in Chapter 2) were submitted to both PrankWeb and FTSite, highlighting the residues forming the predicted active sites. Protein database (PDB) files were submitted to both servers, the resulting predictions were observed in PyMOL. These were aligned in MUSCLE (Multiple Sequence Comparison by Log-Expectation) (Madeira *et al.*, 2019) and a consensus of the binding region indicated the most probable location of the AttA active site. This was further supported by the Hidden Markov Model (HMM) of Arthropod attacin sequence generated in Pfam (El-Gebali *et al.*,

2019). This illustrated the level of entropy at specific codons, as well as conserved regions across the amino acid sequence. The HMM of *G. m. morsitans* AttA was produced using Skylign (available at: <http://skylign.org>) (Wheeler *et al.*, 2014) and an alignment of reference sequences from Chapter 2 and wild sequences generated in this study (Chapter 3).

Prediction of the *Glossina* Def active site was undertaken by aligning the previously published Def active sites from *Allomyrina dichotoma* (AAB36306) and *Oryctes rhinoceros* (BAA36401) to the *G. m. morsitans* Def sequence identified in chapter 3. This alignment was conducted using MUSCLE (Madeira *et al.*, 2019) and indicated the residues likely involved in the Def active site.

Prediction of the active site within the wild protein variants was undertaken in the same way. Structural predictions (generated in section 5.2.5) were submitted to both PrankWeb and FTSite. The resulting structures illustrated any variation in binding pockets between the structures of protein variants.

### 4.3: Results

Non-synonymous nucleotide variations within AttA and Def, identified in Chapter 3, influence amino acid variation within the protein sequence. Protein variants in this chapter will be named according to the amino acid substitutions responsible for variation, i.e., Def-E42 indicates the substitution of Glycine (Gly) for Glutamic acid (Glu) reduce at codon 42. Those sequences that contain no amino acid substitutions were named simply, AttA and Def.

Four non-synonymous mutations were observed within the AttA nucleotide sequence at codons 35, 43, 49 and 86 (Fig. 4.1A). An alignment of amino acid sequences showed the presence of five AttA protein variants: AttA, AttA-N35, AttA-T43, AttA-D49 and AttA-A86. Defensin nucleotide sequences exhibited five non-synonymous mutations at four codons, 18, 42, 45 and 82 (Fig. 4.1B). Codon 18 was found to exhibit two variants with a C-G and C-T substitution resulting in the presence of Ser18 and Ile18 variants respectively. Translation of the nucleotide sequence indicated the presence of 11 Def protein variants: Def, Def-S18, Def-I18, Def-E42, Def-F45, Def-I82, Def-S18/E42, Def-S18/F45, Def-S18/I82, Def-I18/F45 and Def-S18/F45/I82.

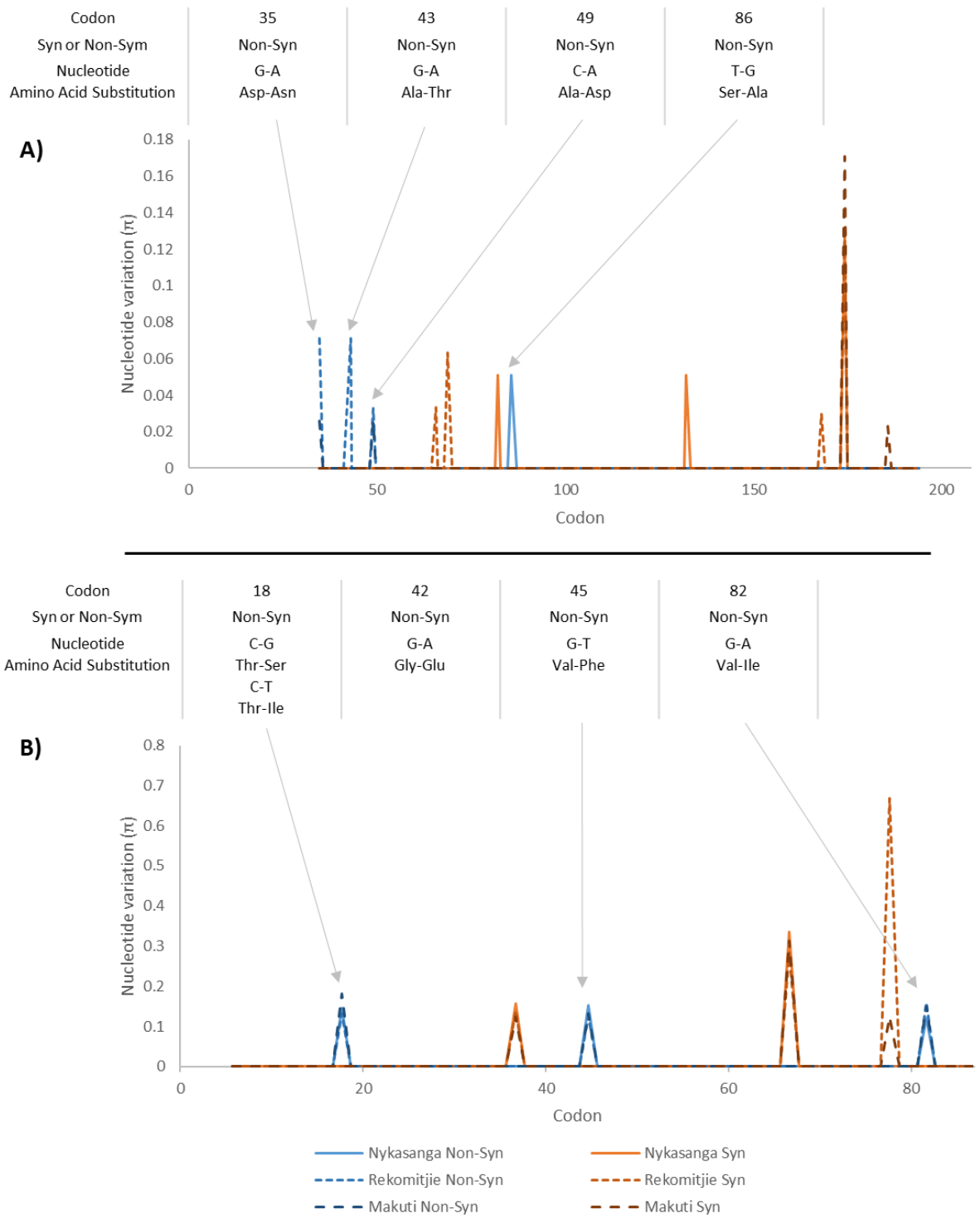


Figure 4.1: All synonymous and non-synonymous mutations within the AttA (A) and Def (B) genes. Sliding window graphs were produced in the previous Chapter (Chapter 3, section 3.2.8).

### 4.3.1: Protein variation and *Trypanosoma* infection

A direct comparison of the number of trypanosome infected samples exhibiting each protein variant illustrated a clear correlation between sample size and infection (Fig. 4.2). The imbalance of samples exhibiting AttA protein variants makes this correction inevitable ( $R^2 = 0.998$ ) (Fig. 4.2A) however, a strong correlation was also observed between the number infected samples within Def protein variants ( $R^2 = 0.896$ ) (Fig. 4.2B). Therefore, unsurprisingly as the number of samples exhibiting a protein variant increase so does the number of infected samples.

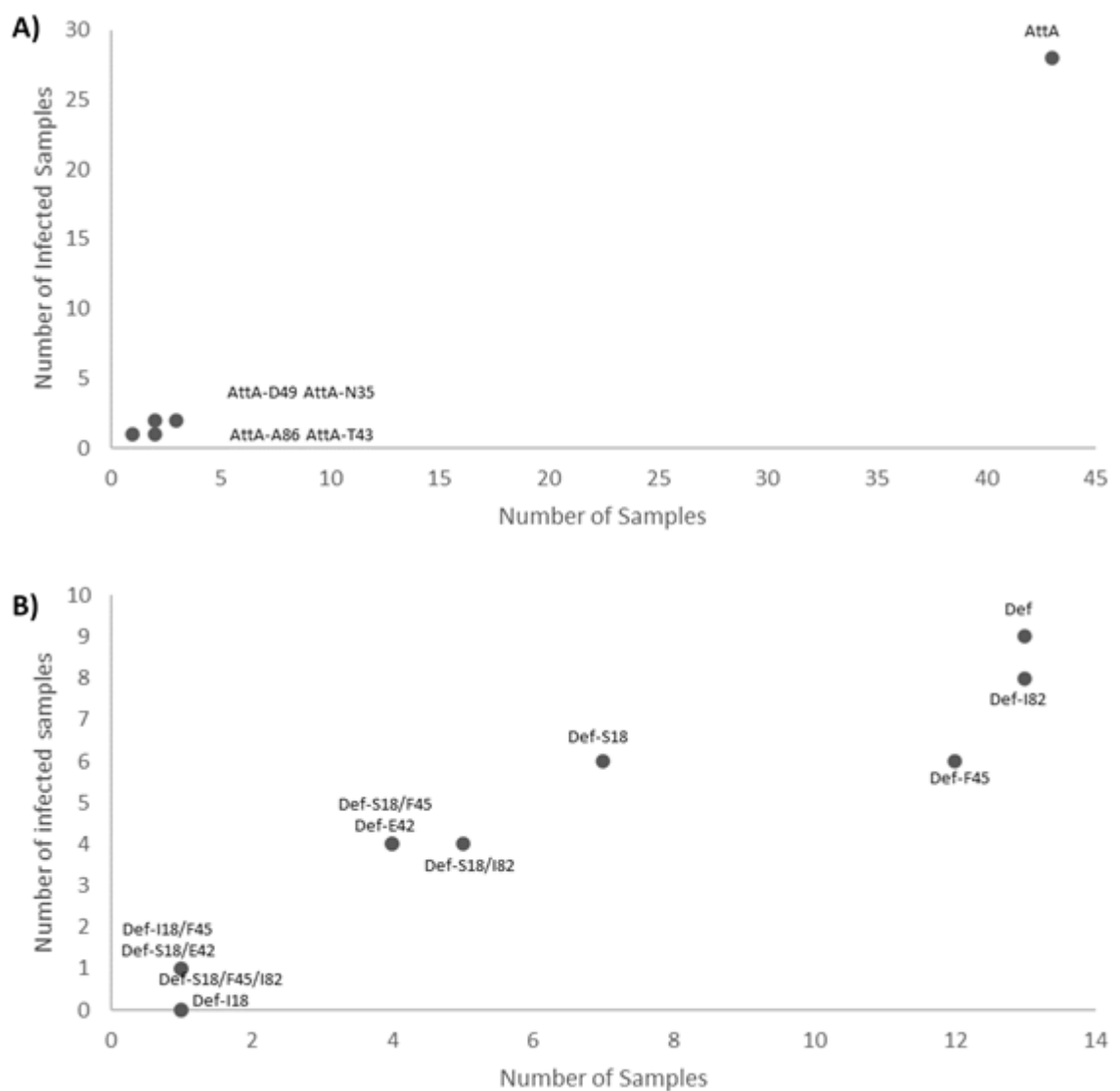


Figure 4.2: A comparison of the number of samples exhibiting each protein variant, and the number of infected samples. A) compares the AttA variants, while Def is shown on graph B.



Interestingly though, a comparison of infection frequency within protein variants shows two differing results (Fig. 4.3). Attacin-A indicated a very minor negative correlation between sample size and infection frequency (Fig. 4.3A). However, regression analysis does not suggest there is any relationship between sample size and infection ( $R^2 = 0.0892$ ;  $P > 0.05$ ). Defensin on the other hand, indicates a slight increase in infection frequency as sample size increase (Fig. 4.3B). Although, regression analysis once again did not indicate any relationship between infection and sample size ( $R^2 = 0.00225$ ;  $P > 0.05$ ).

Curiously, if only variants observed in two or some samples are considered the general trend shifts to a clear negative relationship between sample size and infection frequency (Fig. 4.3B), which is strongly supported by regression analysis ( $R^2 = 0.794$ ;  $P = 0.007$ ). Perhaps, more remarkably is the observation that three of four high frequency variants (except for Def-E42) exhibited the S18 variation, while the low frequency variations all exhibit the wild T18 variant (Fig. 4.3B).

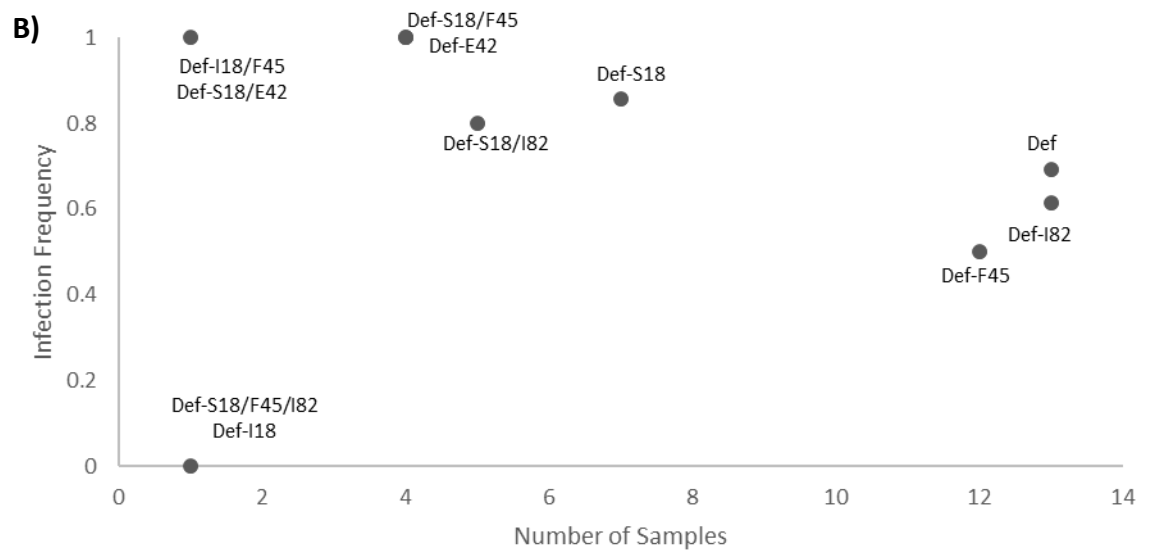
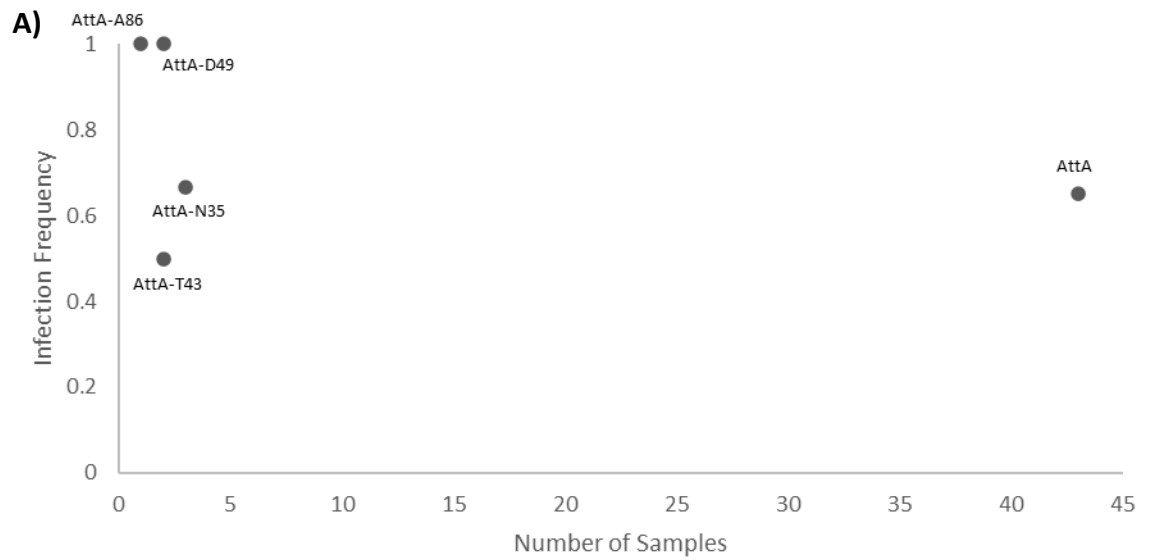


Figure 4.3: A comparison of the number of samples exhibiting each protein variant, and the frequency of infection within each variant. A) Illustrates the relationship between the number of samples expressing each AttA variant and infection rates. B) Shows the relationship between the number of samples expressing each Def variant and infection frequency.

## 4.3.2: Indicators of natural selection

### 4.3.2i: Z-tests

An initial estimation of natural selection was conducted using a Z-test in MEGAX (Kumar *et al.*, 2018). This indicated the presence of both positive and purifying selection, as well as neutral evolution within the sample population (Table 4.1).

When submitted to the MEGA analysis, AttA showed no indication of either positive or purifying selection with all P-values above the statistically significant  $P = .05$  threshold. Interestingly, neutrality was not detected either, though the average P values was lower than that observed for positive and purifying selection.

In contrast, the Def gene indicated far greater selective pressure within the sample population. Neutrality was detected between Haplotypes 3 and 13 ( $P = .045$ ), and Haplotypes 8 and 23 ( $P = .04$ ). Evidence of positive selection was indicated between Haplotypes 3 and 13 ( $P = .023$ ), 15 ( $P = .042$ ) and 17 ( $P = .042$ ), in addition to Haplotypes 7 and 9 ( $P = .042$ ), and Haplotypes 5 and 10 ( $P = .042$ ). Evidence of purifying/negative selection was identified within considerably more haplotypes than positive or neutral selection. Purifying selection was identified between Haplotypes 1 and 11 ( $P = .039$ ); as well as Haplotype 2 and haplotypes 11 ( $P = .031$ ), 12 ( $P = .042$ ) and 25 ( $P = .039$ ). Purifying selection was indicated between Haplotypes 5-10 and Haplotype 23-25 ( $P < .045$ ). Although the P-values corresponding to Z-values of purifying selection between Haplotypes 2 and 12; 5 and 25; 8 and 24; 9 and 25; and 10 and 24, were not statistically significant, they were threshold ( $.051 \leq P \leq .065$ ). These threshold values suggest that selective pressures between these haplotypes by influence evolution although not statistically significant in the current data set.

Table 4.1: A matrix of identified Def haplotypes and indicated selection between them. Lower left shows values indicating Positive selection while purifying is indicated by values in the top right. Any statistically significant Z-scores ( $P < .05$ ) are highlighted. Where neutrality was detected, it is represented by \*.

Haplotype	Purifying Selection											
	1	2	3	4	5	6	7	8	9	10	11	12
1		0.25	0.17	0.06	0.17	0.12	0.12	0.08	0.17	0.12	0.04	0.06
2	1.00		0.06	0.17	0.47	0.36	0.36	0.25	0.47	0.36	0.03	0.04
3	1.00	1.00		0.47	0.17	0.23	0.12	0.17	0.30	0.23	0.36	0.47
4	1.00	1.00	1.00		0.25	0.16	0.36	0.25	0.25	0.36	0.12	0.08
5	1.00	1.00	1.00	1.00		1.00	1.00	1.00	1.00	1.00	0.09	0.06
6	1.00	1.00	1.00	1.00	0.16		1.00	1.00	1.00	1.00	0.06	0.04
7	1.00	1.00	1.00	1.00	0.16	0.08		1.00	1.00	1.00	0.06	0.09
8	1.00	1.00	1.00	1.00	0.08	0.16	0.16		1.00	1.00	0.04	0.06
9	1.00	1.00	1.00	1.00	0.08	0.16	0.04	0.08		1.00	0.09	0.06
10	1.00	1.00	1.00	1.00	0.04	0.08	0.08	0.16	0.16		0.06	0.09
11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		1.00
12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.16	
13	1.00	1.00	0.02	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
14	1.00	1.00	0.43	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.08	0.16
15	1.00	1.00	0.04*	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	1.00	1.00	0.08	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
17	1.00	1.00	0.04	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
18	0.08	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
19	1.00	1.00	0.08	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
21	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
22	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
23	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
24	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 4.1 cont.: A matrix of identified Def haplotypes and indicated selection between them. Lower left shows values indicating Positive selection while purifying is indicated by values in the top right. Any statistically significant Z-scores ( $P < .05$ ) are highlighted. Where neutrality was detected, it is represented by \*.

Haplotype	Purifying Selection												
	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0.17	0.09	0.12	0.08	0.12	1.00	0.17	0.36	0.36	0.25	0.08	0.17	0.12
2	0.12	0.06	0.09	0.06	0.09	0.25	0.06	0.12	0.12	0.08	0.06	0.06	0.04
3	1.00	1.00	1.00	1.00	1.00	0.17	1.00	0.36	0.36	0.25	0.17	0.17	0.12
4	0.25	0.12	0.25	0.25	0.16	0.06	0.25	0.12	0.17	0.17	0.06	0.06	0.09
5	0.17	0.09	0.17	0.17	0.12	0.17	0.17	0.09	0.12	0.12	0.04	0.04	0.06
6	0.12	0.06	0.12	0.12	0.08	0.12	0.12	0.06	0.09	0.09	0.03	0.03	0.04
7	0.23	0.12	0.17	0.12	0.17	0.23	0.23	0.12	0.12	0.09	0.03	0.06	0.04
8	0.17	0.09	0.12	0.08	0.12	0.17	0.17	0.09	0.09	0.06	0.02	0.04	0.03
9	0.08	0.04	0.17	0.17	0.12	0.17	0.17	0.09	0.12	0.12	0.04	0.04	0.06
10	0.12	0.06	0.17	0.12	0.17	0.23	0.23	0.12	0.12	0.09	0.03	0.06	0.04
11	0.36	1.00	0.25	0.16	0.25	0.09	0.36	0.17	0.17	0.12	0.16	0.36	0.25
12	0.25	1.00	0.25	0.25	0.16	0.06	0.25	0.12	0.17	0.17	0.25	0.25	0.36
13		0.16	1.00	1.00	1.00	0.17	1.00	0.36	0.47	0.47	0.17	0.17	0.23
14	1.00		0.36	0.36	0.25	0.09	0.36	0.17	0.23	0.23	0.36	0.36	0.47
15	0.08	1.00		1.00	1.00	0.17	1.00	0.36	0.25	0.36	0.12	0.17	0.17
16	0.08	1.00	0.16		1.00	0.17	1.00	0.36	0.36	0.25	0.08	0.17	0.12
17	0.16	1.00	0.16	0.16		0.12	1.00	0.25	0.36	0.36	0.12	0.12	0.17
18	1.00	1.00	1.00	1.00	1.00		0.08	0.16	0.25	0.25	0.17	0.08	0.12
19	0.08	1.00	0.08	0.08	0.16	1.00		0.16	0.25	0.25	0.17	0.08	0.12
20	1.00	1.00	1.00	1.00	1.00	1.00	1.00		1.00	1.00	0.36	0.16	0.25
21	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.16		1.00	0.36	0.25	0.25
22	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.16	0.16		0.25	0.25	0.16
23	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		1.00	1.00
24	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.08		1.00
25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.16	0.16	

#### 4.3.2ii: Codon based selection: HyPhy, FEL, SCAL, FUBAR and MEME

Having obtained an insight into the selective interactions between each haplotype, codon-based selection analysis was conducted employing HyPhy, FEL, SCAL, FUBAR and MEME to assess which codons were under selection. Each of these methodologies utilise the dN – dS statistical test to screen for signs of positive and purifying selection. Positive values indicated an abundance of non-synonymous mutations and positive selection, while negative values indicate purifying selection (Fig 4.4). While all these methods can be used to predict positive selection MEME cannot be used to accurately indicate purifying selection.

There was no indication of positive selection at any of the mutation sites within the AttA gene fragment. However, HyPhy, SLAC, FEL and FUBAR all indicated varying degrees of purifying selection at the synonymous sites (Fig. 4.4A). SLAC indicated no significant purifying selection at any synonymous sites. FEL indicated two statistically significant sites ( $P < .05$ ) of purifying selection at codons 174 ( $P = .006$ ) and 186 ( $P = .034$ ). Finally, FUBAR indicated significant purifying selection at all synonymous sites (posterior probability  $> 0.9$ ). While all synonymous sites were statistically significant, codon 174 showed a strong purifying pressure, with posterior probability value 0.99.

HyPhy, SLAC, FEL and MEME all indicated that the four non-synonymous mutations were under positive selection, however only codon 18 was found to be statistically significant by all methods ( $P = 0.043, 0.034, 0.012$  and  $0.02$  respectively) (Fig. 4.4B). FUBAR identified three statistically significant (posterior probability value  $> 0.9$ ) sites of positive selection at codons 18, 42 and 45 (Fig. 4.4B). While codon 82 was also identified to be under positive selection, though the posterior probability value of 0.869 narrowly missed 0.9 significance cut off.

Purifying selection was identified at all synonymous substitution sites with varying statistical significance. SLAC identified purifying selection at codons 68 and 78 with significant  $P$  values of  $P = 0.041$  and  $0.001$  respectively (Fig. 4.4B). Codon 67 also presented signs of purifying selection though the  $P$  value was fractionally insignificant ( $P = 0.0532$ ). FEL indicated statistically significant purifying selection at all synonymous substitution sites, codon 37  $P = 0.035$ , codons 67 and 68  $P = 0.011$  and  $0.019$ , respectively, and codon 78  $P = 0.001$ . FUBAR indicated the presence of purifying selection at all four sites with a

high posterior probability of > 0.9. While all posterior probability values were high, codon 78 illustrated a posterior probability value of 1, suggesting a strong purifying selection pressure at codon 78 (Fig. 4.4B).

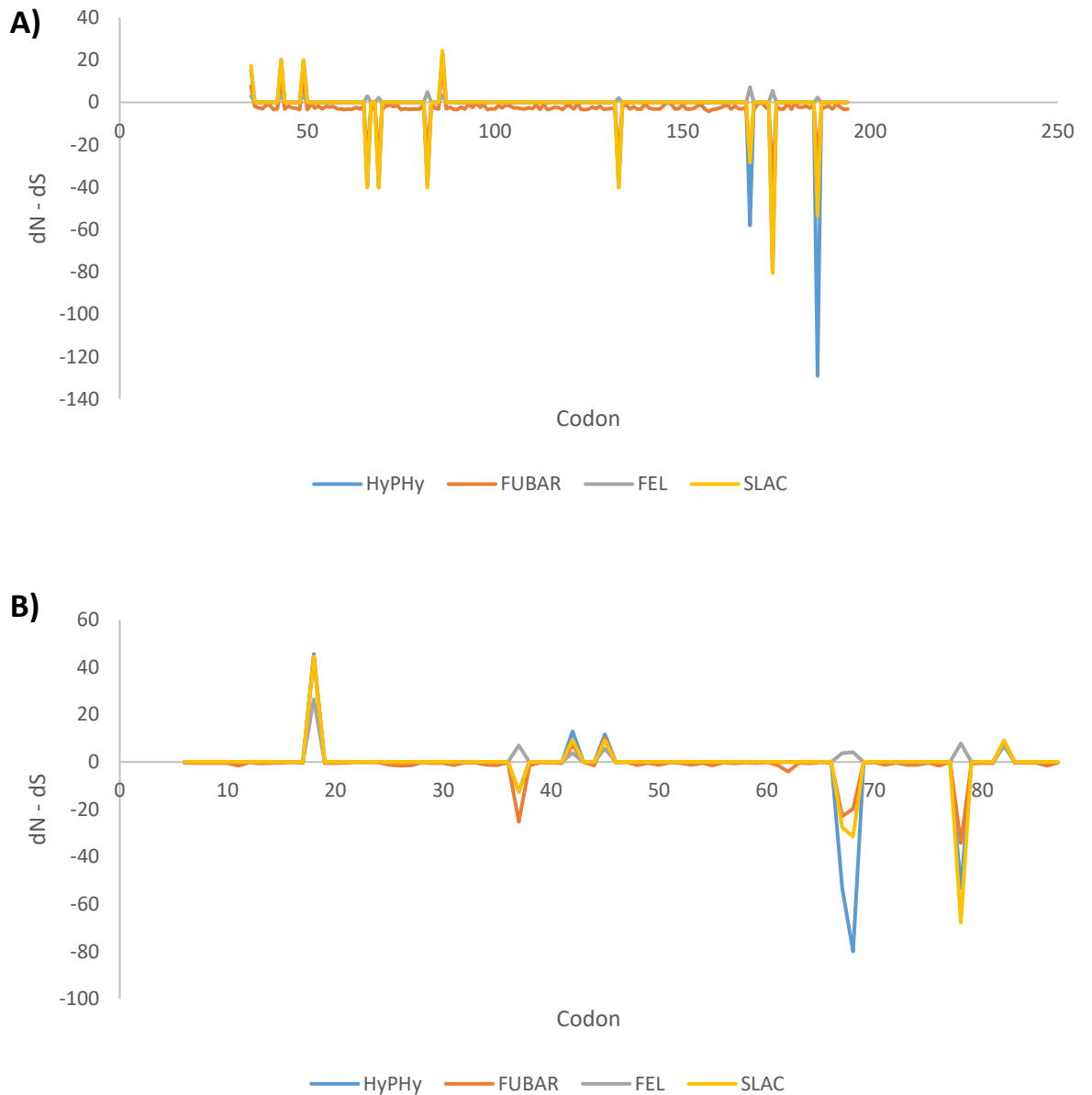


Figure 4.4: A graph showing the dN – dS values at each codon of a protein sequence fragment: A) AttA; B) Def. Positive values indicate an overabundance of non-synonymous mutations and positive selection, while negative values indicate purifying selection and an abundance of synonymous mutations. Each line represents a different methodology.

### 4.3.3: Characterisation of significant codons.

TreeSAAP identified seven significant property changes across the four non-synonymous sites within the AttA fragment. Changes to the:  $\alpha$ -helical tendencies, composition, Compressibility, polar requirement, polarity, turn tendencies and power to be at the C-terminal, of these all but composition were identified as being radical changes (TreeSAAP categories  $6 \leq 8$ ) (Table 4.2). The most radical of these changes was detected at Codon 49 where the Asp49 variant resulted in four radical changes, compared to just two at Ala86 and one at each of Asn35 and Asp49 (Table 4.2).

Defensin exhibited 24 significant property changes across the five non-synonymous sites with radical changes in:  $\alpha$ -helical tendencies, bulkiness, buriedness, coil tendencies, compressibility, equilibrium constant ( $K_c$ ), long-range non-bonded energy, surrounding hydrophobicity and thermodynamic transfer hydrophobicity (Table 4.2). The vast majority of these radical changes are observed in the Ile18 variant, which exhibited eight radical property changes. Three radical property changes were also observed in the Glu42 variant (Table 4.2) Interestingly, TreeSAAP observed no radical changes in the Ser18, Phe45 and Ile82 variations.



Table 4.2: All radical amino acid property changes in both AttA and Def. The property change, category ( $6 \leq 8$ ) and *P*-value of each change are given.

Codon variant	Property	Category	P-value
<b>AttA</b>			
<b>N35</b>	$\alpha$ -helical tendencies	6	0.01
<b>T43</b>	Compressibility	6	0.05
	Polar requirement	7	0.01
	Polarity	6	0.05
	Turn Tendencies	6	0.01
<b>D49</b>	Power to be at the C-terminal	6	0.01
<b>A86</b>	$\alpha$ -helical tendencies	6	0.01
	Turn tendencies	6	0.01
<b>Def</b>			
<b>S18</b>	N/A	N/A	N/A
<b>I18</b>	Bulkiness	6	0.05
	Buriedness	6	0.05
	Coil tendencies	6	0.01
	K <sub>c</sub> (ionization of COOH)	7	0.01
	Long-range non-bonded energy	6	0.05
	Solvent accessible reduction ratio	8	0.05
	Thermodynamic transfer hydrophobicity	7	0.001
	Surrounding hydrophobicity	6	0.05
<b>E42</b>	$\alpha$ -helical tendencies	8	0.05
	Coil tendencies	6	0.01
	Compressibility	7	0.001
<b>F45</b>	N/A	N/A	N/A
<b>I82</b>	N/A	N/A	N/A

Interestingly, despite the number of radical changes detected by TreeSAAP no deleterious amino acid substitutions were found by PROVEAN (Fig. 4.5). The four amino acid substitutions identified within AttA showed an average PROVEAN score of -1.187, with the lowest score (-1.463) being observed at the Asp49 codon (Fig. 4.5A). Defensin illustrated a higher average PROVEAN score of -0.668, however the PROVEAN score exhibited by the Phe45 substitution was considerably lower at -2.223 despite showing no radical property changes (Fig. 4.5B; Table 4.2).

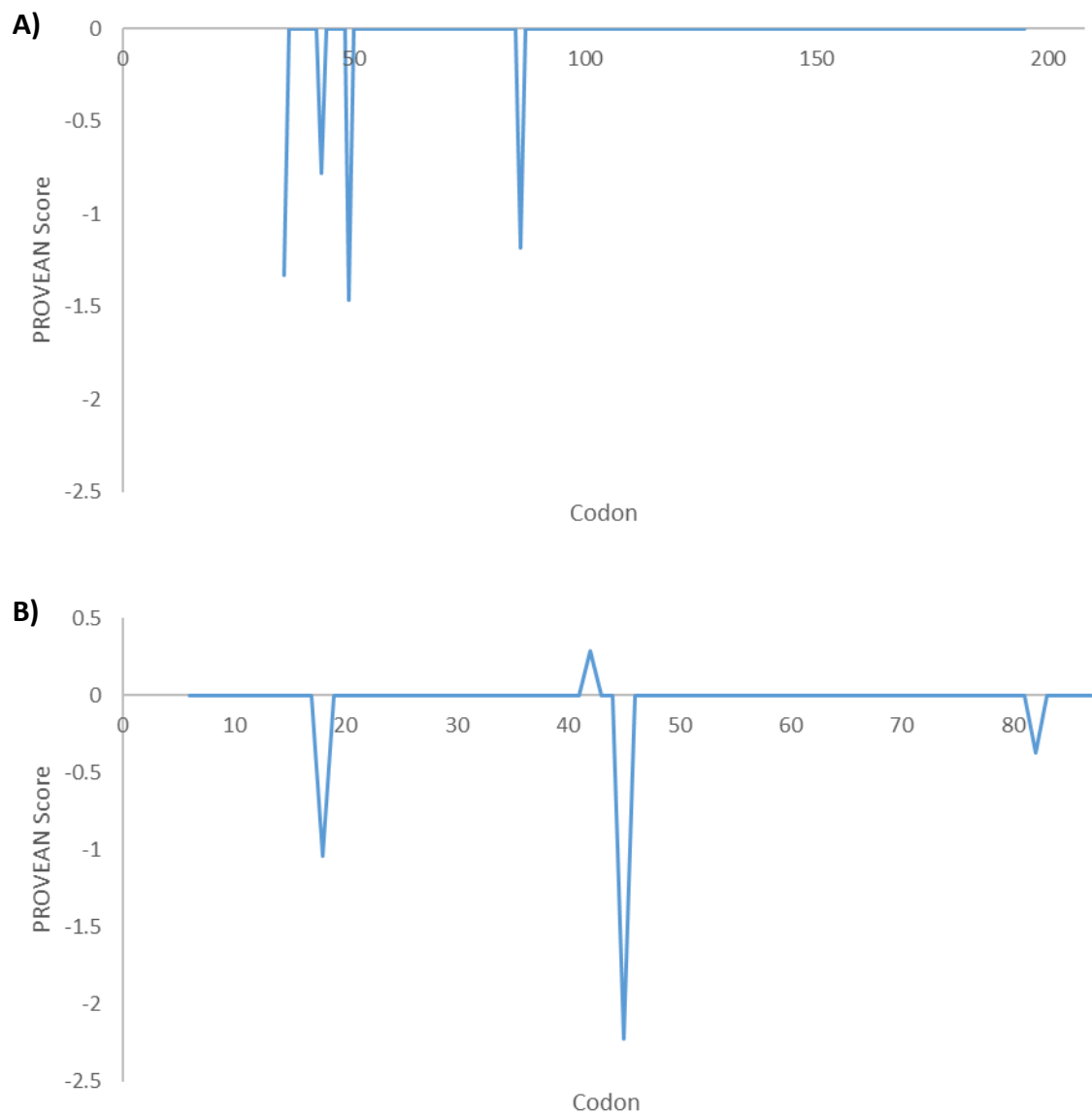


Figure 4.5: PROVEAN scores of each codon along the AttA (A) and Def (B) protein fragments. Amino acid substitutions with a score  $\leq -2.5$  are considered deleterious, any score  $> -2.5$  is considered neutral.

#### 4.3.4: Structural and functional analysis

While the TreeSAAP results illustrated the theoretical impacts of each amino acid substitution, the physical structural changes of each protein variant were examined using I-Tasser structural prediction software. This illustrated the differences between the five AttA variants and the 11 Def variants (Tables 4.3 and 4.4).

#### 4.3.5: Attacin-A

##### 4.3.5i: Prediction and impacts of variation on protein structure

The AttA structures predicted from the wild *G. m. morsitans* samples illustrated similar characteristics to the structures predicted in Chapter 2 (section 2.3A.7iii). The secondary structure consists of a coiled N-terminal leading to a series of anti-parallel  $\beta$ -sheets, which form a concave structure terminating in a final coiled section. The wild attacin structure (Table 4.3: AttA) exhibited seven consecutive anti-parallel  $\beta$ -sheets, with extended coiled structures for both terminals. The number of  $\beta$ -sheets forming the main body of the proteins was found to differ between variants, with AttA-A86 exhibiting six  $\beta$ -sheets in total, separated into two clusters of three anti-parallel sheets, while AttA-N35 and AttA-D49 both exhibit ten consecutive  $\beta$ -sheets (Table 4.3). While the majority of AttA variants all exhibit the same coiled terminals, AttA-A86 exhibited a single  $\alpha$ -helix between codons N14 and V17 (Table 4.3).

As indicated by the secondary structures the surface structure of *G. m. morsitans* AttA exhibits an open channel structure with the N- and C-terminals both curving inwards (Table 4.3, Supplementary Figure 8 Appendix 5), this presents a potential location for the active sites of AttA within this channel. Differences between the AttA variants surface structure appear to be primarily restricted to the distance between the N- and C-terminals, resulting in either a tighter or looser channel. Both AttA-N35 and T43 variants appear to show a looser channel shape with more space between the terminal coils (Table 4.3, Supplementary Figure 8 Appendix 5), while AttA-A86 and N187 illustrate tighter channels. Interestingly, AttA-D49 appears to illustrate both of these traits opening one end of the channel while closing the other (Table 4.3, Supplementary Figure 8 Appendix 5). Variants exhibiting a more open channel structure could allow easier access to the active sites, though the lack of positive selection within the AttA sample suggests this is unlikely.

Table 4.3: Predicted 3D structures of each of the *G. m. morsitans* Def variants. The name of each variant, the haplotypes exhibiting each structure and the amino acid substitutions responsible for the variation are given. PDB files were produced using the I-TASSER server (Yang and Zhang, 2015) and visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre). All coil structures are shown in green;  $\alpha$ -helices are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

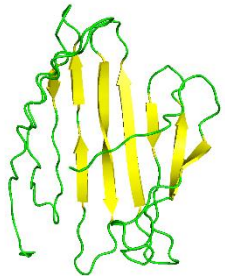
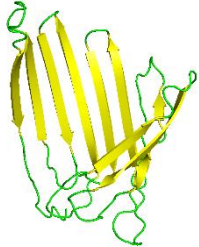
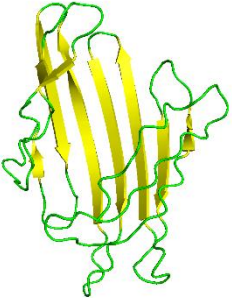
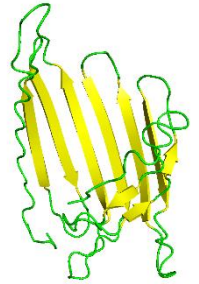
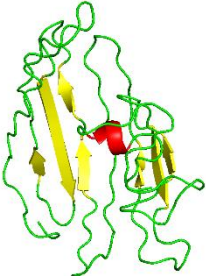
Variant	haplotypes	Amino Acid substitution	Structure	Variant	haplotypes	Amino Acid substitution	Structure
AttA	2, 3, 4, 5 and 11	N/A		AttA-N35	8 and 9	D35 → N35	
AttA-T43	6 and 7	A43 → T43		AttA-D49	10	A49 → D49	

Table 4.3 cont.: Predicted 3-Dimensional structures of each of the *G. m. morsitans* Def variants. The name of each variant, the haplotypes exhibiting each structure and the amino acid substitutions responsible for the variation are given. PDB files were produced using the I-TASSER server (Yang and Zhang, 2015) and visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre). All coil structures are shown in green;  $\alpha$ -helices are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

Variant	haplotypes	Amino Acid substitution	Structure
AttA-A86	1	S86 $\rightarrow$ A86	

Principle component analysis (PCA) shows a clear separation of the AttA variant from the other isoforms (Fig. 4.6), AttA-A86 also shows a clear separation from the other variants while AttA-N35, AttA-T43 and AttA-D49 form a cluster. The separation of AttA-A86 is likely due to the lower number of observed  $\beta$ -sheets and the presence of a  $\alpha$ -helix at the N-terminal (Table 4.3). Similarly, the divergence of AttA from the other variants is likely due to the lower number of  $\beta$ -sheets exhibited by the protein (Table 4.3).

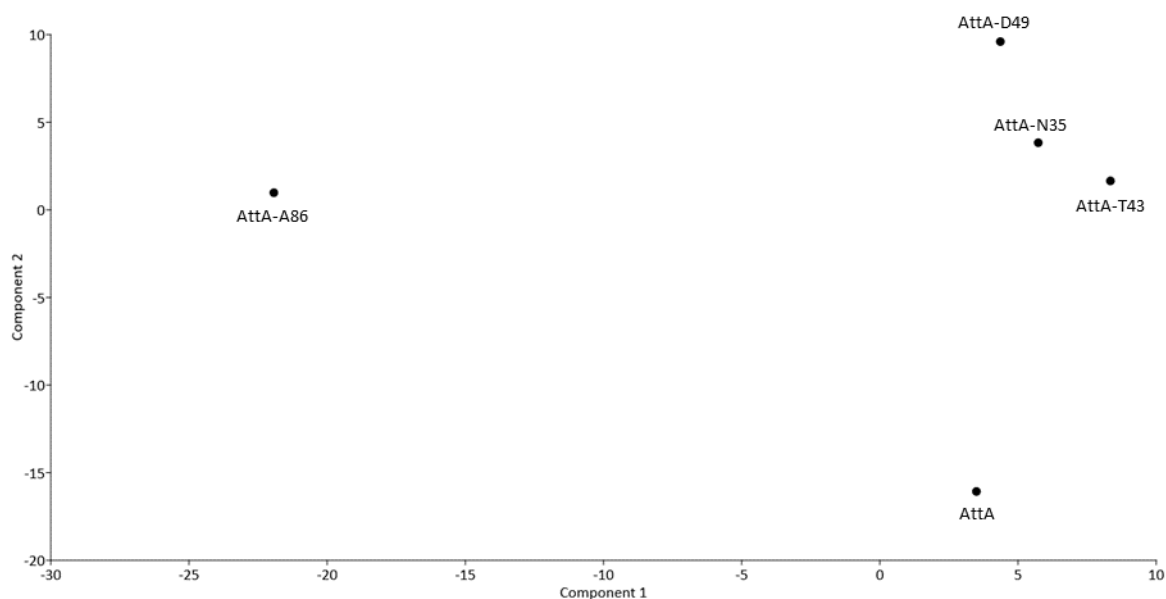


Figure 4.6: Principle component analysis (PCA) plot of wild attacin variants using the first and second Principle components (Eigenvalues: PC1 = 153.66 (40.26 % variance); PC2 = 92.22 (24.16 % variance)). Z-values were calculated, and a matrix produced using DALI (Holm, 2020). PCA analysis was conducted in PAST3 (Hammer *et al.*, 2001).

#### 4.3.5ii: Prediction of and variation within the attacin-A active site

While PROVEAN indicated there was no impact on protein function as a result of the observed amino acid substitutions, structural variation could inhibit or promote ligand binding. The active site of *G. m. morsitans* was predicted to be between Pro65 and His121 (Fig. 4.7). This region was consistently predicted as a binding region in all three of the *Morsitans* group species, by both PrankWeb and FTSite (Fig. 4.7). The predicted binding site appears to interact with residues at the N-terminal to stabilise the binding region and are arranged in four clusters of residues within the identified region (Fig. 4.7A). This was supported by the Arthropod Attacin N- and C-domain HMM logos, which indicated two peaks of high entropy at the start and end of this region, with a low level of entropy between (Fig. 4.7B).

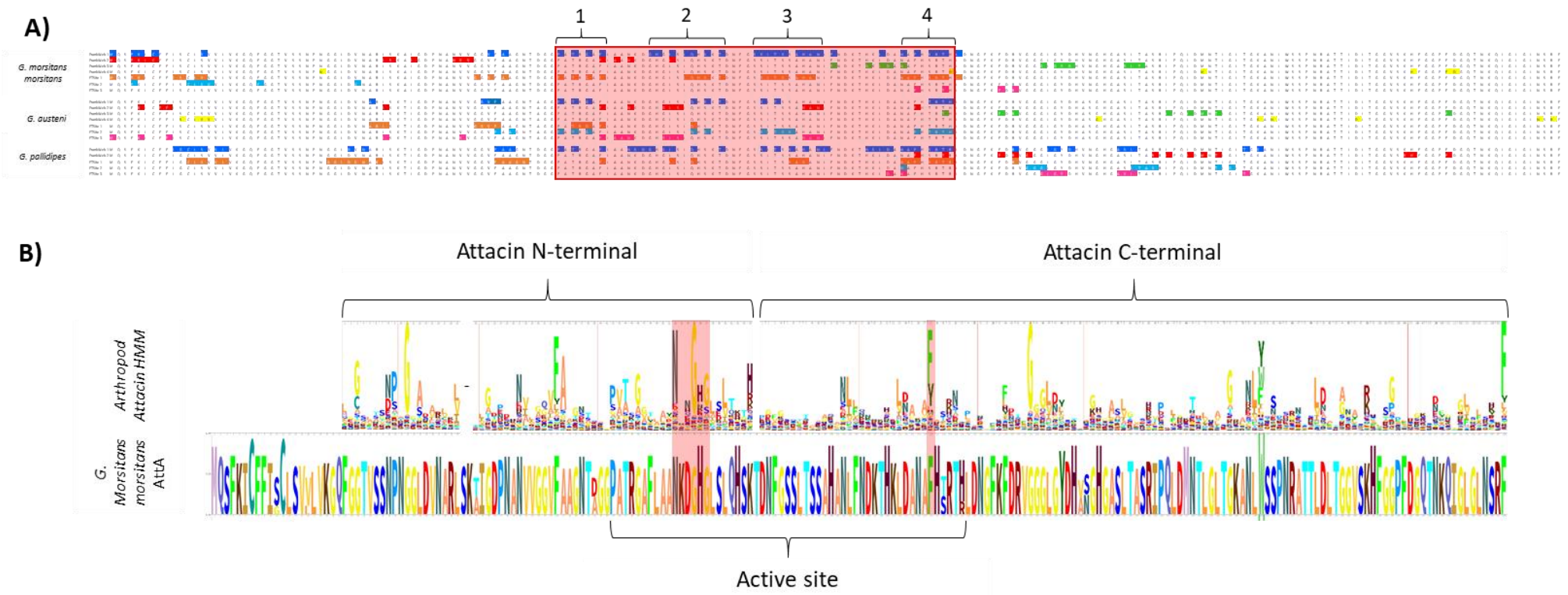


Figure 4.7: A) The alignment of AttA transcripts from *G. m. morsitans*, *G. austeni* and *G. pallidipes*. Highlighted residues show the amino acids identified as being involved in a predicted binding site. The most lightly conserved position of the AttA binding site is shown in the red box. B) The Hidden Markov Model of the two Arthropod Attacin domains (N- and C-terminal domains) aligned to a simple HMM of the *G. m. morsitans* AttA sequence. The areas highlighted in red indicated the areas of high entropy within the predicted binding site. The Arthropod HMM logo was downloaded from Pfam prior to alignment.

Predictions of binding sites within the wild AttA variants indicated some variation away from the predicted region observed in figure 4.7. Three wild variants (AttA, AttA-N35 and AttA-T43) exhibited a number of smaller predicted sites than those observed in the reference sequences (Fig. 4.8). Despite this, similarities can be observed between each variant and the reference predictions, primarily the recognisable four cluster pattern within the Pro65 and His121 region (Fig. 4.7). Although these clusters are not as clearly defined, this is possibly due to the absence of stabilising residues at the N-terminal. Interestingly, both AttA-D49 and AttA-A86 exhibited considerably larger active sites extending beyond His121 and encompassing the majority of the concave protein surface (Fig. 4.8).



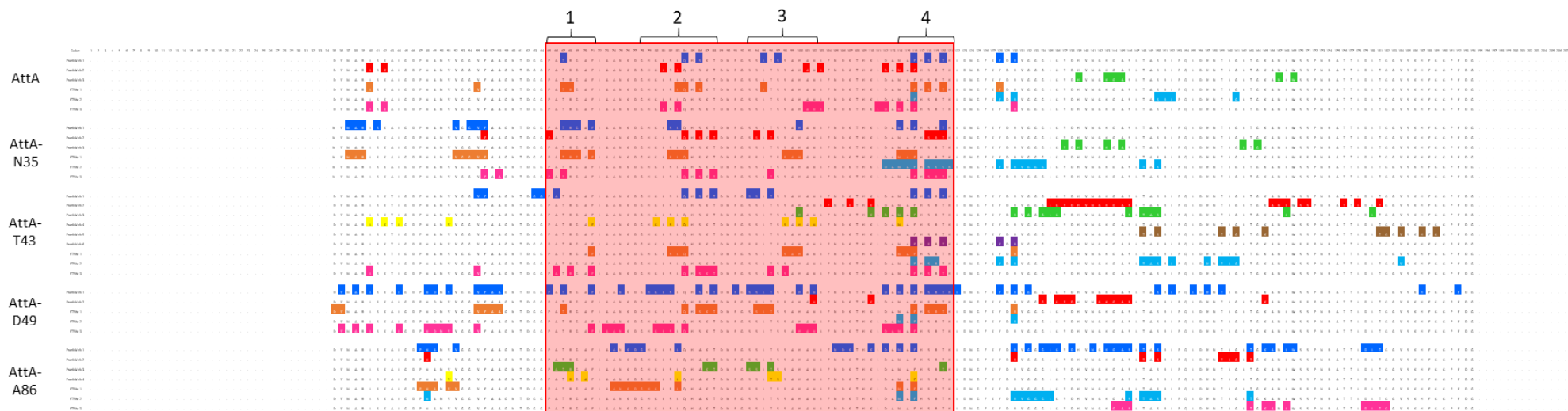


Figure 4.8: The alignment of all wild *G. m. morsitans* AttA protein variants showing the predicted active site region in the red box. Coloured residues were identified as part of predicted active sites with each colour denoting a different binding site or region.

Both PrankWeb and FTSite predicted three potential binding sites within the wild *G. m. morsitans* AttA variant. PrankWeb sites 1 and 2 showed a high degree of similarity to FTSite predictions 1 and 3 (Fig. 4.9), while FTSite site 2 was located at the C-terminal rather than within the predicted binding pocket and PrankWeb site 3 was discounted as this was located on the convex surface of the protein near the C-terminal (Fig. 4.9). PrankWeb site 1 was found to be ten residues long (T67, Q84, S86, L95, S97, F116, S118, T120, F128 and R130) and shared eight of these with FTSite 1 which was 12 residues in length (L40, V55, T67, R68, L83, Q84, S86, L95, F116, S118, T120 and F128) and followed a similar clustering to that observed previously. PrankWeb 2 and FTSite 3 shared nine residues (L40, K42, L81, L83, A101, L103, D112, N114 and F116), these also followed a similar clustering to that observed in figure 4.9.

The predicted binding sites within the AttA-N35 variant showed some similarity to those of AttA, however, they were generally larger (Fig. 4.9 and 4.10). PrankWeb site 1 and FTSite 1 shared 16 residues (N37-R39, V52, G54-F56, T67-G69, F71, S82, L83, H100, N114 and F116 = 18) and clearly follow a similar structure to that observed in the AttA reference sequences. PrankWeb site 2 and FTSite 3 also show a high level of conservation sharing nine residues between them (F56, P65, Q84, S86, T88, T96 and S118-T120). Again, these predicted sites follow the four-cluster pattern, though not as clearly. As observed in the 'wild' AttA variant PrankWeb site 3 was exhibited on the convex surface and was excluded from the study (Fig. 4.10).

The AttA-T43 variant exhibited a far greater degree of variation between predicted binding sites than the other samples. Interestingly, PrankWeb identified six potential binding sites of which only two, PrankWeb sites 1 and 4, show any clear similarity to the predicted binding site of *G. m. morsitans* (Fig. 4.9 and 4.11). While there is less conservation between PrankWeb and FTSite predictions, both PrankWeb site 1 and FTSite 4 consist of 14 residues and share a total of 9 residues between them (V55, A66, Q84, S86, T88, T96, F116, S118 and T120). Additionally, PrankWeb site 4 and FTSite 1 share six sites also fall into the expected active site region. All other predicted sites in AttA-T43 are found after the predicted binding region (Fig. 4.11).

PrankWeb identified two potential binding sites in AttA-D49, site one was found to be 43 residues long and covered the majority of the inner concave protein surface, while site two

was located primarily between G134 and S146 (Fig. 4.12). PrankWeb site 1 showed a high level of similarity to both FTSite 1 and 3, while FTSite 2 comprised just three residues (N114, F116 and R130). FTSite prediction 1 and 3 share a combined 31 residues with PrankWeb site 1 prediction which does cover show a large degree of similarity to the predicted AttA binding sites identified in figure 4.9, however, the size of this predicted binding site is considerably larger.

Individually none of the predicted active sites within the final wild AttA variant (AttA-A86) shows much similarity between the binding sites observed in other AttA variants (Fig.4.13). However, a combination of PrankWeb sites 1 and 3 does show a resemblance to larger binding site observed in (Fig. 5.4). Rather interestingly, the majority of binding sites are predicted to be in the second half of protein, while most AttA variants do exhibit binding sites at in this area, the vast majority are found in the first half.

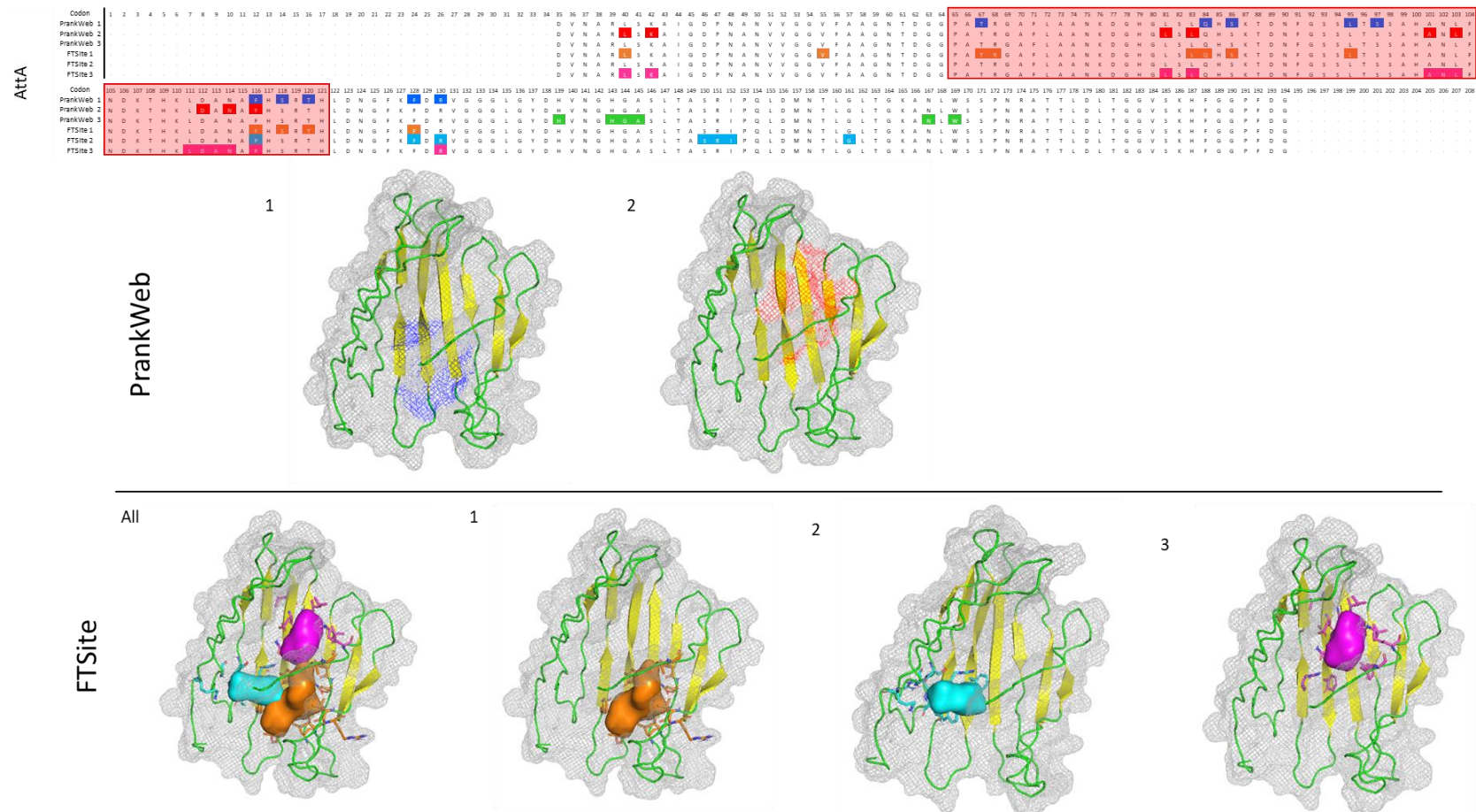


Figure 4.9: An alignment of the wild *G. m. morsitans* AttA amino acid sequence, highlighting all residues within predicted active sites. The boxed area shows previously predicted active site within AttA. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment. FTSite models shoe the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green and  $\beta$ -sheets in yellow with direction shown by the arrow.







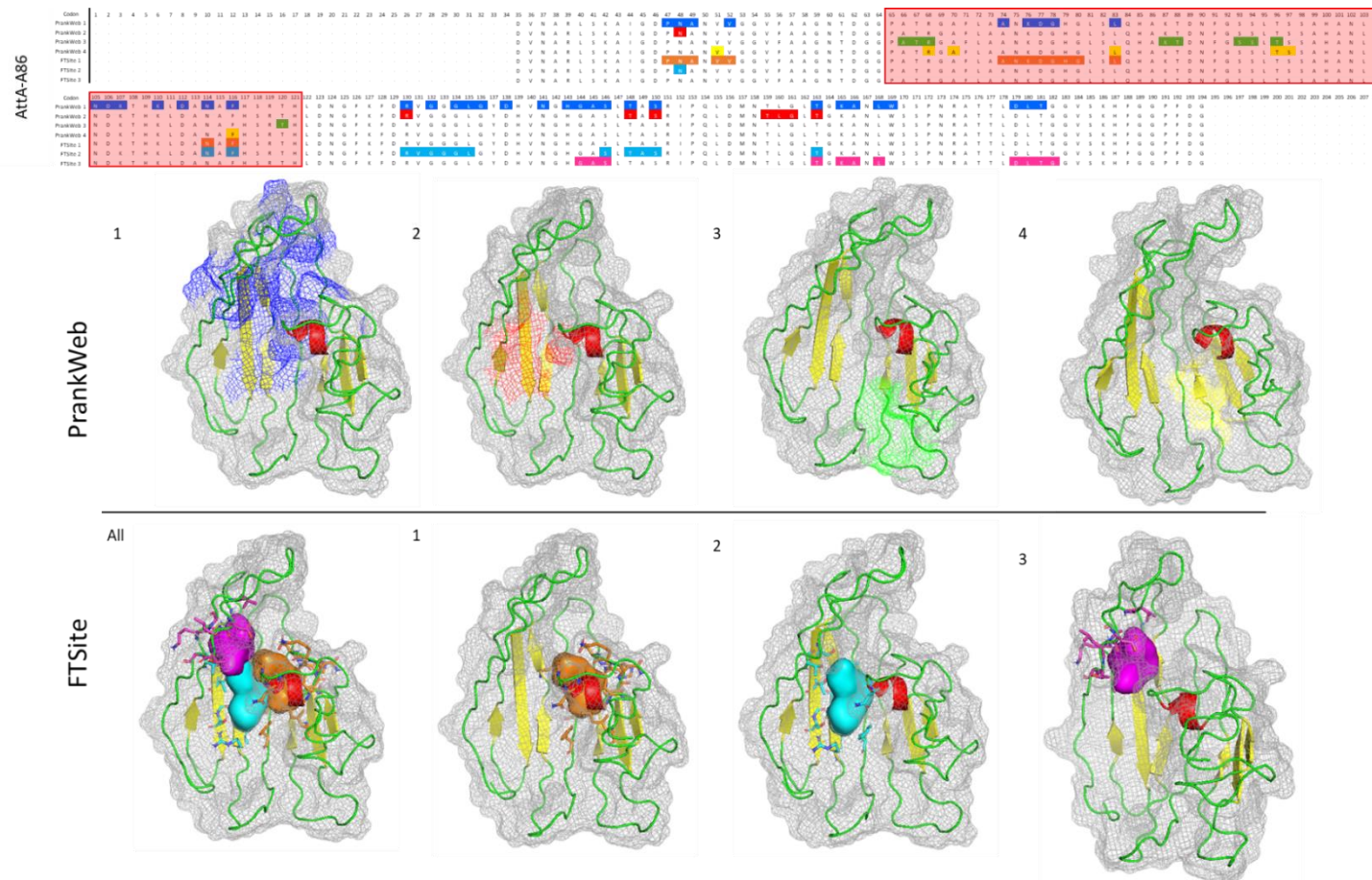


Figure 4.13: An alignment of the wild *G. m. morsitans* Atta-A86 variant amino acid sequence, highlighting all residues within predicted active sites. The boxed area shows previously predicted active site within AttA. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment. FTSite models show the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green,  $\alpha$ -helices are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.



#### 4.3.6: Defensin

##### 4.3.6i: Prediction of protein secondary structure

The Def protein variant exhibited the expected secondary structure of an insect defensin, clearly illustrating the characteristic  $\alpha$ -helix and anti-parallel  $\beta$ -sheets C-terminal structure. Two other protein variants, Def-S18/I82 and Def-S18/E42, also illustrated this characteristic C-terminal structure. Def-S18/I82 exhibited a similar structure to that observed in the Def protein variant, though in a chiral orientation. While C-terminal of Def-S18/E42 exhibited the characteristic structure, the N-terminal comprised primarily of random coil structures rather than the helical structure seen in the Def (Table 4.4).

The remaining eight structures showed considerable variation. Most notable is the absence of the cytosine stabilised  $\alpha$ -helix and anti-parallel  $\beta$ -sheets C-terminal structure. While all but two of the structures indicated varying degrees of a helical structure at the C-terminal there is no indication of anti-parallel  $\beta$ -sheets. The two final structures showed no helical C-terminal structure, Def-I18 maintained the anti-parallel  $\beta$ -sheets but substituting the helix for coils, while Def-F45 exhibited a coiled structure throughout the C-terminal (Table 4.4).

Variations of protein secondary structures could lead to alterations in the protein surface and ultimately impact functionality. Protein surface variation indicates that there is considerable variation between the Def protein variants (Supplementary Figure 9 Appendix 5). The predicted surface structure of Def showed a partially enclosed channel running through the protein between the N and C-terminals, this was also observed in two other Def variants, Def-I18 and Def-E42.

No other Def variants exhibited this channel, however, Def-I82 and Def-S18/I82 both exhibited a large pocket structure between the N and C-terminals (Supplementary Figure 9 Appendix 5). This pocket is observed in a similar location to the channel, it is possible therefore that the Val82/Ila82 substitution is responsible for this opening of the channel into a large pocket. Additionally, Def-S18 exhibited two large pockets, one between the N and C-terminals regions and another running along the top of the surface (Supplementary Figure 9 Appendix 5). All other Def variants showed signs of pockets around the C-terminal though all were considerably smaller than those observed previously.

Table 4.4: Predicted 3-Dimensional structures of each of the *G. m. morsitans* Def variants. The name of each variant, the haplotypes exhibiting each structure and the amino acid substitutions responsible for the variation are given. PDB files were produced using the I-TASSER server (Yang and Zhang, 2015) and visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre). All coil structures are shown in green;  $\alpha$ -helices are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

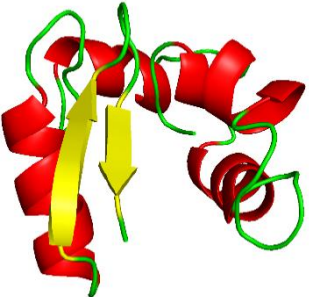
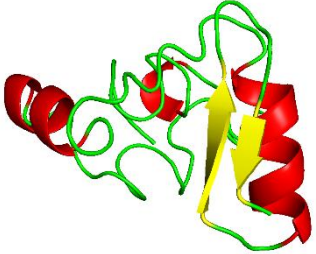
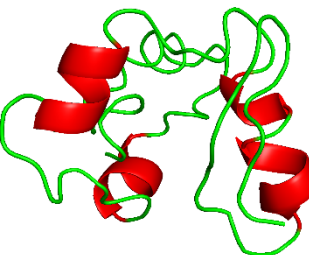
Variant	Haplotypes	Amino Acid substitution	Structure	Variant	Haplotypes	Amino Acid substitution	Structure
Def	4, 6,12 and 17	N/A		Def-S18/182	7	T18 → S18 V82 → I82	
	1, 8, 11, 16 and 23	T18 → S18					Def-E42

Table 4.4 cont.: Predicted 3D structures of each of the *G. m. morsitans* Def variants. The name of each variant, the haplotypes exhibiting each structure and the amino acid substitutions responsible for the variation are given. PDB files were produced using the I-TASSER server (Yang and Zhang, 2015) and visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre). All coil structures are shown in green;  $\alpha$ -helices are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

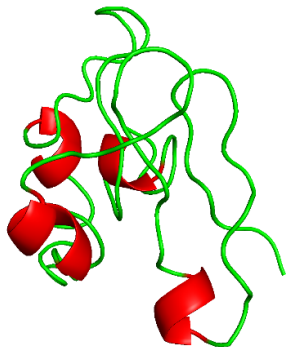
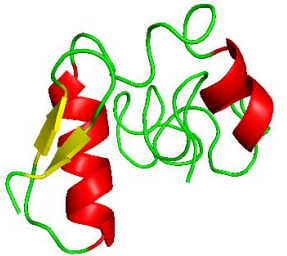
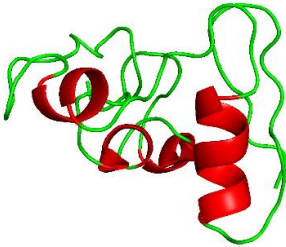
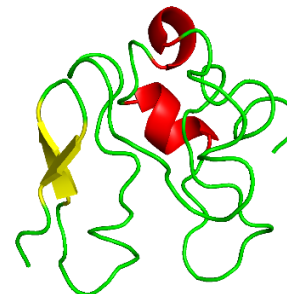
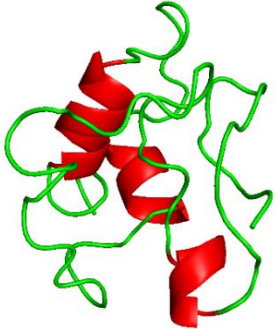
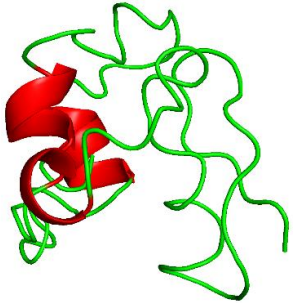
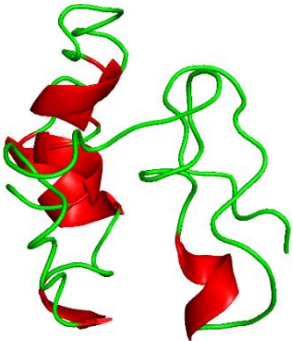
Variant	Haplotypes	Amino Acid substitution	Structure	Variant	Haplotypes	Amino Acid substitution	Structure
Def-I18/F45	21	T18 → I18 V45 → F45		Def-S18/E42	10	T18 → S18 G42 → E42	
							
Def-I82	5	V82 → I82		Def-I18	15	T18 → I18	

Table 4.4 cont.: Predicted 3D structures of each of the *G. m. morsitans* Def variants. The name of each variant, the haplotypes exhibiting each structure and the amino acid substitutions responsible for the variation are given. PDB files were produced using the I-TASSER server (Yang and Zhang, 2015) and visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre). All coil structures are shown in green;  $\alpha$ -helices are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

Variant	Haplotypes	Amino Acid substitution	Structure	Variant	Haplotypes	Amino Acid substitution	Structure
Def-S18/F45/I82	3	T18 → S18		Def-F45	18, 19, 20, and 24	V45 → F45	
		V45 → F45					
Def-S18/F45	2, 22 and 25	T18 → S18 V45 → F45					

The Def protein variants formed two clusters when submitted to PCA (Fig. 5.10). The largest of these clusters consists of five Def variants (Def, Def-S18/E42, Def-I18, Def-I18/F45 and Def-I82), four of these variants exhibit a full or partial cytosine stabilised  $\alpha$ -helix and anti-parallel  $\beta$ -sheets C-terminal structure (Table 4.4). However, Def-I18/F45 initially appears to have a close structural appearance to the other F45 variants. Interestingly, the other three F45 variants (Def- F45, Def-S18/F45 and Def-S18/F45/I82) form a cluster separate from the other wild variants (Fig. 4.14). Table 5.3 illustrates that these variants all exhibit a similar structure that varies considerably from that observed in the other Def variants. The final cluster consists of Def-S18, Def-E42 and Def-S18/I82 (Fig. 4.14). Two of these (Def-S18 and Def-S18/I82) show a full or partial cytosine stabilised  $\alpha$ -helix and anti-parallel  $\beta$ -sheets C-terminal structure (Table 4.4), though in a chiral orientation to that observed in Def. Def-E42 does not however, and its inclusion in the cluster does not appear to be purely based on structure (Table 4.4 and Fig. 4.14).

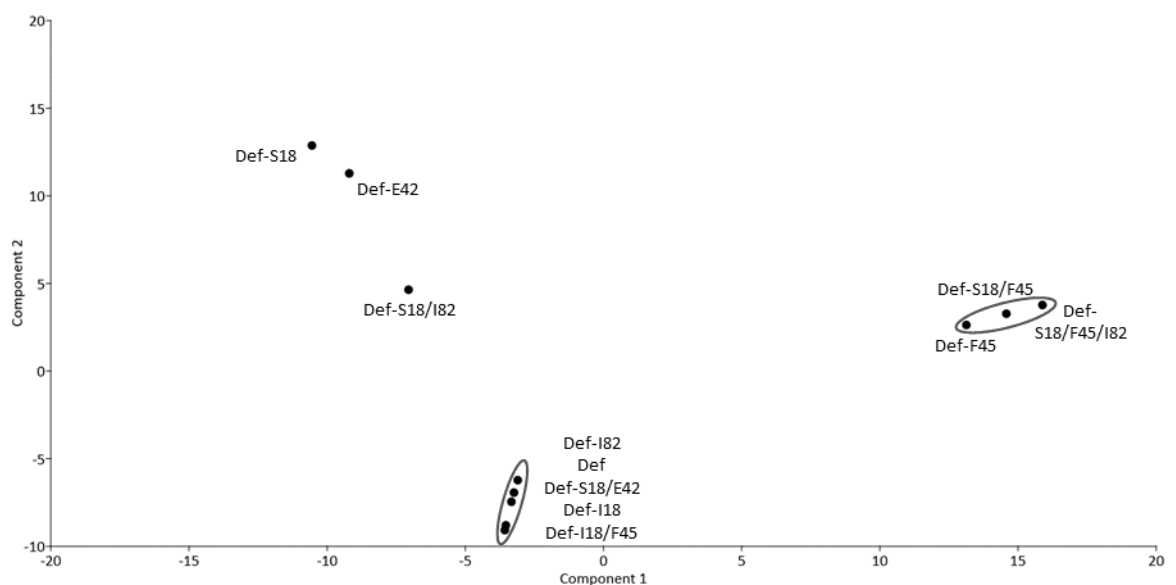


Figure 4.14: Principle component analysis (PCA) plot of wild defensin variants using the first and second Principle components (Eigenvalues: PC1 = 93.94 (19.63 % variance); PC2 = 64.87 (13.55 % variance)). Z-values were calculated, and a matrix produced using DALI (Holm, 2020). PCA analysis was conducted in PAST3 (Hammer et al., 2001).

#### 4.3.6ii: *Glossina* defensin active site identification and variation

An alignment of defensin sequences from *A. dichotoma* (AAB36306) and *O. rhinoceros* (BAA36401) to *G. m. morsitans* defensin sequence, identified in Chapter 3, indicated a relatively high conservation of residues within the C-terminal defensin domain (Fig. 4.15).

This region of conserved residues corresponds with the previously published active sites of both *A. dichotoma* and *O. rhinoceros* defensin proteins, as highlighted in figure 4.15 (Ishibashi *et al.*, 1999; Saido-Sakanaka *et al.*, 1999).

### A)

```

G. morsitans morsitans                MKFYLVLAFLTLCAVAVTALPAGDETRIDLETLEEDLRLVDGAQVTGELKRDKRVTCTNIG
Allomyrina dichotoma (AAB36306.1)  -----VTCDLL
Oryctes rhinoceros (BAA36401.1)    MSRFIVFAFIVAMCIAHSLAAPAPEA-----LEASVIRQKRLTCDLL
                                         :*: :

```

```

G. morsitans morsitans                EW-----VCVAHCNSKSKKSGYCSRGVICYCTN
Allomyrina dichotoma (AAB36306.1)    SFEAKGFAANHSLCAAHCLAI GRRGGSCERGVICRE
Oryctes rhinoceros (BAA36401.1)    SFEAKGFAANHSLCAAHCLAI GRKGGACQNGVVCRR
                                         .:          :*.*** : ...* *.*** *

```

### B)

```

G. brevipalpis                MKYYVSTAIVALFAVAVTCLPTGEDPVNNLETLEQDLKPDNVEQVPVELRRDKRVTCSIGEWVCVGHNSMGKKSGYCSRGVICYCKN
G. fuscipes fuscipes          MKFYLVLAFLTLFAVAVTALPAGDETRINLETLEQDLTLVDAHQVTGELKRDKRVTCTNIGEWACVAHCNAKSKKSGYCSRGVICYCTS
G. palpalis gambiensis       MKFYLVLAFLTLFAVAVTALPAGDEPGTNLETLEQDLTLVDAHQVTGELKRDKRVTCTNIGEWACVAHCNAKSKKSGYCSRGVICYCTS
G. austeni                   MKFYLVLAFLTLFAVAVTASPAGDETRIDLESLEPDLRLVDADQATGELKRDKRVTCTNIGEWACVAHCNSKSKKSGYCSRGVICYCTN
G. morsitans morsitans        MKFYLVLAFLTLCAVAVTALPAGDETRIDLETLEEDLRLVDGAQVTGELKRDKRVTCTNIGEWVCVAHCNSKSKKSGYCSRGVICYCTN
G. pallidipes                MKFYLVLAFLTLFAVAVTALPAGDETRIGLETLEEDLRLVDVQVTGELKRDKRVTCTNIGEWVCVAHCNSKSKKSGYCSRGVICYCTN
G. Swynnertoni               MKFYLVLAFLTLFAVAVTALPAGAE TRIDLETVEEDLRLVDVQVTGELKRDKRVTCTNIGEWVCVAHCNSKSKKSGYCSRGVICYCTN
                                         **:*. : *::* *****. *:* :. .**::* ** : *.. **.******.***.***.***: .*****.*****..

```

Figure 4.15: A) An alignment of *G. m. morsitans* Def amino acid sequences, identified in Chapter 3, to *A. dichotoma* (AAB36306) and *O. rhinoceros* (BAA36401) Def sequences. The boxed area indicated the region previously document as the active site within both AAB36306 and BAA36401. B) An alignment of all predicted Def sequences identified in Chapter 3, highlighting the location of and conservation of amino acids within the predicted active site. \* = complete conservation, : = residues with Gonnet PAM 250 score > 0.5, . = residues with residues with Gonnet PAM 250 score < 0.5 and a gap = no similarity.

Having obtained an approximate location of the active site of *Glossina* Def, the *G. m. morsitans* amino acid sequences were run through two online servers, FTSite and PrankWeb to verify and illustrate the location of the active site. This analysis identified six potential active sites, three from PrankWeb and three from FTSite (Fig. 4.16). When aligned these sites appear to confirm the active site of *Glossina* Def is located around the same C-terminal region as identified in *A. dichotoma* and *O. rhinoceros* (Fig. 4.16). Furthermore, the binding sites seem to be dependent upon interactions with N-terminal residues, though whether for stability or functionality remains undetermined.

Visualisation of the active sites identified by FTSite suggest that only site 1 (Fig. 4.16: FTSite 1) has been accurately predicted as both sites 2 and 3 are located away from the N and C-terminal binding pocket (Fig. 4.16: FTSite 2 and 3). While FTSite 1, is an almost identical match to PrankWeb site 3 (Fig. 4.16). The first site predicted by PrankWeb, appears to show the most similarity to the previously published active sites, encompassing the majority of the N and C-terminal binding pocket (Fig. 4.16: PrankWeb 1). While PrankWeb Site 2 appears to interact primarily with the N-terminal rather than the C-terminal (Fig. 4.16: PrankWeb 2).

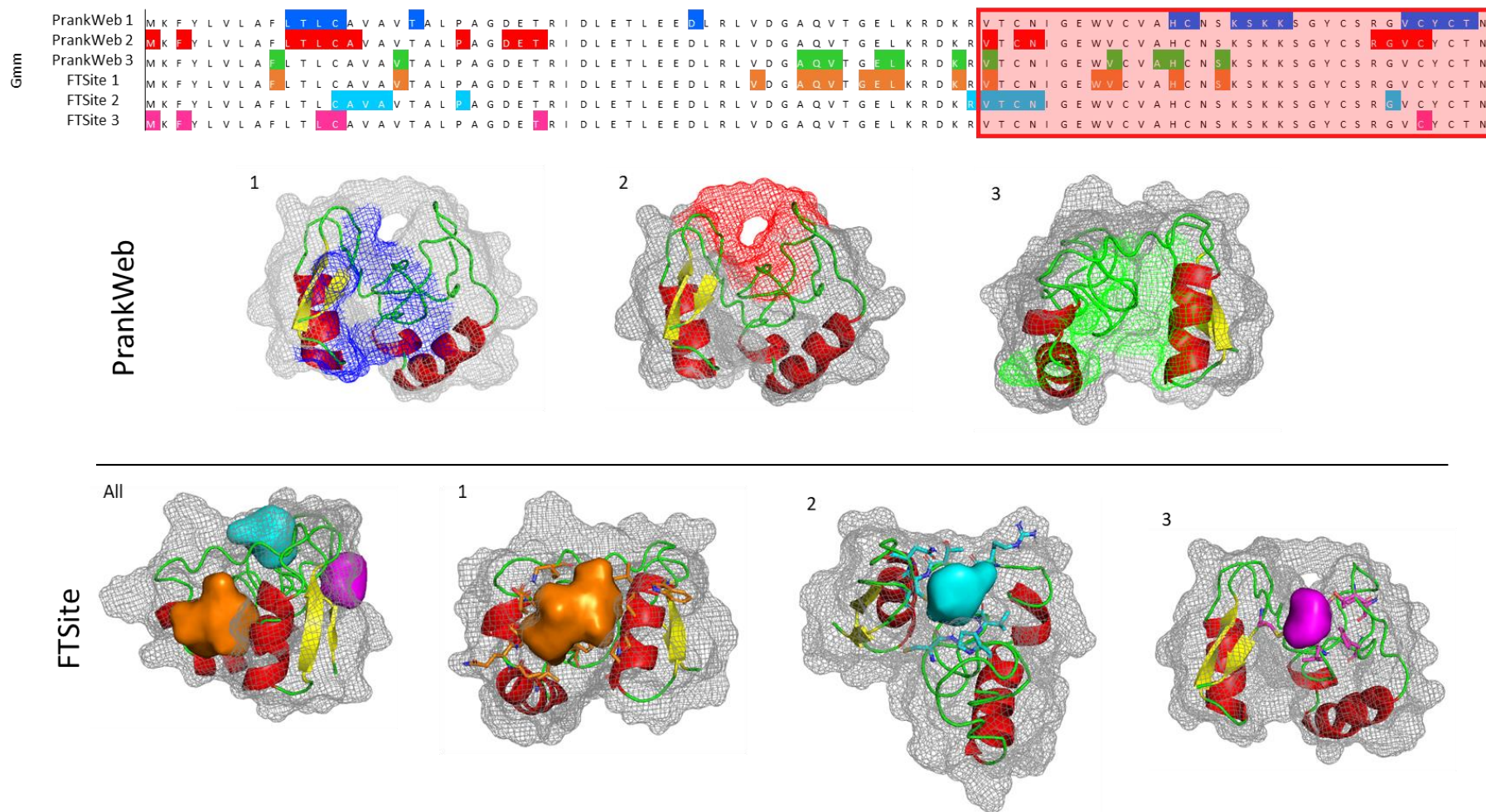


Figure 4.16: An alignment of the *G. morsitans* Def amino acid sequences, identified in Chapter 3, highlighting all residues within predicted active sites. The boxed area shows previously document as the active site within AAB36306 and BAA36401. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment. FTSite models shoe the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green,  $\alpha$ -helicase are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.



PrankWeb predicted a single bonding site within the Def structure (Fig. 4.17), located along the N/C-terminal channel identified in section 4.3.3i. This binding site is formed of 25 residues (V6-A8, T11, V15, A16, 20L, A22-D24, I28, D29, L33, D36, K53, V55, H67, C68, K71, K73, C83-N87) from both the N and C-terminals (Fig. 4.17: PrankWeb 1). FTSite predicted a further three potential active sites within the Def (Fig. 4.17), of these Sites 1 and 2 were found near the documented active site of the Def protein. Site 1 was found to be a result of interactions between seven amino acid residues (L20, A22, I28, Y84, C85, T86 and N87) (Fig. 4.17: FTSite 1), while Site 2 was formed by 12 residues (T11, A16, L20, I28, D29, L30, V55, V63, H67, C68, K71 and Y84) (Fig. 4.17: FTSite 2). Site 3 was discounted from the analysis as this did not indicate any interaction with the C-terminal residues or fall within the established active site region (Fig. 4.17: FTSite 3).

Both PrankWeb and FTSite predicted three potential binding sites within the Def-F45 sequence (Fig. 4.18). Interestingly, both servers predicted almost identical binding sites, with PrankWeb site 1 matching FTSite 3, PrankWeb site 2 matching FTSite 1, and PrankWeb site 3 matching FTSite 2 (Fig. 4.18). PrankWeb site 1 and FTSite 3 rely on the interaction of 13 residues, sharing 11 of these (L20, E25, T26, L28, T32, D36, L37, W62, H67, V82 and C83) between the N and C-terminal regions of the protein structure (Fig. 4.18: PrankWeb 1 and FTSite 3). PrankWeb site 2 and FTSite 1 both indicate interactions between 12 residues (L7, Q44, F45, G47, E48, L49, D52, V55, V63, V65, N69 and S70). Additionally, PrankWeb site 2 predicts further interaction with S75 and FTSite 1 with L39 and T46. This binding site is located centrally between the C-terminal and the third  $\alpha$ -helix (Fig. 4.18: PrankWeb 2 and FTSite 1). Finally, PrankWeb site 3 and FTSite 2 indicate interactions of ten residues (V55, T56, C57, V63, C64, V65, S75, G76, Y77 and C78) predominately found in the C-terminal arthropod defensin domain (Fig. 4.18: PrankWeb 3 and FTSite 2).

As in Def-F45, three potential active sites were identified by both PrankWeb and FTSite within Def-S18, Def-I18, Def-I82, Def-I18/E42 and Def-I18/F45/I82. Def-S18 showed a high level of similarity between PrankWeb site 1 and FTSite 3 and PrankWeb site 2 and FTSite 1 (Fig. 4.19). PrankWeb site 1 consists of 16 residues while FTSite 3 consists of 13 residues, of which 11 (L7, F9, T46, E48, L49, D52, T56 and K74-Y77) were exhibited in both. PrankWeb site 2 consists of 17 residues compared to 11 residues in FTSite 1, however of these 10 (L10, V17, R27, I28, L33, L37, E61, W62, V65 and N69) are shared in both predictions. PrankWeb site 3 is exclusively at the C-terminal and within the predicted active site region, however,

there is no indication of interaction with the N-terminal residues and is located on the external surface rather than centrally like other predicted sites (Fig. 4.19). Furthermore, FTSite 2 shows signs of interactions between both terminals, however, the binding site is located on the dorsal surface of the protein rather than between the N and C-terminals (Fig. 4.19).

Def-I18 showed considerably less similarities between predicted sites, with only PrankWeb site 1 and FTSite 1 shows any clear conservation. In this case, all 12 of the residues identified in PrankWeb site 1 were also present in FTSite 1 (I28, T32, L33, D36, V63, C64, A66, N69, K74, S75, Y77 and C78). While all other predicted active sites utilised residues in both the N- and C-terminals none were located in a similar location to the previously predicted actives (Fig. 4.20).

The predicted sites within Def-I82 showed a higher proportion of residues at the N-terminal than other Def variants (Fig. 4.21). PrankWeb 1 and FTSite 2 both show similarities, PrankWeb site 1 consists of 15 residues while FTSite 2 consists of 13 residues and they share 10 (A19-P21, I28, D29, L33 and I82-C85). Additionally, there is some similarity between PrankWeb site 2 and FTSite 3, where all six of the FTSite 3 residues are also exhibited by PrankWeb site 2 (F9, C13, A66, H67, S70 and K71). However, FTSite 1 does not exhibit any residues within the previously established active site and PrankWeb site 3 shows some characteristics of the previously established bind sites though it is quite small compared to previously predicted active sites (Fig. 4.21).

Predicted active sites within Def-S18/E42 showed some conservation between PrankWeb site 1 and FTSite 1 and PrankWeb site 3 and FTSite 3 (Fig. 4.22). PrankWeb site 1 is comprised of 17 of which 11 (L12, C13, E42, T46, V55, H67, C68, G81-C83 and C85) are shared with FTSite 1, between the N and C-terminals. PrankWeb site 3 and FTSite 3 share eight residues however, this site is situated on the dorsal surface of the protein away from the established location. FTSite 2 is also located on the dorsal surface while PrankWeb site 2 is located on the opposite site to PrankWeb site 1 (Fig. 4.22), therefore these are unlikely to be accurate predictions of the Def active site.

PrankWeb site 1 and FTSite 1 predict the same binding pocket within Def-S18/F45/I82 (Fig. 4.23), though PrankWeb site 1 is slightly larger (14 residues) compared to FTSite 1 (11 residues) though nine of these are exhibited by both predicted sites (K50, V55, C57, V63,

A66, N69 and S75-Y77). This predicted binding site is located in a similar position to the previously observed binding sites within Def. However, none of the other predicted sites show much similarity and are all, with the exception of FTSite 2, located away from predicted binding site (Fig. 4.23).

Only two predictions made by PrankWeb with the Def-E42 structure (Fig. 4.24). However, PrankWeb site 1 shows a conservation between both FTSite 2 and 3. Of the 27 residues predicted to be utilised by PrankWeb site 1, 19 are shared between either or both FTSite 2 and 3 (L10, L12, V15, A16, A 19, P21, T26-D29, E35, D36, R54-N58, G60 and W62). This binding pocket is located between the N and C-terminals in the area previously established. Although PrankWeb site 2 and FTSite 1 also show a high level of similarity sharing 11 residues in total (A8, F9, V45, E48, L49, D52, V55, E61 and S75-Y77). This binding pocket is located on the opposed side of the protein away from the stable binding pocket (Fig. 4.24).

PrankWeb predicted four binding sites within Def-I18/F45, while FTSite identified three potential sites (Fig. 4.25). Of these PrankWeb site 1 and FTSite 3 indicated some similarity, both consist of ten residues of which seven are shared (E48, L49, C64, A66, N69, S75 and Y84). Furthermore, PrankWeb site 2 and FTSite 2 showed similarities also sharing 7 of their eight residues (F9, L12, H67, 68, K71, K73, C85 and T86) (Fig. 4.25). However, neither of these sites are located between the N and C-terminals as would be expected. While neither of the other PrankWeb sites are found in this region, FTSite 1 is and appears to illustrate the most likely location for the Def-I18/F45 active site (Fig. 4.25).

Unlike the Def-I18/F45, Def-S18/F45 shows a high level of conservation between the predicted binding site (Fig. 4.26). PrankWeb site 1 and FTSite 2 are almost identical sharing 15 residues between them (F9, L39, A43-L49, W62-C64, A66, H67 and S70). Equally however, PrankWeb site 2 and FTSite 1 and 3 also show high similarity and also share 15 residues (L10, V17, S18, G23, T26, R27, D36, L37, L39, V65, H67 and R80-C85) between the three predicted sites. Given the high level of similarity between the predicts and the residues forming them either of these pockets could represent the active site of Def-S18/F45 (Fig. 4.26).

PrankWeb predicted three active sites within the final Def variant, Def-S18/I82, while FTSite only predicted two (Fig. 4.27). PrankWeb site 1 unutilised 23 residues while FTSite 1 consists of 25 though they share 19 (V6, T11, L12, A16, L20, A22, G23, T32, D36, L37, E48,

K50, V55, W62, V65, A66 and K73-S75), this pocket is similar to that observed previously in other Def variants. While PrankWeb site 3 and FTSite 2 do appear to predict the same bind pocket it is located on the dorsal surface away from the established active site location and they do not seem to interact with the predicted active sites region. Furthermore, PrankWeb site 2 does not interact with the N-terminal residues and appears to be solely based around the exterior surface of the C-terminal (Fig. 4.27).

Def	PrankWeb 1	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	T	A	L	P	A	G	D	E	T	R	I	D	L	E	T	L	E	E	D	L	R	L	V	D	G	A	Q	V	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	K	S	G	Y	C	S	R	G	V	C	Y	C	T	N
	FTSite 1	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	T	A	P	A	G	D	E	T	R	I	D	L	E	E	D	L	R	L	V	D	G	A	Q	V	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	K	S	G	Y	C	S	R	G	V	C	Y	C	T	N				
	FTSite 2	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	T	A	P	A	G	D	E	T	R	I	D	L	E	E	D	L	R	L	V	D	G	A	Q	V	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	K	S	G	Y	C	S	R	G	V	C	Y	C	T	N				
	FTSite 3	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	T	A	L	P	A	G	D	E	T	R	I	D	L	E	E	D	L	R	L	V	D	G	A	Q	V	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	K	S	G	Y	C	S	R	G	V	C	Y	C	T	N			

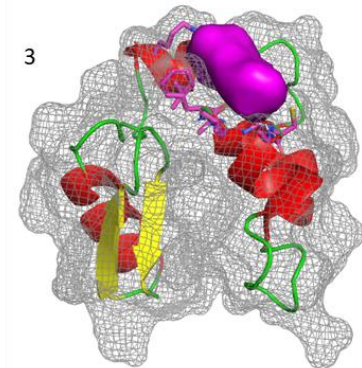
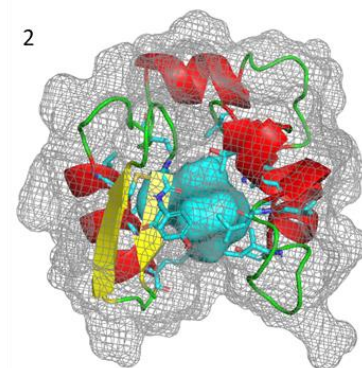
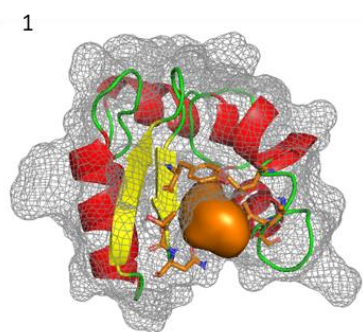
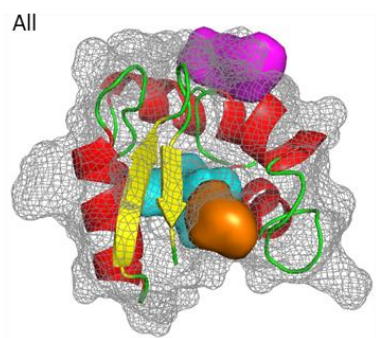
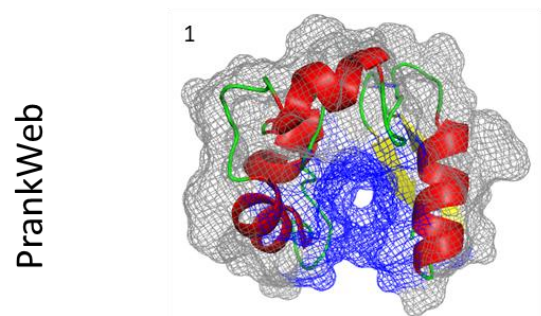


Figure 4.17: An alignment of the wild, natural variation of Def amino acid sequences highlighting all residues within predicted active sites. The boxed area shows previously document as the active site within AAB36306 and BAA36401. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment. FTSite models show the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green,  $\alpha$ -helicase are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

Def-F45

```

PrankWeb 1  - - - - V L A F L T L C A V A V T A L P A G D E T R I D L E T L E E D L R L V D G A Q F T G E L K R D K R T C N I G E W V C V A H C N S K S K K S G Y C S R G V C Y C T N
PrankWeb 2  - - - - V A F L T L C A V A V T A L P A G D E T R I D L E T L E E D L R L V D G A Q F T G E L K R D K R T C N I G E W V C V A H C N S K S K K S G Y C S R G V C Y C T N
PrankWeb 3  - - - - V L A F L T L C A V A V T A L P A G D E T R I D L E T L E E D L R L V D G A Q F T G E L K R D K R T C N I G E W V C V A H C N S K S K K S G Y C S R G V C Y C T N
FTSite 1    - - - - V A F L T L C A V A V T A L P A G D E T R I D L E T L E E D L R L V D G A Q F T G E L K R D K R T C N I G E W V C V A H C N S K S K K S G Y C S R G V C Y C T N
FTSite 2    - - - - V L A F L T L C A V A V T A L P A G D E T R I D L E T L E E D L R L V D G A Q F T G E L K R D K R T C N I G E W V C V A H C N S K S K K S G Y C S R G V C Y C T N
FTSite 3    - - - - V L A F L T L C A V A V T A L P A G D E T R I D L E T L E E D L R L V D G A Q F T G E L K R D K R T C N I G E W V C V A H C N S K S K K S G Y C S R G V C Y C T N

```

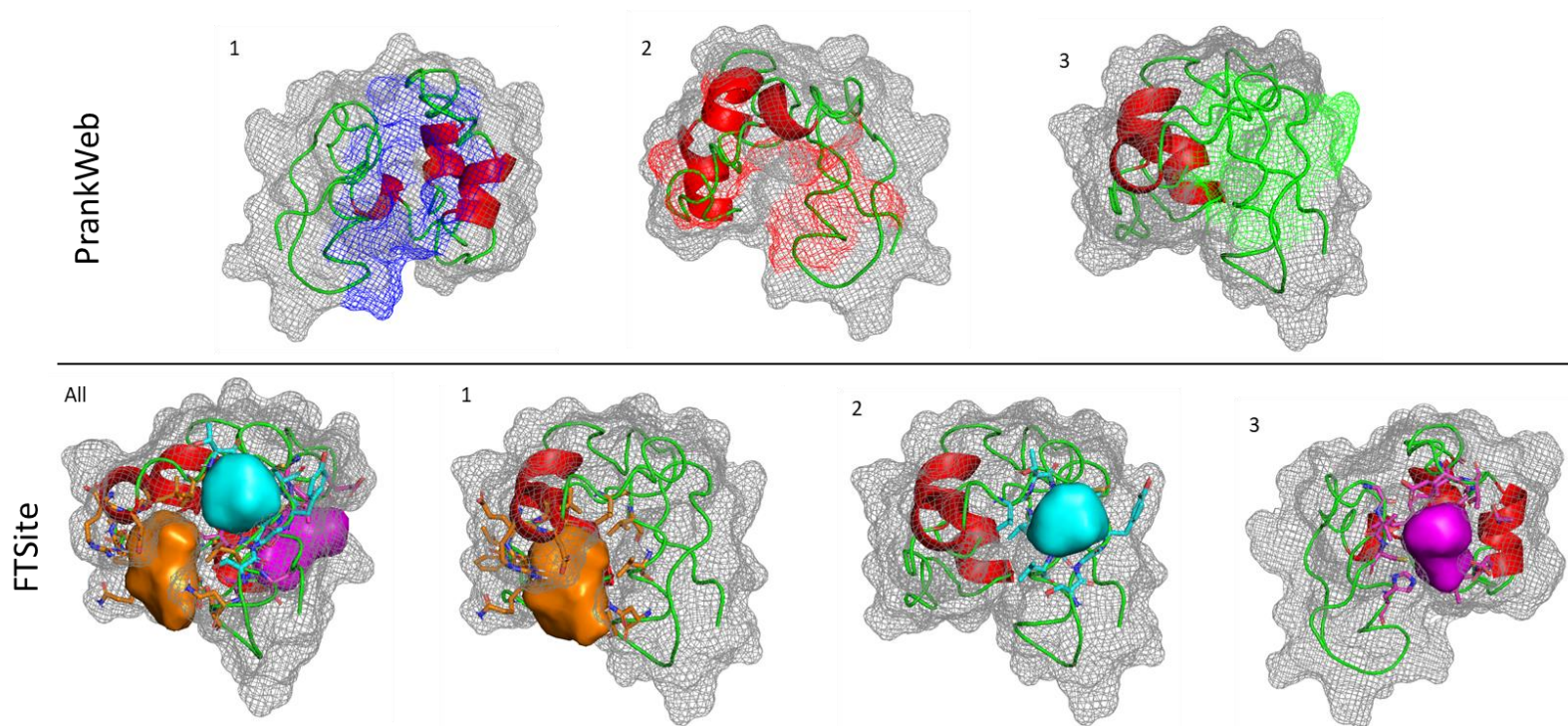


Figure 4.18: An alignment of the Def-F45 variation of Def amino acid sequences highlighting all residues within predicted active sites. The boxed area shows previously document as the active site within AAB36306 and BAA36401. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment. FTSite models show the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green,  $\alpha$ -helicase are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

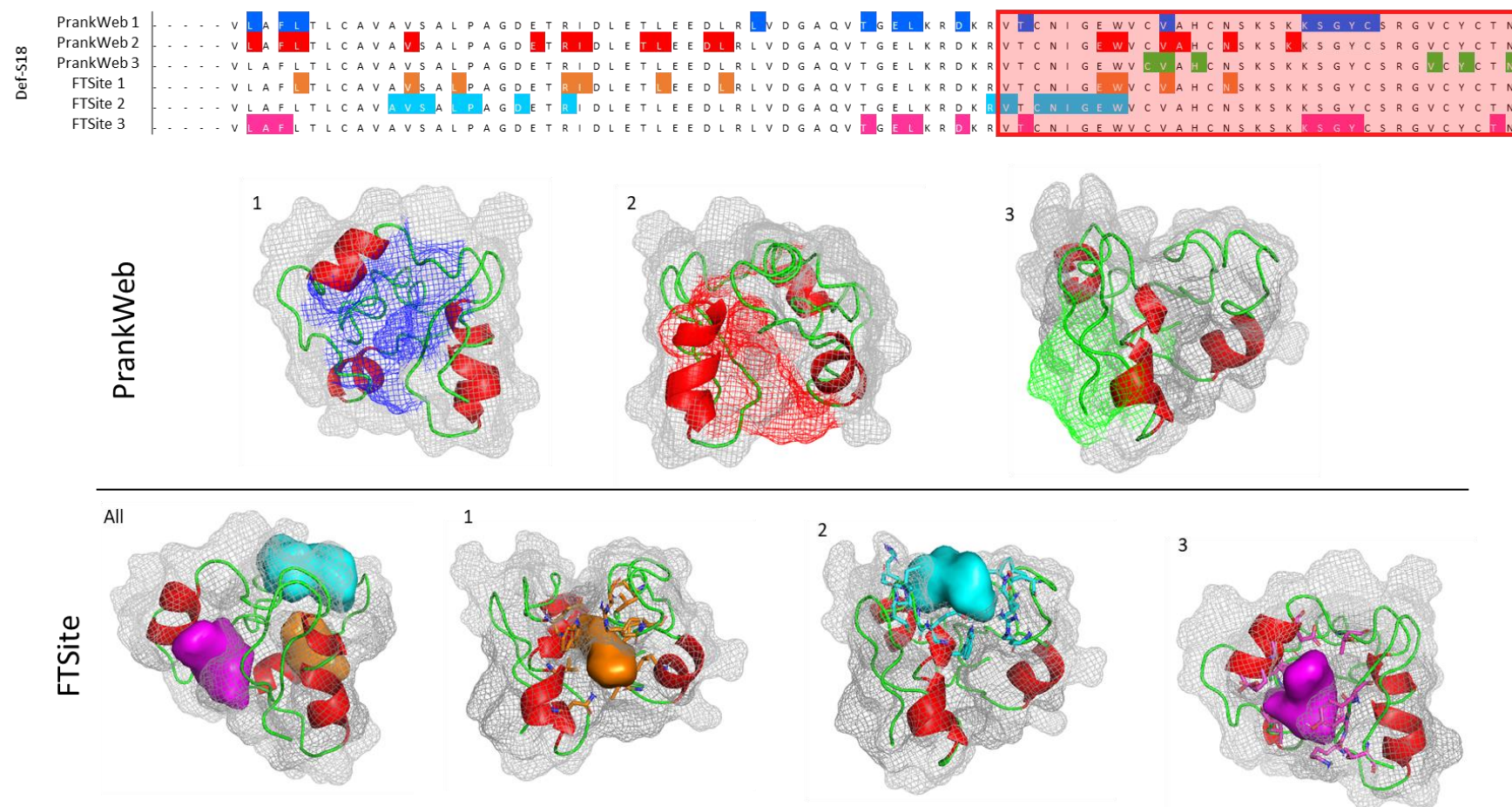


Figure 4.19: An alignment of the wild Def-S18 variant amino acid sequences highlighting all residues within predicted active sites. The boxed area shows previously document as the active site within AAB36306 and BAA36401. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment above. FTSite models show the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green,  $\alpha$ -helicase are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

Def-I18

PrankWeb 1	- - - - V L A F L T L C A V A V I A L P A G D E T R I D L E T L E E D L R L L V D G A Q V T G E L K R D K R	V T C N I G E W V C V A H C N S K S K K S G Y C S R G V C Y C T N
PrankWeb 2	- - - - V L A F L T L C A V A V I A L P A G D E T R I D L E T L E E D L R L L V D G A Q V T G E L K R D K R	V T C N I G E W V C V A H C N S K S K K S G Y C S R G V C Y C T N
PrankWeb 3	- - - - V L A F L T L C A V A V I A L P A G D E T R I D L E T L E E D L R L L V D G A Q V T G E L K R D K R	V T C N I G E W V C V A H C N S K S K K S G Y C S R G V C Y C T N
FTSite 1	- - - - V L A F L T L C A V A V I A L P A G D E T R I D L E T L E E D L R L L V D G A Q V T G E L K R D K R	V T C N I G E W V C V A H C N S K S K K S G Y C S R G V C Y C T N
FTSite 2	- - - - V L A F L T L C A V A V I A L P A G D E T R I D L E T L E E D L R L L V D G A Q V T G E L K R D K R	V T C N I G E W V C V A H C N S K S K K S G Y C S R G V C Y C T N
FTSite 3	- - - - V L A F L T L C A V A V I A L P A G D E T R I D L E T L E E D L R L L V D G A Q V T G E L K R D K R	V T C N I G E W V C V A H C N S K S K K S G Y C S R G V C Y C T N

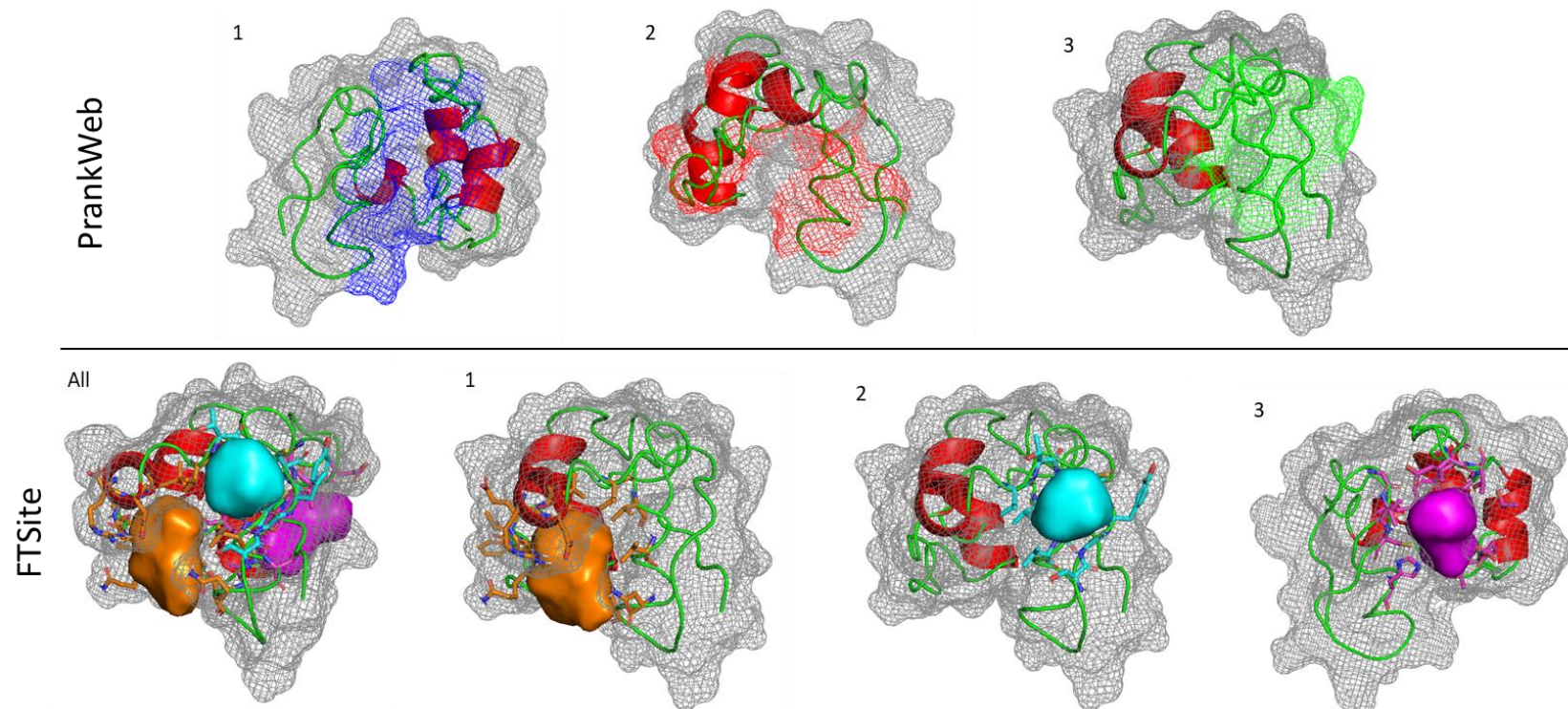


Figure 4.20: An alignment of the Def-I18 variation of Def amino acid sequences highlighting all residues within predicted active sites. The boxed area shows previously document as the active site within AAB36306 and BAA36401. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment. FTSite models show the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green,  $\alpha$ -helicase are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.



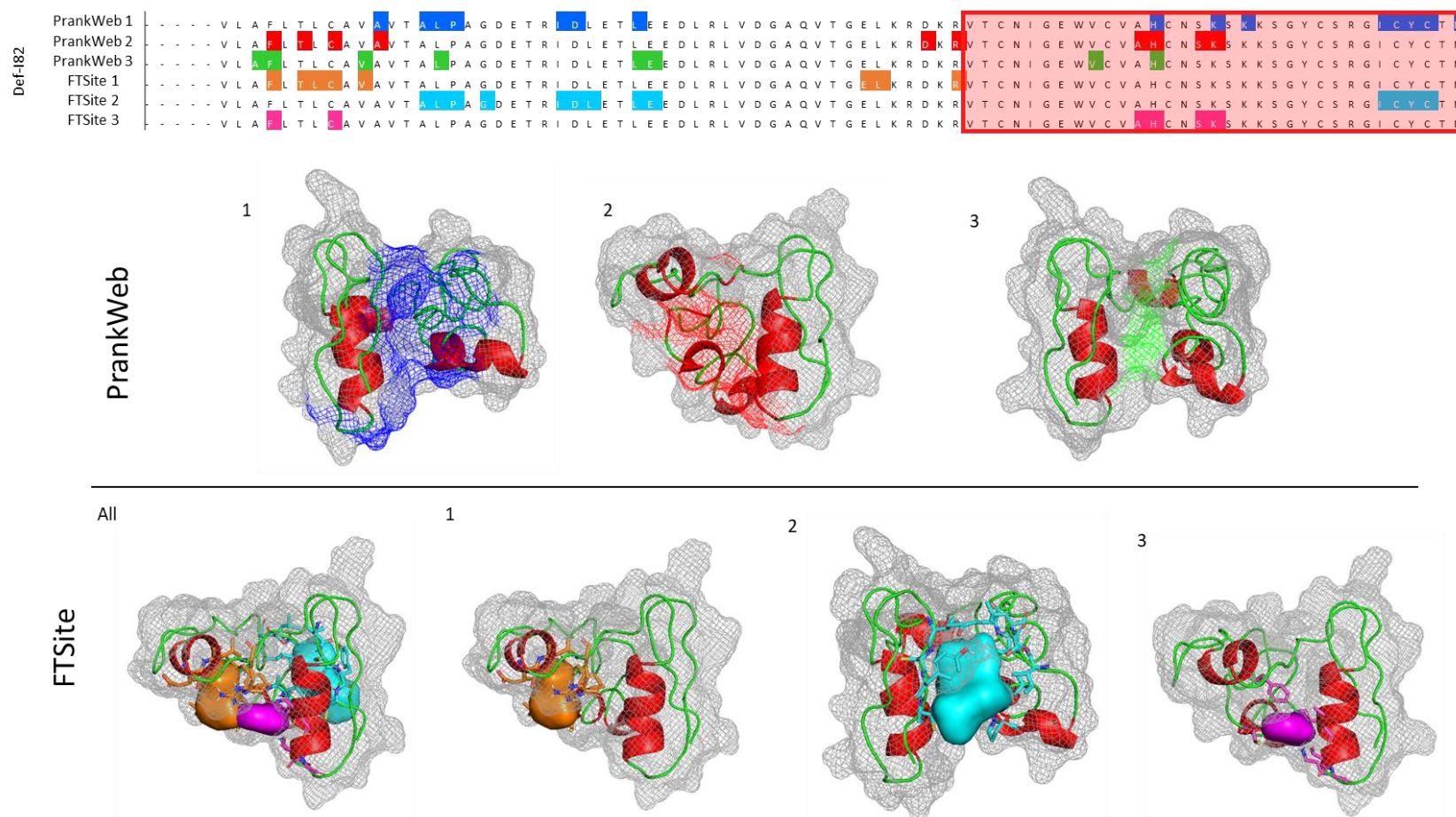


Figure 4.21: An alignment of the Def-182 variation of Def amino acid sequences highlighting all residues within predicted active sites. The boxed area shows previously document as the active site within AAB36306 and BAA36401. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment. FTSite models show the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green,  $\alpha$ -helicase are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

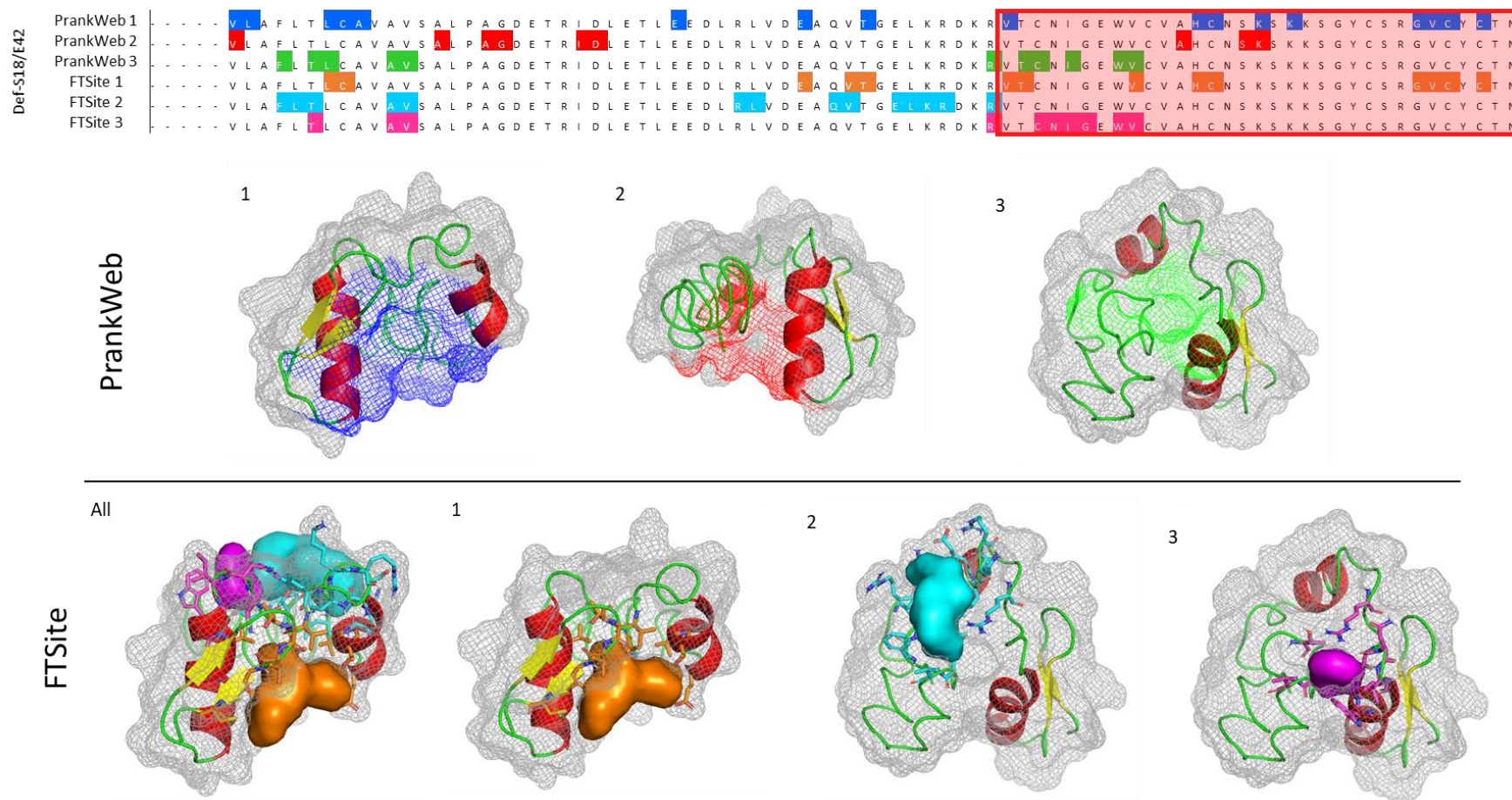


Figure 4.22: An alignment of the Def-S18/E42 variation of Def amino acid sequences highlighting all residues within predicted active sites. The boxed area shows previously document as the active site within AAB36306 and BAA36401. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment. FTSite models show the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green,  $\alpha$ -helicase are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

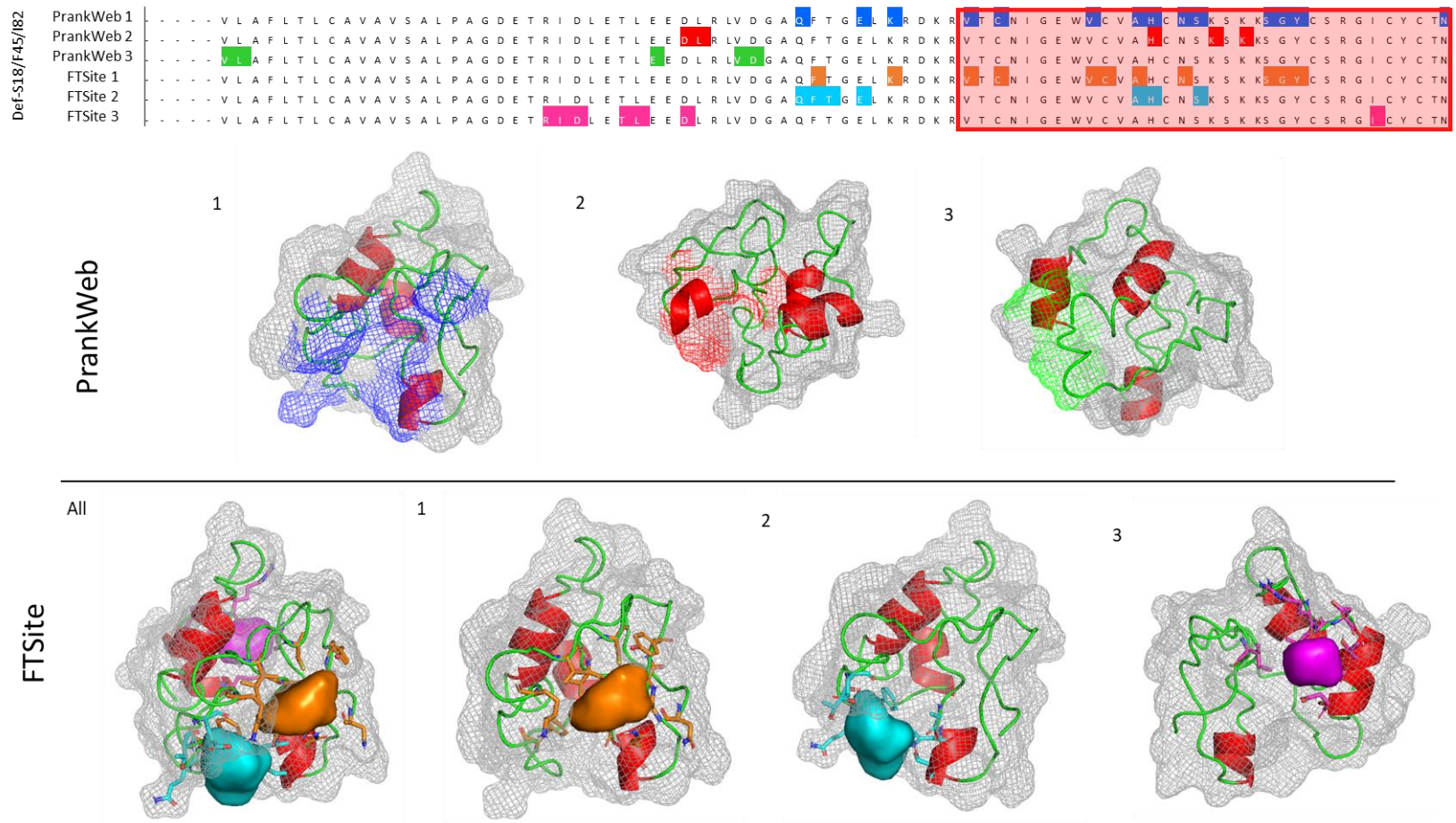


Figure 4.23: An alignment of the Def-S18/F45/I82 variation of Def amino acid sequences highlighting all residues within predicted active sites. The boxed area shows previously document as the active site within AAB36306 and BAA36401. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment. FTSite models show the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green,  $\alpha$ -helicase are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

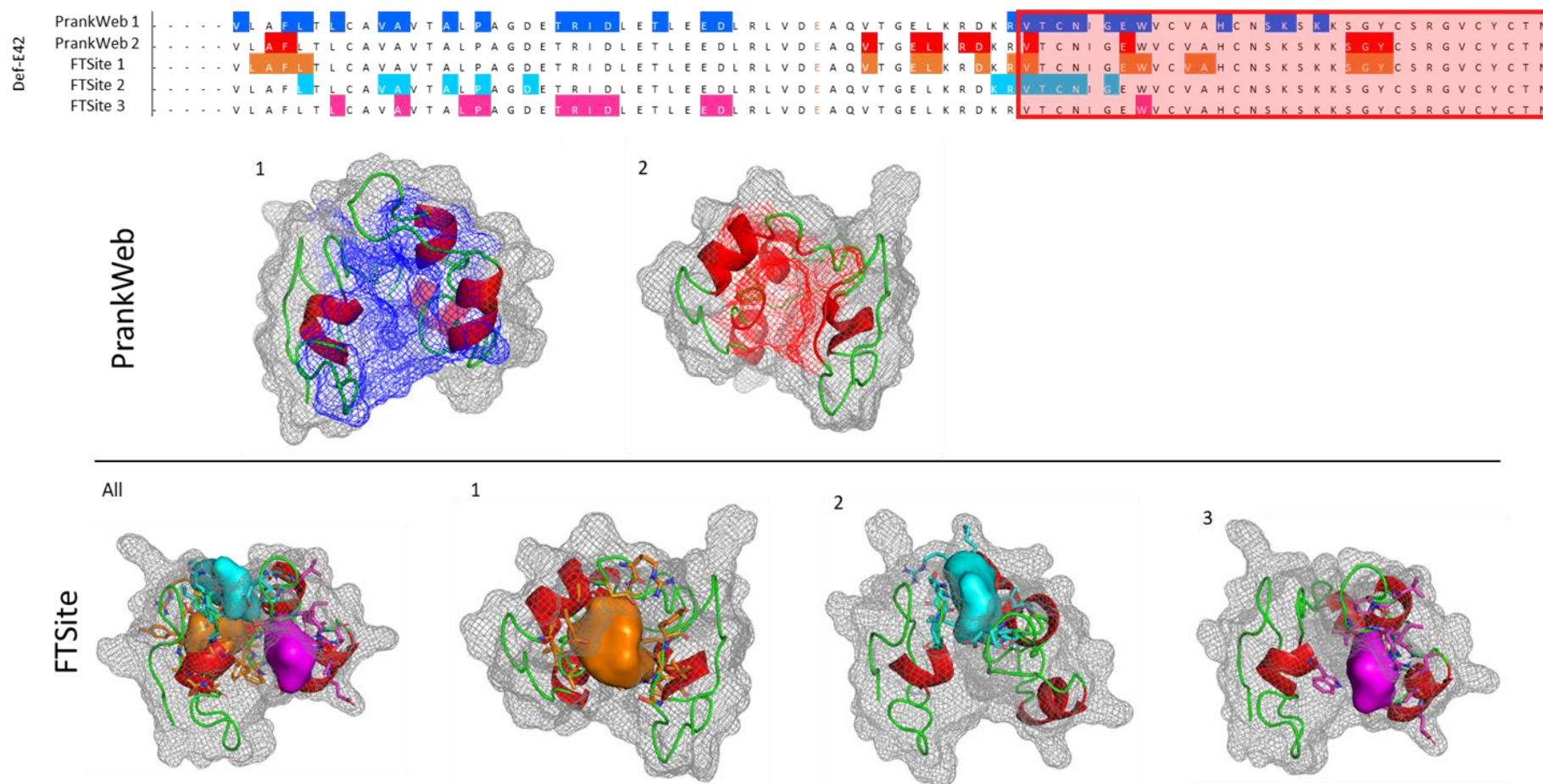


Figure 4.24: An alignment of the Def-E42 variation of Def amino acid sequences highlighting all residues within predicted active sites. The boxed area shows previously document as the active site within AAB36306 and BAA36401. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment. FTSite models show the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green,  $\alpha$ -helicase are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

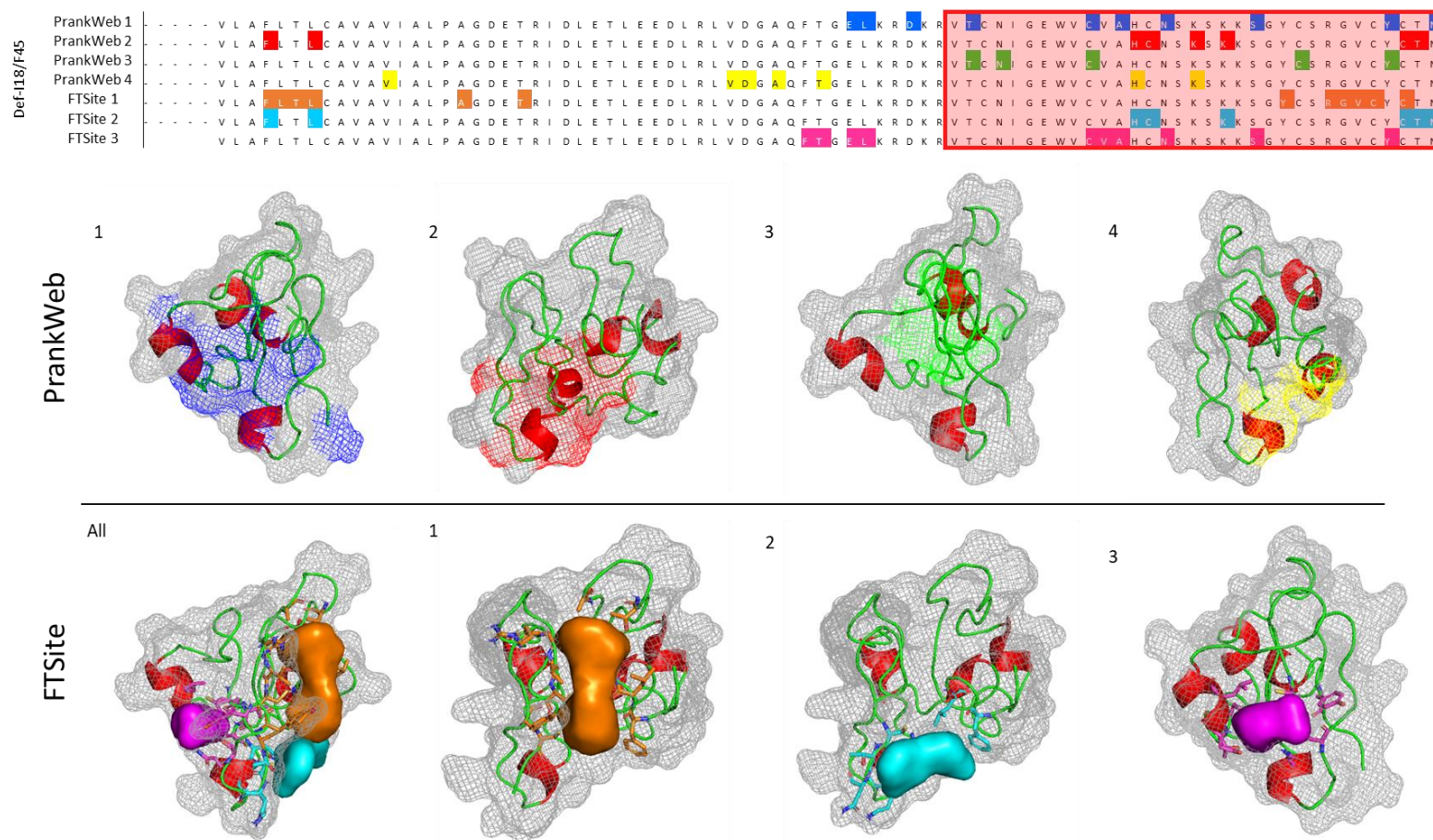


Figure 4.25: An alignment of the Def-I18/F45 variation of Def amino acid sequences highlighting all residues within predicted active sites. The boxed area shows previously document as the active site within AAB36306 and BAA36401. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment. FTSite models show the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green,  $\alpha$ -helicase are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

Def-S18/F45

PrankWeb 1	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	S	A	L	P	A	G	D	E	T	R	I	D	L	E	T	L	E	E	D	L	R	L	V	D	G	A	Q	F	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	S	G	Y	C	S	R	G	V	C	Y	C	T	N
PrankWeb 2	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	S	A	L	P	A	G	D	E	T	R	I	D	L	E	T	L	E	E	D	L	R	L	V	D	G	A	Q	F	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	S	G	Y	C	S	R	G	V	C	Y	C	T	N
FTSite 1	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	S	A	L	P	A	G	D	E	T	R	I	D	L	E	T	L	E	E	D	L	R	L	V	D	G	A	Q	F	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	S	G	Y	C	S	R	G	V	C	Y	C	T	N
FTSite 2	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	S	A	L	P	A	G	D	E	T	R	I	D	L	E	T	L	E	E	D	L	R	L	V	D	G	A	Q	F	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	S	G	Y	C	S	R	G	V	C	Y	C	T	N
FTSite 3	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	S	A	L	P	A	G	D	E	T	R	I	D	L	E	T	L	E	E	D	L	R	L	V	D	G	A	Q	F	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	S	G	Y	C	S	R	G	V	C	Y	C	T	N

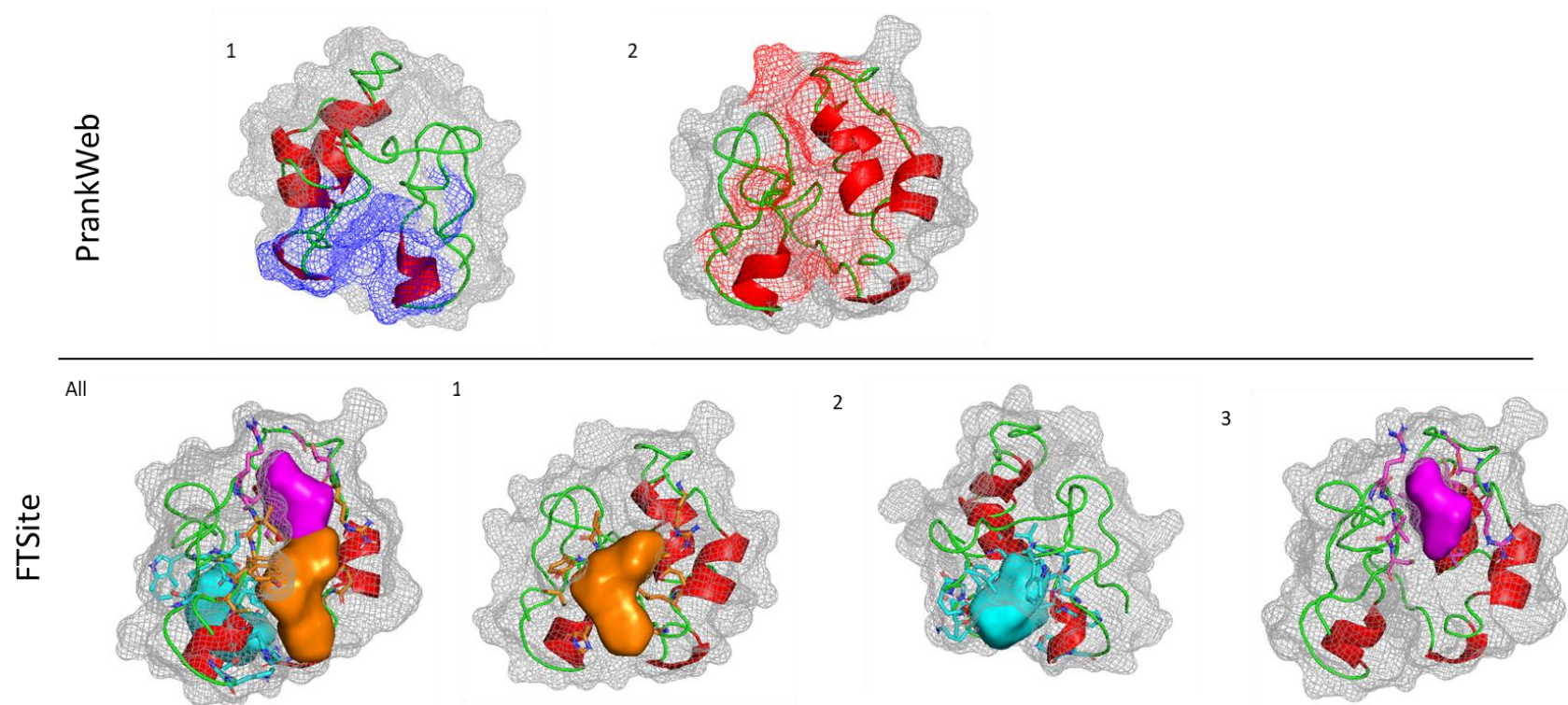


Figure 4.26: An alignment of the Def-S18/F45 variation of Def amino acid sequences highlighting all residues within predicted active sites. The boxed area shows previously document as the active site within AAB36306 and BAA36401. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment. FTSite models show the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green,  $\alpha$ -helicase are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

Def-S18/I82

PrankWeb 1	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	S	A	L	P	A	G	D	E	T	R	I	D	L	E	T	L	E	E	D	L	R	L	V	D	G	A	Q	V	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	K	S	G	Y	C	S	R	G	I	C	Y	C	T	N
PrankWeb 2	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	S	A	L	P	A	G	D	E	T	R	I	D	L	E	T	L	E	E	D	L	R	L	V	D	G	A	Q	V	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	K	S	G	Y	C	S	R	G	I	C	Y	C	T	N
PrankWeb 3	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	S	A	L	P	A	G	D	E	T	R	I	D	L	E	T	L	E	E	D	L	R	L	V	D	G	A	Q	V	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	K	S	G	Y	C	S	R	G	I	C	Y	C	T	N
FTSite 1	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	S	A	L	P	A	G	D	E	T	R	I	D	L	E	T	L	E	E	D	L	R	L	V	D	G	A	Q	V	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	K	S	G	Y	C	S	R	G	I	C	Y	C	T	N
FTSite 2	- - - -	V	L	A	F	L	T	L	C	A	V	A	V	S	A	L	P	A	G	D	E	T	R	I	D	L	E	T	L	E	E	D	L	R	L	V	D	G	A	Q	V	T	G	E	L	K	R	D	K	R	V	T	C	N	I	G	E	W	V	C	V	A	H	C	N	S	K	S	K	K	S	G	Y	C	S	R	G	I	C	Y	C	T	N

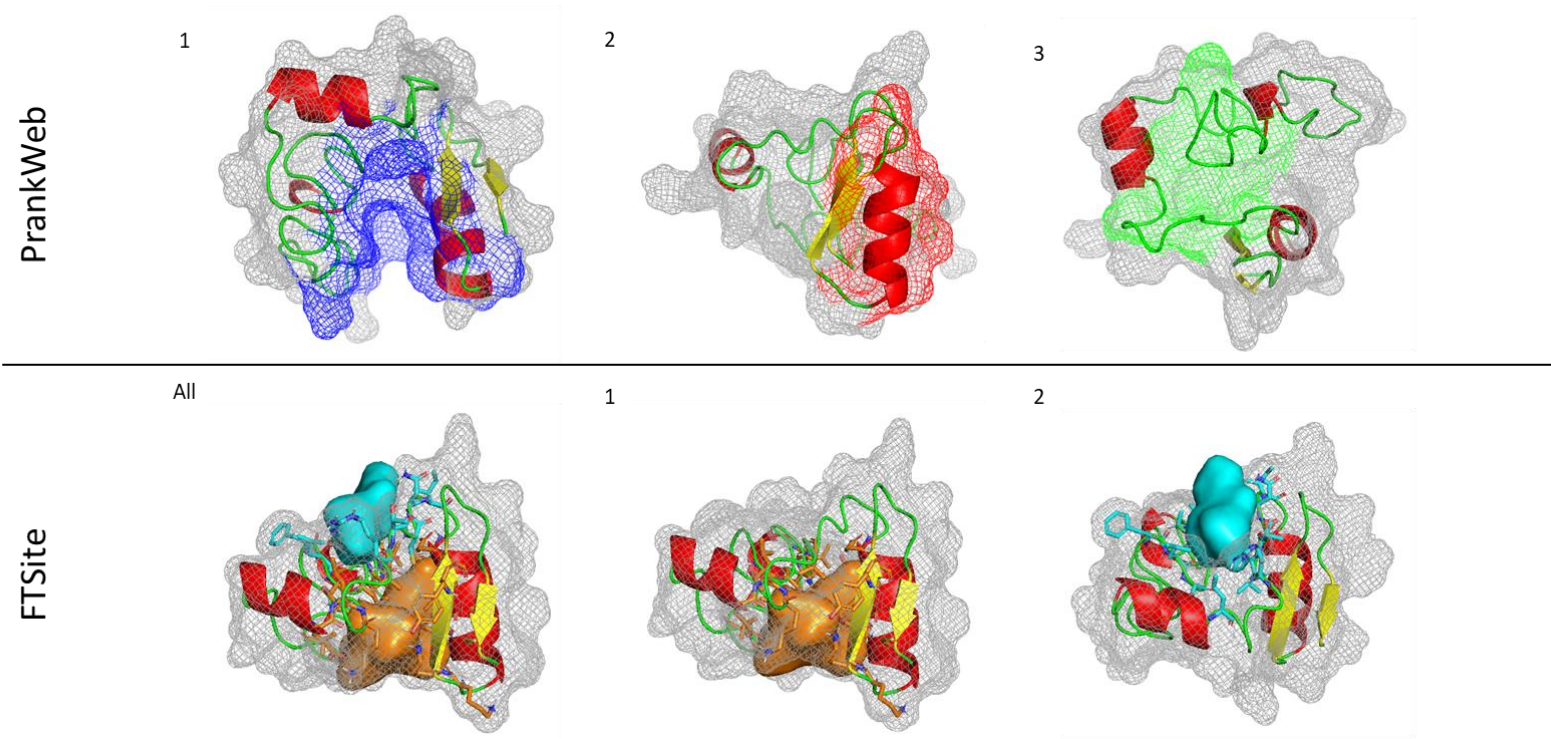


Figure 4.27: An alignment of the Def-S18/I82 variation of Def amino acid sequences highlighting all residues within predicted active sites. The boxed area shows previously document as the active site within AAB36306 and BAA36401. A model of each active site is also presented, PrankWeb predictions are shown on the top row and FTSite predictions below. All PrankWeb predictions illustrate the region of the binding site in the corresponding colour to the alignment. FTSite models show the ligand bind site in relation to the binding pocket, the colour corresponds with the alignment above. All models were visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre), all coil structures are shown in green,  $\alpha$ -helicase are shown in red and  $\beta$ -sheets in yellow with direction shown by the arrow.

## 4.4: Discussion

This chapter aimed to assess the structural and functional consequences of nucleotide variation in relation to trypanosome infection and protein evolution. As expected AttA and Def indicated different evolutionary histories: AttA continued to exhibit minimal variation within the *G. m. morsitans* subpopulation, exhibiting just five protein variants. Whereas Def exhibited 11 protein variants and with showed signs of positive selection indicating the adoption of an advantageous allele.

Chapters 2 and 3 suggested that genetic variation of AttA was under balancing selection to maintain multiple functional alleles within the population (Clark and Wang, 1997; Woolhouse *et al.*, 2002; Chapman *et al.*, 2019). This chapter presented no clear evidence to dispute this observation, as protein variation was negligible and the infection rate within the AttA variant was lower than the overall population infection rate (Chapter 3). Furthermore, there was no indication of positive selection and only inconclusive evidence for purifying selection within the amino acid sequence. This could be an indication of negative frequency-dependent selection whereby protein variant fitness increases as frequency decreases within the population, however, this remains undetermined due to the small population size.

Interestingly, the lack of clear purifying selection suggests that the low nucleotide variation of AttA is maintained by alternative processes. The presence of one abundant protein variant could be the result of a founding effect following a population bottleneck within the sample population (Raupach *et al.*, 2010), while it is also indicative of a concerted evolutionary history (Liao, 1999; Moran *et al.*, 2008). Both of these observations support the previous chapter results, reinforcing the concepts of concerted evolution within the *Glossina* attacin gene family and recent population expansion within the sample population.

To the best of the author's knowledge, this is the first time the three-dimensional structure and active site of AttA has been examined to any great extent. Structural variation was found to be minimal, with just one protein variant illustrating clear variation. The variation that was detected likely results from the changes in amino acid properties observed between variants. It is hard to decipher whether variation within the predicted active sites



is a result of amino acid variation or the absence of stabilising residues in the N-terminal as a result of PCR amplification and sequencing.

Previous chapters have indicated that Def is regulated by elevated levels of natural selection (Chapters 2) and that nucleotide variation within the wild *G. m. morsitans* population is regulated by balancing selection (Chapter 3). A culmination of the results in this chapter substantiates this observation and further concludes that *Def* is currently under both balancing and positive selection. Firstly, the results suggest that there is relationship between trypanosome infection frequency and the three common protein variants. Secondly, significant positive selection as detected at codon 18; suggesting that Def is under positive frequency-dependent selection (Ayala and Campbell, 1974), and therefore, undergoing the adoption of an advantageous allele under the Red Queen arms race (Woolhouse *et al.*, 2002). Interesting, this would initially appear to contradict the observation of balancing selection within the sample population, however, the full extent of amino acid variation must be considered.

Amino acid variation at codon 18 is complicated due to the presence of three amino acid variants. No radical changes in amino acid biochemical properties were identified between Ser18 and Thr18, though there was a clear differentiation between infection rates between the two groups. The second mutation at codon 18 (Ile18) illustrated multiple radical property changes though due to a small sample size direct comparison of infection rates was impossible. Therefore, given the empirical evidence of both positive and balancing selection with Def, it can be hypothesised that the adoption of an advantageous Thr18 codon via positive frequency-dependent selection and the Red Queen arms race is occurring concomitantly with balancing selection to maintain three functional proteins in the population. To the best of the author's knowledge, this phenomenon is not well document in current literature.

The structural variation within Def was far greater than that observed in AttA. While this is less surprising within the more variable N-terminal, the rigid cytosine stabilised  $\alpha$ -helix and anti-parallel  $\beta$ -sheet structure of the insect defensin C-terminal is absent in eight of the wild variants (Def-S18, Def-42, Def-I82, Def-I82, Def-F45, Def-S18/F45, Def-I18/F45 and Def-S18/F45/I82) (Varkey *et al.*, 2006; Altincicek and Vilcinskis, 2007; Wiesner and

Vilcinskas, 2010). Although, this is likely due to inaccuracies during structural prediction, rather than a true structural mutation.

Interestingly, structural variation could also support the observations of balancing and positive selection with Def. Protein structures form three clear clusters, one of which contained primarily highly susceptible sample. This is it feasible that structures within this cluster exhibit a disadvantageous structural characteristic, therefore positive selection could promote the fixation of more advantageous structures and balancing selection is being employed to maintain the preferable protein structures within the population. However, the exact nature of this structural variation remains elusive.

There is little variation within the functional region of the amino acid sequences, therefore it can be assumed that functional variance is a result of structural change rather direct mutation. Indeed, this supports the observations of Zuckerkandl and Pauling (1965), and Bloom and Arnold (2009) that the majority of non-synonymous mutations have minimal direct impact on functionality.

## 4.5: Conclusion

In this chapter, structural variation within *G. m. morsitans* AttA and Def was shown to be evolving under different selective pressures. AttA shows minimal signs of selective pressure, and the limited protein variation supports the previous findings of chapters 2 and 3 that AttA has evolved under concerted evolution, resulting in a reduction genetic variation that is not explained by purifying selection. Additionally, the observed population expansion event (Chapter 3) has resulted in a founding event within the AttA protein, implementing one dominant protein variant throughout the population. Sharply contrasted to this is the observation of contemporary direction selection within Def. Positive frequency-dependent selection and the Red Queen arms race seem to be driving the adoption of the advantageous Thr18 allele within the population, while balancing selection, as identified in Chapter 3, is maintaining the expression of three functional protein variants, Def, Def-F45 and F-I82, within the population.

The full implications of these findings for tsetse evolution and the current understanding of tsetse-trypanosome interactions requires further research. However, this is the first time

that direction selection has been observed within this relationship, indicating that the interspecies arms race and coevolution are playing an active role in the ongoing evolution of *G. m. morsitans*.

## 5: The identification and characterisation of Toll-like receptor protein families within the *Glossina* genome assemblies

### 5.1: Introduction

Toll-like receptors (TLRs) are one of two vital stimulators of the innate immune response, first identified within *Drosophila* (Anderson *et al.*, 1985). They have since been identified in both vertebrate and invertebrate taxa and are responsible for stimulating the expression of antimicrobial proteins in response to pathogen invasion (Hopkins *et al.*, 2005; Coscia *et al.*, 2011). Toll-like receptors play a critical role in both the innate immune system and embryo development, providing vital cues for dorsal/ventral differentiation within the early stages of growth (Anderson *et al.*, 1985; Jang *et al.*, 2006). Coscia *et al.* (2011) noted that TLRs are highly conserved across all classes sharing common ancestors and architecture, though TLRs do appear to have evolved separately in invertebrates and vertebrates.

Following the initial characterisation of Toll in *Drosophila melanogaster* (Anderson *et al.*, 1985; Valanne *et al.*, 2011; Levin and Malik, 2017), a total of nine TLR genes (TLR1-9) have been identified within the *D. melanogaster* genome. There are clear signs of orthology between the TLRs identified in *D. melanogaster* and other dipteran families, including mosquitoes (Christophides *et al.*, 2002). Within the *Anopheles* genus orthologs of TLR1, 5-9 have been identified, along with two additional TLR genes, TLR10 and TLR11. Furthermore, a recent study investigating the evolution of key proteins within the insect TLR pathway (Lima *et al.*, 2021) indicated a total of seven TLR orthologues within the *Glossina brevipalpis* genome and six within the *Glossina f. fuscipes*. Orthologues of *D. melanogaster* TLR2, 6, 7, 8 and 9 were detected in both species, while two orthologues of either TLR1, 3, 4 or 5 were detected in *G. brevipalpis*, while a single orthologue in *G. f. fuscipes* (Lima *et al.*, 2021). The different number of TLR genes within these species may indicate evolution in response to pathogen specific interactions (Hill *et al.*, 2019)

Species-specific evolution has been observed previously within the class Insecta (Coscia *et al.*, 2011) and is supported further by the presence of additional TLR1 and TLR5 genes within the *Anopheles* genome, resulting in TLR1A/B and TLR5A/B variants (Christophides *et*

*al.*, 2002). Furthermore, despite the clear orthology between TLR genes in dipteran genera, high levels of amino acid divergence have been observed between *D. melanogaster* and mosquitoes indicating high levels of adaptive selection within species (Schlenke and Begun, 2003; Sackton *et al.*, 2007; Juneja and Lazzaro, 2009; Kafatos *et al.*, 2009).

Although TLRs are highly conserved across both vertebrates and invertebrates, they appear to have evolved independently of each other (Luo and Zheng, 2000). The evolutionary history of arthropod TLRs has been well documented (Christophides *et al.*, 2002; Levin and Malik, 2017), and show different levels of similarity and divergence between taxa and gene families. The general structure of observed phylogenies forms two genetically distinct clades, one containing TLRs 1 and 3-5, while the other contains TLRs 2 and 6-9 (Fig. 5.3) (Levin and Malik, 2017). A close evolutionary relationship has been documented between TLRs 2 and 7, and 3 and 4 in all arthropod taxa (Fig. 5.3), while a subclade off the *Culicidae* TLR1A/B and TLR5A/B genes illustrates independent evolution of TLR families within mosquitoes (Fig. 5.3B) (Christophides *et al.*, 2002; Levin and Malik, 2017).

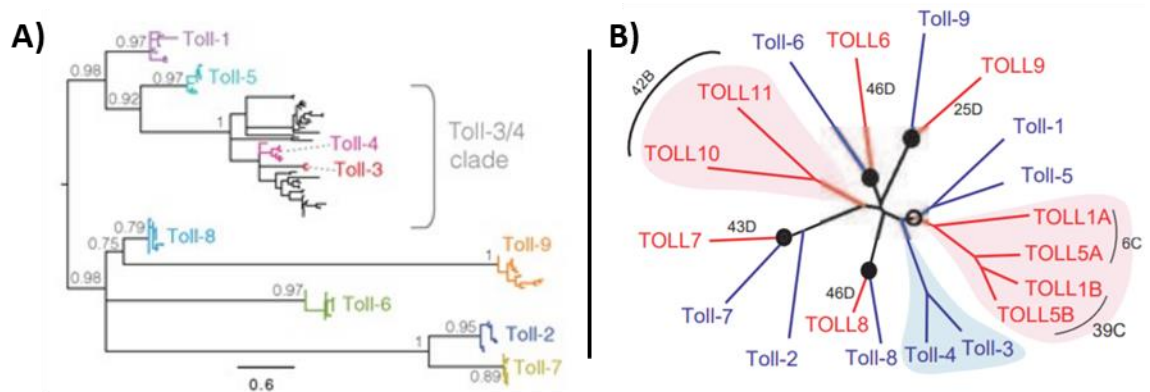


Figure 5.1: The phylogeny of TLR gene within arthropod taxa. A) The genetic diversity of the TIR domains within 12 *Drosophila* species (Levin and Malik, 2017). B) The evolutionary history of TLR genes in Culicidae (Red) and other arthropod species (Blue) (Christophides *et al.*, 2002).

The structure of TLR proteins has been extensively documented, monomers adopt a horseshoe shape however, the natural formation of both homo and heterodimers (Khan *et al.*, 2004; Jin *et al.*, 2007) results in the characteristic “M” shape most often associated with TLR proteins (Fig. 5.2). As a Type-1 glycoprotein (Goldstein, 2007), each TLR monomer comprises three distinct domains: an extracellular receptor domain, a transmembrane domain, and a distinctive intercellular Toll/Interleukin-1 receptor (TIR). The largest, the N-terminal extracellular domain (ectodomain), typically composed of parallel  $\beta$ -sheets on the concave surface with helices forming the convex outer surface and consisting of 16-25

leucine-rich repeats (LRR) domains. LRR domains are approximately 24 to 29 amino acids in length and contains one of two conserved amino acid motifs XLXXLXX and X $\phi$ XX $\phi$ XXXXFXXLX ( $\phi$  = hydrophobic residue; X = any residue) (Uematsu and Akira, 2008) (Fig. 5.2A).

Following the ectodomain is a single helical transmembrane region; this traverses the phospholipid bilayer connecting the ectodomain to the TIR domain (Bell *et al.*, 2005). The TIR domain is essential to the successful transmission of the immune response, binding to and releasing MyD88, SARM, TRIF, TRAM and MAL enabling the continuation of the signalling cascade (O'Neill and Bowie, 2007). Structurally, the TIR domain consists of five central parallel  $\beta$ -sheets, surrounded by  $\alpha$ -helices and a characteristic single "BB loop" (Fig. 5.2B) (Xu *et al.*, 2000; Khan *et al.*, 2004).

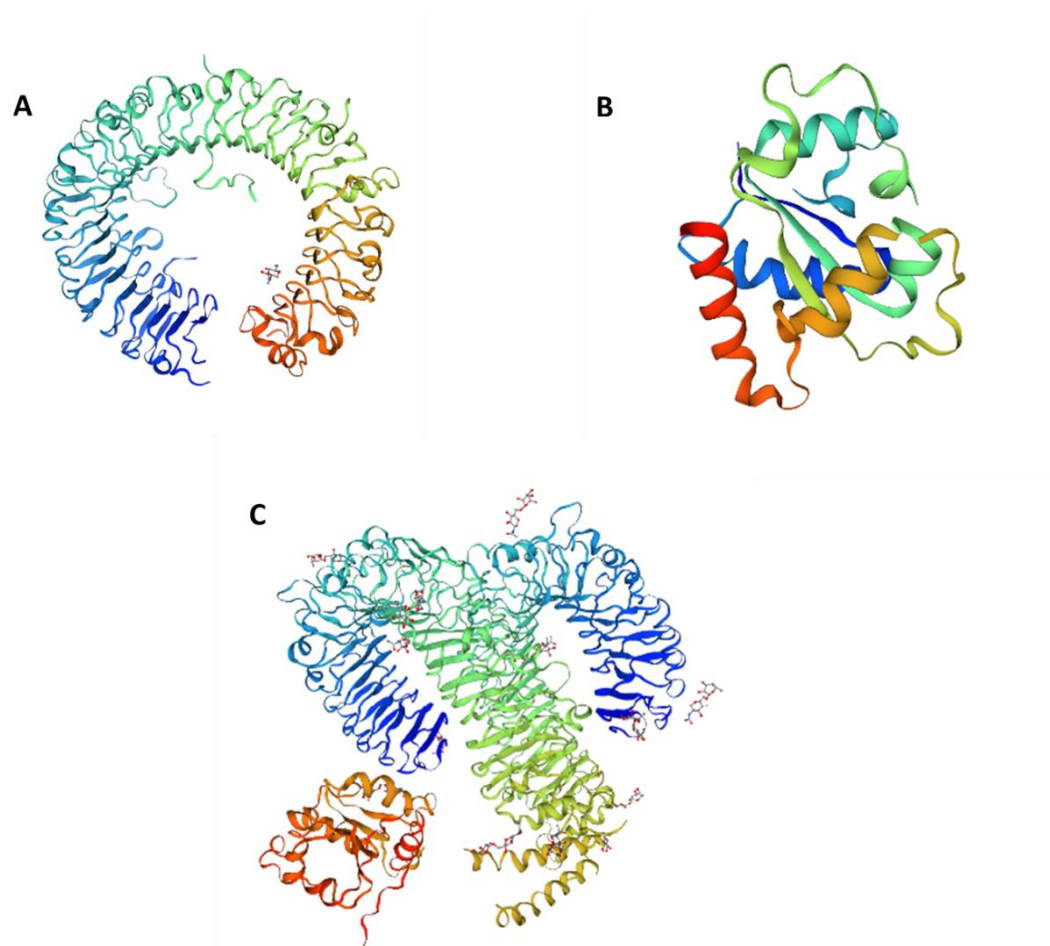


Figure 5.2: The overall structure of TLR proteins illustrating all three subdomains. A) The ectodomain of TLR 9 as a monomer (Ohto *et al.*, 2015). B) The structure of the TIR domain, comprising the central  $\beta$ -sheets surrounded by  $\alpha$ -helices as recorded by Xu *et al.* (2000). C) The homodimer structure of TLR5 exhibiting all three subdomains. The ectodomain can be seen in green and blue; the transmembrane region in yellow and the TIR in red (Zhou *et al.*, 2012). All images were produced using SWISS-MODEL (Guex *et al.*, 2009; Waterhouse *et al.*, 2018).

While the structure and function of TLRs is conserved among all classes, the mode of action varies between vertebrates and invertebrates, as described in detail in Chapters 1, the following is a recap of this crucial immune pathway within invertebrates. The binding of endogenous ligand proteins, such as Spätzle (Spz) (Weber *et al.*, 2003), to stimulate the TLR pathway and immune response (Stein and Nüsslein-Volhard, 1992). The cleaving of Spz by the proteolytic cascade release of the pro-domain, exposing the Spz C-terminal, thereby enabling binding with the TLR extracellular domain (Lemaitre *et al.*, 1996; Weber *et al.* 2003; Tanji *et al.*, 2007; Arnot *et al.* 2010; Valanne *et al.*, 2011). Upon Spz-TLR binding, the TIR domain binds to Myeloid differentiation primary response 88 (MyD88), which subsequently binds to Tube and Pelle, forming the MyD88-Tube-Pelle heterotrimeric complex (Horng and Medzhitov, 2001; Sun *et al.*, 2002; Tauszig-Delamasure *et al.*, 2002). This complex is vital for the degradation of the Dorsal/Dif-Cactus (Cact) complex, enabling the nuclear translocation of Dorsal/Dif, which in turn results in the synthesis of AMPs, such as Attacin, Defensin and Diptericin (Wu and Anderson 1998; Akira *et al.*, 2006; Valanne *et al.*, 2011).

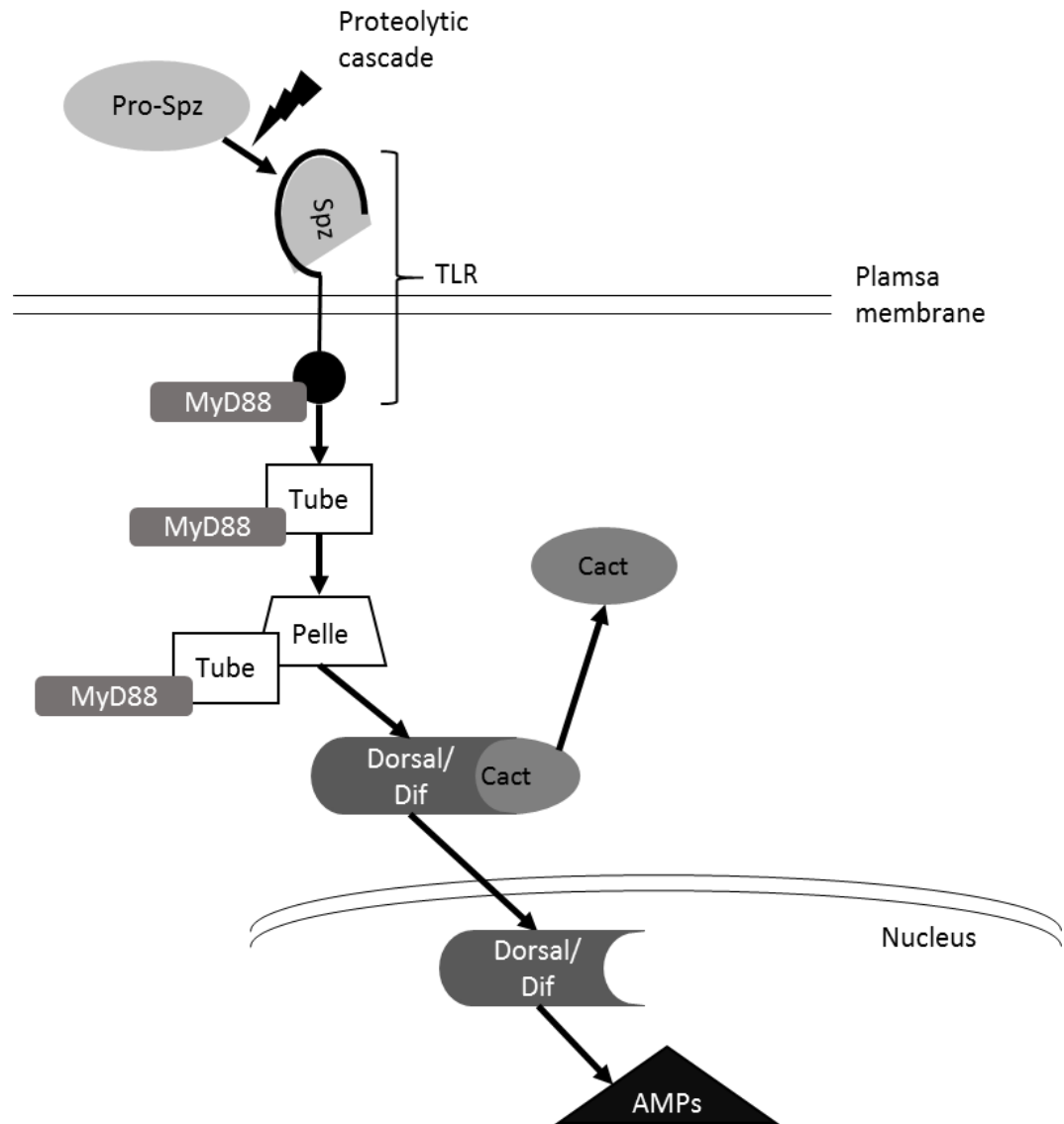


Figure 5.3: A model of the Spz-TLR signaling cascade in *D. melanogaster* (adapted from the model published by Lemaitre et al., 1996).

### 5.1.1: Aims and Objectives

The aim of this chapter is to assess genetic and structural variation between the TLR genes within the *Glossina* genome. Building on the observations of Lima *et al.* (2021), the aim is to identify all TLR genes within the *Glossina* genomes and assess the evolution history, nucleotide and structural variation between them.

To achieve this, the members of the TLR gene family within the *Glossina* genome must first be identified and characterised using similar method to those used in chapter 2.

TLR orthologues from *D. melanogaster*, *S. calcitrans*, *M. domestica*, *Ae. aegypti* and *An. gambiaensis* will be used as control sequences to search the six *Glossina* genomes available



on VectorBase for related species (Giraldo-Calderón *et al.*, 2015). Mapping and characterising the gene structures with the genomes will help to demonstrate the evolution of these genes within *Glossina*, *Drosophila*, *Musca* and *Stomoxys*. The identification of variation between haematophagic genera and other dipteran taxa may indicate species or lifestyle specific evolution. To understand the evolutionary history of the *Glossina* TLR proteins, standard phylogenetic analysis will be used to assess the evolutionary relationship both within and between TLR families.

Given the previously documented conservation of TLR gene interspecies nucleotide analysis can help to highlight areas of conservation and variation within each gene. Furthermore, this may provide an indication to the selective pressures influencing TLR evolution. Finally, prediction of the three-dimensional protein structures of each protein will be undertaken and structural variation will be assessed.

## 5.2: Methods

### 5.2.1: Identification of TLR genes within *Glossina* and other dipteran genera

Previously identified TLR genes from *D. melanogaster*, *M. domestica*, *S. calcitrans*, *An. gambiae* and *A. aegypti* retrieved from VectorBase (Giraldo-Calderón *et al.*, 2015) and FlyBase (Thurmond *et al.*, 2019) were used to mine the available *Glossina* genome assemblies for orthologs. tblastn searches were conducted in both NCBI and VectorBase (Giraldo-Calderón *et al.*, 2015) using the protein sequences of reference genes shown below (Table 5.2). tblastn searches within the available *Glossina* genomes on VectorBase yielded orthologous transcripts, with the most statistically relevant E values and highest identity match chosen as the most likely match for the gene.

Table 5.1: Reference genes for each TLR identified within related dipteran species. The TLR target gene is given, along with the species it was identified within, the gene accession number and the database used to generate a sequence.

TLR gene	Species	Gene accession number	Database
<i>TLR1</i>	<i>D. melanogaster</i>	FBng0262473	FlyBase
<i>TLR2</i>	<i>D. melanogaster</i>	FBgn0004364	FlyBase
<i>TLR3</i>	<i>D. melanogaster</i>	FBgn0015770	FlyBase
<i>TLR3</i>	<i>M. domestica</i>	MDOA009473	VectorBase
<i>TLR4</i>	<i>D. melanogaster</i>	FBgn0032095	FlyBase
<i>TLR5</i>	<i>D. melanogaster</i>	FBgn0026760	FlyBase
<i>TLR5</i>	<i>M. domestica</i>	MDOA007681	VectorBase
<i>TLR6</i>	<i>D. melanogaster</i>	FBgn0036494	FlyBase
<i>TLR7</i>	<i>D. melanogaster</i>	FBgn0034476	FlyBase
<i>TLR8</i>	<i>D. melanogaster</i>	FBgn0029114	FlyBase
<i>TLR9</i>	<i>D. melanogaster</i>	FBgn0036978	FlyBase
<i>TLR10</i>	<i>An. gambiae</i>	AGAP011187	VectorBase
<i>TLR10</i>	<i>A. aegypti</i>	AAEL00400	VectorBase
<i>TLR11</i>	<i>A. aegypti</i>	AAEL009551	VectorBase
<i>TLR13</i>	<i>S. calcitrans</i>	SCAU006576	VectorBase

### 5.2.2: Transcript and domain structure

VectorBase transcripts were used to create maps of each gene transcript, illustrating the position of introns, exons, and protein domains along the CDS. Protein domains were identified using the Simple Modular Architecture Research Tool (SMART) (Letunic and Bork, 2017) and Pfam (El-Gebali *et al.*, 2019) online software, and mapped on to the transcript map using the 'Splice variant' information from VectorBase (Giraldo-Calderón *et al.*, 2015).

### 5.2.3: Phylogenetic analysis

Amino acid sequences for each gene CDS were aligned using the MUSCLE (Multiple Sequence Comparison by Log-Expectation) online sequence alignment tool (Madeira *et al.*, 2019) prior to analysis. Phylogenetic analysis was conducted in MEGAX (Kumar *et al.*, 2018), both the neighbour-joining and maximum likelihood methods were used. The Poisson model was used for the neighbour-joining method, while a model test indicated that the Le-Gascuel 2008 model with Gamma distribution was best suited for the Maximum-likelihood method. 1000 bootstrap replicates were used to ensure the most statistically significant tree was constructed. Furthermore, all sites with less than 50% coverage were excluded from the final tree. Sequences for *M. domestica*, *S. calcitrans*, *L. cuprina*, *D. melanogaster*, *A. aegypti* and *An. gambiae* were included as outgroups where appropriate.

### 5.2.4: Interspecies variation

Nucleotide variation across the *Glossina* TLR genes was compared using DnaSP (V6) (Rozas *et al.*, 2017), using the same method as described in Chapter 2, section 2.2.6. The window size was increased to 50 and step length to 10, each primary protein domain was highlighted to illustrate genetic variation within the CDS.

Principle Component Analysis was conducted to provide an estimation of evolutionary relationships. The pairwise distance was calculated in MEGAX (Kumar *et al.*, 2018), using the Poisson correction model (Zuckerkand and Pauling, 1965). All sites with less than 50 % coverage were eliminated. A matrix was then constructed and PCA run in PAST3 (Hammer *et al.*, 2001). For further detail please refer too chapter 2, section 2.2.7.

HyPhy was not used in this analysis as the aim was to assess genetic variation between TLR genes within the *Glossina* genus rather than selection.

### 5.2.5: Three-dimensional protein prediction

Prediction of the protein three-dimensional structure was undertaken using the I-TASSER online software (Yang and Zhang, 2015) as described in chapter 2 section 2.2.8, and SWISS-MODEL (available at: <https://swissmodel.expasy.org>) (Guex *et al.*, 2009; Waterhouse *et al.*, 2018). Templates for SWISS-MODEL structural predictions were produced protein models from *D. melanogaster* TLR1 and *G. m. morsitans* TLR2. All models were visualised and aligned using PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre).

Comparison of 3D protein structures was undertaken using DALI online server (Holm, 2019). Protein database (PDB) files were uploaded to the server and an All-vs-All analysis was undertaken (Chapter 2, section 2.2.8). This produced a heatmap and distance matrices used to produce a second PCA in the same manner as described above (refer to section 2.2.7).

## 5.3: Results

Nine predicted TLR genes were identified within the available *Glossina* genome assemblies. Of these, six genes, namely TLRs1-2, 6-9 were found to contain all three structural domains characteristic of TLRs. While the remaining three genes identified, TLR 3, 5 and 13, exhibited partial structural domains. All nine genes were identified within the *Morsitans* group species and the Palpalis group species *G. f. fuscipes*, and eight genes were identified within *G. palpalis gambiensis* and *G. brevipalpis*.

### 5.3.1: Gene structure and characterisation

In order to characterise the predicted TLR genes, their structure and coding domains were mapped using the transcripts found on VectorBase. Protein domains within the coding sequence were identified using SMART (Letunic and Bork, 2017) and Pfam (El-Gebali *et al.*, 2019) searches to further identify and characterise the transcript.

#### 5.3.1i: Toll-like receptor 1

Toll-like receptor 1 genes showed a greater degree of variation across the *Glossina* genus. Predicted TLR1 transcripts within *G. austeni*, *G. pallidipes* and *G. f. fuscipes* are all of a similar length and structure, however, the *G. f. fuscipes* transcript contains an additional intron within the TIR domain (Fig. 5.4). Despite the much larger CDS observed within the *G. m. morsitans* transcript (Fig. 5.4), the encoded domains are identical to that of the other *Morsitans* group species and *G. f. fuscipes*. The transcripts of both *G. palpalis gambiensis* and *G. brevipalpis* were found to be considerably shorter than those identified in other *Glossina* species though both contain a N-terminal intron across an LRR domain. While *G. f. fuscipes* is the only species to exhibit an intron within the TLR domain it is possible that this is specific to species within the Palpalis group, though without further analysis of the *G. p. gambiensis* gene this cannot be confirmed.

The coding structure of the predicted *TLR1* gene consists of 13 consecutive LRR domains, followed by a LRR C-terminal flanking region (LRR-CT) and LRR N-terminal flanking region (LRR-NT), a further two LRRs domains and a LRR-CT then precede the Transmembrane and TIR domains (Fig. 5.4). Initial searches within the *G. pallidipes* predicted TLR1 transcript were missing the transmembrane domain, however, examination of the CDS showed that

the domain was present within the sequence. The CDS of *G. brevipalpis* was found to code for nine consecutive LRR domains, followed by a LRR-CT, the transmembrane domain and TIR domains. While *G. palpalis gambiensis* was found to be missing several domains observed in other species. The transcript codes for seven LRR domains and a single LRR-CT region, however, several LRR domains, the transmembrane domain and the characteristic TIR domains are missing (Fig. 5.4).

The structure within *Glossina* species is similar to that observed within other dipteran genera, though the exact number of LRR regions does vary between genera as does the structure and position of non-coding regions. Despite this, the structure of the region appears to be predominantly conserved across dipteran genera with a N-terminal intron splitting the LRR sequence in all species, notably however, no other dipteran species exhibit the same intron as *G. f. fuscipes* within the TIR region.

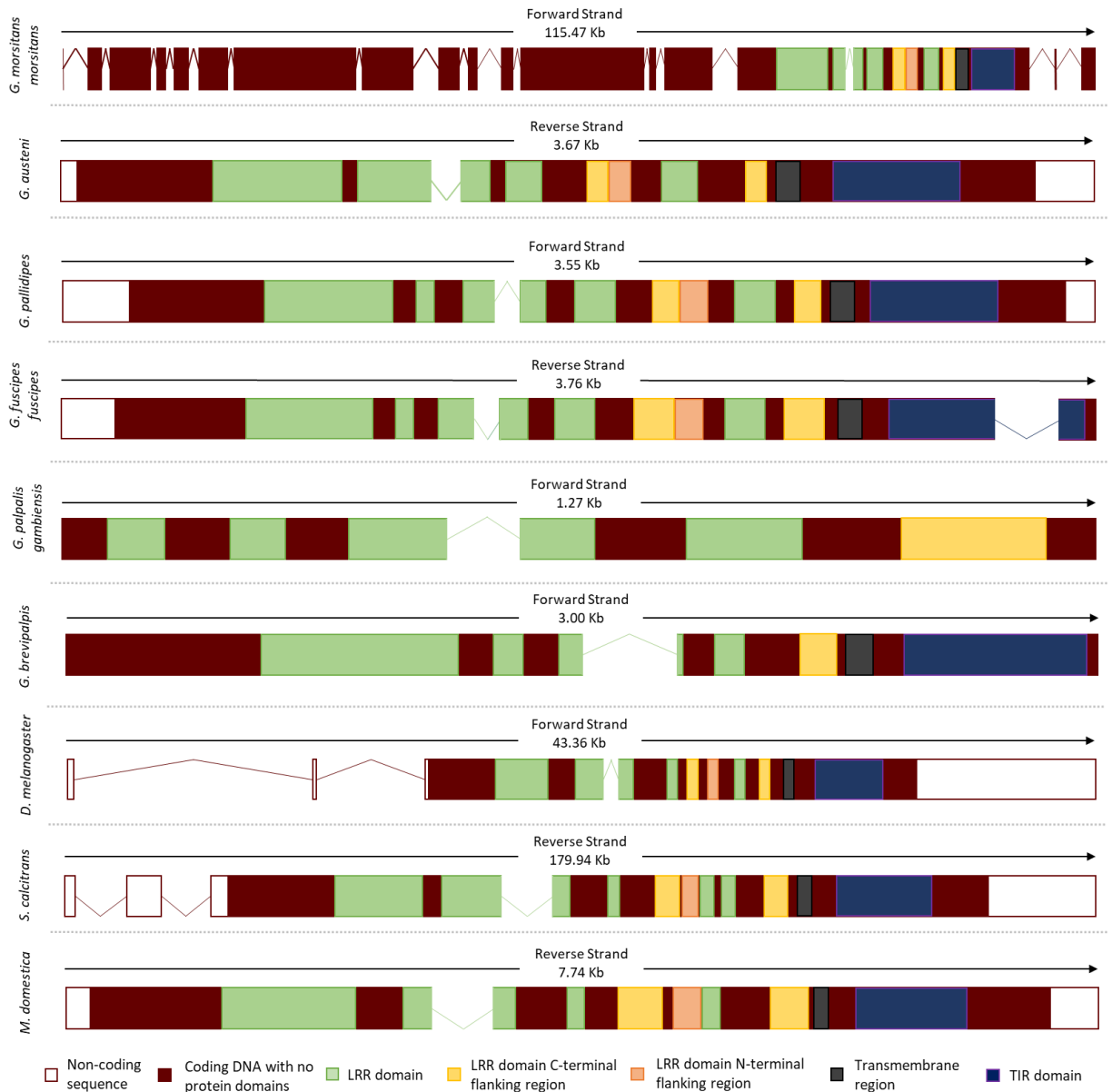


Figure 5.4: A full gene alignment of *TLR1* genes within the *Glossina* genome assemblies with reference genes of *D. melanogaster*, *S. calcitrans* and *M. domestica*. Drawn in Microsoft PowerPoint and adapted from figure produced from VectorBase and FlyBase (not to scale). Exons can be seen in the boxed areas, while introns are represented by linear areas. The coding strand and full transcript length are given above each gene. Alignments were constructed based on VectorBase and FlyBase transcripts, and SMART domains search results for the corresponding amino acid sequences. Encoded protein domains are denoted by colour within the coding regions of each gene. *Glossina* species used the transcripts of the genes seen in Table 5.3, while *M. domestica* and *S. calcitrans* used VectorBase genes MDOA005484 and SCAU010787 transcripts respectively, *D. melanogaster* used FlyBase transcript FBgn0262473.

### 5.3.1ii: Toll-like receptor 2

The structure of the predicted *TLR2* transcripts within the *Glossina* genomes is almost identical between species, with 19 consecutive LRR domains preceding a LRR-CT and LRR-NT domain, four further LRR domains, the transmembrane and TIR domains follow these. The exception to this is *G. brevipalpis*, which contains 18 LRR domains rather than the 19 observed in the other *Glossina* species.

This structure also appears to be conserved across the Dipteran genera, with *D. melanogaster*, *M. domestica* and *S. calcitrans* showing an almost identical structure to that observed within *Glossina* species. However, as with *TLR1*, some variation can be observed in the presence of noncoding region at the N- and C-terminal and the number of LRR present in each dipteran *TLR2* gene. Despite this, the overall structure of the genes is identical with all protein domains being coded for on a single exon (Fig. 5.5).



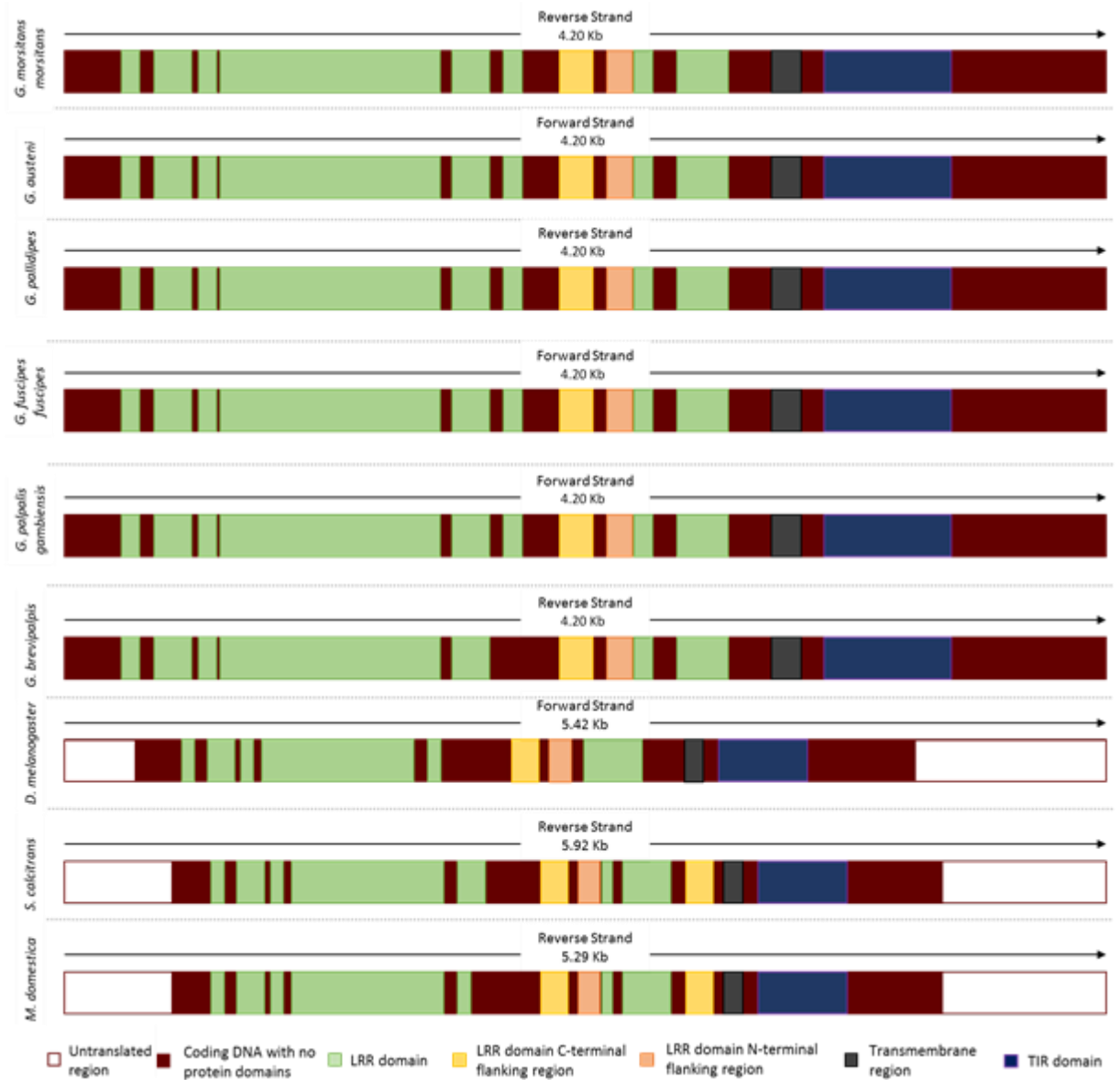


Figure 5.5: A full gene alignment of *TLR2* genes within the *Glossina* genome assemblies with reference genes of *D. melanogaster*, *S. calcitrans* and *M. domestica*. Drawn in Microsoft PowerPoint and adapted from figure produced from VectorBase and FlyBase (not to scale). Exons can be seen in the boxed areas, while introns are represented by linear areas. The coding strand and full transcript length are given above each gene. Encoded protein domains are denoted by colour within the coding regions of each gene. *Glossina* species used the transcripts of the genes seen in Table 5.3, while *M. domestica* and *S. calcitrans* used VectorBase genes MDOA015408 and SCAU008704 transcripts respectively, *D. melanogaster* used FlyBase transcript FBgn0004364.

### 5.3.1iii: Toll-like receptor 3

Predicted *TLR3* genes within the *Glossina* genus exhibit two structures, one consisting of 13 exons and the other 14 (see section 5.3.2). However, the protein domains within each of these genes is identical with 25 LRR regions preceding a single LRR-CT domain, though there is no indication of a transmembrane or TIR domain within any of the predicted *Glossina* genes (Fig. 5.6). The highest level of similarity between *Glossina* species can be seen between *G. austeni* and *G. pallidipes*, each consisting of 13 introns with a similar distribution of protein domains throughout the 3,429 bp CDS. As both of these species belong to the Morsitans group this similarity is unsurprising, however, the variation between these species and *G. m. morsitans* indicates that similarity within the *Glossina* species groups is not guaranteed.

The structure observed within the *Glossina* species is similar to that identified within *S. calcitrans* and *M. domestica*, however, this structure varies considerably to the gene annotation within the *D. melanogaster* genome, which contains three LRR-NT domains and a TIR domain, though the transmembrane region is still missing (Fig. 5.6).

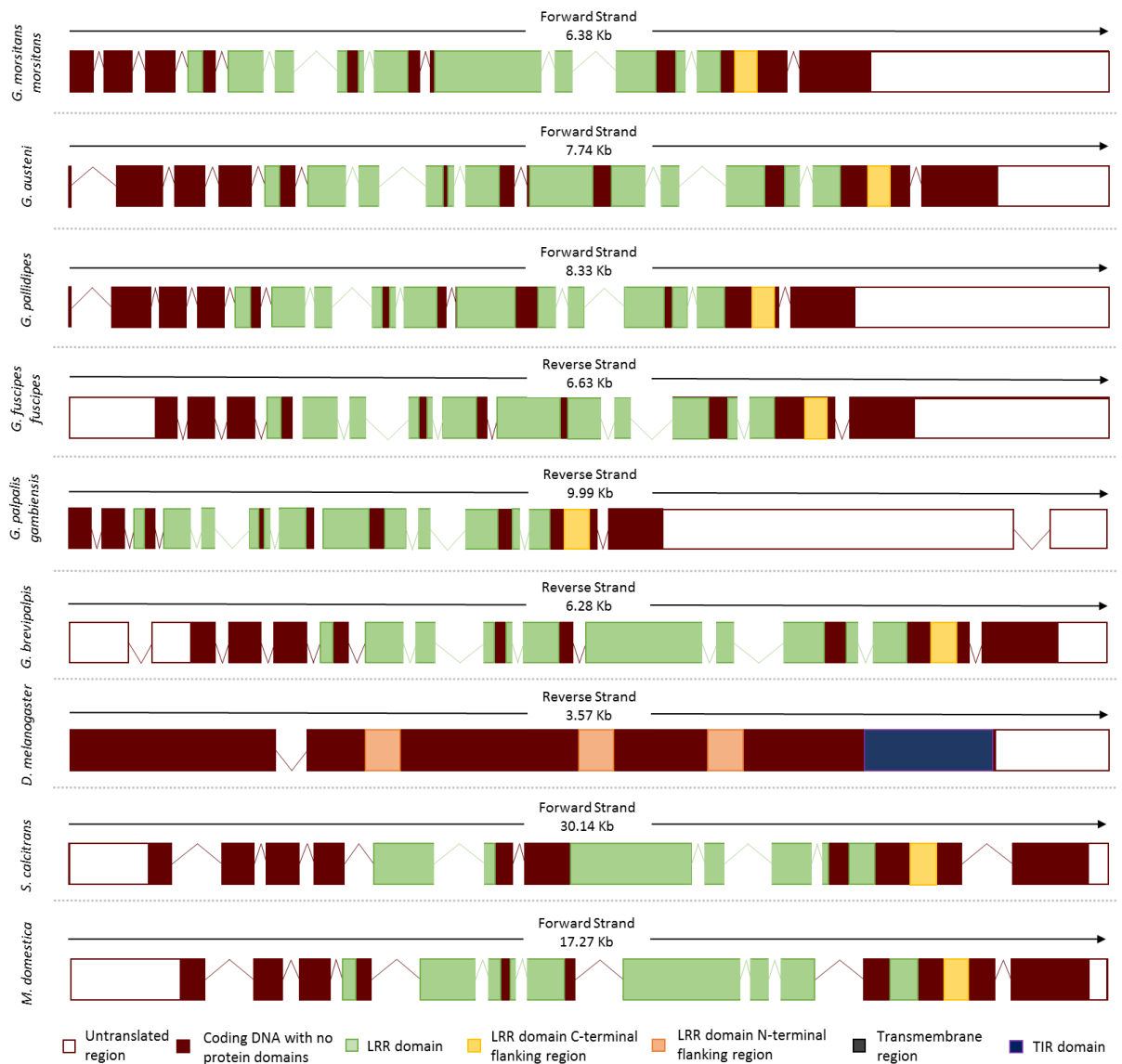


Figure 5.6: The gene alignment of *TLR3* genes within the *Glossina* genome assemblies with reference genes of *D. melanogaster*, *S. calcitrans* and *M. domestica*. Drawn in Microsoft PowerPoint and adapted from figure produced from VectorBase and FlyBase (not to scale). Exons can be seen in the boxed areas, while introns are represented by linear areas. The coding strand and full transcript length are given above each gene. Encoded protein domains are denoted by colour within the coding regions of each gene. *Glossina* species used the transcripts of the genes seen in Table 5.3, while *M. domestica* and *S. calcitrans* used VectorBase genes MDOA015408 and SCAU008704 transcripts respectively, *D. melanogaster* used FlyBase transcript FBgn0004364.

#### 5.3.1iv: Toll-like receptor 5

The predicted structure of *Glossina TLR5* genes is almost identical across all species. *Glossina m. morsitans*, *G. austeni* and *G. brevipalpis* share identical gene structures exhibiting a single exon, while the CDS of *G. pallidipes*, *G. palpalis gambiensis* and *G. f. fuscipes* TLR5 contain introns near the N-terminal. The distribution of protein domains within the *Glossina TLR5* coding regions is also relatively uniform with five LRR domains preceding an LRR-CT and the transmembrane region. The exception to this being *G. palpalis gambiensis* which only codes for three LRR regions and the transmembrane region. The TIR region was missing from all the predicted *Glossina TLR5* genes (Fig. 5.7).

This structure is almost identical to *M. domestica TLR5*, though the orthologue within *M. domestica* contains a large N-terminal noncoding region. This conservation is not seen in other dipteran genera though, both *D. melanogaster* and *S. calcitrans* show significant difference to other genera (Fig. 5.7). *Drosophila melanogaster TLR5* contains two exons and with the ectodomain containing four LRRs, an LRR-NT domain and a further LRR, prior to the transmembrane and TIR domains. The predicted *S. calcitrans* gene consists of eight exons and codes for a single LRR-NT, three LRRs and the transmembrane domain.

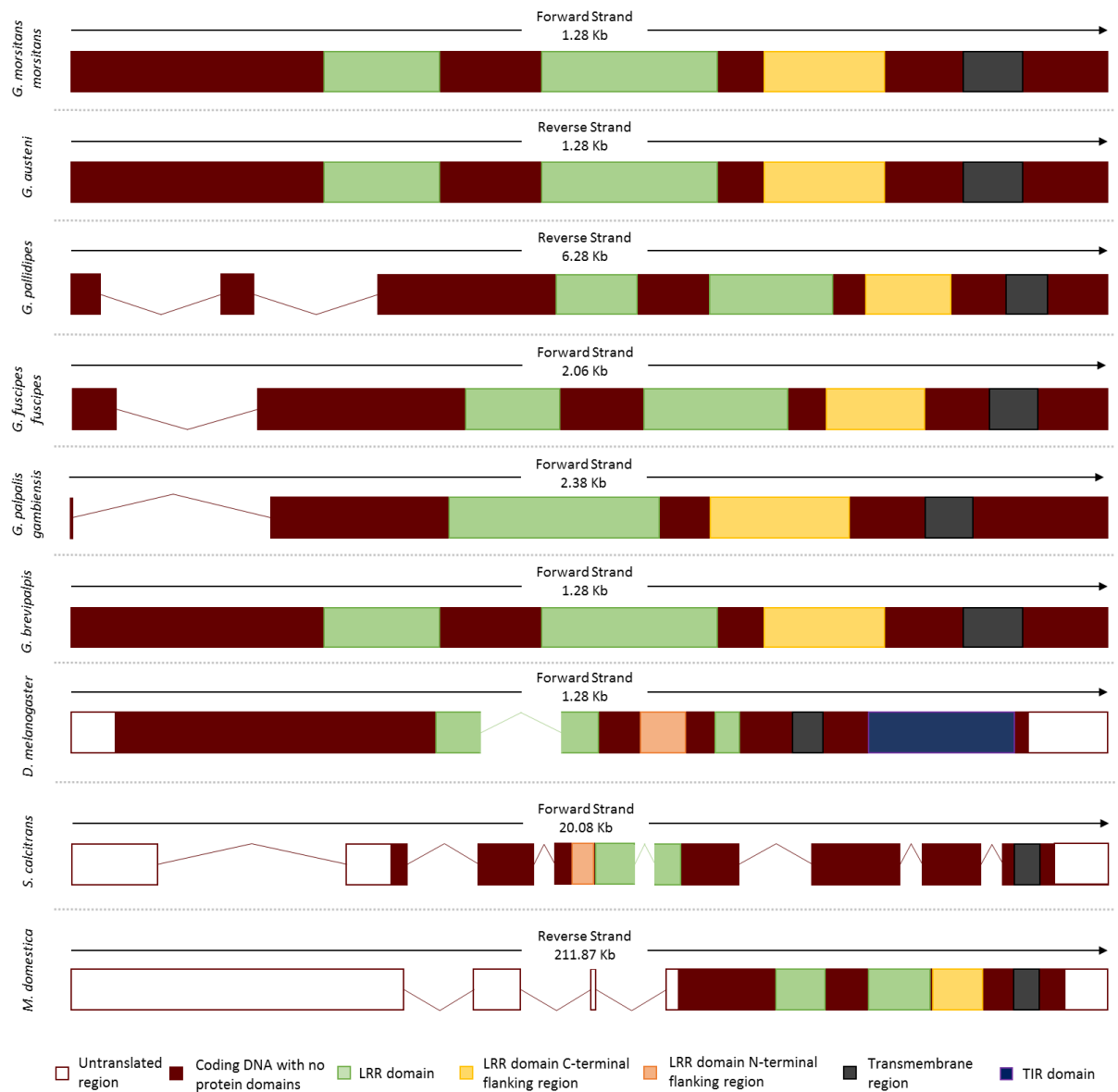


Figure 5.7: The gene alignment of predicted *TLR5* genes within the *Glossina* genome assemblies with reference genes of *D. melanogaster*, *S. calcitrans* and *M. domestica*. Drawn in Microsoft PowerPoint and adapted from figure produced from VectorBase and FlyBase (not to scale). Exons can be seen in the boxed areas, while introns are represented by linear areas. The coding strand and full transcript length are given above each gene. Encoded protein domains are denoted by colour within the coding regions of each gene. *Glossina* species used the transcripts of the genes seen in Table 5.3, while *M. domestica* and *S. calcitrans* used VectorBase genes MDOA007681 and SCAU005919 transcripts respectively, *D. melanogaster* used FlyBase transcript FBgn0026760.

### 5.3.1v: Toll-like receptor 6

The predicted *TLR6* genes illustrated two primary structures, either being coded by a single exon (*G. m. morsitans*, *G. pallidipes* and *G. palpalis gambiensis*) or containing a single C-terminal intron, present in *G. austeni*, *G. f. fuscipes* and *G. brevipalpis* (Fig. 5.8). Protein domains within these genes are highly conserved across the *Glossina* genus, 19 consecutive LRR domains precede a single LRR-NT and three more LRRs, the transmembrane and TIR domains follow these. Interestingly, some variation was observed within *G. f. fuscipes* which exhibited 20 rather than 19 LRRs, and *G. brevipalpis* which exhibited an additional LRR-CT domain (Fig. 5.8).

The single exon gene structure is also observed across the other dipteran genera, showing a highly conserved structure within dipteran TLR6 (Fig. 5.8). Though TLR6 genes identified within *Drosophila*, *Musca* and *Stomoxys* exhibit both N- and C-terminal coding regions absent from the predicted *Glossina* genes. Interestingly, *G. brevipalpis* is the only species to exhibit an LRR-CT in this gene.

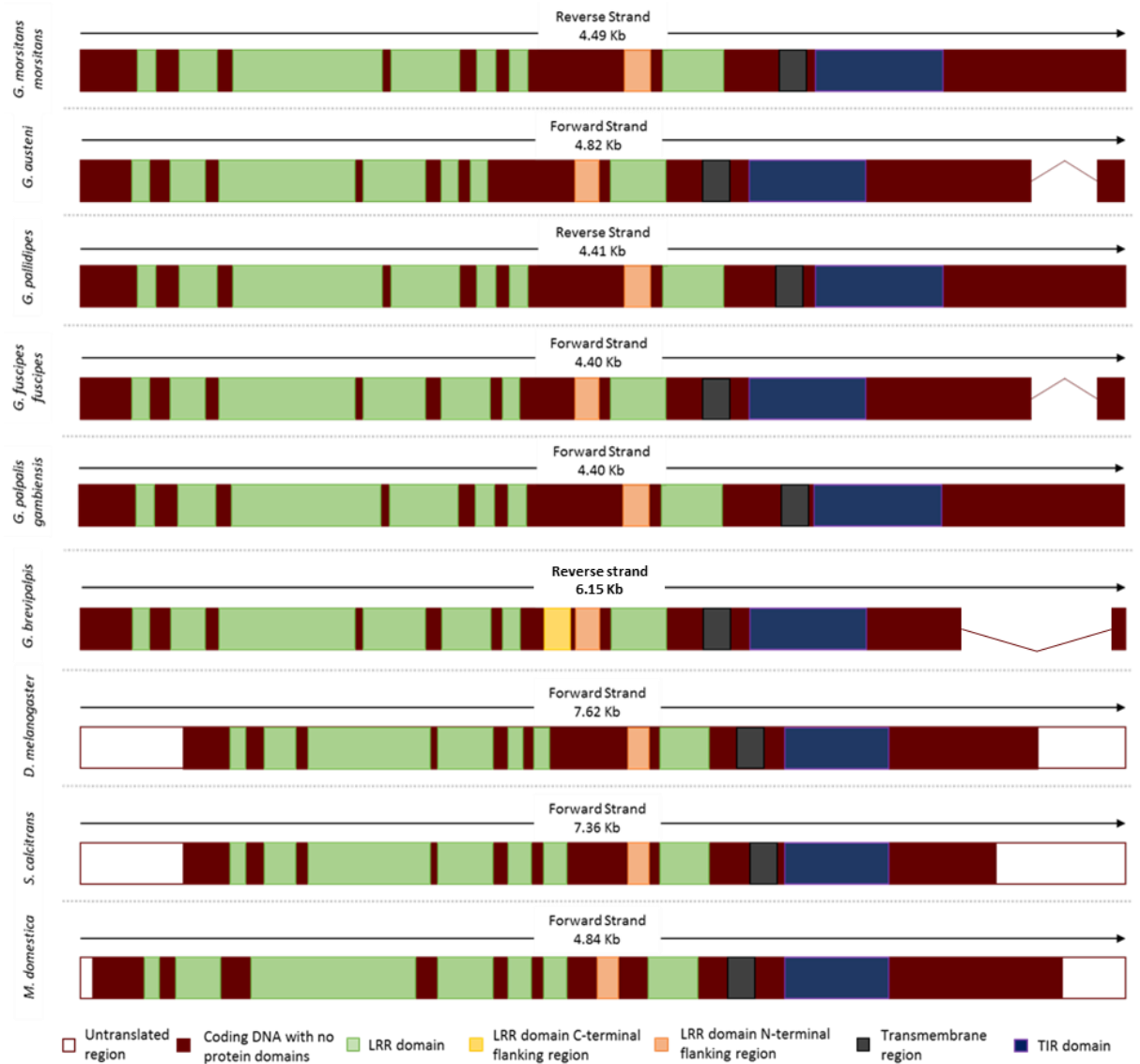


Figure 5.8: The gene alignment of predicted *TLR6* genes within the *Glossina* genome assemblies with reference genes of *D. melanogaster*, *S. calcitrans* and *M. domestica*. Drawn in Microsoft PowerPoint and adapted from figure produced from VectorBase and FlyBase (not to scale). Exons can be seen in the boxed areas, while introns are represented by linear areas. The coding strand and full transcript length are given above each gene. Encoded protein domains are denoted by colour within the coding regions of each gene. *Glossina* species used the transcripts of the genes seen in Table 5.3, while *M. domestica* and *S. calcitrans* used VectorBase genes MDOA006425 and SCAU006979 transcripts respectively, *D. melanogaster* used FlyBase transcript FBgn0036494.

### 5.3.1vi: Toll-like receptor 7

The structure of the predicted *Glossina TLR7* genes is almost identical across the dipteran genera with only *G. palpalis gambiensis* showing any variation. This structure consists of a single exon, between 4.41 and 4.5 Kb long. *Glossina palpalis gambiensis* shows slight variation with a single C-terminal intron being present within the gene. The protein coding domains within the predicted *TLR7* genes were found to be identical. Eighteen LRR regions precede an LRR-CT, an LRR-NT and four more LRRs before the transmembrane and TIR domains (Fig. 5.9).

This single exon structure observed in the predicted *Glossina TLR7* gene is identical to that of other Dipteran genera. However, while the protein domains mirror that of *S. calcitrans*, they differ slightly to that of *D. melanogaster* and *M. domestica*. This suggests some variation between haematophagic and other dipteran species. However, the primary difference between *Glossina TLR7* and other dipteran species appears to be the lack of untranslated regions within the *TLR7* genes (Fig. 5.9).



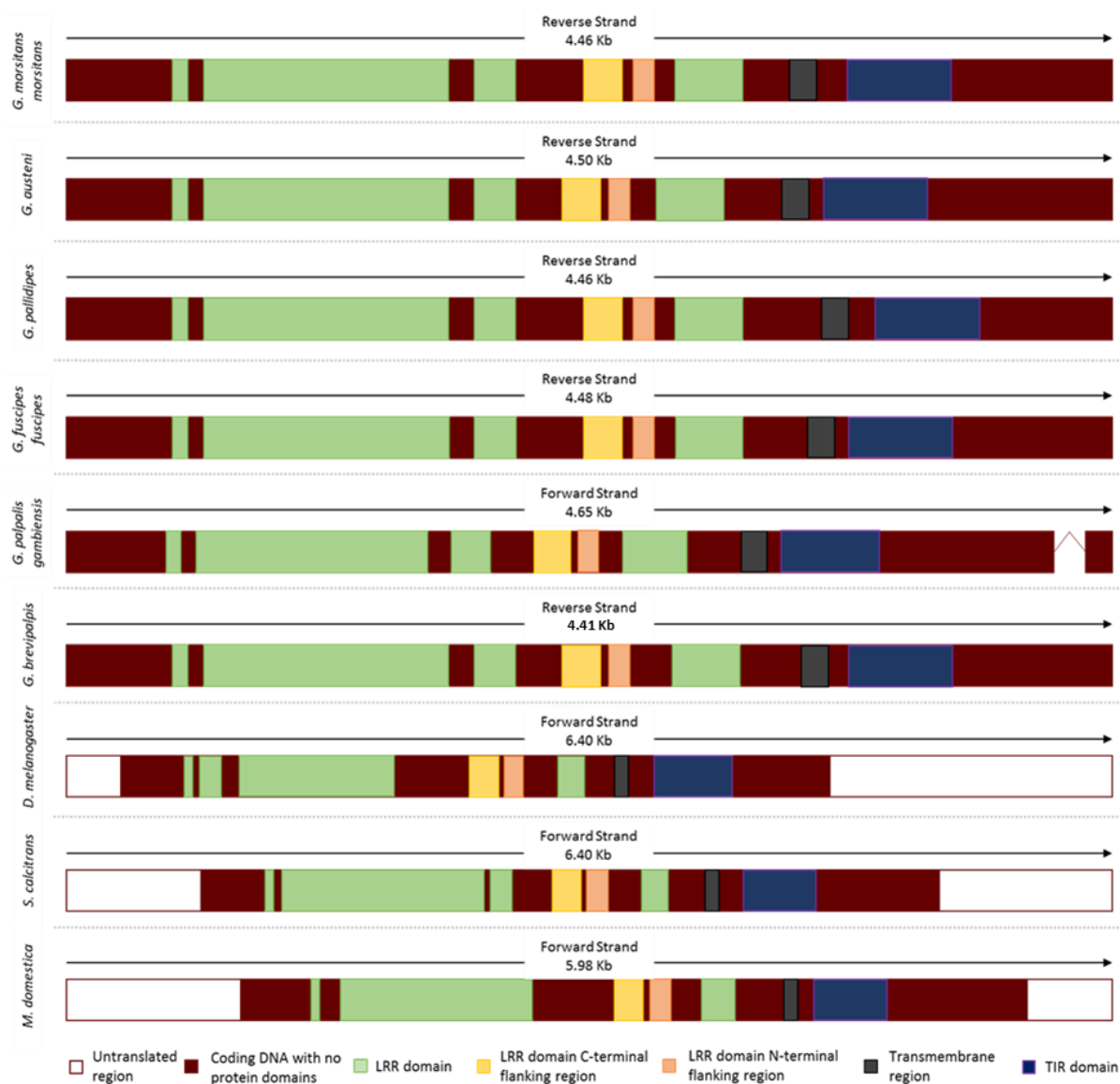


Figure 5.9: The gene alignment of predicted *TLR7* genes within the *Glossina* genome assemblies with reference genes of *D. melanogaster*, *S. calcitrans* and *M. domestica*. Drawn in Microsoft PowerPoint and adapted from figure produced from VectorBase and FlyBase (not to scale). Exons can be seen in the boxed areas, while introns are represented by linear areas. The coding strand and full transcript length are given above each gene. Encoded protein domains are denoted by colour within the coding regions of each gene. *Glossina* species used the transcripts of the genes seen in Table 5.3, while *M. domestica* and *S. calcitrans* used VectorBase genes MDOA015329 and SCAU008297 transcripts respectively, *D. melanogaster* used FlyBase transcript FBgn0034476.

### 5.3.1vii: Toll-like receptor 8

As with gene structure seen in section 5.3.2, the protein domains coded for within each of the predicated *TLR8* genes is identical. Twenty consecutive LRRs precede an LRR-CT and LRR-NT region. These are followed by a further five LRRs and a LRR-CT before the transmembrane and TIR region (Fig. 5.10). This structure differs from the TLR8 gene of other dipteran genera which either contain fewer LRRs (*S. calcitrans* and *M. domestica*) or are missing both LRR-CT regions (*D. melanogaster*). Despite this however, the genetic structure is similar featuring a single exon and a CDS of approximately 4.1Kb, though the genes from other dipteran genera all exhibit noncoding regions at the N and C-terminals of the gene (Fig.5.10).

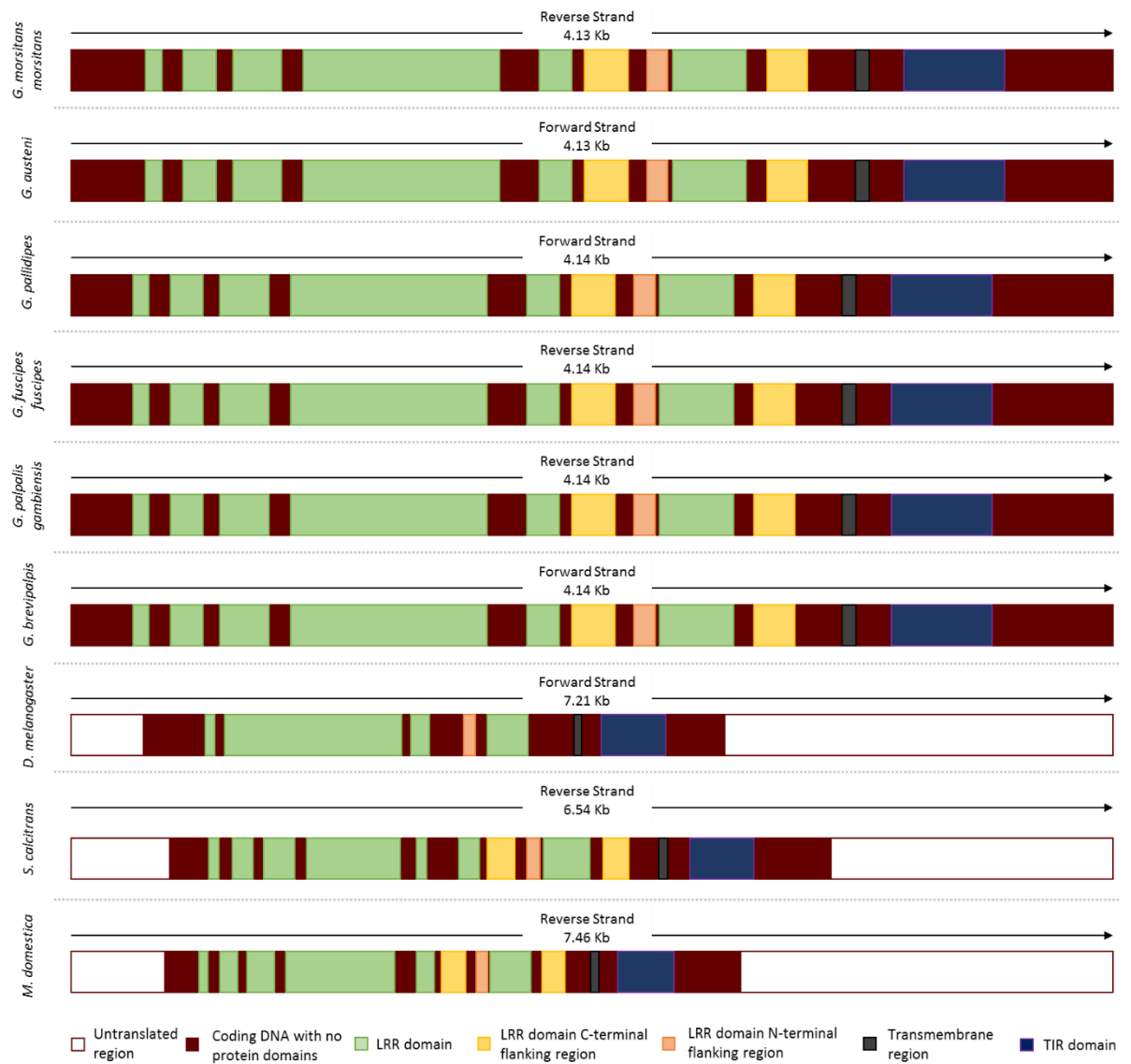


Figure 5.10: The gene alignment of predicted *TLR8* genes within the *Glossina* genome assemblies with reference genes of *D. melanogaster*, *S. calcitrans* and *M. domestica*. Drawn in Microsoft PowerPoint and adapted from figure produced from VectorBase and FlyBase (not to scale). Exons can be seen in the boxed areas, while introns are represented by linear areas. The coding strand and full transcript length are given above each gene. Encoded protein domains are denoted by colour within the coding regions of each gene. *Glossina* species used the transcripts of the genes seen in Table 5.3, while *M. domestica* and *S. calcitrans* used VectorBase genes MDOA005627 and SCAU003362 transcripts respectively, *D. melanogaster* used FlyBase transcript FBgn0029114.

### 5.3.1viii: Toll-like receptor 9

The structure of the predicted *TLR9* genes appears to be conserved across all species with five exons across the transcript. Notably two introns are seen within the TIR domain, a feature that appears to be characteristic of dipteran *TLR9* genes. The predicted *TLR9* genes contain a fewer number of LRRs than expected with just five LRR regions (three in the *G. austeni* orthologue) preceding the transmembrane region and the TIR. While this number seems small compared to other predicted *TLR* genes, this is consistent with other dipteran species although the distribution of the LRRs varies in other species. The presence of a large noncoding region at the C-terminal, present in three of the *Glossina* species, appears to be unique to the *Glossina* genus. However, this is not present in all species with *G. pallidipes* and *G. palpalis gambiensis* following a very similar genetic structure to that observed in *M. domestica* (Fig. 5.11).

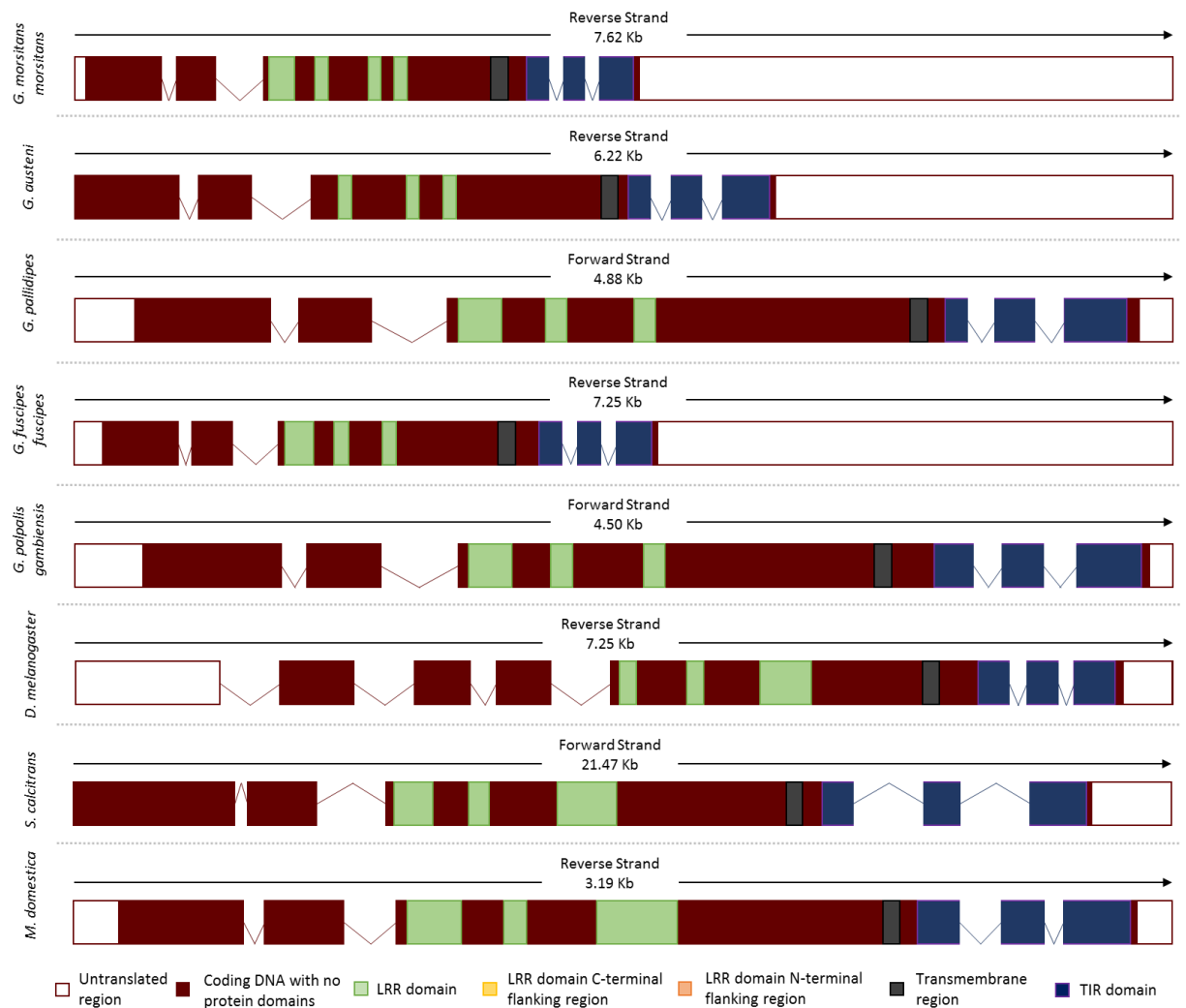


Figure 5.11: The gene alignment of predicted *TLR9* genes within the *Glossina* genome assemblies with reference genes of *D. melanogaster*, *S. calcitrans* and *M. domestica*. Drawn in Microsoft PowerPoint and adapted from figure produced from VectorBase and FlyBase (not to scale). Exons can be seen in the boxed areas, while introns are represented by linear areas. The coding strand and full transcript length are given above each gene. Encoded protein domains are denoted by colour within the coding regions of each gene. *Glossina* species used the transcripts of the genes seen in Table 5.3, while *M. domestica* and *S. calcitrans* used VectorBase genes MDOA002537 and SCAU004205 transcripts respectively, *D. melanogaster* used FlyBase transcript FBgn0036978.

### 5.3.1ix: Toll-like receptor 13

As noted in section 5.3.2, the structure of predicted *TLR13* genes appears to be specific to the *Glossina*. All species feature two conserved introns within the gene, the first starting at nucleotide 605 near the N-terminal, and one within the LRR domain sequences at nucleotide 1,491. However, members of the Palpalis group feature an additional N-terminal intron at the tenth nucleotide, while Fusca group species contain a further intron prior to that seen in the Palpalis group (Fig. 5.12).

The protein domains within each gene are identical with a total of nine LRRs being observed, these are split into a group of four and a group of five with the latter being split across the C-terminal intron. Neither the transmembrane nor TIR region was identified in any of the predicted *TLR13* genes. This structure is similar to that observed within *S. calcitrans* though differs to that of *M. domestica*, which consists of eight LRRs, split into three groups rather than two (Fig. 5.12).

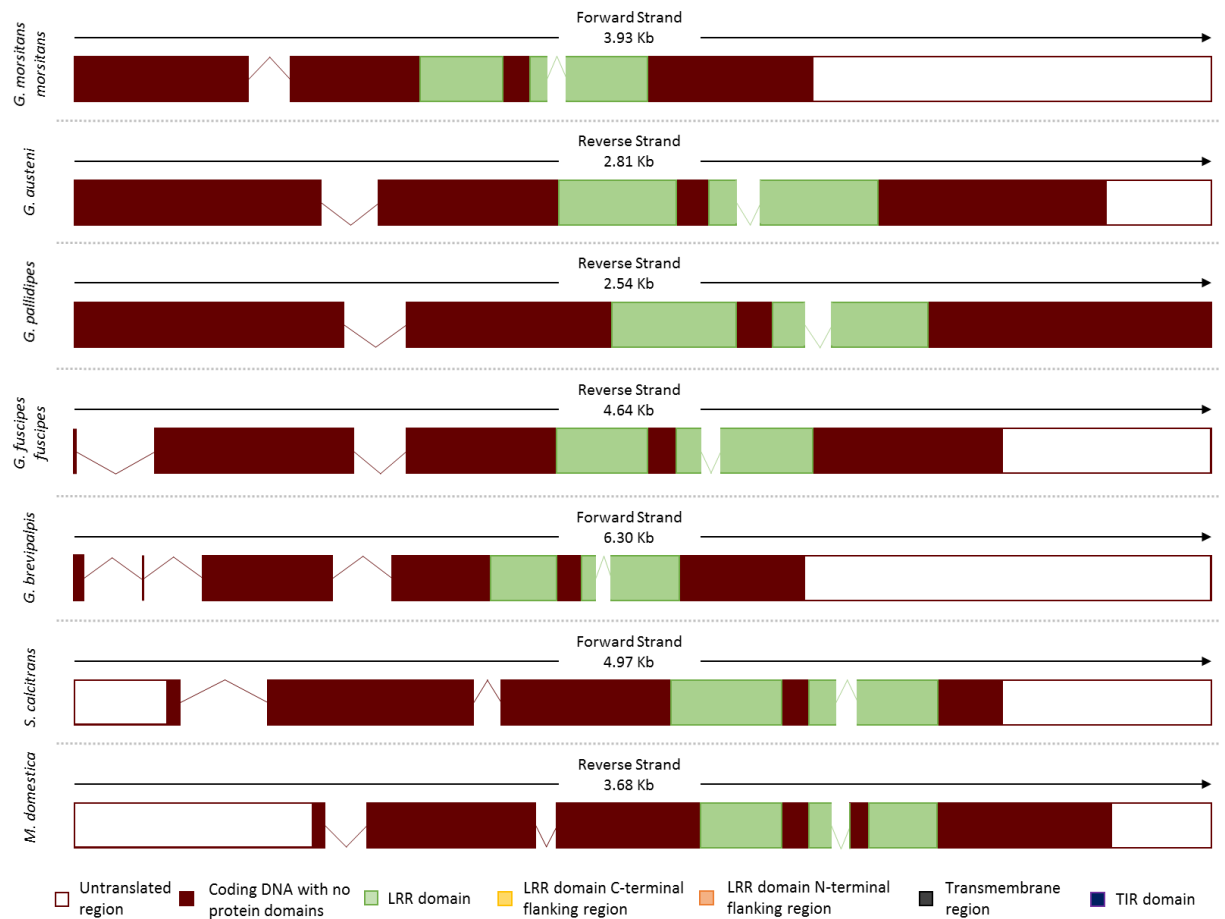


Figure 5.12: The gene alignment of predicted *TLR13* genes within the *Glossina* genome assemblies with reference genes of *D. melanogaster*, *S. calcitrans* and *M. domestica*. Drawn in Microsoft PowerPoint and adapted from figure produced from VectorBase and FlyBase (not to scale). Exons can be seen in the boxed areas, while introns are represented by linear areas. The coding strand and full transcript length are given above each gene. Encoded protein domains are denoted by colour within the coding regions of each gene. *Glossina* species used the transcripts of the genes seen in Table 5.3, while *M. domestica* and *S. calcitrans* used VectorBase genes MDOA002744 and SCAU006576 transcripts respectively.

### 5.3.2: Phylogenetic analysis

Phylogenetic analysis illustrated that the predicted TLR genes fall into five primary clades, following the established TLR evolutionary history (Christophides *et al.*, 2002; Levin and Malik, 2017). The TLR2\_7 subclade, along with TLRs 6, 8, the mosquito TLRs 10, and 11, occupy the largest clade, while complete dipteran TLRs 1 and 3-5, including the mosquito (Culicidae) TLR1/5 subclade form the second largest clade (Fig. 5.13). The *TLR9* clade exhibited the earliest divergence from the other TLR genes observed in previous literature (Fig. 5.13). The three partial genes form two separate clades that diverge from the primary topology: with the *Glossina* TLR5 genes forming a distinct clade divergent from the TLR3 and TLR13 clade (Fig. 5.13).

The Maximum Likelihood and Neighbour-Joining methodologies produced similar topologies. The Maximum Likelihood method exhibited seven clades containing *Glossina* TLR genes that followed the established TLR evolutionary history (Fig. 5.3 and 5.4A). The TLR2\_7 subclade (Clade I), TLR6 and TLR8 (Clades II and III), formed the largest clade within the tree which also contained the mosquito TLR10 and TLR11 genes. This clade is supported by strong bootstrap values (>75) at all but one of the primary nodes. Within each TLR gene the *Glossina* species follow the established evolutionary pattern of the genus forming three clades for each of the Morsitans, Palpalis and Fusca groups (Dyer *et al.*, 2008). The second largest clade contains/ed the predicted *Glossina* and other dipteran species TLR1 gene, as well as complete *TLR3* and *TLR5* genes and dipteran *TLR4* genes. The primary TLR1 clade (Clade IV) contains all Brachycera families and shows clear evidence of the *Glossina* evolutionary history within the clade. The Culicidae TLR1/5 subclade, and *Drosophila* TLRs 3, 4 and 5 subclades show clear divergence from the other TLR genes within this clade. Bootstrap values in this clade were generally strong (>75), though some nodes were poorly supported (<75). The last of the full predicted genes, *TLR9*, formed a distinct divergent clade (Clade V) from the other TLR genes as observed in previous TLR evolutionary histories (Ref). This clade illustrates the expected evolutionary history of dipteran genera and is supported by strong bootstrap values at most primary nodes (Fig. 5.13A).

The three partial TLR genes identified (TLRs 3, 5 and 13), formed two distinctly separate clades divergent from the main topology of the phylogenetic tree (Fig. 5.13A). *Glossina* and *Musca* partial *TLR5* genes form a separate clade, while *S. calcitrans* *TLR5* exhibits a clear



evolutionary deviation. Interestingly, the speciation of the *Glossina* genus shows some variation from the established evolutionary history, with the Morsitans group comprising two clades with *G. m. morsitans* and *G. pallidipes* forming a divergent subclade from *G. austeni*, though divergence within the *Glossina* subclade is not well supported, with bootstrap values <50 (Clade VI, Fig. 5.13A).

The Neighbour-joining method exhibited two major differences to the maximum-likelihood, firstly the presence of a combined TLR6/8 subclade rather than two distinct subclades. Secondly, the TLR9 shows a much closer relationship to the *TLR1* in the neighbour-joining method (Fig. 5.13B). The largest clade is separated into two clades, with *TLR1* and 9 forming a subclade (Subclades III and IV, Fig. 5.13B) apart from TLRs 2, 6-8. Structurally the TLR2\_7 subclade and TLRs 6 and 8 (Subclades I and II respectively, Fig. 5.13B) are almost identical, with all *Glossina* species forming a separate subclade from the other dipterans. However, there is some variation within the clades regarding the position of each *Glossina* species, with some species from the Morsitans and Palpalis groups grouping together, it should be noted however, that not all nodes are strongly supported with bootstrap values ranging between 48 and 87.

Toll-like receptor 9 (Subclade III, Fig. 5.13B) was predicted in five *Glossina* species though no orthologue was identified within *G. brevipalpis*. Figure 5.4 indicates that TLR9 follows the basic species evolution observed in other TLR genes. Predicted *Glossina* genes form two subclades with the Palpalis and *Morsitans* group species separated with strong nodal support (100), while other dipteran species show the expected divergence observed previously. The expected TLR1/5 subclade is separated into two further subclades with a separate *TLR1* (Subclade IV) grouping with the *Culicidae* TLR1/5 subclade (Fig. 5.13B) as observed by Christophides *et al.*, (2002). Predicted *Glossina* *TLR5* genes form a separate clade (Clade V, Fig. 5.4B)) with the previously identified *M. domestica* gene, though identified *D. melanogaster* and *S. calcitrans* TLR5 genes can be observed within different clades. Both the *TLR1* and *TLR5* clades (Clades IV and V, Fig. 5.13B)) support the *Glossina* species established by Dyer *et al.* (2008), though it is not supported by strong bootstrap values in TLR5 (<90).

As in the maximum likelihood tree the remaining two partially predicted genes, *TLR3* and 13, form a subclade (Fig. 5.13B). Phylogenetic analysis of TLR3 illustrates that while the

*Glossina*, *Musca* and *Stomoxys* TLR genes form a clade, *D. melanogaster* TLR3 forms a separate sub-clade with *D. melanogaster* and *A. aegypti* TLR4. Toll-like receptor 13 mirrored the expected speciation of the *Glossina* genus, though the gene was not identified within *G. palpalis gambiensis*. A clear division is visible between the Morsitans and Palpalis group clades, supported by strong bootstrap values (100). While the Fusca group forms a sister clade, again supported bootstrap values of 100 (Fig. 5.13B).

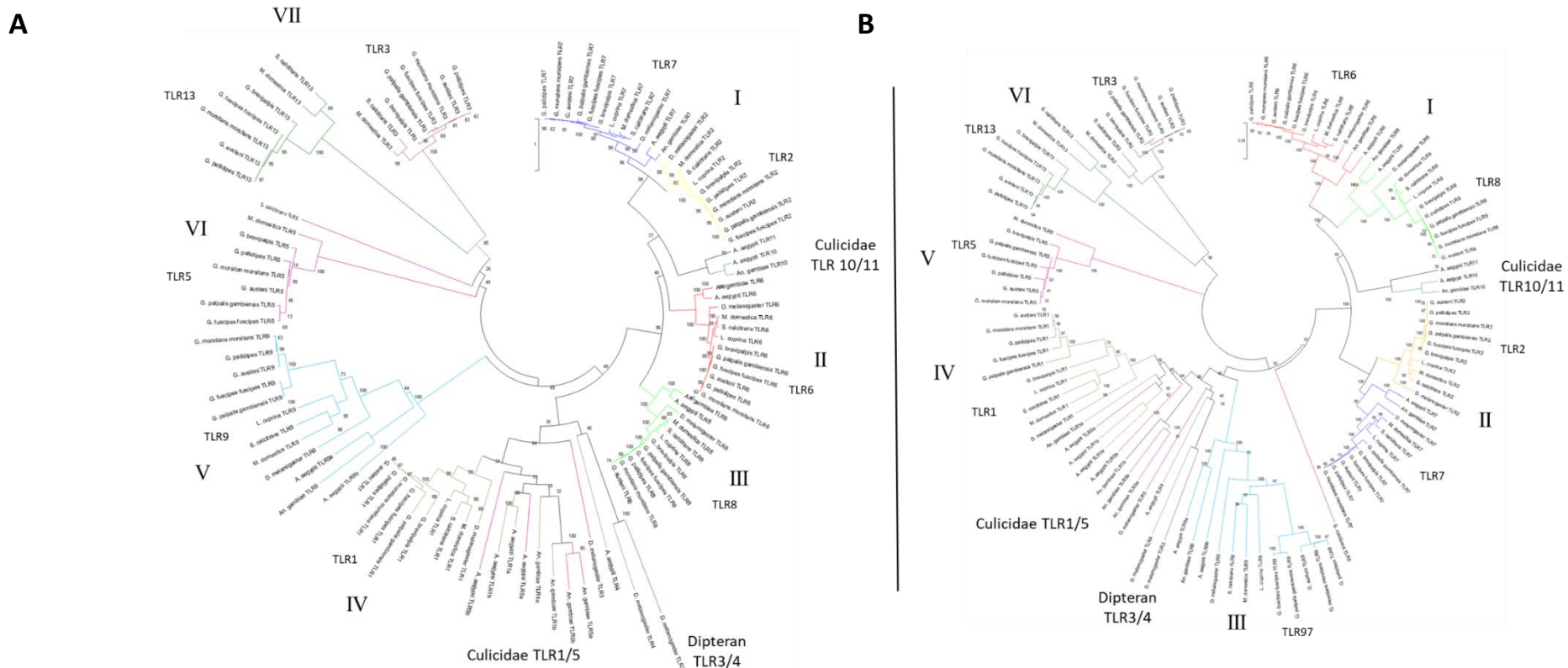


Figure 5.13: Evolutionary analyses were conducted in MEGAX (Kumar *et al.*, 2018), using both the Maximum Likelihood (A) and the Neighbour-Joining (B) methods (Saitou and Nei, 1987) and a bootstrap test with 1000 replicates (Felsenstein, 1985). A) The Le-Gascuel 2008 model (Le and Gascuel, 2008) and discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 1.3828)). B) The evolutionary distances were computed using the Poisson correction method (Zuckerkand and Pauling (1965) and are in the units of the number of amino acid substitutions per site. Both trees are drawn to scale, with branch lengths measured in the number of substitutions per site. All positions with less than 50% site coverage were eliminated. There were a total of 1153 positions in the final dataset. Each clade contains all identified *Glossina* genes, with the available gene sequences for *M. domestica*, *S. calcitrans*, *L. cuprina*, *D. melanogaster*, *A. aegypti* and *An. gambiae* included as outgroups for each gene. Each TLR gene is represented by colour; *TLR1* = gold, *TLR2* = yellow, *TLR3* = brown, *TLR4* = silver, *TLR5* = pink, *TLR6* = red, *TLR7* = blue, *TLR8* = light green, *TLR9* = light blue, *TLR10* = turquoise, *TLR11* = black and *TLR13* = green.

### 5.3.3: Interspecies variation

#### 5.3.3i: Inter-gene diversity

When variation was assessed across the full length of predicted TLR genes nucleotide diversity was found to be low ( $\pi < 0.08$ ) in all genes, though dN/dS was found to be higher in TLRs 1 and 13, and considerably higher in *TLR9* than other predicted genes (Table 5.2). Nucleotide variation in the LRR receptor domain was found to be higher than the TIR domain, though *TLR1* exhibited greater variation within the TIR domain than the LRR (Table 5.2 and Fig. 5.14). Variation between *Glossina* species within TLRs 2, 6, 7 and 8 was found to be consistent across the gene with only a minimal increase in variation within the LRR domain compared to the TIR (Table 5.2 and Fig. 5.14). In TLRs 2, 7 and 8 all the major points of nucleotide variation were found outside of the primary protein domains, while TLR6 did exhibit a large peak of nucleotide variation within the LRR domain (Fig. 5.14). Interestingly, the increased variation observed within the TIR domain of TLR1 appears to be the result of a large point of variation around nucleotide 2985 (Fig. 5.14). The increased nucleotide variation within the TLR9 observed in table 5.4, was present across the full gene, though the two most diverse area of the gene are outside of the LRR and TIR domains (Fig. 5.14). Furthermore, the dN/dS values within the TLR9 gene are considerably larger than those observed in other TLR genes (Fig. 5.14).

Table 5.2: Genetic variation across the *Glossina* within each TLR gene. The number of orthologues compared, the number of segregating sites between them, the nucleotide diversity ( $\pi$ ) and the average number of nucleotide difference (K) between the orthologues is given. dN/dS values were calculated to give an indication of synonymous and non-synonymous across orthologues. \* = The predicted *TLR1* from *G. palpalis gambiensis* was removed due to the difference in CDS length and missing TIR.

TLR gene	Number of <i>Glossina</i> orthologues	Segregating sites	Nucleotide Diversity ( $\pi$ )	Average number of Nucleotide Difference (K)	dN	dS	dN/dS
Full gene							
<i>TLR1</i>	5*	285	0.05134	151.5	0.0329	0.2736	0.1202
<i>TLR2</i>	6	567	0.05557	233.4	0.00433	0.27189	0.0159
<i>TLR3</i>	6	556	0.07043	232	0.02255	0.29568	0.0763
<i>TLR5</i>	6	152	0.06763	60.667	0.01234	0.35659	0.0346
<i>TLR6</i>	6	654	0.06465	269.4	0.00945	0.31034	0.0305
<i>TLR7</i>	6	498	0.04862	198.667	0.00746	0.21278	0.0351
<i>TLR8</i>	6	389	0.03739	154.467	0.00399	0.17023	0.0234
<i>TLR9</i>	5	266	0.05852	141.5	0.04693	0.11267	0.4165
<i>TLR13</i>	5	403	0.07707	177.8	0.03126	0.30547	0.1023
TIR domain							
<i>TLR1</i>	5*	62	0.08617	26.8	0.03557	0.29894	0.119
<i>TLR2</i>	6	51	0.05024	21	0	0.25201	0
<i>TLR3</i>	N/A	N/A	N/A	N/A	N/A	N/A	N/A
<i>TLR5</i>	N/A	N/A	N/A	N/A	N/A	N/A	N/A
<i>TLR6</i>	6	54	0.05534	22.467	0.00274	0.29233	0.0094
<i>TLR7</i>	6	62	0.05805	24.267	0.00257	0.26269	0.0098
<i>TLR8</i>	6	50	0.04935	20.333	0.00189	0.24559	0.0077
<i>TLR9</i>	5	26	0.03034	13.5	0.01223	0.1092	0.112
<i>TLR13</i>	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 5.2: Genetic variation across the *Glossina* within each TLR gene. The number of orthologues compared, the number of segregating sites between them, the nucleotide diversity ( $\pi$ ) and the average number of nucleotide difference (K) between the orthologues is given. dN/dS values were calculated to give an indication of synonymous and non-synonymous across orthologous. \* = The predicted *TLR1* from *G. palpalis gambiensis* was removed due to the difference in CDS length and missing TIR.

TLR gene	Number of <i>Glossina</i> orthologues	Segregating sites	Nucleotide Diversity ( $\pi$ )	Average number of Nucleotide Difference (K)	dN	dS	dN/dS
LRR region							
<i>TLR1</i>	5*	186	0.08186	85.3	0.02878	0.32365	0.0889
<i>TLR2</i>	6	362	0.0614	149.267	0.00459	0.31721	0.0143
<i>TLR3</i>	6	401	0.07046	167.267	0.01775	0.32711	0.0543
<i>TLR5</i>	6	110	0.07238	44.733	0.0119	0.41817	0.0285
<i>TLR6</i>	6	403	0.06779	168.267	0.00765	0.34616	0.0221
<i>TLR7</i>	6	320	0.05389	130.2	0.00481	0.2685	0.0179
<i>TLR8</i>	6	238	0.03646	95.733	0.00352	0.16693	0.0211
<i>TLR9</i>	5	50	0.04281	26.2	0.02389	0.11955	0.1998
<i>TLR13</i>	5	108	0.06616	47.7	0.02334	0.26933	0.0867

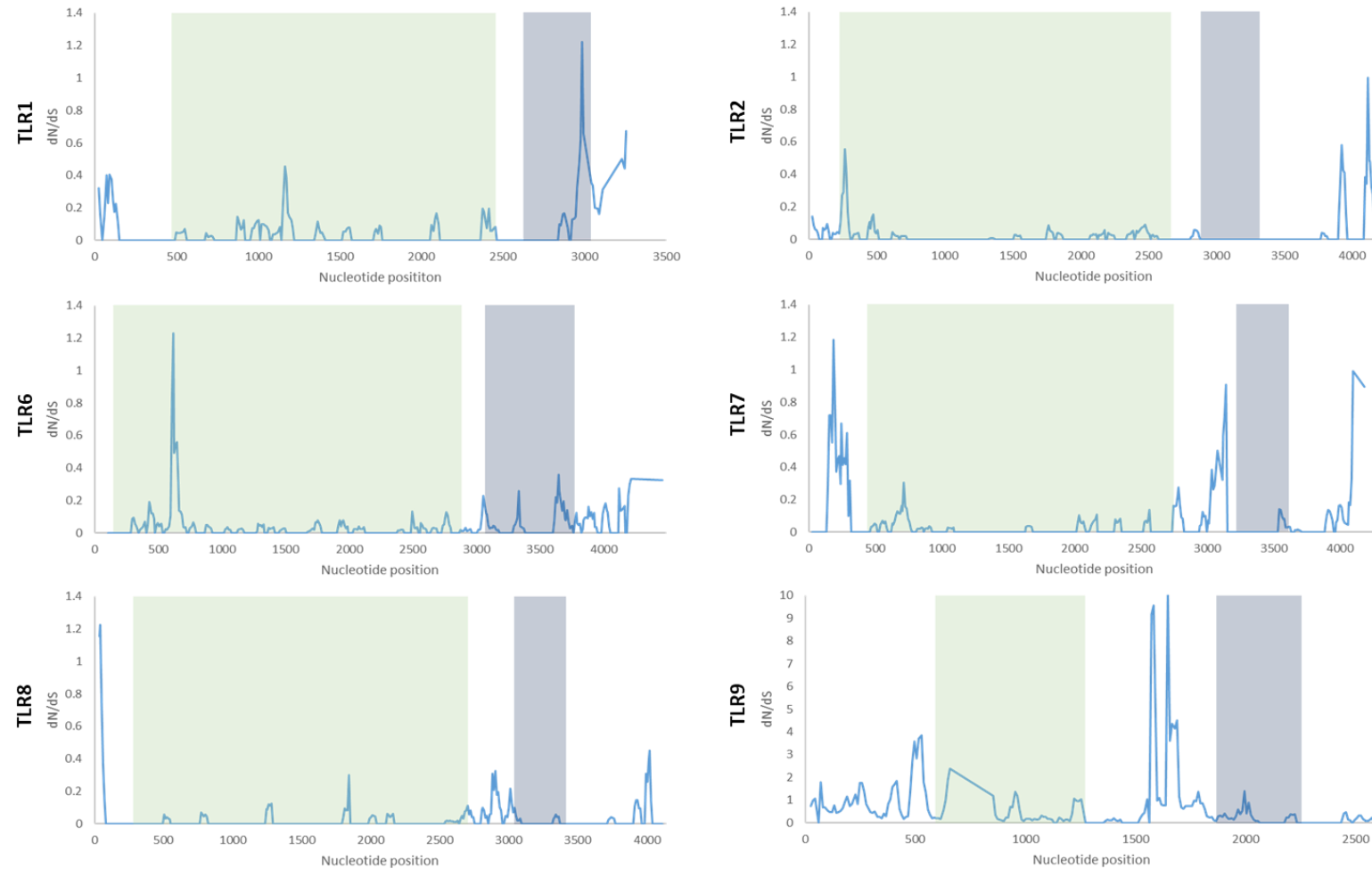


Figure 5.14: Sliding window analysis showing dN/dS variation across the predicted TLR genes within the *Glossina* spp. Sliding window analysis was run in DnaSP (version 6), using all complete predicted TLR genes. The window size = 50 while step size = 10. The green shaded area indicates the region there LRR domains were coded, while the blue area shows the TIR domain.

### 5.3.3ii: Pairwise distance

Principle component analysis (PCA) offers a different approach to the estimation of the evolutionary relationships between each predicated gene above by compare the pairwise distance ( $P$ ) of amino acid substitutions between sequences.

This analysis produced a PCA plot (Fig. 5.15) showing a similar clustering of sequences to that observed within the phylogenetic analysis in section 5.3.1. Notably, all predicated genes formed clusters within the established gene families. As with the phylogenetic analysis (Fig. 5.14), specific clusters can be observed between TLR2 and 7 and TLR6 and 8, both forming clusters around (-1.4,-0.05) and (-1.3,0.1), respectively (Fig. 5.15). As in the phylogenetic analysis the remaining genes form distinct gene clusters. Toll-like receptor 1 shows some variation within the cluster though the majority of *TLR1* genes clustering around (0.55, -0.2). Predicted TLR9 genes show as similar divergence from TLR1 within the PCA, forming a cluster at (0.3,0.9) signifying the subclade observed in figure 5.14. Toll-like receptor 5 formed a separate cluster apart from all over predicted TLR genes at (1.1, 1.45), this divergence is in agreement with the observations from the phylogenetic analysis. Finally, PCA shows a much greater degree of divergence between the *TLR3* and *TLR13* families than was implied by the neighbour-joining phylogenetic analysis, though is similar to divergence observed within the Maximum Likelihood tree (Fig.14A).



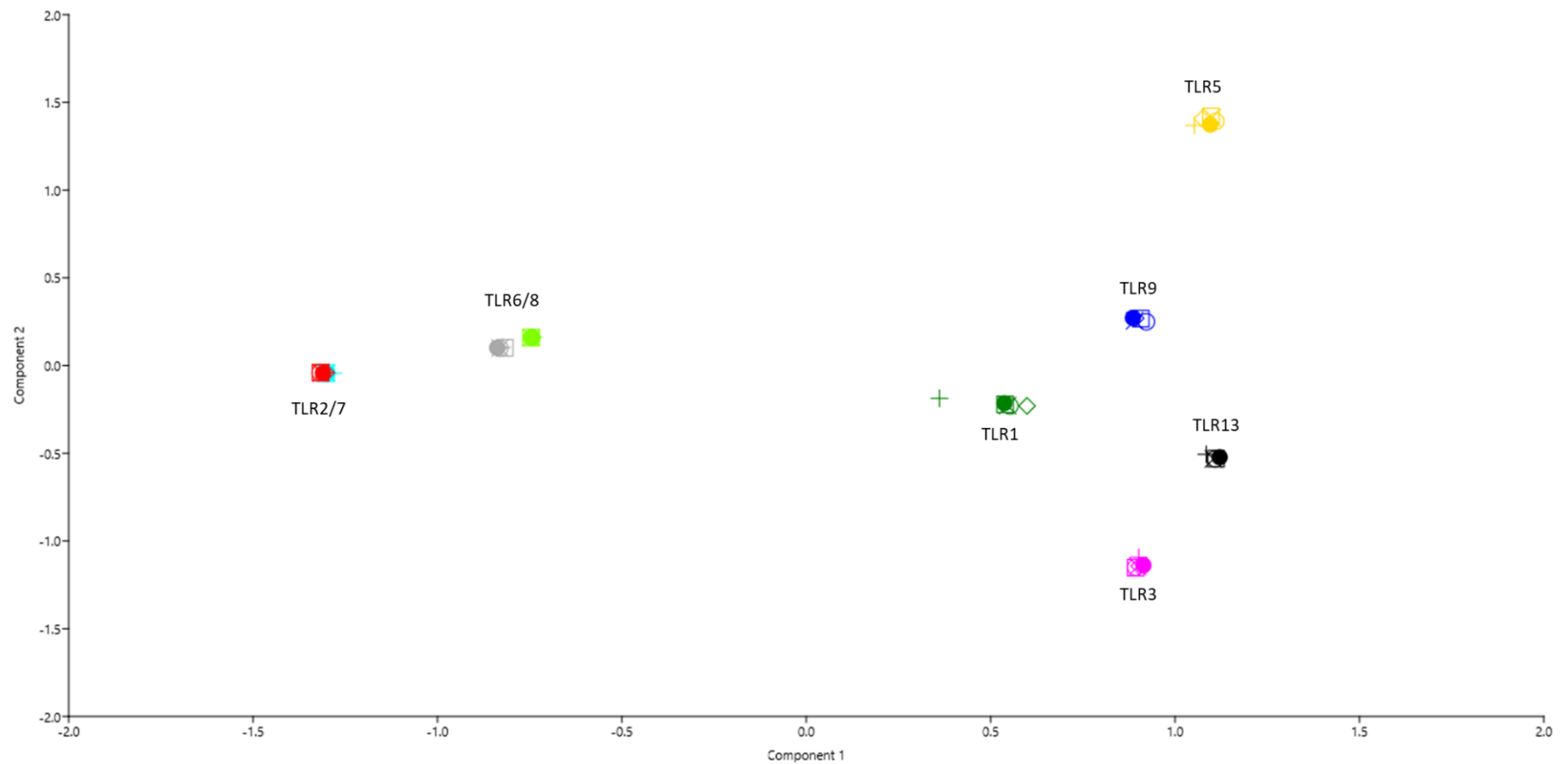


Figure 5.15: Principle component analysis (PCA) plot of all predicted TLR gene within the *Glossina* genus using the first and second Principle components (Eigenvalues: PC1 = 1.01337 (35.943 % variance); PC2 = 0.42556 (13.834 % variance). Pairwise distance estimations were conducted in MEGA7 (Kumar et al., 2016), using pairwise distance. All sites with less than 50 % coverage were eliminated. A distance matrix was produced in Microsoft Excel and PCA analysis was conducted in PAST3 (Hammer et al., 2001). Individual predicted TLR genes are show by plots, species is denoted by shape: *G. austeni* = X; *G. brevipalpis* = +; *G. f. fuscipes* = ●; *G. m. morsitans* = □; *G. pallidipes* = ○; *G. palpalis gambiensis* = ◇. While gene families are denoted by colour. *TLR1* = Green, *TLR2* = Aqua, *TLR3* = Pink, *TLR5* = Gold, *TLR6* = Grey, *TLR7* = Red, *TLR8* = Light Green, *TLR9*= Blue and *TLR13* = Black.

#### 5.3.4: 3-dimensional protein structure analysis

Protein modelling was performed on all predicted TLR proteins synthesised by the predicted genes. This enabled a comparison of protein structures between species and also enabled further analysis of the relationship between genes within the TLR families at a protein level.

Of the genes characterised above, six were observed to contain all three characteristic protein subdomains and showed a relatively high degree of similarity regarding the protein domains coded within the gene families. Modelling of the 3-dimensional protein structure further supported the observations made above, with all genes shown to produce a full TLR protein (Fig 5.16).

The extracellular domain illustrates the characteristic horseshoe shape, with  $\beta$ -sheets on the concave surface and helices/coils forming the convex outer surface. The number of helices varies between proteins, with the majority of the convex structure comprising coiled structures rather than helices. The extended helix forming the transmembrane region was clearly visible following ectodomain prior to the TIR domain. The TIR region of each protein contained the expected  $\beta$ -sheets surrounded by helices (Fig. 5.16). The overall structure of each protein family appears to be conserved across the *Glossina*, while there is some variation in additional coiled structures branching off the extracellular domain in several protein families, this do not appear to affect the over structure of the protein.

Variation within the TLR2, 6 and 8 protein models appeared to be minimal, with all species following a similar structure (Fig. 5.16). However, there is a greater degree of variation within the TLR 1, 7 and 9 models. The predicted protein structure of GPPI000821 (*G. palpalis gambiensis* TLR1), showed a high similarity to the extracellular domains of the TLR1 proteins despite a much shorter CDS, and as expected, did not code for the full extracellular region or the transmembrane and TIR regions. Interestingly, *G. brevipalpis* was observed to diverge from other TLR1 proteins with a slightly twisted structure visible in figure 5.16. Predicted protein structures within the TLR7 family show a conserved extracellular domain though there appears to be a fair greater amount of random coil structures prior to the transmembrane region and TIR domains. Finally, TLR9 shows a variation of the position of the transmembrane region within the *G. palpalis gambiensis* protein model, though the

extracellular and TIR domains shows high similarity and alignment to the other *Glossina* proteins. (Fig. 5.16).

Unlike the previous predicted TLR genes, both TLR3 and TLR13 genes were found to be missing both the transmembrane and the TIR domains. This again was apparent in the 3-dimensional protein model (Fig. 5.17). All predicted TLR3 genes were found to code for 25 LRR domains while TLR13 contained LRR domains, comprising the extracellular domain of the TLR protein, this is reinforced by the protein model which illustrated a clear horseshoe shape seen in all other TLR proteins (Fig. 5.17). Unlike other predicted TLR proteins, TLR3 shows a clear presence of several helices on the convex surface of the extracellular domain, rather than the coiled structures observed in other proteins. While there is little variation between protein models, some branching strands can be observed, specifically in the structures from the Palpalis group species. However, these appear to have little to no direct effect upon the overall structure of the protein. As with TLR3 and TLR13, predicted TLR5 genes did not code for a full TLR protein. The genetic structure observed in section 5.3.3iv illustrated the presence of several LRR domains and the transmembrane region of the protein was encoded, which was also observed within the predicted protein models (Fig. 5.16). The evidence of the extracellular domain is clearly visible, with eight parallel  $\beta$ -sheets forming the concave surface and four helices being present of the convex surface though is considerably shorter than those observed in other TLR proteins. The transmembrane helix is also present and highly conserved across all *Glossina* species (Fig. 5.17).

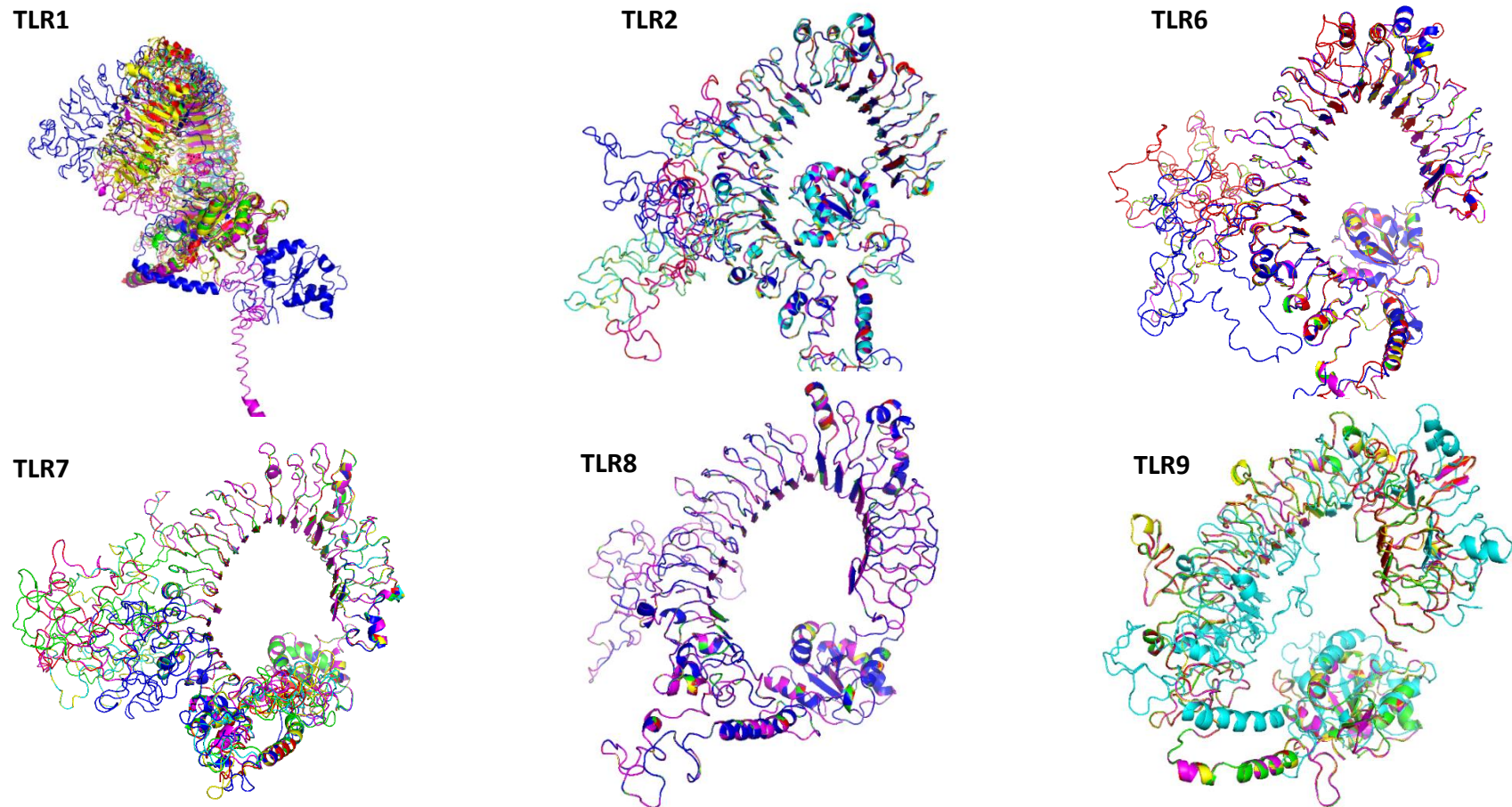


Figure 5.16: Structural alignment of complete predicted *Glossina* TLR proteins. Proteins structures were produced using SWISS-MODEL (Guex et al., 2009; Waterhouse et al., 2018) and models visualised and aligned in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre). Extracellular domains can be seen by the characteristic horseshoe shape, with consecutive  $\beta$ -sheets on the concave surface and helices/coils forming the convex outer surface, the helical transmembrane region can be observed at the base of each model, with the TIR domains. Model colour denotes the species of the predicted protein structure Red = *G. pallidipes*; Green = *G. austeni*; Purple = *G. m. morsitans*; Yellow = *G. f. fuscipes*; Cyan = *G. palpalis gambiensis*; Blue = *G. brevipalpis*.



Figure 5.17: Structural alignment of partial predicted *Glossina* TLR proteins. Proteins structures were produced using SWISS-MODEL (Guex et al., 2009; Waterhouse et al., 2018) and models visualised and aligned in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre). Extracellular domains can be seen by the characteristic horseshoe shape, with consecutive  $\beta$ -sheets on the concave surface and helices/coils forming the convex outer surface, the helical transmembrane region can be observed in TLR5. Model colour denotes the species of the predicted protein structure: Red = *G. pallidipes*; Green = *G. austeni*; Purple = *G. m. morsitans*; Yellow = *G. f. fuscipes*; Cyan = *G. palpalis gambiensis*; Blue = *G. brevipalpis*.

#### 5.3.4i: 5.3.4i: 3-dimensional structure comparison

While visual comparison of the predicted protein structures illustrated several variations both between and within TLR families, structural comparison was conducted to provide further insight into the relationship of protein structures.

Previous genotypic comparisons had grouped the predicted TLR genes into five main clades (Fig. 5.14) or groups (Fig. 5.15), structural comparison illustrated similar groupings. The predicted protein structures can be separated into two major groupings, those exhibiting a complete TLR protein and those exhibiting a partial protein. This variation was clearly visible when all structures were compared in an All-vs-All analysis, the heatmap below (Fig. 5.18) shows a clear separation of complete and partial TLR proteins. Within each of these groups there is further variation, TLR5 protein models form a group separate from other partial genes with Z-values averaging 25.3 between TLR5 structures but decreasing to a maximum of 18.8 between *G. palpalis gambiensis* TLR5 and *G. brevipalpis* TLR1. Toll-like receptor 3 and TLR13 also group, though it is clear the TLR3 protein family has a more conserved structure with Z-values ranging between 29 and 51.5 within the TLR3 family. It should be noted that the *G. palpalis gambiensis* TLR1 is also observed within the group.

Toll-like receptor 9 proteins show a high level of conserved structure, being the only complete protein to form an exclusive group, though Z-values are lower than those seen in TLR3 (ranging from 33.1 to 39.1). While TLR6 and TLR8 proteins show a higher degree of conservation between the two families, TLRs1, 2 and 7 do not form individual clades rather conservation appears to be random both within and between species and genes. Interestingly, *G. f. fuscipes* TLR2 appears to bare no conservation with any other TLR protein ( $Z = 0.1$ ), contradicting the observations made above (Fig. 5.18).

Further analysis was conducted using PCA to assess the relationship between the structures and visualise the relationship between each TLR protein (Fig. 5.19). As with the heatmap (Fig 5.17), this illustrated the presence of specific TLR5 (-70, -45) and TLR9 (-15, -15) groupings, however, as with the previous PCA (Fig. 5.15), TLR3 and TLR13 showed a greater degree of variation than expected. The remaining TLR proteins show no clear groupings with all forming a large group. The PCA analysis once again illustrated the two notable outliers within the protein structure. *Glossina palpalis gambiensis* TLR1 is found between

the TLR13 and TLR3 groups, while *G. f. fuscipes* TLR2 can be overserved as a clear outlier (Fig. 5.19).

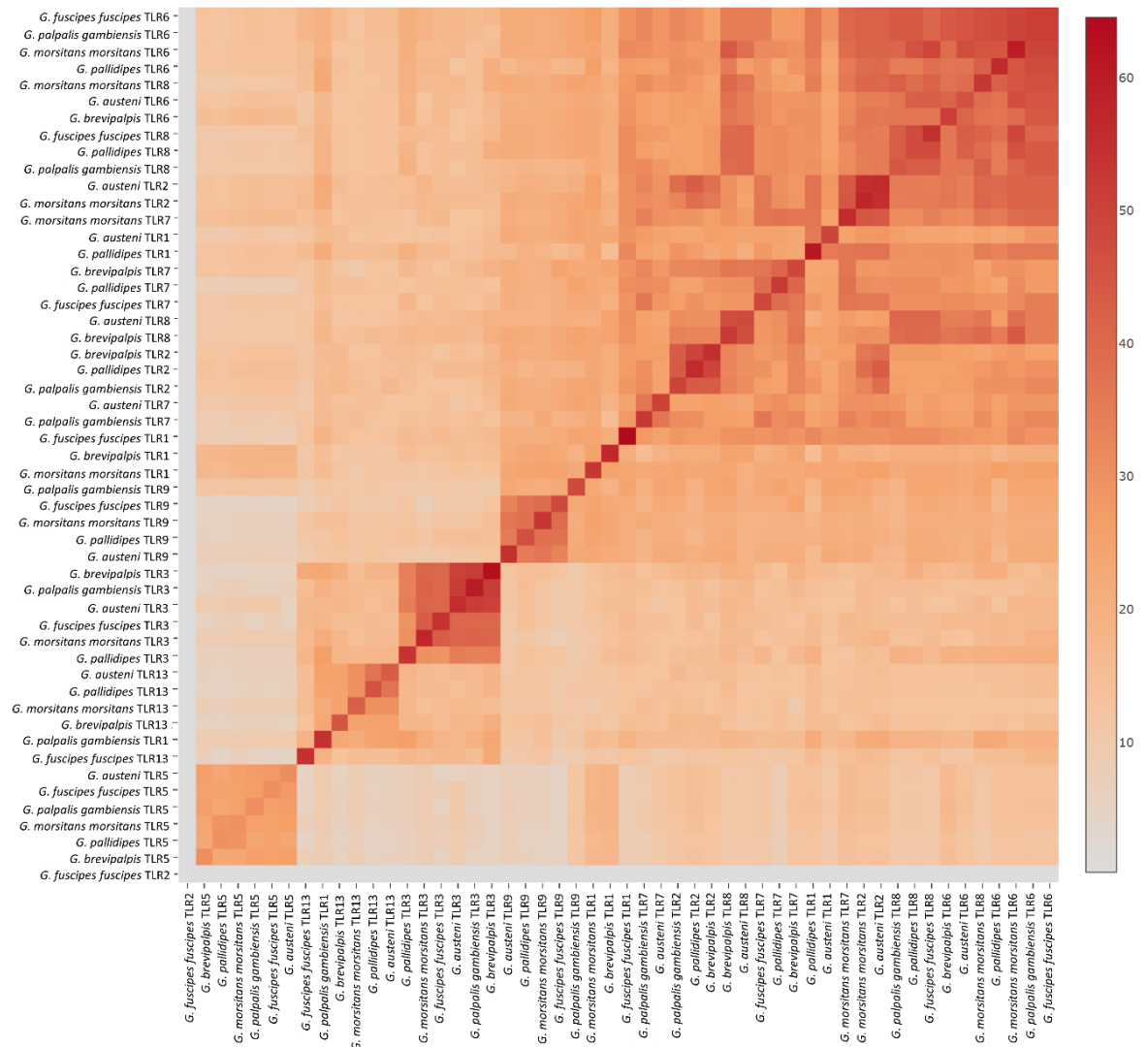


Figure 5.18: A heatmap comparing the conservation of TLR protein structures across all identified *Glossina* TLR genes. The heatmap was constructed using DALI server (Holm, 2019). Colour representing the D-value as estimated by DALI, higher Z-values, and thus conservation, are shown by the deep red colouration.

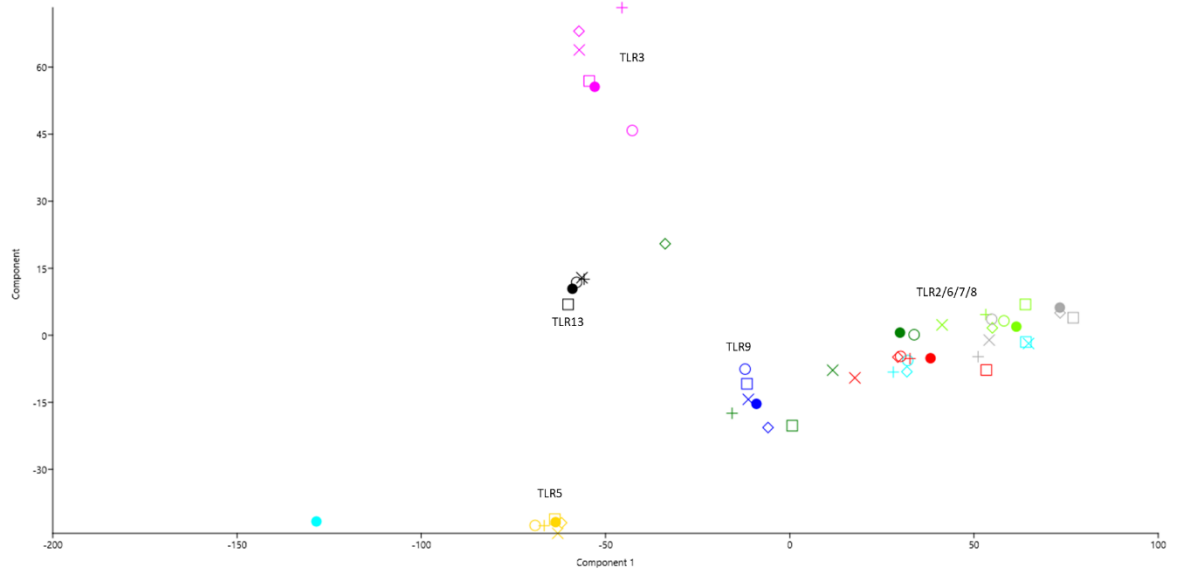


Figure 5.19: Principle component analysis (PCA) plot of all TLR protein structures within the *Glossina* genus using the first and second principle components (Eigenvalues: PC1 = 2770 (57.293 % variance); PC2 = 755.793 (15.632 % variance). Z-values were calculated, and a matrix produced using DALI (Holm, 2019). PCA analysis was conducted in PAST3 (Hammer et al., 2001). Individual predicted TLR genes are show by plots, species is denoted by shape: *G. austeni* = X; *G. brevipalpis* = +; *G. f. fuscipes* = ●; *G. m. morsitans* = □; *G. pallidipes* = ○; *G. palpalis gambiensis* = ◇. While gene families are denoted by colour. TLR1 = Green, TLR2 = Aqua, TLR3 = Pink, TLR5 = Gold, TLR6 = Grey, TLR7 = Red, TLR8 = Light Green, TLR9= Blue and TLR13 = Black.



## 5.4: Discussion

The aim of this chapter was to identify and characterise the TLR genes and proteins present within the *Glossina* genome assemblies. Nine predicted TLR genes were identified within the *Glossina* genus, of which six (TLR1, 2, 6-9) were seen to code for full TLR proteins, while three (TLR3, 5 and 13) were seen to code for partial TLR proteins. No orthologues of the *Drosophila* TLR4 or *Anopheles* TLR10 and TLR11 genes were identified within the *Glossina* genomes. The identification of potential TLR13 genes within the *Glossina* genus, shows a distinct variation from the *Drosophila* and *Culicidae* genera, though TLR13 has been previously annotated within the dipteran genera *Musca* and *Stomoxys*.

Of the predicted TLR genes identified, *Glossina* TLR1, 2, 6-9 appear to follow evolution previously observed within other dipteran genera (Christophides *et al.*, 2002; Levin and Malik, 2017). Both Christophides *et al.*, (2002) and Levin and Malik (2017), observed the presence of a TLR2\_7 clade within the dipteran TLR families, this was mirrored by those genes predicted within the *Glossina* genome. Predicted genes for *TLR6* and *TLR8* also formed a subclade showing high conservation between the two genes, this was not observed in any previous literature, with previous studies showing a close evolutionary relationship between *TLR8* and *TLR9* (Levin and Malik, 2017). Predicted genes for *TLR1* form a subclade off the *Culicidae* TLR1A/B and TLR5A/B clade, again this was observed by Christophides *et al.*, (2002), suggesting that those predicted as *TLR1* are members of that gene family. The observation of a clade containing *TLR1* and *TLR9* is unexpected as all previous phylogenetic analysis of dipteran TLR genes suggest that *TLR9* is divergent from *TLR8* rather than *TLR1* (Christophides *et al.*, 2002; Levin and Malik, 2017). The phylogenetic analysis indicates that the majority of the predicted TLR genes have been correctly identified, however, the separation of TLRs 3, 5 and 13 suggest that further analysis is required. This is further supported by pairwise distance PCA analysis, where all genes show a low divergence, while forming clusters that mirror the clades produced by the phylogenetic analysis.

The conserved characteristic genetic structure of TLR genes can also be observed within each of the predicted TLR1, 2, 6-9 genes. The identification of extracellular, transmembrane and TIR domains supports their identification as TLRs as does the conserved nature of these structures when compared to other dipteran TLR genes. Interestingly, there does appear

to be some variation between haematophagic species and other dipteran species especially within TLR1, 2 and 7. Predicted *TLR1* genes contain two LRRs between the LRR-NT and second LRR-CT domains which appear to be only present within haematophagous genera (*Glossina*, *Stomoxys*, *Aedes* and *Anopheles*). While *TLR2* genes of haematophagous species also contain a higher number of LRR domains in the initial extracellular domains, with *S. calcitrans* and *Glossina* spp. containing 18 or 19 LRRs (prior to the LRR-CT) compared to the 17 found in *D. melanogaster* and *M. domestica*. Finally, the structure of *Glossina TLR7* genes is almost identical to that of *S. calcitrans* but differs to that of *D. melanogaster* and *M. domestica*, suggesting some variation between haematophagous and other dipteran species. While there is no clear explanation for this, it is likely a result of evolutionary adaptations to control pathogens ingested during a blood meal.

Literature regarding the variation of specific genes within haematophagous, herbivorous, and nectarivorous species is surprisingly scarce. However, several studies have documented the proteomic variation in haematophagous saliva and compared to other insect species, describing the presence of haematophagous-specific genes for anticoagulants and variations in immune gene expression (Andrade *et al.*, 2005; Ware and Luck, 2017; Arcà and Ribeiro, 2018). However, whether there are indeed gene variants specific to haematophagous species remains unclear.

The 3-D protein models of TLR1, 2, 6-9 show that a complete TLR protein is coded for by each gene. These models form the characteristic shape of TLR proteins, with all three protein domains being present. Direct comparison of the structures illustrates a similar pattern to that of the phylogenies, forming a clear cluster containing TLRs1,2, 6-9. While TLR9 structures can be observed to form a tighter cluster compared to other proteins, the majority of proteins structures show no clear separation into gene families.

Of the gene families discussed above, *TLR1* shows a greater degree of variation within the gene family. Firstly, the pairwise distance PCA plot illustrated a relatively large degree of variation between *G. brevipalpis TLR1* and other predicted *TLR1* genes. However, this variation is in keeping with the established speciation of the *Glossina* genus (Dyer *et al.*, 2008). The divergence of the Fusca group species from the Morsitans and Palpalis groups suggests that the variation observed within the *TLR1* gene family may be a result of species-specific mutation, common in TLR genes (Coscia *et al.*, 2011), within the *Glossina* Fusca

group. In order to assess this, the prediction of additional Fusca group TLR1 genes would be required, followed by further evolutionary analysis within the *Glossina* genus.

PPI000821 (*G. palpalis gambiensis*) illustrated little variation to other *TLR1* genes within either the phylogenetic or pairwise distance analysis, however the CDS length and protein coding regions demonstrated considerable variation from other members of this family. The CDS of GPPI000821, obtained from VectorBase, is significantly shorter than those observed in other species at only 1,185 bp, this is likely a result of missing sequence data within the *G. palpalis gambiensis* genome. This gene is found on Scaffold3740 which is just 1,274 nucleotides in length and there is a significant amount of missing sequence data either side of this scaffold. Therefore, it is highly likely that the N and C-terminals of GPPI000821 are in this missing data, meaning that the missing transmembrane and TIR regions are in fact present within the genome. Gaps in genome assemblies are common with multiple studies attempting to address them using recent technical advancements in next-generation sequences (NGS) (Tørresen *et al.*, 2017; Utturkar *et al.*, 2017; Peona *et al.*, 2021). Interestingly, improved assembly of the Atlantic Cod genome resulted in the discovery of new tandem repeats (Tørresen *et al.*, 2017). Therefore, it is possible that improved assembly of the *Glossina* genome will help verify the observations above.

Contrary to *G. palpalis gambiensis*, the predicted *G. m. morsitans TLR1* gene contains an exceptionally long CDS at 9,165 nucleotides. While GMOY011790 does encode a full TLR protein, the protein domains are only observed near the C-terminal of the CDS. It is likely that this increased CDS is a mistake made during the annotation of GMOY011790, as no protein domains can be found prior to the start of the TLR extracellular domain. Whether the N-terminal section is either a noncoding sequence or if there is a mistake within the genome sequence remain to be determined. Despite this variation within the predicted *TLR1* gene family, all six genes (TLR1, 2, 6-9) illustrate significant conservation with orthologous in the TLR superfamily and should be consider as such.

Interestingly, Lima *et al.* (2021) reported two TLR paralogues within *G. brevipalpis* where only one was fully identified in this study. A search of TLR genes within the *G. brevipalpis* genome indicated another annotated Toll orthologue in GBRI007354, however, this gene only codes for the LRR and transmembrane domains, unlike GBRI007308 (identified above) which codes for a full TIR protein and is missing the TIR domain. Therefore, GBRI007354

cannot encode a functional TLR gene on its own, despite having strong sequence similarities to both *D. melanogaster* and other *Glossina TLR1* genes. Curiously, the previous annotated gene, GBRI007364, codes for a transmembrane and a TIR domain, though there is no indication of an LRR region. As such, it is possible that a second *TLR1* orthologue is present within the *G. brevipalpis* genome though the current annotation of the genome does not show this. If the *G. brevipalpis* genome does indeed encode two *TLR1* paralogous it could be a unique adaption of the Fusca group as no species from either the *Morsitans* or *Palpalis* groups have indicated this.

Phylogenetic analysis placed the predicted *Glossina TLR5* gene in a separate clade within *M. domestica* and *S. calcitrans* (Fig. 5.3). However, the clade shows a conserved nature within the predicted genes with a low pairwise distance between genes, as illustrated by the PCA plot. However, this variation from the predicted TLR genes and the *TLR5* gene previously identified within *D. melanogaster* may be explained by the genetic structure of the genes. All six genes have a considerably shorter CDS than other TLR genes and are missing the TIR domain. Interestingly, the CDS length of the predicted *Glossina TLR5* genes is similar to that of the *D. melanogaster* orthologue, therefore it is likely that the absences of the TIR domain are the cause of this variation.

Structurally, *Glossina TLR5* show a high conservation (Fig. 5.17 and 5.18) with the C-terminal of the extracellular domain and the transmembrane region clearly visible following protein modelling (Fig. 5.16). This suggests that *TLR5* may be present within the genome, though poor genome annotation may prevent accurate identification. Therefore, considerable additional research is required before *TLR5* can be confirmed or rejected as present within the *Glossina* genome.

The final two TLR gene families identified, *TLR3* and *TLR13* form a separate clade with a high variance between the families (Fig. 5.3 and 5.12). Each gene follows the speciation of the *Glossina* genus (Dyer *et al.*, 2008) and shows little to no pairwise distance between identified genes (Fig. 5.13). As previously stated, the variation between *Glossina TLR3* gene and that of *D. melanogaster* is apparent in the number of LRRs present and lack of transmembrane and TIR domains. Orthologues for predicted *Glossina TLR3* gene can be observed in both *M. domestica* and *S. calcitrans* though there is no literature to support their identification. Toll-like receptor 13 has no orthologue within the *Drosophila* genome,

however, has previously been annotated within the *S. calcitrans* genome on VectorBase, though as with *TLR3* there is no literature to support this annotation. Both *TLR3* and *TLR13* code for a full LRR receptor protein, however, with the absence of a transmembrane and TIR domain, in addition to the highly conserved nature of LRR proteins (Dhulkotia *et al.*, 2000; Helft *et al.*, 2011) it is likely that these predicted genes are, in fact, not members of the TLR superfamily.

## 5.5: Conclusion

The identification of six TLR families within the *Glossina* genomes confirms the observation of Lima *et al.*, (2021). However, further research is required to fully characterise the three partial genes hypothesised to be orthologous of TLRs 3, 5 and 13, while the evidence suggests that these are not true TLRs they could represent pseudo genes or other related LRR proteins. These results establish the fundamental aspects of TLR genomic analysis within *Glossina* species and the intraspecies variation must also be consider (see chapter 6). Interestingly the present of TLR2, 6 and 9 presents the opportunity for future analysis to determine if trypanosomal identification in *Glossina* is similar to that of triatomines or if they are following a separate signalling pathway.

## 6: Intraspecies variation of the Toll-like Receptor 2 gene within wild *Glossina morsitans morsitans*, and the association with endosymbiont and trypanosome infection.

### 6.1: Introduction

The TLR signalling pathway is critical to establishing an effective immune response to invasive pathogens and has been described in detail previously in this study (see Chapters 1 and 5). Therefore, it is understandable that majority studies concerning genetic variation within TLR genes has been directed to determining the consequence for susceptibility and resistance to infection (Christophides *et al.*, 2002; Sackton *et al.*, 2007; Cuscó *et al.*, 2014; Antonides *et al.*, 2019). The highly polymorphic nature of TLRs is similar to the rate of nucleotide variation observed in genes under pathogen associated coevolution and purifying selection (Wlasiuk and Nachman, 2010; Netea *et al.*, 2012). However, it is unlikely that the evolution of immune signalling pathways within the Diptera is driven purely by pathogen-host coevolution (Anderson and May, 1982) as coevolution primarily affects proteins that directly interact with one another, which is not the case for dipteran TLR proteins (Janeway and Medzhitov, 2002). Therefore, it is important to understand what selective pressure are driving TLR evolution within Dipteran spp.

The intraspecies evolution of TLR genes at a population level has been thoroughly described across multiple taxa, with numerous publications describing the evolutionary and functional impacts of polymorphisms within TLRs across mammalian (Tschirren *et al.*, 2013; Cuscó *et al.*, 2014), avian (Antonides *et al.*, 2019), fish (Palti, 2011) and insect (Christophides *et al.*, 2002; Evans *et al.*, 2006; Sackton *et al.*, 2007; Zou *et al.*, 2007) species. Polymorphisms within mammalian TLRs have been directly associated with susceptibility to infection (Schröder and Schumann, 2005; Cheng *et al.*, 2007; Misch and Hawn, 2008; Netea *et al.*, 2012; Fels Elliott *et al.*, 2017). However, these occurrences are rare and under strict purifying selection, preventing fixation of these variants and reducing the impact on the population as a whole (Netea *et al.*, 2012).

Despite the strong purifying selection exerted on TLR polymorphism, variation associated with susceptibility to specific pathogens between populations supports the observation of episodic positive selection within TLR genes (Wlasiuk and Nachman, 2010). One such example can be observed in the prevalence of a Gly299 variant of human TLR4 between African and European populations. This variant offers an advantageous resistance to the malarial parasite *Plasmodium falciparum*, though increases susceptibility to gram-negative sepsis and is therefore, prevented from reaching fixation by purifying selection (Ferwerda *et al.*, 2007). However, while this variant has been practically eliminated from the European population, it has been recorded in approximately 15% of the African population (Ferwerda *et al.*, 2007; Netea *et al.*, 2012). Interestingly, genetic variation within TLR genes of the avian species bananaquit (*Coereba flaveola*), was also found to be directly correlated to malarial infection within a population (Antonides *et al.*, 2019). The authors' observed that infected groups exhibited a higher number of alleles at a lower frequency, including several unique alleles, than uninfected groups, suggesting that natural selection was favouring resistant alleles while maintain genetic variation within the population.

Studies regarding TLR variation within Dipteran taxa is lacking. Studies into the sequence variation of mosquitoes immune receptor found that these conserved genes were often under purifying selection to eliminate sequence variation, as observed in mammalian genes (Little and Cobbe, 2005). While there are multiple accounts of rapid evolution of immune genes within the *Drosophila* immune signalling pathways, including TLRs, these serve as general overview rather than in-depth assessment of one pathway specifically (Schlenke and Begun, 2003; Sackton *et al.*, 2007; Juneja and Lazzaro, 2009). One fascinating explanation for majority of adaptive mutations within signalling pathways was presented by Begun and Whitley (2000), who suggested the possibility of direct interference of immune signalling by a pathogen. This phenomenon has been observed specifically in bacterial pathogens that can inject interference proteins directly into host cells inhibiting immune signalling (Salyers *et al.*, 1994; Schmid-Hempel, 2008).

The importance of bacterial endosymbiosis to the *Glossina* cannot be understated, therefore these bacteria must be able to avoid, the primary symbiont within the *Glossina* genus, *Wigglesworthia glossinidia*, a gram-negative bacterium which plays a critical key role in sexual and immunological development of juvenile tsetse (Pais *et al.* 2008). As these are gram-negative bacteria, the stimulation of the TLR pathway should result in an immune

response to clear the bacteria however, this does not appear to be the case. An in depth review of bacterial endosymbionts within arthropods suggested there are two primary methods utilised by endosymbionts to minimise the impact of the host immune system on symbiont populations: location and immune evasion (Hurst and Darby, 2009). Intracellular infections, like that of *W. glossinidia* within the tsetse bacteriome (Aksoy, 1995), do not directly interact with immune system. However, intracellular defences such phagolysosome and high levels of reactive oxygen species can still result in clearance of the bacteria from cells (Urban *et al.*, 2006).

The second method described by Hurst and Darby (2009), proposes that extracellular endosymbionts may avoid detection by the host immune system altogether, this was observed within the *Glossina* milk glands where *W. glossinidia* live extracellularly (Pais *et al.*, 2008; Wang *et al.*, 2009). Wang *et al.*, (2009) observed a correlation between elevated levels of peptidoglycan recognition protein (*PGRP-LB*) within female *Glossina* samples, and the population size of *W. glossinidia* within the fly. The PGRPs are a large conserved protein family that stimulate the immune pathways within insect species (Dziarski and Gupta, 2006). However, PGRP-LB has been observed to remove free PGN within both *Drosophila* (Zaidman-Rémy *et al.*, 2006) and *Glossina* (Wang *et al.*, 2009), effectively masking the presence of *W. glossinidia* and preventing a full immune response to the endosymbionts.

Literature on the interactions between trypanosomes and TLRs within *Glossina* species is currently severely lacking. However, four TLRs (*TLR2*, *TLR4*, *TLR6* and *TLR9*) have been reported to recognise ligands from the genus *Trypanosoma*, specifically *Trypanosoma cruzi* (the causative agent of Chagas Disease) within their triatomine vector (Bafica *et al.*, 2006; Uematsu and Akira, 2008; Kumar *et al.*, 2009). Given the similarity of other protein interactions in response to trypanosomal infection, and the conserved nature of TLRs, it is plausible that similar interactions may be present between African trypanosomes and the *Glossina* genus (Ursic-Bedoya *et al.*, 2011). Several trypanosomal PAMPs have been recognised to activate the TLR pathway, primarily glycosylphosphatidylinositol (GPI) anchors. Glycosylphosphatidylinositol anchors recognition has been observed to active TLR2, TLR4 and TLR6 (Campos *et al.*, 2001), while glycoinositolphospholipids were observed to be recognised by TLR4 but not TLR2. It should be noted however, that the identification of GPI anchors by TLR6 is reliant upon the formation of a heterodimer with TLR2 (Uematsu



and Akira, 2008). TLR9 was also observed to play critical role in *T. cruzi* resistance (Bafica *et al.*, 2006).

Indeed the concept that *T. brucei* and other African *Trypanosoma* species may exhibit a similar response in other hosts is supported by the findings of Drennan *et al.*, (2005) who found that *T. brucei* infection within mice was MyD88 dependent. As MyD88 is the primary intercellular adaptor protein within the TLR pathway, this suggests that trypanosomal infections stimulate a similar immunological response in both mammalian and insect hosts. As MyD88 is conserved within the TLR signalling pathway it strongly suggests that the TLR signalling cascade is also vital within the *Glossina* response to parasitic infection.

### 6.1.1: Aims and Objectives

This is the first study of its kind to assess a TLR within a wild *G. m. morsitans* population. Toll-like receptor 2 was selected due to the critical role it plays in the detection of four PAMPs associated with *T. cruzi* within the triatomine vector (Uematsu and Akira, 2008; Kumar *et al.*, 2009), and is therefore, likely to exhibit a similar response within *Glossina* in response to African trypanosomes. Although *T. cruzi* differs substantially from African *Trypanosoma* spp., similar PAMPs are likely responsible for the detection of African and American trypanosome spp. The central region of the extracellular region was targeted for sequencing as polymorphic sites tend to cluster in protein receptor regions (Sackton *et al.*, 2007). Therefore, this section of the extracellular region is likely to contain the higher concentration of polymorphic sites.

In this chapter we aim to address objective 5 (section 1.5): To evaluate the intraspecies variation of TLR2 and assess the impact of selection upon both the structure and function, in relation to symbiont and trypanosome infection.

Following the methodology described in chapter 3, genetic variation was assessed within *TLR2*, using the same gDNA extracted from the tsetse collected from the three subpopulations in Northern Zimbabwe. The evolutionary history of the *TLR2* sequences was measured using phylogenetic methodologies, while allele diversity was assessed using haplotype analysis. The frequency of synonymous and nonsynonymous polymorphic sites was assessed to illustrate intraspecies nucleotide variation. As before, gene flow,

population genetics and tests of neutrality were conducted to detect recent population changes and indicate divergence from neutrality indicative of natural selection.

Protein structural and functional variation and natural selection were assessed using much of the methodology described in chapter 5. Three dimensional modelling of protein variants was undertaken using the I-Tasser server (Yang *et al.*, 2015) to assess for and compare any structural deviations observed in the sample. While both Z-score and HyPHY was utilised to assess the current selective nature of the TLR2 genes.

Finally, the relationship between TLR2 genetic variation, endosymbionts and trypanosome infection was assessed. *Wigglesworthia glossinidia* and trypanosome screening was conducted in Chapter 4 and this chapter uses those same results. Direct comparisons of TLR2 variation and *W. g. morsitans* variation was achieved as described in Chapter 4 by overlapping haplotype networks. The same process of assessing parasite infection by comparing the TLR2 haplotype network to infection rather than geographical location.

## 6.2: Materials and Methods

### 6.2.1: Tsetse samples collection and gDNA extraction

Please refer to Chapter 3 (sections 3.2.1 and 3.2.2) for details on the collection of wild tsetse samples and gDNA extraction.

### 6.2.2: Primer design and Polymerase Chain Reaction

Primers were designed using NCBI Primer BLAST (Ye *et al.*, 2012, *Glossina* genome data on VectorBase (Giraldo-Calderón *et al.*, 2015) was mined for relevant sequences to provide a template for primer design (Table 6.1). This study targets the TLR ectodomain as the receptor region of the protein, as the size of the target fragment exceeded 1500 bp two overlapping fragments were targeted to produce a single longer gene.

Polymerase Chain Reaction (PCR) was conducted to amplify the target fragment of TLR2. Amplification was conducted using 12.5 µl of DreamTaq™ PCR master mix (2X DreamTaq buffer, 0.4 mM of each dNTP, 4mM MgCl<sub>2</sub>) (Thermo Scientific, UK), 1 µl each of gene-specific forward and reverse primers and 1 µl of gDNA, with a final reaction volume of 25 µl made up with PCR grade water. Cycling conditions were set at an initial denaturation at 94 °C for five minutes, followed by 35 cycles of 94 °C for 30 seconds, 57°C for 30 seconds and 72°C for one minute, a final 10 minute extension was conducted at 72°C before being held at 4°C.

Gel electrophoresis was performed to determine the success of PCR amplification. 5 µl of PCR products was mixed with 1.2 µl 6x Gel loading dye (Thermo Scientific, UK) and GelRed™ Stain (Cambridge Bioscience, UK), 1Kb Hyperladder (Bioline, UK) was used as marker. All samples were run on a 1% agarose at 90 V for 45 minutes prior to visualisation on an UV transilluminator (gel images available in Supplementary Figures 6-7 Appendix 4).

Table 6.1: *TLR2* prime information: nucleotide sequence, length, melting temperature (T<sub>m</sub>), G-C percentage and expected fragment size of each primer used in this study.

Target gene/fragment	Name	Sequence (5'-3')	Length (bp)	T <sub>m</sub> (°C)	GC%	Fragment size	Designed by
<i>TLR2-A</i>	<b>TLR2-AF</b>	ATCATAAGTCAGGTGCAGTC	20	54.86	45.00	1042bp	This work
<i>TLR2-A</i>	<b>TLR2-AR</b>	CAACGCCATTTGGGTAAT	20	55.11	40.00		This work
<i>TLR2-B</i>	<b>TLR2-BF</b>	CGATTGGCCATATAGAGGAT	20	54.15	45.00	961bp	This work
<i>TLR2-B</i>	<b>TLR2-BR</b>	TCCATTGAACAGTCGCATT	19	55.10	42.11		This work

### 6.2.3: Sanger sequencing and sequence analysis

Sanger sequencing was conducted by the DNA sequencing facility at the Natural History Museum, London as detailed in Chapter 3, section 3.2.5. Sequence chromatograms were read using SnapGene software (from GSL Biotech; available at [snapgene.com](http://snapgene.com)). Sequences were then aligned using the MUSCLE (Multiple Sequence Comparison by Log-Expectation) sequence alignment tool (Madeira *et al.*, 2019). The two TLR2 fragments (A and B) were aligned to form one continuous sequence for analysis.

Of the 63 samples, 62 full TLR2 fragments were successfully constructed. These were then submitted to phylogenetic and population genetic analysis.

### 6.2.4: Intra-species phylogenetic analysis

Phylogenetic analysis was conducted to estimate the evolutionary history of each gene within the sample population. A sequence length of 1,701 bp was used and Phylogenetic analysis was conducted using MEGAX (Kumar *et al.*, 2018), Neighbour-joining trees were constructed using the Jukes-Cantor model with 1,000 bootstrap replicates. Maximum-likelihood trees were constructed using the model of best fit, estimated using the MEGAX 'Find Best NDA/Protein Models' function. This indicated that Tamura 3-parameter model + Gamma distribution with Invariant sites (T3+G+I) best suited the data. 1,000 bootstrap replicates.

### 6.2.5: Haplotype analysis

Further to phylogenetic analysis, the haplotype variation of each gene was analysed to provide further support the predicted evolution. Haplotype data files were generated in DnaSP (version 6) (Rozas *et al.*, 2017), and TCS haplotype network was produced in PopART (Leigh and Bryant, 2015) using the method described in Chapter 3, section 3.2.7.

### 6.2.6: Intra-species nucleotide variation analysis

Having predicted the evolutionary and haplotype variation within each gene, the sites of DNA polymorphisms and nucleotide variation at those points was also investigated. Nucleotide variation ( $\pi$ ) was calculated using the 'DNA polymorphism' function in DnaSP (version 6) (Rozas *et al.*, 2017) (Refer to Chapter 3, section 3.2.8). The region of analysis

was set between base pairs 1,108 and 2,809, codon specific sliding window analysis was conducted (window size = 3; step size = 3). Nucleotide variation ( $\pi$ ) is estimated using equation 1 in Appendix 2 (Nei, 1987, equation 10.5 or 10.6; Nei and Miller, 1990, equation 1).

#### 6.2.7: dN/dS: Synonymous vs non-synonymous variation

Having established the presence of DNA polymorphisms within the gene fragments, the nature of these mutations was also established in order to predict the potential functional and structural impacts on the protein synthesis. The presence of synonymous (dS) and non-synonymous (dN) mutations was accessed in DnaSP (version 6) (Rozas *et al.*, 2017) as specified in section 3.2.9, while region of analysis was specified between base pairs 1,108 and 2,809. Pi ( $\pi$ ) values were calculated using equation 1 (Nei, 1987, equation 10.5 or 10.6; Nei and Miller, 1990, equation 1). dN/dS ratios (also referred to as Ka/Ks) were calculated using the 'Pi(a)/Pi(s) and Ka/Ks ratios' function, with all parameters remaining the same as above.

#### 6.2.8: Gene flow analysis

Given the severe lack of literature concerning wild tsetse populations, population genetic analysis was conducted to provide a novel insight into the *G. m. morsitans* population. Gene Flow and Genetic Differentiation analysis was conducted in DnaSP (version 6) (Rozas *et al.*, 2017) and while a Mantel-test was conducted using PAST3 (Hammer *et al.*, 2001) as described previously in 3.2.10.

Genetic distance between sample was calculated using Pairwise-distance (P-distance) in MEGAX (Kumar *et al.*, 2018) this was compared to geographical distance to assess the relationship of genetic variation between subpopulations.

#### 6.2.9: Demographic change and test for neutrality: Pairwise mismatch, Tajima's D, Fu's $F_s$ and Coalescent Simulation

The impacts of population genetics on genetic variation were considered by examining the demographic change and neutrality of the genes. Tests of neutrality, Tajima's D (Tajima, 1989) and Fu's  $F_s$  (Fu, 1997), pairwise mismatch (Watterson 1975; Slatkin and Hudson 1991;

Rogers and Harpending 1992) and raggedness ( $r$ ) (Harpending, 1994) analysis were undertaken in DnaSP (version 6) (Rozas *et al.*, 2017) as described in section 3.2.11.

#### 6.2.10: Recombination analysis

As described in section 3.2.12, recombination was assessed within the *TLR2* sample to assess the impacts on nucleotide diversity and the implications for the interpretation of pairwise mismatch analysis. This was conducted using the two methods (DnaSP V6 and GARD) detailed in 3.2.12.

#### 6.2.11: Indication of Selection: Z-tests and HyPHY based analysis

The presence of selective pressure acting upon the *TLR2* genes was assessed using the two methods described in sections 4.2.3 and 4.2.4. MEGAX (Kumar *et al.*, 2018) used the standard statistical Z-test test to assess deviations from the population mean neutral selection (see section 4.2.3). Hypothesis testing using Phylogenies (HyPHY) was also employed (Pond and Muse, 2005), this was run on both MEGA7 (Kumar *et al.*, 2016) and Datamonkey online servers (Pond and Frost, 2005; Weaver *et al.*, 2018) as described in section 4.2.4.

#### 6.2.12: Three-dimensional protein modelling and function impact

Structural predictions of the *TLR2* protein variants were made using the I-TASSER online server (Zhang, 2008; Roy *et al.*, 2010; Yang *et al.*, 2015), models with highest TM- and C-score was selected as the more reliable structure. All secondary structure models were visualised, using PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre).

The functional impacts of amino acid variants were predicted using Simple Modular Architecture Research Tool (SMART) (Letunic and Bork, 2017) and PROVEAN (Choi, 2012; Choi *et al.*, 2012) as described in sections 5.2.2 and 4.2.6 respectively.

#### 6.2.13: *Wigglesworthia* haplotype variation and population genetics

Given the obligatory nature of symbiosis between *W. glossinidia* and *Glossina* spp. the *Wigglesworthia 16S* gene was submitted to the same population genetic analysis. PCR amplification and sequencing of *W. g. morsitans 16S* rRNA gene was conducted as described in sections 3.2.3 to 3.2.5. Amplification was successful in 34 of the 63 (53.97%)

*G. m. morsitans* samples. Eleven samples were successfully amplified from both the Nykasanga and Rekomitjie collection sites, and 12 sequences were obtained in samples from Makuti. This equates to 78.57% of Nykasanga samples, 55% from Rekomitjie and only 42.86% of the Makuti samples. A multiple sequence alignment was conducted as described in section 3.2.5 with a final sequence length of 1180bp. In order to assess the relationship between *G. m. morsitans* and *W. g. morsitans* the same analytical methods were used, haplotype analysis was conducted as described in section 3.2.7, while gene flow, Tajima's D, Fu's  $F_s$  and demographic change were all assessed as described above in section 3.2.10 and 3.2.11.

#### 6.2.14: Association of AMP and symbiont nucleotide variation

The association between *AttA* and *Def* nucleotide variation and that observed within *W. g. morsitans 16S* was assessed using a standard Pairwise-distance analysis in MEGAX (Kumar *et al.*, 2018). The *P*-distance between tsetse AMP sample was calculated and plotted against the corresponding *W. g. morsitans 16S* *P*-distance to illustrate the relationship between the two, a mantel-test was also conducted as described in section 3.2.10. This would give an insight into the influence of endosymbiosis on genetic variation



## 6.3: Results

### 6.3.1: Intra-species genetic variation and population genetics of wild *Glossina morsitans morsitans* TLR2 gene

#### 6.3.1i: Phylogenetic analysis

Phylogenetic analysis exhibited two slightly different topologies between the Neighbour-Joining and Maximum-Likelihood methods (Fig. 6.1). The Maximum-Likelihood tree exhibited five primary clades, subdivided into several subclades (Fig. 6.1A). Clade I consisted of three subclades of which one (clade I.a) contains samples solely from Makuti. Interestingly, clade I.b consisted of four samples of three from Rekomitjie and just one sample, M146, from Makuti. Two other clades (II.a and V.d) exhibited samples from just two collection locations, II.a exhibited two samples from Makuti and one from Nykasanga, while clade V.d contained two samples from both Makuti and Rekomitjie. All other clades contain either a sample from a single location or a mix of all three locations. However, all bootstrap values are low, ranging between 0 and 59, suggesting that this tree is not strongly supported.

In contrast, the neighbour-joining method (Fig. 6.1B) indicated eight primary clades containing varying numbers of samples. Interestingly, a greater number of observed clades were seen to contain samples from one location. Clades I.a, II and IV.a contain a combined total of 12 samples solely from Makuti, while clade I.c contains three samples solely from Rekomitjie. This high proportion of geographically isolated clades suggests that there is a relationship between geographical location and nucleotide variation when analysed using the bootstrap method. The remaining clades all contain samples from all three collection sites, in a more expansive tree than that produced by the maximum-likelihood method (Fig. 6.1A). Once again, the bootstrap values showed a large range in values between 0 and 93. Three nodes, one in each of clades I.d, III.c and IV.a were found to be significant with bootstrap values of 87, 93 and 79 respectively (Fig. 6.1B).

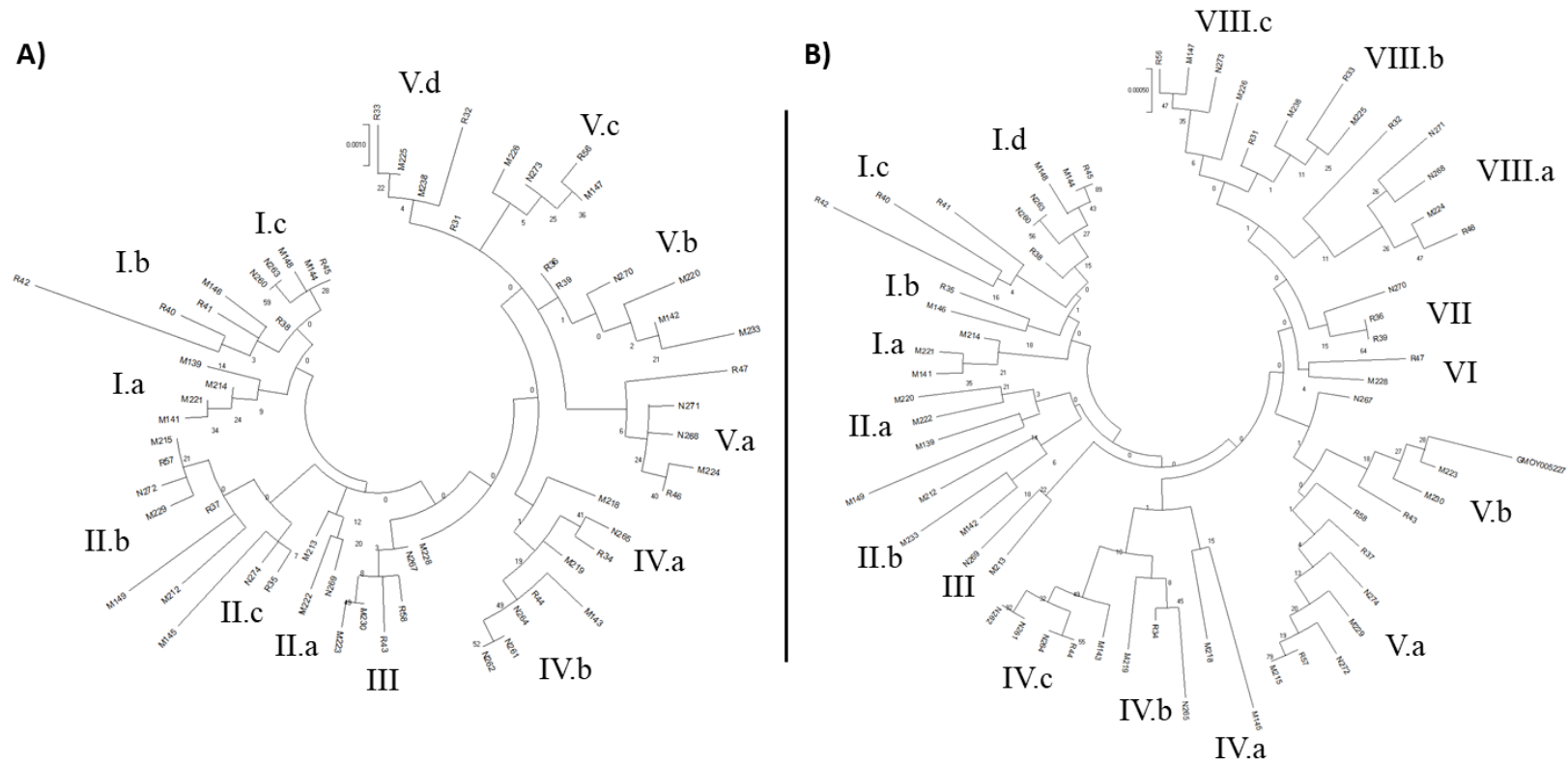


Figure 6.1: Evolutionary analysis of the *TLR2* gene fragments was conducted using MEGAX (Kumar *et al.*, 2018). A) A Maximum Likelihood tree produced using the Tamura 3-parameter model (Tamura, 1992). The tree with the highest log likelihood (-2827.31) is shown. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.1897)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 49.35% sites). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 62 nucleotide sequences. B) Neighbour-Joining method was used to predict evolutionary history (Saitou and Nei, 1987), the evolutionary distances were computed using the Jukes-Cantor method (Jukes and Cantor, 1969) and are in the units of the number of base substitutions per site. The optimal tree is shown with the sum of branch length = 0.05487047. Both trees are drawn to scale, with branch length measured in the number of substitutions per site. Codon positions included were 1st+2nd+3rd+Noncoding, while all positions with less than 95% site coverage were eliminated. 1000 bootstrap replicates were used.

### 6.3.1ii: Haplotype diversity

A total of 56 haplotypes were observed within the sample population ( $H = 56$ ) with a very high haplotype diversity ( $H_d$ ) of 0.997. There is no sign of a common haplotype within the TLR2 samples, with just six shared haplotypes in the sample population, two shared between Makuti and Rekomitjie, one shared between Nykasanga and Rekomitjie, and three location specific haplotypes containing two samples (two from Nykasanga and one from Rekomitjie). A simple AMOVA test showed no significant relationship between geographical location and genetic variation within the TLR2 gene ( $\phi_{st} = 0.01409$  ( $P = 0.187$ )).

### 6.3.1iii: Nucleotide diversity

Sliding window analysis illustrated a total of 22 polymorphic sites within the TLR2 fragment, with an overall nucleotide variation ( $\pi$ ) across all samples of  $\pi = 0.00377$ . These were unevenly distributed throughout the fragment, with 16 being found in the first half of the fragment (Fig.3A) and the remaining six in the second half. Both Nykasanga and Rekomitjie samples exhibited 19 of the 22 polymorphic sites, while the Makuti population exhibited all 22. The polymorphic site at nucleotide 1806 was observed in both the Rekomitjie and Makuti populations, while the site at nucleotide 2364 was observed in Nykasanga and Makuti. Interestingly, two polymorphic sites at nucleotides 1227 and 2706 were exhibited by a single sample from Makuti (Fig. 6.3A). Of these 22 polymorphic sites, 21 were found to be synonymous mutations with just one non-synonymous variation being observed at nucleotide 2364 (Fig. 6.3B).

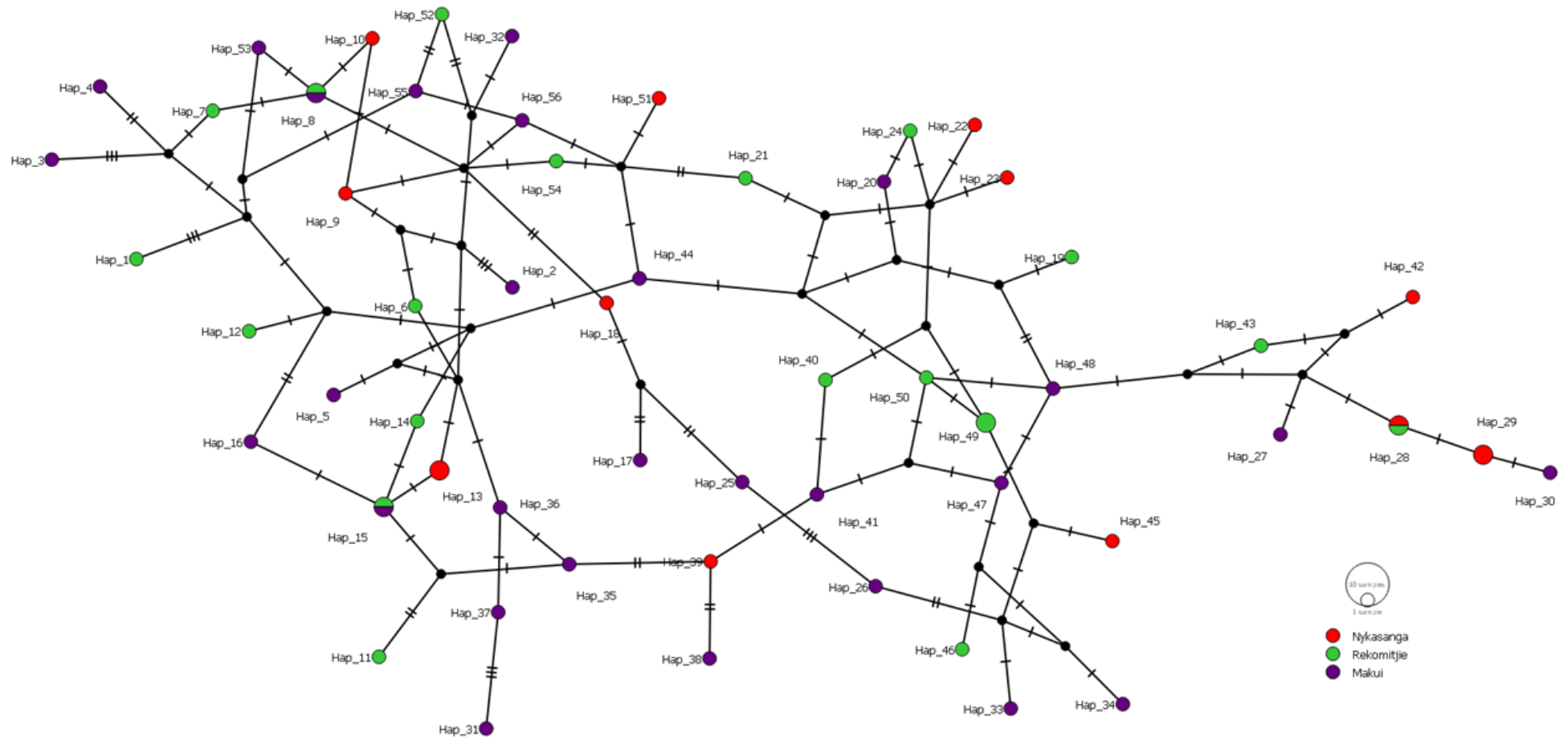


Figure 6.2: TCS haplotype networks of the *TLR2* gene fragment within the sample population. The circle size represents of the number of samples within a haplotype, while the colours represent the geographical location of each sample. Black lines crossing a branch indicate the number of nucleotide mutations between haplotypes and solid black circles signify inferred or missing haplotype. All networks were produced in PopArt (Leigh and Bryant, 2015).

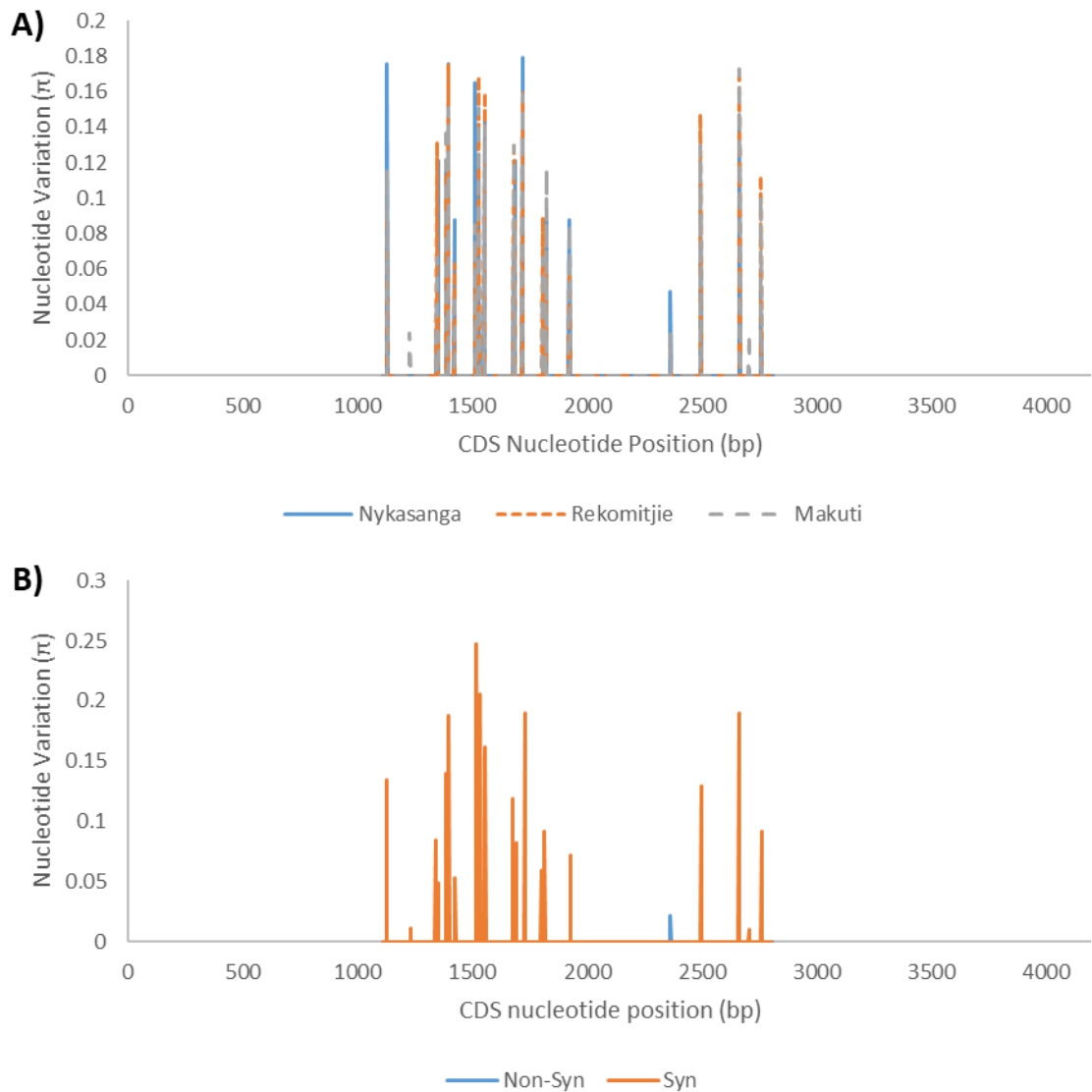


Figure 6.3: Sliding window graphs of the *G. m. morsitans* *TLR2* gene fragment; A) illustrates the distribution of polymorphic sites throughout the gene fragment. Each subpopulation is represented by the colour of the line. B) illustrates the characteristic (Synonymous and Non-synonymous) of each polymorphic site. Produced in DnaSP V6 (Rozas et al., 2017),  $\pi$  ( $\pi$ ) represents nucleotide diversity within the fragment against the nucleotide position of the mutation. The coding region of the fragment was set between nucleotides 1108 - 2809. Window size = 3, step size = 3.

### 6.3.1iv: Population dynamics and tests of neutrality

Gene flow analysis indicated a high level of gene flow between subpopulations with both haplotype and nucleotide statistics (Hs and Ks respectively) indicating high diversity between populations, though the number of substitutions between populations was low (Dxy = 0.004). The fixation index (Fst) was low (Fst ≈ 0) further supporting the hypothesis of a panmictic population presented in Chapter 4 (Table 6.2).

Table 6.2: The gene flow results for TLR2 gene fragment across all collections localities. Values are rounded to three d.p where possible. M = Makuti; N = Nykasanga and R = Rekomitjie. Hs: Haplotype statistic. Ks: Nucleotide statistic. F<sub>st</sub>: Fixation index. Dxy: The average number of nucleotide substitutions. Da: The net nucleotide substitution per site between populations. For equations used see section Appendix 2.

Pop. 1	Pop. 2	Hs	Ks	F <sub>st</sub>	Dxy	Da
N	R	0.988	6.36	0.024	0.004	0.00009
N	M	0.993	6.44	0.013	0.004	0.00005
R	M	0.997	6.41	-0.012	0.004	-0.00004

A direct comparison of geographical distance and Fst between collection sites indicated a strong positive correlation between distance and Fst ( $R^2 = 0.857$ ), however, this was found to be insignificant ( $P > 0.05$ ) interestingly, no relationship was observed between genetic distance (P-distance) and geographical distance.

Test for neutrality (Tajima's D and Fu's Fs) indicated mixed results. Tajima's D was found to be positive ( $D = 1.16044$ ) indicating balancing selection within the gene fragment. While Fu's Fs produced a very strongly negative result ( $F_s = -71.477$ ) indicating that the population expansion or recovery event. Though both results were statistically insignificant ( $P > 0.05$ ), they do support the previous observations made in Chapter 4.

Pairwise mismatch, under the presumption of no recombination, indicated a population expansion event, with a high observed frequency of variation but a relatively low observed pairwise difference (Fig. 6.4). Raggedness (r) was found to be low ( $r = 0.0133$ ,  $P > 0.05$ ) supporting the observation of a recent population expansion event. When r was calculated using Coalescent theory, under the presumption of free recombination, very little change was observed, with an average r value of 0.01816 and a 95% confidence interval between 0.00954 and 0.03382.

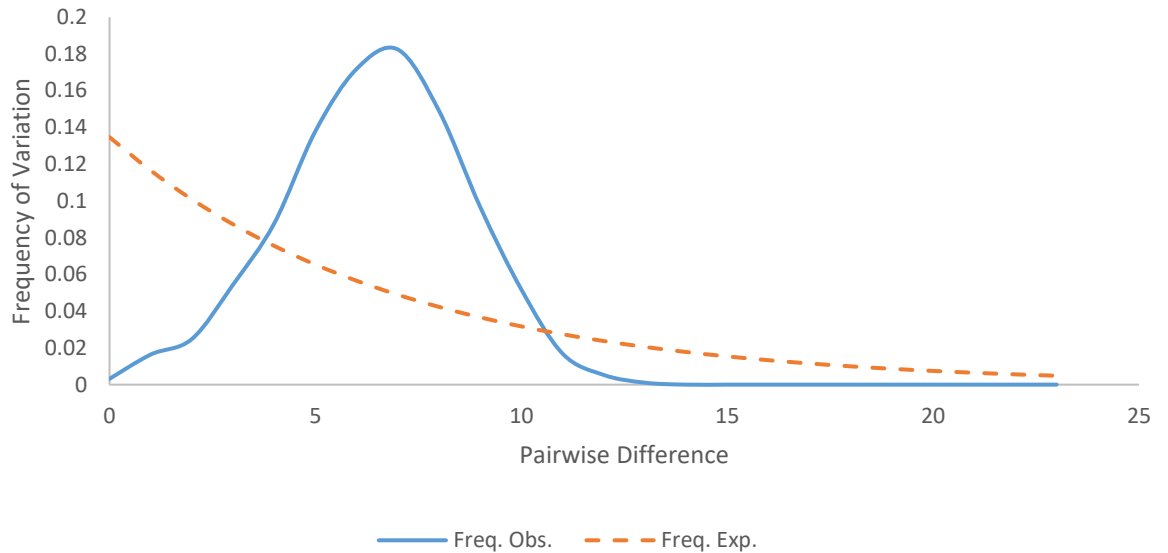


Figure 6.4: Pairwise mismatch analysis of the TLR2 gene fragment showing the observed and expected frequencies of nucleotide variation.

### 6.3.1v: Recombination

Recombination was detected by both DnaSP and GARD (Datamonkey) though the number of breaking points was found to differ. DnaSP predicted 10 recombination breaking points across the TLR2 gene fragment ( $R_m = 10$ ) though the recombination parameter pre gene was found to be much higher ( $R = 619$ ). Recombination was also detected by GARD which indicated two sites (nucleotides 1421 and 1718) of recombination within the gene, though a total of 21 potential breakpoints were observed.

Table 6.3: All breaking points detected by DnaSP (V6), the nucleotide position and substitution are shown, as is the nature (synonymous or non-synonymous) of the mutation. The inferred region of recombination is also given between two nucleotide sites.

Nucleotide Position	Nucleotide Substitution	Synonymous or non-synonymous	Between Sites
<b>1,128</b>	T → C	Synonymous	1,128 – 1,344
<b>1,344</b>	T → C	Synonymous	1,344 – 1,386
<b>1,395</b>	G → A	Synonymous	1,395 – 1,512
<b>1,527</b>	G → A	Synonymous	1,527 – 1,554
<b>1,680</b>	T → C	Synonymous	1,680 – 1,719
<b>1,719</b>	G → A	Synonymous	1,719 – 1,806
<b>1,824</b>	T → C	Synonymous	1,824 – 1,923
<b>1,923</b>	C → T	Synonymous	1,923 – 2,496
<b>2,496</b>	C → T	Synonymous	2,496 – 2,667
<b>2,667</b>	A → G	Synonymous	2,667 – 2,760

### 6.3.2: The impacts of genetic variation on structure, functionality, and selection

#### 6.3.2i: Protein structure and functionality

The identification of a single non-synonymous mutation within the TLR2 gene fragment suggested that structural variation within the wild *G. m. morsitans* TLR2 gene would be minimal. Initial modeling results for the two TLR2 variants showed little structural variation, with both variants illustrating the expected concave 'horseshoe' shape (Table 6.4). The common wild type TLR2 structure comprised of two groups of parallel  $\beta$ -sheets forming the concave surface of the protein, the largest of these consists of 11 sheets running from the N-terminal to the midpoint of the sequence. The second group consists of six  $\beta$ -sheets and terminates at the C-terminal of the protein fragment. A total of five  $\alpha$ -helices were observed within the fragment structure, two at either terminal with fifth situated approximately half way in the structure.

Interestingly, the TLR2-T788 variant exhibited two potential structures with similar statistical significance (Table 6.4). While the fundamental structure of these two models is identical to that observed in the TLR2 structure, the composition of each varies slightly. The number of  $\beta$ -sheets comprising the concave surface of the protein varies between the two variants with TLR2-T788(a) exhibiting groups of 11 and five  $\beta$ -sheets, compared to the groups of 12 and seven observed in TLR2-T788(b). The number and position of  $\alpha$ -helices on the convex surface also varies with TLR2-T788(a) exhibiting a total of eight helices, while TLR-T788(b) exhibiting four (Table 6.4).

While, the T788 variation had some minimal impact on the structure of the TLR2 protein, no variation was observed in encoded domains. As described in chapter 2, *G. m. morsitans* TLR2 consists of 25 LRRs preceding the transmembrane and TIR domains, the gene fragment amplified was found to encode 14 complete and one partial LRR domains. These 14 LRR domains was found to match the position of those identified with the *G. m. morsitans* TLR2 protein in Chapter 2, suggesting that the T788 variation has no functional impact on the TLR2 protein. This observation was also found to be neutral by PROVEAN with a score of -0.186.



Table 6.4: Predicted 3-Dimensional structures of each of the *G. m. morsitans* TLR2 variants. The name of each variant, the haplotypes exhibiting each structure and the amino acid substitutions responsible for the variation are given. PDB files were produced using the I-TASSER server (Yang and Zhang, 2015) and visualised in PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre). All coil structures are shown in green,  $\alpha$ -*helices* are shown in red and  $\beta$ -sheets in yellow with direction depicted by the arrow.

Variant	Haplotypes	Amino Acid substitution	Structure
TLR2	1-9, 11-48 and 50-56	N/A	
TLR2-T788(a)	10 and 49	A788 → T788	
TLR2-T788(b)	10 and 49	A788 → T788	

### 6.3.2ii: Inferred natural selection

Initial screens for selective pressures were conducted using a Z-test in MEGAX (Kumar *et al.*, 2018), rather surprisingly this indicated that purifying selection was occurring between all haplotypes with a high statistical significant ( $P < 0.05$ ), while neutrality selection was also detected between most all haplotype. There was no significant indication of positive selection between any haplotypes, though haplotypes 10 and 49 showed insignificant signs of positive selection ( $P > 0.1$ ).

The detected purifying selection was further investigated using the codon-based methods: HyPhy, FEL, FUBAR, MEME and SLAC. As described in Chapter 4, these methods used the  $dN - dS$  statistical test to screen of signs of positive and purifying selection. As expected, multiple points of purifying selection were detected across the gene fragment, with many sites being identified by more than one method (Fig. 6.5), interestingly one point of statistically significant positive selection was also detected. FEL detected 15 statistically significant sites of pervasive purifying selection ( $P \leq 0.05$ ), of which ten were found to be highly significant ( $P \leq 0.01$ ). SLAC identified 13 significant sites of purifying selection ( $P \leq 0.05$ ), with eight being significant to  $P \leq 0.01$ . FUBAR identified 18 sites of purifying selection, with posterior probabilities greater than 0.9. Furthermore, 16 of these sites were still significant the posterior probability was increased to 0.99. Interestingly, FUBAR also identified one significant point of positive selection (posterior probability = 0.9) at codon 788, though this was not supported by MEME which identified no significant sites of positive selection.

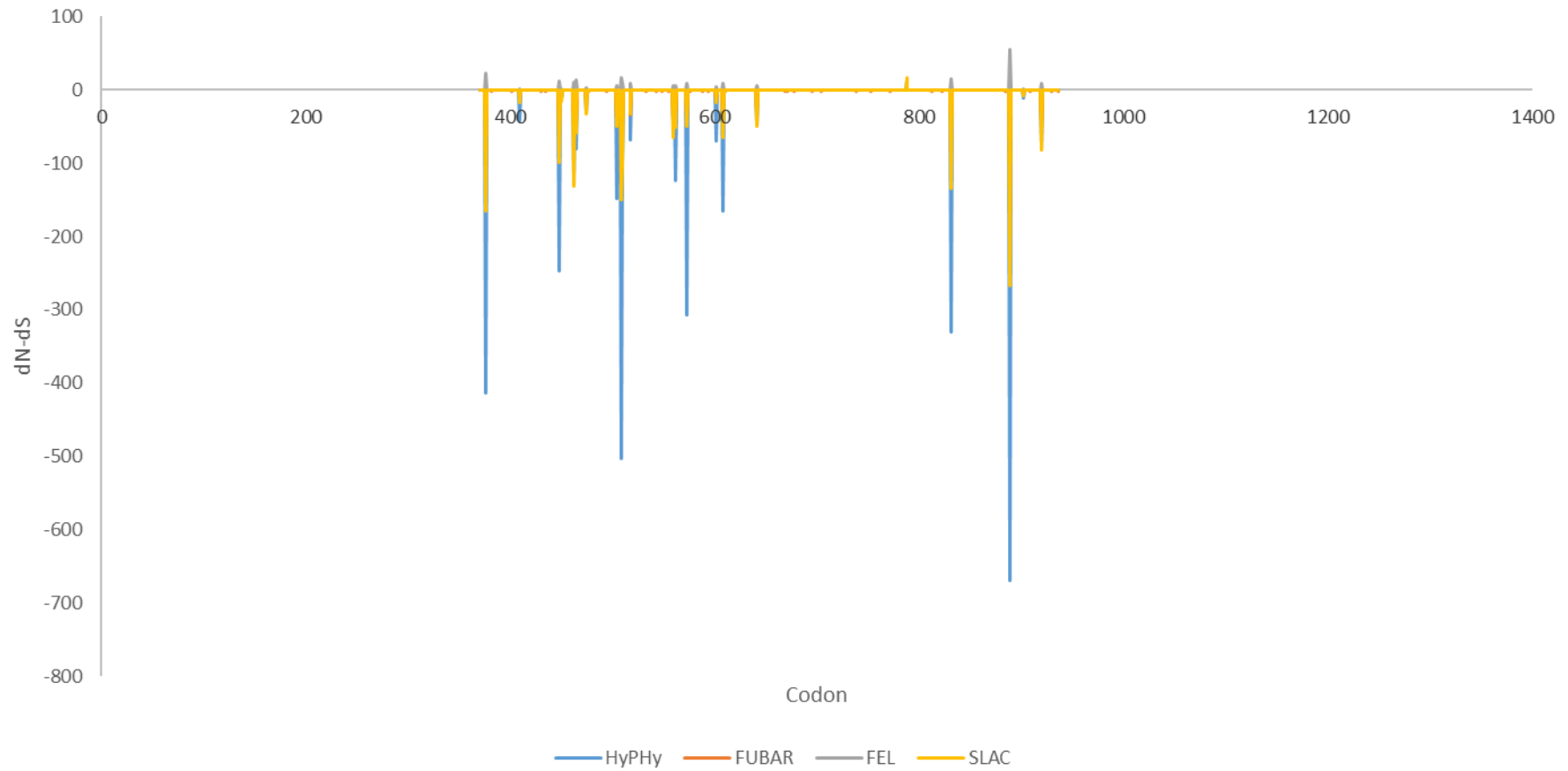


Figure 6.5: A graph showing the dN – dS values at each codon of the TLR2 sequence fragment. Positive values indicate an overabundance of non-synonymous mutations and positive selection, while negative values indicate purifying selection and an abundance of synonymous mutations. Each line represents a different methodology.

### 6.3.3: *Wigglesworthia* endosymbiosis and *Trypanosoma* infection

#### 6.3.3i: *Wigglesworthia* endosymbiosis and TLR2 genetic variation

Fifty-six TLR2 haplotypes were observed within the sample population (section 6.3.1ii) compared to the 15 *W. glossinidia 16S* haplotypes observed in the 34 successfully amplified samples (Chapter 4, section 4.3.3i), as such 26 TLR2 haplotypes contained no identified *W. glossinidia 16S* haplotype (Fig. 6.8A). Of the six shared TLR2 haplotypes (Haps 15, 28 and 49) one was exhibited by a single corresponding *W. glossinidia 16S* haplotype, while the remaining three (Haps 8, 13 and 29) contained two identified *16S* haplotypes (Fig. 6.6A). No direct relationship was observed between the two genes, which was supported by the results of a simple AMOVA ( $\phi_{st} = -0.00927$  ( $P = 0.587$ )). *Wigglesworthia glossinidia 16S* haplotype 2 was also absent from the analysis as the corresponding sample did not produce a full TLR2 gene fragment.

When samples with no corresponding *W. glossinidia 16S* haplotype were removed from the analysis, the network formed four distinct clusters (Fig. 6.6B). Cluster I consisted of five TLR haplotypes (Haps 28, 29, 30, 42 and 43), totalling six samples, of which two thirds were found to exhibit *W. glossinidia 16S* haplotype 12. Cluster II consists of nine single sample TLR2 haplotypes (Haps 23, 33, 34, 39, 41, 45, 46, 49 and 50), again two thirds of these samples were found to exhibit *W. glossinidia 16S* haplotype 12. However, this percentage drops in Clusters III and IV, Cluster III consists of seven TLR2 haplotypes (Haps 2, 3 and 6-10) exhibited by eight samples, of which only half were found to exhibit *W. glossinidia 16S* haplotype 12. Finally, cluster 4 contains six TLR2 haplotypes (Haps 11, 13, 14, 16, 20 and 31) exhibited by a total of seven samples, over half of these also exhibit novel *W. glossinidia 16S* haplotypes (Fig. 6.6B). However, despite this smaller sample size a simple AMOVA still shows no significant relationship between the two haplotype groups ( $\phi_{st} = -0.06427$  ( $P = 0.817$ )).

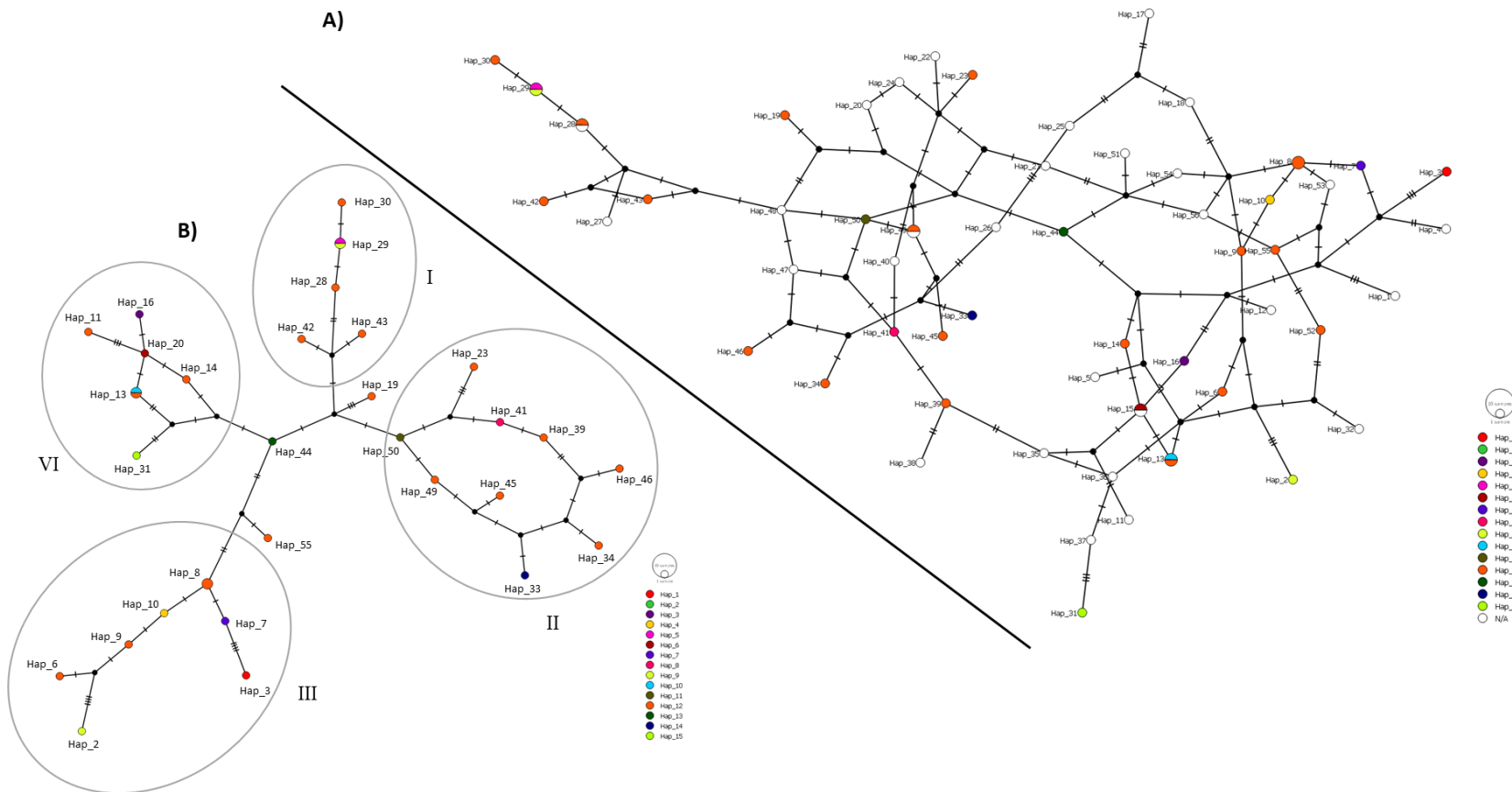


Figure 6.6: TCS haplotype networks for the *G. m. morsitans* TLR2 gene, showing the frequency of *Wigglesworthia 16S* haplotypes within the exhibited haplotypes. Produced in PopART, the circle size represents the number of samples exhibiting a haplotype, while the colour represent the presence of *Wigglesworthia* haplotypes within the samples. White filled sections represent the samples that failed to amplify during PCR, while black circles represent missing or inferred haplotypes. Black lines crossing a branch indicate the number of nucleotide mutations between haplotypes. A) Illustrates the full haplotype network adapted from the TCS network produced previously (Figure 6.2). B) Illustrates the TLR2 haplotype network when all samples that failed to produce a *W. glossinidia 16S* have been removed. Clusters are highlighted within the grey circles.

### 6.3.3ii: *Trypanosoma* infection and TLR2 genetic variation

An extensive screening of all tsetse samples for *Trypanosoma spp.* was conducted in Chapter 4 (section 4.3.4), this identified an unusually high infection percentage of 69.35% (43/62 flies) within the sample population. A high level of these were found to be mixed infection with two or more *Trypanosoma spp.* identified within a single tsetse sample. The presence of the veterinary important *T. vivax* was confirmed via sequencing; while *T. brucei* was not confirmed via sequencing, gel electrophoresis suggested it could be present in the sample population (Chapter 3, section 3.3.4).

A direct comparison of the TLR2 genetic variation and trypanosome infection illustrated no clear relationship (Fig. 6.7). However, given the extent of the genetic variation observed within the TLR2 gene fragment this is not surprising. This observation is supported by a simple AMOVA test that showed no relationship between genetic variation and infection ( $\phi_{st} = 0.01629$  ( $P = 0.97$ )).

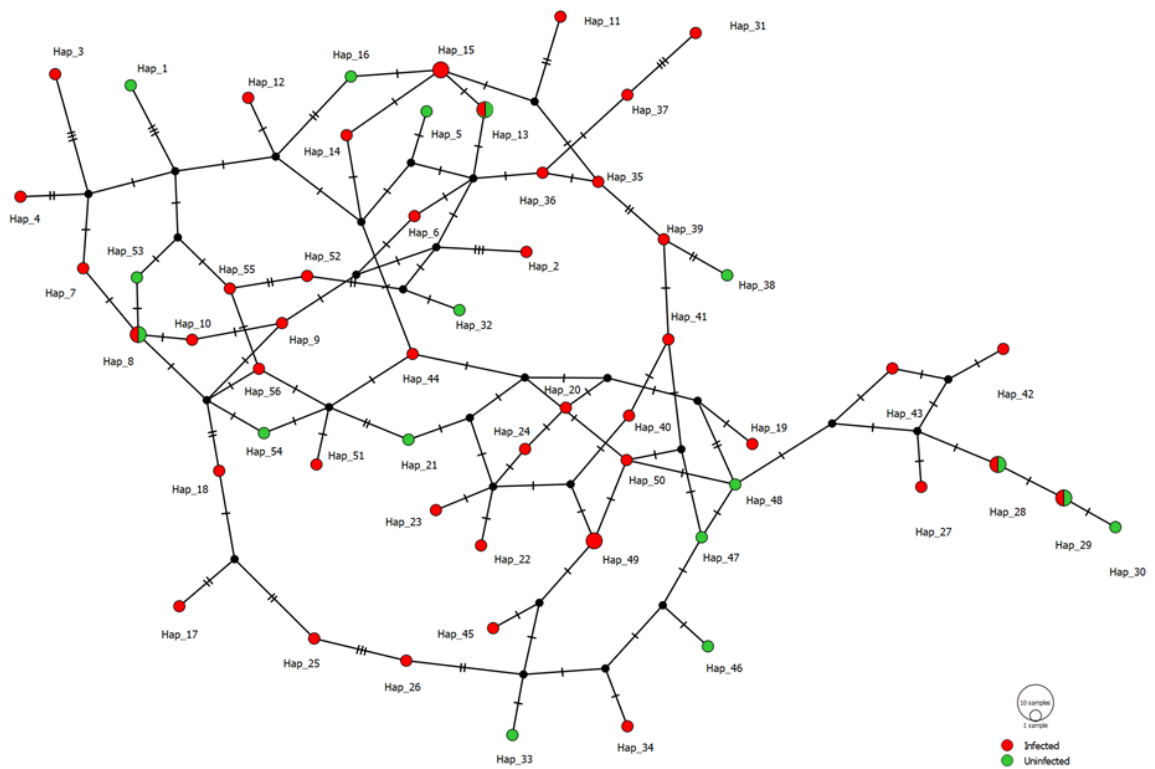


Figure 6.7: TCS haplotype network of the TLR2 gene fragment showing the frequency of infected and uninfected samples within each haplotype. Produced in PopART, the circle size represents the number of samples exhibiting a haplotype, while the colours represent the infection status of samples within the haplotype. Black circles represent inferred or missing haplotypes, while black lines crossing a branch indicate the number of nucleotide mutations between haplotypes.

### 6.3.3iii: Comparison of the TLR2 and symbiont genetic variation

The relationship between *G. m. morsitans* TLR2 and *W. g. morsitans* 16S genetic variation, illustrated similar results to those observed between *Def* and the symbiont 16S (see section 3.3.5). Although a positive correlation was observed between *Def* and *W. g. morsitans* 16S, the positive correlation observed between TLR2 and symbiont 16S was more substantial ( $R^2 = 0.114$ ) and just exceeded the statistical threshold ( $P = 0.051$ ). This positive correlation was also exhibited by a graphical comparison of genetic distance (Fig. 6.8).

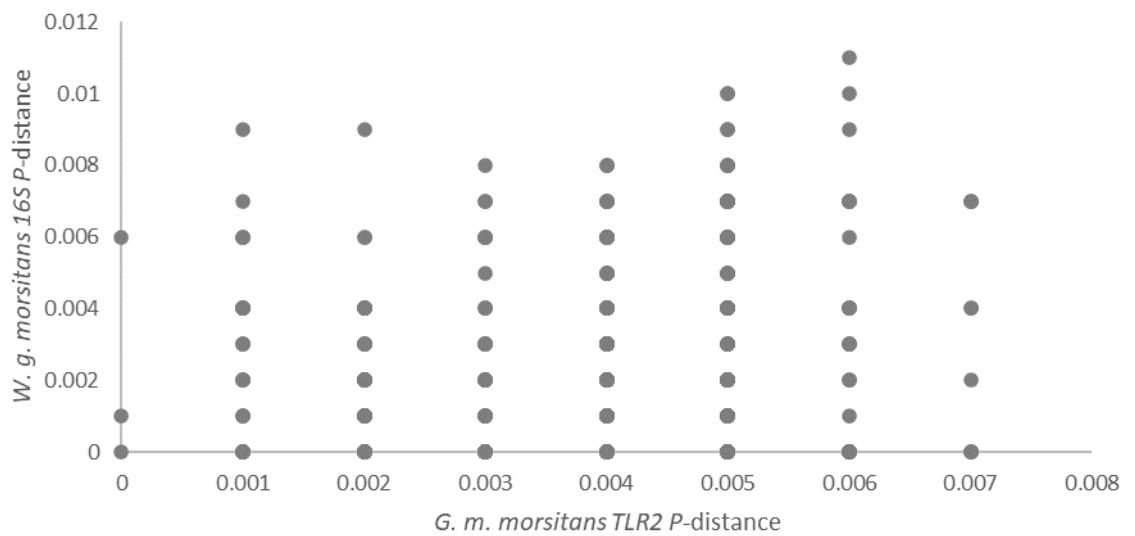


Figure 6.8: The association of genetic P-distance between TLR2 genes and successful *W. g. morsitans* amplification. P-distance was calculated in MEGAX (Kumar *et al.*, 2018), and graphs were plotted Microsoft excel.

## 6.4: Discussion

This chapter aimed to address the paucity of comprehensive data concerning the intraspecies variation and evolution of TLR genes within a dipteran species. The results presented above provide a unique insight into the genetic variation and evolution of *TLR2* within a wild *G. m. morsitans* population, and the relationship of this gene with both the endosymbiont *W. g. morsitans* and trypanosomes infection.

The LRR region of *G. m. morsitans TLR2* was found to be highly polymorphic, exhibiting a 1.29 % mutation rate (22 polymorphic sites over the 1701 bp fragment), which was found to be comparable to other TLR intraspecies studies in mammals (Ferwerda *et al.*, 2007; Netea *et al.*, 2012). This high level of polymorphic sites was further exhibited by the high number of haplotypes exhibited within the sample population. Interestingly, these result are similar to those observed by Antonides *et al.*, (2019), supporting the author's observation that allele variation maybe mediated by negative-frequency dependent selection. Additionally, the indication of balancing selection by Tajima's D seemingly supports this nucleotide diversification (Tajima, 1989).

The identification of significant purifying selection at multiple polymorphic sites within the *G. m. morsitans TLR2* gene supports the previous observation, maintaining genetic variation through purifying selection. Therefore, it can be implied that, as in *Drosophila* spp., *Glossina TLR2* is under strict Darwinian selection to eliminate nucleotide variants or promote advantageous variants (Sackton *et al.*, 2007) as no clear relationship between either specific alleles or polymorphic sites and infection rates was observed, unlike those described by Antonides *et al.*, (2019). Curiously, Sackton *et al.* (2007) also comment that positively selected sites clustered in recognition regions, such as the LRRs in TLRs, which indicated that coevolution drives the adaptive evolution of the immune system. Only one polymorphic site was inconclusively suggested to be under positive section the *TLR2* fragment, and direct comparison to pathogens requires further analysis. The employment of a larger sample population would enable the concepts of negative-frequency dependent selection and Darwinian evolution to be investigated further within *G. m. morsitans TLR2*.

The population expansion event, detailed in detail in Chapter 3, was also indicated by the results of *TLR2*. A high level of gene flow was observed between all three subpopulations,



while Fu's  $F_s$ , pairwise mismatch and raggedness all strongly indicated a population expansion event (Rogers and Harpending, 1992; Harpending, 1994; Fu, 1997). This high level of gene flow illustrate within the sample population is supported by, and likely results from, the free interbreeding observed between subpopulations. Furthermore, as described in both AMPs in Chapter 3, haplotypic analysis of TLR2 supports the observation of a population expansion, though whether the extent of the variation is a result of either balancing/negative-frequency dependent selection or a population recovery remains undetermined. Interestingly, the comparison genetic distance and geographic distance illustrated a neutral correlation supporting the observation of free interbreeding and no geographic specific variation.

Despite the high level of nucleotide variation, amino acid variation was minimal, with just a single non-synonymous polymorphism observed in two samples. This variation had minimal impact on the structure of the LRR region, though given the conservation of TLR structure this is to be expected. The impact of this mutation appears to have no functional impact, as both SMART and PROVEAN indicated no variation in protein domains or delicious functional changes. As deletions TLR variations are strictly controlled by purifying selection (Netea *et al.*, 2012) and no such selective pressure was observed at this polymorphic site, this further supports the hypothesis that *G. m. morsitans* TLR2 is evolving under balancing and negative frequency-dependent selection (Burdon *et al.*, 2013; Unckless and Lazzaro, 2016b).

As stated previously, no clear relationship was observed with trypanosome infection and TLR variation. Though given the extent of genetic variation, lack of amino acid variation and high infection rate this was expected. Equally, a similar observation was made between TLR2 and symbiont variation. However, a larger comparable symbiont population may provide more comprehensive results as currently, 46.42 % of the identified TLR2 haplotypes exhibited no corresponding 16S haplotypes.

When all samples with no identified *W. glossinidia* 16S haplotypes were removed from the analysis, some apparent relationships can be observed. Of the four observed in figures 6.8B, a possible association between genetic variation and trypanosome infection can be observed within Cluster I. The segregation of this cluster from the rest of the network by two missing haplotypes, suggests a great genetic distance between sample. Furthermore,

the lower infection rate in this cluster is indicative of the phenomenon observed by Antonides *et al.*, (2019). Perhaps most significantly however, is that is trend observed within Cluster I is, to a less degree, also exhibited in figure 6.9, suggesting the rare polymorphisms separating these samples from much of the population do lend some immunological advantage to the samples exhibiting TLR2 haplotypes 28-30. This divergence appears to be the result of C/A nucleotide polymorphism at nucleotide 1348. While this polymorphism is synonymous, there is growing evidence that synonymous mutations are not always neutral (Lebeuf-Taylor *et al.*, 2019). In this case, is possible that this polymorphism offers a unique advantage in combatting pathogen infection.

Interestingly, a positive correlation was observed between genetic distance of *TLR2* and *W. g. morsitans 16S* suggesting that divergence of one influence the other. The recognition of gram-negative bacteria by TLR2 has been recorded on multiple occasions (Ulevitch and Tobias, 1999; Holden *et al.*, 2017), as *Wigglesworthia* is a gram-negative genera of it is possible that genetic variation within the symbiont may require variation of the receptor however this is currently purely hypothetical in this system.

## 6.5: Conclusion

The principle aim of this chapter was to evaluate the intraspecies variation of Toll-Like Receptor 2 in a wild *Glossina* population, to assess the impacts of selection on TLR2, and to investigate the relationship between TLR2 diversity and endosymbiont and trypanosome infection. This chapter presents a novel insight into the nucleotide variation of the TLR2 ectodomain, high levels of nucleotide polymorphisms were observed across the protein fragment, though all but one was found to be synonymous. Given the lack of non-synonymous mutations, structural variation was minimal however, the true extent of the synonymous mutations requires further consideration. The *Glossina* TLR2 gene appears to be under negative selection to reduce genetic variation, as has been well documented previously other insects (Little and Cobbe, 2005; McCann *et al.*, 2012), suggesting that despite the expensive nucleotide diversity this is being reduced following a rapid expansion.

Furthermore, if the observation made within this chapter are indeed directly comparable to those made by Antonides *et al.* (2019), it is, to my knowledge, the first time the TLR

variation has been linked to trypanosome resistance in any dipteran species. However, considerable further research is required to confirm this speculation, though should this prove to be accurate it could present a significant development in the understanding of tsetse innate immunity.

## 7: General discussion and conclusions

The aim of this thesis was to provide novel insights into tsetse-symbiont-trypanosome interactions using evolutionary methods. Consequently, we also address the paucity of research regarding the inter and intraspecies variation of immune genes within the *Glossina* genus. To achieve this, five primary objectives were specified and addressed in chapters 2-6. Chapters 2-4 focus solely on the inter and intraspecies analysis of the AMP families, attacin and defensin: Chapters 5 and 6 focus on the *Glossina* TLR families and intraspecies and population genetics of TLR2.

Previous research indicated that the *Glossina* genome contained five attacin homologues arranged into an attacin cluster consisting of three *AttA* homologues and a single *AttB* and *AttD* gene (Wang *et al.*, 2008). The later observations of Trappeniers *et al.*, (2019) identified and characterised four of these genes within the *G. m. morsitans* genome (see chapter 2). Using these identified genes as a template, mining for attacin orthologues within the six available *Glossina* genomes identified 24 novel attacin orthologues, including the missing *AttA* homologue within *G. m. morsitans*. Additionally, a single novel defensin orthologues were identified within the six available *Glossina* genomes. The difference between a gene family cluster and an individual gene (as seen in attacin and defensin respectively), as well as the difference in nucleotide variation suggested that attacins and defensin have been subject to differing evolutionary histories and are likely maintained via varying selective pressures.

The evolutionary history of each gene was explored further in chapter 3, where the contrasting intraspecies nucleotide variation observed between *AttA*, and *Def* supported the emerging theme of differing evolutionary patterns. *AttA* exhibited low levels of nucleotide and haplotype diversity within the sample population, while *Def* showed elevated levels of variation and indications of balancing selection. Interestingly, both AMPs and *COI* indicated the presence of a recent population expansion event, resulting in high levels of gene flow and indicating interbreeding between subpopulations within a panmictic population. As expected, comparisons of wild *G. m. morsitans* mtDNA and the endosymbiont *W. g. morsitans* indicated a high level of similarity, though given the nature of the endosymbiotic lifestyle it is likely the evidence for a population expansion event must

be treated with some caution. No clear relationship was observed between tsetse AMP and symbiont genetic variation, suggesting that the two aspects evolve independently of each other (see section 7.1).

A surprisingly high infection rate (69.35 %) was observed within the sample population. Infection and AMP nucleotide variation showed no association, however, this was not the case between *W. g. morsitans* and infection, where haplotype variation indicated two hypothetical outcomes (Chapter 3, section 3.4, and section 7.1 below).

As maybe expected from the results in chapter 3, amino acid variation was also varied between AttA and Def. AttA shows five protein variants, of which the wild type AttA was common, being exhibited by 43 samples, while the other four were exhibited by the remaining eight samples. Defensin on the other hand showed 11 variants, 61.29 % of the samples, or 38 specimens, exhibited the three common variants (Chapter 4). Interestingly, two conclusions could be drawn when comparing protein variants to infection frequency: firstly, as expected, there is a positive correlation between infection and variant frequency, i.e.: as the number of samples exhibiting a variant increase, so does the number of infected samples. Secondly, there is a negative correlation between variant frequency and infection frequency, i.e.: as the number of samples exhibiting a variant increase, the percentage of infected sample decreases.

Screens for selection support the previous theme of differing evolutionary history and pressures between the AMPs. AttA illustrated marginal balancing selection, though the low protein variance does not strongly support this. Conversely, the low variation does support the hypothesis of concerted evolution within the attacin family, though whether variants are purely a result of evolution or a founding effect within an expanding population remains unclear. On the other hand, the elevated level of nucleotide and amino acid variation identified within *Def* suggest the gene was subject to positive directional selection at codon 18, and balancing selection between three protein variants (see chapters 3 and 4). The indication of positive selection and an increase in specific alleles within the sample population presents signatures of the Red Queen effect. Amino acid variation appears to have minimal impact on the AttA tertiary structure, while Def appears to show unexpectedly high level of structure variation especially in the conserved C-terminal region.

Prior to this study literature regarding TLR families within the *Glossina* genus was limited. The results within this thesis support this observations of a recent publication stating the presence of six TLR families within the *Glossina* genome (Lima *et al.*, 2021). Furthermore, this study highlighted the presence of three additional TLR-like genes. As expected, the majority of predicted TLR genes followed the established evolutionary history of other dipteran TLRs, though the exceptions to this are the three partially identified genes, TLRs 3, 5 and 13 (Chapter 5). Interspecies variation of individual TLR families indicated a high level of conservation within the predicted gene family and between similar families such as, TLRs 2/7 and 6/8. Protein structure indicated a far greater degree of variation within the full TLR proteins, though a high level of conservation was observed within partial TLR proteins.

Intraspecies variation and population genetics of the TLR2 gene showed an unexpectedly elevated level of nucleotide variation within the sample population. Of the 22 polymorphic sites identified one was non-synonymous, illustrating the expectedly high conservation frequently observed within the TLR genes. Test for gene flow, population genetics and neutrality indicated comparable results with both AMPs and *COI*. Despite the elevated nucleotide and haplotype diversity, there was evidence of interbreeding between the subpopulations following a population expansion event, and balancing selection influencing the evolution of TLR2 within the *G. m. morsitans* population. Amino acid and protein variation was negligible; the single amino acid substitution had a minor impact on predicted protein structure and no observable impact on functional domains within the receptor region of the protein. Fifteen polymorphic sites indicated statistically significant signs of purifying selection, while codon 788 indicated inconclusive signs of positive selection.

The association between TLR2 nucleotide variation, symbiont variation and infection produced inconclusive results. Given the high level of nucleotide variation within the TLR2 fragment this was expected, however the clustering of TLR2 haplotypes following the removal of samples with no corresponding symbiont sequence did present results that are more comparable to those observed within *AttA* and *Def*. Equally, comparisons of infection genetic and proteins variation and infection show no clear relationship requires further analysis.

## 7.1: Limitations of this study

The empirical nature of this research means that several limitations must be considered when interpreting the results within this study. Sampling bias and size, methodological limitations and time restraints have, to varying extent, dictated the progression and outcome of this research. This section will explore these limitations in greater detail and highlight their impact on the results and reach of this thesis.

One of the most fundamental limitations of the study of wild populations is sample size and bias, ensuring a balanced and random sample population is utilised is vital to understanding the natural population dynamic. Although only 63 wild *G. m. morsitans* specimens were utilised within this study, it is often possible to assess genetic diversity with a specific population with few samples ( $n > 20$ ). Therefore, it is likely the 63 wild *G. m. morsitans* samples used offer an accurate insight in the evolutionary and population genetic history of sample population. However, as all samples were teneral males, having consumed at least one blood meal, infection rates could be inflated. Given that juvenile male tsetse' are the most susceptible to trypanosome infection, it is possible that the unexpectedly high infection rate of is a consequence of the sample dynamics (Distelmans *et al.*, 1982; Otieno *et al.*, 1983). Although, multiple attempts were made to acquire additional samples from either the same or comparable populations it was not possible within the time available. Interestingly, the more susceptible male samples may emphasise the results in this thesis as the of potential relationships between genetic variation and trypanosome resistance would be more consequential.

The methodologies used in this study all have inherent limitations. While PCR is widely used in molecular studies it is not without limitations (Wintzingerode *et al.*, 1997; Smith and Osborn, 2009; Garibyan and Avashia, 2013). Nonspecific primer binding, whether to contaminant DNA or similar sequence fragments, within the sample can result nonspecific amplification and inaccurate results, the amplification of *AttB* in 12 of the samples in this study clearly emphasised this. Pool screening methods, such as that used to screen of *W. g. morsitans* and trypanosome species, have additional limitations and can be influenced by the ratio DNA concentrations within the sample (Boakye *et al.*, 2007). This the case of this study, a high concentration of *G. m. morsitans* DNA compared to symbiont or trypanosome DNA could overwhelm the primers making amplification ineffective.

The process of predicting protein structure is naturally difficult to accurately perform. Protein prediction can be undertaken using two primary methods: the first, '*ab initio*,' produces a model from the principles of the amino acids present in the absence of empirical evidence. The second, is Template-Bases Modelling (TBM) which aligns the query sequence to known amino acid sequences and corresponding protein structures to produce a structural prediction (Contreras-Moreira *et al.*, 2005; Zhang, 2008; Yang *et al.*, 2015). In this study two servers were used to predict protein structures: I-TASSER uses a combination of both the TMB and '*ab initio*' methods to produce protein models (Zhang, 2008; Roy *et al.*, 2010; Yang *et al.*, 2015). SWISS-MODEL on the other hand, only utilises the TMB method (Guex *et al.*, 2009; Waterhouse *et al.*, 2018). The primary limitations of the TMB method arise from inaccurate template selection and alignment, however this can be partially accounted for by providing appropriate templates for modelling. Additionally, the use of *ab initio* methods between template clusters can help to reduce alignment mismatches (Contreras-Moreira *et al.*, 2005). Although, these methods are not as reliable as laboratory methods such as crystallization structural analysis, recent developments in prediction software means that the models produced in this study likely represent a general insight into structural variation within a wild population.

The final major limitation to this study was time. While many of the practical challenges of this work could have been overcome by the development of new primers or alternative methodologies this was not possible in the time scale of this study. As such, the results generated herein provide the first insights into an evolutionary approach to understanding the interactions of the tsetse-symbiont-trypanosome triplet.

## 7.2: Impacts, implications, and future research

The results presented within this thesis lay the foundations for further evolutionary and genomic studies into the interactions between wild tsetse, trypanosomes, and symbionts, and for the investigation of novel genetic control methods for African trypanosomiasis. Future research in this area can be separated into three primary areas: Population genetics, coevolutionary and protein interactions.

Understanding the population genetics of wild tsetse flies is essential to controlling the spread of vector species as well as highlighting the impact and effectiveness of recent and



ongoing control programs. As highlighted by this study and previous publications, tsetse populations can rebound following the termination of control measures (Turner and Brightwell, 1986; Hargrove, 2000; Shereni *et al.*, 2016; Shereni *et al.*, 2021), The indication of population expansion event in the Hurungwe region of Zimbabwe through molecular techniques, supports the increasing capture rate of *G. m. morsitans* and *G. pallidipes* during a census within the vicinity of Makuti between 2005 and 2019 (Shereni *et al.*, 2021). However, whether this is a consequence of environmentally stimulated population migration (climate change), reinfestation following the termination of control measures in 2001 (Shereni *et al.*, 2016) or a combination of both remains undetermined. Shereni *et al.*, (2016) stated further that recorded cases of HAT in the Hurungwe region had increased from zero between 1994 and 2004 to 28 between 2005 and 2015. Although at the time of writing, the authors stated that the exact cause of this resurgence was unknown, this time scale directly coincides with increase in vector population observed by Shereni *et al.*, (2021) and reinforced by results within the study.

The evidence of tsetse reinfestation and the considering re-emergence of HAT within the collection area clearly emphasise the necessity for continued observation and control measures within high-risk areas. The ability of tsetse populations to recover from intensive population control strategies over the course of approximately 15 years, not only presents implications of the Control of African trypanosomiasis, but for all vector borne pathogens and exhibits an effective reminder for the need to develop novel control methods. Current control methods have two primary shortcomings as described in Chapter 1. Firstly, the cost: the ensuing cost of this continuous control has a large impact on the economy of countries within the tsetse belt. In Zimbabwe alone, the cost of treating the ~30,000 Km<sup>2</sup> currently infested by tsetse flies over a 20 year period, with no discount, would be between US\$26,820,000 for insecticide treated cattle and US\$349,980,000 for 10 traps/ Km<sup>2</sup> (Shaw *et al.*, 2013; Shereni *et al.*, 2021). And secondly, the ecological impact. As targeted control methods (i.e.: tsetse specific trapping) is often one of the more expensive control methods, the use of insecticide treated cattle and sprays are commonly employed. Given the generic nature of these treatments it is inevitable that other insect populations will be affected by this form of control. Increasing global concern over the decline in insect populations (Forister *et al.*, 2019; Simmons *et al.*, 2019), partnered with the economic impact of prolonged control requires novel methods be explored.

One alternative method is the genetic control of trypanosomes, using the innate immune response of tsetse to *Trypanosoma* spp. to eliminate the parasite before transmission to the mammalian host. To the author's knowledge this is the first study to examine the evolutionary history of a wild tsetse population, as well as the potential impact on trypanosome infection and symbiont variation. To date studies into the interaction between tsetse, symbionts and pathogens have been conducted on lab-bred colonies, while the effect of nucleotide variation has not been considered.

This study presented several novel observations that expand the current understanding of tsetse genomics and trypanosome-tsetse-symbiont interactions. One of the primary aims of this thesis was to identify and characterise select immune genes within the *Glossina* genomes to provide a foundation for future genomic research. The identification of 24 attacin, six defensin novel orthologues within the six *Glossina* genomes reinforces the observation of Hao *et al.* (2001) and Wang *et al.* (2008), who first described the attacin and *Def* gene families within the *Glossina* genus. This study demonstrates that the attacin cluster within the *G. m. morsitans* genome is replicated throughout the *Glossina* genus, albeit with slight structural variation between species. The previous description of *Glossina Def* by Hao *et al.* (2001) as an individual gene within the genome was substantiated, identifying a single *Def* coding gene within each of the six *Glossina* genomes examined. Furthermore, the conformation of six TLR family orthologues within the *Glossina* genome supports the observation of TLR1, 2 and 6-9 with species of the Palpalis and Fusca group (Lima *et al.*, 2021). Potential orthologous for *Drosophila* TLRs 3 and 5, and the Muscidae TLR13 were also identified though further research is required to characterise these genes fully. The methodologies utilised for gene identification in this study could be employed across all genomes to characterise genes of interest and provide a more in-depth analysis of genes of interest.

The evolutionary history of *AttA* and *Def* differs at a fundamental level. The presence of tandem repeats of the *AttA* within the attacin cluster and the low levels of intraspecies nucleotide variation could be indicative of either a mariner transposition (Wang *et al.*, 2008) or concerted evolution within the attacin gen family (Liao, 1999) (see chapters 2 and 3), though it has been impossible to fully explore either of these hypothesis within this study. Conversely the individual *Def* gene identified within the *Glossina* genome, suggests that Darwinian selection strongly influences the evolution of *Def* (Sackton *et al.*, 2007). The

evolutionary conservation of TLR genes has been well documented and is maintained via purifying selection similar to that observed within *Def*.

While this fundamental evolution analysis is vital to understanding genes, the functional characteristics of *AttA* and *Def* remain unclear. The mode of action of *AttA* has never been fully documented and this study presents the first in depth structural analysis. It has been eluded that *AttA* acts in a similar method to CPPs, penetrating the cell membrane interrupting internal pathways (Bulet *et al.*, 1999), the structures observed within this study show some support for this given their similarity to the structure of the CPP penetratin (Magzoub *et al.*, 2001, 2002; Su *et al.*, 2008; Eiríksdóttir *et al.*, 2010), though this remain unconfirmed. If *AttA* is to be considered as potential genetic control target, considerable further research is required understand the structure and function. Protein crystallisation would provide considerably more insight into the structure of *AttA* and would in turn provide additional information into the mode of action.

Information on the structure defensin is considerably more abundant however, functional information remains theoretical. However, perhaps the most fascinating aspect of the results within the study is the observation and influence of balancing and positive frequency dependent selection on the Thr18 isoforms of *G. m. morsitans Def*. This has several implication for both coevolution potential genetic control targets. Firstly, these results are a clear indication of an existent arms race between the *G. m. morsitans* population and pathogens. The adoption of an apparently advantageous allele into the population to reduce infection rates is indicative of the Red Queen arms race and ongoing evolution. Secondly, this presents a potential target for novel genetic controls of trypanosomiasis. The indication of a significantly lower infection rate of sample exhibiting the Thr18 isoform, compared to the unexpectedly high infection rate other the other isoforms suggests a natural advantage in combating trypanosome infection. However, further research is required to assess this further both *in vitro* and *vivo*. *In vivo*, the examination of a larger and mixed sex population sample could provide greater significance to the observations made in this study. Furthermore, the isolation and purification of the *Def* isoforms for testing against cultured African trypanosomes *in vitro*, would provide a definitive answer as to the difference in effective elimination of trypanosomes.

A previous study exploring the antibacterial and antiparasitic properties of the AMP Prolixicin utilised cloned recombinant bacterial colonies to express and isolate Prolixicin before exposing Gram-negative, Gram-positive and cultured *T. cruzi* to the protein (Ursic-Bedoya *et al.*, 2011). Therefore, a similar *in vitro* analysis of the Thr18 and Ser18 Def isoforms from recombinant bacterial colonies, could provide an experimental comparator to the observations made within this thesis. This would be critical to determining the potential of the *Def* gene as a novel genetic control method.

The identification of TLR2, 6 and 9 orthologues within the *Glossina* genome suggests that these genes may play an important role in the identification of trypanosomes, as they do in the detection of *T. cruzi* in triatomines (Bafica *et al.*, 2006; Uematsu and Akira, 2008; Kumar *et al.*, 2009). While the noticeable absence of TLR4 shows clear signs of genera specific evolution. The objective of this study was to establish which TLR genes were present within the *Glossina* genome, it is now critical that trypanosome specific ligands are identified, and research conducted to characterise which TLR proteins are responsible for the identification of trypanosome infections.

The influence of *Wigglesworthia* symbiosis on the *Glossina* genetic evolution and the impact of genetic variation on their relationship has not been comprehensively assessed. Our results suggest that the *Glossina* AMPs and symbionts evolve independently of each other, and that nucleotide variation within both organisms has no clear relationship. Due to time restraints and a small sample available, the relationship between symbionts and protein variation could not be explored however, initial results shows limited relationship between the success of *W. g. morsitans 16S* amplification and protein variation. While the biological reasons for this observation remain unclear, the implication for tsetse-symbiont interactions justify exploring this concept in future studies.

Interestingly, symbiont variation indicated three intriguing prospects in relation to trypanosome infection. Increased susceptibility to infection could be the consequence from two proposed processes: firstly, that symbiont variation may inhibit effective development of tsetse immune system (Kikuchi, 2009; Symula *et al.*, 2011; Weiss *et al.*, 2012; Sassera *et al.*, 2013). Secondly, that bacterial variation could result in more effective evasion of the innate immune system and thus reduce regular expression of AMPs (see Chapter 3, section 3.4 for full details). Alternatively, resistance to infection could also be

influenced by symbiont variation as suggested by the presence of two uninfected genetic distant samples (Chapter 3), though further analysis of this observation was impeded by the sample size.

This observation of potential symbiont driven resistance is vital to the concept genetic control. The implementation of symbiont mediated control has been extensively documented in control of dengue fever in *Ae. aegypti* and *Ae. albopictus* (Anders *et al.*, 2020; Bradly *et al.*, 2020; Khadka *et al.*, 2020). The use of *Wolbachia* strains was found to reduce dengue incidences by 73 % (Anders *et al.*, 2020) and provide a significant economic saving, reducing both control and treatment costs (Bradly *et al.*, 2020). While considerable further research is required to substantiate this observation within the wild tsetse population, it does present an exciting potential target for symbiont driven control.

This evolutionary approach to understanding pathogen host interactions presents novel insights into this complex topic. This is first study to present evidence of both balancing selection and the Red Queen hypothesis influencing the evolution of an immune gene within a single population. The implications of this remain to be fully determined however, further studies into the interactions of *Glossina* Def and trypanosomes would prove a unique insight in the outcome on infection rates, genetic variation, and the trypanosome population.

### 7.3: Conclusion

This thesis aimed to establish a fundamental understanding of the evolutionary history of three immune genes with the *Glossina* genus. By providing insights into the relationship between tsetse immune gene evolution, bacterial symbionts, and trypanosome infection to inform potential molecular approaches to trypanosome control. The results herein offer a novel insight into the evolution of AMPs and immune genes within the *Glossina* genus and the impacts of variation on trypanosomes infection and *W. g. morsitans* symbiosis. Establishing a solid understanding of the elementary aspects of immune evolution is critical to developing novel control methods and this thesis establishes this for three important immune genes, AttA, Def and TLR2. Furthermore, the techniques and methodologies employed in this study offer a framework for future genetic studies. Although the suitability for AttA as genetic control target remains unanswered, the possibility of employing both

Def and the endosymbiont *W. g. morsitans* in control strategies is promising. Considerable further research is required to establish the feasibility of genetic trypanosome control, however, the results presented in this study suggest that it is a very real possibility.

## Appendix

### Appendix 1: Tables of identified attacin and defensin genes

Supplementary Table 1: All predicted defensin genes within the *Glossina* species. The gene name (where applicable), scaffold, nucleotide position, introns, exons, CDS length and coding strand are given below.

Species	VectorBase Gene Name	Contig number	Nucleotide position	Exons	Introns	CDS length (bp)	Strand
<i>G. m. morsitans</i>	N/A	scf7180000644371	18,960-19,280	1	0	264	Forward
<i>G. austeni</i>	GAUT030101	Scaffold36	354,886-363,695	4	3	942	Forward
<i>G. pallidipes</i>	GPAI019770	Scaffold235	146,189-151,142	4	3	420	Forward
<i>G. f. fuscipes</i>	GFUI031425	Scaffold40	1,059,512-1,064,539	4	3	534	Reverse
<i>G. p. gambiensis</i>	GPPI029745	Scaffold228	261,990-262,253	1	0	264	Reverse
<i>G. brevipalpis</i>	GBRI000865	Scaffold0	2,773,250-2,776,202	2	1	321	Forward

Supplementary Table 2: All predicted attacin genes within the each *Glossina* species. The predicted attacin paralogue, given gene name (where applicable), scaffold, nucleotide position, introns, exons, CDS length and coding strand are given below. \* indicates partial genes.

Species	Predicted gene	VectorBase Gene Name	Contig number	Nucleotide position	Exons	Introns	CDS length (bp)	Strand
<i>G. m. morsitans</i>	AttA	GMOY01052 1	scf718000065214 9	638,573 - 639,494	2	1	627	Reverse
	AttA	N/A	scf718000065214 9	639,765 - 640,499	2	1	618	Forward
	AttA*	GMOY01052 2	scf718000065214 9	661,930 - 663,204	2	1	228	Reverse
	AttB	GMOY01052 3	scf718000065214 9	663,489 - 664,306	2	1	160	Forward
	AttD	GMOY01052 4	scf718000065214 9	671,559 - 672,730	2	1	564	Forward
<i>G. austeni</i>	AttA	GAUT047990	Scaffold7	939,695 - 940,613	2	1	627	Forward
	AttA	GAUT047992	Scaffold7	936,195 - 937,092	2	1	627	Reverse
	AttA*	GAUT048001	Scaffold7	899,231 - 915,116	3	2	369	Forward
	AttB*	GAUT048006	Scaffold7	911,050 - 925,383	2	1	351	Reverse
	AttD	GAUT047991	Scaffold7	902,389 - 903,322	2	1	564	Reverse
<i>G. pallidipes</i>	AttA*	GPAI040769	Scaffold62	1,402,317 - 1,410,531	3	2	309	Reverse
	AttA*	N/A	Scaffold62	1,405,591 - 1,405,827	1	0	234	Forward
	AttA	GPAI040759	Scaffold62	1,424,965 - 1,425,723	2	1	627	Reverse
	AttB	GPAI040754	Scaffold62	1,428,374 - 1,429,213	2	1	627	Forward
	AttD	GPAI040752	Scaffold62	1,435,469 - 1,437,551	2	1	564	Forward
<i>G. f. fuscipes</i>	AttA*	GFUI014661	Scaffold1	900,431 - 908,035	2	1	291	Reverse
	AttA*	GFUI014668	Scaffold1	897,808 - 904,537	2	1	333	Forward
	AttA*	N/A	Scaffold1	932,923 - 933,640	N/A	N/A	N/A	Forward
	AttB	GFUI014658	Scaffold1	929,105 - 930,185	2	1	627	Reverse
	AttD	GFUI014660	Scaffold1	882,095 - 894,922	4	3	1,515	Reverse
<i>G. p. gambiensis</i>	AttA*	N/A	Scaffold114	885,827 - 836,405	N/A	N/A	N/A	Reverse
	AttA*	GPPI020332	Scaffold114	331,398 - 343,842	2	1	408	Forward
	AttA/AttB*	N/A	Scaffold114	365,281 - 364,822	N/A	N/A	N/A	Forward
	AttD	GPPI020339	Scaffold114	882,095 - 894,922	4	3	564	Reverse
<i>G. brevipalpis</i>	AttA*	GBRI004567	Scaffold118	537,998 - 528,505	1	0	327	Reverse
	AttA*	N/A	Scaffold118	494,995 - 495,333	N/A	N/A	N/A	Reverse
	AttB	GBRI004559	Scaffold118	502,411 - 503,419	2	1	627	Reverse
	AttD	GBRI004558	Scaffold118	365,281 - 364,822	3	2	807	Reverse



## Appendix 2: Statistical equations

### Equation 1: Nucleotide variation ( $\pi$ )

Nucleotide variation ( $\pi$ ) in a randomly mating population can be described in three ways: Nei's (1987) equations 10.5 and 10.6, defined respectively as:

$$\hat{\pi} = \frac{n}{n-1} \sum_{ij} \hat{x}_i \hat{x}_j \pi_{ij}$$

and

$$\hat{\pi} = \sum_{i<j} \frac{\pi_{ij}}{n_c}$$

where  $\pi_{ij}$  is the proportion of nucleotide differences in the  $i$ th and  $j$ th type of the sequences.  $n$  represents the number of DNA sequences being examined,  $\hat{x}_i$  is the  $i$ th type of DNA sequence in the sample, while  $n_c$  the total number of DNA sequence comparisons. However, in equation 10.6  $i$  and  $j$  indicate the  $i$ th and  $j$ th sequences rather than the type of the sequences (Nei, 1987).

The third definition is given by Nei and Miller (1990) in equation 1:

$$\hat{\pi} = 2 \sum_{i<j} \frac{\hat{d}_{ij}}{[n(n-1)]}$$

where  $\hat{d}_{ij}$  represents the estimated number of nucleotide variation between sequences  $i$  and  $j$  and  $n$  represents the number of sequences examined (Nei and Miller, 1990). This simplified method and is not used by DnaSP, though estimates using this method will be similar.

### Equation 2: Pairwise distance ( $P$ -distance)

$P$ -distance is the proportion ( $P$ ) of polymorphic sites between two sequences defined as:

$$P = \frac{n_d}{n}$$

where  $n_d$  is the number of polymorphisms and  $n$  is the total number of nucleotides in the sequence.

### Equation 3: Haplotype-based statistics (Hs)

The haplotype-based statistics (Hs) is a weighted estimate of the average haplotype diversity in a subpopulation (Hudson, Boos *et al.*, 1992), defined in equation 3a as:

$$Hs = \sum_{i=1}^L w_i H_i$$

where  $L$  is the number of locations and  $w_i$  is the weighting factor for population  $i$ .  $H_i$  is defined by Hudson, Boos *et al.*, (1992) equation 4.

### Equation 4: Nucleotide-based statistic (Ks)

Hudson, Boos *et al.*, (1992) also provide a definition for a weighted nucleotide-based statistic (Ks), defined in equation 10 as:

$$Ks = wK_1 + (1 - w)K_2$$

where  $w$  is in the interval (0,1) and  $K_1$  and  $K_2$  represent populations (Hudson, Boos *et al.*, 1992).

### Equation 5: Fixation index (Fst)

The Fixation index (Fst) was estimated using the definition given by Hudson, Slatkin *et al.*, (1992) equation 3:

$$F_{st} = 1 - \frac{H_w}{H_b}$$

Where  $H_w$  and  $H_b$  represents the mean number of differences between sampled sequences within and between subpopulations respectively (Hudson, Slatkin *et al.*, 1992).

### Equation 6: The average number of nucleotide substitutions per site between populations (Dxy)

The average number of polymorphisms per nucleotide between two populations was defined by Nei (1987) in equation 10.20 as:

$$\hat{d}_{XY} = \sum_{ij} \hat{x}_i \hat{y}_j d_{ij}$$

where the two populations are defined as X and Y, and the sample frequencies of the  $i$ th haplotype of each population is designated as  $\hat{x}_i$  and  $\hat{y}_j$ . The number of nucleotide substitutions between the  $i$ th haplotype from X and the  $j$ th haplotype from Y is denoted by  $d_{ij}$  (Nei, 1987).

### **Equation 7: The number net nucleotide mutation per site between populations (Da)**

Nei (1987) also defined the total number of polymorphisms per site between populations in equation as 10.21 as:

$$\hat{d}_A = \hat{d}_{XY} - \left( \frac{\hat{d}_X + \hat{d}_Y}{2} \right)$$

where  $\hat{d}_{XY}$  is the result of equation 10.20 (as defined above), while  $\hat{d}_X$  and  $\hat{d}_Y$  are defined by equation 10.19 (Nei, 1987).

### **Equation 8: Pairwise mismatch**

Pairwise mismatch can be calculated using one of two equations. Firstly, as  $Q(i)$  as defined by Slatkin and Hudson (1991), equation 1:

$$Q(i) = \frac{1}{1 + \theta} \left( \frac{\theta}{1 + \theta} \right)^i$$

Secondly, as defined by Rogers and Harpending (1992) equation 3:

$$\hat{F}_i \approx \frac{\theta^i}{(\theta + 1)^{i+1}} = \hat{F}_0 (1 - \hat{F}_0)^i$$

In both equations  $\theta$  is equal to  $2N\mu$ .  $N$  represents the haploid population size and  $\mu$  is the mutation rate per generation (Slatkin and Hudson, 1991).

### **Equation 9: Raggedness (r)**

Raggedness ( $r$ ) was calculated as defined by Harpending (1994), equation 1:

$$r = \sum_{i=1}^{d+1} (x_i - x_{i-1})^2$$

Where  $d$  is the maximum of differences in the mismatch distribution of  $x$  (Harpending 1994).

### Equation 10: Tajima's $D$ statistic

Tajima defined his  $D$  statistic as follows (Tajima, 1989: *equation 38*):

$$D = \frac{d}{\sqrt{\hat{V}(d)}} = \frac{k - \frac{S}{a_1}}{e_1 S + e_2 S(S - 1)}$$

Where  $a_1$ ,  $\hat{V}(d)$ ,  $e_1$  and  $e_2$  were calculated using equations 3, 35, 36 and 37 respectively within the same publication.  $k$  is average number of pairwise nucleotide differences between DNA samples as defined by equations 10 and 11.

### Equation 11: Fu's $F_s$

Fu's  $F_s$  was calculated as described by (Fu, 1997) in equation 1:

$$F_s = \ln\left(\frac{S'}{1 - S'}\right)$$

Where  $S'$  is defined as the probability of having no fewer than  $k_0$  alleles in a random sample (Fu, 1997).

### Equation 12: Genetic recombination

Genetic recombination was calculated as defined by Hudson (1987):

$$C = 4Nc$$

Where  $N$  represents population size and  $c$  is the recombination rate (Hudson, 1987).

### Equation 13: Z-test

Was conducted in MEGAX using the following equation:

$$Z = \frac{(d_N - d_S)}{\sqrt{\text{Var}(d_S) + \text{Var}(d_N)}}$$

Where  $d_s$  is the number of synonymous mutations per synonymous site, and  $d_N$  is the number of non-synonymous mutations per non-synonymous site. The variance is represented by  $\text{Var}(d_S)$  and  $\text{Var}(d_N)$ .

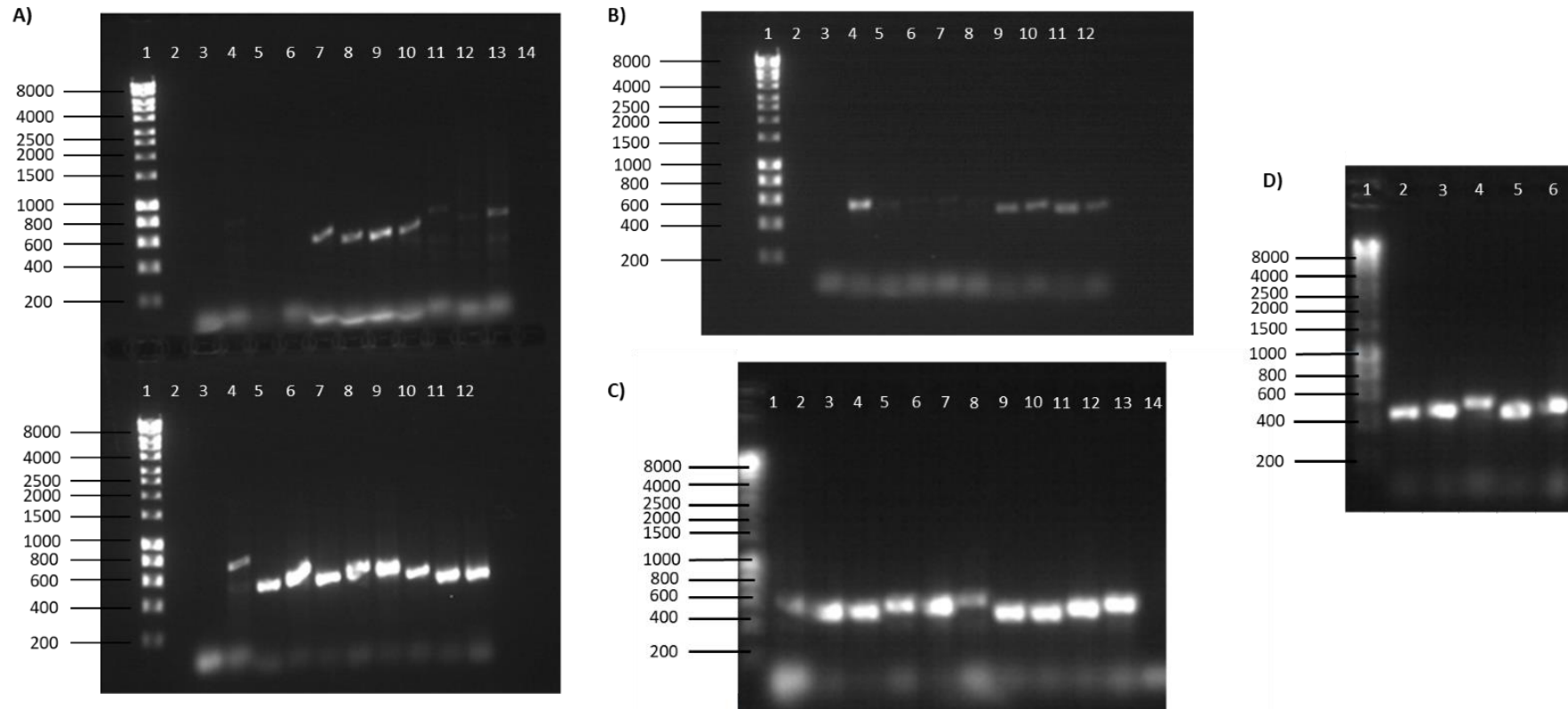
### Appendix 3: *Glossina* genomes information

Supplementary Table 3: The genome number/Assembly identification, structural variation and size of the genomes utilised in this project are given below.

Species	Genome Number/Assembly ID	Structural Annotation Version	Genome Size (Mbp)
<i>G. austeni</i>	GCA_000688735.1	GausT1.8	370.26
<i>G. brevipalpis</i>	GCA_000671755.1	GbreI1.8	315.35
<i>G. f. fuscipes</i>	GCA_000671735.1	GfusI1.8	374.77
<i>G. m. morsitans</i>	GCA_001077435.1	GmorY1.11	366.20
<i>G. p. gambiensis</i>	GCA_000688715.1	GpalI1.8	357.33
<i>G. pallidipes</i>	GCA_000818775.1	GpapI1.5	380.10

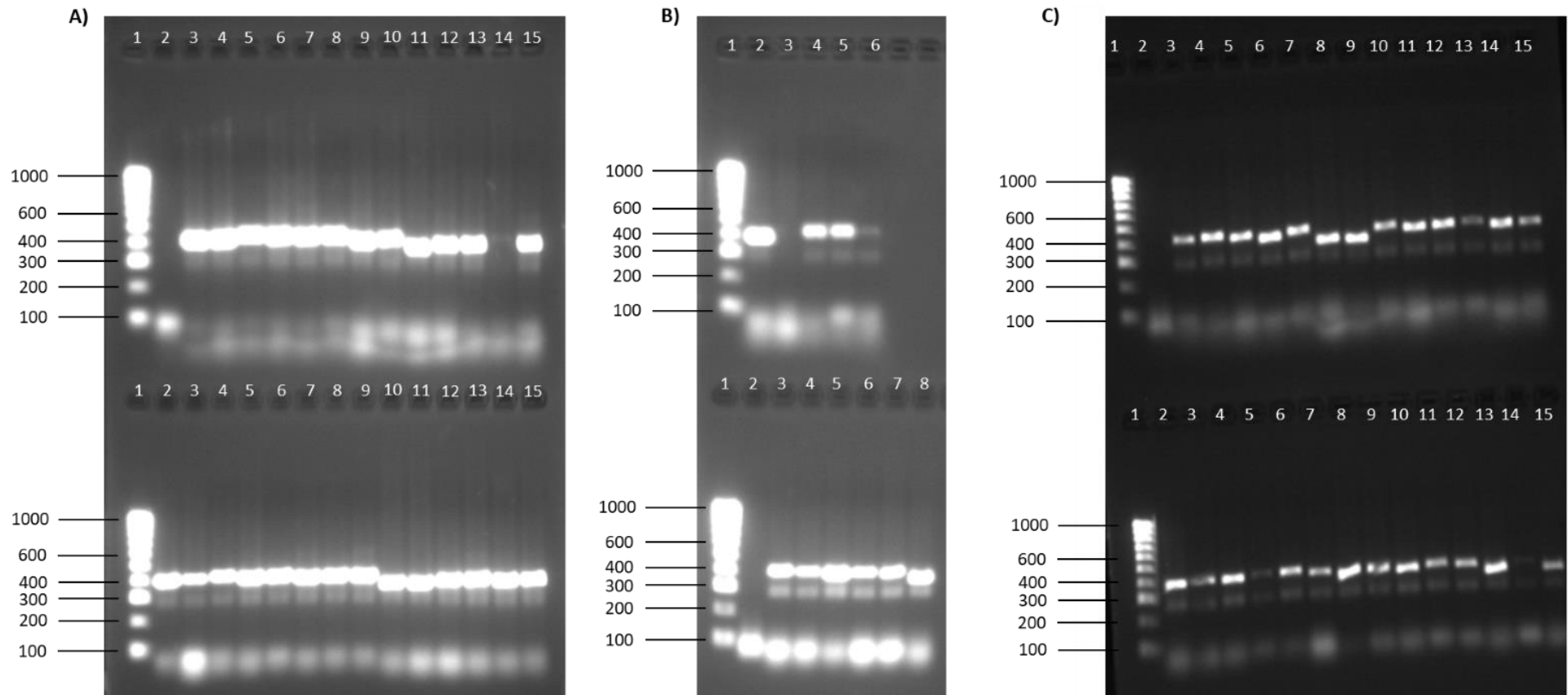
## Appendix 4: Gel images

*G. m. morsitans* AttA:



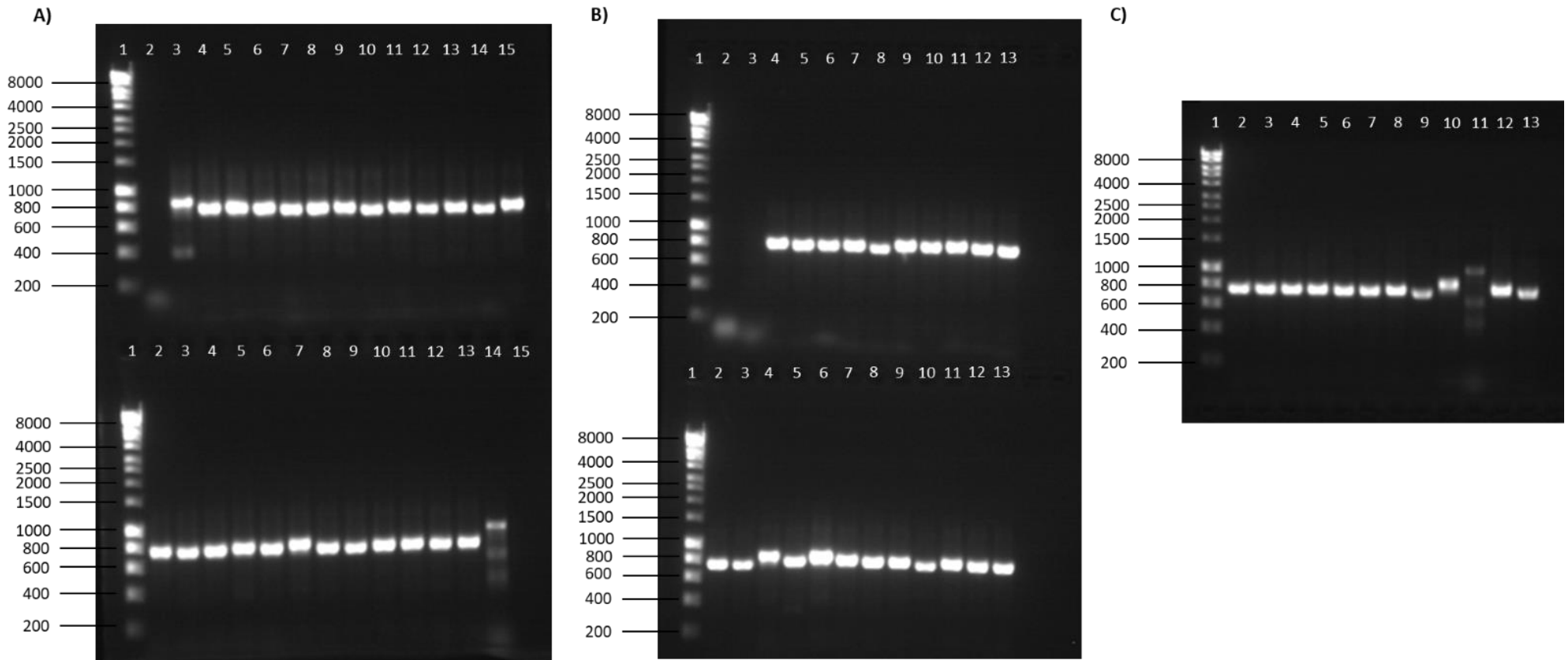
Supplementary Figure 1: Images of the gel electrophoresis results for the *G. m. morsitans* AttA PCR amplification. Gels A and B were run in this study, while gels C and D were run by Akuzike Kalizang'oma as part of his MSc dissertation. All gels were run on a 1% agarose gel using 1Kb Hyperladder (Bioline, UK) (lane 1), negative controls were run in lane 2 (Gels A and B) and lane 14 (Gel C), wild *G. m. morsitans* samples were run in the remaining lanes. The expected amplicon size is  $\approx 580$  bp.

*G. m. morsitans* Def:



Supplementary Figure 2: Images of the gel electrophoresis results for the *G. m. morsitans* Def PCR amplification. All gels were run as part of this study, using a 1.5% agarose gel and 100bp Hyperladder (Bioline, UK) (lane 1). Negative controls were run in lane 2 (Gels A and C top layer, gel B bottom layer), wild *G. m. morsitans* samples were run in the remaining lanes. The expected amplicon size is  $\approx$  400 bp.

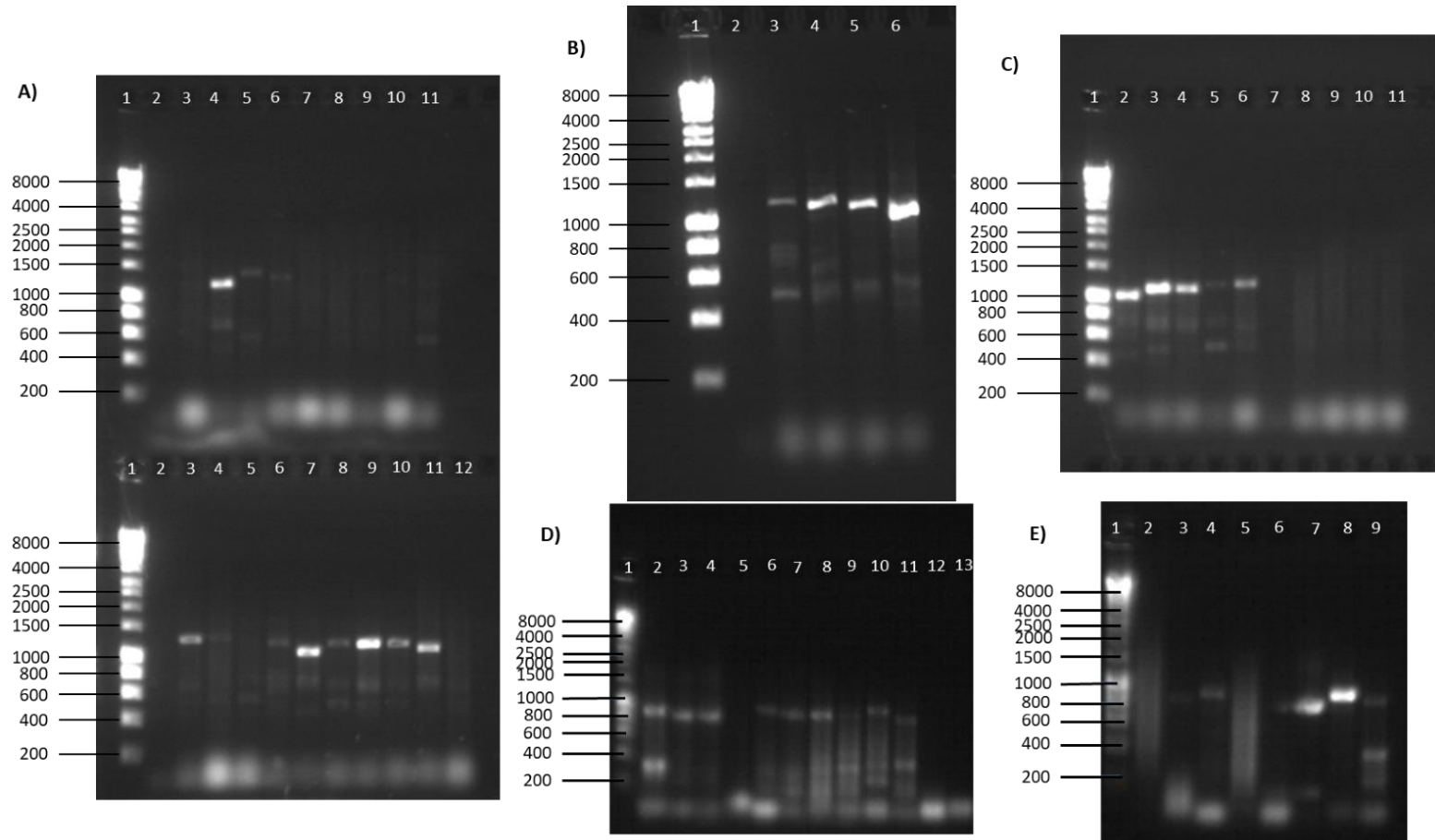
*G. m. morsitans* COI:



Supplementary Figure 3: Images of the gel electrophoresis results for the *G. m. morsitans* COI PCR amplification. All gels were run as part of this study, using a 1% agarose gel and 1Kb Hyperladder (Bioline, UK) (lane 1). Negative controls were run in lane 2 (Gels A and B top layer), wild *G. m. morsitans* samples were run in the remaining lanes. The expected amplicon size is  $\approx 780$  bp.

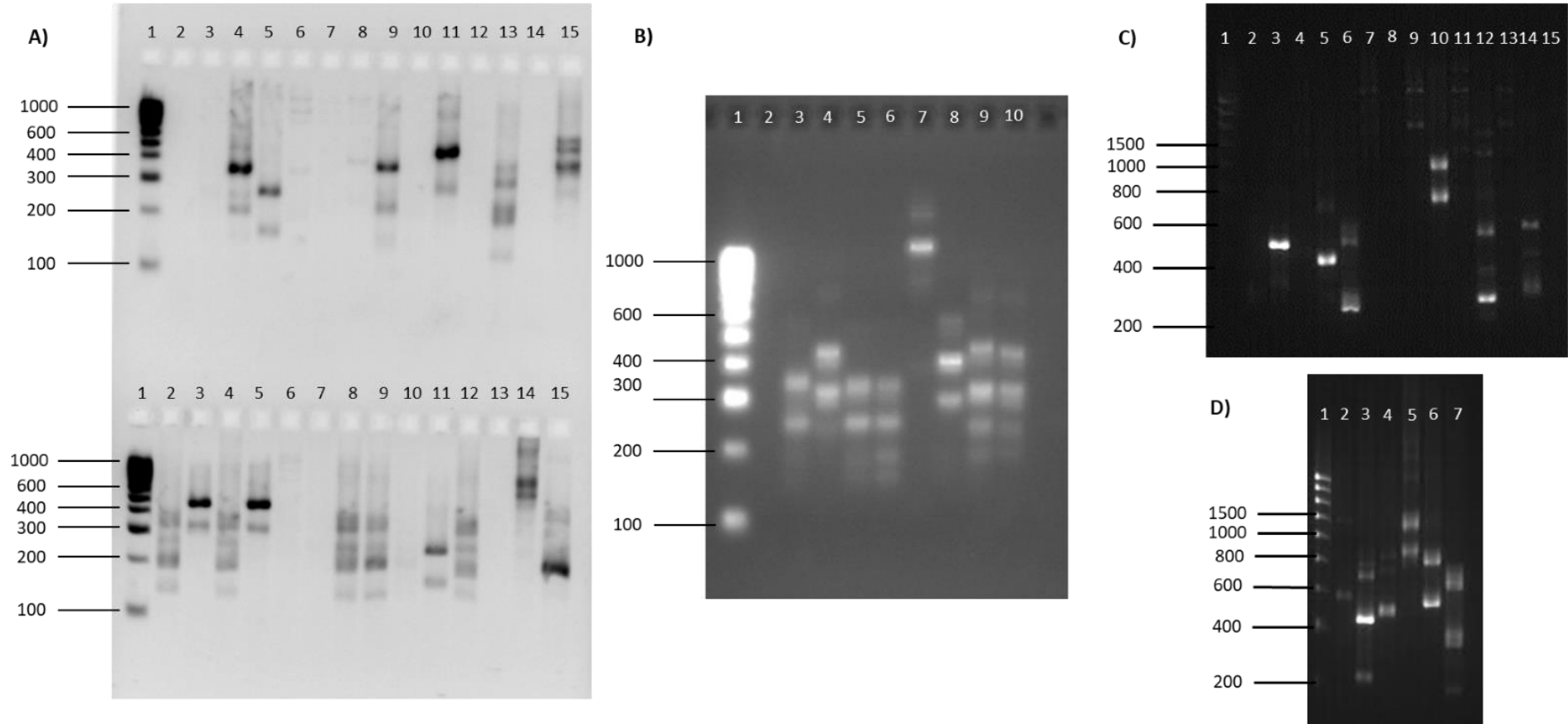


*Wigglesworthia* 16S:



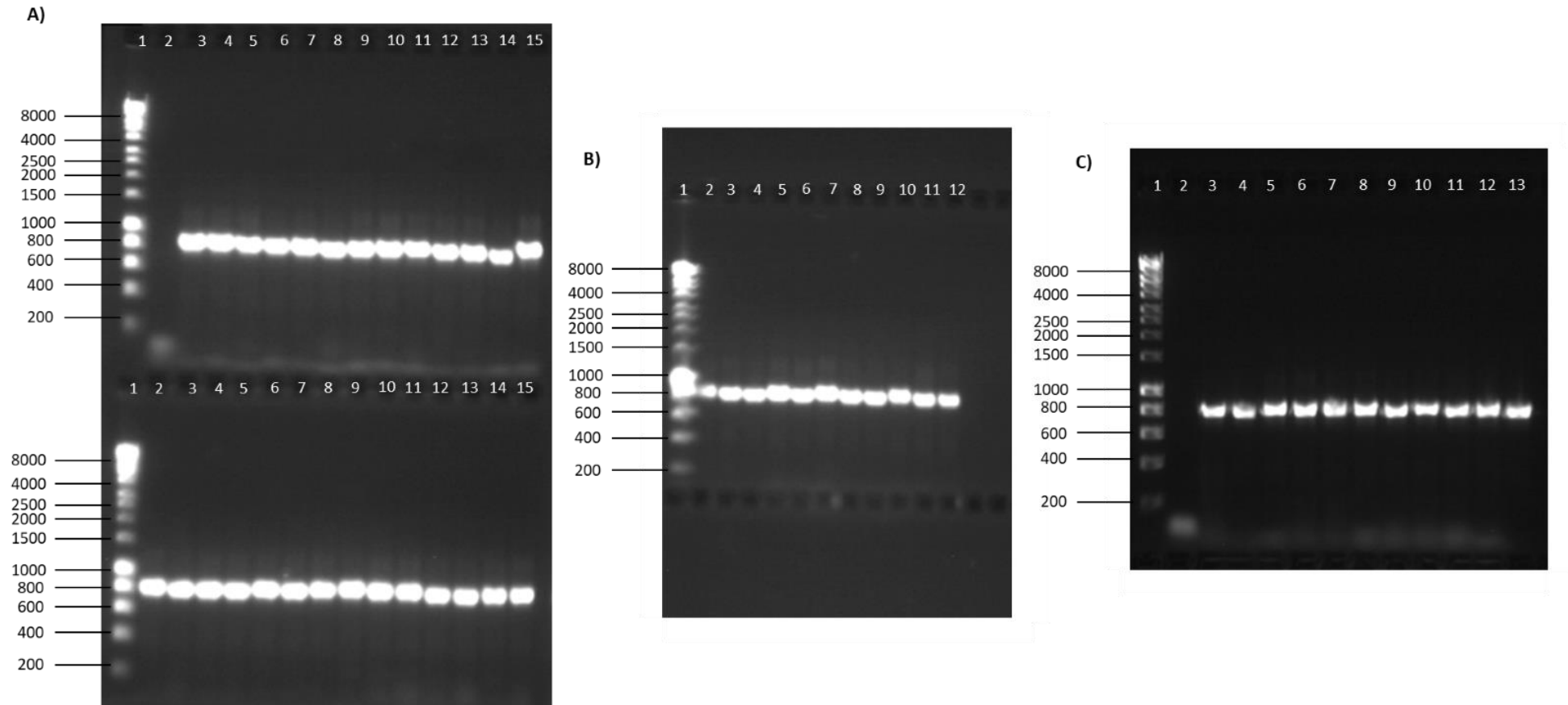
Supplementary Figure 4: Images of the gel electrophoresis results for the *W. g. morsitans* 16S PCR amplification. Gels A, B and C were run in this study, while gels D and E were run by Akuzike Kalizang'oma as part of his MSc dissertation. All gels were run on a 1% agarose gel using 1Kb Hyperladder (Biolone, UK) (lane 1), negative controls were run in lane 2 (Gel A top and bottom layers, Gel B and C) and lane 13 (Gel D), wild *G. m. morsitans* samples were run in the remaining lanes. The expected amplicon size is  $\approx$  1050 bp.

*Trypanosoma ITS:*



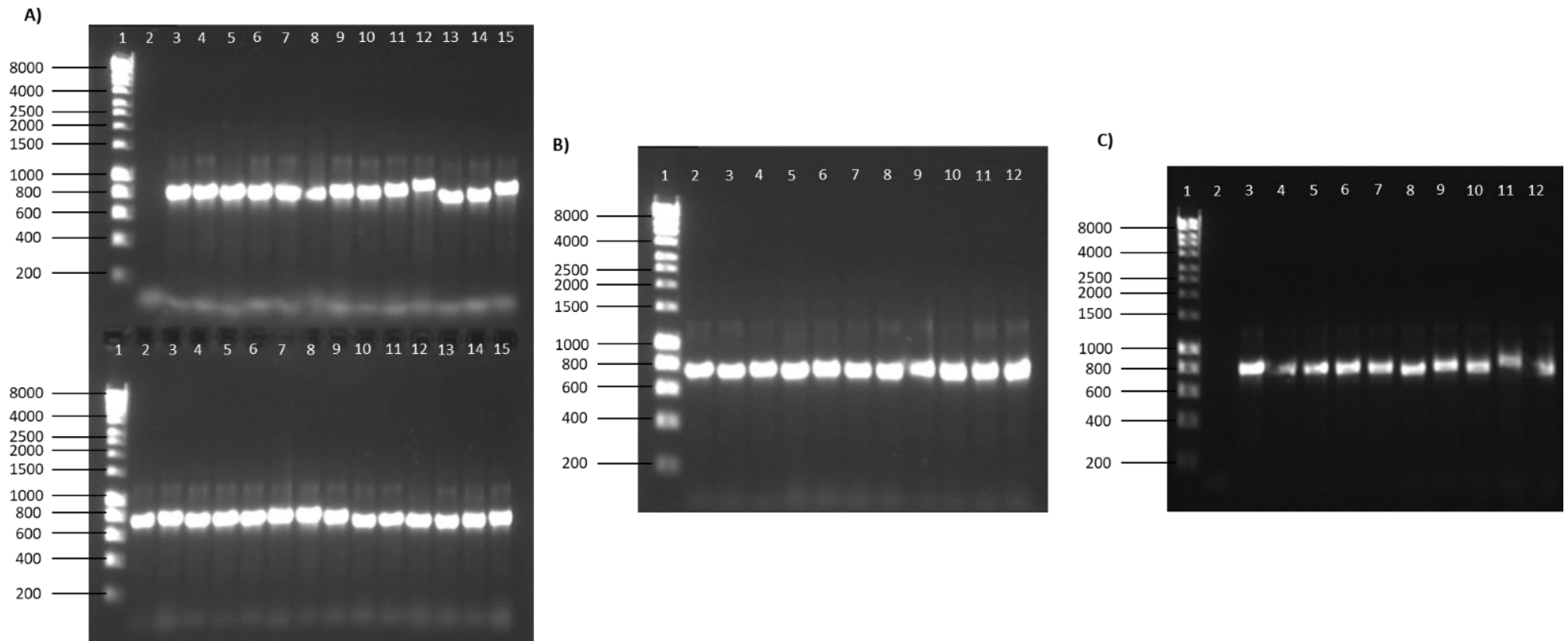
Supplementary Figure 5: Images of the gel electrophoresis results for the *Trypanosoma ITS* amplification. Gels A and B and C were run in this study, while gels C and D were run by Akuzike Kalizang'oma as part of his MSc dissertation. All gels were run on a 2% agarose gel, gels A and B used a 100bp Hyperladder (Bioline, UK) (lane 1) and gels C and D used a 1Kb Hyperladder (Bioline, UK) (lane 1). Negative controls were run in lane 2 (Gel A top layer, Gel B) and lane 15 (Gel D), wild *G. m. morsitans* samples were run in the remaining lanes. The amplicon size varied depending on the *Trypanosoma spp.* present.

*G. m. morsitans* TLR2 Fragment A:



Supplementary Figure 6: Images of the gel electrophoresis results for Fragment A of *G. m. morsitans* TLR2 PCR amplification. All gels were run as part of this study, using a 1% agarose gel and 1Kb Hyperladder (Bioline, UK) (lane 1). Negative controls were run in lane 2 (Gels A top layer and gel C), wild *G. m. morsitans* samples were run in the remaining lanes. The expected amplicon size is  $\approx$  820 bp.

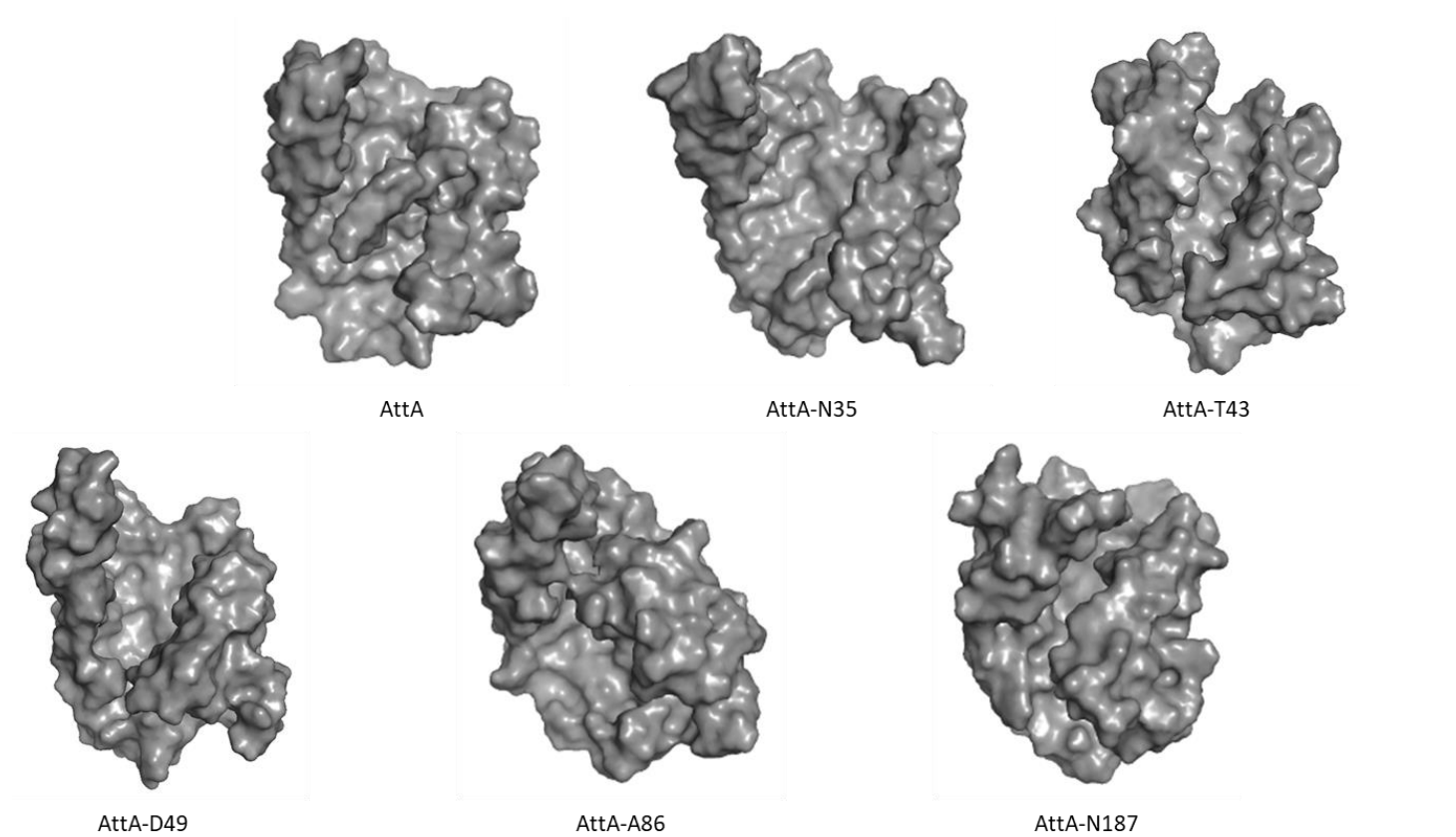
*G. m. morsitans* TLR2 Fragment B:



Supplementary Figure 7: Images of the gel electrophoresis results for Fragment B of *G. m. morsitans* TLR2 PCR amplification. All gels were run as part of this study, using a 1% agarose gel and 1Kb Hyperladder (Bioline, UK) (lane 1). Negative controls were run in lane 2 (Gels A top layer and gel C), wild *G. m. morsitans* samples were run in the remaining lanes. The expected amplicon size is  $\approx$  780 bp.

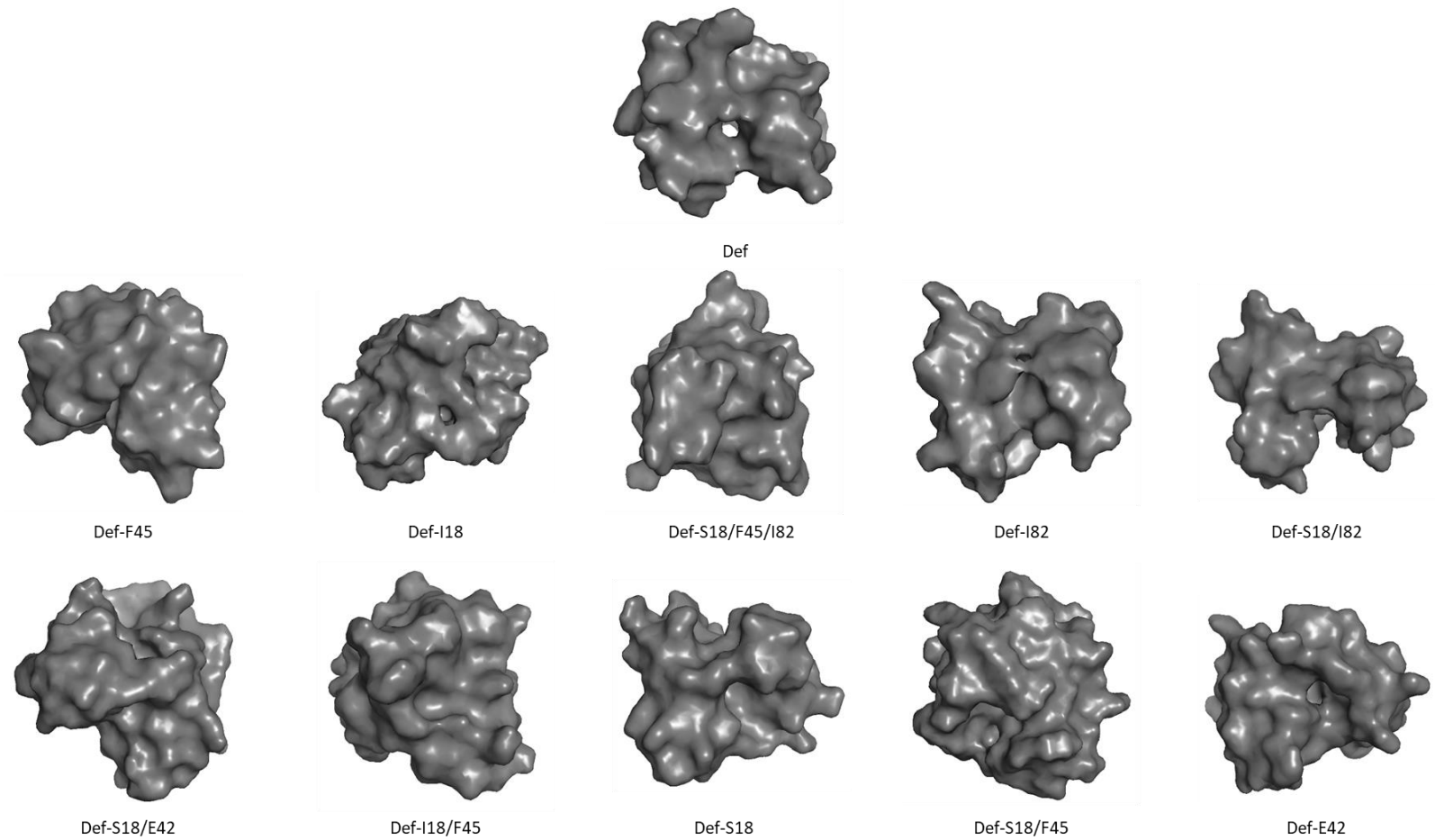
## Appendix 5: Protein surface structure

Attacin-A:



Supplementary Figure 8: The surface structure of wild *G. m. morsitans* AttA protein isoforms. All images were produced and visualised in PyMOL.

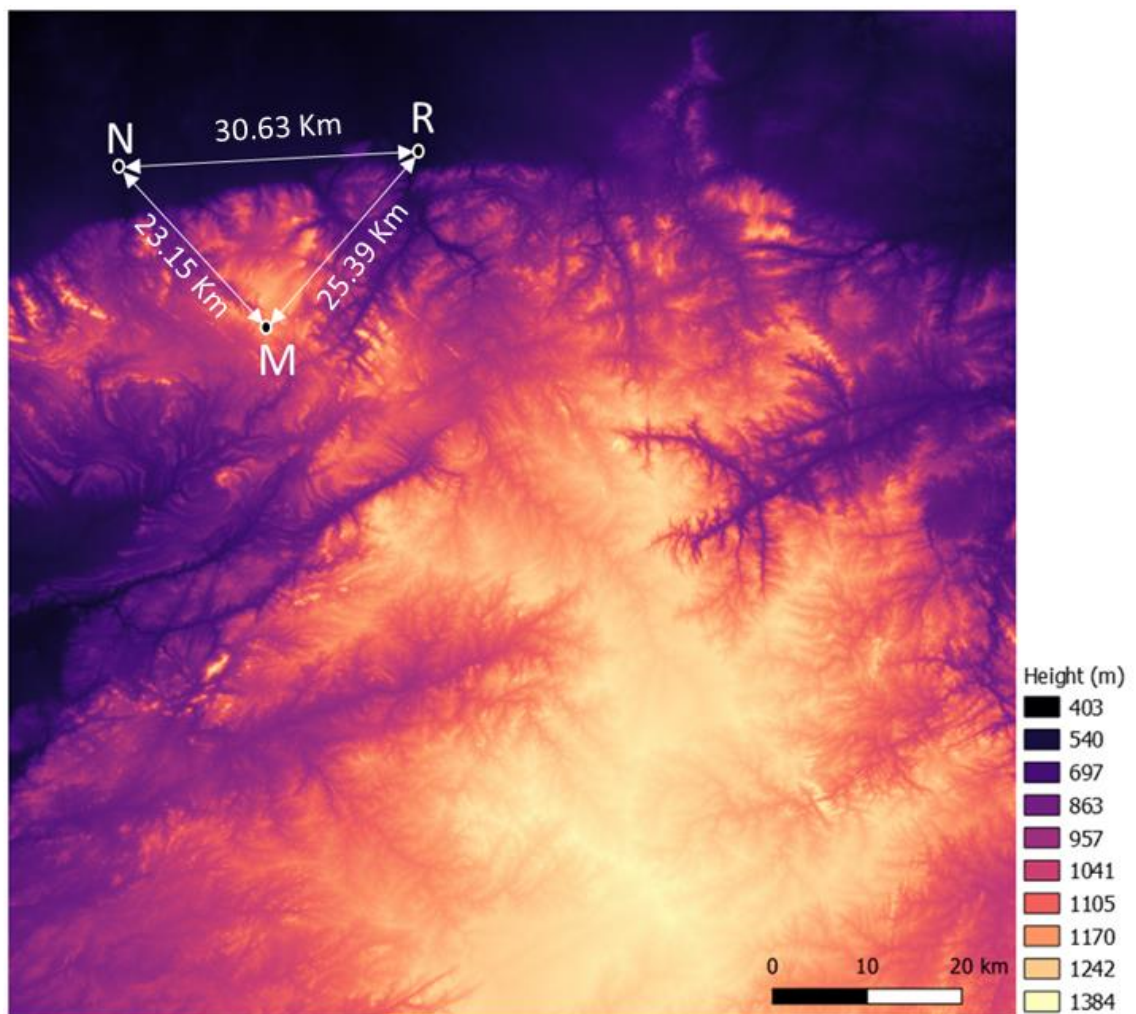
Defensin:



Supplementary Figure 9: The surface structure of wild *G. m. morsitans* Def protein isoforms. All images were produced and visualised in PyMOL.

## Appendix 6: Topology

The prediction of free interbreeding between subpopulations and a panmictic tsetse population was assessed for physical feasibility by examining the geography around the collection locations. The Shuttle Radar Topography Mission (SRTM) Digital elevation model (DEM) (Cowan and Cooper, 2005), SRTM1S17E029V3, was downloaded from United States Geological Survey (USGS) EarthExplorer. A geographical heat map, illustrating altitude and the physical geographic features was produced in QGIS (QGIS Development Team, 2018) using a Raster Data Set allowing elevation to be represented by colour.



Supplementary Figure 10: A geographic heat map illustrating height/altitude (m) and the physical geographical features in the area surrounding the three collection sites (M = Makuti; N = Nykasanga and R = Rekomitjie). The distance between each location is also given (Km). Created on QGIS using DEM data generated from the SRTM.

## List of references

- Abbot, P. and Moran, N.A. 2002. Extremely low levels of genetic polymorphism in endosymbionts (Buchnera) of aphids (Pemphigus). *Molecular Ecology*, 11 (12), pp.2649–2660.
- Abd-Alla, A.M.M., Bergoin, M., Parker, A.G., Maniania, N.K., Vlak, J.M., Bourtzis, K., Boucias, D.G. and Aksoy, S. 2013. Improving Sterile Insect Technique (SIT) for tsetse flies through research on their symbionts and pathogens. *Journal of Invertebrate Pathology*, 112 (SUPPL.1), pp.S2–S10.
- Adams, E.R., Malele, I.I., Msangi, A.R. and Gibson, W.C. 2006. Trypanosome identification in wild tsetse populations in Tanzania using generic primers to amplify the ribosomal RNA ITS-1 region. *Acta Tropica*, 100 (1–2), pp.103–109.
- Ageitos, J.M., Sánchez-Pérez, A., Calo-Mata, P. and Villa, T.G. 2017. Antimicrobial peptides (AMPs): Ancient compounds that represent novel weapons in the fight against bacteria. *Biochemical Pharmacology*, 133, pp.117–138.
- Agosta, S.J., Janz, N. and Brooks, D.R. 2010. How specialists can be generalists: Resolving the ‘parasite paradox’ and implications for emerging infectious disease. *Zoologia*, 27 (2), pp.151–162.
- Akira, S., Uematsu, S. and Takeuchi, O. 2006. Pathogen recognition and innate immunity. *Cell*, 124 (4), pp.783–801.
- Akoda, K., Van Den Bossche, P., Marcotty, T., Kubi, C., Coosemans, M., De Deken, R. and Van Den Abbeele, J. 2009. Nutritional stress affects the tsetse fly’s immune gene expression. *Medical and Veterinary Entomology*, 23 (3), pp.195–201.
- 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), pp.848–851.
- Aksoy, S., Chen, X. and Hypsa, V. 1997. Phylogeny and potential transmission routes of midgut-associated endosymbionts of tsetse (Diptera: Glossinidae). *Insect Molecular Biology*, 6 (2), pp.183–190.



- Alam, U., Medlock, J., Brelsfoard, C., Pais, R., Lohs, C., Balmand, S., Carnogursky, J., Heddi, A., Takac, P., Galvani, A. and Aksoy, S. 2011. Wolbachia Symbiont Infections Induce Strong Cytoplasmic Incompatibility in the Tsetse Fly *Glossina morsitans*. *PLoS Pathogens*, 7 (12), pp.e1002415.
- Allcock, A.L. and Strugnell, J.M. 2012. Southern Ocean diversity: new paradigms from molecular ecology. *Trends in Ecology & Evolution*, 27 (9), pp.520–528.
- Altincicek, B. and Vilcinskas, A. 2007. Analysis of the immune-inducible transcriptome from microbial stress resistant, rat-tailed maggots of the drone fly *Eristalis tenax*. *BMC Genomics*, 8 (1), pp.326.
- Anderson, R.M. and May, R.M. 1982. Coevolution of hosts and parasites. *Parasitology*, 85 (2), pp.411–426.
- Anderson, K. V., Jürgens, G. and Nüsslein-Volhard, C. 1985. Establishment of dorsal-ventral polarity in the *Drosophila* embryo: Genetic studies on the role of the Toll gene product. *Cell*, 42 (3), pp.779–789.
- Andrade, B.B., Teixeira, C.R., Barral, A., Barral-netto, M. and Reitor Miguel Calmon, A. 2005. Haematophagous arthropod saliva and host defense system: a tale of tear and blood. *Anais da Academia Brasileira de Ciencias*, 77 (4), pp.665–693.
- Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science*, 181 (4096), pp.223.
- Antonides, J., Mathur, S., Sundaram, M., Ricklefs, R. and DeWoody, J.A. 2019. Immunogenetic response of the bananaquit in the face of malarial parasites. *BMC Evolutionary Biology* 2019 19:1, 19 (1), pp.1–12.
- Arcà, B. and Ribeiro, J.M. 2018. Saliva of hematophagous insects: a multifaceted toolkit. *Current Opinion in Insect Science*, 29, pp.102–109.
- Arnot, C.J., Gay, N.J. and Gangloff, M. 2010. Molecular Mechanism That Induces Activation of Spätzle, the Ligand for the *Drosophila* Toll Receptor. *Journal of Biological Chemistry*, 285 (25), pp.19502–19509.
- Austen, E.E. 1911. *A handbook of the tsetse-flies*. British Museum (Natural History).

- Ayala, F.J. and Campbell, C.A. 1974. Frequency-Dependent Selection. *Annual Review of Ecology and Systematics*, 5 (1), pp.115–138.
- Bafica, A., Santiago, H.C., Goldszmid, R., Ropert, C., Gazzinelli, R.T. and Sher, A. 2006. Cutting Edge: TLR9 and TLR2 Signaling Together Account for MyD88-Dependent Control of Parasitemia in *Trypanosoma cruzi* Infection. *The Journal of Immunology*, 177 (6), pp.3515–3519.
- 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 (SUPPL.1), pp.S116–S122.
- Bell, J.K., Botos, I., Hall, P.R., Askins, J., Shiloach, J., Segal, D.M. and Davies, D.R. 2005. The molecular structure of the Toll-like receptor 3 ligand-binding domain. *Proceedings of the National Academy of Sciences*, 102 (31), pp.10976–10980.
- Bennett, G.M. and Moran, N.A. 2015. Heritable symbiosis: The advantages and perils of an evolutionary rabbit hole. *Proceedings of the National Academy of Sciences*, 112 (33), pp.10169–10176.
- Beschin, A., Van Den Abbeele, J., De Baetselier, P. and Pays, E. 2014. African trypanosome control in the insect vector and mammalian host. *Trends in Parasitology*, 30 (11), pp.538–547.
- Bing, X., Attardo, G.M., Vigneron, A., Aksoy, E., Scolari, F., Malacrida, A., Weiss, B.L. and Aksoy, S. 2017. Unravelling the relationship between the tsetse fly and its obligate symbiont *Wigglesworthia* : transcriptomic and metabolomic landscapes reveal highly integrated physiological networks. *Proceedings of the Royal Society B: Biological Sciences*, 284 (1857), pp.20170360.
- Bloom, J.D. and Arnold, F.H. 2009. In the light of directed evolution: pathways of adaptive protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 106 (1), pp.9995–10000.
- Bloom, J.D., Drummond, D.A., Arnold, F.H. and Wilke, C.O. 2006. Structural Determinants of the Rate of Protein Evolution in Yeast. *Molecular Biology and Evolution*, 23 (9), pp.1751–1761.

- Bolintineanu, D.S., Langham, A.A., Davis, H.T. and Kaznessis, Y.N. 2007. Molecular dynamics simulations of three protegrin-type antimicrobial peptides: interplay between charges at the termini,  $\beta$ -sheet structure and amphiphilic interactions. *Molecular simulation*, 33 (9–10), pp.809.
- Bonmatin, J., Genest, M., Petit, M., Gincel, E., Simorre, J.P., Cornet, B., Gallet, X., Caille, A., Labbé, H., Vovelle, F. and Ptak, M. 1992. Progress in multidimensional NMR investigations of peptide and protein 3-D structures in solution. From structure to functional aspects. *Biochimie*, 74 (9–10), pp.825–836.
- Bonmatin, J.M., Bonnat, J.L., Gallet, X., Vovelle, F., Ptak, M., Reichhart, J.M., Hoffmann, J.A., Keppi, E., Legrain, M. and Achstetter, T. 1992. Two-dimensional  $^1\text{H}$  NMR study of recombinant insect defensin A in water: Resonance assignments, secondary structure and global folding. *Journal of Biomolecular NMR*, 2 (3), pp.235–256.
- Booker, T.R., Jackson, B.C. and Keightley, P.D. 2017. Detecting positive selection in the genome. *BMC Biology* 2017 15:1, 15 (1), pp.1–10.
- Borst, P. and Cross, G.A.M. 1982. Molecular basis for trypanosome antigenic variation. *Cell*, 29 (2), pp.291–303.
- Boulanger, N., Brun, R., Ehret-Sabatier, L., Kunz, C. and Bulet, P. 2002. Immunopeptides in the defense reactions of *Glossina morsitans* to bacterial and *Trypanosoma brucei* infections. *Insect Biochemistry and Molecular Biology*, 32 (4), pp.369–375.
- Boulanger, N., Bulet, P. and Lowenberger, C. 2006. Antimicrobial peptides in the interactions between insects and flagellate parasites. *Trends in Parasitology*, 22 (6), pp.262–268.
- Brogden, K.A. 2005. Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nature Reviews Microbiology*, 3 (3), pp.238–250.
- Brooker, S. 2010. Estimating the global distribution and disease burden of intestinal nematode infections: Adding up the numbers – A review. *International journal for parasitology*, 40 (10), pp.1137.
- Brun, R., Blum, J., Chappuis, F. and Burri, C. 2010. Human African trypanosomiasis. *The Lancet*, 375 (9709), pp.148–159.

- Buchon, N., Poidevin, M., Kwon, H.M., Guillou, A., Sottas, V., Lee, B.L. and Lemaitre, B. 2009. A single modular serine protease integrates signals from pattern-recognition receptors upstream of the *Drosophila* Toll pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 106 (30), pp.12442–7.
- Bulet, P., Hetru, C., Dimarcq, J.L. and Hoffmann, D. 1999. Antimicrobial peptides in insects; structure and function. *Developmental and Comparative Immunology*, 23 (4–5), pp.329–344.
- Burdon, J.J., Thrall, P.H. and Ericson, L. 2013. Genes, communities & invasive species: understanding the ecological and evolutionary dynamics of host–pathogen interactions. *Current Opinion in Plant Biology*, 16 (4), pp.400–405.
- Caljon, G., De Vooght, L. and Van Den Abbeele, J. 2014. The Biology of Tsetse–Trypanosome Interactions. *Trypanosomes and Trypanosomiasis*. Vienna: Springer Vienna, 41–59.
- Campos, J.L., Zhao, L. and Charlesworth, B. 2017. Estimating the parameters of background selection and selective sweeps in *Drosophila* in the presence of gene conversion. *Proceedings of the National Academy of Sciences*, 114 (24), pp.E4762–E4771.
- Capewell, P., Cren-Travaillé, C., Marchesi, F., Johnston, P., Clucas, C., Benson, R.A., Gorman, T.A., Calvo-Alvarez, E., Crouzols, A., Jouvion, G., Jamonneau, V., Weir, W., Lynn Stevenson, M., O’Neill, K., Cooper, A., Swar, N.R.K., Bucheton, B., Ngoyi, D.M., Garside, P., Rotureau, B. and MacLeod, A. 2016. The skin is a significant but overlooked anatomical reservoir for vector-borne African trypanosomes. *eLife*, 5 (September 2016).
- Cecchi, G., Mattioli, R.C., Slingenbergh, J. and De la Rocque, S. 2008. Land cover and tsetse fly distributions in sub-Saharan Africa. *Medical and Veterinary Entomology*, 22 (4), pp.364–373.
- Chanie, M., Adula, D. and Bogale, B. 2013. Socio-Economic Assessment of the Impacts of Trypanosomiasis on Cattle in Girja District, Southern Oromia Region, Southern Ethiopia. *Acta Parasitologica Globalis*, 4 (3), pp.80–85.
- Chapman, Joanne R, Hill, T. and Unckless, R.L. 2019. Balancing Selection Drives the Maintenance of Genetic Variation in *Drosophila* Antimicrobial Peptides. *Genome*

*Biology and Evolution*, 11 (9), pp.2691–2701.

Chapman, Joanne R., Hill, T., Unckless, R.L. and Wayne, M. 2019. Balancing Selection Drives the Maintenance of Genetic Variation in *Drosophila* Antimicrobial Peptides. *Genome Biology and Evolution*, 11 (9), pp.2691–2701.

Charlesworth, D. 2006. Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLOS Genetics*, 2 (4), pp.e64.

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), pp.49–58.

Cheng, P.L., Eng, H.L., Chou, M.H., You, H.L. and Lin, T.M. 2007. Genetic polymorphisms of viral infection-associated Toll-like receptors in Chinese population. *Translational Research*, 150 (5), pp.311–318.

Cheng, Q., Ruel, T.D., Zhou, W., Moloo, S.K., Majiwa, P., O’neill, S.L. and Aksoy, S. 2000. Tissue distribution and prevalence of *Wolbachia* infections in tsetse flies, *Glossina* spp. *Medical and Veterinary Entomology*, 14 (1), pp.44–50.

Cherry, J.L. 2010. Expression Level, Evolutionary Rate, and the Cost of Expression. *Genome Biology and Evolution*, 2, pp.757–769.

Choi, Y. 2012. A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein. *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine - BCB '12*. 2012. New York, New York, USA: ACM Press, 414–417.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. and Chan, A.P. 2012. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE*, 7 (10), pp.e46688.

Chong, R.A. and Moran, N.A. 2016. Intraspecific genetic variation in hosts affects regulation of obligate heritable symbionts. *Proceedings of the National Academy of Sciences*, 113 (46), pp.13114–13119.

Christophides, G.K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S., Blass, C., Brey, P.T., Collins, F.H., Danielli, A., Dimopoulos, G., Hetru, C., Hoa, N.T., Hoffmann, J.A.,

- Kanzok, S.M., Letunic, I., Levashina, E.A., Loukeris, T.G., Lycett, G., Meister, S., Michel, K., Moita, L.F., Müller, H.M., Osta, M.A., Paskewitz, S.M., Reichhart, J.M., Rzhetsky, A., Troxler, L., Vernick, K.D., Vlachou, D., Volz, J., von Mering, C., Xu, J., Zheng, L., Bork, P. and Kafatos, F.C. 2002. Immunity-Related Genes and Gene Families in *Anopheles gambiae*. *Science*, 298 (5591), pp.159–165.
- Christophides, G.K., Vlachou, D. and Kafatos, F.C. 2004. Comparative and functional genomics of the innate immune system in the malaria vector *Anopheles gambiae*. *Immunological Reviews*, 198, pp.127–148.
- Chu, D. and Wei, L. 2019. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. *BMC Cancer* 2019 19:1, 19 (1), pp.1–12.
- Clark, A.G. and Wang, L. 1997. Molecular Population Genetics of *Drosophila* Immune System Genes. *Genetics*, 147, pp.713–724.
- Cociancichs, S., Ghazio, A., Hetru, C., Hoffmanns, J.A. and Letelliers, L. 1993. *Insect Defensin, an Inducible Antibacterial Peptide, Forms Voltage-dependent Channels in Micrococcus luteus*.
- Cornet, B., Bonmatin, J.M., Hetru, C., Hoffmann, J.A., Ptak, M. and Vovelle, F. 1995. Refined three-dimensional solution structure of insect defensin A. *Structure*, 3 (5), pp.435–448.
- Courtin, D., Berthier, D., Thevenon, S., Dayo, G.K., Garcia, A. and Bucheton, B. 2008. Host genetics in African trypanosomiasis. *Infection, Genetics and Evolution*, 8 (3), pp.229–238.
- Courtin, F., Rayaissé, J.B., Tamboura, I., Serdébéogo, O., Koudougou, Z., Solano, P. and Sidibé, I. 2010. Updating the Northern Tsetse Limit in Burkina Faso (1949–2009): Impact of Global Change. *International Journal of Environmental Research and Public Health* 2010, Vol. 7, Pages 1708-1719, 7 (4), pp.1708–1719.
- Cudic, M., Bulet, P., Hoffmann, R., Craik, D.J. and Otvos Jr, L. 1999. Chemical synthesis, antibacterial activity and conformation of dipterin, an 82-mer peptide originally isolated from insects. *European Journal of Biochemistry*, 266 (2), pp.549–558.

- Cuscó, A., Sánchez, A., Altet, L., Ferrer, L. and Francino, O. 2014. Non-synonymous genetic variation in exonic regions of canine Toll-like receptors. *Canine Genetics and Epidemiology* 2014 1:1, 1 (1), pp.1–12.
- Dale, C. and Maudlin, I. 1999. *Sodalis* gen. nov. and *Sodalis glossinidius* sp. nov., a microaerophilic secondary endosymbiont of the tsetse fly *Glossina morsitans morsitans*. *International Journal of Systematic Bacteriology*, 49 (1), pp.267–275.
- De Deken, R. and Bouyer, J. 2018. Can sequential aerosol technique be used against riverine tsetse? *PLOS Neglected Tropical Diseases*, 12 (10), pp.e0006768.
- de Jong, M.A., Wahlberg, N., van Eijk, M., Brakefield, P.M. and Zwaan, B.J. 2011. Mitochondrial DNA Signature for Range-Wide Populations of *Bicyclus anynana* Suggests a Rapid Expansion from Recent Refugia. *PLoS ONE*, 6 (6), pp.e21385.
- Desquesnes, M. and Dia, M.L. 2003a. Mechanical transmission of *Trypanosoma congolense* in cattle by the African tabanid *Atylotus agrestis*. *Experimental Parasitology*, 105 (3–4), pp.226–231.
- Desquesnes, M. and Dia, M.L. 2003b. *Trypanosoma vivax*: Mechanical transmission in cattle by one of the most common African tabanids, *Atylotus agrestis*. *Experimental Parasitology*, 103 (1–2), pp.35–43.
- Dhople, V., Krukemeyer, A. and Ramamoorthy, A. 2006. The human beta-defensin-3, an antibacterial peptide with multiple biological functions. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1758 (9), pp.1499–1512.
- Dimarcq, J.L., Hoffmann, D., Meister, M., Bulet, P., Lanot, R., Reichhart, J.M. and Hoffmann, J.A. 1994. Characterization and transcriptional profiles of a *Drosophila* gene encoding an insect defensin. A study in insect immunity. *European Journal of Biochemistry*, 221 (1), pp.201–209.
- Dimopoulos, G., Richman, A., Muller, H.M. and Kafatos, F.C. 1997. Molecular immune responses of the mosquito *Anopheles gambiae* to bacteria and malaria parasites. *Proceedings of the National Academy of Sciences*, 94 (21), pp.11508–11513.
- Dipeolu, O.O. and Adam, K.M.G. 1974. On the use of membrane feeding to study the development of *Trypanosoma brucei* in *Glossina*. *Acta Tropica*, 31 (3), pp.185–201.

- Distelmans, W., D'haeseleer, F., Kaufman, L. and Rousseeuw, P. 1982. The susceptibility of *Glossina palpalis palpalis* at different ages to infection with *Trypanosoma congolense*. *Annales de la Société Belge de Médecine Tropicale*, 62 (1), pp.41–47.
- Drennan, M.B., Stijlemans, B., Van Den Abbeele, J., Quesniaux, V.J., Barkhuizen, M., Brombacher, F., De Baetselier, P., Ryffel, B. and Magez, S. 2005. The Induction of a Type 1 Immune Response following a *Trypanosoma brucei* Infection Is MyD88 Dependent. *The Journal of Immunology*, 175 (4), pp.2501–2509.
- Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O. and Arnold, F.H. 2005. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences*, 102 (40), pp.14338–14343.
- Dushay, M.S., Roethele, J.B., Chaverri, J.M., Dulek, D.E., Syed, S.K., Kitami, T. and Eldon, E.D. 2000. Two attacin antibacterial genes of *Drosophila melanogaster*. *Gene*, 246 (1–2), pp.49–57.
- Dyer, N.A., Lawton, S.P., Ravel, S., Choi, K.S., Lehane, M.J., Robinson, A.S., Okedi, L.M., Hall, M.J.R., Solano, P. and Donnelly, M.J. 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. *Molecular Phylogenetics and Evolution*, 49 (1), pp.227–239.
- Dziarski, R. and Gupta, D. 2006. The peptidoglycan recognition proteins (PGRPs). *Genome Biology* 2006 7:8, 7 (8), pp.1–13.
- Ebner, M. 2006. Coevolution and the Red Queen effect shape virtual plants. *Genet Program Evolvable Mach*, 7, pp.103–123.
- Echave, J. and Wilke, C.O. 2017. Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. <https://doi.org/10.1146/annurev-biophys-070816-033819>, 46, pp.85–103.
- Echave, J., Spielman, S.J. and Wilke, C.O. 2016. Causes of evolutionary rate variation among protein sites. *Nature reviews. Genetics*, 17 (2), pp.109.
- Eiríksdóttir, E., Konate, K., Langel, Ü., Divita, G. and Deshayes, S. 2010. Secondary structure of cell-penetrating peptides controls membrane interaction and insertion. *Biochimica*



*et Biophysica Acta (BBA) - Biomembranes*, 1798 (6), pp.1119–1128.

- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., Sonnhammer, E.L.L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S.C.E. and Finn, R.D. 2019. The Pfam protein families database in 2019. *Nucleic Acids Research*, 47 (D1), pp.D427–D432.
- El-Sayed, N.M., Hegde, P., Quackenbush, J., Melville, S.E. and Donelson, J.E. 2000. The African trypanosome genome. *International Journal for Parasitology*, 30 (4), pp.329–345.
- Elbers, A.R.W., Koenraadt, C.J.M. and Meiswinkel, R. 2015. Mosquitoes and Culicoides biting midges: vector range and the influence of climate change. *Rev. Sci. Tech. Off. Int. Epiz*, 34 (1), pp.123–137.
- Engström, P., Carlsson, A., Engström, A., Tao, Z.J. and Bennich, H. 1984. The antibacterial effect of attacins from the silk moth *Hyalophora cecropia* is directed against the outer membrane of *Escherichia coli*. *The EMBO Journal*, 3 (13), pp.3347–3351.
- Esterhuizen, J., Rayaisse, J.B., Tirados, I., Mpiana, S., Solano, P., Vale, G.A., Lehane, M.J. and Torr, S.J. 2011. Improving the Cost-Effectiveness of Visual Devices for the Control of Riverine Tsetse Flies, the Major Vectors of Human African Trypanosomiasis. *PLoS Neglected Tropical Diseases*, 5 (8), pp.e1257.
- Esterhuizen, J., Njiru, B., Vale, G.A., Lehane, M.J. and Torr, S.J. 2011. Vegetation and the importance of insecticide-treated target siting for control of *Glossina fuscipes fuscipes*. *PLoS Neglected Tropical Diseases*, 5 (9).
- Evans, J.D., Aronstein, K., Chen, Y.P., Hetru, C., Imler, J.L., Jiang, H., Kanost, M., Thompson, G.J., Zou, Z. and Hultmark, D. 2006. Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Molecular Biology*, 15 (5), pp.645–656.
- Feasey, N., Wansbrough-Jones, M., Mabey, D.C.W. and Solomon, A.W. 2010. Neglected tropical diseases. *British Medical Bulletin*, 93 (1), pp.179–200.
- Feeney, W.E., Welbergen, J.A. and Langmore, N.E. 2012. The frontline of avian brood parasite–host coevolution. *Animal Behaviour*, 84 (1), pp.3–12.
- Fels Elliott, D.R., Perner, J., Li, X., Symmons, M.F., Verstak, B., Eldridge, M., Bower, L.,

- O'Donovan, M., Gay, N.J. and Fitzgerald, R.C. 2017. Impact of mutations in Toll-like receptor pathway genes on esophageal carcinogenesis. *PLOS Genetics*, 13 (5), pp.e1006808.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17 (6), pp.368–376.
- Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39 (4), pp.783–791.
- Ferwerda, B., McCall, M.B.B., Alonso, S., Giamarellos-Bourboulis, E.J., Mouktaroudi, M., Izagirre, N., Syafruddin, D., Kibiki, G., Cristea, T., Hijmans, A., Hamann, L., Israel, S., Elghazali, G., Troye-Blomberg, M., Kumpf, O., Maiga, B., Dolo, A., Doumbo, O., Hermsen, C.C., Stalenhoef, A.F.H., Van Crevel, R., Brunner, H.G., Oh, D.Y., Schumann, R.R., De La Rúa, C., Sauerwein, R., Kullberg, B.J., Van Der Ven, A.J.A.M., Van Der Meer, J.W.M. and Netea, M.G. 2007. TLR4 polymorphisms, infectious diseases, and evolutionary pressure during migration of modern humans. *Proceedings of the National Academy of Sciences*, 104 (42), pp.16645–16650.
- Forister, M.L., Pelton, E.M. and Black, S.H. 2019. Declines in insect abundance and diversity: We know enough to act now. *Conservation Science and Practice*, 1 (8), pp.1–8.
- Fu, Y.X. 1997. Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. *Genetics*, 147 (2), pp.915–925.
- Funk, Daniel J., Wernegreen, J.J. and Moran, N.A. 2001. Intraspecific variation in symbiont genomes: Bottlenecks and the aphid-Buchnera association. *Genetics*, 157 (2), pp.477–489.
- Funk, Daniel J, Wernegreen, J.J. and Moran, N.A. 2001. *Intraspecific Variation in Symbiont Genomes: Bottlenecks and the Aphid-Buchnera Association*.
- Gandon, S. 2002. Local adaptation and the geometry of host-parasite coevolution. *Ecology Letters*, 5 (2), pp.246–256.
- Ganz, T. 2003. The Role of Antimicrobial Peptides in Innate Immunity. *Integrative and Comparative Biology*, 43 (2), pp.300–304.
- Gao, J.M., Qian, Z.Y., Hide, G., Lai, D.H., Lun, Z.R. and Wu, Z.D. 2020. Human African

trypanosomiasis: the current situation in endemic regions and the risks for non-endemic regions from imported cases. *Parasitology*, 147 (9), pp.922.

Geiger, A., Ravel, S., Mateille, T., Janelle, J., Patrel, D., Cuny, G. and Frutos, R. 2006. Vector Competence of *Glossina palpalis gambiensis* for *Trypanosoma brucei* s.l. and Genetic Diversity of the Symbiont *Sodalis glossinidius*. *Molecular Biology and Evolution*, 24 (1), pp.102–109.

Giraldo-Calderón, G.I., Emrich, S.J., MacCallum, R.M., Maslen, G., Dialynas, E., Topalis, P., Ho, N., Gesing, S., Madey, G., Collins, F.H. and Lawson, D. 2015. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Research*, 43 (D1), pp.D707–D713.

Giraldo-Calderón, G.I., Emrich, S.J., MacCallum, R.M., Maslen, G., Emrich, S., Collins, F., Dialynas, E., Topalis, P., Ho, N., Gesing, S., Madey, G., Collins, F.H., Lawson, D., Kersey, P., Allen, J., Christensen, M., Hughes, D., Koscielny, G., Langridge, N., Gallego, E.L., Megy, K., Wilson, D., Gelbart, B., Emmert, D., Russo, S., Zhou, P., Christophides, G., Brockman, A., Kirmizoglou, I., MacCallum, B., Tiirikka, T., Louis, K., Dritsou, V., Mitraka, E., Werner-Washburn, M., Baker, P., Platero, H., Aguilar, A., Bogol, S., Campbell, D., Carmichael, R., Cieslak, D., Davis, G., Konopinski, N., Nabrzyski, J., Reinking, C., Sheehan, A., Szakonyi, S. and Wieck, R. 2015. VectorBase: An updated Bioinformatics Resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Research*, 43 (D1), pp.D707–D713.

Gómez-Valero, L., Latorre, A., Gil, R., Gadau, J., Feldhaar, H. and Silva, F.J. 2008. Patterns and rates of nucleotide substitution, insertion and deletion in the endosymbiont of ants *Blochmannia floridanus*. *Molecular Ecology*, 17 (19), pp.4382–4392.

Guex, N., Peitsch, M.C. and Schwede, T. 2009. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis*, 30 (SUPPL. 1).

Gunne, H., Hellers, M. and Steiner, H. 1990. Structure of preproattacin and its processing in insect cells infected with a recombinant baculovirus. *European Journal of Biochemistry*, 187 (3), pp.699–703.

Haddrill, P.R., Loewe, L. and Charlesworth, B. 2010. Estimating the Parameters of Selection

- on Nonsynonymous Mutations in *Drosophila pseudoobscura* and *D. miranda*. *Genetics*, 185 (4), pp.1381–1396.
- Hammer, D.A.T., Ryan, P.D., Hammer, Ø. and Harper, D.A.T. 2001. *Past: Paleontological Statistics Software Package for Education and Data Analysis*.
- Hao, Z., Kasumba, I., Lehane, M.J., Gibson, W.C., Kwon, J. and Aksoy, S. 2001. Tsetse immune responses and trypanosome transmission: Implications for the development of tsetse-based strategies to reduce trypanosomiasis. *Proceedings of the National Academy of Sciences of the United States of America*, 98 (22), pp.12648–12653.
- Hargrove, J.W. 1988. Tsetse: the limits to population growth. *Medical and Veterinary Entomology*, 2 (3), pp.203–217.
- Hargrove, J.W. 2000. A theoretical study of the invasion of cleared areas by tsetse flies (Diptera: Glossinidae). *Bulletin of Entomological Research*, 90 (3), pp.201–209.
- Hargrove, J.W., Omolo, S., Msalilwa, J.S.I. and Fox, B. 2000. Insecticide-treated cattle for tsetse control: the power and the problems. *Medical and Veterinary Entomology*, 14 (2), pp.123–130.
- Harpending, H.C. 1994. Signature of Ancient Population Growth in a Low-Resolution Mitochondrial DNA Mismatch Distribution - ProQuest. *Human Biology*, 66 (4), pp.591–600.
- Harrison, P.M., Milburn, D., Zhang, Z., Bertone, P. and Gerstein, M. 2003. Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Research*, 31 (3), pp.1033–1037.
- Hashimoto, C., Hudson, K.L. and Anderson, K. V. 1988. The Toll gene of *Drosophila*, required for dorsal-ventral embryonic polarity, appears to encode a transmembrane protein. *Cell*, 52 (2), pp.269–279.
- Hedengren, M., Borge, K. and Hultmark, D. 2000. Expression and Evolution of the *Drosophila* Attacin/Diptericin Gene Family. *Biochemical and Biophysical Research Communications*, 279 (2), pp.574–581.
- Hide, G. 1999. History of sleeping sickness in East Africa. *Clinical microbiology reviews*, 12 (1), pp.112–25.

- Hill, T., Koseva, B.S. and Unckless, R.L. 2019. The Genome of *Drosophila innubila* Reveals Lineage-Specific Patterns of Selection in Immune Genes. *Molecular Biology and Evolution*, 36 (7), pp.1405–1417.
- Holden, J.A., O'Brien-Simpson, N.M., Lenzo, J.C., Orth, R.K.H., Mansell, A. and Reynolds, E.C. 2017. *Porphyromonas gulae* Activates Unprimed and Gamma Interferon-Primed Macrophages via the Pattern Recognition Receptors Toll-Like Receptor 2 (TLR2), TLR4, and NOD2. *Infection and Immunity*, 85 (9).
- Holley, A.S. 2011. In Vitro Determination of Canine Immunological Responses Due to Exposure to Recombinant *Wolbachia* Surface Protein. Auburn University.
- Holm, L. 2020. DALI and the persistence of protein shape. *Protein Science*, 29 (1), pp.128–140.
- Horn, D. 2014. Antigenic variation in African trypanosomes. *Molecular and Biochemical Parasitology*, 195 (2), pp.123–129.
- Horng, T. and Medzhitov, R. 2001. *Drosophila* MyD88 is an adapter in the Toll signaling pathway. *Proceedings of the National Academy of Sciences*, 98 (22), pp.12654–12658.
- Hoskin, D.W. and Ramamoorthy, A. 2008. Studies on Anticancer Activities of Antimicrobial Peptides. *Biochimica et biophysica acta*, 1778 (2), pp.357.
- Hotez, P.J. and Kamath, A. 2009. Neglected Tropical Diseases in Sub-Saharan Africa: Review of Their Prevalence, Distribution, and Disease Burden. *PLOS Neglected Tropical Diseases*, 3 (8), pp.e412.
- Hu, C. and Aksoy, S. 2006. Innate immune responses regulate trypanosome parasite infection of the tsetse fly *Glossina morsitans morsitans*. *Molecular Microbiology*, 60 (5), pp.1194–1204.
- Huang, H.W. 2000. Action of Antimicrobial Peptides: Two-State Model. *Biochemistry*, 39 (29), pp.8347–8352.
- Hudson, R.R. 1987. Estimating the recombination parameter of a finite population model without selection. *Genetical Research*, 50 (3), pp.245–250.
- Hudson, R.R., Boos, D.D. and Kaplan, N.L. 1992. A statistical test for detecting geographic

- subdivision. *Molecular Biology and Evolution*, 9 (1), pp.138–151.
- Hudson, R.R., Slatkin, M. and Maddison, W.P. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132 (2), pp.583–589.
- Hultmark, D., Engström, A., Andersson, K., Steiner, H., Bennich, H. and Boman, H.G. 1983. Insect immunity. Attacins, a family of antibacterial proteins from *Hyalophora cecropia*. *The EMBO Journal*, 2 (4), pp.571–576.
- Hurst, G. and Darby, A.C. 2009. The inherited microbiota of arthropods, and their importance in understanding resistance and immunity Gut microbiota of the agricultural pest *Bactrocera oleae* View project.
- Hwang, P.M. and Vogel, H.J. 1998. Structure-function relationships of antimicrobial peptides. *Biochemistry and Cell Biology*, 76 (2–3), pp.235–246.
- Imler, J.L. and Bulet, P. 2005. Antimicrobial Peptides in *Drosophila*: Structures, Activities and Gene Regulation. *Mechanisms of Epithelial Defense*. Basel: KARGER, 1–21.
- Imler, Jean-Luc and Hoffmann, J.A. 2000. Signaling mechanisms in the antimicrobial host defense of *Drosophila*. *Current Opinion in Microbiology*, 3 (1), pp.16–22.
- Imler, J L and Hoffmann, J.A. 2000. Toll and Toll-like proteins: an ancient family of receptors signaling infection. *Reviews in immunogenetics*, 2 (3), pp.294–304.
- Janeway, C.A. and Medzhitov, R. 2002. Innate immune recognition. *Annual Review of Immunology*, 20, pp.197–216.
- Jang, I.H., Chosa, N., Kim, S.H., Nam, H.J., Lemaitre, B., Ochiai, M., Kambris, Z., Brun, S., Hashimoto, C., Ashida, M., Brey, P.T. and Lee, W.J. 2006. A Spätzle-Processing Enzyme Required for Toll Signaling Activation in *Drosophila* Innate Immunity. *Developmental Cell*, 10 (1), pp.45–55.
- Jendele, L., Krivak, R., Skoda, P., Novotny, M. and Hoksza, D. 2019. PrankWeb: a web server for ligand binding site prediction and visualization. *Nucleic Acids Research*, 47 (W1), pp.W345–W349.
- Jiggins, F.M. and Hurst, G.D.D. 2003. The Evolution of Parasite Recognition Genes in the Innate Immune System: Purifying Selection on *Drosophila melanogaster*

Peptidoglycan Recognition Proteins. *Journal of Molecular Evolution*, 57 (5), pp.598–605.

Jimenez, M.J., Arenas, M. and Bastolla, U. 2018. Substitution Rates Predicted by Stability-Constrained Models of Protein Evolution Are Not Consistent with Empirical Data. *Molecular Biology and Evolution*, 35 (3), pp.743–755.

Jin, M.S., Kim, S.E., Heo, J.Y., Lee, M.E., Kim, H.M., Paik, S.G., Lee, H. and Lee, J.O. 2007. Crystal Structure of the TLR1-TLR2 Heterodimer Induced by Binding of a Tri-Acylated Lipopeptide. *Cell*, 130 (6), pp.1071–1082.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P. and Hassabis, D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021 596:7873, 596 (7873), pp.583–589.

Juneja, P. and Lazzaro, B.P. 2009. Population genetics of insect immune responses. *Insect Infection and Immunity: Evolution, Ecology, and Mechanisms*. Oxford University Press, 206–224.

Kafatos, F., Waterhouse, R., Zdobnov, E. and Christophides, G. 2009. Comparative genomics of insect immunity. *Insect Infection and Immunity*. 86–105.

Kambris, Z., Blagborough, A.M., Pinto, S.B., Blagrove, M.S.C., Godfray, H.C.J., Sinden, R.E. and Sinkins, S.P. 2010. Wolbachia Stimulates Immune Gene Expression and Inhibits Plasmodium Development in *Anopheles gambiae*. *PLoS Pathogens*, 6 (10), pp.e1001143.

Kasozi, K.I., Zirintunda, G., Ssempijja, F., Buyinza, B., Alzahrani, K.J., Matama, K., Nakimbugwe, H.N., Alkazmi, L., Onanyang, D., Bogere, P., Ochieng, J.J., Islam, S., Matovu, W., Nalumenya, D.P., Batiha, G.E.S., Osuwat, L.O., Abdelhamid, M., Shen, T., Omadang, L. and Welburn, S.C. 2021. Epidemiology of Trypanosomiasis in Wildlife—Implications for Humans at the Wildlife Interface in Africa. *Frontiers in Veterinary Science*, 8, pp.565.

- Key, F.M., Teixeira, J.C., de Filippo, C. and Andrés, A.M. 2014a. Advantageous diversity maintained by balancing selection in humans. *Current Opinion in Genetics and Development*, 29, pp.45–51.
- Key, F.M., Teixeira, J.C., de Filippo, C. and Andrés, A.M. 2014b. Advantageous diversity maintained by balancing selection in humans. *Current Opinion in Genetics & Development*, 29, pp.45–51.
- Khan, J.A., Brint, E.K., O’Neill, L.A.J. and Tong, L. 2004. Crystal Structure of the Toll/Interleukin-1 Receptor Domain of Human IL-1RAPL. *Journal of Biological Chemistry*, 279 (30), pp.31664–31670.
- Khush, R.S. and Lemaitre, B. 2000. Genes that fight infection: what the Drosophila genome says about animal immunity. *Trends in Genetics*, 16 (10), pp.442–449.
- Kikuchi, Y., Hosokawa, T., Nikoh, N., Meng, X.Y., Kamagata, Y. and Fukatsu, T. 2009. Host-symbiont co-speciation and reductive genome evolution in gut symbiotic bacteria of acanthosomatid stinkbugs. *BMC Biology*, 7 (1), pp.2.
- Kosakovskiy, S.L. and Frost, S.D.W. 2005. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Molecular Biology and Evolution*, 22 (5), pp.1208–1222.
- Kozakov, D., Grove, L.E., Hall, D.R., Bohnuud, T., Mottarella, S.E., Luo, L., Xia, B., Beglov, D. and Vajda, S. 2015. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature Protocols*, 10 (5), pp.733–755.
- Krafsur, E.S. 2009. Tsetse flies: Genetics, evolution, and role as vectors. *Infection, Genetics and Evolution*, 9 (1), pp.124–141.
- Kramer, L., Grandi, G., Leoni, M., Passeri, B., McCall, J., Genchi, C., Mortarino, M. and Bazzocchi, C. 2008. Wolbachia and its influence on the pathology and immunology of *Dirofilaria immitis* infection. *Veterinary Parasitology*, 158 (3), pp.191–195.
- Kristofich, J., Morgenthaler, A.B., Kinney, W.R., Ebmeier, C.C., Snyder, D.J., Old, W.M., Cooper, V.S. and Copley, S.D. 2018. Synonymous mutations make dramatic contributions to fitness when growth is limited by a weak-link enzyme. *PLoS Genetics*,



14 (8).

- Kumar, H., Kawai, T. and Akira, S. 2009. Pathogen recognition in the innate immune response. *Biochemical Journal*, 420 (1), pp.1–16.
- Kumar, S., Stecher, G. and Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, 33 (7), pp.1870–1874.
- Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, 35 (6), pp.1547–1549.
- Langley, P.A., Hargrove, J.W., Mauchamp, B., Royer, C. and Oouchi, H. 1993. Prospects for using pyroproxyfen-treated targets for tsetse control. *Entomologia Experimentalis et Applicata*, 66 (2), pp.153–159.
- Lazzaro, B.P. and Clark, A.G. 2001. *Evidence for Recurrent Paralogous Gene Conversion and Exceptional Allelic Divergence in the Attacin Genes of Drosophila melanogaster*.
- Le, S.Q. and Gascuel, O. 2008. An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, 25 (7), pp.1307–1320.
- Leak, S. 1999. *Tsetse biology and ecology: their role in the epidemiology and control of trypanosomosis*, 1st ed. Centre for Agriculture and Bioscience International.
- Lebeuf-Taylor, E., McCloskey, N., Bailey, S.F., Hinz, A. and Kassen, R. 2019. The distribution of fitness effects among synonymous mutations in a gene under directional selection. *eLife*, 8.
- Lehane, M.J., Aksoy, S. and Levashina, E. 2004. Immune responses and parasite transmission in blood-feeding insects. *Trends in Parasitology*, 20 (9), pp.433–439.
- Lehmann, T., Hume, J.C.C., Licht, M., Burns, C.S., Wollenberg, K., Simard, F. and Ribeiro, J.M.C. 2009. Molecular evolution of immune genes in the malaria mosquito *Anopheles gambiae*. *PLoS ONE*, 4 (2).
- Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, Jean-Marc and Hoffmann, J.A. 1996. The Dorsoventral Regulatory Gene Cassette *spätzle/Toll/cactus* Controls the Potent

- Antifungal Response in *Drosophila* Adults. *Cell*, 86 (6), pp.973–983.
- Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, Jean Marc and Hoffmann, J.A. 1996. The Dorsoventral Regulatory Gene Cassette *spätzle/Toll/cactus* Controls the Potent Antifungal Response in *Drosophila* Adults. *Cell*, 86 (6), pp.973–983.
- Leulier, F. and Lemaitre, B. 2008. Toll-like receptors - Taking an evolutionary approach. *Nature Reviews Genetics*, 9 (3), pp.165–178.
- Levin, T.C. and Malik, H.S. 2017. Rapidly Evolving Toll-3/4 Genes Encode Male-Specific Toll-Like Receptors in *Drosophila*. *Molecular Biology and Evolution*, 34 (9), pp.2307–2323.
- Liao, D. 1999. Concerted evolution: molecular mechanism and biological implications. *American Journal of Human Genetics*, 64 (1), pp.24.
- Lima, L.F., Torres, A.Q., Jardim, R., Mesquita, R.D. and Schama, R. 2021. Evolution of Toll, Spatzle and MyD88 in insects: the problem of the Diptera bias. *BMC Genomics* 2021 22:1, 22 (1), pp.1–21.
- Little, T.J. and Cobbe, N. 2005. The evolution of immune-related genes from disease carrying mosquitoes: Diversity in a peptidoglycan- and a thioester-recognizing protein. *Insect Molecular Biology*, 14 (6), pp.599–605.
- Lloyd, L.L. 1930. Some factors influencing the trypanosome infection rate in tsetse flies. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 23 (5), pp.533–542.
- Lorch, M., Mason, J.M., Clarke, A.R. and Parker, M.J. 1999. Effects of core mutations on the folding of a  $\beta$ -sheet protein: Implications for backbone organization in the I-state. *Biochemistry*, 38 (4), pp.1377–1385.
- Lorch, M., Mason, J.M., Sessions, R.B. and Clarke, A.R. 2000. Effects of mutations on the thermodynamics of a protein folding reaction: Implications for the mechanism of formation of the intermediate and transition states. *Biochemistry*, 39 (12), pp.3480–3485.
- Losos, G.J. and Ikede, B.O. 1972. Review of Pathology of Diseases in Domestic and Laboratory Animals Caused by *Trypanosoma congolense*, *T. vivax*, *T. brucei*, *T. rhodesiense* and *T. gambiense*. *Veterinary Pathology*, 9 (1), pp.1–79.

- Lundkvist, G.B., Kristensson, K. and Bentivoglio, M. 2004. Why Trypanosomes Cause Sleeping Sickness. *Physiology*, 19 (4), pp.198–206.
- Luo, C. and Zheng, L. 2000. Independent evolution of Toll and related genes in insects and mammals. *Immunogenetics*, 51, pp.92–98.
- Luque-Ortega, J.R., Hof, W. van't, Veerman, E.C.I., Saugar, J.M. and Rivas, L. 2008. Human antimicrobial peptide histatin 5 is a cell- penetrating peptide targeting mitochondrial ATP synthesis in Leishmania. *The FASEB Journal*, 22 (6), pp.1817–1828.
- Lynch, M. 2010. Evolution of the mutation rate. *Trends in Genetics*, 26 (8), pp.345–352.
- Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D. and Lopez, R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, 47 (W1), pp.W636–W641.
- Magzoub, M., Kilk, K., Eriksson, L.E.G., Langel, Ü. and Gräslund, A. 2001. Interaction and structure induction of cell-penetrating peptides in the presence of phospholipid vesicles. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1512 (1), pp.77–89.
- Magzoub, M., Eriksson, L.E.G. and Gräslund, A. 2002. Conformational states of the cell-penetrating peptide penetratin when interacting with phospholipid vesicles: effects of surface charge and peptide concentration. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1563 (1–2), pp.53–63.
- Mahalingam, B., Louis, J.M., Hung, J., Harrison, R.W. and Weber, I.T. 2001. Structural implications of drug-resistant mutants of HIV-1 protease: High-resolution crystal structures of the mutant protease/substrate analogue complexes. *Proteins: Structure, Function, and Bioinformatics*, 43 (4), pp.455–464.
- Marcos, M.L. and Echave, J. 2020. The variation among sites of protein structure divergence is shaped by mutation and scaled by selection. *Current Research in Structural Biology*, 2, pp.156–163.
- Masocha, W. and Kristensson, K. 2012. Passage of parasites across the blood-brain barrier. *Virulence*, 3 (2), pp.202–212.
- Matthews, K.R., Ellis, J.R. and Paterou, A. 2004. Molecular regulation of the life cycle of African trypanosomes. *Trends in Parasitology*, 20 (1), pp.40–47.

- McCandlish, D.M. and Stoltzfus, A. 2014. Modeling evolution using the probability of fixation: History and implications. *Quarterly Review of Biology*, 89 (3), pp.225–252.
- McCann, H.C., Nahal, H., Thakur, S. and Guttman, D.S. 2012. Identification of innate immunity elicitors using molecular signatures of natural selection. *Proceedings of the National Academy of Sciences*, 109 (11), pp.4215–4220.
- Mellanby, K. 1937. Water and fat content of tsetse flies. *Nature*, 139 (3525), pp.883.
- Meyer, A., Holt, H.R., Oumarou, F., Chilongo, K., Gilbert, W., Fauron, A., Mumba, C. and Guitian, J. 2018. Integrated cost-benefit analysis of tsetse control and herd productivity to inform control programs for animal African trypanosomiasis. *Parasites & Vectors* 2018 11:1, 11 (1), pp.1–14.
- Mihok, S., Maramba, O., Munyoki, E. and Kagoiya, J. 1995. Mechanical transmission of *Trypanosoma* spp. by African *Stomoxynae* (Diptera: Muscidae). *Tropical Medicine and Parasitology*, 46 (2), pp.103–105.
- Misch, E.A. and Hawn, T.R. 2008. Toll-like receptor polymorphisms and susceptibility to human disease. *Clinical Science*, 114 (5), pp.347–360.
- Moran, Y., Weinberger, H., Sullivan, J.C., Reitzel, A.M., Finnerty, J.R. and Gurevitz, M. 2008. Concerted Evolution of Sea Anemone Neurotoxin Genes Is Revealed through Analysis of the *Nematostella vectensis* Genome. *Molecular Biology and Evolution*, 25 (4), pp.737–747.
- Morrison, L.J., Vezza, L., Rowan, T. and Hope, J.C. 2016. Animal African Trypanosomiasis: Time to Increase Focus on Clinically Relevant Parasite and Host Species. *Trends in parasitology*, 32 (8), pp.599–607.
- Mosser, D.M. and Edelson, P.J. 1984. Activation of the alternative complement pathway by *Leishmania* promastigotes: parasite lysis and attachment to macrophages. *Journal of immunology (Baltimore, Md. : 1950)*, 132 (3), pp.1501–5.
- Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K. and Kosakovsky Pond, S.L. 2012. Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLoS Genetics*, 8 (7), pp.e1002764.
- Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S.L. and

- Scheffler, K. 2013. FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. *Molecular Biology and Evolution*, 30 (5), pp.1196–1205.
- Muse', S. V and Gaut, B.S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11 (5), pp.715–724.
- Myllymäki, H., Valanne, S. and Rämetsä, M. 2014. The Drosophila Imd Signaling Pathway. *The Journal of Immunology*, 192 (8), pp.3455–3462.
- Ndeledje, N., Bouyer, J., Stachurski, F., Grimaud, P., Belem, A.M.G., Molélé Mbäindingatoloum, F., Bengaly, Z., Oumar Alfaroukh, I., Cecchi, G. and Lancelot, R. 2013. Treating Cattle to Protect People? Impact of Footbath Insecticide Treatment on Tsetse Density in Chad. *PLoS ONE*, 8 (6).
- Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia university press.
- Nei, M. and Miller, J.C. 1990. A Simple Method for Estimating Average Number of Nucleotide Substitutions Within and Between Populations From Restriction Data. *Genetics*, 125 (4), pp.873–879.
- Netea, M.G., Wijmenga, C. and O'Neill, L.A.J. 2012. Genetic variation in Toll-like receptors and disease susceptibility. *Nature Immunology*, 13 (6), pp.535–542.
- Newstead, R. 1924. *Guide to the Study of Tsetse-Flies.*, 1st ed. University Press of Liverpool, Ltd., Hodder & Stoughton, Ltd.
- Ngan, C.H., Hall, D.R., Zerbe, B., Grove, L.E., Kozakov, D. and Vajda, S. 2012. FTSite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics*, 28 (2), pp.286–287.
- Niaré, O., Markianos, K., Volz, J., Oduol, F., Touré, A., Bagayoko, M., Sangaré, D., Traoré, S.F., Wang, R., Blass, C., Dolo, G., Bouaré, M., Kafatos, F.C., Kruglyak, L., Touré, Y.T. and Vernick, K.D. 2002. Genetic loci affecting resistance to human malaria parasites in a West African mosquito vector population. *Science*, 298 (5591), pp.213–216.
- Nicolas, P. 2009. Multifunctional host defense peptides: intracellular-targeting antimicrobial peptides. *The FEBS Journal*, 276 (22), pp.6483–6496.

- O'Neill, L.A.J. and Bowie, A.G. 2007. The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. *Nature Reviews Immunology*, 7 (5), pp.353–364.
- Oberle, M., Balmer, O., Brun, R. and Roditi, I. 2010. Bottlenecks and the maintenance of minor genotypes during the life cycle of *Trypanosoma brucei*. *PLoS Pathogens*, 6 (7), pp.1–8.
- Ohto, U., Shibata, T., Tanji, H., Ishida, H., Krayukhina, E., Uchiyama, S., Miyake, K. and Shimizu, T. 2015. Structural basis of CpG and inhibitory DNA recognition by Toll-like receptor 9. *Nature*, 520 (7549), pp.702–705.
- Okonechnikov, K., Golosova, O. and Fursov, M. 2012. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, 28 (8), pp.1166–1167.
- Ooi, C.P., Haines, L.R., Southern, D.M., Lehane, M.J. and Acosta-Serrano, A. 2015. Tsetse GmmSRPN10 Has Anti-complement Activity and Is Important for Successful Establishment of Trypanosome Infections in the Fly Midgut. *PLoS Neglected Tropical Diseases*, 9 (1).
- Otieno, L.H., Darji, N., Onyango, P. and Mpanga, E. 1983. Some observations on factors associated with the development of *Trypanosoma brucei brucei* infections in *Glossina morsitans morsitans*. *Acta Tropica*, 40 (2), pp.113–120.
- Otvos, L. 2005. Antibacterial peptides and proteins with multiple cellular targets. *Journal of Peptide Science*, 11 (11), pp.697–706.
- 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), pp.5965.
- Palti, Y. 2011. Toll-like receptors in bony fish: From genomics to function. *Developmental & Comparative Immunology*, 35 (12), pp.1263–1272.
- Pasvol, G., Weatherall, D.J. and Wilson, R.J.M. 1978. Cellular mechanism for the protective effect of haemoglobin S against *P. falciparum* malaria [22]. *Nature*, 274 (5672), pp.701–703.
- Peona, V., Blom, M.P.K., Xu, L., Burri, R., Sullivan, S., Bunikis, I., Liachko, I., Haryoko, T.,

- Jønsson, K.A., Zhou, Q., Irestedt, M. and Suh, A. 2021. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol Ecol Resour*, 21, pp.263–286.
- Percoma, L., Sow, A., Pagabeleguem, S., Dicko, A.H., Serdebéogo, O., Ouédraogo, M., Rayaissé, J.B., Bouyer, J., Belem, A.M.G. and Sidibé, I. 2018. Impact of an integrated control campaign on tsetse populations in Burkina Faso. *Parasites & Vectors*, 11 (1), pp.1–13.
- Petersen, F.T., Meier, R., Kutty, S.N. and Wiegmann, B.M. 2007. The phylogeny and evolution of host choice in the Hippoboscoidea (Diptera) as reconstructed using four molecular markers. *Molecular Phylogenetics and Evolution*, 45 (1), pp.111–122.
- Pond, S.L.K. and Frost, S.D.W. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, 21 (10), pp.2531–2533.
- Pond, S.L.K. and Muse, S. V. 2005. HyPhy: Hypothesis Testing Using Phylogenies. *Statistical Methods in Molecular Evolution*. New York: Springer-Verlag, 125–181.
- Portelli, S., Phelan, J.E., Ascher, D.B., Clark, T.G. and Furnham, N. 2018. Understanding molecular consequences of putative drug resistant mutations in Mycobacterium tuberculosis. *Scientific Reports 2018 8:1*, 8 (1), pp.1–12.
- Qin, Y.J., Buahom, N., Krosch, M.N., Du, Y., Wu, Y., Malacrida, A.R., Deng, Y.L., Liu, J.Q., Jiang, X.L. and Li, Z.H. 2016. Genetic diversity and population structure in *Bactrocera correcta* (Diptera: Tephritidae) inferred from mtDNA *cox1* and microsatellite markers. *Scientific Reports*, 6 (1), pp.1–10.
- Raupach, M.J., Thatje, S., Dambach, J., Rehm, P., Misof, B. and Leese, F. 2010. Genetic homogeneity and circum-Antarctic distribution of two benthic shrimp species of the Southern Ocean, *Chorismus antarcticus* and *Nematocarcinus lanceopes*. *Marine Biology 2010 157:8*, 157 (8), pp.1783–1797.
- Ravi, C., Jeyashree, A. and Devi, R. 2011. *Antimicrobial Peptides from Insects: An Overview*.
- Reddy, K.V.R., Yedery, R.D. and Aranha, C. 2004. Antimicrobial peptides: premises and promises. *International Journal of Antimicrobial Agents*, 24 (6), pp.536–547.
- Richman, A.M., Dimopoulos, G., Seeley, D. and Kafatos, F.C. 1997. Plasmodium activates

- the innate immune response of *Anopheles gambiae* mosquitoes. *EMBO Journal*, 16 (20), pp.6114–6119.
- Rio, R.V.M., Symula, R.E., Wang, J., Lohs, C., Wu, Y.N., Snyder, A.K., Bjornson, R.D., Oshima, K., Biehl, B.S., Perna, N.T., Hattori, M. and Aksoy, S. 2012. Insight into the Transmission Biology and Species-Specific Functional Capabilities of Tsetse (Diptera: Glossinidae) Obligate Symbiont *Wigglesworthia*. *mBio*, 3 (1).
- Rispe, C. and Moran, N.A. 2015. Accumulation of Deleterious Mutations in Endosymbionts: Muller's Ratchet with Two Levels of Selection. <https://doi.org/10.1086/303396>, 156 (4), pp.425–441.
- Roditi, I. and Lehane, M.J. 2008. Interactions between trypanosomes and tsetse flies. *Current Opinion in Microbiology*, 11 (4), pp.345–351.
- Rogers, A.R. and Harpending, H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, 9 (3), pp.552–569.
- Rolff, J. and Reynolds, S.E. 2009. Introducing insect infection and immunity. In: Rolff, J. and Reynolds, S.E. (eds.). *Insect Infection and Immunity: Evolution, Ecology, and Mechanisms*. Oxford University Press, 1–9.
- Roy, A., Kucukural, A. and Zhang, Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, 5 (4), pp.725–738.
- Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J.C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S.E. and Sánchez-Gracia, A. 2017. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Molecular Biology and Evolution*, 34 (12), pp.3299–3302.
- Sachs, J. and Malaney, P. 2002. The economic and social burden of malaria. *Nature* 2002 415:6872, 415 (6872), pp.680–685.
- Sackton, T.B., Lazzaro, B.P., Schlenke, T.A., Evans, J.D., Hultmark, D. and Clark, A.G. 2007. Dynamic evolution of the innate immune system in *Drosophila*. *NATURE GENETICS* |, 39.
- Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4 (4), pp.406–425.



- Salyers, A., Whitt, D. and Whitt, D. 1994. *Bacterial pathogenesis: a molecular approach*.
- Sanger, F., Nicklen, S. and Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74 (12), pp.5463–5467.
- Sassera, D., Epis, S., Pajoro, M. and Bandi, C. 2013. Microbial symbiosis and the control of vectorborne pathogens in tsetse flies, human lice, and triatomine bugs. *Pathogens and Global Health*, 107 (6), pp.285–292.
- Sawai, M. V., Waring, A.J., Kearney, W.R., McCray, P.B., Forsyth, W.R., Lehrer, R.I. and Tack, B.F. 2002. Impact of single-residue mutations on the structure and function of ovispirin/novispirin antimicrobial peptides. *Protein Engineering, Design and Selection*, 15 (3), pp.225–232.
- Schlenke, T.A. and Begun, D.J. 2003. Natural Selection Drives Drosophila Immune System Evolution. *Genetics*, 164 (4), pp.1471–1480.
- Schmid-Hempel, P. 2008. Parasite immune evasion: a momentous molecular war. *Trends in Ecology & Evolution*, 23 (6), pp.318–326.
- Schröder, N.W.J. and Schumann, R.R. 2005. Single nucleotide polymorphisms of Toll-like receptors and susceptibility to infectious disease. *The Lancet Infectious Diseases*, 5 (3), pp.156–164.
- Schwarz, R. and Dayhoff, M. 1979. Matrices for Detecting Distant Relationships. *Dayhoff, M., Ed., Atlas of Protein Sequences*. National Biomedical Research Foundation, 353–358.
- Senger, K., Armstrong, G.W., Rowell, W.J., Kwan, J.M., Markstein, M. and Levine, M. 2004. Immunity Regulatory DNAs Share Common Organizational Features in Drosophila. *Molecular Cell*, 13 (1), pp.19–32.
- Shaw, A.P.M., Torr, S.J., Waiswa, C., Cecchi, G., Wint, G.R.W., Mattioli, R.C. and Robinson, T.P. 2013. Estimating the costs of tsetse control options: An example for Uganda. *Preventive Veterinary Medicine*, 110 (3–4), pp.290–303.
- Shereni, W., Anderson, N.E., Nyakupinda, L. and Cecchi, G. 2016. Spatial distribution and trypanosome infection of tsetse flies in the sleeping sickness focus of Zimbabwe in Hurungwe District. *Parasites & Vectors* 2016 9:1, 9 (1), pp.1–9.

- Shereni, W., Neves, L., Argilés, R., Nyakupinda, L. and Cecchi, G. 2021. An atlas of tsetse and animal African trypanosomiasis in Zimbabwe. *Parasites and Vectors*, 14 (1), pp.1–10.
- Siewert, K.M. and Voight, B.F. 2017. Detecting Long-Term Balancing Selection Using Allele Frequency Correlation. *Molecular Biology and Evolution*, 34 (11), pp.2996.
- Simard, F., Licht, M., Besansky, N.J. and Lehmann, T. 2007. Polymorphism at the defensin gene in the *Anopheles gambiae* complex: Testing different selection hypotheses. *Infection, Genetics and Evolution*, 7 (2), pp.285–292.
- Simarro, P.P., Cecchi, G., Franco, J.R., Paone, M., Diarra, A., Ruiz-Postigo, J.A., Fèvre, E.M., Mattioli, R.C. and Jannin, J.G. 2012. Estimating and Mapping the Population at Risk of Sleeping Sickness. *PLOS Neglected Tropical Diseases*, 6 (10), pp.e1859.
- Simmons, B., Balmford, A., ... A.B.E. and and 2019, undefined. 2019. Worldwide insect declines: an important message, but interpret with caution. *Wiley Online Library*, 9 (7), pp.3678–3680.
- Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H. and Flook, P. 1994. Evolution, Weighting, and Phylogenetic Utility of Mitochondrial Gene Sequences and a Compilation of Conserved Polymerase Chain Reaction Primers. *Annals of the Entomological Society of America*, 87 (6), pp.651–701.
- Slatkin, M. and Hudson, R.R. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129 (2), pp.555–562.
- Smith, T.K., Bringaud, F., Nolan, D.P. and Figueiredo, L.M. 2017. Metabolic reprogramming during the *Trypanosoma brucei* life cycle. *F1000Research*, 6, pp.683.
- Soares, M.P. and Yilmaz, B. 2016. Microbiota Control of Malaria Transmission. *Trends in Parasitology*, 32 (2), pp.120–130.
- Stein, D. and Nüsslein-Volhard, C. 1992. Multiple extracellular activities in *Drosophila* egg perivitelline fluid are required for establishment of embryonic dorsal-ventral polarity. *Cell*, 68 (3), pp.429–440.
- Steiner, H., Hultmark, D., Engström, Å., Bennich, H. and Boman, H.G. 1981. Sequence and specificity of two antibacterial proteins involved in insect immunity. *Nature* 1981 292:5820, 292 (5820), pp.246–248.

- Stephens, N.A., Kieft, R., MacLeod, A. and Hajduk, S.L. 2012. Trypanosome resistance to human innate immunity: targeting Achilles' heel. *Trends in Parasitology*, 28 (12), pp.539–545.
- Su, Y., Mani, R., Doherty, T., Waring, A.J. and Hong, M. 2008. Reversible Sheet–Turn Conformational Change of a Cell-Penetrating Peptide in Lipid Bilayers Studied by Solid-State NMR. *Journal of Molecular Biology*, 381 (5), pp.1133–1144.
- Sun, H., Bristow, B.N., Qu, G. and Wasserman, S.A. 2002. A heterotrimeric death domain complex in Toll signaling. *Proceedings of the National Academy of Sciences*, 99 (20), pp.12871–12876.
- Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. and Lehner, B. 2014. Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. *Cell*, 156 (6), pp.1324–1335.
- Suzuki, Y. and Gojobori, T. 1999. A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, 16 (10), pp.1315–1328.
- Swanevelder, Z.H., Surridge, A., Venter, E. and Botha, A.M. 2010. Limited Endosymbiont Variation in *Diuraphis noxia* (Hemiptera: Aphididae) Biotypes From the United States and South Africa. *Journal of Economic Entomology*, 103 (3), pp.887–897.
- Symula, R.E., Marpuri, I., Bjornson, R.D., Okedi, L., Beadell, J., Alam, U., Aksoy, S. and Caccone, A. 2011. Influence of Host Phylogeographic Patterns and Incomplete Lineage Sorting on Within-Species Genetic Variability in *Wigglesworthia* Species, Obligate Symbionts of Tsetse Flies. *APPLIED AND ENVIRONMENTAL MICROBIOLOGY*, 77 (23), pp.8400–8408.
- Tabachnick, W.J. 2010. Challenges in predicting climate and environmental effects on vector-borne disease epistemics in a changing world. *Journal of Experimental Biology*, 213 (6), pp.946–954.
- Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123 (3).
- Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+ C-content biases. *Molecular Biology and*

*Evolution*, 9 (4), pp.678–687.

- Tanji, T., Hu, X., Weber, A.N.R. and Ip, Y.T. 2007. Toll and IMD Pathways Synergistically Activate an Innate Immune Response in *Drosophila melanogaster*. *Molecular and Cellular Biology*, 27 (12), pp.4578–4588.
- Tauszig-Delamasure, S., Bilak, H., Capovilla, M., Hoffmann, J.A. and Imler, J.L. 2002. *Drosophila* MyD88 is required for the response to fungal and Gram-positive bacterial infections. *Nature Immunology*, 3 (1), pp.91–97.
- Taylor, A.W. 1932. The Development of West African Strains of *Trypanosoma gambiense* in *Glossina tachinoides* under Normal Laboratory Conditions, and at Raised Temperatures. *Parasitology*, 24 (3), pp.401–418.
- Tennessen, J.A. 2005. Molecular evolution of animal antimicrobial peptides: widespread moderate positive selection. *Journal of Evolutionary Biology*, 18 (6), pp.1387–1394.
- Thornton, P.K., Robinson, T.P., Kruska, R.L., Jones, P.G., McDermott, J.J. and Reid, R.S. 2006. Cattle trypanosomiasis in Africa to 2030.
- Tiede, S., Cantz, M., Spranger, J. and Bräulke, T. 2006. Missense mutation in the N-acetylglucosamine-1-phosphotransferase gene (GNPTA) in a patient with mucopolidosis II induces changes in the size and cellular distribution of GNPTG. *Human mutation*, 27 (8), pp.830–831.
- Tirados, I., Esterhuizen, J., Kovacic, V., Mangwiro, T.N.C., Vale, G.A., Hastings, I., Solano, P., Lehane, M.J. and Torr, S.J. 2015. Tsetse Control and Gambian Sleeping Sickness; Implications for Control Strategy. *PLOS Neglected Tropical Diseases*, 9 (8), pp.e0003822.
- Toh, H., Weiss, B.L., Perkin, S.A.H., Yamashita, A., Oshima, K., Hattori, M. and Aksoy, S. 2006. Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Research*, 16 (2), pp.149–156.
- Torrent, M., Pulido, D., Rivas, L. and Andreu, D. 2012. Antimicrobial Peptide Action on Parasites. *Current Drug Targets*, 13 (9), pp.1138–1147.
- Tørresen, O.K., Star, B., Jentoft, S., Reinart, W.B., Grove, H., Miller, J.R., Walenz, B.P., Knight,

- J., Ekholm, J.M., Peluso, P., Edvardsen, R.B., Tooming-Klunderud, A., Skage, M., Lien, S., Jakobsen, K.S. and Nederbragt, A.J. 2017. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics*, 18 (1), pp.1–23.
- Trappeniens, 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, pp.1650.
- Travis, J. 1993. Tracing the immune system's evolutionary history. *Science*, 261 (5118), pp.164.
- Tschirren, B., Andersson, M., Scherman, K., Westerdahl, H., Mittl, P.R.E. and Råberg, L. 2013. Polymorphisms at the innate immune receptor TLR2 are associated with *Borrelia* infection in a wild rodent population. *Proceedings of the Royal Society B: Biological Sciences*, 280 (1759).
- Tseng, S.P., Wetterer, J.K., Suarez, A. V., Lee, C.Y., Yoshimura, T., Shoemaker, D. and Yang, C.C.S. 2019. Genetic Diversity and *Wolbachia* Infection Patterns in a Globally Distributed Invasive Ant. *Frontiers in Genetics*, 10, pp.838.
- 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, *Acyrtosiphon pisum*. *Molecular Ecology*, 11 (10), pp.2123–2135.
- Turner, D.A. and Brightwell, R. 1986. An evaluation of a sequential aerial spraying operation against *Glossina pallidipes* Austen (Diptera: Glossinidae) in the Lambwe Valley of Kenya: aspects of post-spray recovery and evidence of natural population regulation. *Bulletin of Entomological Research*, 76 (2), pp.331–349.
- Uematsu, S. and Akira, S. 2008. Toll-Like Receptors (TLRs) and Their Ligands. *Handbook of Experimental Pharmacology*. Springer, Berlin, Heidelberg, 1–20.
- Ulevitch, R.J. and Tobias, P.S. 1999. Recognition of Gram-negative bacteria and endotoxin by the innate immune system. *Current Opinion in Immunology*, 11 (1), pp.19–22.
- Unckless, R.L. and Lazzaro, B.P. 2016a. The potential for adaptive maintenance of diversity in insect antimicrobial peptides. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371 (1695), pp.20150291.

- Unckless, R.L. and Lazzaro, B.P. 2016b. The potential for adaptive maintenance of diversity in insect antimicrobial peptides. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371 (1695), pp.20150291.
- Unckless, R.L., Howick, V.M. and Lazzaro, B.P. 2016. Convergent Balancing Selection on an Antimicrobial Peptide in *Drosophila*. *Current Biology*, 26 (2), pp.257–262.
- Ung, M.U., Lu, B. and McCammon, J.A. 2006. E230Q mutation of the catalytic subunit of cAMP-dependent protein kinase affects local structure and the binding of peptide inhibitor. *Biopolymers*, 81 (6), pp.428–439.
- Urban, C.F., Reichard, U., Brinkmann, V. and Zychlinsky, A. 2006. Neutrophil extracellular traps capture and kill *Candida albicans* yeast and hyphal forms. *Cellular Microbiology*, 8 (4), pp.668–676.
- Ursic-Bedoya, R., Buchhop, J., Joy, J.B., Durvasula, R. and Lowenberger, C. 2011. Prolixicin: a novel antimicrobial peptide isolated from *Rhodnius prolixus* with differential activity against bacteria and *Trypanosoma cruzi*. *Insect Molecular Biology*, 20 (6), pp.775–786.
- Utturkar, S.M., Klingeman, D.M., Hurt, R.A. and Brown, S.D. 2017. A case study into microbial genome assembly gap sequences and finishing strategies. *Frontiers in Microbiology*, 8 (JUL), pp.1272.
- Valanne, S., Wang, J.H. and Rämet, M. 2011. The *Drosophila* Toll Signaling Pathway. *The Journal of Immunology*, 186 (2), pp.649–656.
- Vale, G.A., Lovemore, D.F., Flint, S. and Cockbill, G.F. 1988. Odour-baited targets to control tsetse flies, *Glossina* spp. (Diptera: Glossinidae), in Zimbabwe. *Bulletin of Entomological Research*, 78 (1), pp.31–49.
- Vale, G.A., Hargrove, J.W., Lehane, M.J., Solano, P. and Torr, S.J. 2015. Optimal Strategies for Controlling Riverine Tsetse Flies Using Targets: A Modelling Study. *PLOS Neglected Tropical Diseases*, 9 (3), pp.e0003615.
- Vale, G.A., Hargrove, J.W., Chamisa, A., Grant, I.F. and Torr, S.J. 2015. Pyrethroid Treatment of Cattle for Tsetse Control: Reducing Its Impact on Dung Fauna. *PLOS Neglected Tropical Diseases*, 9 (3), pp.e0003560.
- Van Valen, L. 1973. A new evolutionary law. *Evolutionary Theory*, 1, pp.1–30.

- Varkey, J., Singh, S. and Nagaraj, R. 2006. Antibacterial activity of linear peptides spanning the carboxy-terminal  $\beta$ -sheet domain of arthropod defensins. *Peptides*, 27 (11), pp.2614–2623.
- Vedithi, S.C., Malhotra, S., Das, M., Daniel, S., Kishore, N., George, A., Arumugam, S., Rajan, L., Ebenezer, M., Ascher, D.B., Arnold, E. and Blundell, T.L. 2018. Structural Implications of Mutations Conferring Rifampin Resistance in *Mycobacterium leprae*. *Scientific Reports 2018 8:1*, 8 (1), pp.1–12.
- Venkatesan, M., Westbrook, C.J., Hauer, M.C. and Rasgon, J.L. 2007. Evidence for a Population Expansion in the West Nile Virus Vector *Culex tarsalis*. *Molecular Biology and Evolution*, 24 (5), pp.1208–1218.
- Vreysen, M.J.B., Saleh, K.M., Ali, M.Y., Abdulla, A.M., Zhu, Z.R., Juma, K.G., Dyck, V.A., Msangi, A.R., Mkonyi, P.A. and Feldmann, H.U. 2000. *Glossina austeni* (Diptera: Glossinidae) Eradicated on the Island of Unguja, Zanzibar, Using the Sterile Insect Technique. *Journal of Economic Entomology*, 93 (1), pp.123–135.
- Wachinger, M., Kleinschmidt, A., Winder, D., Von Pechmann, N., Ludvigsen, A., Neumann, M., Holle, R., Salmons, B., Erfle, V. and Brack-Werner, R. 1998. Antimicrobial peptides melittin and cecropin inhibit replication of human immunodeficiency virus 1 by suppressing viral gene expression. *Journal of General Virology*, 79 (4), pp.731–740.
- Wamwiri, F.N. and Changasi, R.E. 2016. Tsetse Flies (*Glossina*) as vectors of human African Trypanosomiasis: A review. *BioMed Research International*, 2016.
- Wang, J., Hu, C., Wu, Y., Stuart, A., Amemiya, C., Berriman, M., Toyoda, A., Hattori, M. and Aksoy, S. 2008. Characterization of the antimicrobial peptide attacin loci from *Glossina morsitans*. *Insect Molecular Biology*, 17 (3), pp.293–302.
- 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), pp.12133–12138.
- Ware, F.L. and Luck, M.R. 2017. Evolution of salivary secretions in haematophagous animals. *Bioscience Horizons: The International Journal of Student Research*, 10, pp.1–

20.

- Watari, A., Iwabe, N., Masuda, H. and Okada, M. 2010. Functional transition of Pak protLynch, M. 2010. Evolution of the mutation rate. *Trends in Genetics*, 26 (8), 345–352. Available from <https://doi.org/10.1016/J.TIG.2010.05.003>. o-oncogene during early evolution of metazoans. *Oncogene* 2010 29:26, 29 (26), pp.3815–3826.
- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., De Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R. and Schwede, T. 2018. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46 (W1), pp.W296–W303.
- Wayne, M.L., Contamine, D. and Kreitman, M. 1996. Molecular population genetics of ref(2)P, a locus which confers viral resistance in *Drosophila*. *Molecular Biology and Evolution*, 13 (1), pp.191–199.
- Weaver, S., Shank, S.D., Spielman, S.J., Li, M., Muse, S. V and Kosakovsky Pond, S.L. 2018. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Molecular Biology and Evolution*, 35 (3), pp.773–777.
- Weber, A.N.R., Tauszig-Delamasure, S., Hoffmann, J.A., Lelièvre, E., Gascan, H., Ray, K.P., Morse, M.A., Imler, J.L. and Gay, N.J. 2003. Binding of the *Drosophila* cytokine Spätzle to Toll is direct and establishes signaling. *Nature Immunology*, 4 (8), pp.794–800.
- Weitz, B. 1963. The feeding habits of *Glossina*. *Bulletin of the World Health Organization*, 28 (5–6), pp.711–729.
- Wheeler, T.J., Clements, J. and Finn, R.D. 2014. Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics*, 15 (1), pp.7.
- Whelan, S. and Goldman, N. 2001. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution*, 18 (5), pp.691–699.
- Wiesner, J. and Vilcinskas, A. 2010. Antimicrobial peptides: The ancient arm of the human immune system. *Taylor & Francis*, 1 (5), pp.440–464.
- Wlasiuk, G. and Nachman, M.W. 2010. Adaptation and Constraint at Toll-Like Receptors in



- Primates. *Molecular Biology and Evolution*, 27 (9), pp.2172–2186.
- Woolhouse, M.E.J., Webster, J.P., Domingo, E., Charlesworth, B. and Levin, B.R. 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genetics*, 32 (4), pp.569–577.
- Wu, L.P. and Anderson, K. V. 1998. Regulated nuclear import of Rel proteins in the *Drosophila* immune response. *Nature*, 392 (6671), pp.93–97.
- Xu, Y., Tao, X., Shen, B., Horng, T., Medzhitov, R., Manley, J.L. and Tong, L. 2000. Structural basis for signal transduction by the Toll/interleukin-1 receptor domains. *Nature*, 408 (6808), pp.111–115.
- Yang, J. and Zhang, Y. 2015. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Research*, 43 (W1), pp.W174–W181.
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. and Zhang, Y. 2015. The I-TASSER Suite: protein structure and function prediction. *Nature Methods*, 12 (1), pp.7–8.
- Yang, L., Harroun, T.A., Weiss, T.M., Ding, L. and Huang, H.W. 2001. Barrel-Stave Model or Toroidal Model? A Case Study on Melittin Pores. *Biophysical Journal*, 81 (3), pp.1475–1485.
- Yeaman, M.R. and Yount, N.Y. 2003. Mechanisms of Antimicrobial Peptide Action and Resistance. *Pharmacological Reviews*, 55 (1), pp.27–55.
- Yi, H.Y., Chowdhury, M., Huang, Y.D. and Yu, X.Q. 2014. Insect antimicrobial peptides and their applications. *Applied Microbiology and Biotechnology*, 98 (13), pp.5807–5822.
- Yue, F., Shi, J. and Tang, J. 2009. Simultaneous phylogeny reconstruction and multiple sequence alignment. *BMC Bioinformatics*, 10 (Suppl 1), pp.S11.
- Zaidman-Rémy, A., Hervé, M., Poidevin, M., Pili-Floury, S., Kim, M.S., Blanot, D., Oh, B.H., Ueda, R., Mengin-Lecreulx, D. and Lemaitre, B. 2006. The *Drosophila* Amidase PGRP-LB Modulates the Immune Response to Bacterial Infection. *Immunity*, 24 (4), pp.463–473.
- Zhang, G. and Ghosh, S. 2002. Negative Regulation of Toll-like Receptor-mediated Signaling by Tollip. *Journal of Biological Chemistry*, 277 (9), pp.7059–7065.

- Zhang, Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9 (1), pp.40.
- Zhang, Y. and Skolnick, J. 2005. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33 (7), pp.2302–2309.
- Zhou, K., Kanai, R., Lee, P., Wang, H.W. and Modis, Y. 2012. Toll-like receptor 5 forms asymmetric dimers in the absence of flagellin. *Journal of Structural Biology*, 177 (2), pp.402–409.
- Zou, Z., Evans, J.D., Lu, Z., Zhao, P., Williams, M., Sumathipala, N., Hetru, C., Hultmark, D. and Jiang, H. 2007. Comparative genomic analysis of the *Tribolium* immune system. *Genome Biology* 2007 8:8, 8 (8), pp.1–16.
- Zuckerandl, E. and Pauling, L. 1965. Evolutionary Divergence and Convergence in Proteins. *Evolving Genes and Proteins*. Elsevier, 97–166.