

# **Comparative structural genomics and phylogenomics of African Trypanosomes**

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

**Ali Hadi Abbas**

**April 2018**



UNIVERSITY OF  
LIVERPOOL

## **Dedication**

**To my Father's soul, my beloved Mother and my lovely wife and children Mohammed and Jannat, to my brothers I dedicate this work.**

## Acknowledgments

Firstly, I'd like to thank my supervisors Neil, Alistair and Christiane for their support throughout this journey, especially Neil Hall and Alistair Darby who gave me the opportunity to work on the cutting age sequencing technology and comparative genomic approaches utilized in this project.

Also, I want to thank ministry of higher education in Iraq, University of Kufa and Faculty of veterinary medicine who gave me this magnificent opportunity to complete my PhD degree abroad and for funding this project. Special thanks, to the head of faculty of veterinary medicine who supported me during the process of scholarship and after.

Many thanks to all CGR team and PhD students in Neil's group who give me support throughout my study. Especially, Margaret for generating PacBio DNA libraries of *T. vivax* genome sequencing and Charlotte for generating Illumina paired-end Reads of *T. congolense* genome sequencing; Richard, Linda Fatima and Laura for their support in data management and annotation. Also, I'd like to thank Institute of Integrative Biology for providing useful modules to develop the foundations for the genomics and bioinformatics knowledge.

I'd like to thank Liam Morrison, Bill Wickstead and their labs for providing the genomic DNA of *T. congolense* IL3000 and *T. vivax* IL1392 and Andrew Jackson for providing the PacBio assembly of *T. congolense* Tc1/148.

Finally, many thanks to my mum, my wife and my brothers who supported and look after me and all family members and friends who supported me even in words throughout this journey.

# Comparative structural genomics and phylogenomics of African Trypanosomes

Ali Hadi Abbas

## Abstract

The pathogens responsible of the majority of African Animal trypanosomiasis (AAT) are *Trypanosoma brucei*, *T. congolense* and *T. vivax*. These three trypanosomes have very different biology both in the mammalian host and the insect vector. These differences are encoded in their genomes and whole genome sequence comparison of high quality genomic data should allow such comparisons to be performed. Whilst the *T.b. brucei* strain TREU927 genome assembly is currently available as a good quality draft, the current versions of *T. congolense* and *T. vivax* are highly fragmented and include large gaps that interrupt genes and physical integrity of genome assembly. Therefore, there is a need to produce high quality *de novo* genome assemblies of both *T. congolense* strain IL3000 and *T. vivax* IL1392. The most appropriate technology for this currently are the third generation sequencing such as PacBio SMRT long reads sequencing. This technology's long reads permit assembly to a high standard with good contig lengths, and more completed gene models. Comparative genomic analysis carried out on *T. brucei* large chromosomes and the new PacBio *de novo* assemblies in this thesis *T. congolense* and *T. vivax* uncovered putative large structural chromosomal rearrangements between the African trypanosomes. The most extensive examples were noticed between *T. vivax* and *T. brucei*, which might reflect the early divergence of former to the most recent divergence of the latter in the phylogenetic tree of African trypanosomes.

Remarkably, analysis of *T. congolense* genome assembly facilitated the full description of (minichromosomes) devoted to harbouring genes utilized mainly in mammalian host immune evasion mechanism. Subsequently, a new gene family consisting of about 30 genes/pseudogenes encode for putative surface proteins (ESAG3- like proteins) was assigned to this trypanosome most likely only found on these chromosomes.

The phylogenomic analysis of free living and parasitic kinetoplastids uncovered possible core kinetoplastids, parasitic and African trypanosome specific gene sets. Likewise, the genomic repertoire of some stage specific proteins like Haptoglobin-Hemoglobin receptor, PAG, BARP, ESAG6/7-transferrin like proteins were also present among analysed African trypanosomes.

In conclusion, this project has provided genomes that give access to new genomic regions especially, the minichromosomes in two strains *T. congolense* and other interesting sequences of repeated nature in MBCs like centromeres, moreover, it shows unprecedented chromosomal rearrangements across the three African trypanosomes. Finally, phylogenomic analysis revealed for the first time the genomic repertoire of Haptoglobin-Hemoglobin receptor in the African trypanosomes.

# Table of contents

<b>Dedication</b> .....	<b>i</b>
<b>Acknowledgments</b> .....	<b>ii</b>
<b>Abstract</b> .....	<b>iii</b>
<b>Table of contents</b> .....	<b>iv</b>
<b>List of figures</b> .....	<b>viii</b>
<b>List of tables</b> .....	<b>xii</b>
<b>List of abbreviations</b> .....	<b>xiv</b>
<b>Chapter 1 Introduction</b> .....	<b>1</b>
<b>1.1 Classification of Trypanosomes</b> .....	<b>1</b>
1.1.1 The genus <i>Trypanosoma</i> .....	1
<b>1.2 <i>Trypanosoma</i> life cycle</b> .....	<b>2</b>
1.2.1 The life cycle of <i>T. brucei</i> .....	3
1.2.2 The life cycle of <i>T. congolense</i> with respect to that of <i>T. brucei</i> .....	4
1.2.3 The life cycle of <i>T. vivax</i> .....	5
<b>1.3 Cell surface proteins and trypanosomes life cycle</b> .....	<b>7</b>
<b>1.4 <i>Trypanosoma</i> cell morphology</b> .....	<b>10</b>
<b>1.5 African Trypanosomes Diseases and impact</b> .....	<b>12</b>
1.5.1 Human African trypanosomiasis .....	12
1.5.2 Animal African trypanosomiasis .....	13
1.5.3 Treatment of Trypanosomiasis .....	13
1.5.4 AAT economic impact .....	14
<b>1.6 Trypanosome genome and chromatin architecture</b> .....	<b>14</b>
1.6.1 Nucleosomes .....	14
1.6.2 Organization of the chromatin in the eukaryotic nucleus .....	15
1.6.3 Trypanosome genomes .....	16
<b>1.7 Chromosomal rearrangements in eukaryotes</b> .....	<b>24</b>
1.7.1 Forms of chromosomal rearrangements .....	24
1.7.2 Genomic DNA sequences related to chromosomal rearrangements ...	25
1.7.3 Impact of chromosomal rearrangements on gene expression .....	31
1.7.4 Previous comparative genomic analyses of African trypanosomes .....	31

1.8	<b>Aims of the thesis</b> .....	<b>32</b>
<b>Chapter 2</b>	<b><i>T. congolense</i> PacBio SMRT genome sequencing</b> .....	<b>35</b>
2.1	<b>Introduction</b> .....	<b>35</b>
2.1.1	The origin of <i>T. congolense</i> strain IL3000 .....	35
2.1.2	Previous genome sequence effort .....	35
2.1.3	Aims and objectives of the chapter .....	36
2.2	<b>Methods</b> .....	<b>38</b>
2.2.1	gDNA QC.....	38
2.2.2	Preparation of gDNA libraries for PacBio sequencing .....	39
2.2.3	Genome databases .....	41
2.2.4	<i>De novo</i> Genome assemblies.....	41
2.2.5	Assessment of the <i>de novo</i> PacBio TcIL3000 genome .....	42
2.2.6	Generation of scaffold level assembly .....	42
2.2.7	Chromosomal–level assembly of MBCs (pseudo-chromosomes) .....	42
2.2.8	Genome Annotation .....	45
2.2.9	Clustering of proteomic data of analysed assemblies .....	48
2.2.10	Gene Ontology enrichment analysis .....	49
2.2.11	Genome visualization.....	50
2.2.12	Synteny to the reference chromosomes .....	51
2.2.13	Plotting annotated features on <i>T. congolense</i> MBCs.....	52
2.2.14	Inference of putative centromeres .....	52
2.2.15	Strand Switch regions on <i>T. congolense</i> PacBio assembly.....	53
2.2.16	Statistical analysis.....	53
2.3	<b>Results and discussion</b> .....	<b>54</b>
2.3.1	RSII PacBio sequencing and <i>de novo</i> assembly .....	54
2.3.2	Genome Assembly .....	54
2.3.3	Generation of assembly scaffolds.....	59
2.3.4	Chromosomal–level Pseudo Scaffolds.....	62
2.3.5	Sequence error correction and gap filling .....	63
2.3.6	Possible Caveats to pseudochromosomes' level assembly .....	68
2.3.7	Genome annotation of <i>T. congolense</i> chromosomal level assembly ...	70
2.3.8	Comparison of the annotation between the two assemblies of <i>T. congolense</i> IL3000 and the reference <i>T. brucei</i> TREU927 .....	71
2.3.9	The new findings in this genome sequencing project .....	75
2.4	<b>Conclusion</b> .....	<b>110</b>
<b>Chapter 3</b>	<b>The structure of <i>T. congolense</i> minichromosomes</b> .....	<b>112</b>

<b>3.1</b>	<b>Background</b> .....	<b>112</b>
<b>3.2</b>	<b>Aims and objectives of the chapter</b> .....	<b>114</b>
<b>3.3</b>	<b>Methods</b> .....	<b>115</b>
3.3.1	<i>Trypanosoma congolense</i> strain IL3000 genomic DNA. ....	115
3.3.2	Genome sequencing.....	115
3.3.3	<i>T. congolense</i> IL3000 Genome Annotation .....	115
3.3.4	Detection of possible <i>T. congolense</i> mini-chromosomes .....	116
3.3.5	Detection of other possible direct repeats .....	117
3.3.6	Sequence comparison analyses and visualization .....	117
3.3.7	Statistical test.....	117
3.3.8	Protein and DNA Sequence alignment .....	118
3.3.9	Genome visualization .....	118
3.3.10	Clustering of proteomic data of analyzed assemblies.....	118
<b>3.4</b>	<b>Results and discussion</b> .....	<b>119</b>
3.4.1	Generic structure of <i>T. congolense</i> IL3000 min-chromosomes .....	119
3.4.2	Central tandem repeat region (region one).....	120
3.4.3	Conserved relatively GC-rich regions (region two) .....	121
3.4.4	Variable subtelomeric regions stretched over 5 kb (region three) .....	121
3.4.5	Presence of an intact telomeric VSG gene.....	122
3.4.6	VSG pseudogenic sequences .....	125
3.4.7	Telomeric Spacer DNA sequence .....	126
3.4.8	Other features in TcMCs subtelomeric region .....	128
3.4.9	<i>ESAG6</i> , <i>ESAG2</i> , and ATP-dependent DEAD/H box RNA helicases..	129
3.4.10	<i>ESAG3</i> gene family in TcMCs .....	130
3.4.11	Telomeric repeats (region four).....	132
3.4.12	Intermediate chromosomes in <i>T. congolense</i> IL3000 (TcICs) .....	133
3.4.13	Comparison of TcMCs between <i>T. congolense</i> IL3000 and Tc1/148	
	134	
<b>3.5</b>	<b>Conclusions</b> .....	<b>140</b>
<b>Chapter 4</b>	<b><i>T. vivax</i> PacBio SMRT genome sequencing</b> .....	<b>141</b>
<b>5.1</b>	<b>Introduction</b> .....	<b>141</b>
4.1.1	<i>Trypanosoma vivax</i> strain used in this sequencing project .....	141
4.1.2	<i>T. vivax</i> previous genome sequencing efforts .....	142
4.1.3	Aims and objectives of this chapter .....	143
<b>4.2</b>	<b>Methods</b> .....	<b>144</b>
4.2.1	<i>T. vivax</i> gDNA and PacBio SMRT libraries preparation .....	144

4.2.2	Genome assembly databases and protein sequences used in this chapter.....	144
4.2.3	<i>De novo</i> genome assemblies .....	144
4.2.4	Assessment of gene models in <i>T. vivax de novo</i> assemblies.....	145
4.2.5	<i>T. vivax</i> IL1392 genome annotation.....	146
4.2.6	BLAST search.....	146
4.2.7	Genome visualization .....	147
4.2.8	Statistical analyses .....	147
4.2.9	Strand Switch regions .....	147
4.2.10	Directional gene clusters.....	147
4.2.11	Clustering of proteomic databases of analysed kinetoplastids .....	147
4.2.12	Obtaining the number of shared genes from OrthoFinder output ...	148
4.2.13	Plots generation of possible genome rearrangements .....	148
4.2.14	Venn Diagrams .....	149
4.2.15	Generation of assembly statistics .....	149
4.2.16	Gene Ontology enrichment analysis .....	149
<b>4.3</b>	<b>Results and discussion.....</b>	<b>150</b>
4.3.1	SMRT sequencing and <i>de novo</i> Genome assembly .....	150
4.3.2	<i>T. vivax</i> PB genome annotation.....	158
4.3.3	Assessment of <i>T. vivax</i> IL1392 PacBio genome annotation and validation .....	160
4.3.4	Possible new findings in this <i>T. vivax</i> PacBio assembly.....	164
<b>4.4</b>	<b>Conclusion .....</b>	<b>177</b>
<b>Chapter 5</b>	<b>Comparative phylogenomic analysis of Kinetoplastids with a focus on African Trypanosomes .....</b>	<b>178</b>
<b>5.1</b>	<b>Introduction.....</b>	<b>178</b>
<b>5.2</b>	<b>Methods .....</b>	<b>181</b>
5.2.1	Selected Kinetoplastids for the phylogenomic analysis .....	181
5.2.2	Clustering of kinetoplastids proteomes.....	181
5.2.3	Generation of phylogenetic trees in this chapter .....	181
5.2.4	Obtaining the designed sequence sets from OrthoFinder output .....	185
5.2.5	Gene Ontology enrichment analysis.....	185
5.2.6	Searching gene bank database .....	186
<b>5.3</b>	<b>Results and discussion.....</b>	<b>186</b>
5.3.1	Kinetoplastids core gene families .....	186
5.3.2	Specific gene sets that have evolved in parasite lineages .....	194

5.3.3	Specific genes set that have evolved in African trypanosomes lineages	198
<b>5.4</b>	<b>Conclusions</b>	<b>217</b>
<b>Chapter 6</b>	<b>Conclusion and future directions</b>	<b>218</b>
<b>6.1</b>	<b>Conclusions</b>	<b>218</b>
6.1.1	Genomic rearrangements and new findings in African trypanosomes	218
6.1.2	<i>T. congolense</i> minichromosomes and potential expression sites	221
6.1.3	Phylogenomic analysis of Kinetoplastids and African Trypanosomes	222
<b>6.2</b>	<b>Future perspective</b>	<b>223</b>
<b>References</b>		<b>224</b>
<b>Appendix A</b>	<b>Perl script and command lines</b>	<b>278</b>
<b>Appendix B</b>	<b><i>T. congolense</i> expression sites draft paper</b>	<b>281</b>
<b>6.3</b>	<b>Parasite stocks and culture</b>	<b>286</b>
<b>6.4</b>	<b>DNA extraction and sequencing</b>	<b>286</b>
<b>6.5</b>	<b>Assembly and annotation</b>	<b>288</b>
<b>6.6</b>	<b>Telomere assembly and annotation</b>	<b>288</b>
<b>6.7</b>	<b>Multiple Sequence Alignment</b>	<b>289</b>
<b>6.8</b>	<b>Phylogenetic Analysis</b>	<b>289</b>
<b>Appendix C</b>	<b>Additional materials of Chapter 4 and 5</b>	<b>319</b>

## List of figures

FIGURE 1.1	AFRICAN TRYPANOSOMES CYCLICAL LIFE CYCLE	9
FIGURE 1.2	3D ULTRASTRUCTURE OF <i>T. BRUCEI</i> CELL	11
FIGURE 1.3	TRANSMISSION ELECTRON MICROSCOPY SHOWS NUCLEAR ULTRASTRUCTURE AND COMPARTMENTALIZATION OF THE <i>T. BRUCEI</i> NUCLEUS	19
FIGURE 2.1	1% AGAROSE GEL OF <i>TcIL3000</i> gDNA	39
FIGURE 2.2	BIOANALYZER ANALYSES OF 20KB AND 10KB gDNA LIBRARIES OF <i>TcIL3000</i> . A) 20 KB LIBRARY (MEAN SIZE) OF 9 KB	40
FIGURE 2.3	MANUAL ANNOTATION WORK FLOW OF <i>T. CONGOLENSE</i> IL3000 PACBIO GENOME ASSEMBLY	46

FIGURE 2.4 <i>TcIL3000</i> SUB READS DISTRIBUTION OF THE 12 SMRT CELLS OUTPUT. .....	54
FIGURE 2.5 BUSCO COMPARISON OF PACBIO CONTIG LEVEL ASSEMBLIES.....	57
FIGURE 2.6 SEQUENCE COMPARISON OF <i>T. CONGOLENSE</i> PACBIO SCAFFOLDS TO THE TB927 CHROMOSOMES SEQUENCES.....	61
FIGURE 2.7 ILLUMINA READS COVERAGE USED FOR ERROR CORRECTION AND GAP FILLING ACROSS FINAL PSEDOCHROMOSOMAL LEVEL <i>TcIL3000</i> PB ASSEMBLY. .....	65
FIGURE 2.8 BUSCO COMPARISON OF CHROMOSOMAL LEVEL <i>T. CONGOLENSE</i> PACBIO ASSEMBLY BEFORE AND AFTER ERROR CORRECTION AND GAP FILLING STEPS.....	66
FIGURE 2.9 BUSCO TOOLS TEST OF CHROMOSOMAL LEVEL ASSEMBLIES OF <i>TB</i> <i>927</i> , <i>T. CONGOLENSE</i> SANGER AND <i>T. CONGOLENSE</i> PACBIO.....	67
FIGURE 2.10 <i>T. CONGOLENSE</i> PSEUDOCROMOSOME 6 INFERRED BY ABACAS1 USING GENOME SYNTENY ACCORDING TO REFERENCE GENOME TB927.....	68
FIGURE 2.11 THE PREDICTED FEATURES ANNOTATED TO THE <i>T. CONGOLENSE</i> PACBIO MBCs. ANNOTATED GENES ON MBCs WERE SHOWED ACCORDING TO THE COLOUR KEY. THE NUMBERS UNDERNEATH IS THE LENGTH SCALE IN 500 BP INCREMENT, NUMBERS EQUAL OR LARGER THAN 1 ARE IN Mb LENGTH. KARYOPLOTEr PACKAGE ON R-PROJECT WAS USED TO GENERATE THIS PLOT ACCORDING TO THE FEATURE COORDINATES STORED IN THE GFF3 FILE. ....	74
FIGURE 2.12 PUTATIVE CENTROMERE REPEATS IN <i>T. CONGOLENSE</i> PACBIO SEQUENCE ASSEMBLY CHROMOSOMES1, CHROMOSOME3, CHROMOSOME6, CHROMOSOME11 AND CHROMOSOME4 (DENOTED BY RED BRACKETS) CHARACTERIZED BY LOW GC CONTENT (GREEN SHADED) AND CONSIST OF 136 BP AND 247 BP (FOR THE CHR4 AND CHR11) IN LENGTH OF SINGLE REPEAT UNIT.....	80
FIGURE 2.13 AN ACT COMPARISON PLOT OF PROPOSED CHROMOSOMAL REARRANGEMENTS BETWEEN TB927 CHROMOSOME ONE (TOP BAR), <i>T.</i> <i>CONGOLENSE</i> PACBIO CHR2 (MIDDLE BAR).....	83
FIGURE 2.14 A CONTIG HAS THE IDENTIFIER (scf180000002797) OF STRAIN Tc1/148 PACBIO ASSEMBLY SHOWING SIMILAR PUTATIVE REARRANGEMENT TO THAT OF STRAIN IL3000. ....	84

FIGURE 2.15 CHROMOSOMAL DISPLACEMENT OF <i>T. CONGOLENSIS</i> PACBIO CHR7 AND Tb927 CHR1 AFFECTED MAINLY PREDICTED GENES ENCODING FOR ALPHA AND BETA TUBULIN ON Tb927 CHROMOSOME ONE. ....	86
FIGURE 2.17 PUTATIVE INTER AND INTRA- CHROMOSOMAL REARRANGEMENTS ON <i>T. CONGOLENSIS</i> PACBIO CHR10.....	89
FIGURE 2.18 GENOMERIBON PLOT OF THE DISPLACEMENT IN <i>T. CONGOLENSIS</i> OF A SEGMENT BELONGING TO Tb927 CHR1 (PINK RIBBONS) ASSOCIATED WITH A REGION THAT HAS PREDICTED GENES LINKED TO CHR11 (ORANGE RIBBONS). 91	
FIGURE 2.19 A VENN DIAGRAM OF ORTHOFINDER CLUSTERING ANALYSIS FOR THE THREE ASSEMBLIES' PROTEOMIC DATA (Tb927, <i>T. CONGOLENSIS</i> PACBIO AND <i>T. CONGOLENSIS</i> SANGER).....	97
FIGURE 2.20 DISTRIBUTION OF NEW GENES OF <i>T. CONGOLENSIS</i> PACBIO ASSEMBLY SHARED WITH Tb927 ON THE PUTATIVE MBCs OF <i>T. CONGOLENSIS</i> PACBIO ASSEMBLY. ....	100
FIGURE 2.21 REVIGO SUMMARY OF GO IDs WITH MOLECULAR FUNCTIONS PINNED TO THE NEW <i>T. CONGOLENSIS</i> PACBIO GENES. ....	103
FIGURE 2.24 A KARYO PLOT OF <i>T. CONGOLENSIS</i> PACBIO MBCs REVEALS THE NEW PROPOSED SINGLE GENE LOCATIONS ON THE MBCs.....	107
FIGURE 2.25 GO TERMS ENRICHMENT ANALYSIS HIGHLIGHT THE PROPOSED CELLULAR PATHWAYS THAT THESE GENES COULD CONTRIBUTE TO.....	108
FIGURE 3.1 A GENERIC MODEL OF A COMPLETE MINI-CHROMOSOME AND THE PROPOSED SUBSETS OF TcMCs ACCORDING TO THEIR SUBTELOMERIC FEATURES (MODELS NOT TO SCALE) .....	120
FIGURE 3.2 GENOME RIBBON DOT PLOT OF NUCMMER COMPARISON OF TWO COMPLETE TcMCs. ....	121
FIGURE 3.4 DENSITY PLOT OF PROTEIN LENGTH OF PUTATIVE TELOMERIC TcMCs VSGs.....	123
FIGURE 3.5 SPACER DNA ALIGNMENT OF TcMCs PSEUDOGENIC VSG GENES ON SUBTELOMERES OF THE MINI-CHROMOSOMES. ....	127
FIGURE 3.6 BOXPLOT OF SPACER DNA LENGTH A CROSS TcMCs TELOMERIC FEATURES. ....	128
FIGURE 3.7 A GENOMERIBBON PLOT OF NUCMMER COMPARISON.....	129
FIGURE 3.8 ALIGNMENT OF TcMCs ESAG3 AND OTHER TRYPANOSOME ESAG3 PROTEIN SEQUENCES. ....	131

FIGURE 3.9 BOXPLOT OF THE LENGTH OF TELOMERIC REPEATS IN TcMCs WITH TWO SUBTELOMERIC FEATURES OF <i>T. CONGOLENSE</i> IL 3000. ....	133
FIGURE 3.10 GENOMERIBBON PLOT OF NUCMMER COMPARISON OF <i>T. CONGOLENSE</i> STRAIN Tc1/148 PUTATIVE MC (BOTTOM THICK BLACK LINE) AND STRAIN IL3000 COMPLETE SEVEN SUGGESTED MCs TOPE MULTICOLOUR PANEL. ....	138
FIGURE 3.11 AN ARTEMIS PLOT OF A COMPLETE MC OF STRAIN Tc1/148 (21,309 BP LONG) SHOWED A NEW MODEL OF MC.....	139
FIGURE 4.1 DISTRIBUTION OF SEQUENCED READS SEQUENCED FROM DNA LIBRARIES INPUT PREPARED FROM GDNA OF <i>T. VIVAX</i> STRAIN IL1392.....	151
FIGURE 4.2 GENE MODEL INTEGRITY ASSESSMENT USING CORE PROTISTS GENE SET APPLIED BY BUSCO TOOLS ACROSS OUR PB ASSEMBLIES OF <i>T. VIVAX</i> . ....	155
FIGURE 4.3 GENOME SYNTENY OF PACBIO <i>DE NOVO</i> ASSEMBLIES' CONTIGS OF <i>T. CONGOLENSE</i> (A) AND <i>T. VIVAX</i> (B) TO THE REFERENCE <i>T. BRUCEI</i> MBCs. .	157
FIGURE 4.4 A VENN DIAGRAM OF ORTHOFINDER PROTEIN CLUSTERS OF Tb927, <i>T. VIVAX</i> PACBIO AND <i>T. VIVAX</i> SANGER ASSEMBLY PROTEOMIC DATABASES. ....	162
FIGURE 4.5 A VENN DIAGRAM OF ORTHOFINDER PROTEIN SEQUENCE CLUSTERING OF <i>T. VIVAX</i> PACBIO PROTEOME AND THE CURRENTLY AVAILABLE SANGER SEQUENCE ASSEMBLY PROTEOME WERE BOTH ANNOTATED USING LOCALLY INSTALLED COMPANION ANNOTATION PIPELINE. ....	163
FIGURE 4.6 SEMANTIC GENE ONTOLOGY ENRICHMENT ANALYSIS PLOT OF <i>T. VIVAX</i> PACBIO SINGLETONS PROTEIN SEQUENCES SHOWS BIOLOGICAL PATHWAYS. ....	166
FIGURE 4.7 SEMANTIC GENE ONTOLOGY ENRICHMENT ANALYSIS PLOT OF <i>T. VIVAX</i> PACBIO SINGLETONS PROTEIN SEQUENCES SHOWS MOLECULAR FUNCTIONS. ....	167
FIGURE 4.8 SEMANTIC GENE ONTOLOGY ENRICHMENT ANALYSIS PLOT OF <i>T. VIVAX</i> PACBIO SINGLETONS PROTEIN SEQUENCES SHOWS CELLULAR LOCALIZATION. ....	168
FIGURE 4.9 GENOMERIBBON PLOT OF THE LONGEST TWO CONTIGS OF <i>T. VIVAX</i> PACBIO ASSEMBLY SHOWED REGIONS OF SYNTENY WITH REFERENCE GENOME ASSEMBLY OF Tb927 VERSION 5.1 PERMITTING ILLUSTRATION OF PREDICTED COORDINATES BY PROMMER. ....	170

FIGURE 5.1 ORTHOFINDER PIPELINE OPTIONS USED IN THE PHYLOGENETIC ANALYSES OF KINETOPLASTID PROTEOMIC DATA. ....	184
FIGURE 5.2 KINETOPLASTID SPECIES TREE WITH A CORRESPONDING HORIZONTAL BAR GRAPH HIGHLIGHTING THE SIZE OF PROTEOME PER SPECIES AND THE DISTRIBUTION OF MAIN GENE CLUSTERS. ....	187
FIGURE 5.3 FUNCTIONAL ENRICHMENT ANALYSIS OF CORE PROTEIN KINETOPLASTIDS OF MULTIPROTEIN CLUSTERS ACROSS DIFFERENT PARASITIC AND THE FREE LIVING KINETOPLASTIDS. ....	190
FIGURE 5.4 FUNCTIONAL ENRICHMENT ANALYSIS OF 1:1 ORTHOLOGUES OF THE PARASITIC KINETOPLASTIDS. ....	195
FIGURE 5.5 FUNCTIONAL ENRICHMENT ANALYSIS OF CLUSTERS WITH VARIABLE NUMBER OF PROTEIN SEQUENCES OF PARASITIC KINETOPLASTIDS. ....	197
FIGURE 5.6 PHYLOGENETIC TREE OF HAPTOGLOBIN-HEMOGLOBIN RECEPTOR IN AFRICAN TRYPANOSOMES. <i>T. BRUCEI</i> HAS A SINGLE GENE ON CHROMOSOME SIX, WHICH SHOWED A MASSIVE GENE LOSS COMPARABLY TO <i>T. CONGOLENSE</i> AND <i>T. VIVAX</i> , WHICH EXHIBITED VARIABLE TANDEM EXPANSIONS. ....	203
FIGURE 5.7 PHYLOGENETIC TREE OF PREDICTED PROCYCLIC ASSOCIATED PROTEIN PAG SEQUENCES OF <i>T. BRUCEI</i> (RED) AND <i>T. CONGOLENSE</i> (GREEN). ....	207
FIGURE 5.8 PHYLOGENETIC TREE OF <i>BRUCEI</i> ALANINE RICH REPEAT BARP AND GLUTAMIC/ACID ALANINE RICH REPEAT GARP SEQUENCES OF <i>T. BRUCEI</i> (RED) AND <i>T. CONGOLENSE</i> (GREEN), RESPECTIVELY. ....	208
FIGURE 5.9 PHYLOGENETIC TREE OF PUTATIVE PILIN ASSEMBLY DOMAIN (PILO) CONTAINING PROTEINS OF <i>T. CONGOLENSE</i> (GREEN) AND <i>T. VIVAX</i> (BLUE) ..	213

## List of tables

TABLE 1.1 AVAILABLE GENOME ASSEMBLIES OF AFRICAN TRYPANOSOMES. ....	32
TABLE 2.1 PACBIO CONTIG LEVEL ASSEMBLIES .....	56
TABLE 2.2 CONTIG LEVEL ASSEMBLY STATISTICS OF <i>T. CONGOLENSE</i> ASSEMBLIES. ....	59
TABLE 2.3 COMPARISON BETWEEN CONTIGS' ASSEMBLY AND SSPACE-LONGREADS SCAFFOLDS' ASSEMBLY. ....	60
TABLE 2.4 INFERENCE OF PSEUDOCHROMOSOME LEVEL ASSEMBLY OF TcIL3000 USING Tb927 AS A REFERENCE GENOME SEQUENCE. ....	63

TABLE 2.5 GENOME FEATURES OF CURRENTLY AVAILABLE DRAFT SANGER ASSEMBLY, PACBIO ASSEMBLY OF <i>T. CONGOLENSE IL3000</i> AND THE REFERENCE <i>T. BRUCEI</i> TREU 927.....	73
TABLE 2.6 SIZE AND COORDINATES OF POSSIBLE CENTROMERE TANDEM REPEATS IN <i>T. CONGOLENSE</i> CHROMOSOMES 1, 3, 4, 6 AND 11. ....	76
TABLE 2.7 COMPARISON BETWEEN DIVERGENT SSRs AND CONVERGENT SSRs.	93
TABLE 2.8 PROPHESED DIRECTIONAL GENE CLUSTERS DGCs PER EACH OF TcIL3000 PREDICTED CHROMOSOMES.....	95
TABLE 2.9 NEW PROPOSED SHARED GENE FAMILIES OF <i>T. CONGOLENSE</i> PACBIO WITH Tb927 ACCORDING TO THE CLUSTERING ACHIEVED BY ORTHOFINDER. .	99
TABLE 2.10 GO SLIM CLASSIFICATION AND COUNTS OF 477 GO TERMS DERIVED FROM NEW 203 <i>T. CONGOLENSE</i> PACBIO GENES SHARED WITH TB927 GENES. ....	102
TABLE 2.11 THE PERCENTAGE OF GO SLIM CLASSIFICATION AND COUNT FOR THE NEW GENE SET OF <i>T. CONGOLENSE</i> PACBIO ORTHOLOGUES TO CORRESPONDING Tb927 GENES COMPARED TO THE ALL <i>T. CONGOLENSE</i> PACBIO GO ANNOTATIONS.....	104
TABLE 3.1 IDENTIFIED GENES IN THE SUBTELOMERIC REGIONS OF <i>TcIL3000</i> MINICHROMOSOMES. ....	124
TABLE 3.2 <i>T. CONGOLENSE</i> STRAIN Tc1/148 PACBIO SEQUENCE CONTIGS WITH PUTATIVE COMPLETE MCs.....	136
TABLE 4.1 <i>T. VIVAX</i> gDNA PACBIO <i>DE NOVO</i> ASSEMBLIES, ASSEMBLY STATISTICS OF DIFFERENT ASSEMBLERS USED TO GENERATE CONTIG LEVEL ASSEMBLAGE. ....	153
TABLE 4.2 COMPARISON BETWEEN PB HGAP2 BASED ASSEMBLY AND SANGER ASSEMBLY OF <i>T. VIVAX</i> USING DIFFERENT ASSEMBLY AND ANNOTATION PARAMETERS.....	159
TABLE 4.3 <i>T. VIVAX</i> CONTIGS THAT SHOWED REGIONS OF SYNTENY TO THE REFERENCE <i>T. BRUCEI</i> TREU927 GENES.....	171
TOTAL OF 82 SSRs (41 DSSRs AND 41 CSSRs) WITH NOTABLY LONGER SEQUENCES OF DSSRs COMPARED TO CSSRs (MEANS= 10,636, 3,625) BP WAS OBSERVED, RESPECTIVELY .....	173
TABLE 4.4 NUMERICAL COMPARISON BETWEEN DSSRs AND CSSRs OF <i>T. VIVAX</i> . ....	174

TABLE 5.1 MEDICALLY AND VETERINARILY IMPORTANT HETEROXENIC KINETOPLASTIDS. ....	180
TABLE 5.2 SELECTED KINETOPLASTIDS FOR PHYLOGENETIC STUDY. ....	183
TABLE 5.3 1:1 ORTHOLOGUES GENE FAMILIES OF CORE KINETOPLASTID PROTEIN SET. ....	188
TABLE 5.4 CLUSTERS OF MULTI-GENE FAMILIES INFERRED FROM PROTEIN PRODUCTS OF SHARED SIMILARITY ACROSS KINETOPLASTIDS. ....	193
TABLE 5.5 N: N SEQUENCE CLUSTERS AMONG PARASITIC KINETOPLASTIDS. ....	196
TABLE 5.6 N: N GENE FAMILIES AMONG AFRICAN TRYPANOSOMES. ....	202
TABLE 5.7 SHARED PROTEIN SEQUENCE CLUSTERS BETWEEN <i>T. BRUCEI</i> AND <i>T. CONGOLENSE</i> .....	206
TABLE 5.8 SHARED PROTEIN SEQUENCE CLUSTERS BETWEEN <i>T. CONGOLENSE</i> AND <i>T. VIVAX</i> . ....	212
TABLE 5.9 SPECIES-SPECIFIC ORTHOGROUPS OF AFRICAN TRYPANOSOMES ( <i>T. BRUCIE</i> , <i>T. CONGOLENSE</i> AND <i>T. VIVAX</i> ) .....	216

## List of abbreviations

<b>AAT:</b>	African Animal Trypanosomiasis.
<b>BARP:</b>	<i>Brucei</i> Alanine Rich Protein.
<b>BARP:</b>	<i>Brucei</i> Alanine- Rich Protein.
<b>BES:</b>	Blood Form Expression Sites.
<b>BFs:</b>	Blood Forms.
<b>Chr:</b>	Chromosome.
<b>CR:</b>	Chromosomal rearrangement.
<b>CSF:</b>	Cerebrospinal Fluid.
<b>CSSRs:</b>	Convergent Strand Switch Regions.
<b>DA:</b>	Diminazene Aceturate.

<b>DSSRs:</b>	Divergent Strand Switch Regions.
<b>EP:</b>	Glutamic acid and Proline amino acids rich repeat protein.
<b>ES:</b>	Expression Sites.
<b>ESAGs:</b>	Expression Site Association Genes.
<b>EtBr:</b>	Ethidium Bromide.
<b>GARP:</b>	Glutamic acid/Alanine- Rich Protein.
<b>GB:</b>	Gega base.
<b>gDNA:</b>	Genomic DNA.
<b>GFF:</b>	Gene File Format.
<b>GO:</b>	Gene Ontology enrichment terms.
<b>GPEET:</b>	Glycine, Proline, two Glutamic acid and threonine amino acids rich repeat protein.
<b>HAT:</b>	Human African Trypanosomiasis.
<b>HGAP2:</b>	Hierarchical Genome Assembly process version 2.
<b>HGAP3:</b>	Hierarchical Genome Assembly process version 3.
<b>HpHbR:</b>	Haptoglobin-Hemoglobin Receptor.
<b>HpHb:</b>	Haptoglobin-Hemoglobin.
<b>IC:</b>	Intermediate Chromosomes.
<b>ISM:</b>	Isometamidium Chloride.
<b>Kb:</b>	Kilobase.
<b>Mb:</b>	Megabase.
<b>MCs:</b>	mini-chromosomes.

**MES:** Metacyclic form expression sites.

**N50:** The length of the sequence when half of the genome lie in this contig length and higher.

**NAHR:** Non-Allelic Homologous Recombination.

**NCBI:** National Centre for Biotechnology Information advances science and health.

**ORFs:** Open Reading Frames.

**Orthogroups:** Protein clusters that were clustered according to protein sequence pairwise similarity resulted from all-vs-all BLASTp similarity search to all species in the analysis.

**PAG:** Procylic Associated Genes.

**PARP:** Procylic Acidic Repetitive Protein.

**PB:** PACBIO (Pacific BioScience).

**PM:** Peritrophic Matrix.

**Pseudo:** pseudogene.

**PTU:** Polycistronic Transcription Units.

**rRNA:** Ribosomal RNA.

**rRNA:** Ribosomal RNA.

**SMRT cells:** Single Molecule Real Time sequencing cells.

**SMRT:** Single Molecule Real Time.

**snoRNA:** Small nucleolar RNA.

**snRNA:** Small nuclear RNA.

**SSRs:** Strand Switch Regions.

**TADs:** Topological Associated Domains.

**Tb927:** *Trypanosoma brucei* strain TRUE 927.

**TbHpHbR:** *T. brucei* Haptoglobin-Hemoglobin Receptor.

**TcHpHbR:** *T. congolense* Haptoglobin-Hemoglobin Receptor.

**TcICs:** *T. congolense* Intermediate- sized chromosomes.

**TcIL3000:** *Trypanosoma congolense* strain IL3000.

**TcMCs:** *T. congolense* mini- chromosomes.

**TcPB:** *Trypanosoma congolense* strain IL3000 PACBIO assembly.

**TcSanger:** *Trypanosoma congolense* strain IL3000 Sanger based assembly (the current reference of *T. congolense* in the database).

**TE:** Transposable Elements.

**TRF:** Tandem Repeat Finder.

**tRNA:** Transfer RNA.

**TvPB:** *Trypanosoma vivax* strain Y486 PACBIO based sequence.

**TvSanger:** *Trypanosoma vivax* strain Y486 Sanger based sequence.

**tVSG:** Telomeric proximal Variant Surface Glycoprotein.

**VSGs:** Variant Surface Glycoproteins.

**WHO:** World Health Organization.

# Chapter 1 Introduction

## 1.1 Classification of Trypanosomes

The kinetoplastids belong to phylum *Euglenozoa* order *Kinetoplastida* which are characterized by the presence of the flagella and the kinetoplast (Hajduk, Siqueira and Vickerman, 1986). They are widespread throughout different environments such as aquatic fresh water and land environments, e.g. bacterivorous kinetoplastid *Bodo saltans* (Deschamps *et al.*, 2011), plant parasites e.g. the genus *Phytomonas* (Camargo, 1999; Stuart *et al.*, 2008; Jaskowska *et al.*, 2015); some are monoxenic insect parasites such as *Crithidia*, *Angomonas* and *Leptomonas* (Maslov *et al.*, 2013). The most medical and veterinary important members are the heteroxenous *Trypanosomatidae* that infect humans and animals causing serious acute or chronic illnesses and often transmitted by insect vectors.

The extracellular trypanosomes cause Human African Trypanosomiasis (HAT) caused by *T. brucei* subspecies *T. b. rhodesiense* and *T. b. gambiense* (Brun *et al.*, 2010; World Health Organization, 2013) and Animal African trypanosomiasis (AAT) which caused by *T. b. brucei*, *T. congolense* and *T. vivax* (Milligan and Baker, 1988). These species are transmitted by the blood feeding tsetse flies of the family *Glossinidae* (Lehane *et al.*, 2004; Watanabe *et al.*, 2014).

### 1.1.1 The genus *Trypanosoma*

This genus includes a number of protozoan parasites most importantly the African trypanosomes. Generally, *Trypanosoma* genus can be divided into two divergent sub genera; the *Salivarian* and *non-Salivarian* (or *Stercorarian*) trypanosomes. Members of these classes could cause human and/or animal diseases in Africa and America, respectively (Hoare, 1972; Haag, O'hUigin and Overath, 1998; Stevens *et al.*, 1999). Both sub genera need two hosts during their life cycle (heteroxenous life cycle): an insect vector and a vertebrate host. However, the mode of transmission and life cycle are different (Stevens and Gibson, 1999; Stevens *et al.*, 1999; Jackson, 2015).

### **1.1.1.1 *Salivarian* trypanosomes**

The HAT and AAT causative trypanosomes are within this subgenus, which involve *T. brucei*, *T. congolense* and *T. vivax* and develop in the tsetse fly (*Glossina* species) as an invertebrate vector in which, trypanosomes undergo developmental stages and a mammalian host (Stevens and Gibson, 1999).

### **1.1.1.2 *Stercorarian* trypanosomes**

The members of this subgenus are found in variety of hosts and could infect birds, reptiles as well as mammals, for example *T. cruzi* that causes Chagas disease in the human host in Central and South of America (Schmuñis, 2013; Truc *et al.*, 2013). The main vector of these trypanosomes is the *Triatomiae* blood sucking insect (kissing bugs). The trypanosomes develop in the alimentary tract of this insect after a blood meal from an infected vertebrate host. Unlike the *Salivarian* trypanosomes, the *Stercorarian* parasites are excreted with faeces of infected vector and contaminate the skin of the human host. They enter the blood stream via the bite wound (Rassi, Rassi and Marin-Neto, 2010; Ramsey *et al.*, 2015).

## **1.2 *Trypanosoma* life cycle**

Generally, the life cycle of African trypanosomes (*T. brucei*, *T. congolense* and *T. vivax*) is “heteroxenic”, which means they complete their life cycle into two hosts; an invertebrate host and a mammalian host. The main African trypanosomes insect vector is the blood feeding males and females of tsetse fly of genus *Glossina*, in which the trypanosomes enter the fly by feeding on blood of an infected animal (Hoare, 1972). In each host, trypanosomes undergo adaptation and developmental stages, during which they proceed to a stationary stage then a proliferative stage to ensure infective population number (Hoare, 1972). Different primary developmental stages occur in the tsetse fly, which terminate in a final form ready to infect a mammalian host (metacyclic form) (Hoare, 1972; Vickerman, 1985; Vickerman *et al.*, 1988; Gardiner, 1989; Matthews, 2005; Fenn and Matthews, 2007).

The main differences in life cycle among these trypanosomes are indeed the insect developmental stages, which differ in all three species of trypanosomes (Hoare, 1972). (Figure 1.1).

### **1.2.1 The life cycle of *T. brucei***

The vertebrate stage starts by the injection of blood infective form (metacyclic trypanosomes) into the blood stream of the mammalian host after a blood meal of an infected tsetse fly. This evokes a local immune reaction in the site of the bite called a “chancre”. The metacyclic form is preadapted to the blood environment by having a single coat of condensed Variant Surface Glycoprotein (VSG) that protects the parasite from the initial host immune response (Turner *et al.*, 1988). After a successful establishment, the metacyclic trypanosomes multiply into long cylindrical blood forms, which in turn undergo different proliferative rounds and each time change its VSG coat by switching from a VSG gene to another to escape the developed immune response through a mechanism called “antigenic variation” (Barry and McCulloch, 2001; Morrison, Marcello and McCulloch, 2009). Consequently, the long cylindrical form transforms into short broad form which might lack the tendency to proliferate and able to infect a tsetse fly when its feeding on an infected animal (MacGregor and Matthews, 2010).

In the tsetse vector the blood meal is firstly stored in the crop then it transfers to the midgut passing the proventriculus, in the midgut the stumpy bloodstream form develops into long procyclic forms within first few hours. These initial forms tend to express stage specific surface protein (procyclins) or procyclic acidic repetitive protein (PARP) which are GPEET and EP-rich proteins named according to the tandem repeats of amino acids on the N-terminus of the surface protein, while the late stage procyclic trypanosomes express just the EP (Roditi and Clayton, 1999). These proteins can resist the proteases and the acidic conditions of the midgut (Liniger *et al.*, 2003). Four days post infection the procyclins cross the peritrophic matrix (PM) and reach to the ectoperitrophic space to advance towards the anterior midgut (Vickerman, 1985). The procyclins penetrate the peritrophic membrane from the site of its origin to reach to the proventriculus at day 6 and start to transfer to mesocyclic

forms (Van Den Abbeele *et al.*, 1999). During their journey to salivary glands, mesocyclics differentiate to epimastigotes (Sharma *et al.*, 2008). In the tsetse salivary glands trypanosomes undergo asymmetrical division into a thin long highly motile form and a short stumpy form; the latter form of trypanosomes tend to attach to the epithelial cells (Van Den Abbeele *et al.*, 1999; Sharma *et al.*, 2008).

The epimastigotes undergo two intermediate stages in order to develop unattached form the metacyclic form shielded with VSG layer and prepared to withstand the new environment in the blood stream (Van Den Abbeele *et al.*, 1999).

### **1.2.2 The life cycle of *T. congolense* with respect to that of *T. brucei***

The blood stage of *T. congolense* appears smaller, shorter and without a prominent undulating membrane in comparison to *T. brucei*, which in turn has two forms, the non-replicating form (stumpy form) and a long slender form. The latter form cannot withstand the proteases of the fly's midgut after the blood meal, while the former can survive and establish successful primary colonization in the fly's midgut (Sbicego *et al.*, 1999). Moreover, *T. congolense* tends to attach strongly to the endothelium of blood vessels, while *T. brucei* could migrate to different tissues which might make the differences in the pathogenicity of the two parasites (Coustou *et al.*, 2010).

After two days in the insect midgut, both trypanosome species undergo developmental changes to differentiate into procyclic forms. During this differentiation, the parasite loses its VSG coat, in comparison to *T. brucei* VSGs, those of *T. congolense* are richer in carbohydrate (Bütikofer *et al.*, 2002; Utz *et al.*, 2006). Both species migrate to reach the mouthparts passing the proventriculus within six days after infection (Peacock *et al.*, 2012); this location serves as a crossroads between the two species in which *T. brucei* initiates an asymmetric division and ends up with two morphologically different progenies. One of them is short and the other is long epimastigote, however such dividing process has not been revealed in *T. congolense*. Whilst the shorter epimastigote of *T. brucei* in the proventriculus is important as they

migrate to the salivary glands to proliferate and localized there (Van Den Abbeele *et al.*, 1999; Sharma *et al.*, 2008), in *T. congolense* the transformation occurs in proboscis and when these trypanomastigotes are attached to the labrum of the proboscis at day 13 after infection (Peacock *et al.*, 2012) they develop to epimastigotes (Hoare, 1972). Furthermore, Jefferies *et al.* (1987) showed that the trypanosomes are also found in the cibarium in both *T. congolense* and *T. vivax*.

Finally, in both *T. congolense* and *T. brucei*, as soon as the epimastigotes attached to the tissue of final localization they replicate and differentiate into mammalian infective form (metacyclics), which are characterized by their own VSG coat to cope with the new environment (Tetley and Vickerman, 1985). In *T. congolense* these epimastigotes develop in the hypopharynx and labrum while for *T. brucei* they develop in salivary glands (Thévenaz and Hecker, 1980). According to this difference between these species a sub classification has been introduced: a subgenus Trypanozoon for those trypanosomes who are replicating in midgut and salivary glands, which involves *T. brucei* and Nannomonas for those who replicate in the midgut and proboscis (i.e. *T. congolense* and *T. vivax*) (Hoare, 1972).

### **1.2.3 The life cycle of *T. vivax***

In the blood stream, *T. vivax* proliferate to generate more elongated cells that are more likely to attach the blood vessels or invade certain tissues like lymph nodes, spleen, myocardiac muscles and central nervous system, occasionally, this might cause acute haemorrhage in the mammalian host (Magona, Walubengo and Odimin, 2008). The tsetse pre-adapted form seems more elongated than the short stumpy form seen in *T. brucei* and it more likely showed affinity to adhere to cellular culture through their flagellum (Gardiner and Wilson, 1987). Furthermore, some morphological changes also observed as the kinetoplast of the *T. vivax* is prominent at the more rounded posterior end and the undulating membrane looks temperately developed (Hoare, 1972).

Since *T. vivax* showed two modes of transmission via invertebrate hosts, its life cycle in this host could be divided into cyclical (when developmental stages occur) (Hoare, 1972) and non-cyclical spread (mechanical transmission) (Desquesnes and Dia, 2003).

### **1.2.3.1 Cyclical transmission**

After the tsetse fly fed experimentally on blood meal of an infected host (Ooi *et al.*, 2016), within first two days the parasites in the foregut are more likely to die, while in the proboscis and cibarium the epimastigote forms started to appear after day 3 of infection (no-procyclic forms noticed) and proliferating through day 3-7 through symmetrical and asymmetrical divisions similar to that of *T. brucei brucei* seen in the salivary gland of the tsetse fly (Rotureau *et al.*, 2012). These divisions occur while the parasite still attached to the mouth parts, with the asymmetrical division ends with the generation of pre-metacyclic forms that could be shed predominantly with the saliva of the infected fly. By the attachment of *T. vivax* to the mouth parts, especially cibarium, it maintain its infection and in the same time releases metacyclic forms that could infect the mammalian host (Ooi *et al.*, 2016). The simple life cycle of *T. vivax* and higher infection rate resulted from better adaptation of this trypanosome to the tsetse fly and might represent an earliest member of African trypanosomes (Ooi *et al.*, 2016). From this life cycle, *T. vivax* seems to be mostly confined to the mouth parts especially, cibarium and proboscis. This ability to stick to the mouth parts seen in *T. vivax* might permit it to maintain on other mode of transmission, which does not involve any developmental progression.

### **1.2.3.2 Mechanical (non-cyclical) transmission of *T. vivax***

Although the mechanical transmission has been established experimentally for African trypanosomes for *T. vivax*, *T. brucei* (Mihok *et al.*, 1995) and *T. congolense* (Sumba, Mihok and Oyieke, 1998), the only successful natural AAT infections outside the tsetse fly zone are linked to *T. vivax*, which is the most prevalent of the infections in East Africa. This might emphasis the natural mechanical role of transmission of *T. vivax* evidenced from the high prevalence

rate of *T. vivax* infections in tsetse-free African regions (Cherenet *et al.*, 2004; Mossaad *et al.*, 2017) and the only African trypanosomes that are able to infect animals outside Africa in South and Central America (Cortez *et al.*, 2006; Rodrigues *et al.*, 2008).

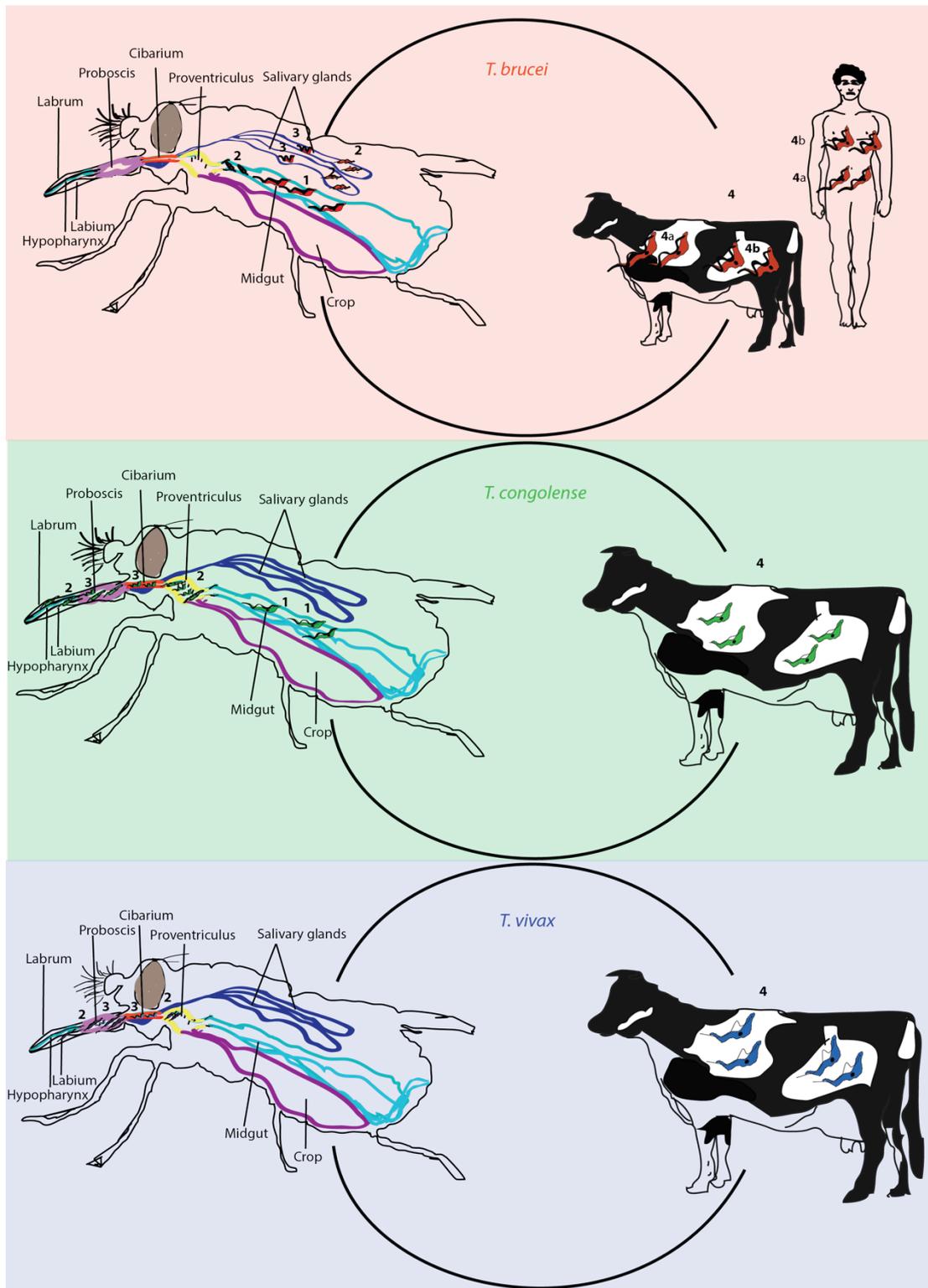
### 1.3 Cell surface proteins and trypanosomes life cycle

The cell surface coating proteins of trypanosomes in blood stage are the VSGs which is the most abundant surface protein. Each VSG coat is replaced by other variant through activation of a new VSG gene in a new telomeric region (Pays *et al.*, 1989; Chaves *et al.*, 1999) and its accompanied with other surface proteins, which transcribed from the same VSG genes containing genomic regions called Blood Form Expression Sites (BES). These genes are mainly being series of Expression Site Association Genes *ESAGs*, which considered as modified VSG sequences. An example of *ESAG* surface proteins are *ESAG 6* and *ESAG 7* –transferrin like receptors that scavenge the ferric ions from the blood stream of the host. *ESAG6/7* genes are present in each BES identified in *T. brucei*, suggesting their importance to the BFs of *T. brucei* (Donelson, 2003; Jackson *et al.*, 2013).

When the parasite enters the insect vector this VSG coat is rapidly replaced with another protecting coat called “procyclins or PARP” which are GPEET and EP- rich proteins named according to the tandem repeats of amino acids on the N- terminus of the surface protein, while the late procyclic stage trypanosomes express just the EP (Roditi and Clayton, 1999). These proteins can resist the proteases and the acidic conditions of the midgut (Liniger *et al.*, 2003). Meanwhile, the VSG is degraded and the main degradation site is the flagellar pocket reviewed in (Field and Carrington, 2009). This degradation process is achieved by a surface protease metalloprotease-B (MSP-B) (LaCount *et al.*, 2003). In *T. brucei* after the midgut stage the parasite migrates to the proventriculus then to the salivary glands of the Tsetse fly, where the procyclins are replaced by *Brucei* Alanine Rich Protein BARP proteins that cover the epimastigote form (Urwyler *et al.*, 2007). Then the parasite undergoes differential replication to generate free vertebrate infecting trypanosomes (Metacyclic form) shielded with a specific VSG coat expressed

from specific *VSG* genes located on particular genomic territories close to the end of the expression sites (MES) (Donelson, 2003; Kolev, Günzl and Tschudi, 2017). In the genomic context, the difference between the BES and MES is the size of the genomic area and the gene contents, as the BES are more likely to be long regions 40-70 Kbp, while the MES length is 3-6 Kbp as reviewed by (Donelson, 2003). The gene content structure of the BES is more complicated and generally involves a combination of 1-12 *ESAG* genes/pseudogene. The subtelomeric *VSG* genes/pseudogenes, *Ingi* elements and a telomeric proximal *VSG* copy these structures are located between a polymerase I promoter and the telomeric repeat units of (TTAGGG) at the end of the chromosome (Pays *et al.*, 1989; Redpath *et al.*, 2000; Donelson, 2003; Hertz-Fowler *et al.*, 2008), while it is more simplified in MES sites to a few *VSG* gene copies with no apparent *ESAG* genes/pseudogenes between the MES promoter and the telomeric repeat (Barry *et al.*, 1998). The metacyclic *VSG* coat could be replaced after the parasite enters the blood stream of a mammalian host with a new blood stage *VSG* gene expressed from one BES from about 20 in *T. b. brucei* estimated in strain Lister 427 (Navarro and Cross, 1996).

The current draft genome sequences of *T. congolense* and *T. vivax* are interrupted with many gaps, so that, the tandem repeat organization of genes that encode for surface coating proteins during different stages of life cycle might be missing. Moreover, it could not show obvious telomeric repeats and clear subtelomeric structures within these chromosomal ends. Therefore, an upgraded genome assembly for these two trypanosomes is essential to uncover such important genomic regions.



**Figure 1.1 African Trypanosomes cyclical life cycle.** Developmental stages in the insect host (tsetse fly): Procyclic form (1), Epimastigote form (2) and Metacyclic form (3) as well as the blood stage forms (4) long cylindrical (4a) and (4b) short stumpy form.

## 1.4 *Trypanosoma* cell morphology

Most of the experimental studies were carried out on *T. brucei*. Hence, most of the description here is regarding this model trypanosome. Generally, the trypanosome cell is an extended hemoflagellate that has a single flagellum, its cytoskeleton is extremely polarized and it is formed by a mesh of microtubules, which in turn could be changeable affecting the size and shape of the cell during different stages of the parasite life cycle (Matthews, 2005). The polarity of the microtubules defines the polarity of the cell as reviewed by (Ooi and Bastin, 2013).

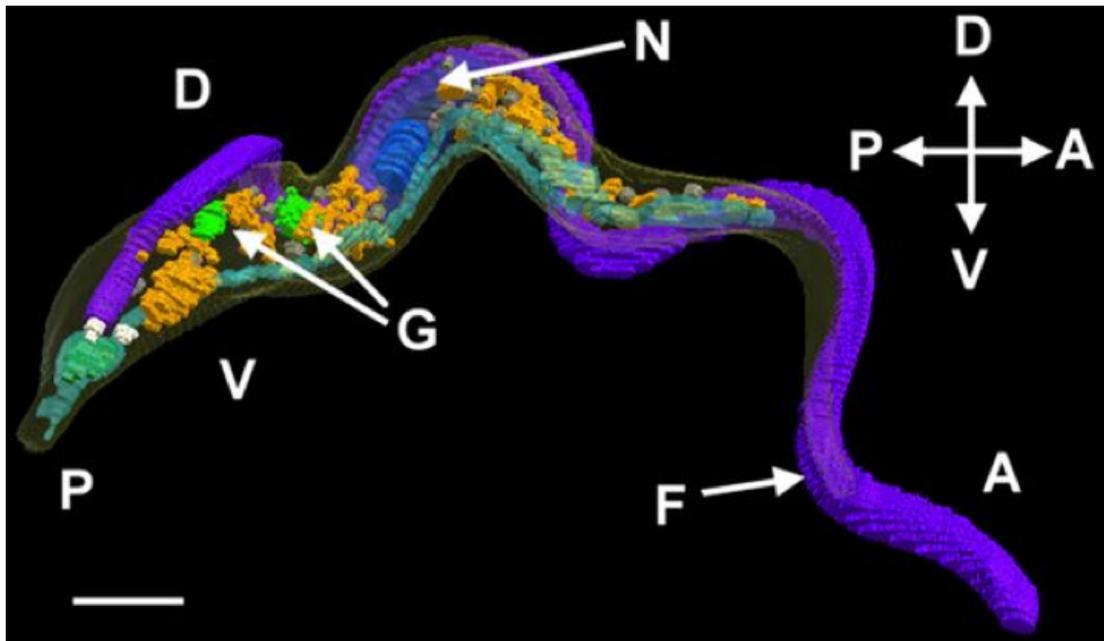
Being elongated, the longitudinal diameter length is range from 14-40  $\mu\text{m}$  (Mogk *et al.*, 2014), while the transverse one about 5  $\mu\text{m}$  (Field *et al.*, 2004). The trypanosome can be easily identified under the light microscope after staining by Giemsa or DAPI staining methods, which could show the main trypanosome cell components (flagellum, the flagellar pocket, the nucleus and a prominent mitochondrion “kinetoplast”) (Field *et al.*, 2004) (**Figure 1.2**).

The flagellum also consists of microtubules; it originates from cytoskeleton region always close to the kinetoplast and preceded by a basal body; the flagellar base consists of doubled sets of nine microtubules surrounding a pair of two central ones and extended with a structure called the paraflagellar rod (PGR), and it emerges from a special cellular compartment called “flagellar pocket” (Vaughan, 2003). Furthermore, the flagellar pocket has an important role of endocytosis and exocytosis of materials (Engstler, 2004) and its known to be the site of VSG protein degradation (Ooi and Bastin, 2013).

The morphology of the trypanosomes is quiet changeable throughout the life cycle in both the insect vector and the mammalian host; moreover, the position of the flagellum origin is changing accordingly and it could be linked to the parasite pathogenicity (Sharma *et al.*, 2009; Rotureau, Subota and Bastin, 2011).

The origin of the flagellum is defined by the position of the kinetoplast In the blood stage form BSF or trypanomastigotes, and the insect midgut forms (procyclic form, Mesocyclic and the free Metacyclic form), the kinetoplast is

located at the posterior end, to the nucleus. By contrast the other insect forms (i.e. proventricular epimastigote and the attached epimastigote forms) have a kinetoplast located at the anterior end (Robinson *et al.*, 1995; Sunter and Gull, 2016). Similar alternate repositioning of kinetoplast-flagellar base is also true for *T. congolense* (Peacock *et al.*, 2012) and *T. vivax* (Ooi *et al.*, 2016). Some morphological difference between the three African trypanosomes were also noticed, as *T. brucei* has two morphologically differed clones (the long cylinder and the short stumpy forms with remarkable long flagellum and undulating membrane), while *T. congolense* showed almost uniform clones with characteristic small flagellum, however, the *T. vivax* blood forms have an intermediate developed flagellum, with slightly bigger fast-moving cells containing a large prominent posterior kinetoplast (Hoare, 1972).



**Figure 1.2 3D ultrastructure of *T. brucei* cell.** Anterior end of the parasite (A), Posterior end (P), Dorsal surface (D), Ventral surface (V), Golgi apparatus (G), Flagellum (F), Nucleus (N), Basal bodies (white colour) and Flagellar pocket (green area on posterior end), the scale bar 1  $\mu\text{m}$ . The figure is adopted from (Hughes *et al.*, 2017).

## **1.5 African Trypanosomes Diseases and impact**

### **1.5.1 Human African trypanosomiasis**

Human African Trypanosomiasis (HAT) or sleeping sickness is endemic in sub-Saharan Africa and transmitted by tsetse fly (*Glossina* species), which is distributed within in a belt in sub-Saharan Africa (Brun *et al.*, 2010). In this region 70 million humans are at risk of infection. During an epidemic in 1998, 83,000 new cases were reported (WHO, 2015). However, the number of annual cases has dropped to 6,314 in 2013, and 3,796 new cases were reported in 2015 (WHO, 2015). It has been considered one of the most neglected tropical diseases (Vanderelst and Speybroeck, 2010).

HAT is caused by two subspecies of *Trypanosoma brucei*, which are geographically separated, these are *T. b. gambiense*, which is the most prevalent in West Africa and responsible for the chronic form of the disease (Giroud *et al.*, 2009; Wastling *et al.*, 2011), while *T. b. rhodesiense* causes an acute form of the illness in East Africa but with a very low rate (Brun *et al.*, 2010).

#### **1.5.1.1 Clinical manifestations of the disease**

The disease can be divided into two stages the haemolytic stage, that be can follow by the meningoencephalitis phase, with fluctuating parasitemia and different organ localization of the trypanosomes. The early stages are characterized by fever, nausea, headache and lethargy; these signs are transient depending on the alternate presence and absence of trypanosomes in the blood stream. The late stage is manifested by localization of trypanosomes in the central nervous system via crossing from the blood stream to the brain through the blood-brain barrier (Kennedy, 2004; Rodgers, 2010). Once it establishes infection in the brain and cerebrospinal fluid (CSF), the host immune response against trypanosomes can have adverse effects on nervous system. Consequently the clinical manifestations appear as anxiety and loss of concentration, which develops into difficulties in speaking and walking and then tendency to sleep (the characteristic sign of the disease) (Kennedy, 2006, 2008).

### 1.5.2 Animal African trypanosomiasis

Animal African Trypanosomiasis (AAT) is also known as “Nagana”, a Zulu word that means “useless or powerless”. Is a disease in domestic animals especially bovine species this is a severe progressive illness and can be fatal; it is characterized by fever, anaemia, weight loss and lack of milk production (Steverding, 2008). Besides the domestic animals, trypanosomes could infect wild animals, however, with low or unrecognized symptoms these animals are potential reservoirs of the parasite (Steverding, 2008).

AAT is caused mainly by *T. congolense*, *T. vivax*, *T. evansi* and *T. brucei* *brucei* and to a lesser extent the *T. brucei* subspecies responsible for HAT; meanwhile, *T. congolense* considered the most virulent one to the domestic animals as it is responsible of severe anaemia (Pinchbeck *et al.*, 2008; Dayo *et al.*, 2010), while variations in disease intensity was also reported among different strains of *T. congolense* (Bengaly *et al.*, 2002; Masumu *et al.*, 2006, 2009; Van Den Bossche *et al.*, 2011). Whilst *T. brucei* causes a mild form of the disease in cattle, the virulence of *T. vivax* fluctuates significantly between west and east Africa, with the western strains being the most virulent and responsible for acute illnesses (Fasogbon, Knowles and Gardiner, 1990).

Remarkably, *T. vivax* is able to cause trypanosomiasis outside the tsetse belt in Africa and further in other continents like South America by potential mechanical transmission through other blood feeding insects such as the stable fly (*Stomoxys*) and horse fly (*Tabanids*) infecting cattle and other domestic animals like camels, water buffaloes, horses, donkeys, dogs, pigs and goats (Fasogbon, Knowles and Gardiner, 1990; Dirie *et al.*, 1993; Osório *et al.*, 2008).

### 1.5.3 Treatment of Trypanosomiasis

Anti-trypanosome drugs are used for the treatment and control of AAT; there are three main effective drugs: isometamidium chloride (ISM), ethidium bromide (EtBr) and diminazene aceturate (DA) which have been used for these purposes for a long time (Kinabo, 1993). DA is used for therapeutic purposes, while ISM is used as prophylactic drug and gives approximately

three months of protection, while EtBr is mainly used for treatment (Sinyangwe *et al.*, 2004). However, due to the mutagenic properties of EtBr, it should be removed from the list, but in fact it is still used in many countries. Almost 35 million doses of trypanocidal drugs are used annually (Geerts *et al.*, 2001). Accordingly, due to the extensive use of these trypanocidal drugs especially during 1960s-1970s, resistance and cross resistance to these drugs were developed and started to be wide spread across Africa (Kinabo, 1993).

Although these trypanosomes mainly cause AAT, interestingly, rare cases of human infection have been reported (Truc *et al.*, 2013).

#### **1.5.4 AAT economic impact**

One-third of the total land in sub-Saharan Africa is infested by *Glossina* species (Mattioli *et al.*, 2004). Within this region, some 46–62 million head of cattle and other livestock species are at risk of trypanosomiasis, which is a great threat on ruminant livestock production (Swallow, 1999).

Animal trypanosomiasis in Africa represent a major limitation on livestock husbandry, while bovine trypanosomiasis in Africa causes 3 million deaths annually. Overall economic losses of fatality and loss of production and cost of treatment of animals was estimated to US\$ 4.5 billion (Mattioli *et al.*, 2004).

The prevalence of the three African trypanosomes in domestic cattle in the Jos Plateau in Africa revealed a mean prevalence of 46.8%, with the lowest infection rate by *T. brucei* 3.2% and the highest by *T. congolense* 27.7% (Majekodunmi *et al.*, 2013).

## **1.6 Trypanosome genome and chromatin architecture**

### **1.6.1 Nucleosomes**

Nuclear DNA in eukaryotes is coiled around special structural proteins called histones to form nucleosomes, this mechanism allows the chromatin to compact the long strands of the DNA within the nucleus, whilst permitting

access for viable processes like DNA replication, DNA repair and transcription (Krude, 1995). There are four highly-conserved so-called “core histones” in eukaryotic cells, H2A, H2B, H3 and H4; each two core histones form an octamer structure that wraps 147 bp of DNA with a spacer DNA ranging 160-250 bp between each two nucleosomes, and each series of nucleosomes is locked by linker histone H1 (Luger *et al.*, 1997). Nucleosome dynamics could arise from alteration in the structure of octamer core histone complex (Zlatanova *et al.*, 2009). H1 histone, which is less conserved than core histones across eukaryotes, is thought to stabilize the chromosome and chromatin condensation by binding to the nucleosomes dyad and spacer DNA (Bednar *et al.*, 1998; Carruthers *et al.*, 1998; Happel and Doenecke, 2009).

Histones that package the DNA threads are called ‘canonical histones’, however, other kinds of histones ‘non-canonical histones’ are involved in different functions such as transcription initiation and termination, chromosome segregation, DNA repair and cell division (Baldi and Becker, 2013). In contrast to genes encoding for canonical histones, which occur in arrays of tandem repeats, non-canonical histone genes tend to occur singly in the genome of eukaryotes (Dalmasso, Sullivan and Angel, 2011).

## **1.6.2 Organization of the chromatin in the eukaryotic nucleus**

During the interphase stage of cell cycle, the chromosomes are positioned in non-random fashion leading to the spatial organization and compartmentalization of the eukaryotic genome. These chromatin features could be found together in different combinations leading to the formation of different topological domains; these domains are distributed in a hierarchical way (de Graaf and van Steensel, 2013).

### **1.6.2.1 Chromosome territories**

Within the nucleus, the chromosomes have their own regions known as “chromosome territories” (Cremer and Cremer, 2010). In yeasts, centromeres of the chromosomes are attached to the spindle pole body, while telomeres are positioned in the periphery and attached to the special lamina fibres of nuclear envelope and the only parts of the chromosome that enter the

nucleolus are rDNA regions. In addition, functionally related gene clusters could be co-located such as tRNA genes and early replication start sites (Tjong *et al.*, 2012).

The active topological associated domains (TADs) of chromatin seem to have hierarchical topology, regions with high number of genes (euchromatin) tend to organize as loosed accessible and transcriptionally active domains while poor genic regions (heterochromatin) are transcriptionally inactive and have AT-rich repeats (Misteli, 2005). In the *Drosophila* genome, an active euchromatin domain is built from multiple genes within a chromosomal region of (10-50 kb). demarcated by insulator binding borders, active histone variant H3K4me3 and/or DNase hypersensitivity. Euchromatin domains showed interchromosomal contact but not with heterochromatin domains (Sexton *et al.*, 2012).

Interchromosomal interactions in *S. cerevisiae* of small chromosomes seems to be at, a higher rate than the intrachromosomal ones; in contrast large chromosomes showed inverse figures (Daniels, Gull and Wickstead, 2010).

### 1.6.3 Trypanosome genomes

The genomes of Trypanosomes are diploid and the parasite *T. brucei* has a haploid genome size of about 35 Mb, consisting of 11 megabase-size chromosomes (MBCs) hosting the housekeeping genes of the organism packed inside a nucleus of ~ 2.5 µm in diameter ( Ersfeld *et al.* 1999;Ogbadoyi *et al.* 2000). However, the genome of *T. cruzi* has 36 pairs of chromosomes and along with *Leishmania major*, which has ~ 28 chromosomes are more likely to reflect ancestral karyotype (El-Sayed *et al.*, 2005). In addition to MBCs, *T. brucei* genome has ~ 100 minichromosomes (MCs), and their size ranged from 30 to 100 kb and ~ 5 intermediate chromosomes (ICs) (200- 900 kb) in size (Hertz-Fowler, 2007). The ICs and MCs are thought to be aneuploid and inherited in a non-Mendelian way (Wells *et al.*, 1987).

The MBCs showed two separate types of protein coding genes; the internal diploid chromosomal regions contain housekeeping genes arranged in DGCs, which are interrupted by non-coding DNA sequences called Strand Switch

regions (SSRs), DGCs polycistronic transcribed by polymerase II (Pol II) and distal regions on both ends of the chromosome (sub telomeric region) considered as a mono-allelic region containing genes encoding mainly for surface proteins (most likely VSGs) and transcribed by polymerase I (Pol I) (Berriman *et al.*, 2002; El-Sayed *et al.*, 2005).

### **1.6.3.1 Transcription of housekeeping genes**

Most protein coding genes of *T. brucei* are arranged in tandem repeat clusters and the vast majority do not have introns and are ordered in long clusters called polycistronic transcription units (PTU) or DGCs (Imboden *et al.*, 1987). Two genes were reported to have introns in *T. brucei*, they are ATP-dependent DEAD/H-box RNA helicase and poly A polymerase (Mair *et al.*, 2000; Berriman, 2005).

The house keeping genes in the core regions of *T. brucei* genome are transcribed from transcriptional start sites and terminate with transcription termination sites in non-coding chromosomal regions on the boundaries of a PTU or DGC (Siegel *et al.*, 2009).

### **1.6.3.2 Transcription of genes encoding for surface proteins**

A unique feature of trypanosome transcription machinery is the employment of Pol I in the transcription of protein coding genes responsible for surface trypanosome protein coat. In eukaryotic cells, polymerase I is devoted to transcribing rRNA genes, while *T. brucei* involved this polymerase in the transcription of blood-stage VSG genes and the procyclins (Horn, 2001).

The procyclic stage genes are arranged in clusters and they are polycistronically transcribed; in *T. brucei* these genes are organized internally on MBC six and ten along with clusters of other protein-coding genes (Daniels, Gull and Wickstead, 2010). However, they are grouped in small pools of 3-6 genes (Liniger *et al.*, 2001). The post transcriptional maturation is crucial for the stage specific transcript maturation, the post-transcription added sequence at the 3' end of procyclins transcript is responsible for the stability of these mRNA during procyclic stage but not in the blood stage (Hotz *et al.*, 1997).

Meanwhile, the VSG genes are localized on the BES and MES, which are located at the ends of the MBCs and ICs and transcribed by polymerase I (Pays *et al.*, 2001).

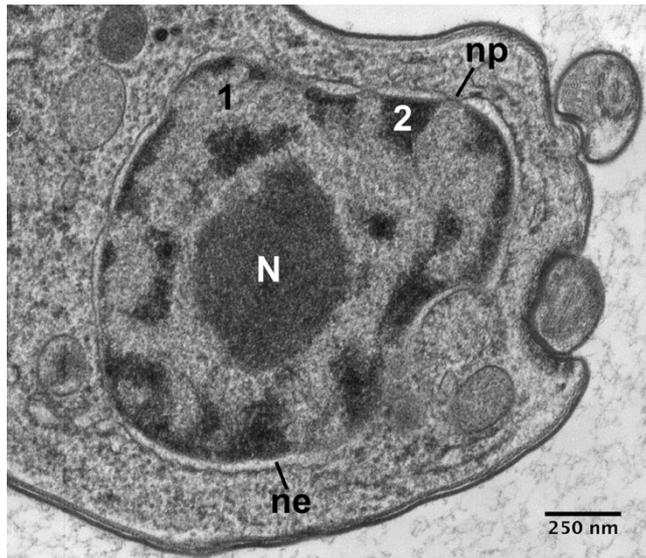
Noteworthy, the size and boundaries of the PTUs yet to be determined into the other two African trypanosomes genomes (*i.e.* *T. vivax* and *T. congolense*).

### **1.6.3.3 Chromatin organization in the Trypanosome nucleus**

Chromatin organization during interphase of *T. brucei* and *T. cruzi* nuclei could be visualized using electron microscopy (Daniels, Gull and Wickstead, 2010). Heterochromatin reflects electron dense areas located peripherally to the nucleus, while euchromatin regions showed electron lucent regions and its position is to associate with the nuclear pores, which in turn are distributed almost consistently along the nuclear envelope (Figure 1.3).

In comparison to mammals, trypanosomes have a high gene density and these genes are arranged in DGCs so that the rate of transcription of MBCs seems to be similar except the silent VSG gene spectrum (Daniels, Gull and Wickstead, 2010).

The chromosomes of trypanosome do not segregate in well-defined detectable chromosomes during mitotic cell division (Wickstead and Gull, 2007). Moreover, trypanosomes chromatin seems to be more diffused packed in the nucleus than the mammalian chromatin (Belli, 2000). However, during different stages of *T. brucei* life cycle, the chromatin shows some changes, especially in the insect stages and mammalian blood-form stage (Belli, 2000).



**Figure 1.3 Transmission electron microscopy shows nuclear ultra-structure and compartmentalization of the *T. brucei* nucleus.** The dense structure (dark colour) in the centre is the nucleolus (N), euchromatin (1) appears as electron-lucent (bright) while heterochromatin looks as electron-dense areas (2) (2, “heterochromatin”) regions in the nucleoplasm. The structures of the nucleus are encased by the nuclear envelope; nuclear pores (np) in the envelope are always coexisting with euchromatin. The *Trypanosomes* sample preparation for transmission electron microscope was done by freezing under high-pressure, then freeze substitution with 1% glutaraldehyde and 2% uranyl acetate in acetone and followed by 2% osmium tetroxide and 2% uranyl acetate in acetone then after immersing in resin. The image courtesy of Catarina Gadelha (University of Cambridge, United Kingdom), (Daniels, Gull and Wickstead, 2010).

As in higher eukaryotes nucleosomes of *T. brucei* and *T. cruzi* also have core histones of H2a, H2B, H3 and H4 or their variants. However, their N-terminal sequences are different (Hecker *et al.*, 1994). Furthermore, the H1 linker histone of these parasites reserves only the C-terminal side of the human H1 (Galanti *et al.*, 1998). It has been noticed that histone variants of H2AZ and H2BV are associated with repetitive DNA like telomeric repeats of MBCs and MCs (Lowell *et al.*, 2005).

Other core histone variants (H3V and H4V) seem to be more associated with possible transcriptional initiation and termination sites in *T. brucei* SSRs

(Siegel *et al.*, 2009) . H3V is also found in subtelomeric and telomeric repeats of MBCs and MCs (Lowell and Cross, 2004; Siegel *et al.*, 2009).

#### **1.6.3.4 Trypanosome chromosomal regions**

Like another eukaryotes *T. brucei* 45S *rRNA* genes loci tend to locate within the nucleolus, while telomeric regions of MBCs and MCs have different localization throughout the life cycle (Chung *et al.*, 1990; Ersfeld and Gull, 1997). During procyclic stage, these regions clustered to the periphery of the interphase nucleus, whilst they distribute within the nucleus in bloodstream-form cells (Chung *et al.*, 1990; Pérez-Morga *et al.*, 2001).

In contrast to higher eukaryotes, trypanosomes don't have obvious chromosomal loops and domains within the context of chromosome territories, because of the high number of genes in MBCs and polycistronic transcription and post transcriptional regulation (Daniels, Gull and Wickstead, 2010).

The distal chromosomal regions (i.e. telomeres and the subtelomeres) (except the active metacyclic expression sites in metacyclic stage cells and the active BES blood form expression sites in the mammalian blood form cells) of MBCs are functionally potential heterochromatin. Moreover, MCs in *T. brucei* are transcriptionally silent (large part of these chromosomes consist of palindromic 177-bp repeat blocks) (Wickstead, Ersfeld and Gull, 2004). Therefore, it is more likely that the electron-dense chromatin located marginally to the *T. brucei* interphase nucleus is associated with minichromosomes (Ogbadoyi *et al.*, 2000). Whilst MCs and telomeric repeats might be packed together to form heterochromatin, the distal subtelomeres of VSG expression sites might denote examples of optional heterochromatin in trypanosomes, as their transcriptional eminence changed through the life cycle of the parasite via *in situ* switching (Daniels, Gull and Wickstead, 2010).

### 1.6.3.5 Nuclear dynamics of interphase Trypanosome

Interphase eukaryotic cells show a highly dynamic nuclear pattern (Misteli, 2001). In a study on *T. cruzi* interphase chromosomes dynamics (Elias *et al.*, 2002), found potential movement of DNA repeats from an arbitrary allocation in G1 phase to the marginal positioning at the early stage of S phase was found, perhaps because of links with replication repositories at the nucleus borders.

Other evidence of genetic loci movements comes from *in situ* exchanging of Blood-form Expression Sites (BESs) its only possible if two expression sites are come together in order to ignite switching (Chaves *et al.*, 1999). In addition, such proximity of BESs is also important for homologous recombination of the VSG genes in silent pool or in other sites with a gene on active BES site (Hertz-Fowler *et al.*, 2008). It has been postulated that the Trypanosomatids conserved dynein light chain LC8 might enables the locating of the BES region (Brandenburg *et al.*, 2007).

### 1.6.3.6 Meiosis in trypanosomes and its role in genome evolution

Sexual reproduction has been identified in different parasitic protozoa including *T. brucei* as reviewed by (Weedall and Hall, 2015). In *T. brucei*, sexual stages seem to occur in the tsetse fly (Jenni *et al.*, 1986) and experimental evidences suggested that the salivary gland (site of metacyclic forms) is the only place where mating is happening (Jenni *et al.*, 1986; Sharma *et al.*, 2008) through forming of gametes (small pear-like cells) from epimastigote cells (Peacock *et al.*, 2014).

Effect of mating on the genome contents were noticed by formation of polyploidy populations showed haploid, diploid, triploid and tetraploid trypanosomes (Gibson *et al.*, 2008). Moreover, an approximate of double fold increase in the amount of minichromosomes was estimated in the progenies in comparison to the parental cells (Wells *et al.*, 1987). More importantly, trials on generating a hybrid between human infective strain *T. b. gambiense* and animal infective strain *T. b. brucei* resulted in generating a new hybrid that carry the human pathogenic gene (SRA gene) (Gibson *et al.*, 2015).

Interestingly, an evidence of seminal event was found in the genome of *T. brucei*, as a large genomic duplication affected MBCs four and eight in the favour of increase diversity and gene dosage of genes involve in host immune evasion mechanism (Jackson, 2007).

#### **1.6.3.7 Role of genomic structural variation in biological systems**

Structural variation (SV) can be defined as a change in DNA sequence of length more than 500 bp (Freeman *et al.*, 2006). Such change can be translocation, duplication, inversion or deletion and it also can be change In copy number variation (CNV), were a change in a DNA segment occur in the genome increase or decrease the number of regions with high similarity > 90%, these changes are relative to a reference genome (Sharp *et al.*, 2005).

DNA structural variations (SVs) has been linked to many human diseases (Weischenfeldt *et al.*, 2013), speciation (Noor *et al.*, 2001) and adaptation (Iskrow, Gokcumen and Lee, 2012). Environmental stresses are elevating the genetic diversity in a number of traits in a population (Bubliy and Loeschcke, 2002).

The role of SVs in adaptation has been studied in different biological systems. For example, the allelic reduction in glucose-6-phosphate dehydrogenase in malaria endemic regions was correlated to relative resistance to *Plasmodium* infection as reviewed by (Ruwende and Hill, 1998). The most important SVs in speciation are the genomic inversions (Feder, Nosil and Flaxman, 2014).

Chromosomal inversions were also linked to adaptation in insects *Anopheles* species reviewed by (Ayala *et al.*, 2014). For example, the adaptation of the *Anopheles gambiae* in different environments with link to chromosomal inversions in 86% in the geographical regions; xeric and nonxeric ;wet and dry; high and low temperature (Bayoh, Thomas and Lindsay, 2001). While, the role of chromosomal inversions in resistance to the insecticides that was used to eradicate Mosquitoes was indicated in three different species of *Anopheles* (*An. gambiae*, *An. arabiensis* and *An. stephensi*) in Ethiopia.

It has been mentioned that the environmental stresses on various organisms might drive increase the activity of the Transposable Elements (TE) in the genome, which could lead to gene mobilisation or genic activation/silencing (by affecting gene promoters), which leads to CNVs, reviewed by (Capy *et al.*, 2000).

Other structural variation in the DNA is the Single Nucleotide Polymorphism (SNP), this change affecting one nucleotide and it is linked to environmental adaptation, which has been studied in different biological systems; for example, in human, SNPs affecting genes correlated to the metabolism of certain food consumption and SNPs related to certain environmental condition have been found (Hancock *et al.*, 2010).

Genomic structural changes were also studied in protozoa: large genomic duplication in MBC four and eight in *T. brucei* genome has led to increase dosage of genes (mainly functional in host-parasite interface) by creating new paralogs, removing others and increase diversity in gene sequences of paralogs between the affected chromosomes (Jackson, 2007). Population structural study on SNPs profile affected isolates of *Plasmodium falciparum* from five different locations in Africa, was noticed in genes responsible of drug resistance and a gene involves in erythrocyte invasion (Duffy *et al.*, 2017). Similar study conducted on *Leishmania donovani* showed distinct parasite population structure, showed the Indian strains are more closely to African strains than the European ones (Downing *et al.*, 2012).

The structural genomic studies on parasites did not approach large genomic SVs. In this project, SVs across African trypanosomes was analysed and showed a cross species variations.

## 1.7 Chromosomal rearrangements in eukaryotes

Eukaryotic genomes are subjected to different evolutionary forces that reshape the chromosomes by deletion, acquisition, modification and or reordering of genetic materials. Outlining these changes is important to realize evolutionary biology (adaptation, survival and species origin) (Eichler and Sankoff, 2003).

### 1.7.1 Forms of chromosomal rearrangements

The chromosomal rearrangements are a collection of events that reorganize the physical collinearity within chromosomes (intra-chromosomal rearrangements) or among chromosomes (inter-chromosomal rearrangements) and it could be on small scale (a few Kilo base pairs) or on large scale affecting entire chromosomes in respect to a reference sequence or another genome of a closely related organism (Harewood and Fraser, 2014).

Translocations that do not alter the number of genes include balanced chromosomal rearrangements such as reciprocal translocations, inversions, fissions and fusions. Inversions arise from reorientation of internal chromosomal segment in regard to flanking regions (intra-chromosomal); these chromosomal inversions have been linked to environmental adaptation and geographic separation in different organisms (Noor *et al.*, 2001; Ayala *et al.*, 2011; Berg *et al.*, 2017). The reciprocal translocations involve exchange of genomic segments between two different chromosomes (inter-chromosomal rearrangements); sometimes the exchange take place from one chromosome (as a donor), while the other chromosome serves as a recipient chromosome generating unbalanced exchange (i.e. non-reciprocal translocation) (Haber and Leung, 1996).

Interestingly, *T. brucei* sometimes uses the reciprocal translocation to maintain the antigenic variation of BFs to evade the host immune factors. This mainly affects the telomere proximal VSG genes by which the old exhausted VSG gene will be replaced by a new VSG gene located on the end of a different MBC, IC or MC in BES or MES. It is also responsible for formation of a mosaic VSG

coat by reciprocal recombination of two regions on two different chromosomes containing VSG pseudogenes or one with a partial VSG gene and the other with VSG pseudogene. These mosaic VSG genes more likely to be expressed during the late infections of the mammalian host (Pays *et al.*, 1985; Morrison, Marcello and McCulloch, 2009).

### **1.7.2 Genomic DNA sequences related to chromosomal rearrangements**

Chromosomal rearrangements have been linked to specific regions on chromosomes; most likely; repetitive DNA sequences and these could be coding or non-coding regions characterized by high sequence similarity (Clancy and Shaw, 2008; Tsai and Lieber, 2010). Repetitive DNA sequences could represent a major feature of the genome of an organism; for example 50% of the human genome is occupied by repeat DNA sequences (Richard, Kerrest and Dujon, 2008).

The DNA sequences that are more likely to be involved in chromosomal structural variations are Transposable Elements (TEs), Segmental Duplications (SDs), Tandem Repeats (TRs) (Clancy and Shaw, 2008; Weckselblatt and Rudd, 2015), *tRNA*, *rRNA*, tandem repeat arrays of paralogous protein coding genes (Richard, Kerrest and Dujon, 2008) and telomeric repeats (Peitl *et al.*, 2002; Krutilina *et al.*, 2003).

#### **1.7.2.1 Transposable elements**

Historically, TEs are wide spread through a wide range of organisms in the prokaryotes and eukaryotes genomes (Capy, 1998). They were first discovered in the maize genome by Barbara McClintock in 1984 (McClintock, 1984). These elements can be classified into two major classes: retrotransposons and DNA transposons (Richard, Kerrest and Dujon, 2008).

Retrotransposons have a specific pattern repeats and lack introns; they are most likely to be positioned adjacent to each other and they can move themselves throughout different genomic regions by copy and paste using RNA mediated transposition that could be reversed by a reverse-transcriptase enzyme to be pasted again in a new location in the genome (Wicker *et al.*,

2007). These TEs could be also subdivided into Long Interspersed Nuclear Elements (LINEs), non- Long Terminal Retroelements (non-LTR) like *ingi*/RIME, Short Interspersed Nuclear Elements (SINEs), and Dictyostelium Intermediate Repeat Sequences (DIRS-like elements) (Wicker *et al.*, 2007).

Non-LTR elements like have extremely diverse sequences and they can insert themselves in many locations in the DNA, more specifically, for example, *rRNA* genes telomeric repeats and in the subtelomeric regions (reviewed by (Craig, 1997)). In *T. brucei* genome, *ingi* (long autonomous retroelements) and RIME (nonautonomous retroelements) are the most abundant mobile genetic elements (Aksoy, 1991). Their localisation in the *T. brucei* genome is not random, the *ingi* element is mostly flanked by two halves of RIME (A and B) on upstream and downstream, respectively and they are more abundant in subtelomeric regions (Bringaud *et al.*, 2004).

Meanwhile, the DNA transposons that move around the genome by a cut and paste mechanism (nonRNA-mediated process), through single strand or double strand break points and a faulty DNA damage repair more likely between two homologous chromatids or two similar regions of two different chromosomes (Wicker *et al.*, 2007).

Due to their distribution throughout the genome and their high sequence similarity, TEs are likely to cause genome instability, which in turn leads to cellular mutations and might evoke cellular defects leading to cancer in humans (Chénais, 2013; Burns, 2017); or it they could serve as an evolutionary driver, causing species adaptation to certain environmental niches or stresses (Capy *et al.*, 2000; Casacuberta and González, 2013).

The genome sequence of *T. brucei* showed presence of TEs in all MBCs they occupy 5% of the genome (Berriman, 2005). LINE, *ingi* and RIME elements are the most abundant, while the LINE are represented mostly as non-long terminal retrotransposons (non-LTR) in trypanosome genome and are able to insert themselves in particular genome sequences such as telomeric or subtelomeric repeats, *rRNA* (Craig, 1997). *Ingi* elements are responsible for

causing indels and rearrangements in *T. brucei* genome (Berriman, 2005; Jackson *et al.*, 2010).

### **1.7.2.2 Segmental duplication**

Segmental duplication (SDs) refers to the duplication of certain genomic regions longer than 1 kbp that exist many times in the genome, sometimes called Low Copy Repeats (LCRs). This special kind of repeat could be found in internal regions of the chromosomes, but in higher eukaryotes they are more likely to exist near the centromeres and sub telomeric regions (Bailey and Eichler, 2006; She *et al.*, 2008).

Their involvement in possible genomic rearrangements might be because of their high sequence similarity (>90%) among different copies distributed in the genome, which gives high probability to induce Non-Allelic Homologous Recombination (NAHR) (Bailey and Eichler, 2006; Mok *et al.*, 2008).

Although they are not very frequent, SDs have more importance towards speciation and environmental adaptation as the gene birth (gene expansion) or gene lose (gene contraction) are more likely to occur within these regions, consequently in favour of certain function or vice versa (Newman *et al.*, 2005; Wilson *et al.*, 2006; Mok *et al.*, 2008; Sun *et al.*, 2010). These regions are more likely to contain genes that are involved in the adaptation of an organism to cope with certain environmental stress (Duda and Palumbi, 1999; Mok *et al.*, 2008; Chang and Duda, 2012).

SDs are linked to chromosomal evolution (Elsik, Tellam and Worley, 2009) and have been correlated with evolutionary chromosomal rearrangements (Murphy *et al.*, 2005).

In parasites, SDs is also linked to the emergence of clusters of species-specific genes that favour certain species of parasites to cope with new hosts or habitats (Sun *et al.*, 2010; DeBarry and Kissinger, 2014). Large segmental duplication events affecting two chromosomes (MBC four and eight in *T. brucei*) resulted in generation of species-specific paralogues in favour of surface proteins (Jackson, 2007), genome duplication was also responsible for

genome evolution and host adaptation on subspecies level of *T. brucei* (Jackson *et al.*, 2010).

### **1.7.2.3 Tandem repeat arrays**

Tandem Repeats (TRs) or sometimes called “simple repeats” are clusters of repeated DNA units that occupy a large section of the chromosome and they are arranged in head-to-tail pattern (direct repeats) or head-to-head (inverted repeats) (Richard, Kerrest and Dujon, 2008).

The ability of TRs to cause genomic instability and genome evolution have been also shown (Armour, 2006). Their likelihood to form DNA secondary structures like crucified strands, hairpin and triplex organisations can induce genomic rearrangements (Bacolla *et al.*, 2008). These arrays of repeats could be subjected to either an array expansion or contraction according to the event (Usdin and Grabczyk, 2000). Moreover, TRs are more likely to have AT-rich motifs and these motifs facilitate double strand breakage leading to potential large chromosomal rearrangements (Richard, Kerrest and Dujon, 2008; Kvikstad and Makova, 2010).

TRs have been also recorded in trypanosomes showed species and strain specificity (Wickstead, Ersfeld and Gull, 2004; Zafra *et al.*, 2011).

### **1.7.2.4 *tRNA* genes, *rRNA* genes and tandem arrays of paralogous genes**

These genes are more likely to cluster in large tandemly repeated genes, which work just like the TRs or SDs in both cases with high sequence similarity (Richard, Kerrest and Dujon, 2008).

*tRNA* genes are crucial for the living cells as they facilitate translation of the cellular proteins; clusters of *tRNA* genes were observed dispersed throughout genomes of prokaryotes and eukaryotes (DeLotto and Schedl, 1984; InoKuchi, 1986; Kuhn, Clarke and Carbon, 1991; Lander *et al.*, 2001; Bermudez-Santana *et al.*, 2010). In *T. brucei* *tRNA* genes are distributed through MBCs, most likely in SSRs and usually in clusters of 2-3 genes (Daniels, Gull and Wickstead, 2010).

Similarly, *rRNA* genes encoding for the ribosome components in eukaryotic genomes are more likely to organize in large arrays of tandemly repeated sequences; the size of a repeated *rRNA* cluster is more likely to be correlated to the genome size (Prokopowich, Gregory and Crease, 2003). The *rRNA* genes are highly conserved, thus sequence polymorphism is minimized and make them resemble the DNA direct repeats (Ganley and Kobayashi, 2007). As in the other eukaryotic organisms, *rRNA* genes in kinetoplastids are also arranged in clusters of tandemly repeated sequences (Ivens *et al.*, 2005; Daniels, Gull and Wickstead, 2010).

In higher eukaryotes, some protein-coding gene families consist of a number of paralogous genes that are most likely to arrange in clusters in the genome; such clusters are variable in their length (gene numbers) among different species and strains (Waterston *et al.*, 2002; Rooney and Ward, 2005). Sequenced trypanosomatid genomes showed that a number of paralogous gene clusters are present in the genome and these clusters are more likely to be generated from segmental duplication of the trypanosome genome (Thomashow *et al.*, 1983; Daniels, Gull and Wickstead, 2010).

Presence of such repeated genomic sequences among trypanosomatid genomes might evoke possible inter-species chromosomal rearrangements and the best way to look at such events could be through the Whole Genome Sequence Comparison (WGC) (Spencer-Smith *et al.*, 2012). While, *T. brucei* genome assembly is available at standard quality (Berriman, 2005), the current draft versions of *T. congolense* and *T. vivax* are highly fragmented. So, the generation of new more contiguated *de novo* genome assemblies of these trypanosomes are required to describe such important possible genomic events across species of African trypanosomes.

#### **1.7.2.5 Telomeric repeats**

The chromosomes of eukaryotes are protected by specific DNA repeats that cap the free ends of chromosomes from both sides preventing them from being targeted by DNases, recombination with eroded DNA elsewhere in the genome (Li *et al.*, 1998) or establishment of chromosomal fusion (Zakian, 1997) and

preserve linear appearance of the chromosomes as well as nuclear localization and separation (Croft *et al.*, 1999). The telomeric caps in vertebrates consist of a tandem repeat of TTAGGG sequence units of varying lengths (Meyne, Ratliff and Moyzis, 1989). The length of telomeric repeats is governed by cellular replication rate, recurrent cell division is the main cause of telomere shortening; however, the length of this repeat region is maintained by a specific enzyme called telomerase that adds more repeat units and preserves the telomere's length (O'Sullivan and Karlseder, 2010). Disruption in the activity of telomerase due to aging or other factors leads to shortening of telomeres, which leads to chromosomal instability and cancer in mammals (Raynaud *et al.*, 2008; Gonçalves dos Santos Silva *et al.*, 2010).

Interestingly, chromosomal ends of *T. brucei* are also capped with a similar telomeric sequence as to vertebrates. The trypanosome genome has retained telomeric ends and proximal subtelomeric regions for a particular purpose, which is to generate antigenic variation throughout the life stages of the parasite more widely during the blood stage. The two genomic regions (*i.e.* subtelomeres and telomeres) are utilized as ES to regulate the expression of VSG protein coat (Donelson, 2003; Hertz-Fowler *et al.*, 2008). These proximal ends of trypanosome chromosomes are primed with sets of VSG, ESAG genes/pseudogenes as well as specific non-coding DNA repeats and a characteristic proximal VSG gene copy adjacent to the telomeric repeats (Donelson, 2003; Hertz-Fowler *et al.*, 2008). Furthermore, the *T. brucei* genome devotes a large set of mini and intermediate chromosomes to increase the repertoire of VSG/Telomere contents in its genome (Wickstead, Ersfeld and Gull, 2004). Switching between the active and silent VSG gene by recombination events could be the main mechanism to maintain renewing of surface antigen (Hertz-Fowler *et al.*, 2008). The length of telomeres at the end of ESs seems crucial in the switching frequency and mechanism, as the shortening of this sequence induces more gene conversion events of the telomeric proximal VSG genes and a surge in double strand breaks at the subtelomeric regions (Hovel-Miner *et al.*, 2012).

### 1.7.3 Impact of chromosomal rearrangements on gene expression

Chromosomal rearrangements that lead to transposition of genomic sequences could affect the expression level of these genes. If the translocation does not affect the physical collinearity of a coding sequence, it still might affect the level of gene expression if such an event translocates the gene or its promoter from its normal position. For example, an inversion in human genome placed a gene encoding for aromatase (*CYP19* gene) downstream to an active promoter, which consequently caused abnormal increase in the expression of this gene leads to a disorder characterized by increase, blood level of oestrogen hormone and a related phenotype (Shozu *et al.*, 2003; Demura *et al.*, 2007). While chromosomal dislocations in the human genome might evoke phenotypic disorders, some unicellular parasites employed similar mechanism to avoid immune response of the mammalian hosts. In *T. brucei* during blood stage, antigenic variation of VSG coat that is in direct contact with immune factors in blood stream of the mammalian host is replaced from time to time in order to avoid the destruction by host immune response. One gene should be activated once at a time to replace the exhausted one, and this occurs from a repertoire of more than 1000 VSG genes (Berriman, 2005; Marcello and Barry, 2007a). Such translocation is happening in BES when reciprocal translocation between two subtelomeric regions is mediated by telomeric repeats and the 70 bp repeat, in which the silent new VSG gene switches the previously active one in an active BES, permitting the expression of new VSG coat to maintain persistent infection (Pays *et al.*, 1985; Pays, Vanhamme and Pérez-Morga, 2004; Hertz-Fowler *et al.*, 2008; Stockdale *et al.*, 2008; Morrison, Marcello and McCulloch, 2009).

### 1.7.4 Previous comparative genomic analyses of African trypanosomes

Previous sequencing efforts have been made on different trypanosomes in Africa (Table 1.1). However, structural comparative genomics are scarce.

Comparative genomic analyses before this study on the most important African trypanosomes (*T. brucei*, *T. congolense* and *T. vivax*) were focused mainly on variant surface glycoproteins VSG genes (Jackson *et al.*, 2012, 2013) and

other surface protein coding genes like *PAGs*, *ESAG6/7* transferrin like receptor, other *ESAGs*, *BARP* and *GARP* (Jackson *et al.*, 2013). The previously generated draft assemblies of *T. congolense* and *T. vivax* revealed the presence of 11 MBCs in synteny with those of *T. brucei* (Jackson *et al.*, 2012). However, these genome assemblies showed high degree of fragmentation with noticeable large BIN especially, in *T. vivax* genome assembly pursue the need to generate more complete genome assemblies for these two trypanosomes.

Since extensive analysis has been conducted on VSG genes and surface protein repertoire comparison between African trypanosomes (Jackson *et al.*, 2012, 2013), the focus should be devoted towards some missing yet important features of *T. congolense* and *T. vivax* genomes.

### Table 1.1 Available genome assemblies of African trypanosomes.

Genome statistics are made from latest genome version on TriTryp database release version 37

[http://tritrypdb.org/common/downloads/Current\\_Release/](http://tritrypdb.org/common/downloads/Current_Release/)

African Trypanosomes	Scaffolds	assembly size bp	N50 size*	GC %	% ambiguous bases	Number of protein coding genes
<i>T. brucei</i> Lister427	32	26,754,408	2,482,252	44.88	3.93	9,313
<i>T. brucei</i> TREU927	131	35,826,294	3,542,885	45.47	0.03	11,703
<i>T. brucei gambiense</i> DAL972	11	22,148,088	2,224,448	47.09	0.17	8,082
<i>T. congolense</i> IL3000	2,839	41,372,041	1,222,280	39.68	17.62	11,792
<i>T. vivax</i> Y486	8,290	47,506,340	21,884	45.72	12.06	12,050
<i>T. rangeli</i> SC58	7,433	14,019,393	2,203	53.25	0.02	7,475
<i>T. grayi</i> SNR4	2,871	20,934,132	16,775	53.59	0.68	10,676
<i>T. theileri</i>	253	29,602,501	517,122	34.62	13.64	11,312
<i>T. evansi</i>	13	25,432,160	2,439,084	46.53	0	10,109

## 1.8 Aims of the thesis

The primary aim of this project is to firstly, generate new improved *de novo* genome assemblies and annotation for two African trypanosomes that cause a huge economic impact on animal husbandry in Africa, these are *T. congolense* and *T. vivax*, as the current available genome assemblies of the two species are highly fragmented and missing important genome linkage data

within MBCs. The new sequencing technology used here PacBio SMRT sequencing could facilitate the bridging of the gaps in the MBCs in current genome assemblies to bring more integrity that might reveal potential new genomic areas on this subset of chromosomes such as protein-coding genes, paralogous/orthologous genes, DGCs, SSRs, completing fragmented gene sequences, centromere repeat sequences and telomeric repeat sequences. Such comprehensive sequences could provide a new information about interspecies differences on genomic level and the consequent effect on cell biology and potential drug targets of different African trypanosomes studied by utilizing comparative whole genome sequence comparison.

Secondly, currently the public databases are missing important genomic data as African trypanosomes evolved subsets of chromosomes, which were previously shown in *T. brucei* to host genes that are correlated with the ability of the parasite to maintain its survivability in the blood stream of mammalian hosts; these subdivisions are intermediate chromosomes and mini chromosomes. Through using third generation long reads SMRT technology, we might be able to rescue this part of the trypanosome genome and provide an integrative, detailed description of its coding and non-coding features, which in turn could provide the scientific community with valuable resources towards more understanding of trypanosome pathogenicity.

Thirdly, it is important to achieve a comparative phylogenomic study using available whole genome putative proteomic databases of different kinetoplastid organisms living in different environmental circumstances from free-living kinetoplastid, non-mammalian parasites and different intracellular and extracellular mammalian parasitic trypanosomatids along with the proteome databases that generated from *T. congolense* and *T. vivax* sequencing and annotation projects. The output of such analysis might answer many very important questions, such as, what is the core gene set that, is shared among all kinetoplastids? What are the genes that enabled parasitic kinetoplastids to be parasites? What is the gene repertoire that characterized the African trypanosomes from the other kinetoplastids? Finally, what are the gene sets that shaped the differences/similarities among three African trypanosomes (*T. brucei brucei*, *T. congolense* and *T. vivax*) more specifically

towards the developmental stages of the life cycle in the insect vector (the tsetse fly)?

Such analysis will provide important information about the parasitic lifestyle of the trypanosomes, valuable catalogues of genes that could be potential drug/vaccine targets in order to overcome these parasites and finally, the genomic innovations, expansions and contractions in gene families among the three African trypanosomes.

## **Chapter 2      *T. congolense* PacBio SMRT genome sequencing**

### **2.1 Introduction**

In this chapter, the third generation PACBIO SMRT sequencing technology was employed to generate a new genome assembly of *T. congolense* to improve the current knowledge of the genome structure and gene content of this trypanosome species.

#### **2.1.1 The origin of *T. congolense* strain IL3000**

*T. congolense* strain IL3000 is a clonal line derived from an ancestral strain called “Transmara I” originally isolated from a bovine host in the Transmara area of Kenya in 1966 (Wellde *et al.*, 1974; Hirumi and Hirumi, 1991). In order to complete its life cycle, *T. congolense* has a heteroxenous life style and needs two hosts: insect and vertebrate hosts to complete its lifecycle.

Insect stages of *T. congolense* IL3000 grow well in culture media (Gray *et al.*, 1984) and the blood form could be *in vitro* cultivated too (Hirumi and Hirumi, 1991). So, this strain has been used widely in many studies, particularly on surface proteins of different stages of its life cycle (Eshita *et al.*, 1992; Bayne *et al.*, 1993; Beecroft, Roditi and Pearson, 1993; Sakurai, Sugimoto and Inoue, 2008; Lane-Serff *et al.*, 2016).

#### **2.1.2 Previous genome sequence effort**

In order to study this economically and biologically important parasite, a genome sequence was previously produced using Sanger sequencing technology (Jackson *et al.*, 2012), which ended by generation of a draft genome sequence for the *T. congolense* strain IL3000, currently available on TritrypDB. The draft genome was organised, using the more complete *T. brucei* genome, into 11 Mega base chromosomes (MBCs) and a “Bin” which in turn was a collection of contigs that could not be assigned to the MBCs. However, the current draft assembly lacks contiguity and is interrupted by many gaps of variable size, resulting in fragmented gene models, missing

paralogues, (especially within tandemly repeated gene arrays) such as: tubulin, rRNA and variant surface glycoprotein. Moreover, some chromosomal territories could not be retrieved like VSG genes expression sites, SSRs regions, intergenic regions and directional gene clusters (DGCs). Finally, it potentially missed repeats or repeat containing features such as Mini Chromosomes (MCs) which could have particular biological functions. These limitations could complicate further downstream analyses such as population genetic studies, drug and vaccine targeting and structural variation studies. These problems with the previous genome version could have resulted from the genome biology of trypanosomes, as it has many repeated gene arrays (Jackson *et al.*, 2007) and the limitation of the Sanger sequencing technology that could not be expected to describe long stretches of repeated DNA (Chaisson, Wilson and Eichler, 2015) or high GC genomic regions (Loomis *et al.*, 2013).

Third generation sequencing technology such as the PacBio RSII SMRT sequencing (used in this study) generate long reads of up to 60 kb, these have the potential to span large DNA segments particularly those of repeated nature, and generate longer contigs that provide more contiguity and unravel potentially important structures (Ferrarini *et al.*, 2013; Rhoads and Au, 2015).

### **2.1.3 Aims and objectives of the chapter**

The main aim of this chapter is to generate an improved genome assembly that could exhibit physical contiguity, leading to a better analysis and understanding of potential large structural variations in comparison to the closest related species *T.b. brucei*, identifying new possible coding or non-coding sequences with a view to the potential biological impacts of the proposed new findings.

In order to achieve these aims we have sequenced, assembled and annotated the genome of *T. congolense* strain IL3000 using PacBio SMRT sequencing and applied comparative bioinformatics analysis approaches. We have also analysed another PacBio SMRT assembly of *T. congolense* strain Tc1/148 generated by Dr. A. Jackson (Institute of Infection and Global Health,

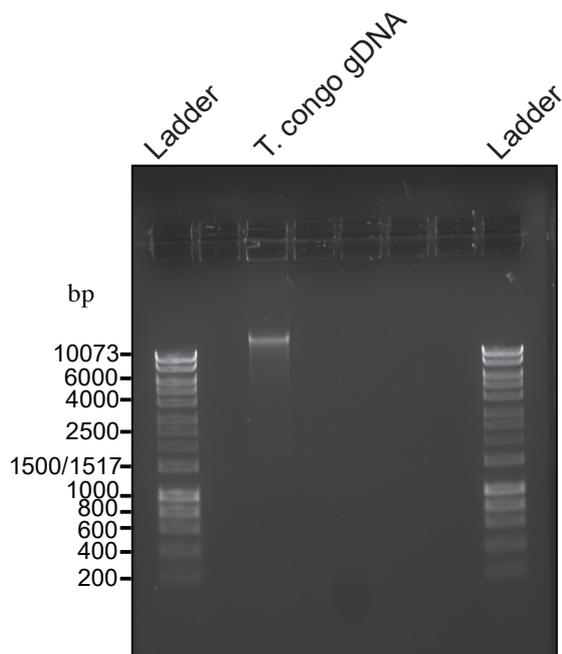
University of Liverpool) specially to support our findings of potential large structural chromosomal rearrangements in this chapter.

## 2.2 Methods

### 2.2.1 gDNA QC

The *T. congolense* strain IL3000 gDNA (15 µg) was kindly provided by Dr Liam Morrison (University of Edinburgh).

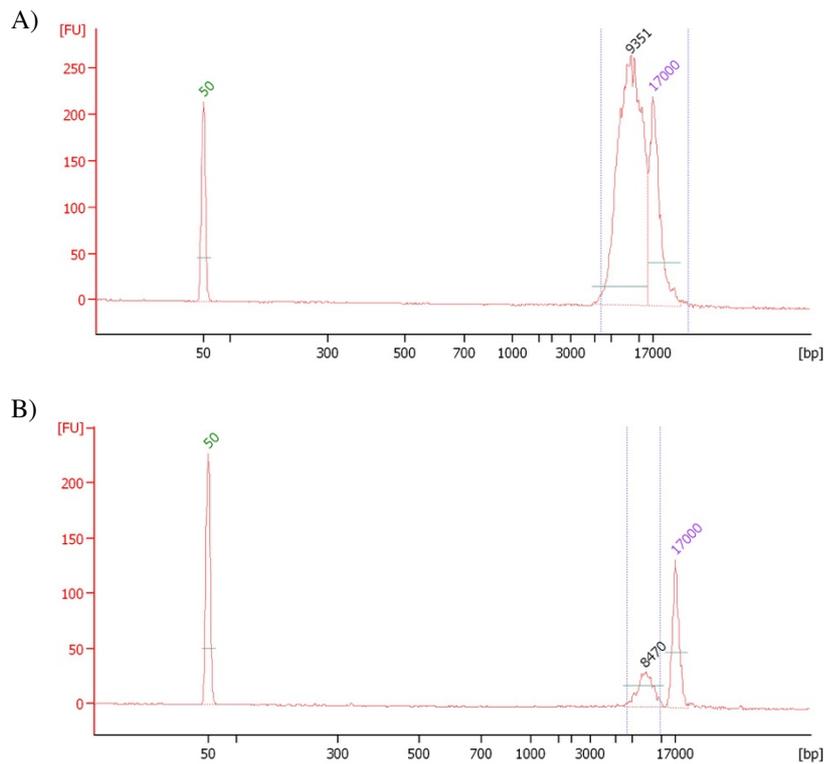
The quality of the provided genomic DNA was further checked in house to ensure the best DNA quality and purity. The presence of possible protein and/or RNA contamination were checked using QUBIT, dsDNA and RNA and protein BR Assay Kits and the results showed that the provided *TcIL3000* gDNA had undetectable amount of protein and/or RNA, which make it suitable for DNA library preparation and genome sequencing assembly later on. Furthermore, the physical integrity of gDNA was investigated by electrophoresis at 75 volts for two hours on a 1% agarose gel. Although the gDNA showed a minor fragmentation, the bulk of it was higher than 10 kb (Figure 2.1).



**Figure 2.1 1% Agarose gel of *TcIL3000* gDNA.** An electrophoretic analysis of *T. congolense* TcIL3000 gDNA (third column from the left) and Hyper ladder I (first and last columns) (Bioline, UK). The gDNA was subjected to electrophoresis for two hours at 75 volts.

### 2.2.2 Preparation of gDNA libraries for PacBio sequencing

Two different PacBio SMRT DNA libraries were prepared one targeted at 10 kb DNA and the other at 20 kb. The libraries were prepared according to the manufacturer instructions from the gDNA strain IL3000. The two different PacBio libraries were assessed before sequencing by Bioanalyzer analysis. The 20 kb library showed relatively longer fragments compared to the 10 kb DNA template (Figure 2.2).



**Figure 2.2 Bioanalyzer analyses of 20Kb and 10Kb gDNA libraries of TcIL3000.** A) 20 kb library (mean size) of 9 kb. The majority of DNA fragments are between 5-17 kb extending towards the upper marker. B) 10 kb library (average size of 8 kb) with a low peak. Numbers in green and purple represent lower and upper standard DNA markers.

The DNA libraries were loaded separately onto the RSII PacBio SMRT sequencer as follows: four cells for the 10 kb and eight cells for the 20 kb library protocol.

### 2.2.3 Genome databases

The genome databases of the reference sequence of:

- *Trypanosoma brucei brucei* strain TREU927 v5.1 GeneDB (<http://www.genedb.org/Homepage/Tbruceibrucei927>)
- *Trypanosoma congolense* strain IL3000 v26 <http://tritrypdb.org/>

The PacBio genome assembly of *T. congolense* strain Tc1/148 was kindly provided without annotation from Dr Andrew Jackson (Institute of Infection and Global Health/University of Liverpool) and is under GenBank accession number NHOR00000000.

### 2.2.4 De novo Genome assemblies

#### 2.2.4.1 SMRT portal analyses version 2.3

More than 1.7 Gb of data were generated and *de novo* assembled using the PacBio Hierarchical Genome Assembly process versions 2 & 3 (HGAP\_Assembler 2 & 3) (Chin et al., 2013) with diploid genome analyses option and a target genome size of 40 Mb, (the estimated genome size for *T. congolense*) with the other parameters left at default. It was then polished with QUIVER (Chin et al., 2013) for error correction. Both tools were implemented in SMART portal version 2.3.

#### 2.2.4.2 Other assemblers

Another long-read PACBIO genome assembler CANU version 1.2 (Koren et al., 2016) was used to generate TcIL3000 genome assemblies from the PacBio RSII SMRT sequencer raw reads “fastq” output files by adopting two protocols:

Firstly, on filtered PacBio reads on full default settings with targeted genome size set to 40 Mb – the resulting assembly was named “CANUdefault”.

Secondly, CANU was allowed to correct and assemble all reads with an assumed error rate of 0.035, the other parameters were left on default – the resulting assembly was named “CANU2”.

## **2.2.5 Assessment of the *de novo* PacBio TcIL3000 genome**

### **2.2.5.1 Genome statistics**

The statistics such as the total size of the assembly, N50 size, GC%, max contig size, mini contig size and gap size of all analysed assemblies were calculated using custom Perl script kindly provided from a PhD student Laura Gardiner.

### **2.2.5.2 Validation of gene models in genome assemblies using BUSCO**

In order to compare between different assemblies for gene completeness, an assembly comparison approach was adopted using conserved core eukaryotic gene subset for “protists” implemented in the BUSCO tool version 2 (Simão *et al.*, 2015). Briefly, this package uses BLASTx sequence similarity to search for protist core genes in the assembly and utilizes AUGUSTUS gene prediction tool (Stanke *et al.*, 2004) for gene model prediction, then uses Hidden Markov Model (HMM) (Söding, 2005) to transfer domains (functional assignment) to the predicted gene set. The final step is to generate statistics on complete, duplicated, fragmented and missing gene models in the tested assembly in comparison to a reference database of 215 conserved protists genes implemented in this tool. The tool was used with the default options and the summary comparison plots were generated using “generate\_plot.py” script implemented in the same tool suit.

### **2.2.6 Generation of scaffold level assembly**

SSPACE\_LongReads package version 1.1 (Boetzer and Pirovano, 2014) was employed to scaffold PacBio contigs based on filtered reads after mapping them back to the contigs generated by HGAP3 assembler using “BLASR” (Chaisson and Tesler, 2012) implemented in SSPACE\_LongReads. The default options were used to generate the scaffold-level assembly of *T. congolense* strain IL3000.

### **2.2.7 Chromosomal-level assembly of MBCs (pseudo-chromosomes)**

The chromosomal level assembly is so important to putatively build chromosome sequences of sequenced genome of an organism based on

genome synteny to a closest organism reference sequence, which could facilitate investigation of the structure of the chromosomes and possible structural variants on a genomic level.

In order to achieve this goal, genome sequences can be aligned to a reference genome of a closely related species. Although, there is a draft reference sequence for *T. congolense* strain IL3000 available ([http://tritrypdb.org/common/downloads/Current\\_Release/TcongolenseIL3000/fasta/data/](http://tritrypdb.org/common/downloads/Current_Release/TcongolenseIL3000/fasta/data/)), it is highly fragmented with large gaps interrupting the contiguity of the pseudo-chromosomes sequences (which represent hypothetical chromosome structure); therefore, the order was based on the *T. brucei* TRUE 927 reference genome. The assumption being that this would provide a better framework for comparative genomic analysis.

The ABACAS tool version 1.3.1 (<https://sourceforge.net/projects/abacas/files/>) was used to build the chromosomal-level pseudo scaffolds. The general principle used was to assume conserved synteny between the query and the reference genome (*T. b. brucei* strain TREU927). Thus, query scaffolds with high similarity to a reference were placed in the corresponding chromosomal-level pseudo scaffold bin. Pseudo scaffold bins were then reoriented, ordered and aligned to the reference MBCs. ABACAS was run using protein based sequence homology implemented using the PROMmer algorithm (Delcher, 2002) of MUMmer package (Kurtz *et al.*, 2004) version 3 to accurately align the query to the reference. The ABACAS 1.3.1 options used were “maxmatch” to increase the alignment sensitivity, “promer”: for amino acids based alignment algorithm, “m”: to print the ordered and oriented contigs into a separate file, “b” generate a file that contains the contigs that not assigned to any of the inferred pseudo scaffolds (the Bin).

The output was then viewed using ACT version (Carver *et al.*, 2005) to check the aligned scaffolds to the reference chromosomes. However, the final output of the query assembly was not yet split into chromosomal-level pseudo scaffolds, so *T. congolense* IL3000 11 MBCs were then obtained using the perl

script “splitABACASunion.pl” implemented in the PAGIT pipeline (Swain *et al.*, 2012).

ABACAS 2 (<https://github.com/satta/ABACAS2>) was also used (locally or as implemented in the COMPANION annotation pipeline (Steinbiss *et al.*, 2016) to generate chromosomal-level assembly. The stand-alone version was run with an identity cut off percentage of 55% and sequence coverage length of 100 bp; these options were chosen in order to increase its sensitivity. While the parameters for the automated annotation pipeline version was used on default settings (i.e. 90%, 500 bp) or with increased sensitivity of 45% and coverage of 200 bp.

### **2.2.7.1 Mapping PacBio reads back to the genome assembly**

In order to check contig and scaffolding integrity, the PacBio reads were mapped to the scaffold assembly and chromosome level assembly BWA version 0.7.5a-r405 BWA-MEM algorithm (Li, 2013) was used because it supports mapping of long reads. Option “M” was used to ensure outputs were compatible with PICARD tools, the resulted alignment file was further processed and indexed using PICARD tools version 1.138 (<http://broadinstitute.github.io/picard/>).

### **2.2.7.2 Illumina paired-end sequencing**

In order to correct the PacBio assembly using Illumina data, gDNA of the *Tc/L3000* isolate used for the PacBio sequencing (section 2.2.1) was submitted to University of Liverpool Centre for Genomic Research (CGR). The insert size target was 250 bp, and paired-end PCR-amplification free DNA libraries were made according to the manufacturer’s protocol. The DNA library was sequenced on the Illumina HiSeq2500. The resulted raw reads were trimmed and quality filtered and the final fasta of filtered reads were received from CGR.

### **2.2.7.3 Error correction and gap filling using PILON tools**

Filtered Illumina paired-ends reads were used via PILON version 1.16 for correction (Walker *et al.*, 2014). This step was adopted to overcome possible

errors due to the relative low average coverage (47-fold) of the PacBio assembly.

PILON error correction and gap filling tool requires mapped short reads to the target assembly. For this purpose, BWA-MEM algorithm (Li, 2013) was used with option “M” to ensure outputs were compatible with PICARD tools version 1.138 ( <http://broadinstitute.github.io/picard/> ).

Generated sorted, duplicates were marked and BAM file was indexed using PICARD tools. These steps were carried out for three PILON iterations.

## **2.2.8 Genome Annotation**

### **2.2.8.1 Automated genome annotation pipeline**

Web based COMPANION (Steinbiss *et al.*, 2016) automated protozoa genome annotation pipeline was adopted. The pipeline automatically annotates protein coding genes and non-coding genes (e.g. *tRNA*, *rRNA* *ncRNA*, etc.).

In brief, the COMPANION pipeline runs the following processes:

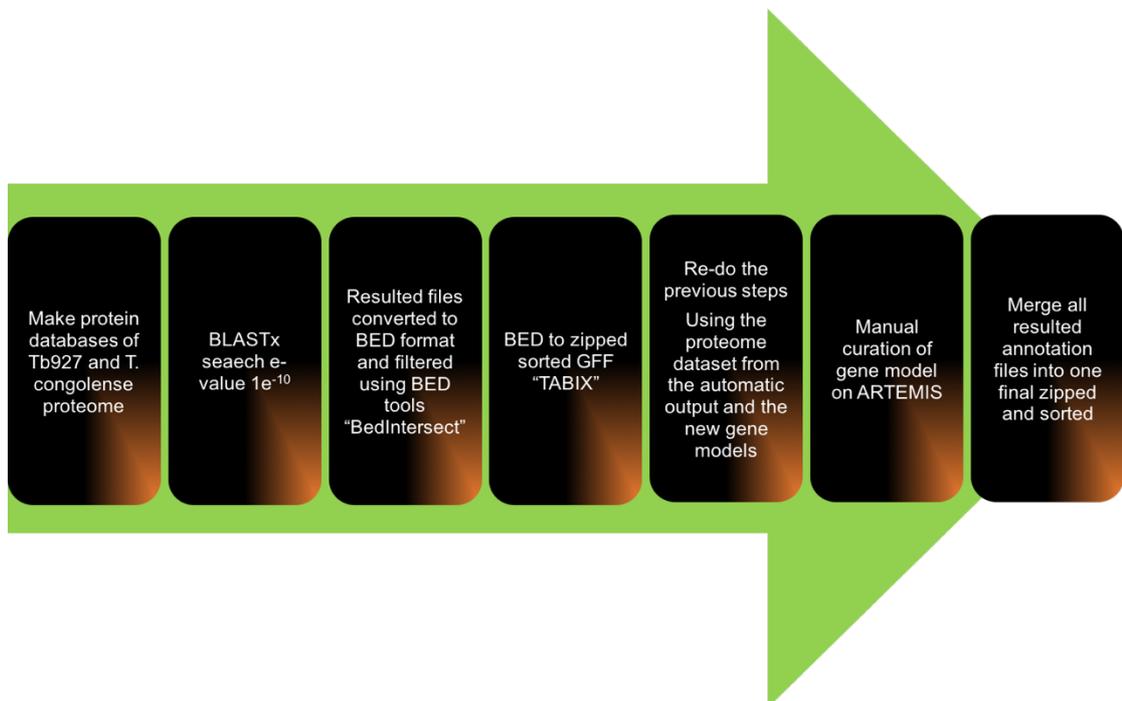
1. Chromosomal level assembly is inferred using ABACAS 2 based on a provided reference.
2. Protein coding sequences are transferred from the reference to the new sequence when there is correspondence using RATT (Otto *et al.*, 2011) (for this step “species” mode was used).
3. *Ab initio* gene calling using AUGUSTUS (Stanke *et al.*, 2004) depending on protein evidence (default threshold of 0.8 was adopted).
4. Protein domain annotation by two means:
  - a. Using HMMER (Johnson, Eddy and Portugal, 2010) to search against Pfam database (Finn *et al.*, 2016).
  - b. Transfer domains from reference proteins using “OrthoMCL” a protein clustering tool (Li, Stoeckert and Roos, 2003).
5. Non-coding protein gene annotation e.g. *tRNA*, *rRNA* and *ncRNA* was implemented using INFERNAL (Nawrocki and Eddy, 2013), HMM homology search and ARAGON (Laslett and Canback, 2004) algorithm of

homology based search query sequences based on prokaryotic, eukaryotic tRNS consensus structure for *tRNA* genes were employed for this task.

6. Finally, output files are generated containing annotated features in GFF3 feature file format, EMBL format and GAF file.

### 2.2.8.2 Manual genome annotation

Manual inspection of the annotation file (predicted features of automated annotation pipeline) was used to check for gene call accuracy using ARTEMIS (Carver *et al.*, 2012), which revealed a number of putative Open Reading Frames (ORFs) that showed significant hits to existing protein-coding genes on the TriTrypDB database (<http://tritrypdb.org/tritrypdb/showQuestion.do?questionFullName=GenomicSequenceQuestions.SequencesBySimilarity>). A workflow was designed (Figure 2.3) to fill these possible annotation gaps using protein evidence.



**Figure 2.3 Manual annotation work flow of *T. congolense* IL3000 PacBio genome assembly.**

**Step one:** Protein databases were created from the reference protein database of *T. b927* and the available protein database of *T. cIL3000* using “makeblastdb” implemented in BLAST+ (Camacho *et al.*, 2009) package version 2.2.28.

**Step two:** The BLASTx search using an e-value cut-off  $1^{-10}$  was applied on the final PacBio assembly against the pre-prepared protein databases with tabulated output formatted files, in order to be able to generate different file formats and to view these hits on the assembly using ARTEMIS.

**Step three:** The tab delimited output files were further converted into “BED” formatted files using custom Perl script in order to be ready for cleaning using BEDTools version 2-2.25.0 (Quinlan and Hall, 2010). This filtering step was essential as the files could be too noisy and too large, which could lead to the sequence browser (ARTEMIS) to be crashed. Option “intersectBed” was used along with option “v” in order to generate output “out.bed” file that contained hits that do not overlap with the automated annotation file output or the hits from the other proteinDB.

**Step four:** BED formatted output files from previous step were then converted into “GFF” file format using custom Perl script (Appendix A 1), which in turn were compressed and sorted by TABIX tools (Li, 2011).

**Step five:** Proteome databases resulting from automated and manually annotated genes were used as a new protein database and another BLASTx search was achieved with the same previous parameters. This step was accomplished to permit the annotation of possible paralogues, which perhaps with sequences perhaps far from evoking significant hits by previous search steps.

**Step six:** the final annotation file was overlaid on the PacBio IL3000 genome sequence, the BLASTx hits were then extended if necessary to the first start codon (i.e. 5' methionine) and/or the stop codon (i.e. 3' TAG/TAA/TGA), conditionally, if the BLASTx hits involved within-frame stop codon/s and/or the hit extended through more than one sequence frame, the annotated model was considered as a pseudogene. Then the gene model was created as a

parent file with a parent new gene ID for the putative new model, its transcript and protein sequence, while the gene name was inherited from the Blast hit. Consequently, an annotation file was generated by ARTEMIS and ready to be merged with other manually created annotation files from other datasets.

**Step seven:** Finally, the annotation files of all manual steps and the automated step were then merged and sorted using “GenomeTools” version 1.5.8 (Gremme, Steinbiss and Kurtz, 2013) with option “sort” enabled to generate the final fixed GFF3 annotation file.

The new PacBio genome assembly of *T. congolense* strain IL3000 was deposited in NCBI with the GenBank accession number PQVL00000000.

### **2.2.9 Clustering of proteomic data of analysed assemblies**

OrthoFinder pipeline version 1.0.0 (Emms and Kelly, 2015) was used for clustering all proteome data sets across all selected species. The rationale beyond using this clustering procedure is to deduce the sequences that shared similarity among the analysed data and clustering them into phylogenetically related clusters (orthogroups), leaving those sequences with no similarity to any other protein sequence as independent sequences (singletons).

Briefly, this pipeline uses BLASTp search to infer the pair-wise sequence similarity score across each pair of proteome datasets with an expectation value cut-off threshold of  $1e10^{-3}$ . However, it considers the bit score rather than e-value score in order to avoid the biases raised from sequence length, which permits to increase the accuracy of calling sequences with shared similarity in a dataset over other currently used methods like OrthoMCL (Emms and Kelly, 2015). Then Markov clustering Algorithm (MCL) (Dongen, 2000) is applied on the Blast results to generate protein clusters according to their similarity. The default value of MCL “-I” was set to (1.5) This could generate main result files including tabulated result files with Orthogroups (rows) containing sequence IDs from species assigned to each cluster (columns). Then a distance matrix was calculated using “dendroblast” by default followed by the “fastme” tool to infer gene trees for all resulted orthogroups.

The default options of the pipeline were used to cluster the proteomic data sets of *T. brucei* TREU927, the *T. congolense* IL3000 Sanger based assembly and the *T. congolense* IL3000 PacBio assembly.

#### **2.2.9.1 Obtaining shared genes from OrthoFinder output**

OrthoFinder pipeline provides statistics on the number of unique and shared genes among the entered datasets so that the number of genes shared among specific combination of assemblies could be obtained.

#### **2.2.9.2 Directional gene clusters**

Clusters of genes on TcIL3000 MBCs skirted by non-coding DNA stretches on both ends of each DGC were viewed on ARTEMIS, and the regions length were calculated according to the first and last protein coding genes coordinates of each cluster.

### **2.2.10 Gene Ontology enrichment analysis**

#### **2.2.10.1 Extraction of the Gene Ontology term (GO) IDs**

*T. brucei* sequence IDs from OrthoFinder results of shared orthogroups with *T. congolense* and for *T. congolense* singleton gene IDs of PacBio assembly assigned by the automatic annotation pipeline were used in the analyses in this chapter. These IDs were extracted and used as a query list to extract the GO IDs from the genome annotation file (in GFF3 format) when available, using a combination of LINUX command lines (Appendix A. 2). The final output of this step is a text file contains a list of GO term IDs that could then be used as an input for the next step.

#### **2.2.10.2 Enrichment analysis of GO terms**

The list of extracted GO terms in the previous step could be redundant and hard to interpret to extract meaningful and prominent functional description to the extracted subset/s. A web based gene function enrichment analysis and visualization tool called “REVIGO” (Supek *et al.*, 2011) and cellular component and molecular function using GO slim classification method implemented in

CateGORizer web application (Hu, Bao and Reecy, 2008) were employed to view each set of GO terms used in this chapter. The principle of REVIGO tool is to use a semantic hierarchical clustering approach of the entered GO terms, removing the redundancy and revealing the parenteral terms and using two dimensions to view these terms in scatterplot or interactive graphs. The size of a sphere or a square in the scatterplot represents the number of texts synced in that common category (i.e. larger square size indicates more generalized term), and the colour refers to the increased “uniqueness” of the term. In this context, there are two contrary relationships in this approach, they are: “uniqueness” and “dispensability” [each one represents a column of values (lies between 0-1) in the REVIGO analysis]. The term “dispensability” signifies the possibility of a term to be similar with other terms in the set of the analysed data. REVIGO categories terms showed more than 0.7 dispensability value with other generalized terms “dispensable terms”, zero dispensability means unfeasibility to rank that term with the other terms which then represents its high uniqueness value. Therefore, the degree of red colour in the scatterplot, the higher probability of uniqueness value of that term.

### **2.2.11 Genome visualization**

Genome inspection for gene model integrity and sequence layout for all manual investigation steps, manual annotation and generation of annotation files from the manually annotated gene models in this chapter were achieved using ARTEMIS version 16.0.0 (Carver *et al.*, 2012).

The sequence comparison tool ACT (Carver *et al.*, 2005) was employed for investigation and exporting plots of proposed chromosomal structural rearrangements among different chromosomes of the compared assemblies.

#### **2.2.11.1 Plots generation of possible genome rearrangements**

In order to view the predicted genomic regions, such as chromosomal translocations between *T. brucei* and *T. congolense*, two tools were used to examine them:

### **2.2.11.2 Plots generated using GenomeRibbon tool**

A web based screening tool of large structural variants “GenomeRibbon” (Nattestad, Chin and Schatz, 2016) was used in order to present possible shared regions between two different sequences, a tab delimited comparison file based on NUCMmer or PROMmer algorithms employed in MUMmer package version 3, which stores the related sequences and the position of the predicted common regions. Here, PROMmer algorithm was evoked with “maxmatch” and the minimum clustering option was set to 200 to increase the search sensitivity. The resultant files were uploaded to the web application and viewed on a web browser.

In order to show which reference genes are possibly affected by the proposed DNA segmental movements, an annotation file should be uploaded to the application, which stores information about the reference features in “BED” file format (<https://genome.ucsc.edu/FAQ/FAQformat.html#format1>). So a six columns file stores required information was generated from the original “GFF” annotation file format of the reference *T.b. brucei* strain TREU 927 using appropriate LINUX command line. Then this file was uploaded to view the reference genes on regions of interest.

### **2.2.12 Synteny to the reference chromosomes**

Inspection/plots generation to show synteny of the assembly contigs/scaffolds/pseudochromosomes to the reference Tb927 MBCs were generated using “mummerplot” tool. This was implemented in MUMmer package version 3 based on PROMmer comparison file output after the “maxmatch” option was enabled and the minimum clustering option was set to 200 to increase the search sensitivity.

#### **2.2.12.1 Plots generated using ACT sequence comparison tool**

Sequences of relevant chromosomes from the two species along with the related pre-prepared comparison file/s were uploaded and viewed. The required plots were then exported from ACT (Carver *et al.*, 2005) version 16.0.0.

Comparison files- are the files that comprise information of shared DNA segments between DNA sequences subjected for evaluation, which make it possible to view by ACT. These files were generated using tBLASTx search implemented in BLAST+ (Camacho *et al.*, 2009) package version 2.2.28. The tab delimited output was enforced with an e-value cut-off limit of  $1e^{-10}$ .

### **2.2.13 Plotting annotated features on *T. congolense* MBCs**

The plots of putative karyotype of *T. congolense* IL3000 PacBio MBCs with the related annotations were generated using R package “KaryoploteR” on R-project (R Core Team, 2016) version 3.3.3. using RStudio version 1.1.3. (RStudio, 2016) <http://www.rstudio.com/>.

#### **2.2.13.1 Venn diagram**

The plots generated to view protein cluster analysis shared between assemblies or assembly specific groups was generated using Rpackage “VennDiagram” (Chen and Boutros, 2011) on R-project using RStudio.1.1.3.

### **2.2.14 Inference of putative centromeres**

#### **2.2.14.1 Inference of possible centromere repeat on TcIL3000 PacBio assembly**

The centromere annotated in the *T. brucei* 927 genome assembly on chromosome one was used as an example for searching for centromeres in the TcIL3000 PacBio assembly. Consequently, a search for analogous region in *T. congolense* pseudochromosome one was performed using ARTEMIS. The region on *T. congolense* MBC one was characterised by simple tandem repeats, the repeat unit length and the GC content were analysed using web based Tandem Repeat Finder (TRF) tool (Gary Benson, 1999) version 4.0, while the coordinates of repeat region in *T. congolense* PacBio assembly were inferred by Red software (Girgis, 2015).

#### **2.2.14.2 Search similar regions on the other *T. congolense* MBCs**

A BLASTn based search with a cut off evalue of  $1e^{-10}$ , tab delimited output file and “megablast” search mode enabled in BLAST+ package (Camacho *et al.*, 2009) version 2.2.28 were conducted based on the region extracted from *T. congolense* MBC one and eleven, which were used as a search database for possible satellite repeats on the genome assembly.

#### **2.2.15 Strand Switch regions on *T. congolense* PacBio assembly**

SSRs were viewed and their DNA sequences were extracted manually using ARTEMIS, while their length was obtained according to the coordinates of the first flanking protein coding features. RepeatMasker package version 4.0.7. was used to search for possible conserved elements within these chromosomal regions. The parameters of “crossmatch”, ‘Trypanosoma brucei’ repeat library database and the ‘gff’ format of result file were allowed.

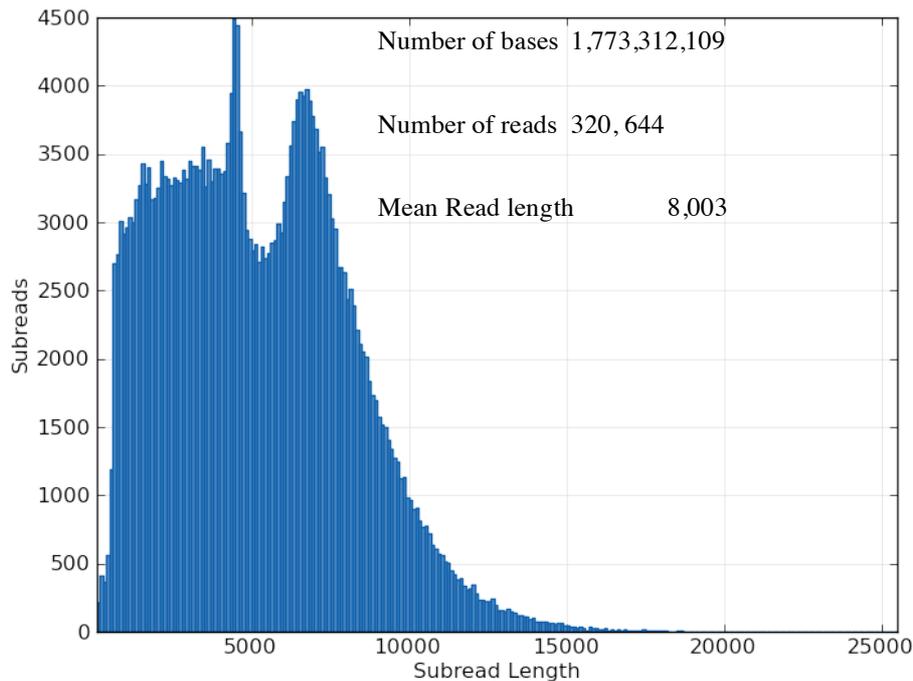
#### **2.2.16 Statistical analysis**

Two sample t-test on R-project (R Core Team, 2016) <http://www.R-project.org>. was used to test statistical significance of the difference in length among divergent and convergent SSRs and forward and reverse of DGCs.

## 2.3 Results and discussion

### 2.3.1 RSII PacBio sequencing and *de novo* assembly

The 12 SMRT cells generated 1.7 Gb of raw data. The PacBio RSII insert size analyses of the prepared DNA libraries have the mean of 8 kb (Figure 2.4).



**Figure 2.4** *TcIL3000* sub reads distribution of the 12 SMRT cells output. The reads length showed two peaks of one about 5 kb and the other 8 kb.

### 2.3.2 Genome Assembly

Of the PB SMRT analysis portal *de novo* assemblers HGAP3 and HGAP2, the latter tool failed to generate a complete assembly dataset as the total genome size from this assembler was only 25 Mb (approximately just slightly over half of the estimated haploid genome size). Therefore, it has been excluded from downstream analysis. In this context, Pacific Bioscience suggest that the HGAP assemblers show better performance when the genome coverage is higher than 50-folds (<https://github.com/PacificBiosciences/Bioinformatics->

[Training/wiki/Large-Genome-Assembly-with-PacBio-Long-Reads](#) ) (Chin et al., 2013). Hence, the sequence coverage could be the most likely reason that halted the HGAP2 assembly from generation of a complete assembly.

CANU assembler was applied on the raw PacBio sequencing data using two modes CANUdefault and CANU2 (see section 2.2.4.2). The classical statistics of assembly metrics showed superiority of both CANU assemblies over the HGAP3 assembly (Table 2.1).

Although a default run of CANU has resulted in the best statistics, it showed highly reduced total assembly size. This could be as a result of strict parameters applied in the default options that trying to generate longer contigs by excluding long repeats (Koren *et al.*, 2016). Repeat sequences are mostly accountable for the assembly fragmentation or even assembly failure, as it is hard to assemble them during the assembly process (Phillippy, Schatz and Pop, 2008; Nagarajan and Pop, 2009). Such interpretation could be further evidenced by our findings of less fragmentation and the highest GC content of the assembly resulting from the default options of CANU over the other assemblies (Table 2.1). Nonetheless, CANU2 with revised parameters showed closer statistics to HGAP3 assembly, yet it had better contiguated sequence, perhaps due to allowing to correct all sequence reads and permit for more reads to be overlapped.

**Table 2.1 PacBio contig level assemblies.** Contigs statistics comparison among different assemblers.

Assembly statistics	HGAP3	CANU2	CANU default
<b>Assembly size bp</b>	39,250,269	39,366,815	30,211,476
<b>Contigs' number</b>	1,541	1,303	344
<b>Max contig size bp</b>	1,475,421	1,880,585	1,866,514
<b>N50 contig size bp*</b>	156,211	231,624	723,744
<b>N25 contig size bp</b>	536,182	972,722	1,167,850
<b>GC percentage</b>	45.89%	46.13%	47.99%

\* N50 contig size: is the length of the contig where half of the assembled genome is in the contigs of this length and above.

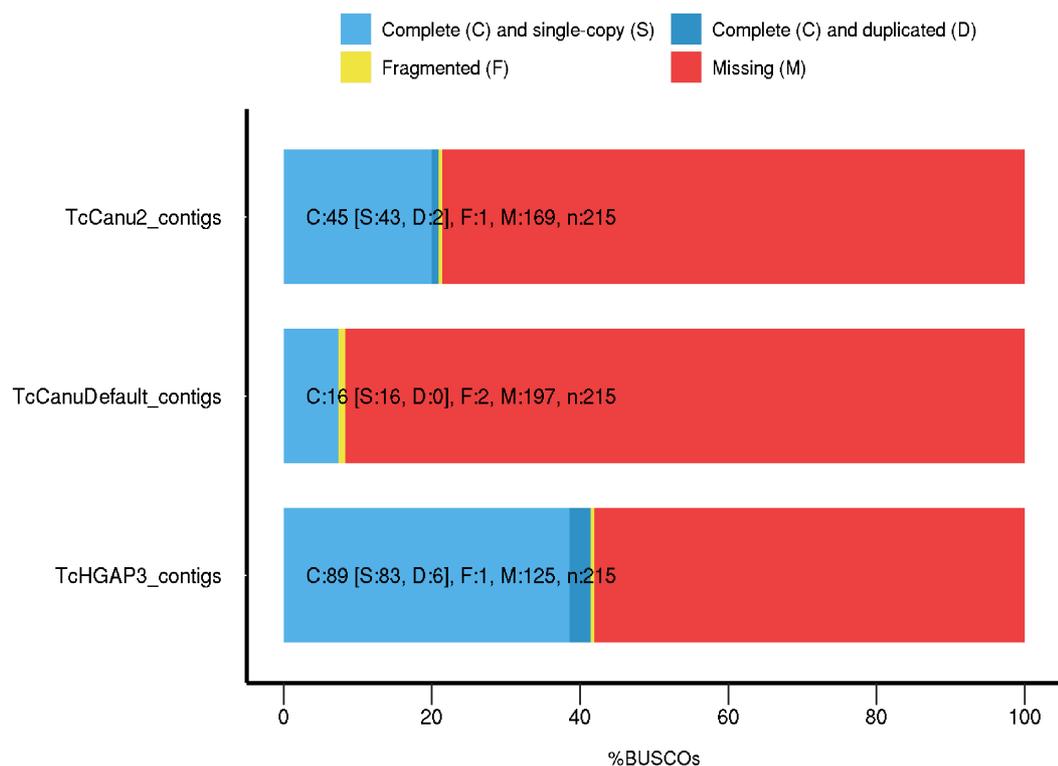
### 2.3.2.1 Assembly Validation

Indeed genome assembly metrics like N50 length are important aspects for assembly validation, yet the assembly has to maintain the integrity of gene models (Ekblom and Wolf, 2014).

Whole genome gene-model integrity testing approach was adopted in order to validate the expected gene models for each PacBio assembly of the *T. congolense* genome (see sections 2.2.5). Interestingly, BUSCO comparison results suggested that the HGAP3 assembly revealed the best gene models over the two CANU assemblies. Astonishingly, it proposed double as the number of complete conserved tested gene set as that of the CANU2 assembly, which in turn has approximately three times more than CANUdefault (Figure 2.5). Moreover, the manual gene-model inspection result was consistent with the BUSCO tools.

The possible reason for the failure of the CANU based assemblies could be due to the relative low genome coverage (47-fold), while the best performance of such assembly approaches requires at least 75-fold genome coverage (Salmela *et al.*, 2017).

Taking together, although the two CANU output displayed the best classical assembly statistics, they indeed failed to present accepted level of complete gene models, which could have a crucial impact on the integrity of the predicted gene models in the annotation step and the following downstream data mining. Therefore, the HGAP3 assembly was decided to be the assembly of choice to undergo the downstream analyses.



**Figure 2.5 BUSCO comparison of PacBio contig level assemblies.** HGAP3 (TcHGAP3\_contig) and two CANU assemblies. Default CANU parameters were used (TcCanuDefault\_contigs). Correction of all PB reads (TcCanu2\_contigs).

### 2.3.2.2 HGAP3 assembly

The total contig number produced by the HGAP3 implemented in SMRT portal version 2.3 was 1,541, with a total consensus assembly size of more than 39Mb and a mean coverage of 47x, half of the contigs (N50) were more than 156 kb in size with a max contig length of 1,475,421 bp while the smallest contig length was 645 bp.

The assembly statistics presented promising results regarding the sequence continuation represented by a relative high N50 contig length, and there were no gaps (the gaps in the sequence symbolized by Ns) in PacBio (PB) contig level assembly. So, there is hope for the possibility of closing many gaps and resolving ambiguous nucleotides in the current reference, as the PB N50 contig length is even higher than the max contig length of Sanger sequence (Table 2.2) with a PB maximum contig size larger than 1.4 Mb. For example, it could fit to the largest pseudochromosome of the reference trypanosomes genome *T.brucei* TREU927 (chromosome 11, 5 Mb in size), it will span over more than a quarter of its total size in one contiguated contig. The importance of having long reads to generate such long contigs using PacBio SMRT sequencing is to unravel important genomic regions, structural variants and possible gap closure. These benefits of SMRT sequencing technology were also referred to by other researchers (Berlin *et al.*, 2015; Rhoads and Au, 2015).

The relative lower GC percent in PacBio assembly is perhaps because of resolving repeat regions of (AT-rich) nature such as those in centromeres (Rovira, Beermann and Edström, 1993; Lamb and Birchler, 2003; Sun *et al.*, 2003) and in the strand switch regions (Obado *et al.*, 2007).

**Table 2.2 Contig level assembly statistics of *T. congolense* assemblies.** Comparison between *T. congolense* IL3000 PacBio HGAP3 assembly and current available Sanger based sequence assembly.

Assembly statistics	<i>T. congolense</i> Sanger	<i>T. congolense</i> PacBio
Number of Contigs	3,101	1,541
Total assembly bases	30,463,056	39,250,269
Max contig length	143,854	1,475,421
N50 length	14,753	156,211
N25 length	29,119	536,182
GC percentage	48.35%	45.89%
Gaps percentage	0.01%	0%

### 2.3.3 Generation of assembly scaffolds

Two assembly contigs or more could be joined altogether to generate longer putative DNA segments called a “Scaffold” if there is enough evidence to allow for such a coupling process, this approach was used to generate more structurally correlated genome sequences, which could be accomplished using suitable tools.

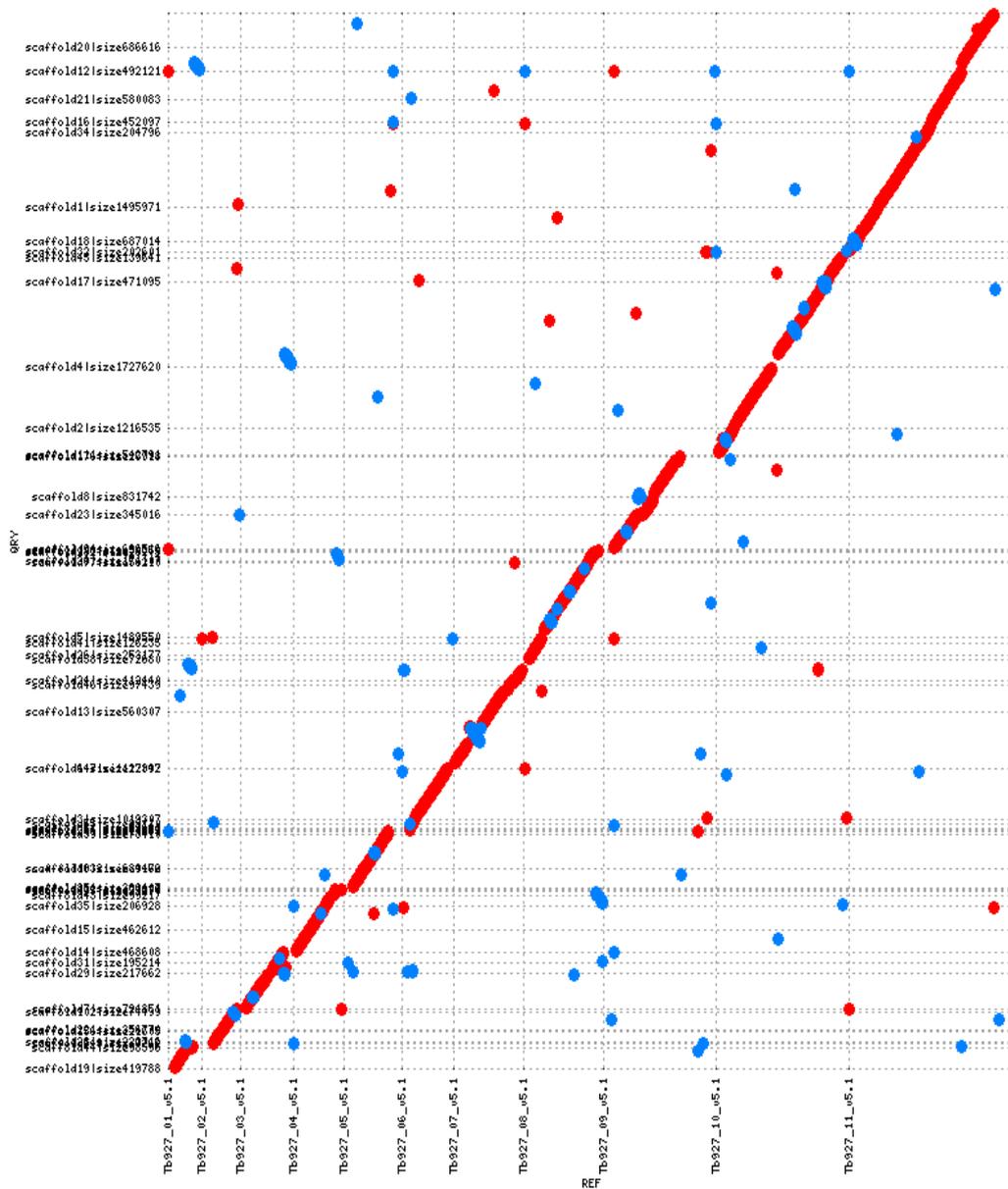
SSPACE-LongReads package was retained (see section 2.2.6); by adopting this approach the assembly was enhanced as 1,016 scaffolds were generated from the original 1,541 contigs and improved the other assembly statistics, maximum contig length was increased by more than 250 kb and N50 length was also increased by 50 kb. However, it introduced gaps in the final scaffolds

assembly. These gaps inflated the total assembly size by more than 400 kb (Table 2.3).

Reordering and aligning the assembly scaffolds to the reference sequence Tb927 pseudochromosomes revealed good synteny to the reference and exhibited some possible conversion points as viewed by MUMmerplot (Figure 2.6).

**Table 2.3 Comparison between contigs' assembly and SSPACE-LongReads scaffolds' assembly.**

<b>Assembly statistics</b>	<b>HGAP3</b>	<b>SSPACE-LongReads</b>
<b>Assembly size bp</b>	39,250,269	39,718,340
<b>Contigs' number</b>	1,541	1,016
<b>Max contig size bp</b>	1,475,421	1,727,620
<b>N50 contig size bp</b>	156,211	251,440
<b>N25 contig size bp</b>	536,182	696,516
<b>GC percentage</b>	45.89%	45.26%
<b>Gaps percentage</b>	0%	1.40%



**Figure 2.6 Sequence comparison of *T. congolense* PacBio scaffolds to the Tb927 chromosomes sequences.** Large scaffolds (Y axis) aligned to the Tb927 sequences of MBCs (X axis) generated by MUMmerplot application using protein sequences similarity for higher sensitivity and cluster size length of 200. Sequence inversions can be seen (blue dots), with scattered dots showing possible repeat or duplicated regions. The Tb927 reference Bin was removed to reduce figure noise.

### 2.3.4 Chromosomal-level Pseudo Scaffolds.

Assigning contigs or scaffolds to a putative haploid chromosomal size genomic segment called a “pseudochromosome” or “pseudo-molecule” has numerous benefits, such as providing the structural bases for comparative genomics studies and spatial information about linear gene order which are highly likely to be co-expressed, especially, in the trypanosomatid organisms when their genes are clustered in arrays called “Directional Gene Clusters (DGCs)”. Structural variants like inversions, indels and intra or inter chromosomal rearrangements between two or more different strains or species could be more easily detected in highly continuous genome sequences (Zapata *et al.*, 2016). Therefore, a reference based synteny orientation and reordering approach was adopted on our scaffold assembly in order to create larger sequences with the aim to have chromosomal sized segments (Assefa *et al.*, 2009) (see section 2.2.7).

Two bioinformatics tools were used to generate chromosome level information. ABACAS 1 showed better genome metrics by assigning more contigs to the predicted pseudo chromosomes with reduced contig size than ABACAS 2 did (Table 2.4). Although ABACAS2 is well suited for the automatic annotation pipelines, it lacks user control over different parameters; most importantly choosing between NUCMmer and PROMmer algorithms (NUCMmer algorithm was enabled by default), which probably make it fit the genome comparison studies among different strains rather than species. Such limitations in ABACAS2 and flexibility of ABACAS1 tools was led to apply the latter for this important step in the genome assembly.

The adopted approach using ABACAS1 generated 11 putative chromosomes showing synteny to the corresponding MBCs of Tb927 reference genome sequence.

**Table 2.4 Inference of pseudochromosome level assembly of TcIL3000 using Tb927 as a reference genome sequence.** Comparison between the output of ABACAS1 and ABACAS2.

	ABACAS1 PsChr*	ABACAS2 PsChr*	ABACAS1 Bin**	ABACAS2 Bin**
<b>Number of Scaffolds***</b>	55	32	961	984
<b>Size in bp</b>	21,276,030	13,649,841	18,442,310	26,071,699
<b>GC %</b>	48.62	48.19	41.39	43.72

\*PsChr: Referred to the (pseudochromosomes) putative chromosomes inferred from 11 haploid MBC of Tb927.

\*\*A location contains all sequences that do not have enough evidence of sharing synteny with the reference. This could include scaffolds/contigs with species-specific genes, repeat sequences and mini-chromosomes for the latter “see chapter three”.

\*\*\*Number of scaffolds assigned to each category (total number of scaffolds was 1,016 “see Table 2.3”).

### 2.3.5 Sequence error correction and gap filling

PacBio SMRT sequencing platform employs polymerase chemistry which increases the life span of this enzyme, allowing it to produce longer reads, which is the powerful feature of this sequencing platform. However, this comes with a cost, as the error rate is considered to be relatively high (up to 15%) (Eid *et al.*, 2009; Salmela *et al.*, 2017). Fortunately, these errors are random and the assembly algorithms can overcome this downside when there is a relatively high coverage rate above 70 fold (Salmela *et al.*, 2017).

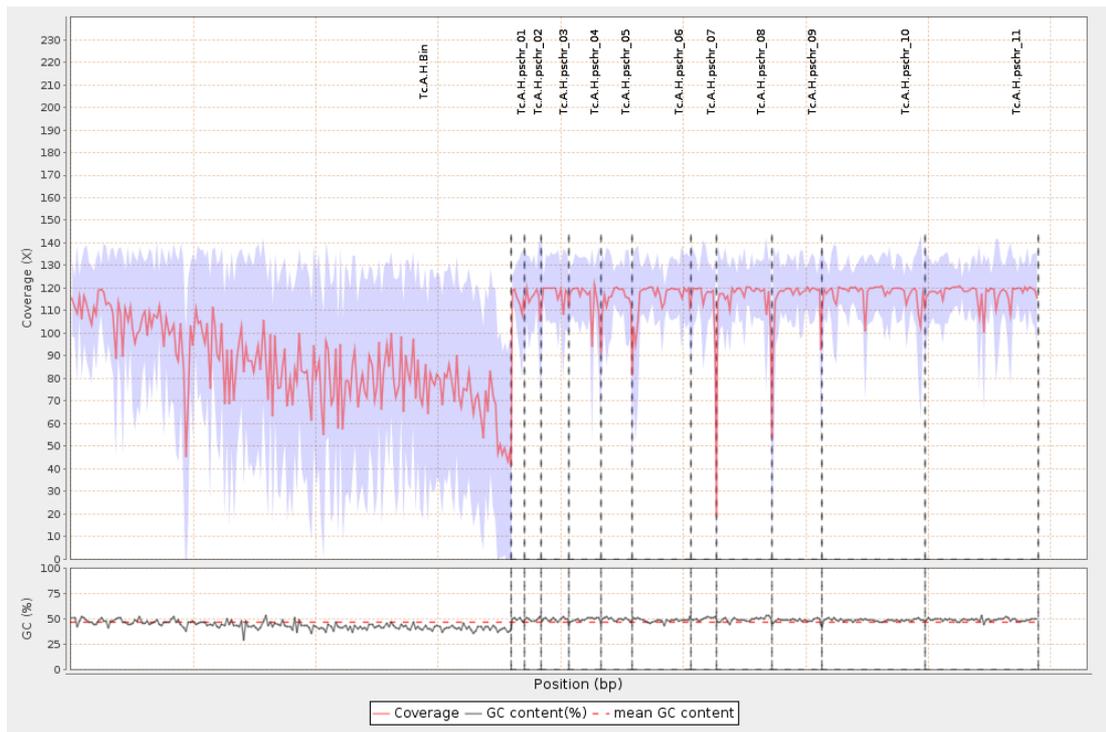
As mentioned earlier the assembly has an overall coverage of 47 fold, and it has been shown that HGAP assembler with a genome coverage less than 50 times could generate higher error rate (Chin et al., 2013; Lee, Gurtowski and Yoo, 2014).

#### **2.3.5.1 Illumina sequencing and data generation of TcIL3000 gDNA**

About 23 Gb of data were generated, with a final total number of reads of 197 million with estimated trimmed mean read length of 121 bp. Those data were utilized for error correction and gap filing of the final *T. congolense* PacBio assembly.

#### **2.3.5.2 Mapping Illumina reads to the final pseudo-chromosome level PB assembly of TcIL3000**

The mean coverage of starting iteration was 103 fold across the PacBio assembly. Interestingly, the chromosomes showed higher and more stable coverage rate 116 comparably to the Bin (Figure 2.7). The relative stable and higher coverage across chromosomes might lie in the diploid nature of the core region of MBCs, while the reduced coverage trend across BIN contigs could be due to the nature of allocated sequences, which mostly showed subtelomeric class (heterochromatin) or mini-chromosomes, which in turn are thought to be aneuploid (Melville, Gerrard and Blackwell, 1999; Wickstead, Ersfeld and Gull, 2004; Berriman, 2005).



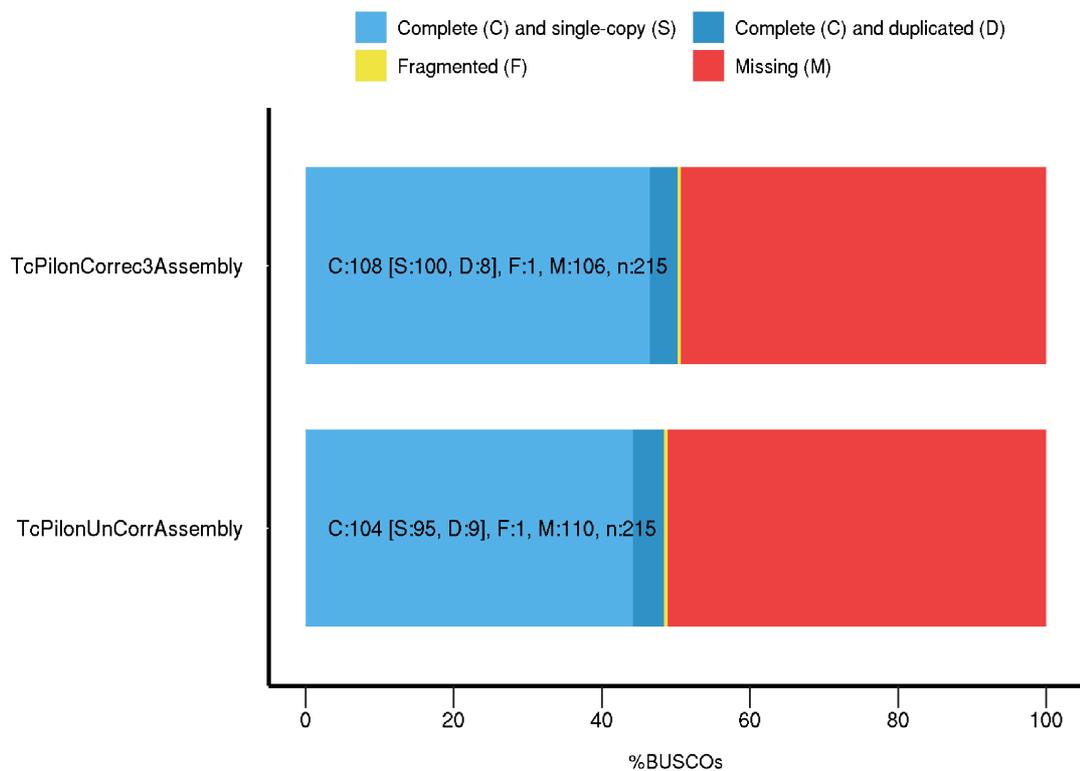
**Figure 2.7** Illumina reads coverage used for error correction and gap filling across final pseudochromosomal level *TcIL3000* PB assembly. Higher and relatively stable coverage is noticeable across 11 chromosomes (right side panel) in comparison to the lower oscillating levels of coverage across the bin (left side panel).

### 2.3.5.3 Error correction and gap filling using PILON

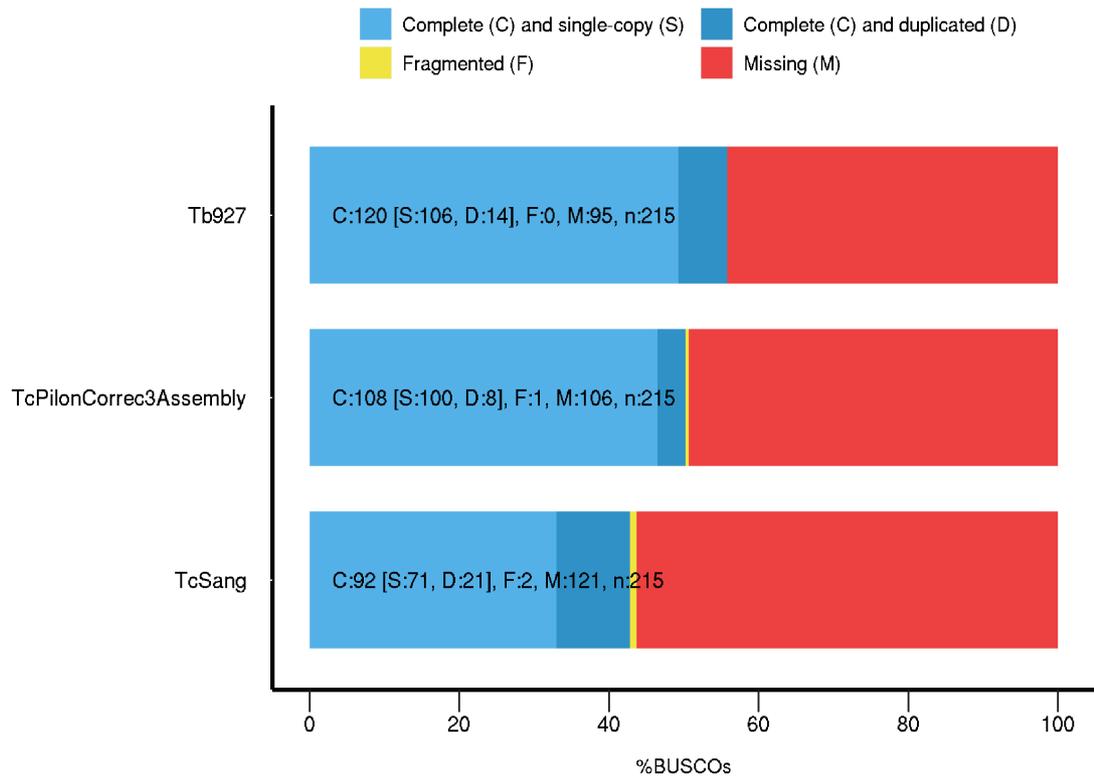
Three consecutive iterations were achieved on the final PB assembly in order to get high per base quality and filling many gaps in the assembly based on paired end reads evidence. 84 gaps were completely closed and more than a total of 160.7 kb of estimated ambiguous nucleotides were resolved.

Gene model integrity was assessed in the final corrected version (third iteration of PILON run) compared to the initial uncorrected input using BUSCO comparison tool. Improved scores were noticeable in the upgraded version as it presented more completed gene models and less missing models (Figure 2.8). This final upgraded assembly was then decided to be the subject of the gene annotation step.

Furthermore, comparisons showed improvements in the annotation of the tested core gene set by unravelling possible missed or incomplete eukaryotic core genes and a reduced number of duplicated models in comparison to the Sanger assembly (Figure 2.9). This result emphasizes the degree of completeness of the gene models and suggesting physical integrity of the PacBio assembly in comparison to the current draft Sanger assembly of TcIL3000.

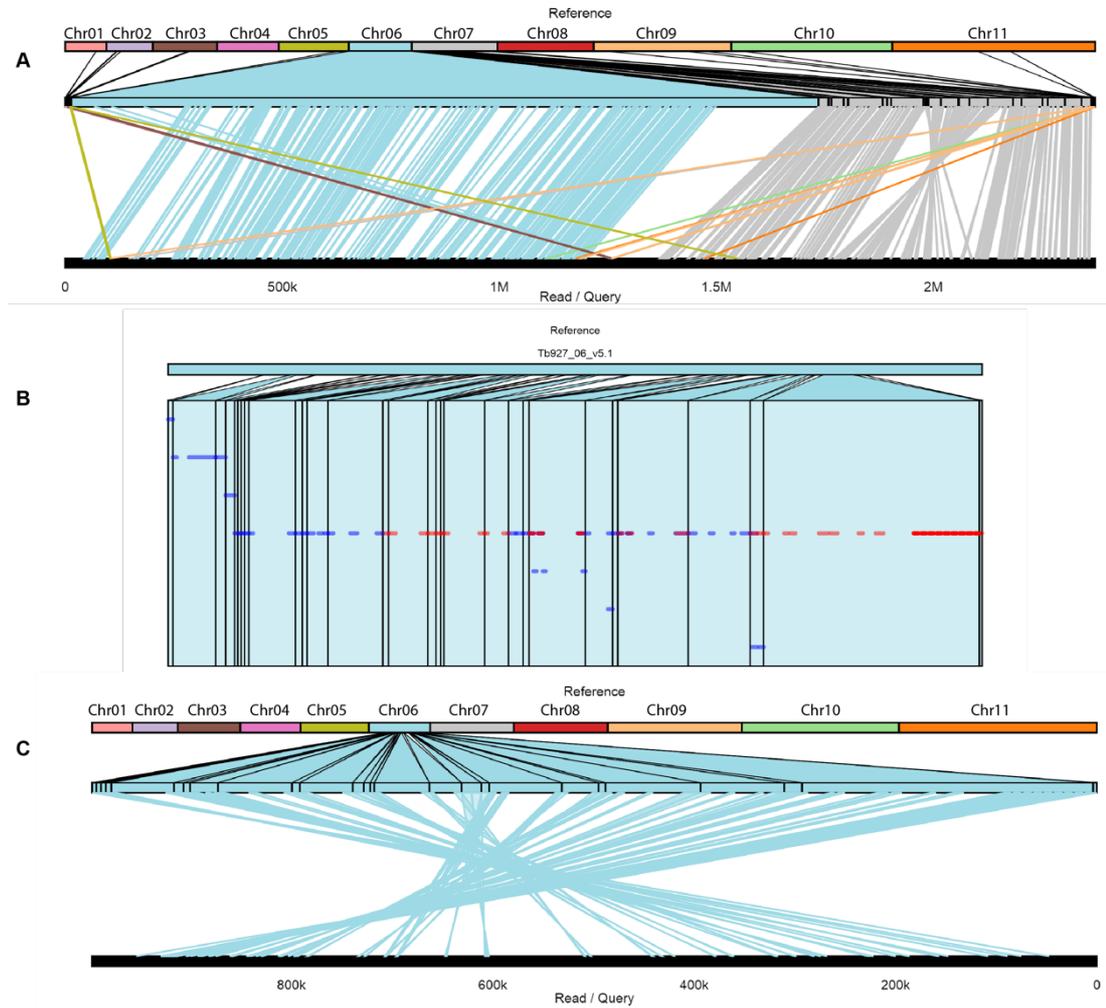


**Figure 2.8 BUSCO comparison of chromosomal level *T. congolense* PacBio assembly before and after error correction and gap filling steps.** The third iteration of consecutive base error correction and gap filling using Pilon (TcPilonCorr3Assembly) compared to the raw initial assembly (TcPilonUncorrAssembly). Default parameters were applied to generate this plot.



**Figure 2.9 BUSCO tools test of chromosomal level assemblies of *Tb 927*, *T. congolense* Sanger and *T. congolense* PacBio.** Improvement in gene models of *T. congolense* PacBio (TcPilonCorrec3Assembly) when compared to current Sanger assembly (TcSang) of the *T. congolense* genome using protists gene set and showing the result of tests for model genome assembly of *T. brucei* TREU927.

### 2.3.6 Possible Caveats to pseudochromosomes' level assembly



**Figure 2.10** *T. congolense* pseudochromosome 6 inferred by ABACAS1 using genome synteny according to reference genome Tb927. Highlighting possible erroneous placements of a contig with synteny to Tb927 chr 7. A) The top multicolour bar represents the reference Tb927 pseudochromosomes, while the bottom thick black line refers to *T. congolense* PacBio chr06, which has a region with synteny to Tb927 chr07 (grey ribbons). B) *T. congolense* PacBio assembly contigs that aligned to the reference chr6. The contig (intermittent blue and red line) is the longest contig >800 kb and it's the last contig assigned entirely to Tb927 chr06. C) The longest and the last contig (bottom black bar) in B showing its relation to Tb927 chr06 (cyan blue ribbons). PROMmer algorithm was adopted to infer synteny to the reference sequence. Plots generated by GenomeRibbon.

The largest scaffolds/contigs of the assembly were allocated to the *T. congolense* inferred 11 MBCs. This is because they showed broad synteny to the corresponding *T. brucei* 927 MBCs, the core region of these chromosomes mainly encodes for housekeeping genes.

Although this approach resulted in relatively good results for the chromosomal internal regions, this approach may result in dubious assignments of contigs especially towards the ends of the chromosomes. Some of these could be due to the low synteny with reference in these more variable regions. Accordingly, these possible problems have been highlighted and classified as follows:

I- Anonymous allocations:

This possible type of dislocation affected a few contigs, which was especially noticed at the end of the pseudochromosomes.

For example, pseudochromosomes 6 where a long contig 700 kb shows synteny revealing its relation to Tb927 chr7 as it been verified and visualized using GenomeRibbon tool (Figure 2.10). Besides that, aligned PacBio reads to this scaffold assembly failed to show connecting reads between the contigs, which is assigned to Tb927 chr6. Furthermore, the previous contig end with a stretch of telomeric repeats, which was thought to be the end of the putative chr6 in *T. congolense* PacBio. Although the reason is unclear to this doubtful placement, this contig has a cluster of hypothetical genes at the 3' end that highly similar to another set at the 5' end of a contig assigned to the 5' end of *T. congolense* PacBio chr7. Therefore, it is possibly a wrong placement resulting from these conflicting sequences or a failure in the locating of this contig to the putative MBC7 by splitting perl script due to this sequence conflict.

Another event was noticed at the 5' end of *T. congolense* PacBio putative chr10. A contig of 186 kb, could also be applied to chr11, as there is no linking read support to the putative *T. congolense* MBC 10 and the putative open reading frame are syntenic with corresponding sequences on the reference MBC 11.

## II- Allocation of false positive syntenic contigs:

Another potential faulty placement of contigs to the MBC that are thought to be partial or complete mini-chromosomes were assigned to the subtelomeric regions of a number of MBCs. This is probably due to the nature of these MCs harbouring subtelomeric features (see chapter 3). Briefly, these contigs have subtelomeric genes (VSG, ESAGs, DEAH/D box RNA helicase) and a long central repeat that spans for more than 10 kb, with possible high similarity to VSG sequences. Therefore, perhaps these contigs are potential false candidates (false positive) for some subtelomeric regions on MBCs (6, 9 and 11).

Although a few proposed miss-assemblies were noticed, the majority were accurate, and this did not have a major effect on the downstream structural analyses as contigs were considered and sequence physical contiguity ensured by mapped overlapped PacBio reads, ensuring physical integrity. The possible structural variations were assessed according mainly to the contigs or in a few cases to the PacBio reads linked scaffolds.

### **2.3.7 Genome annotation of *T. congolense* chromosomal level assembly**

Assigning putative protein coding and non-coding genes sequences is a crucial step in genome projects in order to permit further in depth comparative genomics and to reveal important biological aspects of the organism under study. However, this process is highly complicated and quite laborious. In order to simplify this step in genome projects, a number of automated genome annotation pipelines were developed.

#### **2.3.7.1 Automatic annotation pipeline**

The COMPANION automated annotation pipeline (see section 2.2.8.1) was developed at Sanger Institute specifically for the annotation of protozoan genome projects.

This approach annotated 8,292 putative genes, 2028 putative pseudogenes, 142 *rRNA*, 71 *tRNA*, 3 *snRNA* and 50 *snoRNA* genes (Table 2.5). The *snRNA* and *snoRNA* non-coding genes predicted by COMPANION are currently

absent from *TcIL3000* and more *rRNA* and *tRNA* genes were reported. These non-coding genome features are more likely to contain nucleotide content of a repetitive nature and they tend to be located in repetitive tandem arrays (Dunbar *et al.*, 2000; Liang *et al.*, 2005; Bermudez-Santana *et al.*, 2010). They were found in such clusters in the PacBio assembly.

However, further examination of our *T. congolense* PacBio assembly annotation showed possible open reading frames that were not assigned as potential gene or pseudogene models, so, manual intervention was inevitable to annotate these potential open reading frames.

#### **2.3.7.2 Manual annotation of *T. congolense* PacBio genome assembly**

The manual annotation for predicting possible gene models in PacBio assembly based on protein evidence (see section 2.2.8.2), resulted in adding a total of 1,009 putative protein-coding genes/pseudogenes to the final annotation file (985 potential protein coding genes and 24 were considered pseudogenes). The predicted protein coding and other features of the final *T. congolense* strain IL3000 PacBio assembly are shown (Figure 2.11).

#### **2.3.8 Comparison of the annotation between the two assemblies of *T. congolense* IL3000 and the reference *T. brucei* TREU927**

In general, the number of predicted protein coding genes in the PacBio assembly was less than available in the Sanger version in the current database by approximately, 3,100 (Table 2.2). So, to investigate this difference, assembly and gene model integrity checking was adopted.

Firstly, the available draft Sanger based assembly of *T. congolense* is highly fragmented as evidenced by manual sequence inspection, the gene model integrity testing (Figure 2.9) and assembly statistics. Analysis shows that the higher number of putative gene models in the draft Sanger assembly is partly due to the fragmented genome, which in turn could cause inflation in the total number of suggested gene models, as some of these models could be distributed over different contigs. Such a pattern is evidenced by the detection of more possible duplicated gene models in the Sanger assembly (Figure 2.9).

Analysis of predicted gene lengths between the two assemblies shows that the mean length of *T. congolense* PacBio features was 1,362 bp with a minimum length of 70 bp, while that of *T. congolense* Sanger was (1,338 bp, 35 bp), respectively. Statistical analysis (*t.test*) to compare between means of the two groups revealed a significant difference under confidence interval of 95%, p-value  $< 2.2e^{-16}$ , df = 23,052. Such a significant difference in gene model length gives support to this interpretation between the two assemblies.

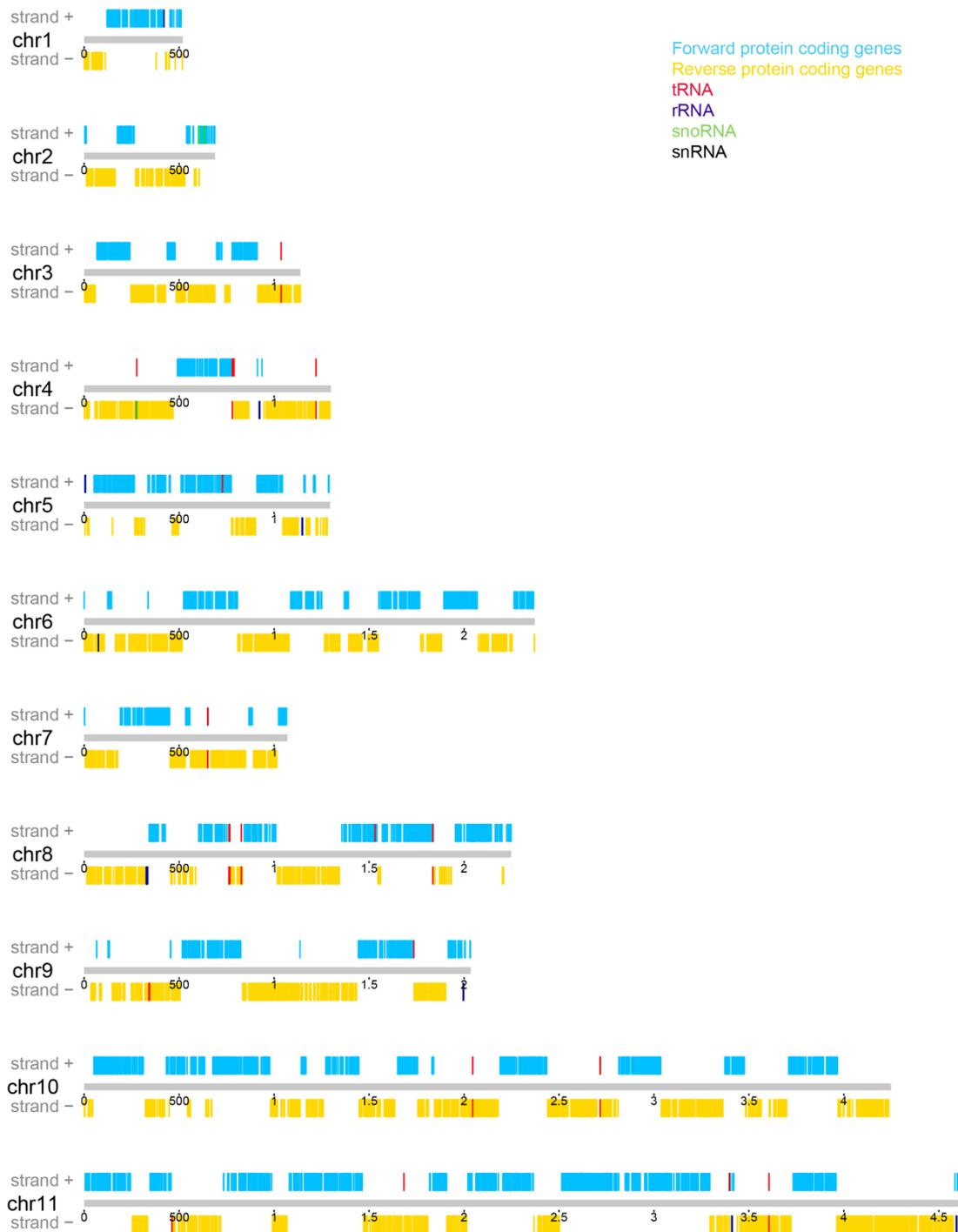
In addition, the number of the annotated fragmented gene models in the Sanger assembly is high (1,267), in comparison to more complete gene models in the PacBio assembly (Table 2.2).

Secondly, the suggested pseudogene models by the annotation pipeline in the PacBio assembly were more than those in the Sanger reference by 1,700, this reduced the difference between the two assemblies to 1400 genes. However, the relatively high pseudogenes predicted in the *T. congolense* PacBio assembly could be either an over estimation of these models suggested by the COMPANION pipeline or an under estimation of these models in the Sanger sequence or perhaps by both. To date, the absence of sequence data of all life stages of this parasite and the fact that the genes are located in core regions and are polycistronic expressed, make it difficult to judge the completeness of the gene models. Furthermore, close inspection to the predicted pseudogenes by the automatic pipeline showed that some of these models are in fact intact gene models, but the pipeline was not able to extend such models to the first methionine or to the next stop codon or both. Such miss-annotation could be fixed manually; however, this approach would be laborious and time consuming when we consider inspecting more than 2,000 pseudogenic features. Unfortunately, I couldn't do this step due to time constraints. Most importantly, the pipeline does transfer protein domains and assign functional annotation to such models; therefore, it would not affect downstream comparative genomic analyses.

Finally, the annotation pipeline with default parameters could also be under estimating the number of coding genes in our version in comparison to the manually curated Sanger assembly, as the AUGUSTUS threshold was set to the default cut-off (0.8), which might prevent many predicted gene models from passing such a stringent threshold. Furthermore, some possible amputated ORFs at the ends of some contigs noticed by manual inspection would not have passed the pipeline filters. The break of such contigs could be due to the repetitive nature of some genes like VSGs and *Ingi* (Berriman, 2005) or in a region of segmental duplication (Jackson *et al.*, 2010; Treangen and Salzberg, 2012; Nattestad, Chin and Schatz, 2016).

**Table 2.5 Genome features of currently available draft Sanger assembly, PacBio assembly of *T. congolense* IL3000 and the reference *T. brucei* TREU 927.**

	Sanger assembly	PacBio assembly	<i>T. brucei</i>
<b>Protein coding genes</b>	11,462	9,233	10, 159
<b>pseudogenes</b>	330	2,046	1, 444
<b>Fragmented genes</b>	1,271	2	488
<b>rRNA</b>	55	142	105
<b>tRNA</b>	59	71	66
<b>snRNA</b>	0	3	6
<b>snoRNA</b>	0	50	320



**Figure 2.11** The predicted features annotated to the *T. congolense* PacBio MBCs. Annotated genes on MBCs were showed according to the colour key. The numbers underneath is the length scale in 500 bp increment, numbers equal or larger than 1 are in Mb length. KaryoploteR package on R-project was used to generate this plot according to the feature coordinates stored in the GFF3 file.

### **2.3.9 The new findings in this genome sequencing project**

The contiguated genome assembly encouraged us to investigate different aspects of *T. congolense* features such as the possibility of finding new putative protein coding genes and other non-coding DNA sequences, such as simple tandem repeats of the centromeres of eukaryotic chromosomes, strand switch regions and directional gene clusters, which are not presented well in the current Sanger assembly. Moreover, to search for the possible interspecies structural chromosomal variations.

#### **2.3.9.1 Putative centromere repeats in MBCs *T. congolense* PacBio assembly**

The PacBio data was searched for putative centromere repeats. These repeats are conserved within eukaryotic chromosomes. However, such repeats are species specific in terms of actual DNA sequence, but they share global characteristics. They are often found as short simple repeats flanking special genomic features. The inference of this type of repeats in the *T. congolense* PacBio MBCs was made by comparing the synteny of these regions with the closely related species *T. brucei* (see methodology section 2.2.14). In the reference species, it consists of AT rich tandem repeat units flanked by specific genes such as rRNA or retrotransposons. Potential centromere repeat sequences were identified in five MBCs in our PB assembly (Table 2.6).

**Table 2.6 Size and coordinates of possible centromere tandem repeats in *T. congolense* chromosomes 1, 3, 4, 6 and 11.**

Chromosome no.	Size of putative Centromeres (kb)	Coordinates*
1	2.2	415,704 : 419,834
3	13.0	725,163 : 738,239
4	1.0	1,237,876 : 1,238,839
6	14.4	106,668 : 121,101
11	15.0	2,174,797 : 2,196,244

\* Coordinates were identified by Repeat Detector (Red) software prediction viewed by ARTEMIS.

### 2.3.9.1.1 Pseudochromosome one

A region located on *T. congolense* pseudochromosome one matched corresponding features of the *T. brucei* TRUE927 chromosome 1 centromere. This segment consisted of repetitive DNA 2 kb in length, contains tandem repeat units of 136 bp long characterized by a low GC content (32%). These features are only present once on the chromosome and flanked by transposable elements stretched along an expanse of 28 kb upstream and an *rRNA* gene downstream (Figure 2.12. A). However, this region was interrupted by a gap, suggesting the presence of a longer centromere sequence than the assembled one. In addition, the BLASTn search of this sequence also showed hits to contigs in the Bin sequence of the assembly. The presence of flanking retroelements have also been reported in eukaryotic centromere repeats (Schueler *et al.*, 2001). The GC content of the corresponding *T. b927* centromere is 33.5% in comparison to 32% of *T. congolense*. The BLAST search did not show sequence similarity in *T. congolense* to the Tb927

centromere sequences. Furthermore, in contrast to *T. congolense* centromere on chromosome one, the equivalent *T. brucei* centromere lacks the consistency of the tandem repeat units, which might suggest inter-species differences regarding this type of repeats.

#### **2.3.9.1.2 Pseudochromosome three**

The inferred centromere sequence in this *T. congolense* PacBio MBC covers 13 kb (GC 28%) located at the beginning of a 214 kb contig (CTG000377) (Figure 2.12. B). A cluster of putative dynamin genes is located upstream to this region, however, it is separated by a physical gap, while in a downstream position within the same contig of these repeats there are genes encoding for surface proteins. Although this repeat is similar to that of chromosome one, it exhibited two repeat units 136 and 408 bp.

The *T. brucei* centromere on the corresponding chromosome presented comparatively higher GC percent (48.5%) and different repeat period (120 bp). These repeat statistics are not only differing from *T. congolense* predicted centromere but also differ from *T. brucei* MBC one. Remarkably, the annotated genes upstream to the centromere were aspartyle aminopeptidase and kinesin, while the downstream showed a long stretch of *rRNA* sub units encoding genes, suggesting different localization of the centromere on this chromosome between the two species (centromeric chromosomal rearrangements). This proposed translocation of *rRNA* genes is not uncommon in eukaryotic genomes as these sequences known to be changing their position, number and distribution even between different strains of the same species due to its repetitive nature, co-localization with satellite repeats and the presence of interspersed retroelements in the centromeres (Schubert and Rieger, 1985; Dubcovsky and Dvorak, 1995).

#### **2.3.9.1.3 Pseudochromosome six**

A 14.4 kb stretch of 136 bp putative centromere repeat with high sequence similarity to those on MBC 1 and 3 on pseudochromosome 6 was identified. However, it showed lower GC content (GC 27%). Upstream, a conserved hypothetical gene and two copies of RNA-dependent DNA-polymerase were

found, while downstream, three copies of conserved hypothetical proteins, and a cluster of predicted genes encoding for cell surface protein cysteine peptidase (Figure 2.12.C). A Similar region on *T. brucei* MBC 6 presented an inconsistent pattern of tandem repeat units that revealed relatively higher GC content (32.4%), which localized between a cluster of *rRNA* on 5' flanking region and two genes encoding for surface proteins (Adenylate cyclase, GRESAG4).

#### **2.3.9.1.4 Pseudochromosome 11**

Putative tandem repeats found on a 2 kb repeat sequence located at the 3' end of a 1.5Mb contig (CTG000474) and a 12 kb region on the 5' end of a 128 kb following contig (CTG000476), this repeat sequence also showed low GC (GC 32%) and different component repeat units of 247 bp, in between those two contigs, a contig of length 5.8 kb containing retrotransposable elements (*ingi*) was noticed and showed different GC content.

The proposed centromere sequences on this MBC did not show sequence similarity to the previously mentioned centromeres of *T. congolense*, suggesting potential differences in basic centromere sequence between *T. congolense* MBCs. Such differences in centromere repeat sequences among different MBCs was also noticed in *T. brucei* (Obado *et al.*, 2007). However, the biological significance of such differences has not been revealed yet.

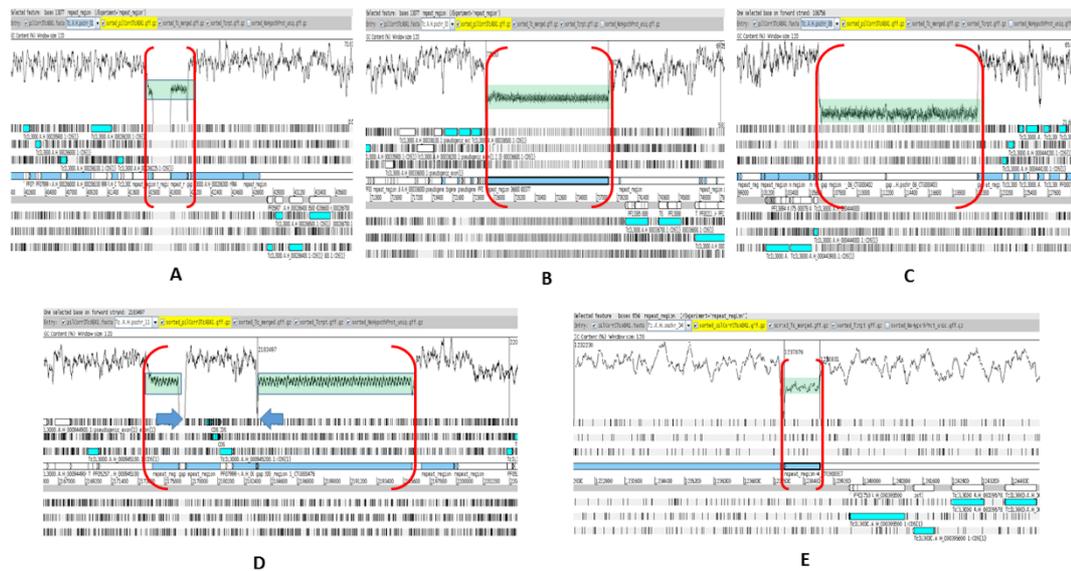
The proposed region was preceded by genes with hypothetical protein product and housekeeping genes, whilst downstream to it there are genes with unidentified protein domains and DEAD box RNA helicase (Figure 2.12.D).

#### 2.3.9.1.5 Pseudochromosome four

Tandem repeats of DNA sequences stretched over a 956 bp region was also noticed on the 5' end of a 35,278 kb long contig (CTG000387) assigned to *T. congolense* MBC 4. This region showed sequence similarity evidenced by BLASTn search, and the GC content (32%) and repeat unit size was similar to the putative centromere on MBC11. This region is followed by predicted genes encodes for proteins related to DNA replication, conserved proteins with unknown function and amino acid transporters (Figure 2.12.E). A similar region on MBC 4 of *T. brucei* flanked by genes encode for surface protein GRESAG4 downstream and a gene encodes for flagellar attachment zone upstream.

In general, the proposed centromere sequences presented in this work were generally syntenic with those of *T. brucei* by the fact of some *rRNA* genes are flanking these repeats on MBCs (1, 2, 3, 6 and 7); Although only chromosome one and six exhibited similar synteny, the BLASTn search to *T. congolense* PacBio assembly identified some contigs with similar sequences in the BIN which also showed *rRNA* flanking sequences, suggesting consistency in the synteny (Obado *et al.*, 2007). Furthermore, presence of retroelements flanking these repeats were found in both *T. brucei* and *T. cruzi* (Akiyoshi and Gull, 2013) and the presence of genes encoding for surface proteins flanking these regions in both African trypanosomes (our independent analysis); however, this seems to be more likely to occur in *T. congolense*, suggesting a general synteny of the centromeres across trypanosomes.

The predicted centromeres here also showed a general trend of AT- rich sequences; a criterion shared among all studied centromeres in eukaryotes (Talbert, Bayes and Henikoff, 2009) and more specifically in *T. brucei*, however, the only exception was those of *T. cruzi*, which further lack tandem repeat merit (Obado *et al.*, 2005).



**Figure 2.12 Putative centromere repeats in *T. congolense* PacBio sequence assembly chromosomes1, chromosome3, chromosome6, chromosome11 and chromosome4 (denoted by red brackets) characterized by low GC content (green shaded) and consist of 136 bp and 247 bp (for the Chr4 and Chr11) in length of single repeat unit. A) a region of 2 kb with GC content of 32%, located between a 28 kb segment of transposable elements up stream to it and an *rRNA* gene downstream. B) Similar region was detected on pseudo-chromosome 3 by BLASTn megablast search with e-value  $1e^{-10}$  covering a region of 13 kb with GC content of 28%, preceded by an immediate gap and a cluster of dynamin genes followed by genes encodes for surface proteins. C) another region of proposed centromere repeats found on pseudo-chromosome 6 extends over 14.4 kb (GC 27% and has similar repeat blocks) in strand switch region positioned between RNA-dependent DNA-polymerase genes on reverse strand and a cluster of hypothetical conserved genes, retroelements and surface protein genes. D) AT rich tandem repeats units of 247 bp on Chromosome 11, interrupted with a contig (contig CTG000475) that has retrotransposons insertion sites (thick blue arrows), showed GC content 32%. E) BLASTn megablast search (e-value  $1e^{-10}$ ) of previous region to the PacBio assembly showed similar DNA pattern of 957 bp in length on pseudo-chromosome 4, exhibiting a low GC percentage (32%).**

Our analysis of these putative centromeres has also highlighted that the sequence similarity and repeat unit size of these regions is not similar among all *T. congolense* MBCs, but is rather between subsets of the MBC. These subsets were MBCs (1, 3 and 6) and MBCs (4 and 11). Similarly, *T. brucei* centromeres supported such findings as the repeat size and sequence of centromere on MBC 3 is different to those on MBCs (4, 5, 8, 9, 10 and 11) (Obado *et al.*, 2007).

To our knowledge, although there is no experimental work done on such sequences in this trypanosome, our analyses of these repeat sequences match the general criteria of being centromeres of eukaryotes and more specifically of trypanosomes due to, synteny, with the close relative *T. brucei* centromeres, make these sequences more likely to be the centromeres of *T. congolense*. This genomic territory has not been revealed before in this organism.

### **2.3.9.2 Putative chromosomal rearrangements in *T. congolense* PacBio MBCs in comparison to the Tb927 MBCs**

An advantage of an assembly containing long contigs is that it has the potential to reveal putative inter and/or intra-chromosomal rearrangements. The longer segments of chromosomes provide more physical integrity and then structural variants could be discovered in the genome under study in comparison to close reference species. Here the focus is on the large inter or intra-chromosomal reorganizations revealed by *T. congolense* PacBio MBCs assembly in comparison to reference *T. brucei* TREU927 MBCs.

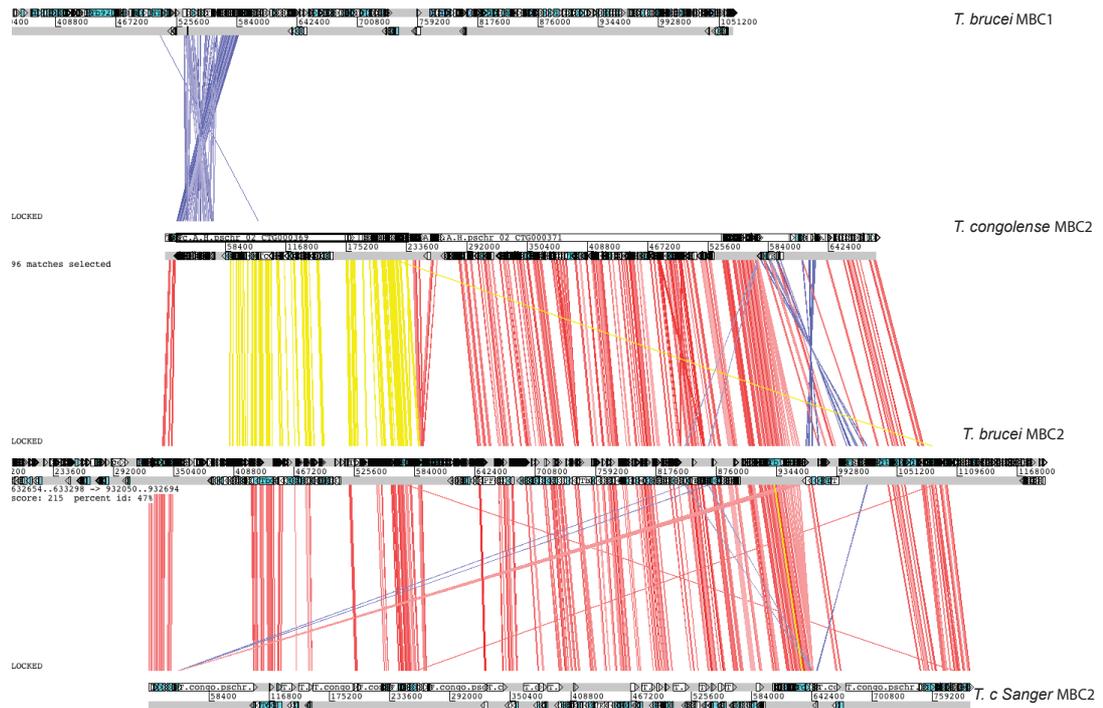
#### **2.3.9.2.1 Chromosomal rearrangements in *T. congolense* PacBio pseudochromosome two**

A contig with the identifier of CTG000369 of >223 kb in length assigned to the predicted MBC two of *T. congolense* PacBio assembly bears putative coding genes that span a region of 47.7 kb on its 5' end, those genes are mostly part of Calpain like group CA members which have a pathological role. Members of alpha and beta tubulin gene family were also detected which have been translocated and inverted in comparison to Tb927 MBC one (Figure 2.13). This

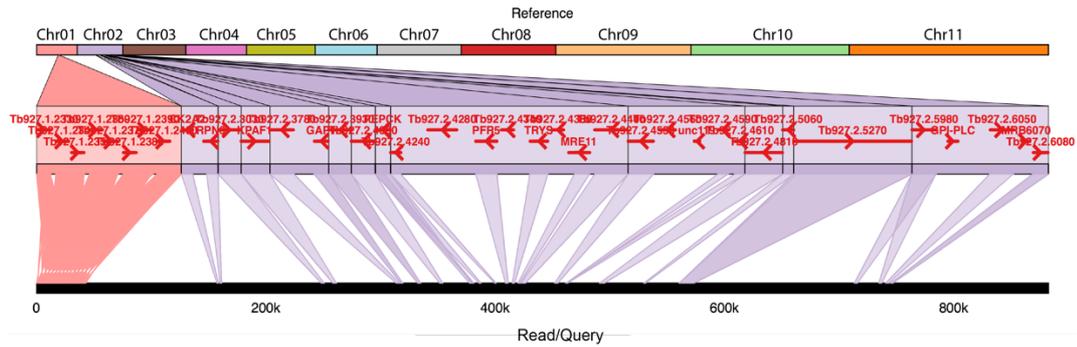
translocated part of the chromosome is not revealed by current Sanger assembly and represented by a big gap in the proposed region of relocation. In order to support our findings, PacBio assembly contigs of strain Tc1/148 were searched for similar exchange to check the proposed scenarios, which showed consistent translocation (Figure 2.14).

An intra-chromosomal rearrangement noticed in chromosome two affected a region of 66 kb towards the 3' end within this area; members of the surface protein family during blood form infection stage in *Tb927* were affected by an inversion in *T. congolense* PacBio. This region is represented by gaps in the current reference of TcIL3000, with one line representing an inverted sequence barely noticed preceding a gap (Figure 2.13).

The rearrangement proposed here might have a significant impact on morphology or pathogenesis between the two trypanosomes, as it affected genes responsible for production of proteins participating in cell morphology structural changes (Hayes *et al.*, 2014) and other members of calpain cysteine peptidase are involved in pathogenicity of trypanosomatids (Branquinha *et al.*, 2013). Furthermore, it could affect cell motility and ciliary activity governed by tubulin genes (Hammond, Cai and Verhey, 2008).



**Figure 2.13** An ACT comparison plot of proposed chromosomal rearrangements between Tb927 chromosome one (top bar), *T. congolense* PacBio chr2 (middle bar). tBLASTx search for sequence similarity with expectation value of  $1e^{-10}$  was applied between each two pairs of chromosomes. The blue crossed lines in the top panel refer to the translocation and inversion in this region in our PacBio assembly; most of the genes affected by this inter chromosomal exchange encode for pathogenic factors (see the text) and are located on one (223 kb) contig CTG000369. Yellow lines refer to the genes on the same contig of *T. congolense* PacBio chr2 that share sequence similarity with *T. brucei* counterparts, which are not affected by the translocation. *T. congolense* Sanger MBC 2 sequence failed to show such an event as a big gap covered this area (bottom panel). The crossed blue lines on the right side of the middle panel refer to intrachromosomal inversion affecting predicted genes on a 66 kb region on *T. congolense* PacBio chr2, known to encode for surface proteins during blood form infection stage in the Tb927 corresponding sequence.



**Figure 2.14 A Contig has the identifier (scf180000002797) of strain Tc1/148 PacBio assembly showing similar putative rearrangement to that of strain IL3000. The Tc1/148 (scf180000002797) (bottom black bar) assigned to *T. brucei* MBC two by PROMmer algorithm of MUMmer package showed a similar region affected by proposed inter chromosomal rearrangement and revealed synteny to a region of MBC one of *T. brucei* sequence assigned to putative *T. congolense* MBC two.**

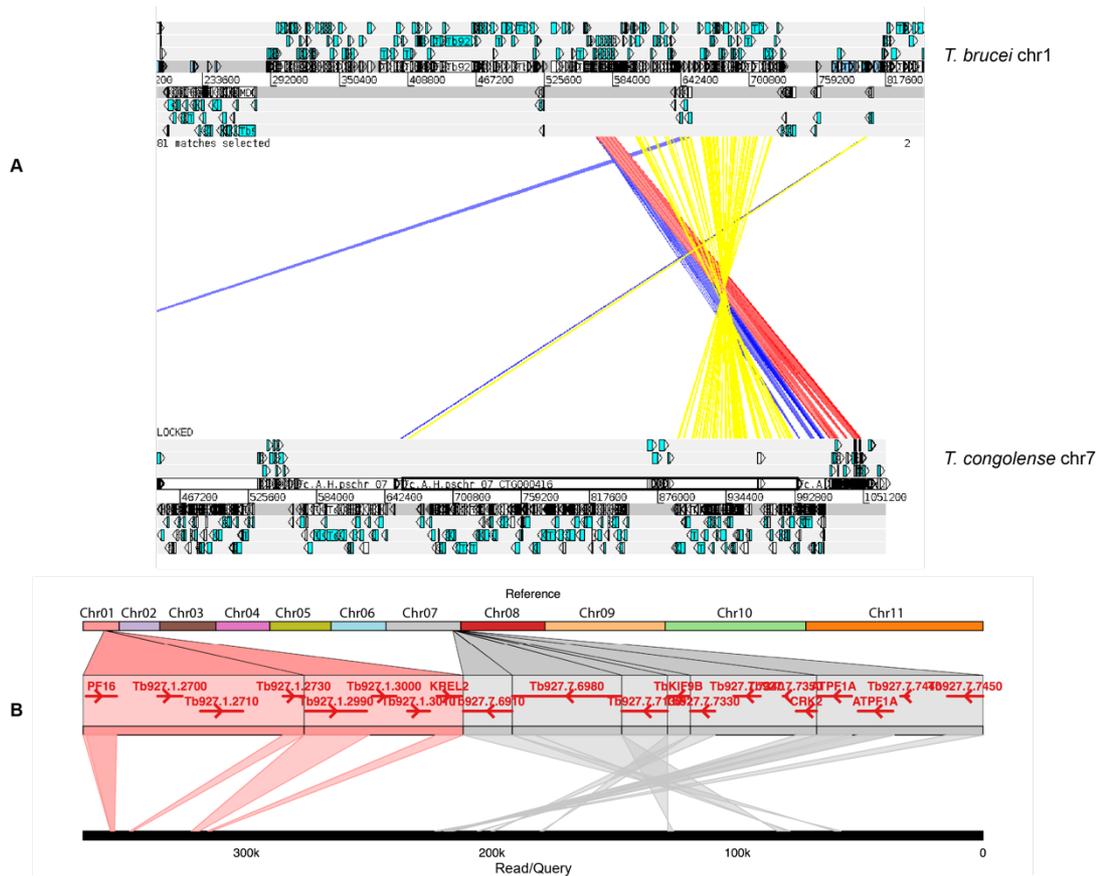
### 2.3.9.2.2 Chromosomal rearrangements in *T. congolense* PacBio MBC 7

Our data revealed a potential displacement on chromosome seven, which acquired a 158 kb segment at the 3' end of Tb927 chr1. The majority of this region is located on contig CTG000416 larger than 336,9 kb, starting with *tRNA* genes and followed by contig of length 23 kb with *alpha*, *beta tubulin* predicted genes and transposable elements (Figure 2.15. A). Tc1/148 also supports this proposed chromosomal structural variation (Figure 2.15. B).

The disrupted region on the bigger contig CTG000416 was inverted and flipped and the following contig CTG000417 has the same trend, while the last set of rearranged genes were in frame to those corresponding features in the Tb927chr1 and encode for *alpha* and *beta tubulin* (Figure 2.15. A).

The presence of this putative translocated region between a cluster of *tRNA* genes and transposons in two strains of *T. congolense* suggests that such genomic changes are in fact real. In other eukaryotes like yeasts, *tRNA* genes were subjected to genomic rearrangements as they represent a cluster of identical sequences in different chromosomes such that they play a role of

exact repeat, which might enhance homologous recombination (Thompson *et al.*, 2003; Noma *et al.*, 2006). Moreover, the break in synteny among trypanosomatids genomes like *T. brucei*, *T. cruzi* and *Leishmania* was also found associated with *tRNA* genes (El-Sayed *et al.*, 2005). Furthermore, the repetitive nature and genome mobilization ability of retroelements were also responsible for large chromosomal rearrangements like translocations and inversions (Han *et al.*, 2009; Weil, 2009).



**Figure 2.15 Chromosomal displacement of *T. congolense* PacBio chr7 and Tb927 chr1 affected mainly predicted genes encoding for alpha and beta tubulin on Tb927 chromosome one.** A) Reoriented genes (Yellow lines) on 3' end of 339 kb TcIL3000 PacBio contig (CTG000416) (lower white arrow denoted by black edges) preceded by *tRNA* genes and followed by a 23 kb contig (CTG000417) with transposable elements, then a set of genes in frame with a corresponding set on Tb927 chr1. Viewed using ACT visualization tool based on tBLASTx search tool. B) Rearrangement affecting similar region on the 3' end of Tc1/148 PacBio contig (Scf7180000002796) over 300 kb in length, viewed using GENOMWRIBBON tool based on the PROMmer algorithm sequence similarity search tool.

#### 2.3.9.2.3 Chromosomal rearrangements in *T. congolense* pseudochromosomes 8

A number of intrachromosomal rearrangements were found (**Error! Reference source not found.**). First, an inversion occurred in genes predicted to have putative function of S-adenosyl-L-methionine-dependent methyltransferase, serine esterase (DUF676), 5S ribosome-binding GTPase and 5'-3' exoribonuclease C. This region is flanked upstream by *tRNA* genes in both species. The Sanger assembly, however, failed to show this rearrangement.

Second, a minor dislocation event affected a gene encoding for protein of unknown function. Both previous rearrangements are on contig CTG000430 with length of 370 kb. The third inversion targeted predicted genes for DNA polymerase I, a flagellar component called DIGIT, hypothetical protein and Flagellar Member 3. The latest gene inversion was also shown by the *T. congolense* Sanger sequence.

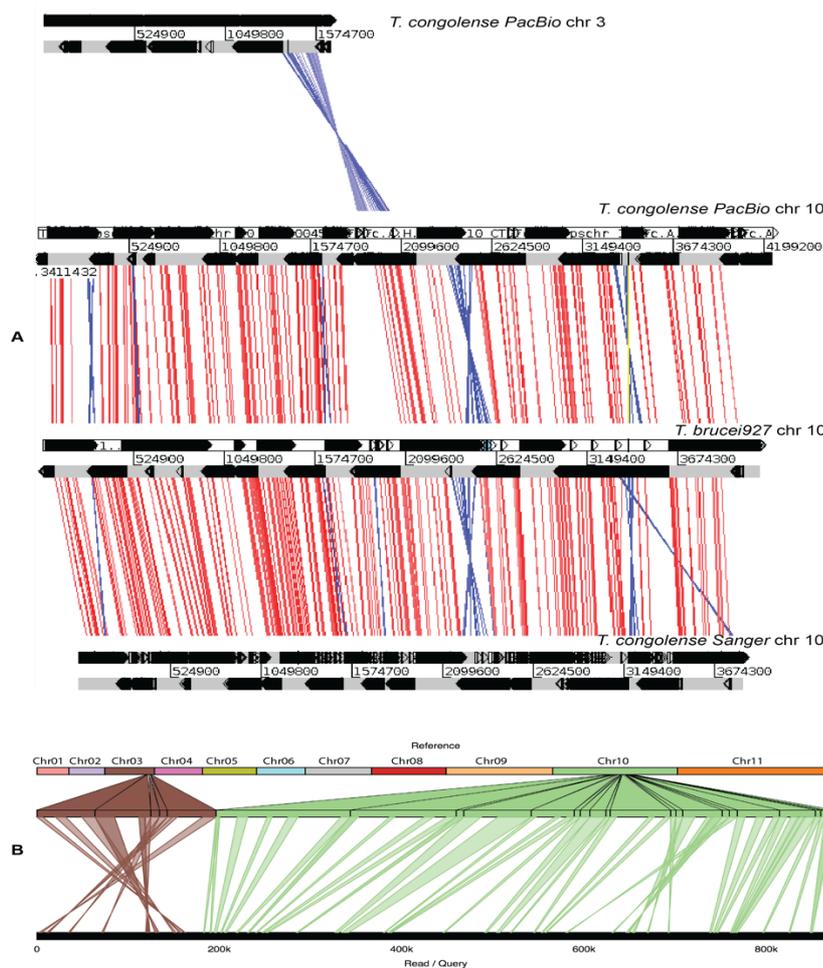
This proposed rearrangement affected genes that encode mainly for structural components, which might also affect phenotypic differences between the two trypanosomes. Presence of *tRNAs* sequences could be the possible origin of such inversions internally to this MBC between the two species.

#### 2.3.9.2.4 Sequence relocations in *T. congolense* PacBio chr 10

A large segment (276 kb) of core genes on the 3' end of *T. brucei* chr 3 was putatively identified on a 873.9 kb long contig CTG000463 allocated to the putative *T. congolense* PacBio chr 10. This displaced, flipped and inverted part of DNA has many predicted housekeeping genes encoding for different proteins with various functions. A large proportion of this region is located on the 5' side at an 874 bp (CTG000463) mainly consist of genes belong to chr10 and it was assigned to chromosome 10, the displaced genes on this contig separated from region of chr10 in both species by a 6.8 kb spacer DNA (a Strand Switch Region) characterized by the presence of three *tRNA* predicted genes. The rest of the dislocated territory located on a 105 kb contig CTG000462 which was scaffolded upstream with previous contig CTG000461 in a way that reflects a complete inversion to the subject (Figure 2.16). This

rearrangement is missing from the Sanger assembly. However, similar putative segmental chromosomal translocations were also found in contig scf718000002591 with a length of 1.9 Mb of strain Tc1/148 PacBio assembly assigned to the reference MBC 10 (Figure 2.16.B).

Five relatively minor intrachromosomal inversions were also noticed on this chromosome in *T. congolense* PacBio. The first location is proposed to affect kinesin, midasin and tetratricopeptide repeat putative genes. The second region affected by such change has predicted genes encoding for proteins with putative vesicle transport domains Vps51/Vps67, U1 small nuclear ribonucleoprotein C, mitochondrial processing peptidase alpha subunit (pseudogene), elongation factor 1-alpha and hypothetical proteins. A third area was noticed to have inverted sequence with genes of putative peptidase C19, ubiquitin carboxyl-terminal hydrolase, protein of unknown function, protein tyrosine phosphatase, a isoleucyl-tRNA synthetase, DNA replication licensing factor MCM8, structural maintenance of chromosome 2, microtubule-associated protein and a hypothetical gene. Finally, this fragment of chr10 in *T. congolense* PacBio has predicted genes encoding for ATP12 chaperone protein and a putative gene of unknown function. Almost all of mentioned intrachromosomal rearrangements were also suggested by Sanger version of the genome assembly on this chromosome (Figure 2.16.A).



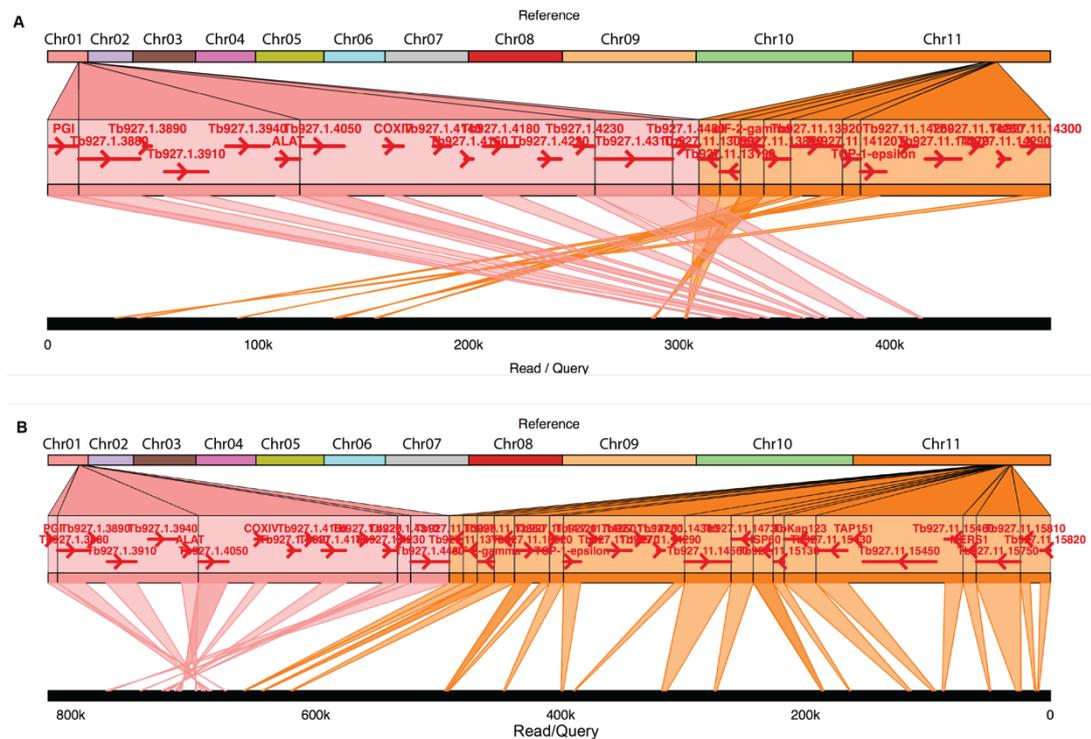
**Figure 2.16 Putative inter and intra- chromosomal rearrangements on *T. congolense* PacBio chr10.** A) An ACT comparison based on tBLASTx search for each pair of MBCs with cut-off evalule  $1e^{-10}$ . The rearrangement between Tb927 chr3 and *T. congolense* PacBio chr10, showing that the shared inverted segment spans 276 kb (a bundle of crossed blue lines top panel). Intra- chromosomal inversions targeted genes on separated regions on this pseudo-chromosome (blue ribbons in the middle panel), which is also showed by *T. congolense* Sanger assembly (blue ribbons in the bottom panel). B) GenomeRibbon plot based on synteny search using PROMmer with maxmatch and option “c” set to 200 showing an 800.7 kb contig (TcAH20) assigned to *T. congolense* PacBio chr10 (bottom black line) starting at 5’ end with a putative inverted segment from Tb927 chr3 (Brown ribbons); separated from Tb927 chr10 genes (green ribbons) by a DNA sequence with three *tRNA* predicted sequences. The upper multi-colour line presents the pseudochromosomes of Tb927 reference sequence.

#### **2.3.9.2.5 Sequence assortments in *T. congolense* PacBio chr 11**

An interchromosomal transposition was noticed to affect a large segment (159 kb) of Tb927 chr1 on 5' end of 475 kb contig (TcAH 2590) allocated to *T. congolense* PacBio chr11, which scaffolded upstream with a contig (CTG000488) containing telomeric repeats on the other end. Downstream to this region, an 8.2 kb non-coding DNA sequence (Characterized by the presence of three predicted *tRNA* genes and a transposon *Ingi2* sequence) in *T. congolense* PacBio chr11 located within the same contig (TcAH 2590) was in synteny with Tb927 chr11 genes (Figure 2.17. A). This region was also supported by Tc1/148 PacBio assembly contig of length 800 kb (Figure 2.17. B).

In line with other rearrangements in previous interchromosomal exchange this region is also inverted. The displaced fragment has a number of predicted genes encoding mainly for surface proteins, ion channel/ calcium-activated BK potassium channel alpha subunit, ABC transporter, integral membrane protein, flagellum attachment zone protein 2, fusaric acid resistance protein-like and other genes encoding for indefinite protein sequences.

The nature of genes and the putative scaffolding according to long reads evidence of this contig (TcAH 2590) to another one ended with telomeric repeats suggests the location of this contig (CTG000488) on the 5' end of *T. congolense* PacBio chr11.



**Figure 2.17 GENOMERIBON plot of the displacement in *T. congolense* of a segment belonging to Tb927 chr1 (pink ribbons) associated with a region that has predicted genes linked to chr11 (orange ribbons). The plot generated is based on shared sequences searched using the PROMmer tool of MUMmer package 3.23 with options of maxmatch and option for synteny alignment “c” set to 200. A) 475 kb *T. congolense* PacBio contig (TcAH 2590) (bottom thick black line) shared sequences (middle horizontal pink stripes) with Tb927 chr1 gene IDs annotation (red arrows) (see text for predicted shared and dislocated genes) on 3’ end of the contig. Shared genes with Tb927 chr 11 (viewed in orange coloured stripes). The top multi-colour bar refers to reference sequences of eleven pseudo-chromosomes of Tb927. B) A Tc1\_148 contig (scf7180000002707) of 800 kb length showed similar rearrangement to IL3000 on the 3’ end of the contig.**

The predicted inter-species large structural variants, most likely interchromosomal inversions, affecting the MBCs of *T. brucei* in comparison to two strains of *T. congolense* have been shown here for the first time.

Our analysis suggests that the interchromosomal rearrangements mainly affected MBC one in *T. brucei*, as shared segments from it were homologous to regions on contigs assigned to *T. congolense* MBCs (2, 7 and 11), which in turn is missing from contigs assigned to the putative *T. congolense* MBC one. This suggests that this chromosome is the most *T. brucei* MBC subjected to the proposed rearrangements most likely affecting a cluster of *alpha* and *beta tubulin* genes. The cluster of tandemly repeated tubulin sequences of high sequence similarity on this chromosome was previously showed to be involved in genetic recombination among different drug resistant *T. brucei* replicas (Gibson and Bailey, 1994).

MBC one has previously been shown to have a high degree of sequence variation up to 30% between different strains and even between the two homologues of *T. brucei* (Melville *et al.*, 1998, 2000); however, these variations were mostly subjected to the subtelomeric region. Gene mobilization on a small scale was also noticed in different *T. brucei* stocks affecting housekeeping genes (normally in the core chromosomal regions) (Gibson and Garside, 1991). Chromosomal structural rearrangements have been also recognized in other protozoa such as *Giardia intestinalis* (Tůmová *et al.*, 2016).

### **2.3.9.3 Strand Switch Regions in *T.c.IL3000***

A Strand Switch Region (SSR) refers to the genomic region located especially in internal chromosomal positions lying between two Directional Gene Clusters (DGCs) in which the direction of transcription process is switched from forward strand to the reverse strand or where transcription termination ends. Thus, they are either divergent sequences dSSRs (when the transcriptional direction is opposite to each other) or convergent cSSRs (when the transcription of two convergent DGCs end). These regions were not studied before in this trypanosome probably due to the lack of contiguity in the current draft

assembly. Herein I am going to describe the putative regions in the *T. congolense* genome.

Genome analysis revealed 75 putative regions (40 divergent and 35 convergent). The former is characterized by longer sequences than the latter under  $p\text{-value} < 9.398e^{-14}$  of 95% confidence. There is a higher probability of presence of transposable elements in dSSRs over cSSRs sites (40%, 11%) respectively, which additionally, have lower GC contents (46.47% and 51.68%), respectively (Table 2.7). It is more likely to have direct repeat motifs of either AT-rich and/or CT-rich repeat units of varying lengths (25-75). Furthermore, predicted *tRNA* features were noticed in these sequences with tendency to locate on cSSRs rather than the dSSRs (Table 2.7).

**Table 2.7 Comparison between Divergent SSRs and Convergent SSRs.**

Category	Divergent SSRs	Convergent SSRs
<b>Number</b>	40	35
<b>Mean length bp</b>	9,803	3,915
<b>Ingi2 elements (percentage)</b>	18 (40%)	5 (11%)
<b>tRNA</b>	0	3
<b>tRNA + Ingi2</b>	0	1
<b>snRNA</b>	1	0
<b>GC%</b>	46.47	51.68

#### 2.3.9.4 Directional Gene Clusters DGCs in *TcIL3000*

The genes of trypanosomatids are arranged in clusters oriented in one direction (either forward or reverse); these genes are more likely to encode for structural proteins and other important proteins for cell survival. Such organization has a special importance, as the genes in these clusters are polycistronic (transcribed in one batch). The size and organization of these gene pools have not been studied before in this species due to the lack of physical sequence integrity.

Our assembly contigs unveiled the gene organization in the majority of these putative regions. The total number of 100 DGCs was assumed from eleven predicted MBCs of *T. congolense* PacBio assembly. However, this number could be an underestimate the actual number, as the DGCs with physical gaps were not included in the analysis and we considered only convergent and divergent clusters.

The number of putative DGCs showed a general trend of increasing count with the chromosome number (size), but an exception was noticed in pseudochromosome four as its DGCs is less than those on chromosome three and two. Whilst the estimated number of forward (fDGCs) is less than that of reverse (rDGCs) (47, 53), respectively, the estimated mean length of both gene-guiding orientations is similar (165,651 bp, 165,595 bp), respective (Table 2.8). fDGCs and rDGCs were sometimes interrupted by one or two predicted *VSG* and or *ESAG* genes located on the alternate strand of the DNA segment. The localization of these clusters is more likely toward the internal part of MBCs leaving the ends of MBCs, for perhaps sub telomeric features.

The SSRs and the DGCs are both related; as the SSRs are localized between each of two DGCs. Our results suggested lower GC content of dSSRs (46.47%), which is highly similar to the corresponding region on chromosome one of *Leishmania major* (Puechberty *et al.*, 2007). In addition, these sequences have been proposed to host polymerase II transcriptional sites, so that they are considered transcriptional initiation sites for both PTUs on single dSSR, which further could host histones in *L. major* and *T. brucei* (Martínez-

Calvillo *et al.*, 2003, 2004; Puechberty *et al.*, 2007; Kolev *et al.*, 2010), respectively. Interestingly, a similar trend in GC content differences between the two regions was also noticed in *Leishmania* chromosomes

**Table 2.8 Prophesied Directional Gene Clusters DGCs per each of TcIL3000 predicted chromosomes.**

Pseudo chromosome	DGCs	fDGCs*	Mean length (bp)	rDGCs**	Mean length (bp)
Chr01	2	1	254, 041	1	111, 491
Chr02	5	2	56, 202	3	144, 377
Chr03	9	4	94, 533	5	131, 843
Chr04	3	1	296, 095	2	261, 082
Chr05	7	4	175, 266	3	71, 099
Chr06	8	2	205, 941	6	105, 447
Chr07	11	5	124, 862	6	125, 176
Chr08	10	5	170, 655	5	150, 334
Chr09	7	4	164, 852	3	288, 906
Chr10	20	10	159, 938	10	195, 304
Chr11	18	9	210, 625	9	209, 119

\*fDGCs: Forward Directional Gene Clusters.

\*\*rDGCs: Reverse Directional Gene Clusters.

### **2.3.9.5 Putative new genes in *T. congolense* PacBio assembly**

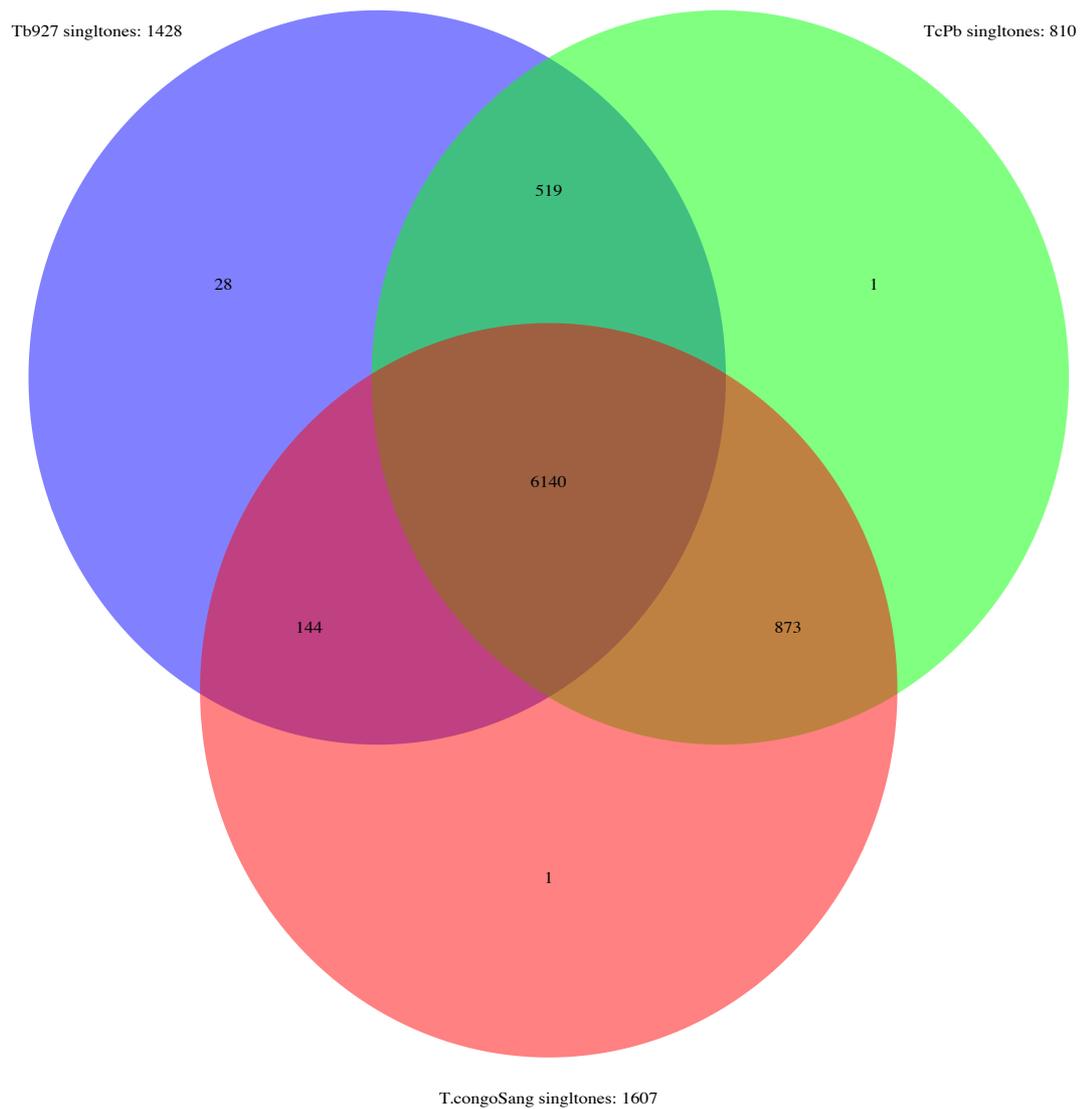
Protein clustering analyses of genome-wide protein products suggested numbers of gene families (orthogroups) that were shared between the three assemblies, between each two and assembly specific orthogroups in addition to the singletons for each assembly are shown in (Figure 2.18). We focused our analysis on predicted new orthologues and unique genes suggested by the *T. congolense* PacBio assembly.

#### **2.3.9.5.1 New TcIL3000 genes shared with Tb927**

The cluster analysis of the TcIL3000 and Tb927 putative proteomic data sets unveiled large groups of new shared gene families with Tb927 rather than more species-specific genes. These were 519 putative shared gene families containing 1,279 genes, from which 1,200 belong to 497 orthogroups and were putatively assigned to functions. From these 604 were *T. congolense* PacBio assembly genes. This cluster of orthogroups showed slightly higher members of *T. congolense* due to paralogs (**Table 2.9**).

##### **2.3.9.5.1.1 Position of the predicted new TcIL3000 genes shared with Tb927 on the MBC**

Data analysis showed that most of the new shared genes are located on the eleven putative MBCs of *T. congolense* PacBio assembly (519/604); furthermore, those genes were more likely on the core regions of MBCs, as evidenced by their location within DGCs suggesting common core gene sets at least between these two trypanosomes (Figure 2.19).



**Figure 2.18** A Venn diagram of OrthoFinder clustering analysis for the three assemblies' proteomic data (Tb927, *T. congolense* PacBio and *T. congolense* Sanger). Number of single genes assigned to each assembly. The new predicted 519 orthogroups shared between *T. congolense* PacBio and Tb927 appear in the shared area between the two proteomes.

### 2.3.9.5.1.2 Investigation of genes shared between Sanger assembly and Tb927

The protein clustering analysis (Figure 2.18) suggesting 144 orthogroups containing 233 sequences shared between *T. congolense* Sanger assembly and Tb927 but not shown by *T. congolense* PacBio proteome set. The missing genes could be due to a number of reasons:

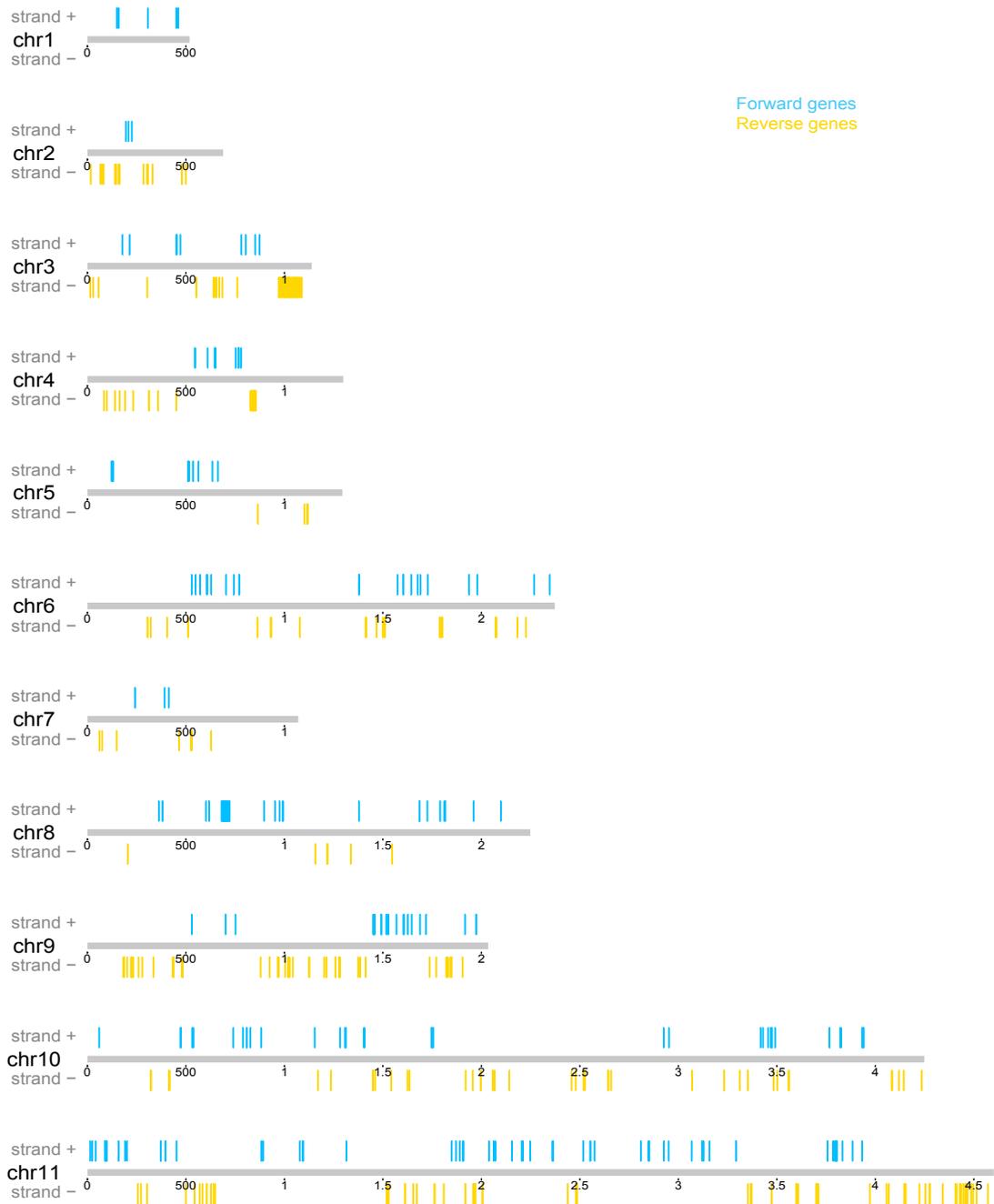
1. Missing sequence from the PacBio assembly
2. Missing gene calls in the annotation or
3. Poor clustering because of sequence
  - a. differences between the assemblies (Li *et al.*, 2012)
  - b. the length of the gene model

Such differences would be expected as the two *T. congolense* assemblies used different sequencing techniques and most importantly, the partial genome Sanger assembly was sequenced to genome coverage of 5x (<http://www.sanger.ac.uk/resources/downloads/protozoa/trypanosoma-congolense.html>), while our final PacBio assembly has a coverage of 47-fold, which was further subjected to error correction step using 110 fold paired ends Illumina HiSeq coverage (as explained earlier in the chapter).

The 233 protein sequences of Sanger assembly shared with *T. brucei* were searched against our PacBio assembly using BLASTx search with a threshold value of  $10^{-5}$ . This search produced 225 significant hits in our PacBio assembly. This result confirms their presence in our assembly, however, the remaining eight sequences were perhaps still missing from PacBio assembly, thus, further investigation was achieved but this time the BLASTx search was done against raw PacBio reads, which in turn showed significant hits towards the eight Sanger sequences, suggesting their absence from the assembly. However, they were found in the raw reads, so perhaps could not be assembled into contigs.

**Table 2.9 New proposed shared gene families of *T. congolense* PacBio with Tb927 according to the clustering achieved by OrthoFinder.**

<b>Category</b>	<b>Count</b>
<b>Total shared gene families</b>	519
<b>Total shared genes</b>	1, 279
<b>TcIL3000 genes</b>	648
<b>Shared gene families with known function</b>	497
<b>Total shared genes with known function</b>	1, 200
<b>TcIL3000 genes with known function</b>	604
<b>Number of Genes on TcIL3000 putative MBCs</b>	519



**Figure 2.19 Distribution of new genes of *T. congolense* PacBio assembly shared with Tb927 on the putative MBCs of *T. congolense* PacBio assembly.** Annotated shared genes on putative *T. congolense* MBCs were showed according to the colour key. The numbers underneath the grey lines represent the length scale in 500 bp increment, numbers equal or larger than 1 are in Mb length. KaryoploteR package on R-project was used to generate this plot according to the feature coordinates stored in the GFF3 file.

### **2.3.9.5.1.3 Functional analysis and Gene Ontology Enrichment analysis of the new shared genes**

In order to investigate functionality of the genes and conduct Gene Ontology enrichment analysis, the genes with known GO IDs were extracted. This provided 203/604 genes with assigned ontology annotation. Within this subset (203) a total number of 477 GO terms were obtained and used as input to GO slim classification which resulted in a final 127 ancestral GO terms (Table 2.10).

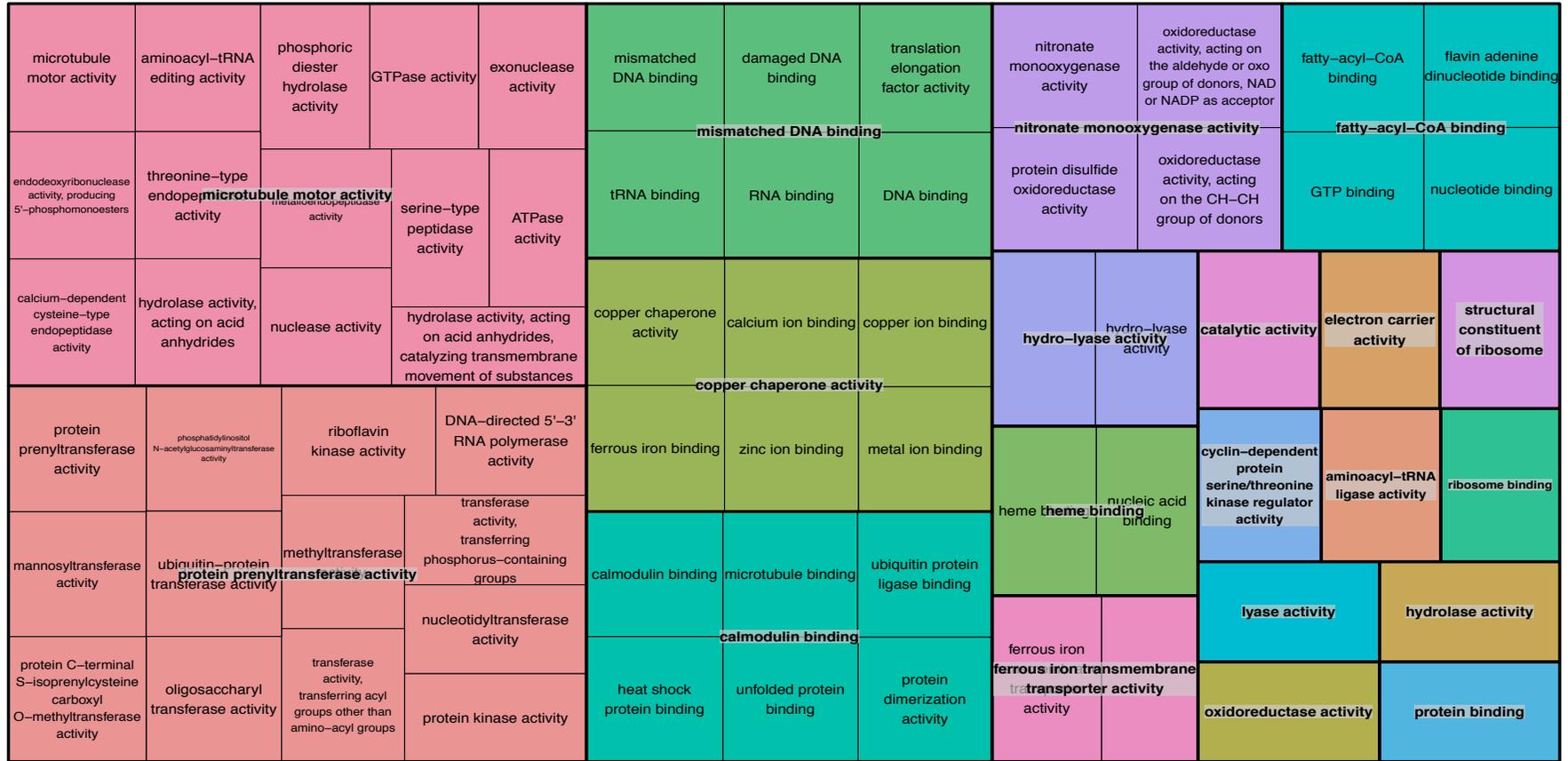
Figure 2.20 summarizes the 225 molecular functions of possible gene functional terms according to REVIGO clustering of GO term IDs.

In order to investigate the importance of this new set of genes compared to the whole genome, the web tool of GO slim classification and count was adopted by collecting the genome wide GO terms and the percentages of the new gene set to the whole picture were calculated. The top ten GO slim classification revealed that three terms lie in the following categories: cell homeostasis, cell replication and calcium ion binding, were fully represented by the new suggested genes. The second set of gene ontology terms resulting from the analysis, signify that 50% of whole genome repertoire is probably involved in four important cellular functions: cytoskeletal protein binding, ribosome, protein kinase activity and motor activity (Table 2.11).

**Table 2.10 GO slim classification and counts of 477 GO terms derived from new 203 *T. congolense* PacBio genes shared with TB927 genes. First ten only are shown.**

<b>GO Class ID</b>	<b>Definitions</b>	<b>Counts</b>	<b>Fractions</b>
<b>GO:0003674</b>	molecular function	78	14.47%
<b>GO:0008150</b>	biological process	65	12.06%
<b>GO:0003824</b>	catalytic activity	45	8.35%
<b>GO:0008152</b>	metabolism	40	7.42%
<b>GO:0005488</b>	binding	29	5.38%
<b>GO:0005575</b>	cellular component	22	4.08%
<b>GO:0009058</b>	biosynthesis	21	3.90%
<b>GO:0016787</b>	hydrolase activity	20	3.71%
<b>GO:0019538</b>	protein metabolism	20	3.71%
<b>GO:0005623</b>	Cell	17	3.15%

Molecular Function treemap



**Figure 2.20** REVIGO summary of GO IDs with molecular functions pinned to the new *T. congolense* PacBio genes. Terms were clustered according to their semantic similarity and the name of each cluster was taken from the most representative member. Each cluster has its own colour and size relative to the number of its members.

**Table 2.11 The percentage of GO slim classification and count for the new gene set of *T. congolense* PacBio orthologues to corresponding Tb927 genes compared to the all *T. congolense* PacBio GO annotations. Top ten values are shown.**

<b>Go IDs</b>	<b>General GO terms</b>	<b>New/all GO IDs %</b>
<b>GO:0019725</b>	cell homeostasis	100
<b>GO:0008283</b>	cell proliferation	100
<b>GO:0005509</b>	calcium ion binding	100
<b>GO:0008092</b>	cytoskeletal protein binding	50
<b>GO:0005840</b>	ribosome	50
<b>GO:0004672</b>	protein kinase activity	50
<b>GO:0003774</b>	motor activity	50
<b>GO:0005739</b>	mitochondrion	37.5
<b>GO:0006629</b>	lipid metabolism	36.8421
<b>GO:0006996</b>	organelle organization and biogenesis	34.7826

#### **2.3.9.5.2 *T. congolense* PacBio assembly unique single genes**

OrthoFinder clustering tool suggested a set of 810 single genes i.e. each single gene has its own orthologous group. Filtering out those genes with unknown function and potential genes from contamination source of origin (genes related to phage and viral replication) resulted in 291 genes with known function that were obtained and subjected to further analyses as follows:

##### **2.3.9.5.2.1 Putative genome localization of singleton genes have known functions in the annotated *T. congolense* PacBio assembly MBCs**

In contrast to the new genes shared with Tb927, singleton genes showed a different figure and were more often found in the contig BIN 221/291, a small number localized in the 70/291 MBCs (Figure 2.21). As most of these genes allocated to the Bin sequence of PacBio assembly, this suggests they were subtelomeric and non-syntenic to *T. brucei* MBCs.

##### **2.3.9.5.2.2 Investigation of Sanger assembly specific singletons**

The protein sequence clustering analysis suggested 1,607 possible sequences were unique to the Sanger assembly (Figure 2.18). An approach to inspect potential non-clustered genes using BLASTx search with e-value cut of  $10^{-5}$  against PacBio assembly was used. The adopted approach resulted in 1,414 significant hits, confirming that a number of genes were present in our assembly but missed during the clusters process. Two possible reasons for this were suggested: 1. 112 gene models from the Sanger data were fragmented and therefore clustered independently; 2. Gene models where either longer or shorter gene models and therefore not clustered by ORTHOFINDER with other protein sequences. This latter reason was the most common due to the sequence differences between the two assemblies (Li *et al.*, 2012). The remaining 193 possible missing genes were BLASTx searched against our PacBio raw reads, which in turn resulted into further 74 significant hits. However, 119 sequences from Sanger assembly singletons as suggested from the cluster analysis were still missing from both PacBio assembly and the raw reads. Accordingly, in depth investigation towards those potential PacBio-missing Sanger singletons was undertaken. The results showed that, the

majority of these putative protein sequences (86/119) revealed significant hits to sequences belong to *Pseudomonas*, *Retroviruses*, *Chlamydia* and *E. coli* when they searched against NCBI non-redundant database, suggesting that these genes were from contaminating sequences in the Sanger assembly. The other sequences (33/119) were short amino acid sequence mainly interrupted by stop codons, signifying protein non-coding genes.

#### **2.3.9.5.3 Enrichment analysis of GO terms of singleton genes**

GO terms IDs were assigned to about 120 single genes of PacBio assembly had in total 229 GO IDs. According to REVIGO those genes could be involved in 67 biological pathways, 14 cellular components and 148 molecular functions.

In general, the main biological pathways that those genes are suggested to be participating in are (Figure 2.22):

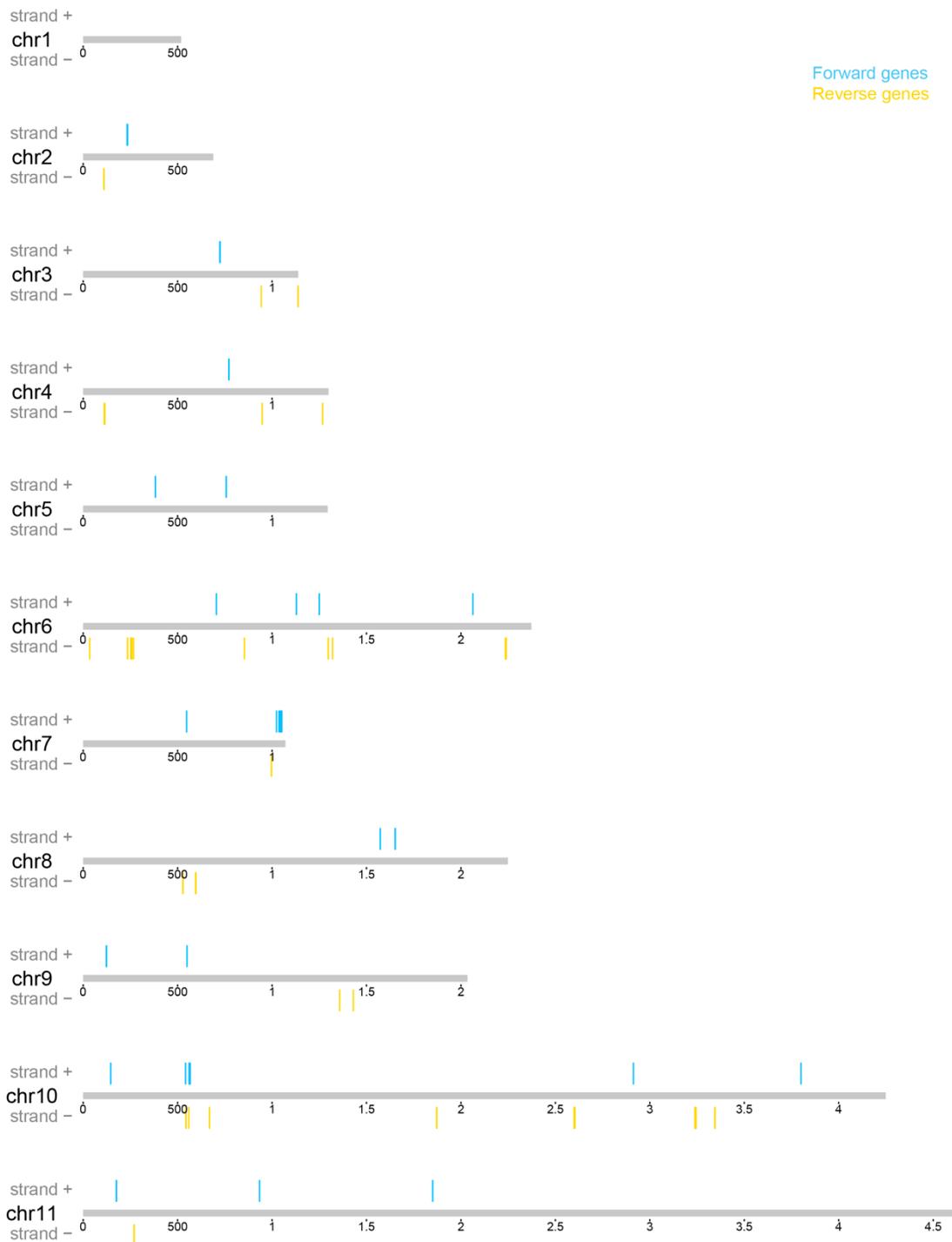
**First**, proteins might have roles in different cellular transport machinery for example intracellular protein transport, trans membrane transport and ferrous ion transport.

**Second**, DNA replication and DNA biology. For example: DNA metabolic process, cyclic nucleotide biosynthetic process, DNA replication initiation and protein methylation.

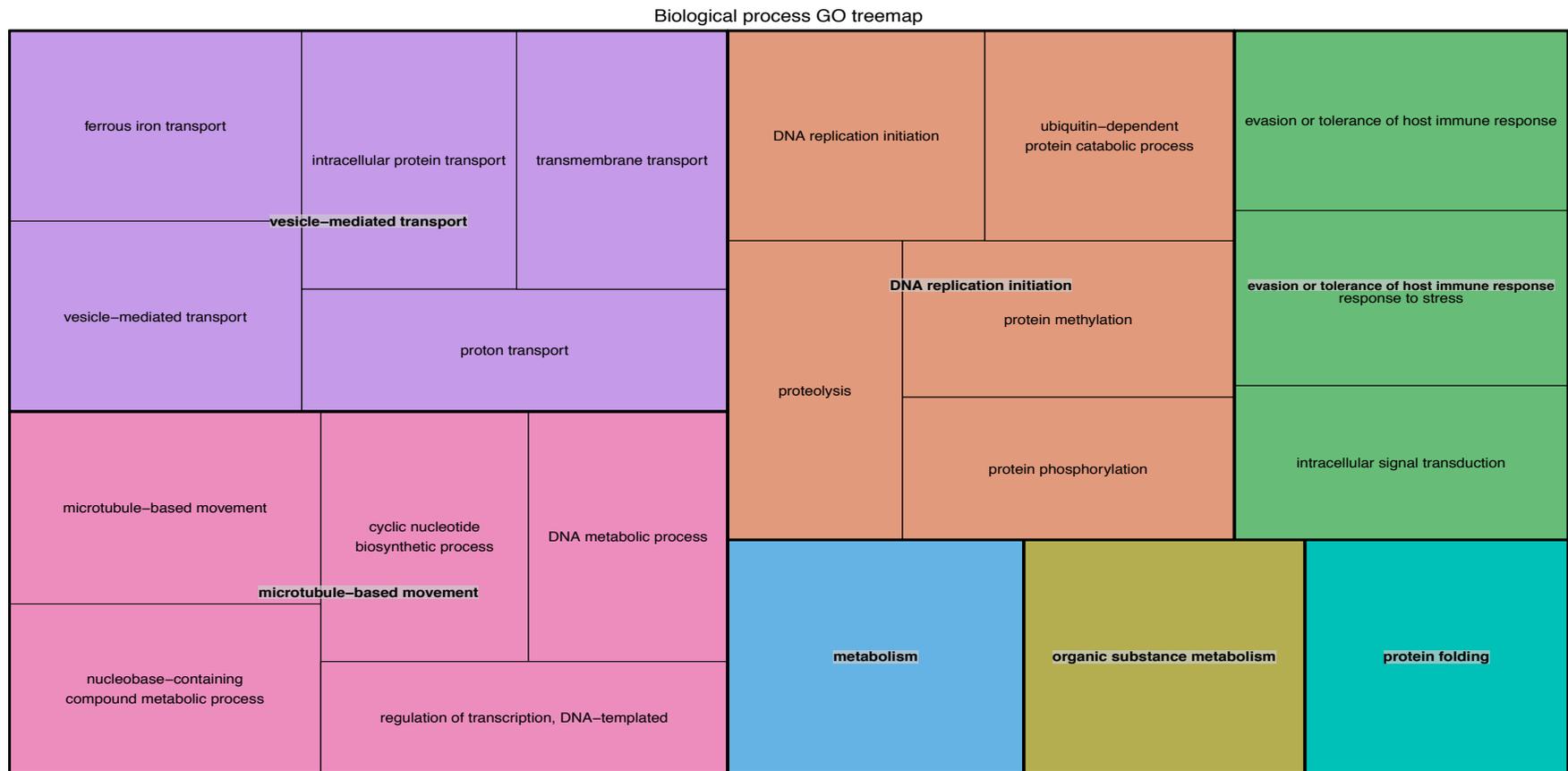
**Third**, Evasion of the host immune system and virulence factors.

The cellular localization of these genes suggested to be more likely encodes for proteins that cover the cell membrane or protein complexes that could attach to it or part of this cellular territory.

Molecular function results of GO terms enrichment analysis proposed varied roles in this category; for example, protease activities, metal binding, transportation, DNA and RNA binding activity.



**Figure 2.21 A karyoplot of *T. congolense* PacBio MBCs reveals the new proposed single gene locations on the MBCs.** The scale is in increment increase of 500 kb. For the viewing clarity of the scale, the length of chromosome of 1 Mb and higher is written as units (1-5).



**Figure 2.22 GO terms enrichment analysis highlight the proposed cellular pathways that these genes could contribute to.** REVIGO semantic algorithm was used to cluster *T. congolense* PacBio single genes, closer terms were put in one group. Each colour-distinguished box includes more related terms.

In general, the new proposed genes in *T. congolense* PacBio assembly showed contrasting tendencies towards genome positioning. Whilst the majority of shared genes with *T. brucei* showed preference localization towards the internal core regions, the singletons were more likely to be in the BIN (most of the contigs in the BIN were not syntenic and not homologous to the *T. brucei* MBCs), which in turn suggests more species specific features to be assigned to the BIN contigs, which might suggest a subtelomeric nature of these contigs (Berriman, 2005; Ivens *et al.*, 2005). The enrichment analysis suggested high probability of both new gene sets to be localized on parts of the *T. congolense* cell membrane with mainly transportation function especially metal and amino acid transportation. Previous publications referred that the transporter proteins are in particular importance for the unicellular parasites as they lack a number of metabolic pathways. This led the protozoa parasite to depend on the host resources via developing transporter system for crucial nutrients like amino acids, folate, important heavy metals (such as iron, copper and calcium) to compensate the lack of synthesis and to suffice the high demands of parasite to establish rapid proliferation and to cope with hostile host environment (Dean *et al.*, 2014). In this context, and most importantly, our data analyses suggested *T. congolense* specific single genes lie in the mentioned category and in immune evasion mechanisms, highlighting the possibility of these genes to be potential candidates for drug targeting and/or vaccine preparation.

## 2.4 Conclusion

This sequencing project has improved the genome assembly of *T. congolense* providing better gene models, more genomic structural integrity and the resolution of many ambiguous regions, enabling us to investigate a number of previously undetermined genome regions in *T. congolense*.

First of all, the analysis revealed inter and intra chromosomal rearrangements of MBCs in the two *T. congolense* strains relative to *T. brucei* genome structure. These structural changes most likely affected segments on putative MBC one. Chromosomal structural variation analysis suggested that MBC one rearranged with other MBC (2, 7, 11) and one more rearrangement between MBC three and MBC 10. These potential genomic translocations were noticed in both *T. congolense* strains (IL3000 and Tc1/148) in comparison to the reference sequence of *T.b. brucei* strain TREU927. These regions were mainly denoted by *tRNA* genes that have been previously considered as a synteny break points across kinetoplastids genomes. Furthermore, large scale chromosomal translocations agreed in two *T. congolense* strains, suggesting that it is more likely to happen on a species scale rather than be strain specific.

The level of sequence contiguity has also enabled us to locate the boundaries of some presumed DGCs on MBCs in *T. congolense* strain IL3000 MBCs, which are particularly important as it represents polycistronic transcriptional units in these trypanosomes. Moreover, a number of strand switch regions with potential features have been proposed for the first time. These regions were previously proven to involve transcriptional initiation (dSSRs) and transcriptional termination (cSSRs) in *T. brucei* and showed a consistent composition of these regions with other kinetoplastid genomes, suggesting a core essential role of these regions to the kinetoplastid genome structure.

There are 648 new genes that clustered with homologous sequences from *T. brucei*, 604 of these genes have known function and the vast majority were annotated on MBCs (not the BIN). Pathway analyses of these genes suggests they mainly have housekeeping roles in the cell. In addition, this analysis implies approximately 800 genes are unique to the *T. congolense* PacBio

assembly from which 291 showed putative functional descriptions, a small number (70 genes) were annotated on the MBCs supporting the possibility of being species/strain-specific genes with functional enrichment analysis referring to their prospective involvement in surface, transporter, and structural proteins. Both shared and singleton genes could be of particular importance as they might be a potential drug and vaccine targets.

Taken all together, the novel findings in this chapter provide the scientific community working on this parasite with data for the study of cell biology, drug discovery, and vaccine design. Additionally, the new data could provide better understanding and more flexibility in exploring genomic regions of interest previously not described.

## Chapter 3      The structure of *T. congolense* minichromosomes

The work in this chapter and the VSG gene expression sites in some *T. congolense* MBCs showing telomeric repeats in both *T. congolense* IL3000 and Tc1/148 are described in a draft paper to be submitted to one of the specialized journals. The draft manuscript is in Appendix B.

### 3.1 Background

In the mammalian host these parasites unleash the expression of genes that encode for variant surface glycoproteins (VSGs), forming a replaceable shield, and by switching this layer through activation of a specific chromosomal territory called blood form expression sites (BES) one at a time, the parasite maintains a persistent infection in the blood stream of the host (Morrison, Marcello and McCulloch, 2009). In order to preserve this survival mechanisms 11 diploid mega-base chromosomes (MBCs) consisting of internal core diploid regions carry housekeeping genes and terminal sub-telomeric regions possess genes mainly encoding for proteins expressed on the parasite cell surface (El-Sayed *et al.*, 2005). In addition, the genome of trypanosomatids has also evolved numerous VSGs archival chromosomes classified into two classes based on their size. One class represented by minichromosomes (MCs) range in their size from 30 to 100 kb and consist of 100 member, while the other 1- 5 intermediate chromosomes (ICs) ranged from 200- 900 kb in the *T. brucei* (Christiane Hertz-Fowler, 2007). The ICs and MCs thought to be aneuploid and inherited in a non-Mendelian way (Wells *et al.*, 1987).

A few studies carried out on the *T. brucei* MCs structure showed that MCs are linear, transcriptionally silent, and a large part of these chromosomes consists of palindromic 177 bp repeat blocks ending with telomeric repeats (TTAGGG) (Wickstead, Ersfeld and Gull, 2004). However, the presence of such chromosomes in *T. vivax* has been estimated to be present in very low copy number (1-2) (Dickin and Gibson, 1989).

Studies on *T. congolense* karyotype revealed the presence of MCs in different strains of this parasite species and characterized by the presence of microsatellite repeats of 369 bp (Garside, Bailey and Gibson, 1994; Shahada *et al.*, 2007). However, their fine sequence and structure are yet to be uncovered.

It has been estimated that these chromosomes comprise 5-10% of the *T. brucei* genome and contain telomeres and telomeric VSG genes, which are devoted to host immune evasion strategy, which in turn is the most important aspect for trypanosomes to maintain their infectivity (Cross, Kim and Wickstead, 2014).

Besides that *T. congolense* is a neglected species, sequencing of highly repetitive genomic regions represents a big challenge among most of the sequencing approaches which often failed to assemble using first and second-generation sequencing (Treangen and Salzberg, 2012). Furthermore, these regions are difficult to clone (Wickstead, Ersfeld and Gull, 2004; Godiska *et al.*, 2009). However, PacBio SMRT sequencing (a third-generation sequencing platform) enables generation of long reads up to 60 kb in length which could be useful to span long repeats in the DNA sequence such as those ones in the minichromosomes.

The sequence and structure of this part of the nuclear genome has not been studied in *T. congolense*. Although the structure of this set of chromosomes in *T. brucei* has been described (Wickstead, Ersfeld and Gull, 2004), the available database lacks this part of the genome as the previous genome sequencing projects of African trypanosomes focused on mega-base chromosomes. Here, we want to highlight this section of the genome for the given reasons and as these chromosomes play an important role during the blood infection stage.

### **3.2 Aims and objectives of the chapter**

The aim of this chapter is to provide a comprehensive description of this important part of the *T. congolense* genome devoted for the parasite pathogenicity and compare these structures where applicable between two *T. congolense* strains.

In order to achieve this goal, *T. congolense* PacBio *de novo* assemblies of *T. congolense* strains (IL3000 and Tc1/148) were searched for the presence of these chromosomes using previous information, manual inspection and appropriate bioinformatic tools.

### 3.3 Methods

#### 3.3.1 *Trypanosoma congolense* strain IL3000 genomic DNA.

The *T. congolense* IL3000 gDNA supplied by Dr. Liam Morrison (Glasgow) was used for whole genome sequencing. This *T. congolense* strain has been isolated from cattle in the Transmara region of Kenya in 1966 and it was derived from strain Transmara I (Hirumi, H., & Hirumi, K. 1991).

#### 3.3.2 Genome sequencing

The detailed Methods for *de novo* genome assembly and annotation are explained in method chapter two section 2.2.4. Briefly, the assembled contigs were scaffolded with SSPACE-LongReads v1-1 into 1,016 scaffolds with an N50 234 kb with max scaffold length of 1.7Mb. The scaffolds were then “polished” PILON v1.16 (>100-fold coverage of paired end Illumina HiSeq reads of the same gDNA) to correct possible errors in the assembly.

The *Trypanosoma brucei brucei* strain TREU927 genome version 5.1 TriTrypDB <http://tritrypdb.org/> was used as a reference to rearrange assembly contigs and infer chromosomal level assembly using ABACAS v1.

#### 3.3.3 *T. congolense* IL3000 Genome Annotation

The procedure of whole genome annotation was explained in detail in chapter two section 2.2.5. Concisely, genome annotation was performed into two stages. In the first stage: a combination of manual and automated gene annotation (COMPANION protozoa annotation pipeline) (Steinbiss *et al.*, 2016) were used; the manual annotation performed by doing a BLASTx search of BLAST+ (Camacho *et al.*, 2009) package release 2.2.28+ against our PB assembly using *TbTRUE927* and the current *T. congolense*IL3000 version-26 protein databases (<http://tritrypdb.org/>). The tabulated BLASTx output files were then further processed to get the hits that do not intersect with automated pipeline output; in this context, these files were converted to a bed format using custom perl script and the latter formatted output was subjected to BEDTools version 2-2.25.0 (Quinlan and Hall, 2010) with intersect option flagged by `-v`. The final output files were converted to a Gene File Format (gff) using

designed perl script. For the easiness of sequence viewing and manipulation on ARTEMIS, the final gff file format were compressed and indexed using TABIX version1.1 (Li, 2011), and the compressed indexed annotation files were viewed using Artemis version 16.0.0 (Kim Rutherford *et al.*, 2000).

Manual curation of the BLASTx hits were carried out by extending the gene models to the start codon (i.e. 5' methionine) and/or extended to the first 3' stop codon if necessary according to the protein evidence.

In the second stage: we considered the fact that many species specific and subtelomeric genes such as VSG genes and genes encode for other variant surface proteins like ESAGs and could be missed by both automated and previous manual annotation approaches. Thus, we have applied a BLASTx search as mentioned earlier but using self-proteomic database (the proteome that generated in the first stage was re-searched against the *T. congolense* PacBio assembly), and this approach enabled us to identify many of subtelomeric genes like VSG genes and a number of ESAG3 genes/pseudogenes.

#### **3.3.4 Detection of possible *T. congolense* mini-chromosomes**

According to the prior description of the structure of mini-chromosomes of the close relative *T. brucei* (Wickstead, Ersfeld and Gull, 2004), this could be summarized by the presence of a central region of tandem-repeated elements flanked by subtelomeric regions carrying silent VSG genes and ending with a stretch of telomeric repeats. We searched *T. congolense* IL3000 PacBio sequence contigs that could not be allocated to the MBCs and contained features that were consistent with them been mini-chromosomes (MCs), Tandem Repeat Finder (G Benson, 1999) (TRF Version 4.09) a web application was implemented to identify the size, frequency and the consensus sequence of tandem repeat unit on candidates that were manually extracted from the sequence using ARTEMIS genome browser.

Then a database was built from these initial repeat sequences in order to search contigs containing sequences similar to these regions using BLASTn of BLAST+ (Camacho *et al.*, 2009) DNA sequence research package

algorithm with the option of “megablast” was enabled with an e-value cut-off  $1^{-10}$ .

### **3.3.5 Detection of other possible direct repeats**

RepeatMasker (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at <http://repeatmasker.org>) version 4-0-7 was used with *T. brucei* rebase version 20170127 database; the search engine was set to “crossmatch” and default setting was used for the other options.

REpeat Detector (Girgis, 2015) (version 05/22/2015), a *de novo* repeat detection tool was used with tabulated output format to identify possible species-specific repeats. Then conventional perl scripts were used to generate gff annotation file format of these regions in order to visualize it in ARTEMIS. This enabled us to identify the sequence similarity among TcMCs subtelomeres with pseudogenic tVSG copies.

### **3.3.6 Sequence comparison analyses and visualization**

A sequence comparison to show regions of shared synteny among different *T. congolense* IL3000 mini-chromosomes (TcMCs) candidate sequences was accomplished using NUCMmer from MUMmer (Kurtz *et al.*, 2004) package version 3.23 with cluster size set to 200 for better visualization. The show-coords option of MUMmer package with ITH flags was used to generate a table file of coordinates of the syntenic regions among putative sequences of TcMCs. The later table was used as an input of Ribbon genome visualization tool (Nattestad, Chin and Schatz, 2016).

### **3.3.7 Statistical test**

One sample t-test in R platform version 3.3.1 (R Core Team, 2016) used to infer the statistical significance of the spacer DNA sequence length located between intact or pseudogenic telomeric VSG genes i.e. the distance between the ends of the pseudogenic or intact copy of VSG gene or other telomeric gene to the first telomeric sequence (TTAGGG), which in turn was manually extracted using Artemis.

### **3.3.8 Protein and DNA Sequence alignment**

The Expression Site Association Gene 3 (ESAG3) protein sequences and the spacer DNA analyzed in this chapter were aligned using MUSCLE (Edgar and Edgar, 2004) implemented in SeaView package version 4 (Gouy, Guindon and Gascuel, 2010). SeaView platform was also used to infer the maximum likelihood phylogenetic tree of *T.c.IL3000* and *T.b.TRUE927* ESAG3 protein sequences which were generated using PhyML version 3 (Stéphane Guindon and Gascuel, 2003) implemented within SeaView with bootstrap iteration 100.

For manual sequence editing, a color-coded amino acid alignment plot to compare between *T. congolense* ESAG3 protein sequences to Tb927 ESAG3 peptides, was generated using GENIOUS version R.9.0 (Kearse *et al.*, 2012).

### **3.3.9 Genome visualization**

Genome inspection for gene model integrity and sequence layout for all manual investigation steps, manual annotation and generation annotation files of the annotated gene models in this chapter were achieved using ARTEMIS version 16.0.0 (Kim Rutherford *et al.*, 2000; Carver *et al.*, 2012).

The sequence comparison tool ACT (Carver *et al.*, 2005) was employed for investigation and exporting plots of proposed chromosomal structural rearrangements among different chromosomes of different assemblies.

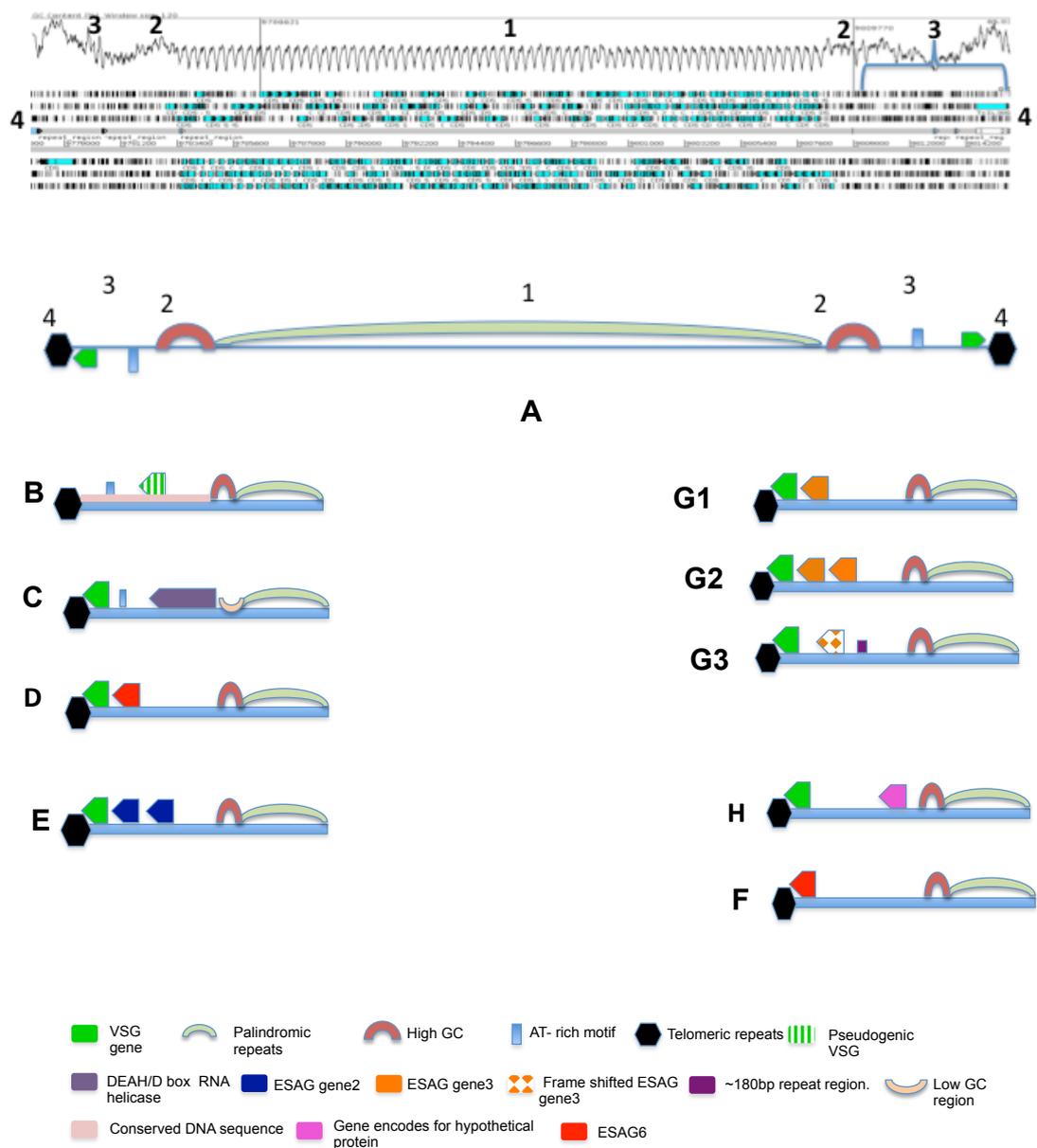
### **3.3.10 Clustering of proteomic data of analyzed assemblies**

See chapter two (section 2.2.9).

### 3.4 Results and discussion

#### 3.4.1 Generic structure of *T. congolense* IL3000 min-chromosomes

Seven putative complete TcMCs sequences and 116 other partial sequences were identified in *T. congolense* IL3000 PacBio assembly that mainly not allocated to megabase sized chromosomes. However, BLASTn search of a characteristic TcMCs repeats revealed that some of these sequences were assigned to a few of MBCs ends. The general structure of TcMCs can be divided into four distinctive regions (Figure 3.1).



**Figure 3.1 A generic model of a complete mini-chromosome and the proposed subsets of TcMCs according to their subtelomeric features (models not to scale).**

A) An ARTEMIS plot (top), a generic cartoon model (bottom) of the putative complete *T. congolense* IL3000 mini-chromosomes

- 1) In which a long stretch of palindromic repeats showed BLASTx hits to VSG genes.
- 2) Conserved regions (1.5-2) kb have relatively higher GC content.
- 3) Highly variable subtelomeric regions 5 kb mainly have distal VSG gene adjacent to the telomeric repeat (3), on both ends of the chromosome.
- 4) Telomeric repeats of 'TTAGGG' units.

B) VSG pseudogene. C) VSG gene accompanied by ATP-dependent DEAD/H box RNA helicase "see text". D) VSG gene with ESAG6-transferrin like. E) VSG with two ESAG2 genes. VSG gene accompanied by ESAG3 genes (a gene family that has not been detected in current reference assembly of *T. congolense* IL3000), G1) one *ESAG3* copies, G2) Two *ESAG3* and G3) one pseudogenic *ESAG3* copies "see the text". H) Gene encodes for a hypothetical protein product with VSG gene. F) an interesting model with only one telomeric ESAG6-transferrin like gene "see the text".

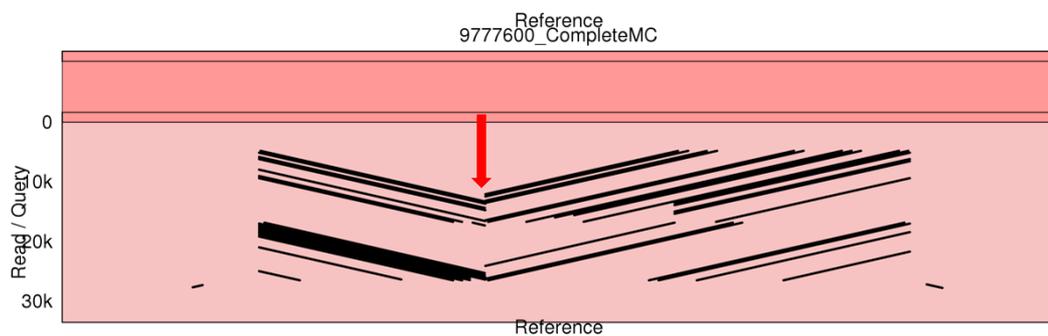
### 3.4.2 Central tandem repeat region (region one)

This region represents 32% - 65% of the total length of the TcMC depending on the size of the TcMC. It is shared among all complete and partial TcMCs and have high AT contents >65%. The tandem repeat units directed from both sides to meet at a point close to the middle of this region is showed by a dot plot view of GenomeRibbon output (Figure 3.2), and could occur in either 369 or 358 (IL3000 and Tc1/148), respectively, with slight differences in consensus sequence among proposed TcMCs. However, the smallest set of these MCs showed tandem repeats from head to tail in one direction.

Surprisingly, the self-proteome BLASTx sequence search showed multiple hits of VSG that have within frame stop codons or not and cover all the six frames translation of the sequence regarding this territory. The length of these hits appears to be 100- 200 amino acids (a quarter to a half-length) of an intact *T.*

*congolense* VSG protein. The biological importance of such findings is yet to be determined.

We propose that if these sequences have biological role/s it could be used as MCs markers or moreover, it might be serve as an archival arsenal for *T. congolense* VSG genes, since it exhibited high sequence conservation among this subset of chromosomes in two *T. congolense* strains and significant hits to VSG gene sequences.



**Figure 3.2 Genome ribbon dot plot of NUCMmer comparison of two complete TcMCs.** Point where tandem repeats of central region from both sides meet (red arrow).

### 3.4.3 Conserved relatively GC-rich regions (region two)

These conserved regions flank the central tandem repeats (region one) over 1.5- 2 kb (Figure 3.1) from both sides. However, the data showed a contrasting but unique GC trend of this sequence when it precedes a specific ATP-dependent DEAD/H box RNA helicase gene family members extending to about 1.7 kb non-coding DNA region (Figure 3.1. C).

### 3.4.4 Variable subtelomeric regions stretched over 5 kb (region three)

This region is located between region two and the telomeric repeat sequences on both ends of the minichromosomes dominated by VSG genes distally and

have almost a conserved length (5 kb) among this category of *T. congolense* genome. This region can be subdivided into different categories according to their feature content (Figure 3.1).

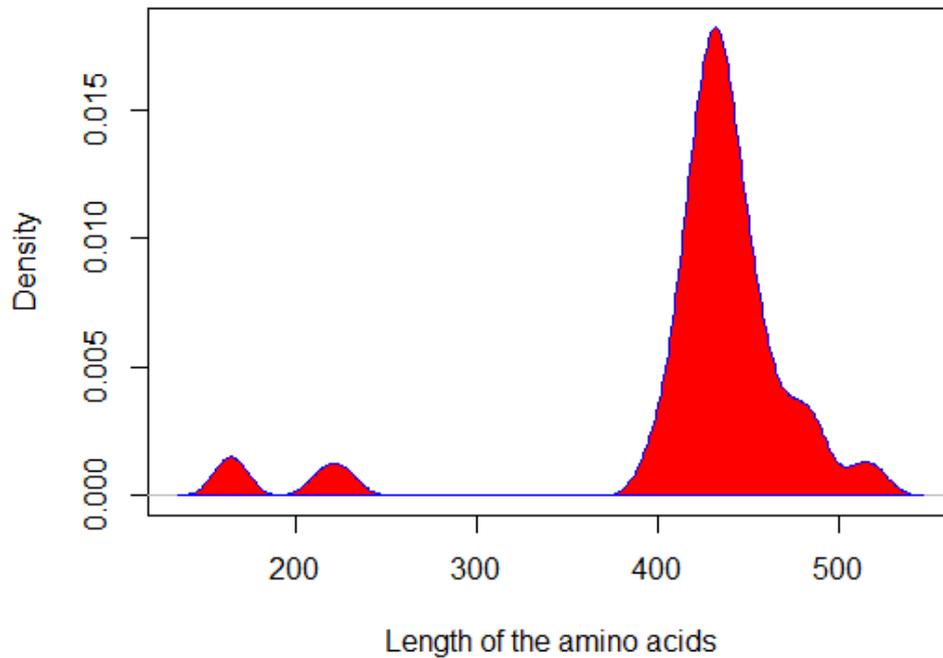
#### **3.4.5 Presence of an intact telomeric VSG gene**

This represents the *VSG* gene that is located closely to the telomeric repeats (the end of TcMC); such feature is termed as “Telomeric VSG (*tVSG*)”. The 3' end of the gene is located within a range of (79- 1823) bp (mean 162) away from the telomeric repeat and this been noticed in 70 sites out of 101 Sites (70%) that showed telomeric repeats at the end.

These *tVSG* gene copies preceded by (19- 45) bp AT-rich or CT-rich repeat motifs 1- 2 kb upstream. The data suggest the presence of only one copy of *tVSG* (i.e. does not accompanied with other protein coding features) in the TcMC subtelomeric area 59 in total. However, one copy or more; intact or pseudogenic version of other features could precede it, such as *ESAG2*, *ESAG3*, *ESAG6* and ATP-dependent DEAD/H box RNA helicase, 13 subtelomeres with telomeric repeat at the end (Table 3.1).

The phylogenetic analysis of telomeric features *tVSG* in TcMCs suggested that the *VSG* genes in TcMCs clustered in variably to orthologous groups and a some of these genes were singletons. Moreover, these genes are more likely to be intact *VSG* copies as their product length distribution showed that the majority of these peptides fall between (400-500) amino acids in length (Figure 3.3).

### Distribution of Telomeric TcMCs VSGs proteins length



**Figure 3.3 Density plot of protein length of putative telomeric TcMCs VSGs.** The majority of VSG protein product lengths lie between 400-500 amino acid in length. The plot was generated using R-package “sm” version 2.2-5.4 <https://CRAN.R-project.org/package=sm>.

**Table 3.1 Identified Genes in the subtelomeric regions of *TcIL3000* mini-chromosomes.** Telomeric VSG genes and other accompanied genes.

Telomeric Feature	Number of sites with this feature	Number of TcMCs with this feature	Notes
<b>Intact VSG copy total</b>	70		59 with only one VSG gene.
<b>ESAG3*</b>	27		7 on contigs ended by obvious telomeric repeats (G1, G2 and G3)**.
<b>ESAG2*</b>	3		One more partial TcMCs but with no obvious telomeric repeats.  Precedes an intact VSG.
<b>ESAG6*</b>	1		
<b>Pseudogenic VSG</b>	26		Highly conserved subtelomeres.
<b>ESAG6 transferrin-like</b>	4		Present as the only telomeric feature
<b>Genes encodes for hypothetical protein</b>	1		Three copies of this gene type.

- Genes accompanying the *tVSGs* in TcMCs proposed models see (Figure 3.1).

### 3.4.6 VSG pseudogenic sequences

In this case the sequences showed frame shifting putative VSG sequences, most likely accompanied by within-frame stop codons (VSG pseudogene). This feature was located almost in the middle of the subtelomeric region, followed by a highly conserved spacer sequence 1800 bp in length till the telomeric repeats. The latter region was characterized by the presence of super conserved AT-rich non-coding repeat motif about 50 bp positioned about 645 bp downstream to the VSG pseudogene. The data also showed that the presence of pseudogenic VSG, 26 (5 in complete TcMC) sites out of 102 (88 partial sequences and 14 from complete TcMCs) total sites (i.e. 26%) with obvious telomeric repeats. Interestingly, this feature was not accompanied by any other possible coding features in this region (Figure 3.1. B). Moreover, the subtelomeres having such features expressed high sequence similarity through the entire subtelomeric region, as is shown by the repeat detector output suggesting potential points of recombination between each other (perhaps to generate mosaic VSG gene product). Such recombination mechanism was able to generate mosaic VSG surface proteins during late infection stage reported in *T. brucei* (Marcello and Barry, 2007b; Hall, Wang and David Barry, 2013) and this also could be the case in *T. congolense*.

### 3.4.7 Telomeric Spacer DNA sequence

This region is located as the DNA sequence between *tVSG* intact (interestingly, we found *ESAG6-like* gene as a telomeric feature see () and (Table 3.1) or pseudogenic copy and the start of telomeric repeats. Hence there needs to be the presence of a telomeric repeat at the very end of the TcMCs partial sequences, so this reduced the total number of contigs to 102 (88 partial sequences and 14 from complete TcMCs) out of 130 predicted sites in total (116 partial sequences and 14 from seven complete TcMCs).

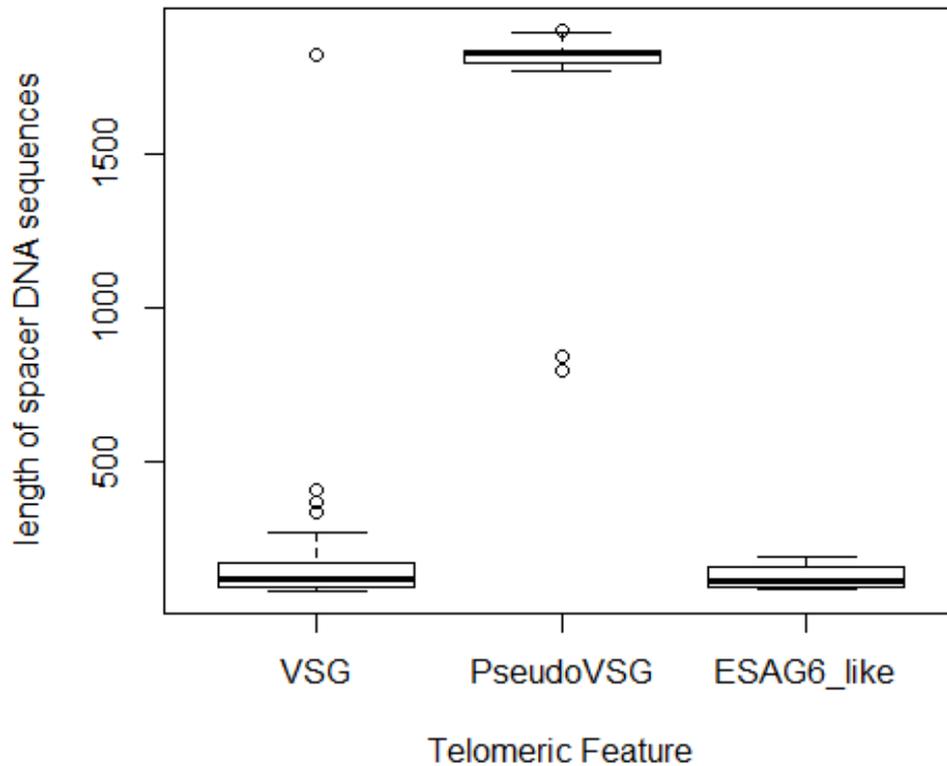
Whilst the spacer length of intact *tVSG* genes range (79- 1,823) bp with mean length of 162, the length of those sequences followed pseudogenic copies of *tVSG* range was (792- 1,898) bp (mean 1,747). The data showed a significant difference ( $p\text{-value} < 2.2e^{-16}$ ) in the length of this region when it is a *VSG* pseudogene site. The boxplot of the lengths of these non-coding DNA sequences belongs to pseudogenic, intact *tVSGs* and *ESAG6-transferrin* like telomeric sites are shown in (Figure 3.5).

As we mentioned earlier, this region is highly conserved among TcMCs subtelomeric regions with *tVSG* pseudogene as is shown in the DNA alignment (Figure 3.4) with an extreme drop in GC content 645 bp downstream to this feature. The common characteristic of this region among telomeric genes is the drop in the GC content.

11017600_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
11073600_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
11360800_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGGTGGCA
11532800_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
11715200_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
11780800_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
11831200_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
12024800_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
12085400_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGGTGGCA
12074400_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
12777600_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
13066400_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
13092800_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
13170400_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
13832000_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
13942400_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
14232800_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
14260000_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
14510400_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGGTGGCA
14816000_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGGTGGCA
15050400_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
15226400_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
15476800_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGGTGGCA
15563200_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGGTGGCA
15631200_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGATGCCA
15818400_pseudo	CAGCCGCCGA	GCAACCTGGG	TGGTCTAAAT	AAAAATGGAG	AATGGTGGCA

**Figure 3.4 Spacer DNA alignment of TcMCs pseudogenic VSG genes on subtelomeres of the mini-chromosomes.** Highly conserved DNA sequence across these sites (54 base window).

### Distance of Telomeric feature from Telomeres

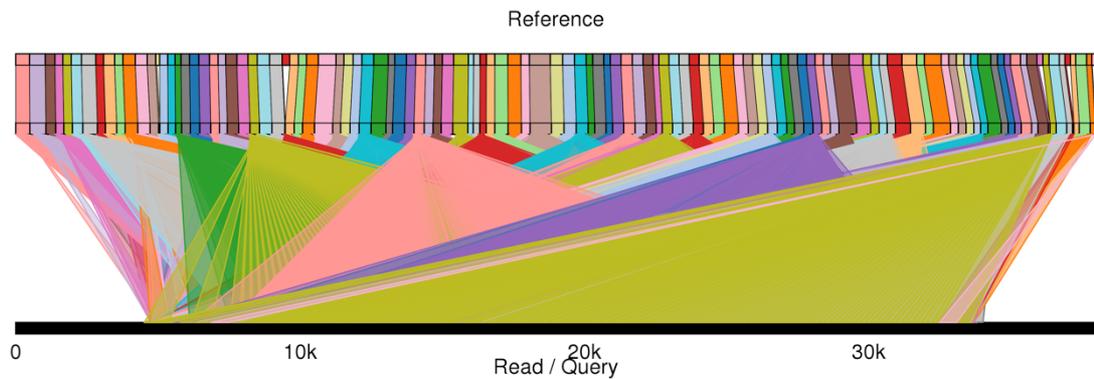


**Figure 3.5 Boxplot of spacer DNA length across TcMCs telomeric features.** The lengths of Spacer DNA from VSG pseudogenes sites showed extreme increase in comparison to the *tVSG* and *tESAG6* genes 'see text'. The plot was generated using R program.

#### 3.4.8 Other features in TcMCs subtelomeric region

This region was considered the most variable region across this set of chromosomes. Unfortunately, the majority of these features were found within the subtelomeres of partial TcMCs sequences, and the partial segments were identified as such due to a breakage in discontinuity of the sequence, but they still have the central palindromic repeats. The NUCMmer of MUMmer package comparison alignment of these sequences to the complete putative TcMC was performed to investigate the presence of central palindromic repeat

characteristic to TcMCs, and the NUCMmer output file was viewed by GENOMERIBBON web application (Figure 3.6).



**Figure 3.6 A GENOMERIBBON plot of NUCMmer comparison.** TcMCs partial sequences (The multi-colour top bar) share the TcMCs characteristic central palindromic repeats with a complete TcMC (bottom black line Query).

#### **3.4.9 ESAG6, ESAG2, and ATP-dependent DEAD/H box RNA helicases**

These genes were also found in the subtelomeres of mega-base chromosomes (MBC) of our PacBio assembly and current Sanger assembly of *T. congolense* IL3000. However, these genes were not noticed in this part of the *T. congolense* genome before. *ESAG2* genes are more likely to occur as two copies accompanying a single intact *VSG* gene throughout the MBCs and in this case TcMCs, while the other genes occur as a single copy accompanying a *tVSG* or being a telomeric feature themselves as in some *ESAG6*- transferrin like genes. Presence of *ESAG6*-like genes in *T. brucei* mini-chromosomes was also reported in low frequency (Cross, Kim and Wickstead, 2014).

Whilst all *tVSG* associated genes are located closely to it, ATP-dependent DEAD/H box RNA helicase genes positioned away from *tVSG* just after the characteristic low GC region that flanks the central tandem repeats (see Figure 3.1. C). However, other copies of this gene on MBCs core regions failed to show similar non-coding conserved sequence and were mostly not related

to VSG genes in terms of genomic localization, suggesting particular importance of this region perhaps as a molecular marker for these genes within telomere/VSG environment.

These relatively long genes occur in their intact form 3, 357 bp or 3, 363 bp in length and were known for their wide range activity on mRNAs from trans-splicing, editing and even silencing of mRNA in both prokaryotes and eukaryotes and more specifically in trypanosomatids (Marchat *et al.*, 2015). Therefore, the presence of these genes with a remarkable non-coding marker might have particular importance to regulate the expressed VSG mRNA from this subset of chromosomes.

Remarkably, all the above sets of genes were noticed in the current *T. congolense* IL3000 Sanger assembly of MBC. However, the *ESAG3* gene family was not characterized before in this species, and we've found it in IL3000 and Tc1/148 PacBio assemblies and exclusively to these chromosomes.

#### **3.4.10 *ESAG3* gene family in TcMCs**

Members of this gene family have been known to encode for highly polymorphic trypanosomal blood stage membrane proteins and located on blood stage expression sites in the *T. brucei* genome (Navarro, 1999; Batram *et al.*, 2014).

We found 27 subtelomeres of TcMCs containing intact one or two copies, or a pseudogene representation mostly frame-shifted along with *tVSG*. Interestingly, the pseudogenes *ESAG3*-like were preceded 500 bp upstream by 180 bp conserved non-coding nucleotides.

Sequence alignment of the protein translation of these predicted *ESAG3* genes of *T. congolense* revealed that they are less divergent in amino acid sequences and more than the corresponding *T. brucei* genes. An exception was, when two *T. congolense* *ESAG3* peptides aligned away from their other *T. congolense* *ESAG3* proteins and close to a cluster of *T. brucei* *ESAG3*. The mapping analysis of these genes showed their location on two putative TcMCs

subtelomeres with G2 model in Figure 3.1 and each one denotes the *ESG3* gene that is located closer to the *tVSG* in both cases. Remarkably, frame-shifted TcMCs *ESAG3* pseudogenes showed truncated versions of these protein sequences (Figure 3.7).



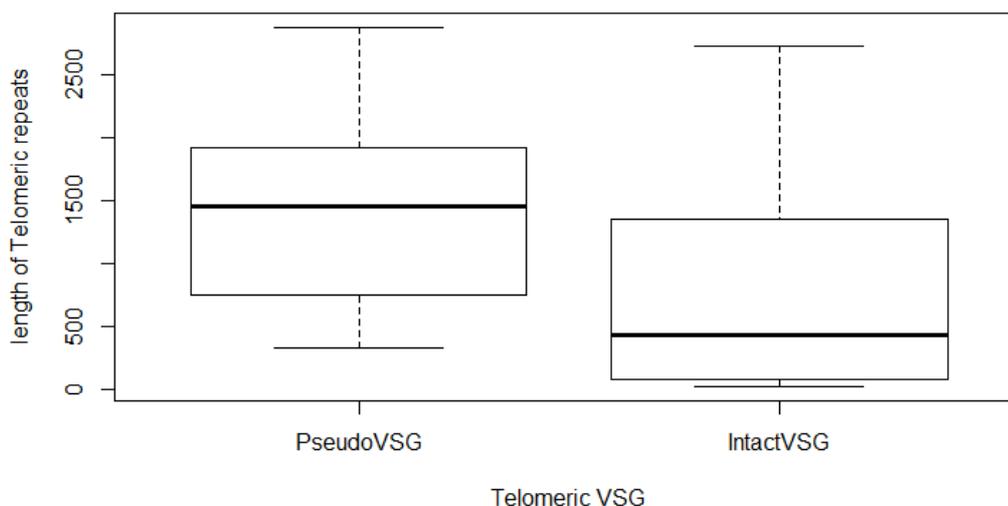
**Figure 3.7 Alignment of TcMCs *ESAG3* and other trypanosome *ESAG3* protein sequences.** Colour coded amino acids revealed more consistency of congolense *ESAG3* proteins (top red box) with apparent truncated proteins (yellow arrows) of frame shifted *ESAG3* pseudogenes. Two sequences were clustered away from *T. congolense* but with *T. brucei* *ESAG3* proteins (lower red G2 box). G3 and G2 referred to the suggested subtelomeric model structure of TcMCs (see **Figure 3.1**).

### 3.4.11 Telomeric repeats (region four)

We noticed the repeats that cap the ends of eukaryotic linear chromosomes in the TcMCs sequences database. The lengths of these telomeric repeats were scored and compared between two major subtelomeric features (intact *VSG* gene and *VSG* pseudogene) of mini-chromosomes.

*Two-way student t* test showed statistical significance differences ( $p$ -value < 0.00034) in the length of these repeats between the two features. Mean length of telomeric repeats linked to *VSG* intact gene was 763.5 bp  $\pm$  97.6, while it was 1406 bp  $\pm$  84.8 of those that showed *VSG* pseudogene (Figure 3.8). These data suggest a correlation between intact *VSG* genes and perhaps shorter accompanied telomeric repeats. In *T. brucei*, the length of telomere was linked to the *VSG* expression mechanism and the presence of shorter sequences were correlated to the laboratory adapted strains (Dreesen and Cross, 2008). Moreover, they were also linked to the *VSG* active expression sites (Dreesen and Cross, 2006). We know that strain IL3000 has been used for a long time in the laboratories and in this respect is similar to *T. brucei*, so we assume that the shorter telomere length that accompanied the predicted intact *VSG* genes might also originate from laboratory adaptation of this strain.

### Length of telomeric repeats of TcMCs according to preceded VSG feature



**Figure 3.8** Boxplot of the length of telomeric repeats in TcMCs with two subtelomeric features of *T. congolense* IL 3000. The pseudogene VSG copies are those VSG genes with frame shift and or within frame stop codons, while the intact copies are those predicted VSG with uninterrupted open reading frames.

#### 3.4.12 Intermediate chromosomes in *T. congolense* IL3000 (TcICs)

Our data suggested the presence of two contigs with palindromic repeats similar to those in TcMCs. However, the repeat unit length exhibited only 358 bp in both contigs. Furthermore, subtelomeres are larger (20 kb) in comparison to 5 kb of the TcMCs and the features content of this region is different, as these subtelomeres have *Ingi2* elements, a blood-form virulence factor cysteine peptidase C, C2H2 two copies zinc finger, genes encoding for unknown proteins, and an intact copy of tVSG. Similarly, intermediate chromosomes in *T. brucei* were also longer and consist of similar centrally positioned repeated DNA core similar to that in MCs (Wickstead, Ersfeld and Gull, 2004).

As these contigs have tandem repeats and four times larger subtelomeric region with different gene contents than those of TcMCs explained above, we

assumed these sequences could belong to possible intermediate chromosomes of *T. congolense* IL3000.

#### **3.4.13 Comparison of TcMCs between *T. congolense* IL3000 and Tc1/148**

The sequence and structure of proposed TcMCs were also analyzed in another *T. congolense* strain (strain Tc1/148) sequenced by collaborators using PacBio sequencing. This assembly showed more completed MCs models with a generalized structure feature contents highly similar to those of TcIL3000 MCs.

A total number of 22 putative complete TcMCs were detected in this strain-assembled contigs with an average length of 26,852 kb; the minimum sequence length was 20,105 kb and the maximum sequence length was 37,974 bp (Table 3.2). This range of length is highly similar to the IL3000 predicted MCs.

The main structure of TcMCs characterized by the presence of four regions (described earlier in this chapter) of this subset of chromosomes is applicable for both strains (Figure 3.1). Here I want to highlight the similarity and possible differences between these two strains with respect to TcMCs:

1. The central palindromic repeats showed high sequence similarity between the two strains (Figure 3.9) in terms of sequence. However, the repeat unit length in Tc1/148 MCs was proposed to be 359 bp in all analysed sequences, while the length of 369 bp is the predominant figure over 359 bp in strain IL3000.
2. In both strains, the genes encode for DEAH/D-box RNA helicase preceded by a conserved DNA stretch of a low GC content about 1.7 kb in length of high sequence similarity between the two strains flanking the central palindromic repeat (Figure 3.1.C).
3. Some Tc1/148 MCs showed putative *ingi2* elements on one sub telomeric region. However, this could not be detected in TcIL3000 MCs, probably because of more fragmented assembly of IL3000 or simply as strain difference.

4. The analysed sub telomeric regions that have a *tVSG* pseudogene (Figure 3.1. B) of MCs of both strains suggested shared related criteria as this region is conserved; there is no other feature other than the pseudo *tVSG* in the relevant subtelomeres, and it is located almost in the middle of the subtelomeric region.
5. Members of *ESAG3* gene family were also detected in investigated Tc1/148 MCs, but lacked the sub-model G2 (Figure 3.1), suggesting a possible variance between strains regarding this gene family.
6. Strain Tc1/148 TcMCs proposed a new subset model characterized by the presence of more likely three genes encoding for an unknown protein product, probably not accompanied by *tVSG* gene and led by a characteristic (1-3) kb of tandem repeat DNA sequence of repeat unit length of 156 bp, located on one subtelomeric region and flanking the central palindromic repeat (Figure 3.10). In this category, BLASTn with megablast search option using evaluate cut off  $e^{-10}$  suggested presence of this model in nine contigs of strain Tc1/148; six of them are claimed as complete MCs and the rest were partial sequences as they don't show telomeric repeats on both ends. All positive contigs showed a central palindromic repeat within their sequences, suggesting that all of detected contigs are within MCs classification. However, a BLASTn search with same options suggested the presence of only one contig in the TcIL3000 assembly of length 5 kb; a contig with the mentioned tandem repeat followed by two hypothetical protein products of predicted genes. We could not link this contig to the MC context because it did not show sequences characteristic of TcMCs.

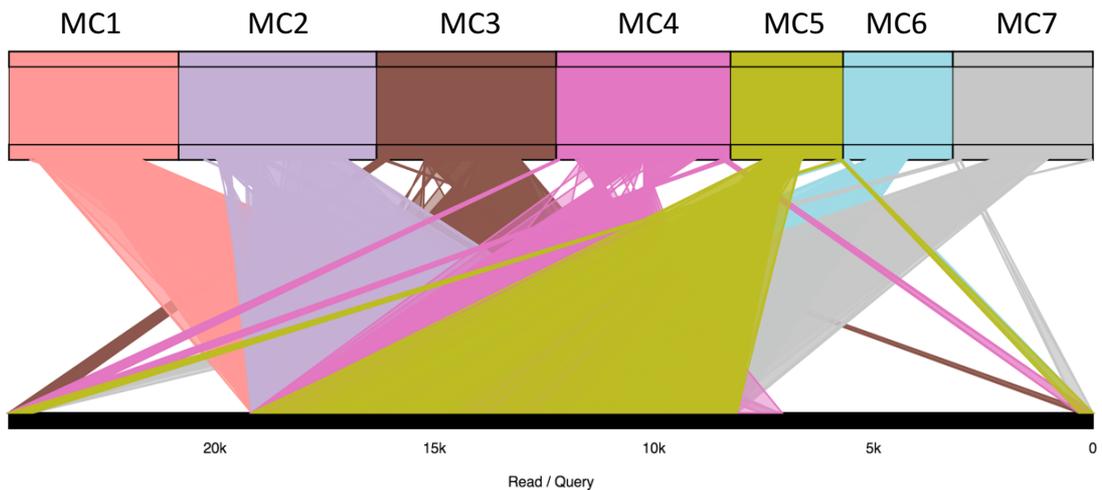
The absence of these important structures from previous sequencing effort could be more likely due to their DNA sequences consisting of long stretches of simple repeated sequences with fluctuating GC patterns, so these regions exhibit particular difficulties in sequencing and genome assembly. Such issues were also highlighted in similar chromosomes in *T. brucei* (Wickstead, Ersfeld and Gull, 2004).

**Table 3.2 *T. congolense* strain Tc1/148 PacBio sequence contigs with putative complete MCs.**

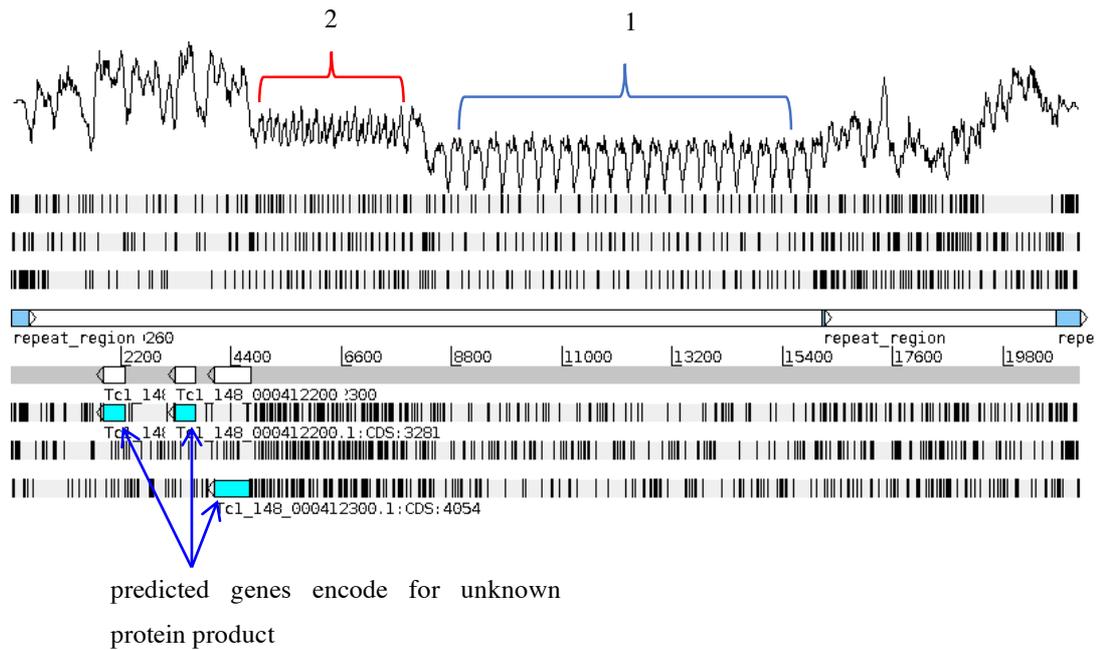
Contig's ID	Length bp	5' end features	3' end features
<b>scf718000002401</b>	35,682	Pseudo <i>tVSG</i>	<i>tVSG</i>
<b>scf718000002408</b>	26,844	<i>tVSG</i>	pseudo <i>tVSG</i>
<b>scf718000002423</b>	33,697	pseudo <i>tVSG</i>	pseudo <i>tVSG</i>
<b>scf718000002425</b>	20,914	pseudo <i>tVSG</i>	<i>tVSG</i>
<b>scf718000002430</b>	37,974	<i>tVSG</i> , <i>DEAH-box</i> <i>helicase</i>	<i>ingi2</i> <i>RNA</i> <i>tVSG</i>
<b>scf718000002447</b>	32,719	<i>tVSG</i>	<i>DEAH-box</i> <i>RNA</i> <i>helicase</i> and <i>tVSG</i>
<b>scf718000002450</b>	27,910	<i>tVSG</i>	1.8 kb tandem RPT* followed by three unknown genes
<b>scf718000002455</b>	21,019	<i>tVSG</i>	Pseudo <i>tVSG</i>
<b>scf718000002462</b>	31,471	<i>tVSG</i>	Pseudo <i>tVSG</i>
<b>scf718000002483</b>	26,203	Pseudo <i>tVSG</i>	Pseudo <i>tVSG</i>
<b>scf718000002484</b>	23,849	<i>tVSG</i> and <i>ESAG3</i>	Tandem RPT* region >1 kb before three unknown genes
<b>scf718000002505</b>	24,693	<i>tVSG</i> , <i>ESAG3</i>	1 kb tandem RPT* region precedes

			three unknown genes
<b>scf718000002508</b>	23,537	<i>tVSG</i>	Short <i>tVSG</i>
<b>scf718000002518</b>	20,105	2.1 kb tandem RPT followed by unknown gene	<i>tVSG</i>
<b>scf718000002528</b>	21,192	<i>tVSG</i>	1-2 kb tandem RPT that precedes three unknown genes
<b>scf718000002529</b>	25,278	<i>tVSG</i> two <i>ESAG2</i>	Pseudo <i>tVSG</i>
<b>scf718000002533</b>	22,990	<i>tVSG</i>	Pseudo <i>tVSG</i>
<b>scf718000002535</b>	26,781	Pseudo <i>tVSG</i>	<i>tVSG</i>
<b>scf718000002542</b>	21,309	Three unknown genes preceded by tandem RPT region	<i>tVSG</i>
<b>scf718000002565</b>	30,758	Transferrin-binding protein like	Pseudo <i>DEAD box RNA helicase, hypothetical and short tVSG</i>
<b>scf718000002597</b>	27,389	<i>tVSG</i>	<i>tVSG</i>
<b>scf718000002748</b>	28,423	Pseudo <i>tVSG</i>	Pseudo <i>tVSG</i>

\*RPT: repeat.



**Figure 3.9** GenomeRibbon plot of NUCmer comparison of *T. congolense* strain Tc1/148 putative MC (bottom thick black line) and strain IL3000 complete seven suggested MCs tope multicolour panel. The broad coloured bundles linked TcIL3000 complete MCs with the Tc1/148 suggested MC sequence (bottom thick black bar), showed shared central tandem repeats between MCs of the two *T. congolense* strains. The deserted region (no coloured ribbons links) refers to the subtelomeric region, while the telomeric sequences of the bottom Tc1/148 sequence's ends are shared with those of MCs belonging to TcIL3000.



**Figure 3.10 An ARTEMIS plot of a complete MC of strain Tc1/148 (21,309 bp long) showed a new model of MC. 1) Region of central palindromic repeats (7.5 kb long) of 359 bp repeat unit. 2) A characteristic tandem repeated DNA sequence span 3 kb with repeat unit length of 156 bp. Three predicted genes with vague function followed region 2. The two repeat regions at the far ends (blue bold arrows) refer to the telomeric repeats that cap both ends of the chromosome.**

### 3.5 Conclusions

This work presented in this chapter highlights an important part of the *T. congolense* nuclear genome, which has special importance for the parasite pathogenicity, especially for blood stage forms. Here, we show for the first time the canonical structure and the possible variants of this subset of chromosomes in two strains of *T. congolense* (IL3000 and Tc1/148).

Our analyses highlighted the existence of defined highly conserved non-coding DNA distinctive to this genomic part, which in turn might suggest high levels of molecular regulation devoted to these chromosomes, especially with the presence of telomeric VSG genes and other coding features known to be expressed on the parasite cell surface and used for immune evasion and parasite survival within blood stream of vertebrate hosts.

Analysing two strains highly supported the undoubted general structure of these chromosomes, yet it revealed the probability of strain specific modifications within TcMCs. The fact that this mini part of trypanosome genome showed here with high degree of specificity in the vicinity of telomere and VSG genes increases its probable importance in the immune evasion strategy of the parasite, and could signify its positioning being a potentially highly regulated surface antigens donor.

This work provided important databases of previously unrecovered genomic territories of this pathogen and establishes the ground information for the professional scientific community towards more understanding of the role of these chromosomes in the life cycle of trypanosomes, while offering potential drug and/or vaccine targets.

## Chapter 4 *T. vivax* PacBio SMRT genome sequencing

### 4.1 Introduction

*T. vivax* is the most predominant trypanosome species in West Africa (Gardiner and Wilson, 1987; Osório *et al.*, 2008). It is mainly transmitted via Tsetse flies of family *Glossinidae*; unlike the other AAT, it can also be mechanically transmitted by other insects like the stable fly (*Stomoxys*) and horse fly (Tabanids), so that (Shaw and Lainson, 1972; Jones and Dávila, 2001), this broad invertebrate host range enabled this species achieve wider geographical expansion even outside Africa like in South America and Mauritius (Hoare, 1972). However, the biological development of *T. vivax* can only occur in the Tsetse flies, and the development takes 3-13 days depending on the strain and the environmental conditions (Gardiner and Wilson, 1987; Osório *et al.*, 2008). This AAT is characterized by a different life cycle in the Tsetse flies, where it completes all developmental stages (epimastigote and metacyclic) in the mouth parts of the fly and does have the midgut proliferation stages as in the other AAT (D'Archivio *et al.*, 2011).

#### 4.1.1 *Trypanosoma vivax* strain used in this sequencing project

*Trypanosoma (Duttonella) vivax* belongs to the phylum *Kinetoplastida* family *Trypanosomatidae* (Rotureau and Van Den Abbeele, 2013). Strain (IL1392) is a clone derived from strain Y486 isolated from infected cows from Zaria in Nigeria (Leeflang, Buys and Blotkamp, 1976; Mooloo, Kutuza and Desai, 1987). This clone was developed to generate *in vitro* epimastigote as well as metacyclic stages of *T. vivax*; an approach that couldn't be achieved with the other strains of this trypanosome (D'Archivio *et al.*, 2011). Moreover, a transfecting metacyclic vector was developed in this strain able to infect small rodents (D'Archivio *et al.*, 2011), which make this strain an important model organism to study this species of trypanosome.

#### 4.1.2 *T. vivax* previous genome sequencing efforts

The studies on this trypanosome genome are scarce, however, studies on expression and purification of VSG genes were achieved (Gardiner *et al.*, 1987; Jackson *et al.*, 2012, 2015). The limited number of studies is related to the difficulties in culturing this microorganism (Gibson, 2012).

The previous studies on the karyotype of this organism showed relatively major differences from *T. brucei*, as it seems to have larger MBCs and it did not separate into obvious 11 MBCs (Van der Ploeg *et al.*, 1984). The presence and the number of possible mini-sized chromosomes was controversial (Van der Ploeg *et al.*, 1984; Dickin and Gibson, 1989).

The current and the only available genome sequence of *T. vivax* was based on Sanger sequencing technology with coverage of 5 times whole genome sequence of strain Y486 (Jackson *et al.*, 2012), which represents the ancestor of strain IL1392. The available draft genome sequence showed 11 pseudo-chromosomes in synteny to the genome sequence of *T. brucei*, and a large Bin sequence contain non syntenic contigs. However, this draft genome sequence is highly fragmented (12,800) contigs with many gaps and fragmented gene models.

### **4.1.3 Aims and objectives of this chapter**

The objective of this sequencing project is to deliver an improved version of the *T. vivax* genome sequence; to generate a *de novo* genome assembly with more physical contiguity in order to permit access to many important genome structures (especially those with repetitive nature); and to provide a genome sequence for a newly developed strain of *T. vivax* that has wide applications in the specialized laboratories. This project will have a big impact on the current efforts to understand the biology and tackle huge issues caused by this pathogen like its pathogenicity, life cycle and developing drug targets.

Accordingly, we have employed third generation PacBio SMRT sequencing technology to achieve a new *de novo* genome assembly of *T. vivax* strain IL1392.

## 4.2 Methods

### 4.2.1 *T. vivax* gDNA and PacBio SMRT libraries preparation

50 µg of high molecular weight of gDNA was prepared from the epimastigote form of strain IL1392, which is a Pasteur Institute clone of strain Y486 generously provided from Dr Bill Wickstead's laboratory/Queen's Medical Centre/ University of Nottingham. 25 µg was submitted to the Centre for Genomic Research (CGR)/ University of Liverpool to prepare 20 kb genomic DNA SMRT libraries. 11 SMRT sequencing RSII cells were primed with DNA libraries for the sequencing project.

### 4.2.2 Genome assembly databases and protein sequences used in this chapter

The genome databases of the reference sequence of *T.b. brucei* strain *TREU927* version 5.1 GeneDB (<http://www.genedb.org/Homepage/Tbruceibrucei927>) were used.

The contig level assembly was retrieved from the original Sanger Institute FTP site of *Trypanosoma vivax* strain Y486 sequencing project (<ftp://ftp.sanger.ac.uk/pub/project/pathogens/Trypanosoma/vivax>) while the annotation statistics were collected from the latest available version on TriTrypDB (<http://tritrypdb.org/common/downloads/>) version 34.

### 4.2.3 *De novo* genome assemblies

#### 4.2.3.1 SMRT portal analyses (Pacific Bioscience) built in assemblers

More than 5.8 Gbases of data were sequenced and the continuous-long-read (CLR) data were mapped and *de novo* assembled using the PacBio Hierarchical Genome Assembly process version 3 (RS\_HGAP\_Assembler. 3) and (RS\_HGAP\_Assembler. 2) (Chin et al., 2013) with diploid genome analyses option on and the target genome size was set to 52 Mbase (Mb), while the other parameters left in default. It was polished then with the Quiver software package; both tools were part of SMRT portal analyses.

#### 4.2.3.2 Locally installed assembler

Another long-reads SMRT sequencing genome assembler, CANU version 1.2 (Koren *et al.*, 2016), was used to generate TcIL3000 genome assemblies from the PacBio RSII SMRT sequencer raw reads “fastaq” output file by adopting two protocols:

First, run the CANU on generated filtered PacBio reads completely on default settings (CANUdefault) with targeted genome size set to 52 Mb.

Second, set the assembler to correct all reads and set the error rate to 0.035, the other parameters were left to default settings (CANU2).

#### 4.2.4 Assessment of gene models in *T. vivax* *de novo* assemblies

In order to compare between different assemblies in terms of gene model completeness, an assembly comparison approach was adopted using conserved core eukaryotic subset “protists” core genes implemented in BUSCO tools (Simão *et al.*, 2015) (see section 2.2.5.2)

##### 4.2.4.1 Chromosomal-level assembly

The standard highly contiguated, closest *T. brucei* TRUE 927 reference assembly was used as a reference to achieve chromosomal-level assembly according to synteny assignments.

Accordingly, ABACAS version 1.3.1. was used. The assembly contigs were reoriented, ordered and aligned to the MBCs of reference strain *T. b. brucei* strain TREU927 using protein based sequence homology. This was implemented in PROMmer algorithm (Delcher, 2002) of MUMmer package (Kurtz *et al.*, 2004) version 3 for the most accurate alignment to the reference chromosomes. ABACAS1 options used were “maxmatch” to increase the alignment sensitivity, “promer”: for amino acids based alignment algorithm, “-m”: to print the ordered and oriented contigs into a separate file, and “-b” generate a file containing the contigs not assigned to any of the inferred pseudomolecules (pseudochromosomes).

The output was then viewed using ACT (Carver *et al.*, 2005) to check the aligned scaffolds to the reference chromosomes.

#### **4.2.5 *T. vivax* IL1392 genome annotation**

COMPANION (Steinbiss *et al.*, 2016) automated protozoa genome annotation pipeline was adopted. Briefly, the pipeline automatically annotates protein coding genes and non-coding genes like tRNA, rRNA ncRNA...etc. It infers chromosomal level assembly using ABACAS 2 based on an available reference. The protein coding sequences were annotated on the new genome sequence by first, transferring reference protein coding sequences when they agreed with query genome possible ORFs using RATT (Otto *et al.*, 2011) (for this step “species” mode was used). Second, *ab initio* gene finding using AUGUSTUS (Stanke *et al.*, 2004) depending on protein evidence was applied (default threshold of 0.8 was adopted). Protein domain annotation was performed by two means: first, using “HMMR” (Johnson, Eddy and Portugaly, 2010) to search against Pfam database (Finn *et al.*, 2016). Second, by transfer of domains from reference proteins using “OrthoMCL” protein clustering tool (Li, Stoeckert and Roos, 2003).

For the protein non-coding gene annotation such as *tRNA*, *rRNA* and *ncRNA*, INFERNAL (Nawrocki and Eddy, 2013) uses HMM homology search and ARAGON (Laslett and Canback, 2004) algorithm of homology-based search query sequences based on prokaryotic, eukaryotic tRNAs consensus structure for *tRNA* genes were employed for this task.

Finally, the output files containing annotated features were produced in GFF3 format feature file, EMBL format and GAF file.

The locally installed COMPANION pipeline on CGR servers was used over the web based, since the latter has a genome size restriction, which could not annotate genomes larger than 62 Mb.

#### **4.2.6 BLAST search**

The BLAST search used in this chapter was applied using BLAST+ (Camacho *et al.*, 2009) package version 2.2.28.

#### **4.2.7 Genome visualization**

Genome inspection for gene model integrity and sequence layout for all investigation steps in this chapter were achieved using ARTEMIS version 16.0.0 (Kim Rutherford *et al.*, 2000; Carver *et al.*, 2012).

The sequence comparison tool ACT (Carver *et al.*, 2005) was employed for investigation the contig assignments to the reference *T. brucei*.

#### **4.2.8 Statistical analyses**

Student two-sample test was used to test statistical significance of the SSRs and DGCs on R-project (R Core Team, 2016) <http://www.R-project.org>.

#### **4.2.9 Strand Switch regions**

SSRs were viewed and their DNA sequences were extracted manually using ARTEMIS, while their length obtained according to the coordinates of the first flanking protein coding features. The conserved region on these regions were predicted by RepeatMasker package version 4.0.7. The parameters of “crossmatch”, repeat library search of ‘Trypanosoma brucei’ and ‘gff’ were allowed.

#### **4.2.10 Directional gene clusters**

The clusters of genes on *T. vivax* IL1392 contigs were viewed on ARTEMIS and the regions length were calculated according to the first and last annotated feature coordinates of each cluster.

#### **4.2.11 Clustering of proteomic databases of analysed kinetoplastids**

OrthoFinder pipeline version 1.0.0 (Emms and Kelly, 2015) was used for clustering all proteome data sets across all selected species. The rationale beyond using this clustering procedure is to deduce the sequences that shared similarity among the analysed data and clustering them into phylogenetically related clusters (orthogroups), leaving those sequences with no similarity to any other sequence as independent sequences (singletons) (for more details see section 2.2.9).

#### **4.2.12 Obtaining the number of shared genes from OrthoFinder output**

OrthoFinder provides statistics on clustered protein sets (for more details see section 2.2.9.1).

#### **4.2.13 Plots generation of possible genome rearrangements**

In order to view the predicted genomic regions such as chromosomal translocations between *T. brucei* and *T. congolense*, GenomeRibbon tools (Nattestad, Chin and Schatz, 2016) was used as follow:

##### **4.2.13.1 Synteny inference to the reference chromosomes**

Inspection and plot generation to show synteny of the assembly contigs to the reference Tb927 MBCs were generated using “mummerplot” tool implemented in MUMmer package version3 based on PROMmer comparison file output, after the “maxmatch” option was enabled and the minimum clustering option was set to 200 to increase the search sensitivity.

##### **4.2.13.2 Plots generated using GenomeRibbon tool**

A web based screening tool of large structural variants “GenomeRibbon” (Nattestad, Chin and Schatz, 2016) was applied in order to present possible shared regions between two different sequences, a tab delimited comparison file based on NUCMmer or PROMmer algorithms employed in MUMmer package version 3, which stores the related sequences and the positions of the predicted common regions were generated. Here, PROMmer algorithm was evoked with the “maxmatch” and the minimum clustering option was set to 200 to increase the search sensitivity. Then resultant files were uploaded to the web application and viewed on web browser.

##### **4.2.13.3 Viewing reference genes affected by the assumed chromosomal translocations in GenomeRibbon plots**

In order to show which reference genes are possibly affected by the proposed DNA segmental movements, an annotation file was uploaded to the application, which stores information about the reference features in “BED” file

format (<https://genome.ucsc.edu/FAQ/FAQformat.html#format1>). Six columns file stores required information was generated from the original “GFF” annotation file format of the reference *T.b. brucei* strain TREU 927 using appropriate LINUX command line. Then this file was uploaded to view the reference genes on regions of interest.

#### **4.2.14 Venn Diagrams**

The plots generated to view protein clustering analysis shared between assemblies or as assembly specific groups was generated using Rpackage “VennDiagram” (Chen and Boutros, 2011) on R-project using RStudio.1.1.3.

#### **4.2.15 Generation of assembly statistics**

The statistics such as the total size of the assembly, N50 size, GC%, max contig size, mini contig size and gap size of all analysed assemblies were calculated using a custom perl script.

#### **4.2.16 Gene Ontology enrichment analysis**

See section 2.2.10.

## **4.3 Results and discussion**

### **4.3.1 SMRT sequencing and *de novo* Genome assembly**

#### **4.3.1.1 RSII SMRT sequencing of *T. vivax* strain IL1392 gDNA**

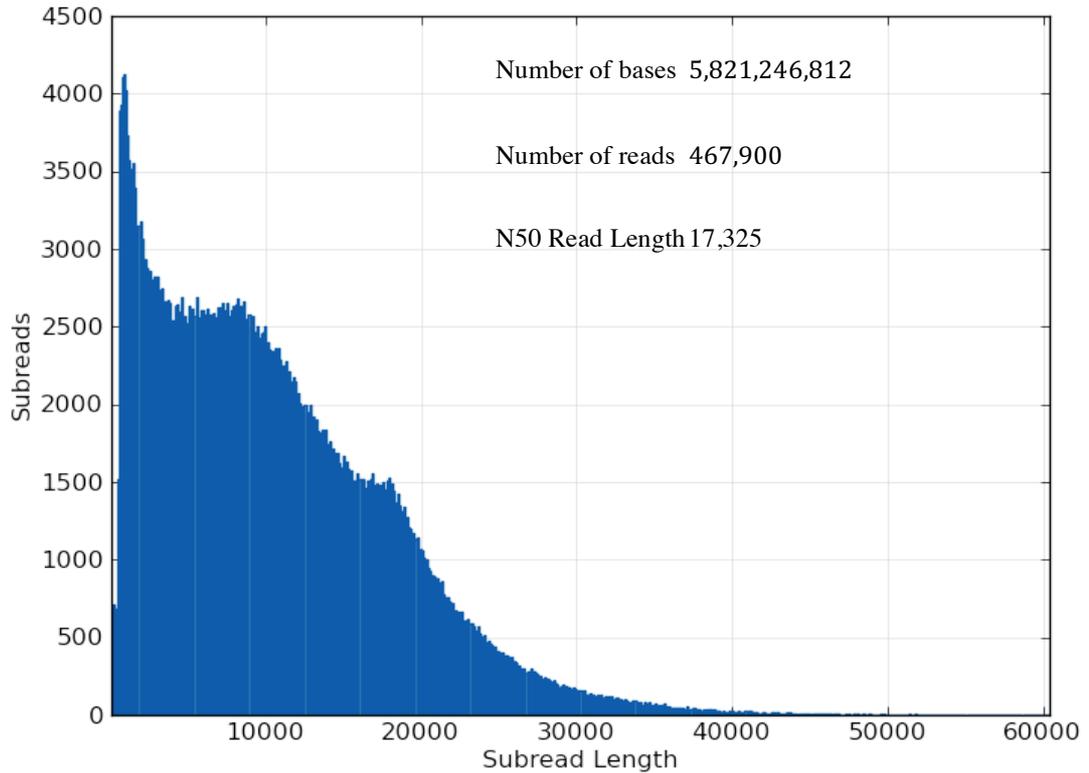
A total of 5.8 Gbases of data were generated from the 11 RSII SMRT cells. The average length of input reads was 12.4 kb and the read N50 length was 17.3 kb (Figure 4.1). The sequence reads' length distribution provided high prospects towards the generation of long continuous contigs, especially for N50 read length and presence of a number of super-long reads up to 50 kb, which could facilitate unveiling of genomic regions of repetitive nature (Figure 4.1).

#### **4.3.1.2 Genome assembly using assemblers in PacBio SMRT portal analysis pipeline**

The sequence reads were subjected to two assemblers of PacBio SMRT portal analysis pipeline (Table 4.1):

First, HGAP assembler version two (HGAP2): This assembler was able to construct a 67.8 Mb *de novo* genome assembly of 773 contigs, with maximum contig length of 2.8 Mb and the contig N50 length was 261 kb. The estimated coverage average was 92.7 fold.

The second assembly algorithm used was HGAP3, which resulted in assembly with different assembly statistics from the previous assembler from the same PacBio RSII SMRT sequencing reads. The number of contigs went up to 876 and the same trend was also noticed with assembly size (74.5) Mb. However, the N50 contig length and maximum contig length showed apposite trend (240 kb, 2.3 Mb), respectively (Table 4.1).



**Figure 4.1** Distribution of sequenced reads sequenced from DNA libraries input prepared from gDNA of *T. vivax* strain IL1392. Showed data size of 5.8 Gb and an average read length of 12.4 kb with N50 read length of more than 17 kb.

#### 4.3.1.3 Stand-alone assemblies using CANU assembler

Two assembly parameters were adopted using this tool. Using CANU with default parameters (CANUdefault), and the other attempt was achieved by manipulating parameters to permit for correcting all PacBio reads and permitting the corrected reads to be assembled (CANU2) (Table 4.1).

As in previous assemblers, these parameters yielded different assembly statistics. The N50 contig length was higher in the assembly generated by applying default parameters of CANU. However, the longest contig length of CANU2 assembly and the overall assembly size showed higher figures when compared to the former CANUdefault assembly (Table 4.1). The huge differences in the assembly statistics using different algorithms, or even the

same algorithm but with different parameters, had a vast effect on the final assembly. One reason behind these differences is that the assembler tries to be more conservative by removing the reads most likely to be of repetitive nature in order to achieve a *de novo* assembly with the least number of contigs and higher N50 contig length. We noticed that the size of unassembled reads of CANUdefault is three times higher than those of CANU2 (35,137; 11,363), respectively. The reads contain potential repeat sequences often to blame for the assembly fragmentation or even mis-assembly (Ekblom and Wolf, 2014; Lian *et al.*, 2014).

The presence of such discrepancies among the generated *de novo* assemblies motivated us to devote further investigation towards each assembly and its feasibility for downstream analysis.

**Table 4.1 *T. vivax* gDNA PacBio *de novo* assemblies, assembly statistics of different assemblers used to generate contig level assemblage.**

	HGAP2	HGAP3	CANU default	CANU2
<b>Number of sequenced bases</b>	5,821,246,812	5,821,246,812	5,821,246,812	5,821,246,812
<b>N50 read length (bp)</b>	17,325	17,325	17,325	17,325
<b>Assembly size (bp)</b>	67,823,889	74, 587,772	49,070,302	51,146,339
<b>Number of contigs</b>	773	876	264	397
<b>Mini contig length</b>	2,953	797	3,475	1,892
<b>Max contig length (bp)</b>	2,897,905	2,319,116	3,213,402	3,385,160
<b>N50 contig length (bp)</b>	261,291	239,911	426,174	240,064
<b>GC %</b>	53.63	53.85	52.51	52.84

#### 4.3.1.4 Choosing the best contig assembly for downstream analysis

The previous section highlighted an apparent huge variation among different assembly algorithms and parameters manifested by variable assembly statistics among assemblies as shown by Table 4.1.

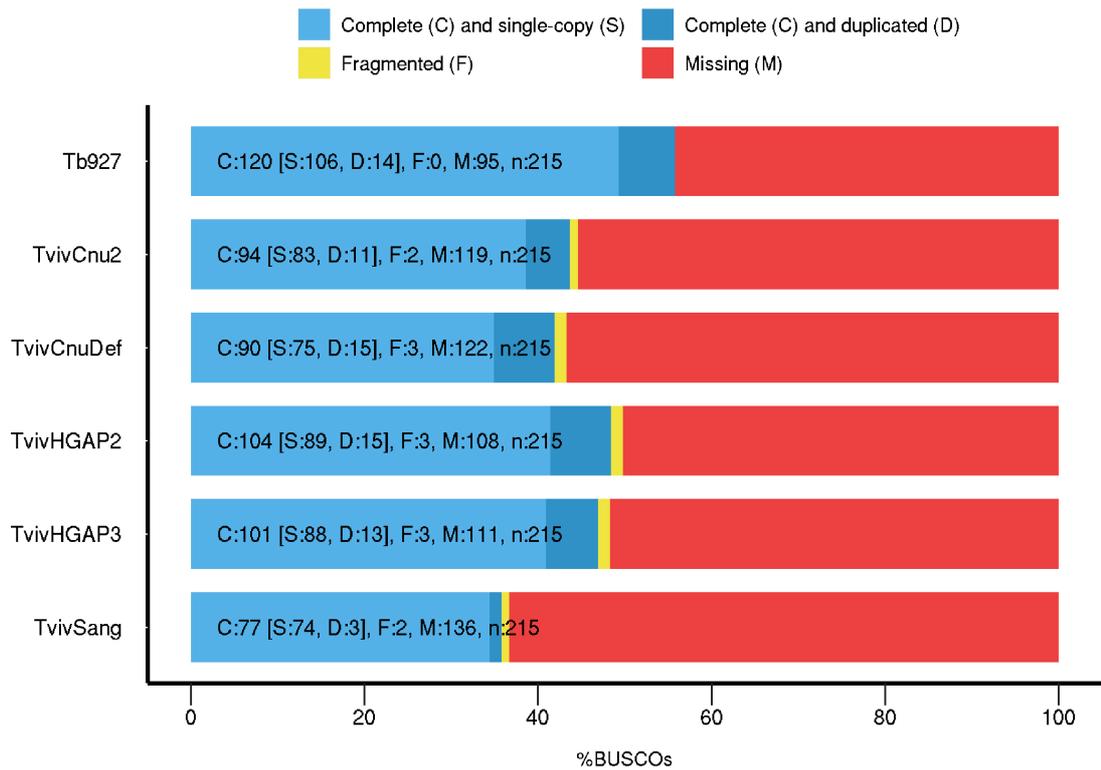
Although the numbers showed a favour towards CANU assembly with default parameters, our previous experience with *T. congolense* PB assembly (Chapter two) showed that these parameters yielded smaller assembly size and unsatisfying gene models, which is also manifested in the *T. vivax* genome project.

Taken together, choosing a tool to assess the gene models besides the assembly parameters could be a good option to aid in choosing the most successful assembly. Thus, the BUSCO comparison tool was adopted to examine the integrity of gene models of each assembly.

Hence, the HGAP assemblies showed better results than the CANU algorithm-based assemblies, with the best showed by the assembly generated by HGAP2, which expressed the highest number of predicted complete gene models (C:104). The same tendency was noticed for the single models (S:89) accompanied by the least missing models among the other assemblies (M:108) (Figure 4.2).

BUSCO analysis results also suggested that all different PacBio assemblies exhibited improved figures over those of Sanger based assembly (current available *T. vivax* sequence) (Figure 4.2).

In a nut shell, the previous data analysis of different *T. vivax* PB assemblies suggesting perhaps a good candidate to be used for the further gene annotations and comparative genomic analyses, as the gene model quality most likely to affect the annotation process and consequently downstream analyses. Therefore, HGAP2 PB assembly was chosen to be the assembly of choice.

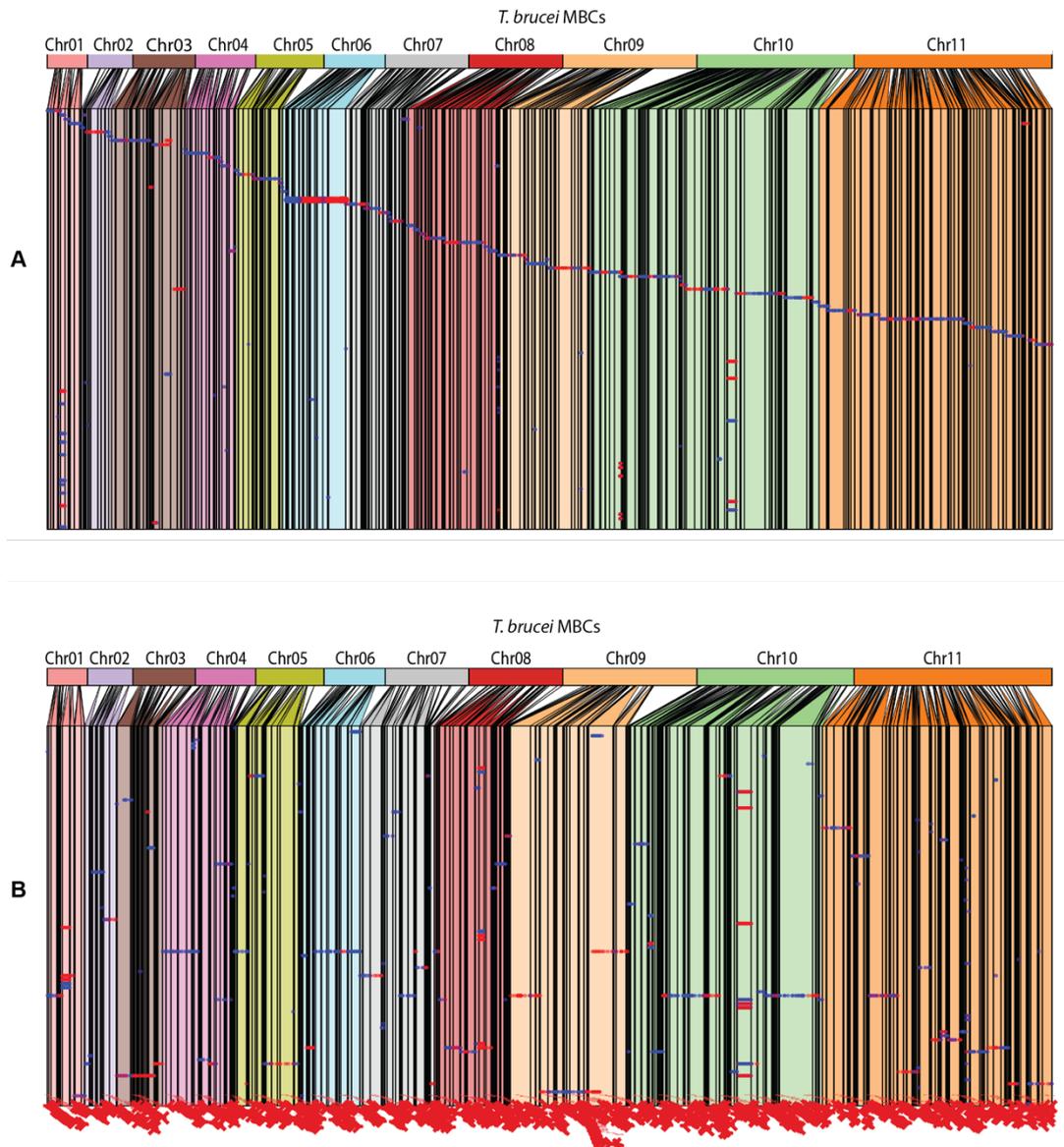


**Figure 4.2 Gene model integrity assessment using core protists gene set applied by BUSCO tools across our PB assemblies of *T. vivax*.** Two CANU assemblies were compared, the first with options set to the correction of all reads and allow to assemble corrected reads (TvivCnu2) and the second CANU assembly done using the default parameters (TvivCnuDef). The other two were HGAP based assemblies [HGAP2 (TvivHGAP2) and HGAP3 (TvivHGAP3)]. The plot highlighted that the HGAP2 assembly showed the best BUSCO stats in terms of the highest number of complete gene models, single genes and the least missing models among the other three *T. vivax* assemblies. Moreover, all PacBio-based assemblies showed better scores than currently available *T. vivax* reference Sanger based assembly. Tb927 genome was used as a standard reference for comparison.

#### 4.3.1.5 Chromosomal-level assembly

The ABACAS contig consignments of *T. vivax* PacBio contigs were checked manually using ACT. The *T. vivax* PacBio contigs showed highly conflicted assignments across different reference MBCs, which make it difficult to appoint certain contigs to specific *T. brucei* MBC, because some long contigs showed synteny to more than one *T. brucei* MBCs, suggesting a probability of large structure rearrangements on chromosomal level between the two species. A similar issue did not arise from the *T. congolense* PacBio genome project as it showed relatively better synteny to the reference chromosomes (Figure 4.3).

Such possible structural translocation would be more obvious in the vicinity of long contigs generated by PacBio SMRT sequencing technology that could reserve more structural integrity, which in turn might harbour more structural variants than the reference (Wu *et al.*, 2017). These putative across species interchromosomal translocations were discussed in detail in section (4.3.4.2).



**Figure 4.3 Genome synteny of PacBio *de novo* assemblies' contigs of *T. congolense* (A) and *T. vivax* (B) to the reference *T. brucei* MBCs. A) *T. congolense* PacBio assembly contigs showed linear alignment to the *T. brucei* MBCs indicating a relative high synteny. B) *T. vivax* PacBio assembly contigs exhibit a scattered distribution, with contigs spanning across different *T. brucei* MBCs referring to less synteny and a distantly related genome. The plots were generated using GenomeRibbon tools.**

### 4.3.2 *T. vivax* PB genome annotation

Noteworthy, PB assembly statistics exhibited higher figures than these of the current draft sequence of *T. vivax*; the number of contigs was highly reduced from 12,283 to 773. The size of largest PB contig is 2.8 Mb in comparison to 55.9 kb of Sanger sequence available assembly. The inferred genome size was larger for PB HGAP2 sequence assembly (67 Mb in comparison to 52 Mb of Sanger sequence assembly) (Table 4.2).

The COMPANION pipeline predicted genomic features on the PacBio HGAP2 assembly of *T. vivax* IL1392 suggested a higher number of genes than the currently available database for *T. vivax* Y486 (18,466, 12,050), respectively. This could be due to the larger assembly size and the noticeable tandem repeated genes of some gene families, which suggest segmental genome duplication that could be missed from the current Sanger based assembly. While the sequences of approximately 2,563 fragmented gene models in Sanger assembly comprise, such features interrupted by physical, within frame gap/s, the PB assembly exhibited absence of such gaps in the predicted genes, yet it has a larger number of genes.

The proposed sequence contiguity could be a reflection of the relatively longer sequenced reads as shown in previous section, and as has been shown by the assembly quality examination of gene model's integrity by testing of conserved eukaryotic genes among different assemblies using BUSCO tool (Figure 4.2), this might have permitted more completed gene models.

The remarkable difference in protein coding gene count between the two assemblies might be linked to the ability of long reads to span over tandemly repeated gene families; a characteristic of kinetoplastid genomes (Berriman, 2005; Ivens *et al.*, 2005; Jackson *et al.*, 2016), which might be represented more extensively in this trypanosome. Furthermore, in other parasitic and non-parasitic organisms, the increase in the number of genes and consequently the genome size is most likely correlated to the adaptation to different environments (Bentkowski, Van Oosterhout and Mock, 2015)(Sundberg and Pulkkinen, 2015).

**Table 4.2 Comparison between PB HGAP2 based assembly and Sanger assembly of *T. vivax* using different assembly and annotation parameters.**

Genomic feature	HGAP2 assembly	Sanger assembly
<b>Total assembly size</b>	67,823,889	51,873,811
<b>Number of contigs</b>	773	12,283*
<b>Size of the largest contig (bp)</b>	2,897,324	55,954*
<b>Size of the smallest contig (bp)</b>	2,953	1,001*
<b>N50 contig length</b>	261,249	6,294*
<b>GC% content</b>	53.64	51.95*
<b>Percentage of unresolved bases (Ns)</b>	0	0.55*
<b>Total number of genes</b>	18,466	12,050
<b>Protein coding genes</b>	18,229	11,004
<b>Number of genes with assigned function</b>	6,609	4,408
<b>Pseudogenes</b>	196	656
<b>Fragmented genes</b>	0	2,563
<b>tRNA genes</b>	268	393
<b>rRNA genes</b>	148	350
<b>snRNA</b>	9	2
<b>snoRNA</b>	34	4

\*Numbers calculated from the contig level assembly at Sanger Institute FTP site (<ftp://ftp.sanger.ac.uk/pub/project/pathogens/Trypanosoma/vivax>).

### **4.3.3 Assessment of *T. vivax* IL1392 PacBio genome annotation and validation**

#### **4.3.3.1 Protein sequences cluster analysis**

The protein sequences predicted by the COMPANION annotation pipeline along with these of Tb927 and *T. vivax* Sanger assembly available on TriTrypDB database were clustered using the OrthoFinder version 1.0.0. algorithm.

The clustering results are summarised in Figure 4.4. A sum of 150 gene families of *T. vivax* PacBio proteomic data were predicted as new to this species and shared with *T. brucei* proteome and 8 groups were suggested to be *T. vivax* PacBio specific groups. However, the resulted clustering analysis revealed a high number of orthogroups of Sanger proteome that were shared with the Tb927 protein sequences (n= 991) and Sanger assembly specific groups (n= 4).

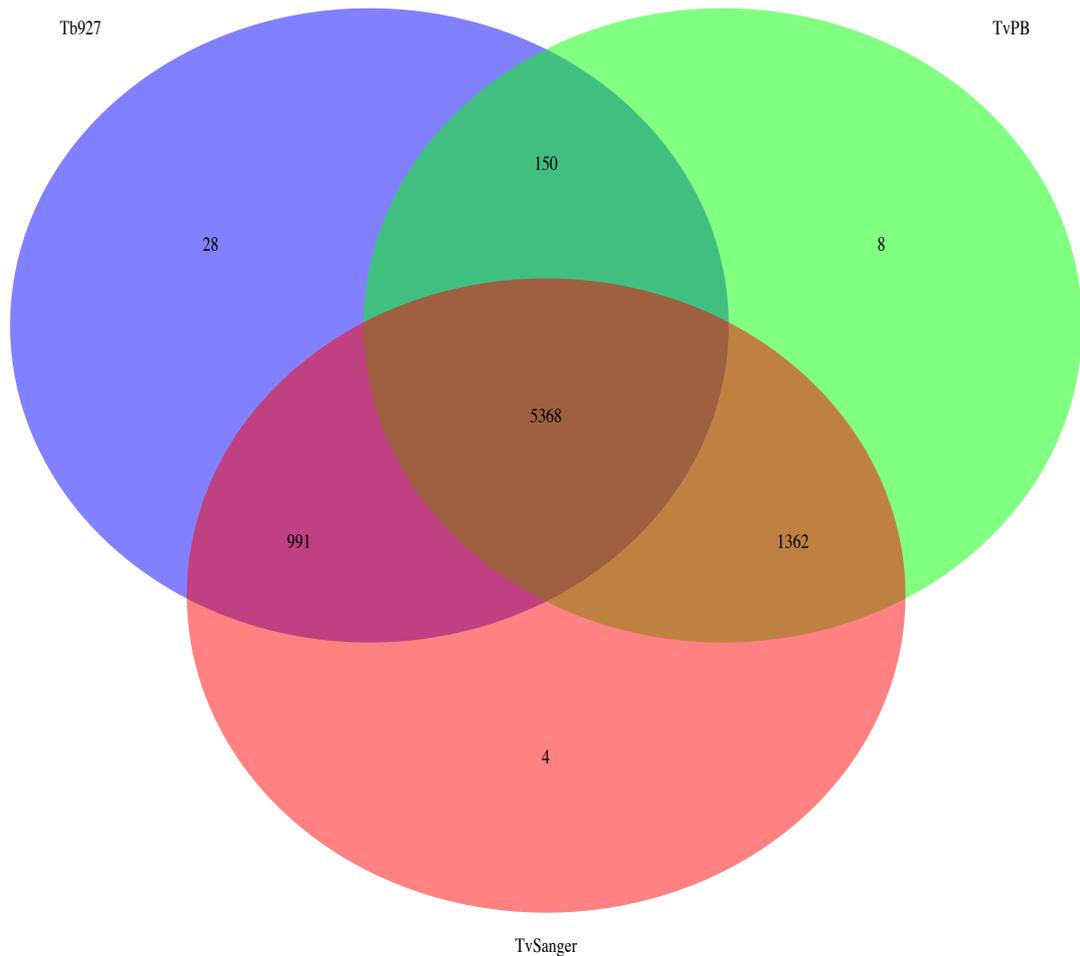
In order to investigate whether *T. vivax* PacBio missing proteins was due to the annotation limitation or an issue in the assembled sequence? First of all, the BUSCO eukaryotic conserved genes proposed a higher score for our draft assembly in comparison to the current draft assembly. While the annotation of available reference sequence was done by extensive manual intervention, the annotation of PB the assembly was achieved using automatic annotation pipeline without further manual intervention.

Thus, to investigate this difference in the annotation two approaches were adopted:

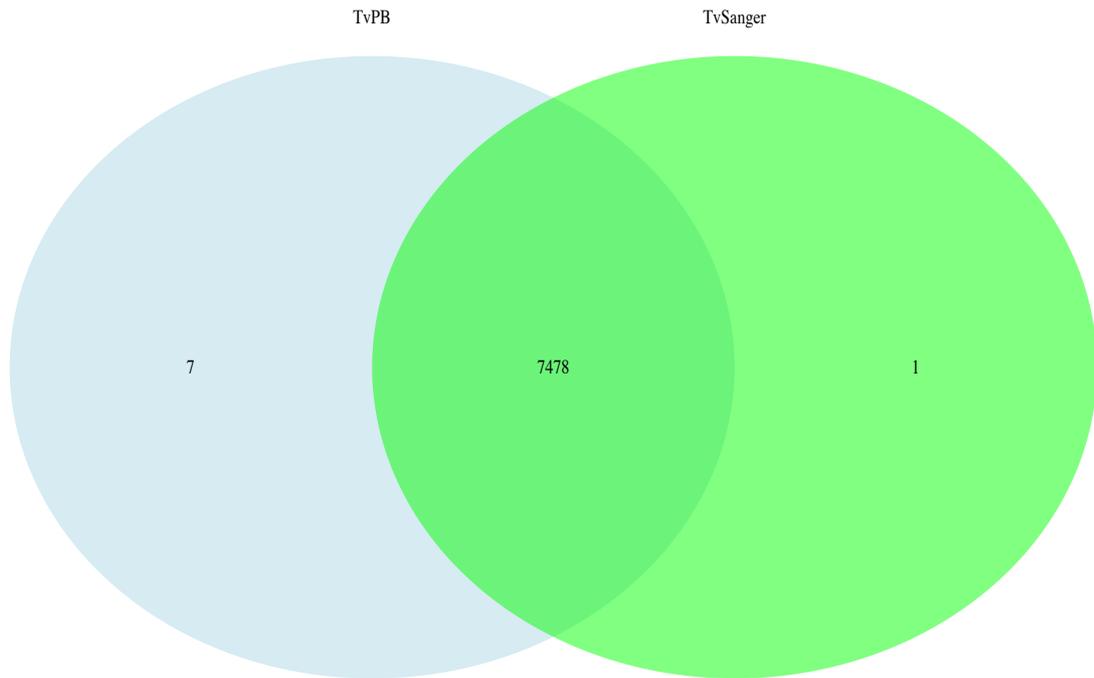
First, local COMPANION annotation process with the same parameters that was applied in PB assembly was applied to *T. vivax* Sanger assembly. This could reduce the probable bias from different annotation methods used to annotate both *T. vivax* assemblies (PB assembly and Sanger assembly). The proteomic data generated by the pipeline were put together with that of PB

assembly, then it was clustered using the protein clustering tool (OrthoFinder) (Figure 4.5).

This approach spotted what was expected and suggested by assembly quality assessment. As most of the gene families from both assemblies were shared and 7 orthogroups (contain 46 genes) were unique to our assembly, two of these groups involved predicted genes as retrotransposons hot spot 4 (RHS4) and the rest were gene families having genes of unassigned function. However, only one orthogroup (consisting of seven members) of *T. vivax* Sanger assembly clustered alone as specific for the latter assembly. A closer look at these proteins revealed that they were annotated as hypothetical proteins and the BLASTp analysis against TriTrypDB database, resulting in hits to fragmented protein sequences of unknown function of *T. vivax*. These results suggested a more likely reason to put them in single group, being fragmented or incomplete sequences.



**Figure 4.4 A VENN diagram of ORTHOFINDER protein clusters of Tb927, *T. vivax* PacBio and *T. vivax* Sanger assembly proteomic databases.** *T. brucei* strain 927(Tb927), blue area has 28 specific gene families, while *T. vivax* has 1,374 specific gene families as predicted from both assemblies. *T. vivax* PacBio assembly (green area) has 8 assembly specific groups of genes and a 150 possible new gene family shared with Tb927 genes. The current *T. vivax* assembly (*T. vivax* Sanger assembly) red area proposed 4 assembly-specific orthogroups and 991 gene families that are shared with members of Tb927 assembly. The clustering algorithm also suggested a number of 5368 orthogroups shared among the three assemblies.



**Figure 4.5 A VENN diagram of ORTHOFINDER protein sequence clustering of *T. vivax* PacBio proteome and the currently available Sanger sequence assembly proteome were both annotated using locally installed COMPANION annotation pipeline.** It proposed only one orthogroup is specific for Sanger assembly (green area), while seven OGs are PB specific (light blue region) and about 7478 shared groups.

Second, we searched the missing proteins from *T. vivax* Sanger assembly against PacBio DNA assembly using BLASTx sequence search with expectation value set to  $1e^{-10}$ . The *T. vivax* Sanger assembly 991 orthogroups consist of 1008 genes, most of which were one to one orthologues with corresponding Tb927 protein sequences. The amino acid sequences of these genes were extracted using relevant command lines on LINUX machine and prepared as a protein database for a BLASTx search against *T. vivax* PacBio assembly. The search resulted in 1,004 non-redundant hits to our draft assembly (Figure 4.4). However, there were four missed protein sequences that have the following IDs (TvY486\_0805550, TvY486\_0805560, TvY486\_0805570 and TvY486\_1004175), presenting short amino acid sequences as follows (three exact copies of protein of unknown function of

length (34aa) and clustered in one group and the fourth was a ribosomal protein (28aa)). Then those sequences were searched against sequencing reads and the non-redundant hits showed significant hits from raw PacBio reads towards these potentially missing sequences from the assembly.

A similar approach was applied to examine the presence or absence of 190 genes in the current *T. vivax* Sanger based assembly lying in the 150 orthogroups shared between the *T. vivax* PacBio assembly and the Tb927 genome (Figure 4.4). The BLASTx search resulted in 157 non-redundant hits to Sanger genome assembly and 33 were absent. 12/33 *T. vivax* PacBio have putative functions assigned to them (five copies of Regulated-SNARE-like domain containing protein, GTP1/OBG/50S ribosome-binding GTPase/Ferrous iron transport protein B, two eukaryotic translation initiation factor eIF2A, Pyridoxal-dependent decarboxylase conserved domain containing protein, and PPPDE putative peptidase domain containing protein).

#### **4.3.4 Possible new findings in this *T. vivax* PacBio assembly**

##### **4.3.4.1 *T. vivax* PacBio single genes**

Protein clustering analysis suggested 7,292 protein sequences that could not have orthologous genes in the three analysed assemblies as they clustered independently (singletons). 6,444 were hypothetical sequences and the remaining 848 genes have putatively assigned functions. 441/848 presented Gene Ontology annotations of 963 of GO term IDs allocated to them by the annotation pipeline.

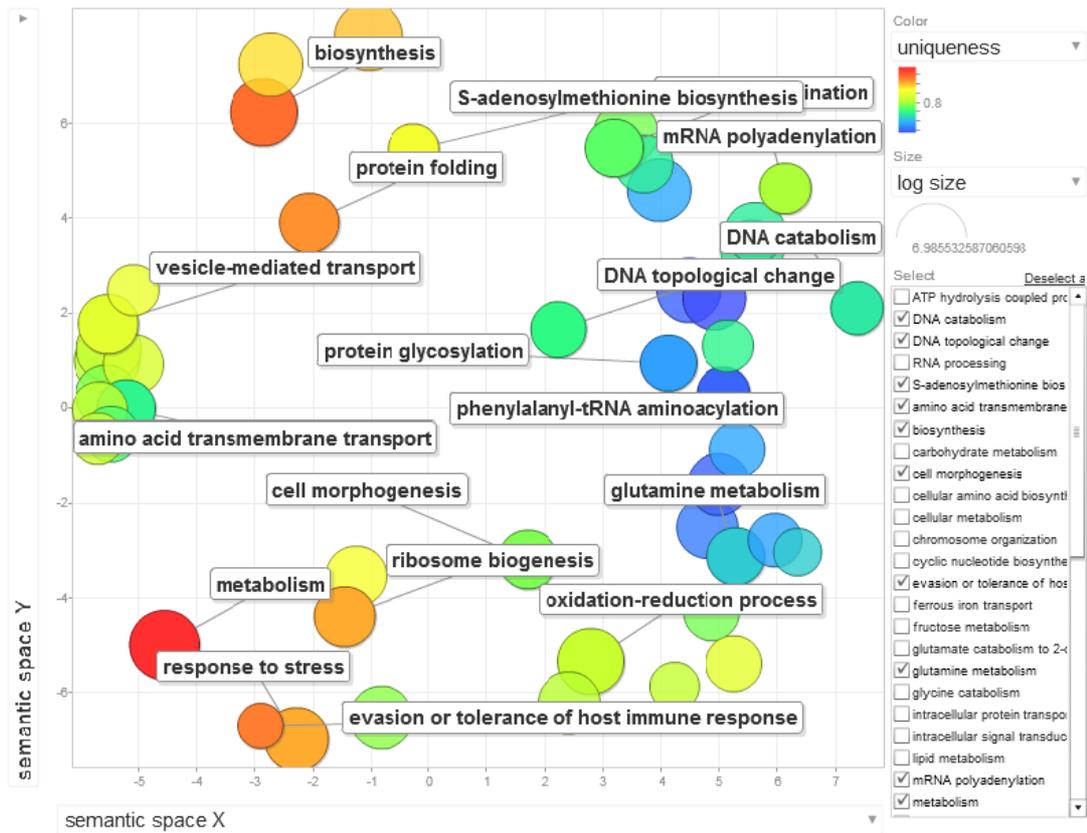
GO terms enrichment analysis by REVIGO was achieved. The analysis suggested high uniqueness values toward pathways involved in protein folding, biosynthetic process, transportation, host immune evasion and ribosome biosynthesis (Figure 4.6).

The possible molecular functions of the GO terms analysis showed in (Figure 4.7) revealed higher specification for ribosomal structural components, phosphopyruvate hydratase enzyme, signal transduction, structural

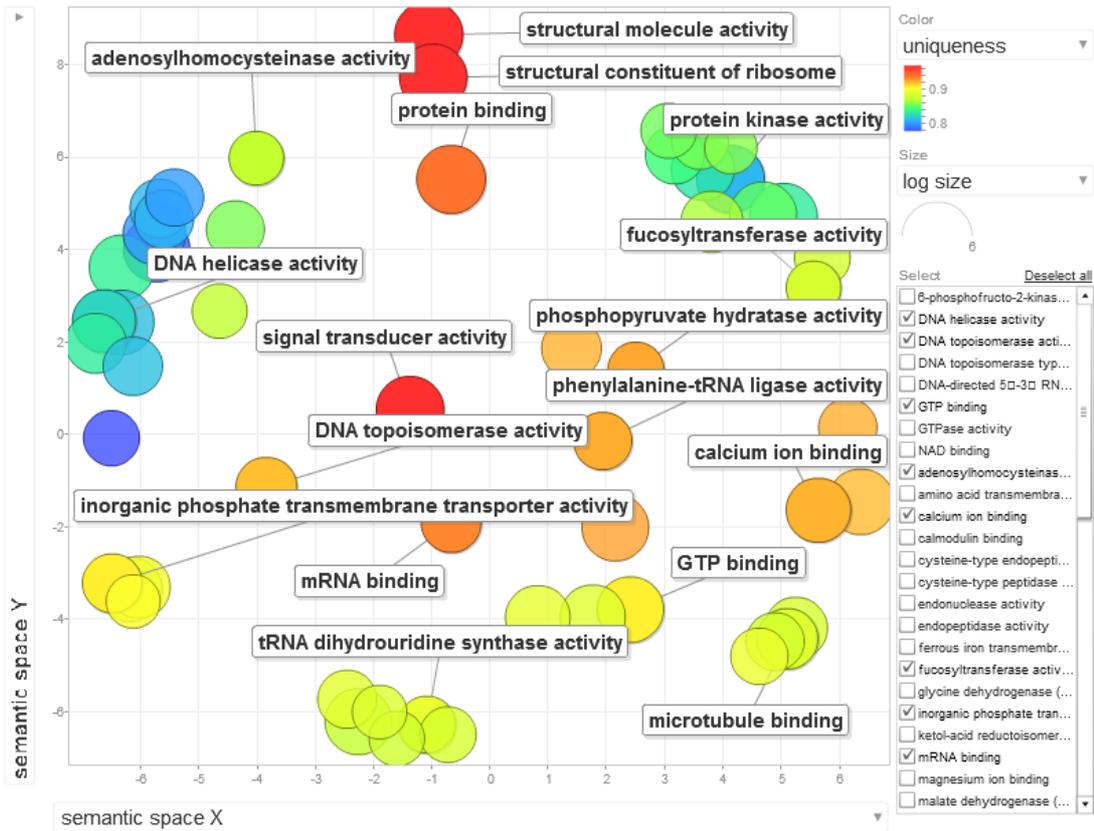
molecules, cellular transmembrane inorganic phosphate transporter and protein binding activity.

The analysis also shines a light on possible cellular localisation of protein products of these genes as shown in Figure 4.8. The ontology terms referred to cell component localisation that showed high speciation specificity and no dispensability are those part of cellular membrane and motile cilium.

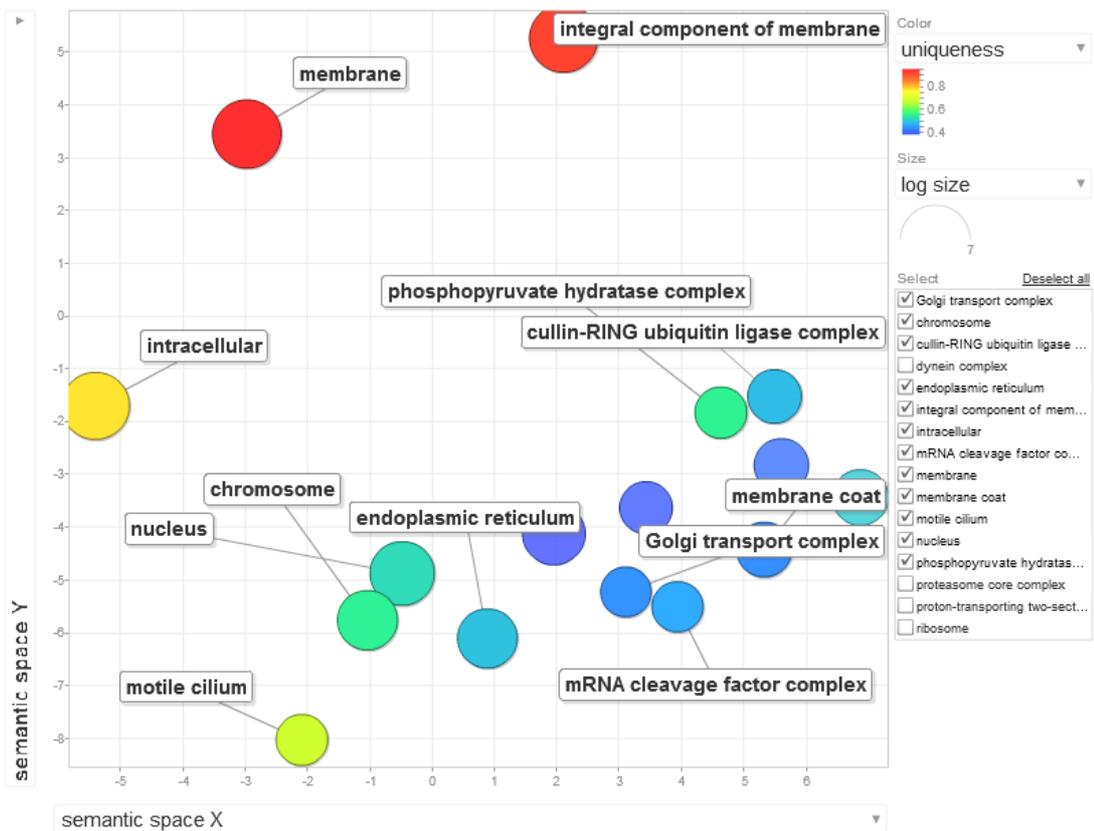
This analysis, along with structural integrity provided by our genome project, suggested new possible pathways or genes involved in metabolic, nutrition and host pathogenicity, which might have particular importance towards drug discovery, vaccine design and more understanding of the biology of this important animal pathogen.



**Figure 4.6** Semantic gene ontology enrichment analysis plot of *T. vivax* PacBio singletons protein sequences shows biological pathways. The redder the colour, the more uniqueness of the term from the other terms. The circle size depends on the number of proteins that have term.



**Figure 4.7 Semantic gene ontology enrichment analysis plot of *T. vivax* PacBio singletons protein sequences shows molecular functions. The redder the colour the more uniqueness of the term. The sphere size depends on the number of proteins sharing the same term.**



**Figure 4.8** Semantic gene ontology enrichment analysis plot of *T. vivax* PacBio singletons protein sequences shows cellular localization. The redder the colour the more uniqueness of the term from the others. The circle size depends on the number of proteins sharing the same term.

#### 4.3.4.2 Predicted massive inter-chromosomal translocations

Trials to infer chromosomal level assembly using ABACAS1 tool could not resolve explicit distinct *T. vivax* MBCs putative territories according to the protein translation of synteny DNA regions with the reference *T. brucei* due to conflicting multiple assignment of *T. vivax* PacBio contigs among different reference MBCs.

In order to investigate whether this conflict in assignment of contigs of PacBio assembly to the reference was due to inter-chromosomal rearrangements or not, we hypothesized that the longest contigs have more probability to show possible rearrangements if present, as the contigs with physical continuity might provide more structural information. Accordingly, the top two largest

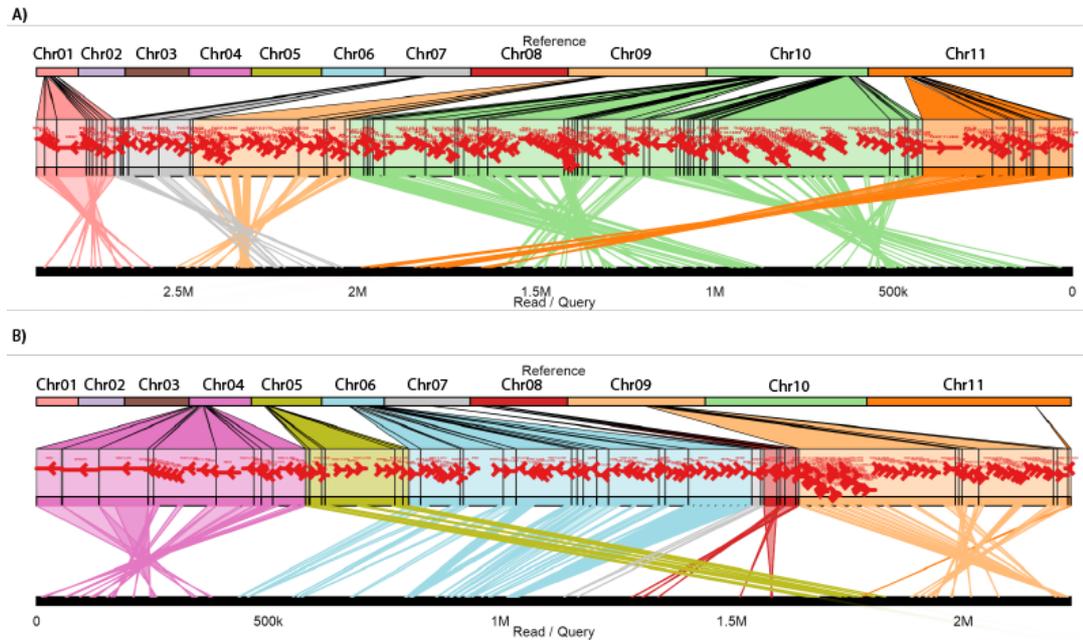
contigs (2.9 Mb, 2.3 Mb) with contig IDs of scf718000002168, scf718000002091 respectively, were chosen for investigation.

Briefly, the largest contig of the assembly suggested synteny regions with a number of pseudochromosomes (1, 7, 9, 10 and 11) of the reference strain Tb927 genome assembly (Figure 4.9). The segment that shares gene sequences with Tb927 chromosome one, stretches over 300 kb, more than 200 kb, 200 kb, 350 kb and over 1.5 Mb, respectively.

The second largest contig showed regions in common with Tb927 MBCs (4, 6, 7, 8, 5 and 9). The largest region spread over 500 kb on the latest contig presented synteny to the reference MBC six.

Such previously unprecedented putative chromosomal translocations showed by the above briefing suggested large-scale interchromosomal translocations in *T. vivax* genome in comparison to the reference chromosomes.

*T. vivax* assembly contigs that showed synteny to the reference *T. b. brucei* TREU927 MBCs as predicted by PROMmer algorithm are shown in Table 4.3.



**Figure 4.9 GenomeRibbon plot of the longest two contigs of *T. vivax* PacBio assembly showed regions of synteny with reference genome assembly of Tb927 version 5.1 permitting illustration of predicted coordinates by PROMmer. A) Ribbons coloured according to the colour of reference chromosome. The longest contig scf718000002168 (2.9 Mb) of *T. vivax* PacBio assembly was proposed to harbour chromosomal regions shared with different chromosomes (1, 9, 7, 11 and 10) of the reference sequence. B) Second max contig scf718000002091 (2.3 Mb) suggested to have chromosomal translocations in comparison to the chromosomes of the reference genome of Tb927 of (chr04, chr06, chr7, chr8, chr05 and chr9), furthermore, shared regions with (chr04, chr08 and chr09) revealed inverted segments in *T. vivax* PacBio contig.**

**Table 4.3 *T. vivax* contigs that showed regions of synteny to the reference *T. brucei* TREU927 genes. First 30 are shown, for the complete table (see appendix C1).**

Tb927 MBC	Tb927 reference genes	T. vivax contig ID	Percent of identity to the reference
Tb927_01_v5.1	ALAT	scf7180000002472	65.76
Tb927_01_v5.1	COXIV	scf7180000002472	77.57
Tb927_01_v5.1	KREL2	scf7180000002124	79.01
Tb927_01_v5.1	KREN1	scf7180000002168	71.07
Tb927_01_v5.1	MDN1	scf7180000002168	60.02
Tb927_01_v5.1	MEAT1	scf7180000002168	84.12
Tb927_01_v5.1	MRPS15	scf7180000002168	87.84
Tb927_01_v5.1	PF16	scf7180000002124	94.06
Tb927_01_v5.1	PGI	scf7180000002472	76.97
Tb927_01_v5.1	PGKA	scf7180000002168	75.41
Tb927_01_v5.1	PGKA	scf7180000002168	75.78
Tb927_01_v5.1	PGKB	scf7180000002168	80.06
Tb927_01_v5.1	PGKB	scf7180000002168	81.06
Tb927_01_v5.1	PGKC	scf7180000002168	78.23
Tb927_01_v5.1	PGKC	scf7180000002168	80.47
Tb927_01_v5.1	RPC128	scf7180000001780	91.19
Tb927_01_v5.1	RPC128	scf7180000002168	88.84
Tb927_01_v5.1	RPC128	scf7180000002168	90.47
Tb927_01_v5.1	Tb927.1.1380	scf7180000002168	79.93
Tb927_01_v5.1	Tb927.1.1560	scf7180000002168	63.89
Tb927_01_v5.1	Tb927.1.1930	scf7180000002168	74.58
Tb927_01_v5.1	Tb927.1.2330	scf7180000002035	100
Tb927_01_v5.1	Tb927.1.2330	scf7180000002035	85.86
Tb927_01_v5.1	Tb927.1.2330	scf7180000002035	91.67
Tb927_01_v5.1	Tb927.1.2330	scf7180000002035	91.85
Tb927_01_v5.1	Tb927.1.2330	scf7180000002035	92.37
Tb927_01_v5.1	Tb927.1.2330	scf7180000002035	93.88
Tb927_01_v5.1	Tb927.1.2330	scf7180000002035	99.44
Tb927_01_v5.1	Tb927.1.2330	scf7180000002035	99.77

Over all, this analysis suggested novel inter-species large structural chromosomal rearrangements, affecting the MBCs of *T. brucei* in comparison with the *T. vivax* genome and to a lesser extent, in two strains of *T. congolense* (have been shown in chapter two). This work showed that across species, genomic exchange widely affected genes of housekeeping nature, involved in a wide spectrum of roles in trypanosomes cell biology like structural proteins, transcriptional roles, metabolic pathways, transportation and cell surface proteins.

The link between genomic translocations and speciation in animals has been established earlier (White, 1969). Moreover, the role of chromosomal rearrangements especially chromosomal inversions, in speciation, phenotypic variations and adaptation to different environments have also been advocated in other eukaryotes like plants (Lowry and Willis, 2010) , and insects like *Anopheles* (Ayala *et al.*, 2011).

Genomic structural rearrangements were likely associated with local adaptation and evoke regions of genomic loci of divergence (Yeaman, 2013). Similar chromosomal inversions is correlated with geographical distribution and phenotypic difference among different higher eukaryotic species (Ullastres *et al.*, 2014; Berg *et al.*, 2016; Zhao *et al.*, 2016). Hence, the predicted genomic inversions in *T. vivax* could be highly linked to its unique life cycle and wider host range in both vertebrate and invertebrate hosts.

These findings also suggested similar translocation and conversions in *T. brucei* MBCs one and three, in a manner almost consistent with those noticed in both strains of *T. congolense*. The predicted consistent rearrangement in *T. congolense* and *T. vivax* affected MBC one in comparison to *T. brucei* MBC one might also propose that the ancestral trypanosomatids might have had a smaller chromosome one as it seems that these translocations have added segments most likely to the internal core region more likely towards *tubulin* gene cluster of the MBC one of *T. brucei*.

Such genomic differences might have shaped the phenotypic differences and the differences in life cycle among African trypanosomes. Here we propose

that the predicted genomic shifts might largely mirror the underlining differences in the life cycle, morphology and host range among African TriTrypanosomes. Remarkably, the degree of the proposed genomic rearrangements is highly mimicking the position of African trypanosomes in the phylogenetic tree with the most extensive case was noticed in *T. vivax* (which diverged earlier and relatively distantly related) to the moderate degree noticed in *T. congolense* (chapter two) that showed closer positioning towards the most recently diverged *T. brucei* (Haag, O'hUigin and Overath, 1998; Stevens and Rambaut, 2001; Stevens *et al.*, 2001; JACKSON, 2015).

The chromosomal structural rearrangements were also observed among different strains in another protozoa like *Giardia intestinalis* (Tůmová *et al.*, 2016).

Our analyses claimed that most of these collinearity breakages happened on strand switch regions, which in most cases contain sequences of repeated nature tRNA or transposable elements. These regions in the *T. vivax* genome will be described in more detail in the next section.

#### **4.3.4.3 Strand switch regions**

The importance of these DNA loci could lie in some features present within their sequences that could favour transcription initiation (DSSRs) or transcription termination (CDGCs). The contigs larger than 500 kb were considered to search for these possible inter-genic clusters as the probability to reveal complete regions is high in long contigs.

Total of 82 SSRs (41 DSSRs and 41 CSSRs) with notably longer sequences of DSSRs compared to CSSRs (means= 10,636, 3,625) bp was observed, respectively. *Two sample t.test* revealed statistical significance of this parameter under  $p < 0.05$  ( $p\text{-value} = 9.028 \times 10^{-14}$ ,  $df = 81$ ). Furthermore, the DSSRs showed more tendency to harbour *Ingi2* elements (Table 4.4). To examine the effect of the presence of *Ingi2* on the length of the SSRs, *one sample t.test* was carried out on the means of DSSRs with and without *Ingi2* and the same approach was performed on CSSRs. In both cases, the presence of the transposable element in SSRs has statistically significant

effect towards the increase of the region length p-value ( $2.632 \times 10^{-11}$ ,  $2.576 \times 10^{-5}$ ), respectively. This, might be an indication of the insertion of these elements in these non-coding genomic regions and such mobile elements are highly active in trypanosomes genome (Khan *et al.*, 2015) and have important roles in remodelling the parasite genomes (Bhattacharya, Bakre and Bhattacharya, 2002).

In contrary to the SSRs length merit, the GC content of the CSSRs presented higher mean number in comparison to DSSRs mean (54.4%, 49.9%), respectively. The same trend was also observed for the presence of tRNA genes in those regions (14.6%, 4.8%), respectively.

Some repeat motifs were found more likely on DSSRs enriched by different nucleotide contents (A, AAT, TTA, TATA or CT) rich motifs of different lengths ranging from 20 -70 nt.

**Table 4.5 Numerical comparison between DSSRs and CSSRs of *T. vivax*.**

	Divergent SSRs	Convergent SSRs
<b>Number</b>	41	41
<b>Mean length bp</b>	10,636	3,625
<b>Ingi2 elements (percentage)</b>	21 (51%)	7 (17%)
<b>tRNA (%)</b>	2 (4.8)	6 (14.6)
<b>tRNA + Ingi2</b>	0	0
<b>snRNA</b>	0	0
<b>GC%</b>	49.9	54.4

#### 4.3.4.4 Directional gene clusters in *T. vivax* IL1392 PacBio contigs

The total number of examined DGCs was 76 (40 fDGCs and 36 rDGCs) the mean length of the all observed DGCs was 192,217 bp (mean length of fDGCs 190,266 and for rDGCs 194,385). However, this number could not reflect the actual DGCs in the genome of *T. vivax*, because we considered the longest possible contigs only. Statistical significance analysis of the means showed no significant differences between the means of fDGC and rDGCs.

Some of these clusters located in the internal parts of contigs showed telomeric repeats at their ends. Whilst these clusters involve mainly housekeeping genes and protein coding genes with unknown function, it occasionally interrupted by 1-6 coding genes for surface proteins such as adenylate cyclase, ESAG5, major surface protease MSP and trans-sialidase, which were located in anti-sense manner to the DGC. A similar layout was also noticed in *T. congolense* and *T. brucei* (Christiane Hertz-Fowler, 2007) suggesting perhaps different transcription of these genes that encodes for surface proteins nonetheless localized in internal locations of MBCs.

The SSRs and the DGCs are both related; as the SSRs are localized between each two DGCs. Our results suggested lower GC content of dSSRs in both *T. congolense* and *T. vivax* (46.47%; 49.9%, respectively, which is highly similar to a corresponding region on chromosome one of *Leishmania major* (Puechberty *et al.*, 2007). In addition, these sequences have been proposed to host polymerase II transcriptional sites, so that it considered transcriptional initiation factors for both PTUs on both ends of this class of spacer genomic regions as most likely to host histones in *L. major* and *T. brucei* (Martinez-Calvillo *et al.*, 2003, 2004; Puechberty *et al.*, 2007; Kolev *et al.*, 2010), respectively. A similar trend in GC content differences between the two regions was also noticed in *Leishmania* chromosomes (Tosato *et al.*, 2001).

Generally, the SSRs in the *T. vivax* genome exhibited high consistency with those of *T. congolense* (chapter two) in terms of the feature length, GC contents and other features. Nonetheless, tRNA sequences were suggested in DSSRs, however still in low numbers, while these were not have observed

in *T. congolense* DSSRs and it also exhibited more *ingi* elements in both *T. vivax* SSRs in comparison to the corresponding regions in *T. congolense* genome, suggesting the possibility that the *T. vivax* genome is highly subjected to genomic re-shaping.

The high likelihood of having retroelements in these regions was also noticed in *T. brucei* and *T. cruzi* (Obado *et al.*, 2007; Macías, López and Thomas, 2016), respectively. We also identified presence of tRNAs most likely in convergent SSRs of *T. congolense* and *T. vivax*. This finding is consistent with corresponding genomic regions in *T. brucei* (Ivens *et al.*, 2005) and it subjected to be a transcription terminator site (Hull *et al.*, 1994; Maree and Patterton, 2014).

These were conserved regions of AT-rich and CT-rich motifs in both analysed genomes, and similar sequences were also proposed to be present in the genome of free living kinetoplastid *Bodo saltans* (Jackson *et al.*, 2016), suggesting a conserved feature in the Kinetoplastid genomes.

The DGCs in *T. vivax* expressed higher figures for these than in *T. congolense* in terms of length and number, suggested an expansion in the core regions of the *T. vivax* genome. This result might be in agreement with the size of the genome and the total number of protein coding genes being both increased in the latter trypanosome.

## 4.4 Conclusion

Over all, the use of third generation SMRT sequencing enabled us to generate a highly contiguated *de novo* genome assembly of *T. vivax* strain IL1392. This is a new strain adapted in the laboratory and able to produce infectious blood forms to the small rodents, which has recently been used in many research groups interested in this organism. Unveiling more genomic regions with physical integrity and better gene models might be provide the scientific community vital genomic resources for this organism.

Accordingly, we have been able to show coding and non-coding important genomic territories of this trypanosome, which emphasized that the increased genome size noticed in this project could be correlated to the increased internal regions with housekeeping nature, which in turn might have suggested larger MBCs in comparison to the other studied trypanosomes. Analysis of these regions also showed the conservation of kinetoplastids' genomic structures.

Our analysis proposed previously unprecedented large-scale genomic translocations affecting core chromosomal regions in comparison to the other African trypanosome *T. brucei* and to a lesser extent *T. congolense*, which could have correlated to the difference in life cycle, cell morphology and host range.

We suggested more predicted genes in this strain, and most likely complete gene models, a high number of singletons (7,000) and high number of genes with unknown functions. More genes in this organism might have enabled it spread over different geographical areas with wide spread host range in comparison to the other African trypanosomes.

Finally, the novel findings in this project could provide the scientific community who are interested in the study of this trypanosome species or comparative studies with other related organisms in different aspects such as cell biology, drug discovery, vaccine design...etc., the genomic resources they need permitting a new opportunity towards better understanding and more flexibility in exploring genomic region of interest than were previously possible.

# Chapter 5      Comparative phylogenomic analysis of Kinetoplastids with a focus on African Trypanosomes

## 5.1 Introduction

The kinetoplastids are characterized by the presence of the flagella and the kinetoplast (Hajduk, Siqueira and Vickerman, 1986). These unicellular eukaryotes include free-living microorganisms, as well as parasites of diverse invertebrate, vertebrate, and plant species. These are widespread throughout different environments such as freshwater aquatic and land environments. For example, the free living aquatic bacterivorous kinetoplastid *Bodo saltans* (Deschamps *et al.*, 2011), plant parasites from the genus *Phytomonas* (Camargo, 1999; Stuart *et al.*, 2008; Jaskowska *et al.*, 2015); or the monoxenic insect parasites such as *Crithidia*, *Angomonas* and *Leptomonas* (Maslov *et al.*, 2013). The medically and veterinarily important members are the heteroxenous Trypanosomatidae that infect humans and animals causing serious acute or chronic illnesses and are often transmitted by a vector (**Table 5.1**).

The African trypanosomes are extracellular and fall into two main disease types those responsible of Human African trypanosomiasis HAT disease, caused by *T. brucei rhodesiense* or *T. b. gambiense* (Brun *et al.*, 2010; World Health Organization, 2013) and Animal African Trypanosomiasis AAT which is caused by *T. b. brucei*, *T. congolense* and *T. vivax* (Milligan and Baker, 1988). These species are all transmitted by the blood feeding flies of the family Glossinidae (Lehane *et al.*, 2004; Watanabe *et al.*, 2014).

From this brief introduction, it is apparent that these organisms have adapted to wide variety of environmental roles, yet they generally share conserved genomic features (Berriman, 2005; El-Sayed *et al.*, 2005).

The advances in DNA sequencing technologies have allowed for the description of a number of different kinetoplastids. Recently, the genome sequence of the free living kinetoplastid *B. saltans* (Jackson *et al.*, 2016); the

closest free living Bodonoid to Trypanosomatids (Doležel *et al.*, 2000; Deschamps *et al.*, 2011) was produced, which could provide a good foundation to track the evolution of parasitic kinetoplastids and describe the core set of genes across free living and parasitic kinetoplastids.

The new draft PacBio assemblies of both *T. congolense* IL3000 and *T. vivax* Y486 derived strain IL1392, along with the publicly available genomes could give new insights into kinetoplastis evolution and specialisation.

The aim of this chapter was to investigate possible similarity or dissimilarity of gene repertoires (Koonin, 2005; Dolinski and Botstein, 2007) of the selected kinetoplastids living in different environments. The two new African trypanosome genome versions of *T. congolense* and *T. vivax* provide more information enabling an improved phylogenetic analysis of the relationship and role in shaping the kinetoplastid evolution. As well as revealing genes linked to lifestyle and pathogenicity, these data have the potential as drug or vaccine targets for pathogenic kinetoplastid therapies.

The objectives of this chapter were:

1. To define orthologous protein sets across the trypanosomes.
2. Describe core genes and patterns of orthologous genes correlated with trypanosome life history traits and parasitic lifestyles.
3. Focus on the orthologous genes that define the African trypanosome species.

The significance of this approach lies in the ability to predict protein orthologues which could have similar functions among different species, due to a shared common ancestor. This approach will also emphasize possible gene expansions and sub-families in certain species compared to the others, possibly to cope with certain demands (Eisen, 1998; Sonnhammer and Koonin, 2002; Zmasek and Eddy, 2002).

**Table 5.1 Medically and veterinarily important heteroxenic kinetoplastids.** Highlighting the most important infective species, their hosts, geographical distribution and micro-environment within their vertebrate hosts. (Human parasites; Animal parasites).

Species name	Geographical distribution	Lifestyle	Main Host(s)		Reference
			Invertebrate	vertebrate	
<i>T. cruzi</i>	South and central America	Intracellular	Reduviidae and Triatominae	Human and mammals	Garza <i>et al.</i> , 2014
<i>Leishmania</i> species	Old and new world	Intracellular	Phlebotomus and Lutzomyia	Human and mammals	Alemayehu and Alemayehu, 2017
<i>T. brucei. gambiense</i> and <i>T.b. rhodesiense</i>	Africa	Extracellular	Glossina	Human	Stuart <i>et al.</i> , 2008; Brun <i>et al.</i> , 2010
<i>T.b. brucei</i>	Africa	Extracellular	Glossina	Domestic animals	Milligan and Baker, 1988
<i>T. congolense</i>	Africa	Extracellular	Glossina	Domestic animals	Gillingwater, Mamabolo and Majiwa, 2010
<i>T. vivax</i>	Old and New world	Extracellular	<i>Glossina</i> and other blood sucking insects	Domestic animals	Osório <i>et al.</i> , 2008

## **5.2 Methods**

### **5.2.1 Selected Kinetoplastids for the phylogenomic analysis**

In addition to our two PacBio assemblies of *T. congolense* and *T. vivax*, five other kinetoplasts were chosen according to their life style and availability and completeness of their genome assemblies (Table 5.2). The proteome datasets of these organisms were inputted in the clustering pipeline.

### **5.2.2 Clustering of kinetoplastids proteomes**

OrthoFinder version 1.0.0 pipeline was used for clustering all proteome data sets across all selected species. The rationale beyond using this clustering tool is to deduce the sequences that shared similarity among the analysed data clustering them into phylogenetically related clusters (orthogroups) leaving those sequences with no similarity to any other sequence as independent sequences (singletons). For the detailed description of the tool (see chapter 2 section 2.2.9).

Default options were applied for protein clustering using OrthoFinder version 1.0.0 and a newer version of this pipeline version 1.1.10 was applied for generation species tree and generation of multiple alignments on the output of version 1.0.0 contains orthogroups using option “-og” and multiple sequence alignment was forced using flag “-M msa” in order to get a result directory with all multiple sequence alignments done by “mafft”, for further sequence alignment inspection if needed (Figure 5.1).

### **5.2.3 Generation of phylogenetic trees in this chapter**

Generation of multiple sequence alignment and gene trees of the orthogroup output was implemented in OrthoFinder pipeline version 1.1.10. The pipeline uses “mafft” (Kato and Standley, 2013) for multiple sequence alignment and FastTree version 2.1.9 (Price, Dehal and Arkin, 2010) with default options for tree generation (for more information see the pipeline manual

<https://github.com/davidemms/OrthoFinder/blob/master/OrthoFinder-manual.pdf>).

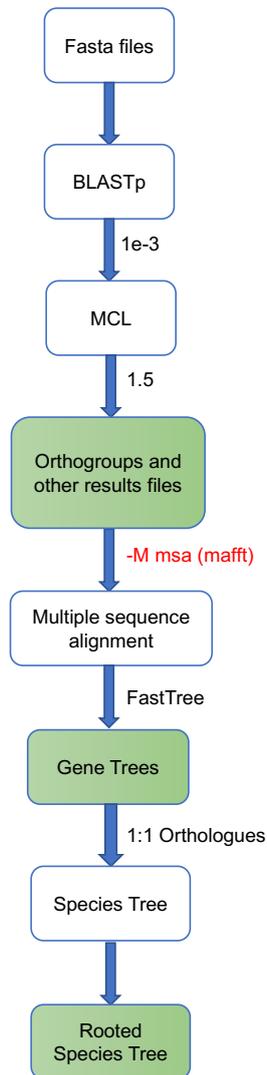
Some phylogenetic trees of gene families in this chapter were also reproduced for robustness using MUSCLE (Edgar and Edgar, 2004) implemented in SeaView package version 4.6.2 (Gouy, Guindon and Gascuel, 2010); and the inference of maximum likelihood phylogenetic tree was inferred PhyML version 3 (Stéphane Guindon and Gascuel, 2003) implemented in SeaView.

The rooted species tree was inferred from gene trees of 1:1 orthogroups automatically from the pipeline by evoking option “-ft”.

FigTree version 1.4.3 (<http://tree.bio.ed.ac.uk/software/figtree/>) was also used to manipulate and view phylogenetic trees.

**Table 5.2 Selected Kinetoplastids for phylogenetic study.** General information about selected species for this analysis. General genome statistics, species names, life cycle profile and their GeneBank accession numbers are shown.

Species name	Life class	Genome			Accession number
		Size (Mbp)	Chromosomes	Number of proteins	
<b><i>Bodo saltans</i></b>	Free living	39.864	Unknown	18,963	GCA_001460835.1
<b><i>Leishmania major</i></b>	Heteroxenous	32.86	36	8,519	ASM272v2
<b><i>Crithidia fasciculata</i></b>	Monoxenous	32.63	30	9,489	GCA_000331325.2
<b><i>Trypanosoma cruzi</i></b>	Heteroxenous	32.53	41	10,338	GCA_000209065.1
<b><i>T. brucei</i></b>	Heteroxenous	35.83	11	10,287	ASM244v1
<b><i>T. congolense</i></b>	Heteroxenous	39.49	11	11,012	This study
<b><i>T. vivax</i></b>	Heteroxenous	67.82	Unknown	19,006	This study



**Figure 5.1 OrthoFinder pipeline options used in the phylogenetic analyses of kinetoplastid proteomic data.** Green boxes refer to main pipeline output, and the generation of multiple sequence alignments option was enforced (red text).

## 5.2.4 Obtaining the designed sequence sets from OrthoFinder output

The main goal of this chapter is to find the putative genomic repertoires shared among kinetoplastids sharing similar lifestyles (i.e. all kinetoplastids (core kinetoplastids), parasitic kinetoplastids, combinations of African trypanosomes and species-specific gene families). Then further sub division into two categories: The groups consisting of only one orthologue for each species (one to one orthologues 1:1) and shared orthologues groups that contain more than one orthologue/paralogue (i.e. multigene families N: N).

UNIX command line functions (Appendix A.3) were used to conditionally call the orthogroups “rows” from the OrthoFinder output “.csv” file for each analysis group mentioned above generating tab delimited text files contain the following information:

- 1- Files of all combinations of kinetoplastid analysis groups containing 1:1 gene families with their functional annotation.
- 2- Files of all combinations of kinetoplastid analysis groups containing N: N gene families, with the number of sequences of each species per orthogroup and their functional annotation.
- 3- African trypanosome species specific gene families.

## 5.2.5 Gene Ontology enrichment analysis

### 5.2.5.1 Extraction of the Gene Ontology term (GO) IDs

Where available, *T. brucei* sequence IDs for an orthogroup were used to extract query lists of GO IDs from the *T. brucei* genome annotation file (in GFF3 format) using a combination of UNIX command lines (Appendix A. 3). The final output of this step was a text file containing a list of GO term IDs that could then be used as an input for the next step.

### 5.2.5.2 Enrichment analysis of GO terms

See section 2.2.10.

### 5.2.6 Searching gene bank database

Specific BLASTp searches were conducted using the NCBI non-redundant protein sequences database with default e-value to identify hits to query FASTA protein sequences.

## 5.3 Results and discussion

Protein clustering results suggested 8,503 orthogroups contained 75.9% of the total number of input genes, from which 2,755 (3.1%) protein sequences were resolved into 115 species-specific gene families. 24.1% of input genes were singletons. (Figure 5.2).

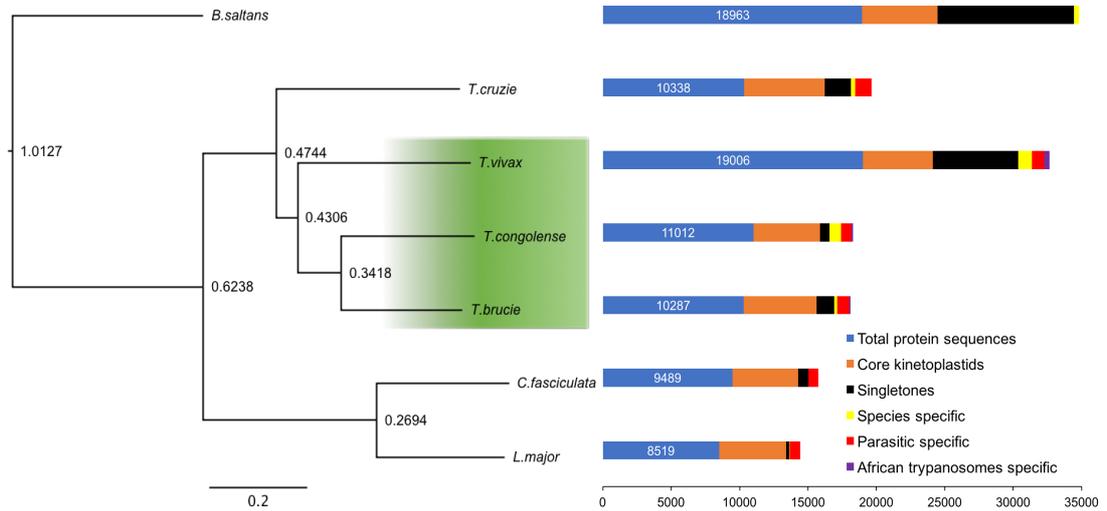
### 5.3.1 Kinetoplastids core gene families

These refer to the protein clusters that have members shared across all examined species. There were 36,411 genes involved in this category distributed over 2,410 1:1 orthologues and 1,753 N: N gene families. *Bodo saltans* and *T. vivax* have the largest predicted proteomes, with both showing the highest values for the unassigned genes with the least unassigned genes found in *L. major*. While *T. vivax* and *T. congolense* presented the highest figures with regard to the species-specific gene families among the analysed kinetoplastids, *L. major* and *C. fasciculata* showed the least (Figure 5.2). The highly species-specific gene families in both PacBio assemblies may reflect the ability of PacBio assemblies to unravel repetitive gene families in these trypanosomes.

#### 5.3.1.1 Tree evolution of parasitism

The species phylogenetic tree was inferred from the 2,410 1:1 orthologous kinetoplastid genes (Appendix C2), *B. saltans* was set as an outgroup and the African trypanosomes were clustered into one clade away from *L. major* and

*C. fasciculata* (Table 5.3) and (Figure 5.2). The tree inference is in agreement with a previous analysis (Jackson, 2015).



**Figure 5.2 Kinetoplastid species tree with a corresponding horizontal bar graph highlighting the size of proteome per species and the distribution of main gene clusters.** Species phylogenetic tree of selected kinetoplastid species deduced from 1:1 orthologous groups across all species as inferred by OrthoFinder. Node ages are shown on each node position. African trypanosomes are highlighted in gradient green box.

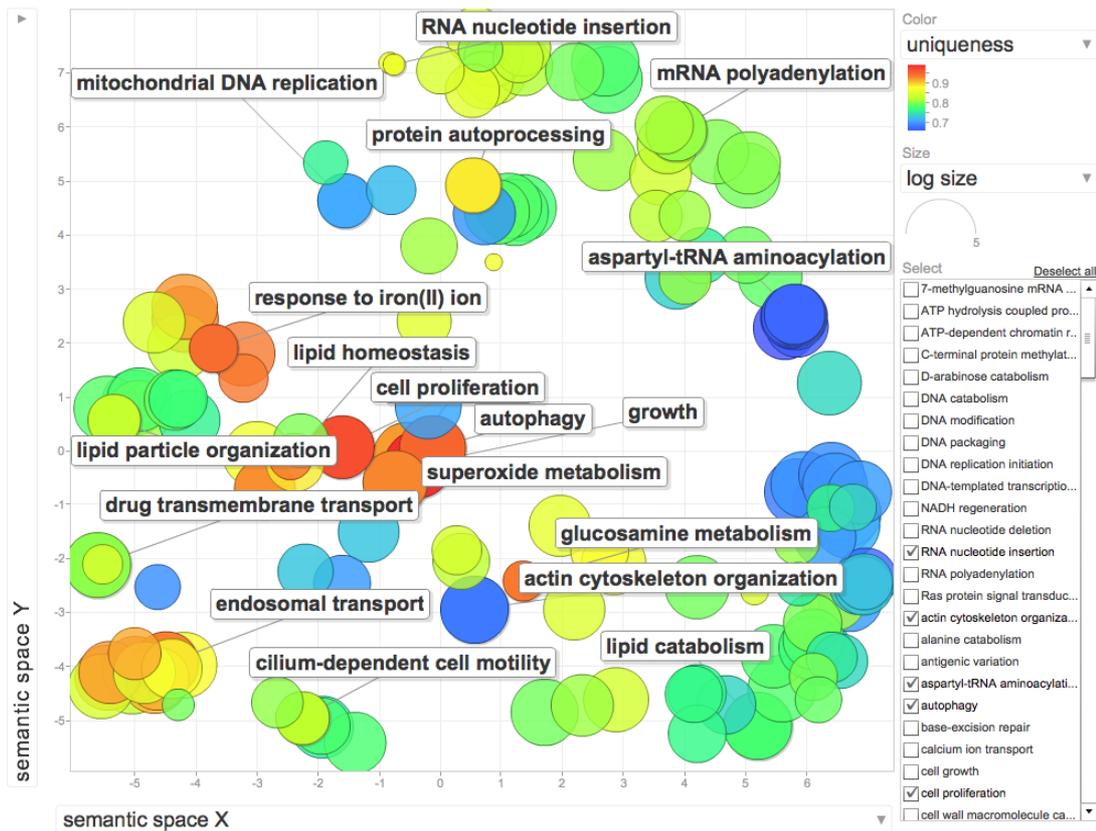
**Table 5.3 1:1 orthologues gene families of core kinetoplastid protein set.** The functional assignment of these families was transferred from the *T. brucei* genes. Only the first ten orthogroups and two fields having orthogroup IDs and the functional depiction were chosen for the viewing clarity reasons, the complete excel sheet is in Appendix C2.

Orthogroup ID	Gene ID T.b. b	Functional Description
<b>OG0003641</b>	Tb11.v5.0553.1	ubiquitin hydrolase, putative
<b>OG0002593</b>	Tb927.1.1000	developmentally regulated phosphoprotein
<b>OG0002594</b>	Tb927.1.1010	E3 ubiquitin-protein ligase KCMF1, putative
<b>OG0004513</b>	Tb927.1.1060	Cell cycle checkpoint protein RAD1-like, putative
<b>OG0004543</b>	Tb927.1.1120	ribosomal RNA-processing protein 8, putative
<b>OG0004551</b>	Tb927.1.1160	kinetoplast ribosomal PPR-repeat containing protein 3
<b>OG0004553</b>	Tb927.1.1210	conserved protein, unknown function
<b>OG0004547</b>	Tb927.1.1220	RWD domain-containing protein
<b>OG0004548</b>	Tb927.1.1230	chaperone protein DnaJ, putative
<b>OG0004489</b>	Tb927.1.1340	Quinonprotein alcohol dehydrogenase-like protein, putative

### **5.3.1.2 Gene duplication in free living kinetoplastid vs parasitic ones**

Genes in free-living kinetoplastid versus parasitic groups were extracted. Where there was at least one sequence per selected species, clustering analysis suggested 1,753 orthologous group (19,541) genes in this category. These sets of clusters are biologically important because they reveal the gene expansion or gene loss of a subset of genes of each species, which might reflect a redundancy or otherwise a shortage in a particular function or biological aspect (Eisen, 1998; Gabaldón and Koonin, 2013).

Functional enrichment analysis suggested that the core genes (i.e. shared among all members of kinetoplastids) clusters are linked to housekeeping pathways like glucosamine metabolism, autophagy, cell proliferation, growth, and endosomal transport. (Figure 5.3).



**Figure 5.3 Functional enrichment analysis of core protein kinetoplastids of multiprotein clusters across different parasitic and the free living kinetoplastids.** The more uniqueness of semantic term the closer to red the colour is, and the size of the circles reflects the number of sequences sharing the same context.

The free-living kinetoplastid in this study showed a vast expansion in a gene family encoding for leucine-rich repeat (LRRP) (OG0000007), almost three times the number of the highest (*T. cruzi*) amongst the parasitic species. The proteins of this family has a LRRP domain, which is most likely involved in protein-protein interactions in a wide range of eukaryotic cell processes, such as regulation of cell cycle, toll-like receptor in T cells and plant innate immunity ( Daher et al., 2007; Athman & Philpott, 2004; McHale, Tan, Koehl, & Michelmore, 2006).

*B. saltans* also showed the highest copy number over the kinetoplastids in protein clusters having functions of calcium/calmodulin-dependent protein

kinase, STE group serine/threonine protein kinase, protein kinase, endosomal integral membrane protein, kinesin, dynein heavy chain, and p-glycoprotein (Table 5.4). The majority of these sets are protein kinases, especially serine/threonine protein kinases, which showed a number of subfamilies in this analysis. They are type of enzymes that phosphorylate proteins at serine or threonine sites and it is a significant post translational modification playing an important part in different aspects of cellular processes and signal transduction in eukaryotic cells ( Parsons, Valentine, & Carter, 1993; Knippschild et al., 2005; Seong & Ha, 2012). The p-glycoprotein (P-gp) is also a protein proven to facilitate the efflux of bacterial toxins and harmful secondary products out of the eukaryotic cells (Lin and Yamazaki, 2003; Amin, 2013). Presumably such extreme expansion in these gene families involved in a wide range of cellular functions, might play a fundamental role in independent life. However, the exact role of such specific protein sets in this organism has not been tested yet.

By contrast, parasitic species have expanded genes participating in host-parasite interactions, like receptor-type adenylate cyclase GRESAG 4 (OG0000011), cysteine peptidase (cathepsin L-like) (OG0000018, OG0000087, OG0000091, OG0000162), Gp63-1 surface protease homolog (Leismanolysin) (OG0000019), Paraflagellar rod protein (OG0000024) and heat shock protein (heat shock protein, putative).

Members of adenylate cyclase GRESAG 4, are a potential blood-form transmembrane receptor that triggers differentiation of trypanosomatids to differentiate within the mammalian blood stream (Gonzales-Perdomo *et al.*, 1988). Interestingly, however, sub groups of this gene family were also reported to be expressed during the insect stage of *T. brucei brucei* (Simo *et al.*, 2010), indicating the versatile yet crucial role of this gene family to ensure successful establishment of these parasites in different hosts.

Cathepsins are also proteases that have an important role during the blood infection stage as they inhibit host immune defenses (Mottram, Brooks and Coombs, 1998; Mottram, Coombs and Alexander, 2004). The other protease (Gp63 enzyme) has been proven to function as a virulence factor by interfering

with complement components of plasma and macrophage receptors (Schlagenhauf, Etges and Metcalf, 1998; Joshi *et al.*, 2002).

The gene family of paraflagellar rod proteins are structural protein that support the flagellum of trypanosomatids, which in turn is an essential component responsible for parasite motility (Santrich *et al.*, 1997; Bastin *et al.*, 1999; Kohl, Sherwin and Gull, 1999). Heat shock proteins were reported to be highly expressed in blood form trypanosomes and they are essential enzymes to withstand the new environment in the blood stream of the mammalian host (Giambiagi-deMarval, Souto-Padrón and Rondinelli, 1996; Folgueira and Requena, 2007).

While the free living kinetoplastid showed expanded gene-families engaging in regulation of broad cellular processes and extruding potential toxic bacterial molecules, parasitic kinetoplastids exhibited a high tendency to expand genes involved in cell surface coat of the parasite perhaps to avoid or modulate hostile environment of their hosts. By looking to Table 5.4 one can see that the parasitic species revealed expansion for gene clusters located on the cell surface and involved in host-parasite relationship.

To sum up, in contrast to the free living kinetoplastids, pathogenic members exhibited redundancy in genes that facilitate their survivability within the hosts.

**Table 5.4 Clusters of multi-gene families inferred from protein products of shared similarity across kinetoplastids.**

Column 1 shows the orthogroup cluster IDs and the last one has the functional assignment of each group extracted according to Tb927 gene annotation. The heat gradient from red to blue shows the size of orthogroup. The first 30 clusters are shown (for complete list see appendix C3).

Orthogroup	<i>Bodo saltans</i>	<i>T. congolense</i>	<i>Crithidia fasciculata</i>	<i>L. major</i>	<i>T. cruzi</i>	<i>T. vivax</i>	<i>T. brucei927</i>	Functional Description
OG0000007	207	7	6	4	76	10	39	leucine-rich repeat protein (LRRP), putative
OG0000011	3	9	7	12	48	101	85	receptor-type adenylate cyclase GRESAG 4, putative
OG0000018	8	3	9	7	53	55	11	cysteine peptidase, Clan CA, family C1, Cathepsin L-like
OG0000019	5	17	2	63	28	7	20	Gp63-1 surface protease homolog, putative
OG0000024	2	3	6	1	77	11	12	Paraflagellar rod protein_
OG0000026	5	4	2	2	80	1	1	Stress responsive A/B Barrel Domain, putative
OG0000027	2	8	8	5	46	14	8	histone H3 variant V
OG0000032	2	49	13	3	1	6	8	pteridine transporter, putative
OG0000033	3	5	18	1	32	16	5	beta tubulin
OG0000038	3	11	4	6	35	7	6	cation transporter, putative
OG0000046	14	9	9	7	8	9	9	dynein heavy chain, putative
OG0000050	8	5	5	3	24	10	7	heat shock 70 kDa protein, putative
OG0000057	1	2	17	1	10	16	11	heat shock protein, putative
OG0000060	5	6	6	8	20	4	5	casein kinase 1, putative
OG0000061	4	11	4	3	6	6	19	glucose transporter, putative
OG0000062	3	1	7	3	22	11	4	elongation factor 1-alpha
OG0000063	4	10	11	8	6	5	6	Calpain-like protein 2
OG0000067	29	5	2	3	2	2	2	Calcium/calmodulin-dependent protein kinase, putative
OG0000070	18	4	4	2	2	8	5	kinesin, putative
OG0000072	3	8	7	5	5	6	8	protein phosphatase 1
OG0000075	1	2	2	1	15	8	9	S-adenosylmethionine synthetase, putative
OG0000076	1	2	5	2	24	2	2	Mitochondrial ribosomal protein L3
OG0000077	7	3	2	5	13	4	4	zinc finger protein family member, putative
OG0000078	1	5	1	2	20	2	7	cyclin-like F-box protein_
OG0000080	15	7	8	1	2	2	2	p-glycoprotein
OG0000082	4	5	5	5	6	6	5	calmodulin
OG0000084	21	6	3	3	1	1	1	STE group serine/threonine-protein kinase, putative
OG0000086	3	9	5	5	2	5	6	fatty acyl CoA synthetase
OG0000087	1	1	1	1	9	21	1	cysteine peptidase C (CPC)
OG0000088	3	1	1	1	1	27	1	ATP-dependent DEAH-box RNA helicase, putative

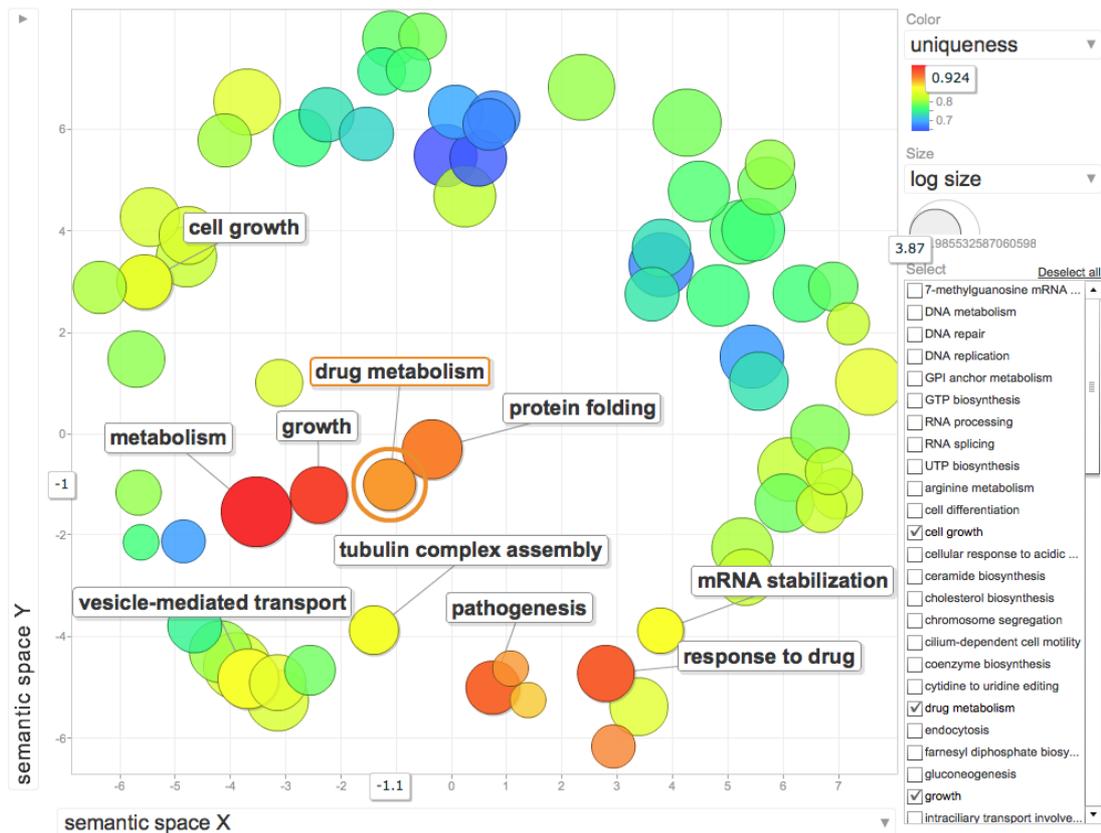
### **5.3.2 Specific gene sets that have evolved in parasite lineages**

There were 673 proposed orthogroups containing 892 sequences conditionally selected as each cluster should have at least one representative sequence from each inputted parasitic kinetoplastids (i.e. all selected species except sequences of *B. saltans*). There were (449) 1:1 orthologous groups and 224 N: N clusters containing 443 sequences.

#### **5.3.2.1 One to one orthogroups across parasitic species**

The enrichment analysis of GO terms assigned to the *T. brucei* polypeptides revealed involvement of these genes in wide cellular processes but with high uniqueness towards protein folding, general metabolic processes, pathogenesis, vesicles mediated transport, growth, response to drugs, drug metabolism and mRNA stabilization (Figure 5.4); the full list of the orthogroups and functional annotation is in (Appendix C4). This analysis highlights that these parasites could share conserved pathways related to common metabolic pathways, proliferation, growth, pathogenesis and interestingly, these pathways that related to drug response and metabolism suggest a selection of potential candidates for multi-species drug targets.

The potential compounds targeting more than one parasitic species have already been investigated by identifying one inhibitor or more to target important proteins related to cell growth or conserved metabolic processes among different kinetoplastids (*Leishmania* and *Trypanosoma* species) (Peña *et al.*, 2015; Qvit *et al.*, 2016).



**Figure 5.4 Functional enrichment analysis of 1:1 orthologues of the parasitic kinetoplastids.** The more uniqueness of semantic term the closer to red colour the circle and its size reflects the number of sequences sharing the same textual context.

### 5.3.2.2 Duplication of genes in the parasitic kinetoplastids

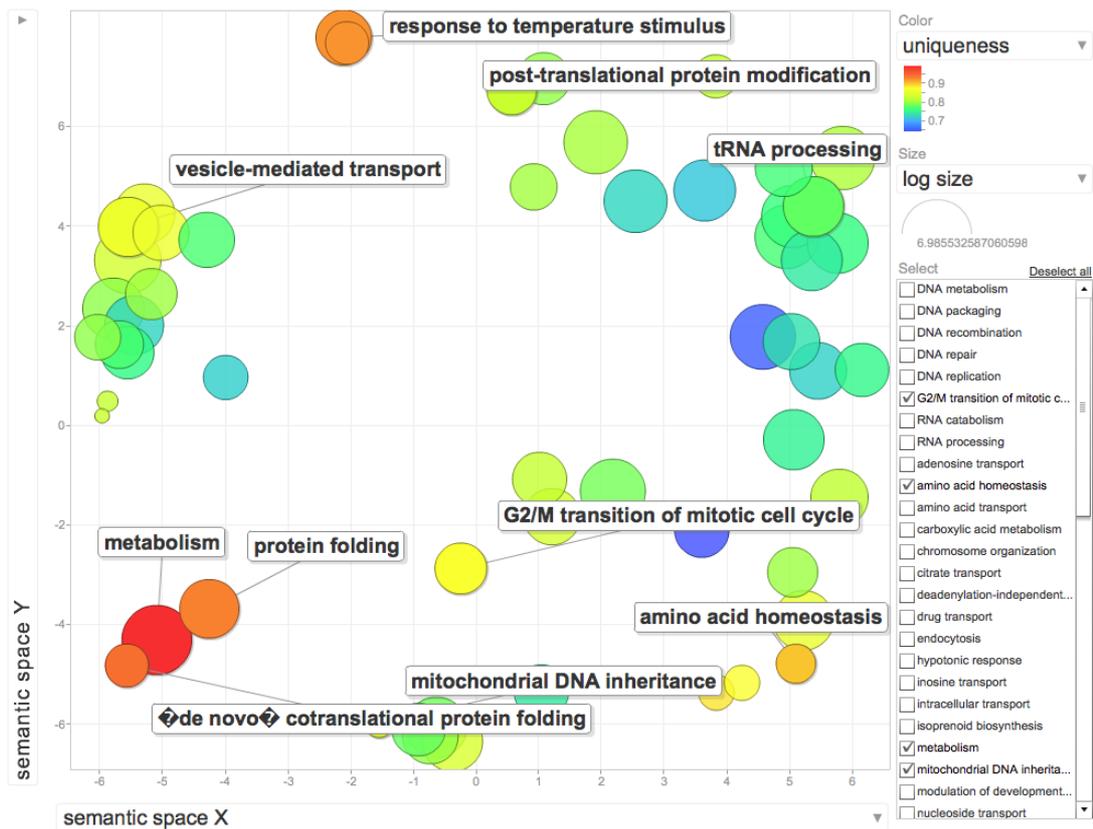
Some kinetoplastid species showed more copies than the others in the number of gene families. In general, the parasites that showed extreme gene expansions in this category are *T. cruzi*, *T. vivax* and *T. brucei* of gene families related to histone variants H4 and H2A, amino acid transporter and mitotubule-associated protein Gb4 (in *T. cruzi* and *T. vivax*); the later protein participating mainly in the movement of the flagellum (Table 5.5).

As in the enrichment analysis of one-one orthologous groups, enrichment analysis of this set showed high values of specificity for GO terms related to metabolism, and protein folding. Moreover, it also highlighted that

kinetoplastids as in the enrichment analysis of one-one orthologous groups, enrichment analysis of this set showed high values of enrichment for GO terms related to metabolism, protein folding. Moreover, it also highlights that kinetoplastids differ in their responses to temperature and amino acids change; which are environmental changes experienced by the parasites during the bottle neck of transmission between insect host to the mammalian host blood stream or vice versa (de Carvalho *et al.*, 1990; Schwede, Kramer and Carrington, 2012; Jimenez, 2014)(Figure 5.5).

**Table 5.5 N: N sequence clusters among parasitic kinetoplastids.** The groups with high variability in sequence contents are shown and the full list can see in Appendix C4. The intensity of the red colour refers to the number of sequences, while blue colour refers to one sequence per species in an orthogroup.

Orthogroup	<i>Crithidia</i>		<i>L. major</i>	<i>T. cruzi</i>	<i>T. vivax</i>	<i>T. brucei</i> 927	Functional description
	<i>T. congolense</i>	<i>fasciculata</i>					
OG0000014	9	7	12	102	24	11	histone H4 variant
OG0000016	6	6	6	105	18	13	histone H2A
OG0000023	1	1	9	51	20	36	hypothetical protein
OG0000030	9	4	3	15	28	27	amino acid transporter 8
OG0000054	9	3	2	42	2	2	phosphate-repressible phosphate permease
OG0000069	4	3	2	12	20	3	mitotubule-associated protein Gb4
OG0000079	9	4	5	1	9	10	major facilitator superfamily
OG0000097	1	1	2	6	21	2	hypothetical protein
OG0000110	11	3	8	2	2	3	Amastin surface glycoprotein, putative
OG0000135	1	1	1	20	1	1	TFIIH basal transcription factor subunit
OG0000152	1	1	1	5	2	13	nucleoside transporter 1
OG0000175	1	1	1	14	3	1	S-adenosylhomocysteine hydrolase
OG0000199	3	5	2	3	3	3	heat shock 70 kDa protein
OG0000215	2	1	1	4	5	5	conserved protein
OG0000216	3	2	2	5	3	3	Basal body protein
OG0000217	1	1	1	11	2	2	hypothetical protein
OG0000218	1	1	1	13	1	1	hypothetical protein
OG0000263	2	2	3	2	4	3	conserved protein, unknown function
OG0000265	2	2	6	2	2	2	calpain-like protein fragment
OG0000312	1	1	10	1	1	1	nucleoside 2-deoxyribosyltransferase
OG0000313	2	2	2	1	4	4	protein kinase



**Figure 5.5 Functional enrichment analysis of clusters with variable number of protein sequences of parasitic kinetoplastids.** The more uniqueness of semantic term the closer to red colour the circle and its size reflects the number of sequences sharing the same textual context.

### **5.3.3 Specific genes set that have evolved in African trypanosomes lineages**

These orthogroups were suggested by the clustering tool as shared among African trypanosomes (AT) (i.e. *T. brucei*, *T. congolense* and *T. vivax*). 590 sequences were contained in 64 gene families, and 41 orthogroups of 1:1 orthologues and 23 clusters contained 467 of N: N sequences from each species. Whilst the *T. vivax* showed vast expansions in a number of orthogroups, 361 sequences (77%), *T. brucei* has the least number of sequences (37).

#### **5.3.3.1 1:1 orthologues containing orthogroups across African trypanosomes**

Although the majority of 1:1 orthogroups are proteins with unknown functions, proteins with phospholipase A1, proteophosphoglycan, aldehyde dehydrogenase, zinc ion binding and transcription factor II a were predicted. The first two proteins related mainly to phospholipid metabolism and CDP-choline pathway, which is part of the “Kennedy pathway”, which plays an important role in the synthesis of glycosylphosphatidylinositol (GPI), an integral part of cellular membrane and considered a vital pathway for *T. brucei* growth in culture (Gibellini, Hunter and Smith, 2008; Farine *et al.*, 2015). In eukaryotic cells interference with the CDP-choline pathway leads to impairment of the cellular proliferation and differentiation and affect the transmission of the movement through cytoplasmic membrane (Fagone and Jackowski, 2013). This suggests potential importance of this pathway to the survival of African trypanosomes, making such genes as potential candidates for drug targeting.

#### **5.3.3.2 Duplication of gene families across African trypanosomes**

There was no large variation in copy number in this set of African trypanosomes. However, *T. vivax* showed a relative massive gene expansion in a number of clusters annotated with unknown function, but most importantly the gene family of the haemoglobin receptor (Haptoglobin-Hemoglobin) and retrotransposons of SLACS (spliced Leader-association conserved sequence) (Table 5.6).

The Haptoglobin-Hemoglobin receptor (HpHbR) facilitates uptake of Haptoglobin-Hemoglobin (HpHbR) complex from mammalian blood (Vanhollebeke *et al.*, 2008) and it has been found differentially expressed in this stage of *T. brucei* life cycle but not during the insect stage. In contrast, proteomic analysis of *T. congolense* life stages showed abundance of this receptor during the insect stage (Eyford *et al.*, 2011). Similarly, to *T. congolense*, *T. vivax* showed higher expression levels of this gene during insect stage but not the blood form stage on the transcriptomic level (Jackson *et al.*, 2015). Immunofluorescent antibodies against TcHpHbR protein revealed an abundant coat of this receptor over the body surface of *T. congolense* including the flagellum of the epimastigote form isolated from the Tsetse fly mouth parts but not from proboscis or midgut (Lane-Serff *et al.*, 2016). In the same study researchers found that the affinity of TcHpHbR to bind to the free heme ~ 1000 times that of its ability to bind the HpHb complexes. Moreover, they also demonstrated that the TcHpHbR receptor is ~ 1000-fold more abundant in epimastigote than the TbHpHbR during the blood form stage. Therefore, it could be renamed as a haemoglobin receptor rather than HpHbR in these trypanosomes.

The fact that this receptor is in both *T. vivax* and *T. congolense* and expressed during their presence in the mouth parts of the Tsetse flies, but not in *T. brucei* which lodge in the salivary glands of the fly and away from blood meals, give them the ability to scavenge free heme liberated from the breakdown of the erythrocytes during blood meal feeding of the Tsetse flies from the vertebrate host.

All these transcriptomic and proteomic evidences point to HpHbR expression and its role in the insect stage of both *T. congolense* and *T. vivax*, however, the current genomic repertoire showed only one orthologue in the latter but not in synteny to the single gene copy in *T. brucei* on chromosome six (gene ID Tb927.6.440). Furthermore, the orthology analysis on TriTrypDB could not suggest a possible orthologue in the current *T. congolense* genome sequence

(<http://tritrypdb.org/tritrypdb/app/record/gene/Tb927.6.440#evolutionary-biology> ). Interestingly, however, searching the amino acid sequence products of the gene family of *T. congolense* in the PacBio assembly against NCBI non-redundant protein sequences database using BLASTp showed significant hits e-values of ( $6e^{-162}$ ,  $4e^{-153}$ ) to the protein sequence accessions BAV81373.1, 4E40\_A, of TcHpHbR from *T. congolense* expressed during epimastigote form (Yamasaki et al., 2016; Lane-Serff et al., 2016). Such independent search results support the reliability of the phylogenetic assignment of this orthogroup.

Remarkably, the *T. congolense* PacBio assembly suggested an expansion of this gene family, six orthologues (5 on the chromosome 10 and single copy on a contig did not allocated to any one of MBC) and more extreme case of 20 orthologues in the *T. vivax* PacBio assembly located on three different contigs of the genome assembly (scf718000001804 (seven tandem copies), scf718000002271 (one copy), scf718000002326 (twelve tandem copies)).

The fact that the species tree (Figure 5.2), showed the divergence of *T. vivax* from the common ancestor prior to that of *T. congolense* and *T. brucei*, which might suggest a scenario of gene loss in both *T. congolense* and in more extreme case *T. brucei*, rather than a gene expansion in *T. vivax* and *T. congolense*. Moreover, TbHpHbR has been modified to increase its affinity to bind to the HpHb complex in the blood of mammalian host, as such an environment perhaps deficient in free heme (Lane-Serff et al., 2016). Even more adaptations have occurred in the subspecies *T. brucei gambiense* that infects human, as the addition of a point mutation in this receptor enabled it to avoid the binding of this receptor to the human innate immunity lipoprotein factors “Trypanosoma Lytic Factor1 and 2” (TLF1 and TLF2), which have the ability to kill other animal-infecting *T. brucei* trypanosomes in the human blood stream (Symula et al., 2012; DeJesus et al., 2013; Higgins et al., 2013).

Phylogenomic analysis of members of this gene family across African trypanosomes revealed a massive gene loss in *T. brucei* represented by a single copy on chromosome six; *T. congolense* orthologues tandemly arranged on chromosome ten, however, one gene located out of this cluster in a contig not assigned to a MBC, curiously, diverged from the other *T.*

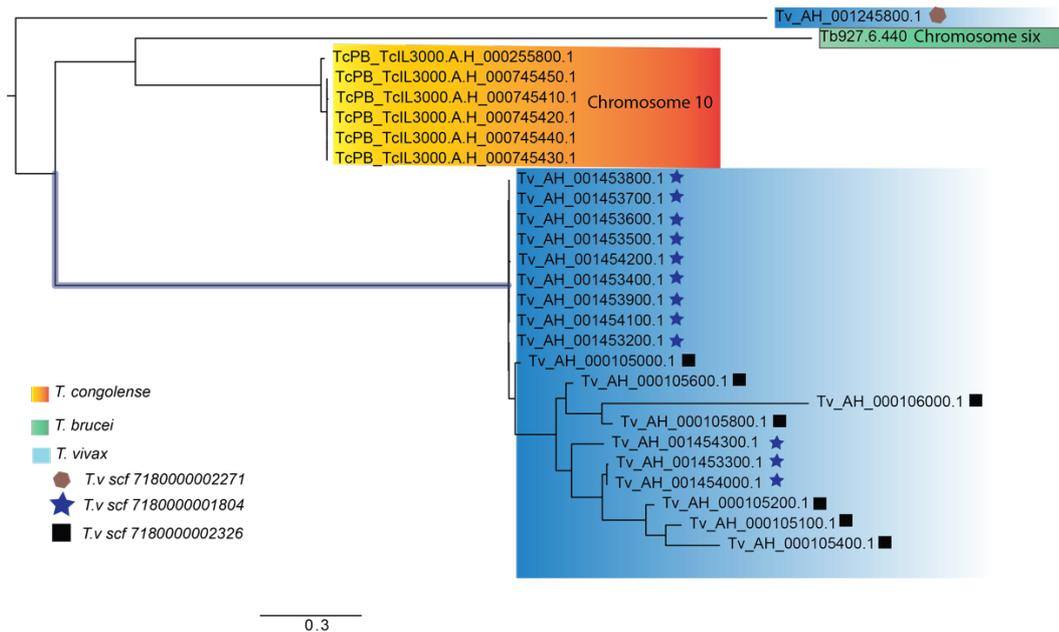
*congolense* specific sequences. *T. vivax* exhibited the most expanded gene family, as it showed 20 putative orthologues distributed as two tandemly allocated gene sets on two different contigs and a single putative gene which strongly diverged from all African trypanosomes sequences in this analysis, which was located in a contig in a DGC and does not contain other members of this gene cluster (Figure 5.6). This phylogenetic analysis showed a trend of gene loss in this gene cluster throughout African trypanosomes. Furthermore, the importance of this receptor for survival within the insect upper digestive system compartments accordingly.

In addition, the presence of five more putative gene copies of orthologues in the genome of *T. congolense* might explain the abundance of this receptor on the surface of epimastigote form in the Tsetse mouth parts in comparison to the protein expression levels of the *T. brucei* orthologue during the blood stage as demonstrated by Lane-Serff *et al.*, 2016. In addition, the results suggest their genome localization is in different a DGC and their sequence diversification especially in *T. vivax*, which might suggest a probability of it being expressed in different stages of the life cycle or perhaps in response to different environmental conditions surrounding the parasite. This and provides the foundations for further investigations about the structure of this receptor in both *T. congolense* and *T. vivax* and its expression.

The tree branch support was noticed to be generally acceptable (>60) for the main branches. However, for internal nodes, support was mostly very low and this could be influenced by a number of factors: firstly, the tandem repeat-like genes within the species and presence of length differences across paralogous copies; this could be reduced by omitting the outliers (extremely short sequences or extremely long sequences). Omitting such sequences might reduce the sensitivity of detecting gene expansion across African trypanosomes. Secondly, it would be beneficial to reduce the gaps in the phylogenetic trees by extensive manual editing and possibly to work on the highly conserved regions (domains), with generation of phylogenetic trees according to these protein domains.

**Table 5.6 N: N gene families among African trypanosomes.** *T. brucei* showed a significant gene loss in a number of shared gene families most importantly, towards the heme receptor (Haptoglobin-Hemoglobin) (highlighted with yellow colour). The intensity of the red colour refers to the number of sequences, while blue colour refers to one sequence per a species in an orthogroup.

Orthogroup	<i>T. vivax</i>	<i>T. congolense</i>	<i>T. brucei</i> 927	Functional description
OG0006896	2	1	1	Adenylate/guanylate cyclase
OG0006332	1	1	1	3 antigenic protein
OG0005729	3	1	1	2 BarP protein
OG0006746	1	2	1	1 cyclophilin-type peptidyl-prolyl cis-trans isomerase
OG0000123	20	6	1	1 haptoglobin-hemoglobin receptor
OG0006743	1	2	1	1 hypothetical PIN domain-containing protein
OG0000017	137	12	1	1 hypothetical protein
OG0000044	47	18	1	1 hypothetical protein
OG0000047	62	1	1	1 hypothetical protein
OG0000373	9	1	1	4 hypothetical protein
OG0005938	1	2	1	3 hypothetical protein
OG0006355	3	1	1	1 hypothetical protein
OG0006747	1	2	1	1 hypothetical protein
OG0006754	2	1	1	1 hypothetical protein
OG0006762	1	1	1	2 hypothetical protein
OG0006819	1	1	1	2 hypothetical protein
OG0006864	2	1	1	1 hypothetical protein
OG0006868	1	1	1	2 hypothetical protein
OG0006888	1	2	1	1 hypothetical protein
OG0006914	1	1	1	2 hypothetical protein
OG0006915	1	2	1	1 hypothetical protein
OG0004841	2	2	1	3 leucine-rich repeat protein (LRRP)
OG0000041	61	7	1	1 SLACS retrotransposable element
<b>total African</b>	<b>361</b>	<b>69</b>	<b>37</b>	<b>467</b>



**Figure 5.6 Phylogenetic tree of Haptoglobin-Hemoglobin receptor in African trypanosomes.** *T. brucei* has a single gene on chromosome six, which showed a massive gene loss comparably to *T. congolense* and *T. vivax*, which exhibited variable tandem expansions. *T. congolense* showed five orthologues tandemly arranged on chromosome 10 and one single gene on the BIN; *T. vivax* exhibited multiple copies, showed a degree of diversification distributed across three different contigs as two tandem repeated clusters and one single and highly diverged rooted out of the other orthologues of this African trypanosomes specific orthogroup.

#### 1.1.1.1 Gene families shared between *T. brucei* and *T. congolense*

The purpose of finding protein clusters shared between *T. congolense* and *T. brucei* is to extract potential useful information about the gene clusters that could be expressed during their shared component of life cycle, as they undergo developmental processes during the procyclic stage of both trypanosomes in the Tsetse midgut, which is a missing chapter from *T. vivax* life cycle, as they stick to the mouth part only during its insect stage (Hoare, 1972). Alternatively, perhaps shared gene families could depict shared armament during the animal infective stage such as variant surface antigens.

The cluster analysis predicted less variation in terms of gene number among 83 shared gene families containing 1, 263 genes in total. 72% of these genes were belong to *T. congolense* proposing a shortage in copy number in the *T. brucei* genome. Examples of gene expansions in *T. congolense* include a cluster of VSG, more putative genes expressed during procyclic stage (PAGs) in the fly than of *T. brucei*, and two clusters of expression site-association gene 2 (*ESAG2*) (Table 5.7). Meanwhile, *T. brucei* showed remarkable expansions in some gene families like those genes encoding for BarP, 75 KDa invariant surface glycoprotein and another cluster of invariant surface glycoprotein (Table 5.7).

Current analysis of orthology suggests there are a few *T. congolense* orthologues to *T. brucei* PAGs family members (TritrypDB). Here, however, our sequence assembly and clustering analyses proposed a gene expansion in this set of genes (or in other words a gene contraction in *T. brucei*) (22, 8) respectively, which was previously shown to be expressed during the insect stage procyclic form of *T. brucei* (Haenni *et al.*, 2006). Whilst most of the *T. brucei* PAGs genes in this cluster are positioned on MBC 10, except one on chromosome 11, two closely related putative *PAG* genes from *T. congolense* were located on MBC two. Meanwhile, the other *T. congolense* members were allocated to the Bin, suggesting their sequence specificity to this African trypanosome (Figure 5.7).

The other surface protein clustered with *PAG* genes are transferrin-like *ESAG6* and *ESAG7*. A putative transferrin-like *ESAG7* on chromosome 11 showed distant association to this group, likewise, two sequences of putative transferrin-like genes on chromosome seven remotely clustered together. Most of the other members of *T. congolense* on the far side of the phylogenetic tree are putative *ESAG6* transferrin-like and displayed variable relevance tendency (Figure 5.7). The other two groups that showed massive gene loss in *T. brucei* are *ESAG2*, which generally presented predominant expression during blood stage metacyclic; however, there are reports indicating some of these sequences could be transcribed during insect PCF (<http://tritrypdb.org/tritrypdb/app/record/gene/Tb927.1.5100#UserComments>); (Nilsson *et al.*, 2010).

The protein cluster that suggested the *T. brucei* gene expansion (two-fold increase) over *T. congolense* is a cluster of stage specific *Brucei* Alanine-Rich Repeat Proteins (BARP) (Table 5.7). These proteins have been shown to be differentially expressed on the *T. brucei* surface of epimastigote form (EPF) in the insect salivary glands 20 times more than the midgut procyclic form (PCF), making it a stage specific coat of this trypanosomes while it is being in the insect salivary glands (Urwyler *et al.*, 2007). Equally, *T. congolense* specific Glutamic acid –Alanine Rich Repeat Protein (GARP) was identified to be expressed on the parasite surface during late PCF in midgut, EPF in the proventriculus and mouth parts of the Tsetse fly (Loveless *et al.*, 2011).

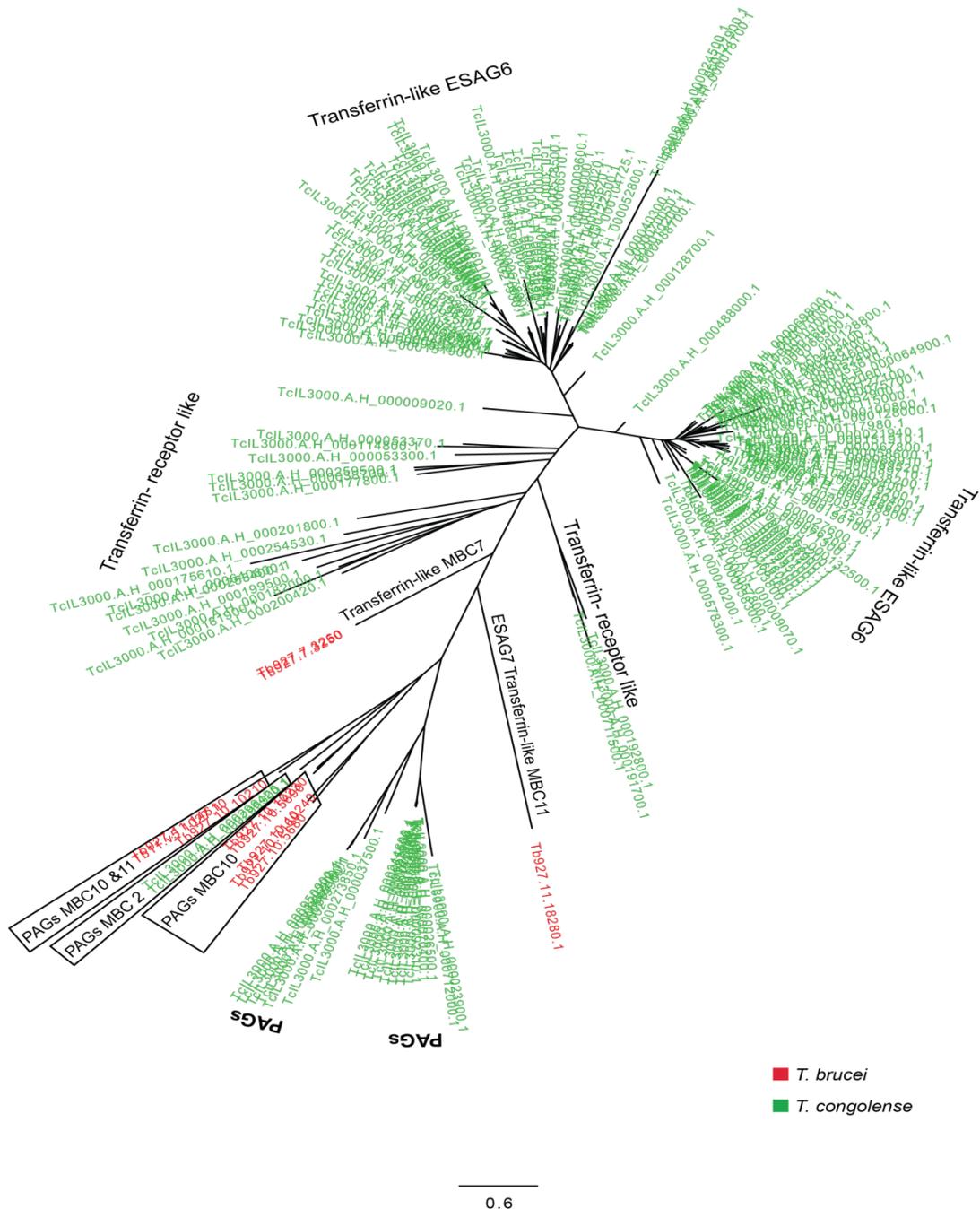
Phylogenetic analysis clustered these proteins into two main branches of weakly similar protein families (BARP and GARP) expressed almost at the same stage of the life cycle of both trypanosomes within the fly but in different body parts (Figure 5.8). Protein sequence similarity in this group were also previously noticed, but on smaller scale between these species by Urwyler *et al.*, 2007).

However, the analysis of these possible gene expansions and contractions was based on the number of genes in each gene family for each species, more robust method is needed to give the accurate inferences using statistical based analysis to assess gene birth and death across species using data from provided phylogenetic tree. Such computational analysis could be achieved by CAFÉ software version 3.0 (De Bie *et al.*, 2006).

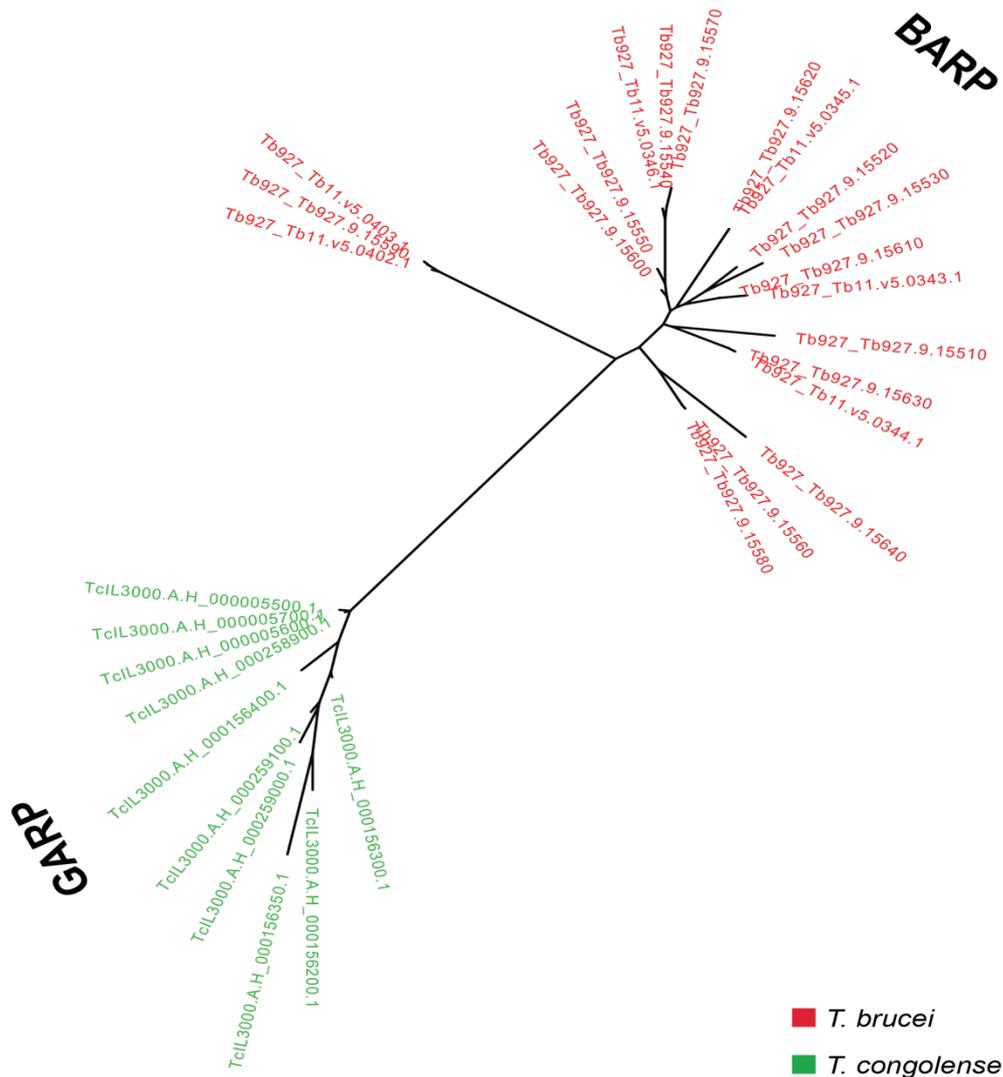
To sum up, this approach analytical proposed putative phylogenetic relationships among gene families, some of which were previously demonstrated to be expressed during certain life stages especially the insect stage in both *T. brucei* and *T. congolense*. This revealed the underlining core genomic repertoire that enabled these two trypanosomes to withstand the new harsh environment in the insect midgut or the upper parts of the insect digestive system, an ability that is probably missing from the other African trypanosome, *T. vivax*.

**Table 5.7 Shared protein sequence clusters between *T. brucei* and *T. congolense*.** *T. brucei* showed a significant gene loss in a number of shared families, most in a VSG cluster OG0000002, PAG and transferrin like cluster (OG0000015), two ESAG2 protein clusters, and a putative gene expansion in BARP (highlighted in yellow colour) and the 75 kDa invariant surface antigen (OG0000104, OG0000186), respectively. The intensity of the red colour refers to the number of sequences, while the blue colour refers to one sequence per species in an orthogroup.

Othogroups	<i>T. congolens</i>	<i>T. brucei</i>	Functional description
OG0000002	407	178	variant surface glycoprotein (VSG)
OG0000015	151	11	procyclin-associated genes PAGs
OG0000020	116	19	expression site-associated gene 2 (ESAG2) protein
OG0000029	86	3	expression site-associated gene 2 (ESAG2) protein
OG0000104	10	20	BarP protein
OG0000186	8	12	75 kDa invariant surface glycoprotein
OG0000310	14	1	hypothetical protein
OG0000445	9	4	hypothetical protein
OG0000618	10	1	hypothetical protein
OG0002120	4	4	invariant surface glycoprotein
OG0002127	7	1	Protein of unknown function (DUF1663)
OG0004804	6	1	hypothetical protein
OG0004836	4	3	hypothetical protein
OG0005702	1	5	hypothetical protein
OG0005932	3	3	cytoskeleton-associated protein
OG0006311	1	4	invariant surface glycoprotein
OG0006741	3	1	hypothetical protein
OG0006745	2	2	hypothetical protein
OG0006748	3	1	hypothetical protein
OG0006750	2	2	acidic phosphatase
OG0006811	1	3	tRNA
OG0007209	1	2	invariant surface glycoprotein
OG0007210	1	2	75 kDa invariant surface glycoprotein
OG0007212	2	1	hypothetical protein
OG0007215	2	1	hypothetical protein
OG0007216	2	1	65 kDa invariant surface glycoprotein-like protein
OG0007228	1	2	hypothetical protein
OG0007245	1	2	hypothetical protein
OG0007255	1	2	Kinesin associated protein



**Figure 5.7** Phylogenetic tree of predicted Procyclic Associated Protein (PAG) sequences of *T. brucei* (red) and *T. congolense* (green). The sequence alignment was achieved using MUSCLE and the tree was generated by FastTree implemented in OrthoFinder version 1.1.10. Two PAG sequences of *T. congolense* clustered in a clade with the corresponding representatives from *T. brucei*, a weak relationship is also noticed with transferrin-like ESAG6 and ESAG7 sequences from both trypanosomes.



**Figure 5.8** Phylogenetic tree of *Brucei* Alanine Rich Repeat BARP and Glutamic/Acid alanine Rich repeat GARP sequences of *T. brucei* (red) and *T. congolense* (green), respectively. The sequence alignment achieved using MUSCLE and the tree generated by FastTree implemented in OrthoFinder version 1.1.10. Sequences of late PCF and EPF of both species during insect stage of the parasite life cycle clustered uniquely on both sides of the tree.

### 5.3.3.3 Shared protein clusters between *T. congolense* and *T. vivax*

Interestingly, the position of *T. congolense* in the species phylogenetic tree is in mid-way between *T. vivax* and *T. brucei* and it shared common features of life cycle with the insect stages of the other African trypanosomes (i.e. *T. brucei* and *T. vivax*), making it a good candidate to demonstrate some shared aspects during a particular stage. In this case, both parasites reside in the mouth parts of the Tsetse fly before infecting a mammalian host, not in the salivary glands as with *T. brucei* (Van Den Abbeele *et al.*, 1999). Hence, shared putative protein families between *T. congolense* and *T. vivax* might suggest some protein families that could show a particular importance during this stage of the life cycle.

The clustering approach of the *T. congolense* PacBio and *T. vivax* PacBio protein products proposes 21 shared orthologous clusters containing 303 genes of equal or variable number of sequences from each species in each single orthogroup. Sequences of *T. vivax* represent 57% of this category. The *T. congolense* genome showed an expansion in *VSG-like GRESAG2*, *GRESAG4*, *VSG*, *papain* family of cysteine proteases and putative elongation factor Tu GTP binding domain containing protein, while the *T. vivax* genome showed possible gene expansion in a family of shared *VSG* genes, Pilo domain containing proteins, and a putative calpain protein (Table 5.8).

There are reports that *GRESAG2* and *GRESAG4* could be expressed during insect stage, particularly *T. congolense* and *T. brucei* (Alexandre *et al.*, 1990; Lopez, Saada and Hill, 2015) and both in latter families (OG0000034, OG0006312), respectively, *T. congolense* showed more sequences than *T. vivax*, suggesting gene expansion in *T. congolense* towards these gene families. In *T. brucei* *GRESAG4* was reported to be expressed on the flagellar membrane during PCF and importantly it was demonstrated to participate in an interesting social behaviour when a clumping of many cells start to be formed in response to certain environmental signals. The group movement reported among trypanosomes at this stage of life cycle (Lopez, Saada and Hill, 2015), if the case for *T. congolense*, having more genes of *GRESAG4* could be one of the prime mechanisms to survive the midgut environment. A

family of proteins (OG0004800) is known for its pathogenicity in *T. congolense* papain cysteine proteases (Lalmanach *et al.*, 2002; Rodrigues *et al.*, 2014). This analysis presents an expansion in a cluster of this pathogenicity factor, in this context, and might suggest an increase in pathogenic repertoire of *T. congolense* over *T. vivax*.

An interesting shared protein family (OG0000065) shows a massive (over 10 times) gene loss that has occurred in the *T. congolense* genome regarding these protein coding genes. Although the function of this gene family is not clear in trypanosomes, domain search suggested a bacterial Pilo-like domain (evalue=  $7.16e^{-03}$ ) containing protein in all four *T. congolense* copies. The bacterial Pilo domain-containing proteins are involved in the assembly of pilin, a protruding structure from the bacterial cell surface responsible for motility and socialising such as flagellar independent movements like sliding and trembling movements (Mattick, 2002; Burrows, 2012). Furthermore, these structures help the bacteria to do many other functions such as adherence to the host cells, protein secretion, DNA acquisition and micro-colony formation (Aas *et al.*, 2002; Kirn, Bose and Taylor, 2003; Han *et al.*, 2007). However, protein domain search of InterproDB (Hunter *et al.*, 2009) of *T. vivax* orthologues suggested the presence of nucleotide transferase domain, which might propose a favourable function of nucleotide transfer in the latter trypanosome.

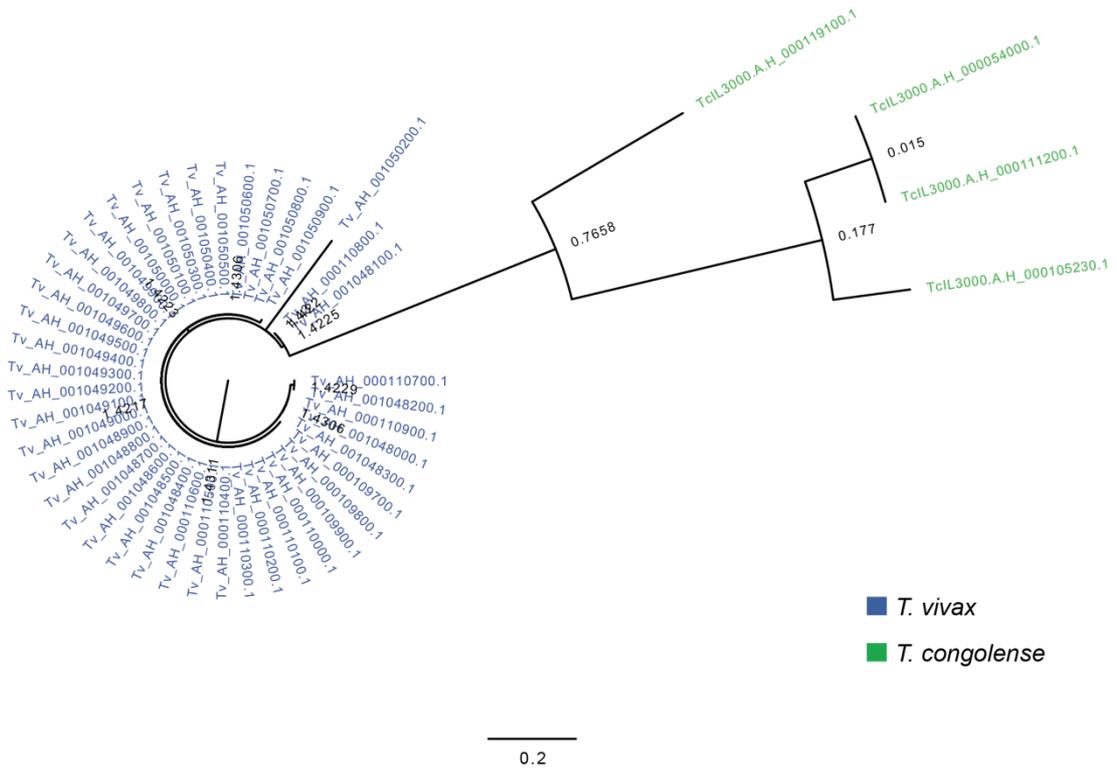
The phylogenetic analysis of this orthogroup suggested that *T. congolense* members were diverged from their ancestral orthologous group of *T. vivax*, which in turn showed a remarkable tandem expansion (Figure 5.9). Moreover, genomic localization of these genes as single instances in three separated positions in *T. congolense* genome and all cases accompanied by VSG genes suggesting three possibilities:

- 1- Co-expression with these variant antigens located on the surface of trypanosomes is highly likely and via similar transcriptional process as they are localized in subtelomeric regions.
- 2- It might suggest a pathogenic role of these genes during the blood form parasites.
- 3- It shows a relative sequence divergence among *T. congolense* members in comparison to the *T. vivax* ones (Figure 5.9), which in turn, however, showed three sets of tandem repeats of highly conserved amino acid sequences located on three different contigs of *T. vivax* PacBio genome assembly and in all three locations, they are followed by genes/pseudogenes encoding for a putative retrotransposons hot spot.

Accordingly, these results suggested *T. congolense* shows a vast copy number reduction in this gene family accompanied by sequence modulation to fulfil a purpose that could be entirely missing from *T. brucei*. However, experimental investigation on this protein family in both trypanosomes is important to investigate their role in these parasites.

**Table 5.8 Shared protein sequence clusters between *T. congolense* and *T. vivax*.** In general: *T. congolense* showed a significant gene expansion in GRESAG2, GRESAG4 and elongation factor Tu (OG0000034, OG0006312, OG0004799) respectively, and gene losses in a VSG cluster, Pilo domain containing protein, calpain protein (OG0000035, OG0000065, OG0000450), respectively. The intensity of the red colour refers to the number of sequences, while blue colour refers to one sequence per species in an orthogroup.

Othogroup	<i>T. congolense</i>	<i>T. vivax</i>	Functional sescription
OG0000034	79	1	GRESAG2
OG0000035	1	79	VSG
OG0000065	4	43	Pilo domain containing protein putative
OG0000134	2	23	hypothetical protein
OG0000450	4	9	calpain protein putative
OG0004796	5	2	VSG
OG0004799	6	1	Elongation factor Tu GTP binding domain containing protein, putative
OG0004800	5	2	Papain family cysteine protease, putative
OG0006312	4	1	Adenylate and Guanylate cyclase catalytic domain containing protein GRESAG4, putative
OG0006313	3	2	Chain A, Crystal Structure Of The N-Terminal Domain Of An Hsp90n, putative
OG0006742	3	1	Tubulin C-terminal domain containing protein, putative
OG0007213	2	1	hypothetical protein
OG0007214	2	1	Hsp90 protein, putative
OG0007266	2	1	PIN domain containing protein, putative
OG0007699	1	1	Hsp90 protein, putative
OG0007702	1	1	Paraflagellar rod protein, putative
OG0007703	1	1	Paraflagellar rod protein, putative
OG0007704	1	1	hypothetical protein
OG0007716	1	1	hypothetical protein
OG0007725	1	1	hypothetical protein
OG0007727	1	1	hypothetical protein
<b>Total</b>	129	174	303



**Figure 5.9** Phylogenetic tree of putative pilin assembly domain (Pilo) containing proteins of *T. congolense* (green) and *T. vivax* (blue). The tree showed high sequence similarity among *T. vivax* genes while it is more diverse in *T. congolense*. Node age is showed on each node.

#### 5.3.3.4 African Trypanosomes predicted species specific gene families

The importance of identifying the species-specific protein families has a particular significance to find candidates for drug or vaccine targets, possible phenotypic variations and candidates for diagnostic tools. The approach adopted for clustering a number of kinetoplastid proteomics data-sets has suggested coinciding equal species specific orthogroups for both *T. brucei* and *T. congolense* (11 each), however, it involved 204 sequences from the former and 857 genes from the latter. Interestingly, there were 997 sequences assigned to 32 *T. vivax* specific gene families (Table 5.9). Apart from unknown function proteins, remarkably, most of the clusters in this category in all three species showed groups of surface proteins like VSGs, ESAG11, Invariant surface glycoproteins and GARP and species-specific transposons.

Each species presented species-specific VSG genes in three groups for both *T. brucei* and *T. congolense*; however, *T. vivax* revealed double figures of independent orthogroups of this family. A similar observation was also reported previously by Jackson *et al.*, 2012 reporting that the *T. vivax* VSG repertoires are more structurally distinct than the other two trypanosomes. Interestingly, however, it showed no noticeable orthologues to the insect stage procyclic-associated transferrin-like proteins. Contrarily, *T. brucei* and *T. congolense* showed both species-specific and shared sequences of this family, suggesting the importance of these surface proteins to survive the Tsetse midgut environment.

Whereas, ESAG11-related sequences are exclusive to *T. brucei* and the large repertoire of *T. congolense* specific VSG-associated genes, *T. vivax* exhibited specifically two heterogeneous GARPs orthogroups. Interestingly, although *T. vivax* has shared a small subset of putative BARP-like proteins (OG0005729) (Table 5.6), with the other African trypanosomes, its repertoire of GARPs proteins exhibited a phylogenic distinctive groups, which surprisingly did not cluster with the closest organism *T. congolense* GARPs. Rather, showed a phylogenic relationship to the *T. brucei* BARPs. Remarkably, Jackson *et al.*, (2015) found one gene of this subset that was favourably transcribed during blood forms of *T. vivax* and other members of this gene family were transcribed

during epimastigote and metacyclic forms, whilst the other African trypanosomes showed copiousness of such sequences during procyclic and epimastigote but not the blood forms. Therefore, these findings along with the Haptoglobin-Hemoglobin comparative phylogenetic analysis might suggest a possible broader role of subsets of this gene family during different stages of the *T. vivax* life cycle.

**Table 5.9 Species-specific orthogroups of African trypanosomes (*T. brucei*, *T. congolense* and *T. vivax*).** *T. vivax* showed the highest number of specific protein families. The majority of these clusters are surface proteins.

<b><i>T. brucei</i></b>		
<b>Orthogroup</b>	<b>Sequences</b>	<b>Functional description</b>
OG0000045	66	variant surface glycoprotein (VSG)
OG0000058	58	hypothetical protein
OG0000176	21	variant surface glycoprotein (VSG)
OG0000237	17	retrotransposon hot spot (RHS) protein
OG0000268	16	variant surface glycoprotein (VSG)
OG0000628	11	65 kDa invariant surface glycoprotein
OG0006993	4	ESAG11-related protein
OG0006994	4	procyclin-associated gene 2-like protein
OG0007448	3	procyclin-associated gene 2 (PAG2) protein
OG0008501	2	hypothetical protein
OG0008502	2	hypothetical protein
<b>Total</b>	<b>204</b>	
<b><i>T. congolense</i></b>		
OG0000006	418	Trypanosomal VSG domain containing protein, putative
OG0000008	300	VSG-associated congolense specific gene
OG0000064	49	Invariant surface glycoprotein, putative
OG0000089	35	Trypanosomal VSG domain containing protein, putative
OG0000262	16	Trypanosomal VSG domain containing protein, putative
OG0000616	11	procyclin-associated gene 2-like protein, putative
OG0000791	10	Retrotransposon hot spot protein, putative
OG0001124	9	hypothetical protein, conserved
OG0006304	5	hypothetical protein, conserved
OG0007698	2	Tubulin/FtsZ family, GTPase domain containing protein, putative
OG0007735	2	Shoulder domain containing protein, putative
<b>Total</b>	<b>857</b>	
<b><i>T. vivax</i></b>		
OG0000021	127	Trypanosomal VSG domain containing protein putative; Pfam
OG0000028	90	Glutamic acid/alanine-rich protein of Trypanosoma putative; Pfam
OG0000036	76	hypothetical protein conserved
OG0000039	72	Trypanosomal VSG domain containing protein putative; Pfam
OG0000042	69	hypothetical protein conserved
OG0000043	69	Trypanosomal VSG domain containing protein putative; Pfam
OG0000056	59	hypothetical protein conserved
OG0000059	57	hypothetical protein conserved
OG0000073	39	Trypanosomal VSG domain containing protein putative; Pfam
OG0000081	37	hypothetical protein conserved
OG0000098	33	Trypanosomal VSG domain containing protein putative; Pfam
OG0000111	29	retrotransposon hot spot protein 4 (RHS4) putative
OG0000128	26	hypothetical protein conserved
OG0000143	24	Trypanosomal VSG domain containing protein putative; Pfam
OG0000165	22	retrotransposon hot spot protein 4 (RHS4) putative
OG0000200	19	Retrotransposon hot spot protein putative; Pfam
OG0000201	19	Invariant surface glycoprotein putative; Pfam
OG0000202	19	hypothetical protein conserved
OG0000267	16	hypothetical protein conserved
OG0000315	15	hypothetical protein conserved
OG0000376	14	Trypanosome variant surface glycoprotein (A-type) putative; Pfam
OG0000626	11	Retrotransposon hot spot protein putative; Pfam
OG0000627	11	Zinc finger C2H2 type
OG0000807	10	hypothetical protein conserved
OG0002169	8	Trypanosomal VSG domain containing protein putative; Pfam
OG0004931	7	retrotransposon hot spot protein 4 (RHS4) putative
OG0006105	6	hypothetical protein conserved
OG0006106	6	Invariant surface glycoprotein putative; Pfam
OG0008485	2	hypothetical protein conserved
OG0008486	2	hypothetical protein conserved
OG0008487	1	hypothetical protein
OG0008490	2	Glutamic acid/alanine-rich protein of Trypanosoma putative; Pfam
<b>Total</b>	<b>997</b>	

## 5.4 Conclusions

The protein clustering analyses in this chapter highlight the important aspects of the kinetoplastid genomes, like the core gene sets shared between all kinetoplastids, but most importantly those belonging to parasitic species and which provide potential gene sets which could serve as potential drug targets to control these pathogens. These genes could also provide the bases for diagnostic tools or vaccine targets.

Notably, the analysis unraveled a possible set of putative genomic loci encoding for the haemoglobin receptor in *T. congolense* and *T. vivax*, which presented a relevance to the *T. brucei* HpHbR. Such putative genes were previously not determined in both first mentioned trypanosomes. However, it was also referred that *T. vivax* lacks the sets of blood form transferrin-like receptors (ESAG6 and ESAG7) found in both *T. brucei* and *T. congolense* and the transferrin like procyclic associated genes of the insect stage. These findings are in congruence with Jackson *et al.*, 2015. Confirming that such gene repertoires might be absent from *T. vivax*, what it could be the alternative for *T. vivax* to obtain this important element? The analysis here suggested only one set of HpHbR putative genes which were highly expanded and showed phylogenetic diversification in the genome of *T. vivax* and were assumed to encode for the haemoglobin receptor, which might propose an expanded role of this gene family in different life stages; however, such suggestions need to be proved experimentally.

Finally, this study also proposes that the genetic repertoires of parasitic kinetoplastids and more specifically African trypanosomes mainly differ in the number of contents and types of the cell surface components, which could be the armament that each species requires in order to withstand the environmental niches, and might play a key role in the differences of the life cycle of each trypanosome.

## Chapter 6 Conclusion and future directions

### 6.1 Conclusions

In this project, *de novo* genome assemblies and annotations of two neglected African Trypanosoma species (two strains of *T. congolense* IL3000, Tc1/148 and *T. vivax* IL1392) were generated, using the latest third generation sequencing technology, PacBio SMRT sequencing, that is able to generate long reads, with has the potential to produce longer assembly contigs with more prospective genomic physical integrity.

#### 6.1.1 Genomic rearrangements and new findings in African trypanosomes

Inter-species chromosomal structural translocations are mostly linked with environmental adaptation and speciation (Ayala *et al.*, 2011; Berg *et al.*, 2017).

*De novo* genome assembly and sequence annotation to reveal possible protein-coding and non-coding genomic features were achieved. The whole genome sequence comparison revealed potential structural genomic variation between *T. congolense* and the closest trypanosome *T. b. brucei* genome assembly (Chapter two), an event that failed to be detected by the current available draft Sanger based sequence assembly of *T. congolense*. These large inter and intra-chromosomal rearrangements between the two trypanosomes were found to be targeted to genes located in the internal chromosomal regions that involve more likely housekeeping genes of some MBCs. Generally, the putative genes affected by the putative chromosomal translocations encode for proteins participating in different roles in the cell, like cell structure and movement (flagellar component and flagellar rod), DNA and RNA modelling, ribosomal, transcription, surface proteins that are involved in the parasite pathogenesis, and metabolite transportation.

The analysis also shed the light on other important chromosomal features of *T. congolense* genome in regard to the sequence structure of the satellite repeats of putative MBCs centromeres, which showed remarkable low GC% and two sequence patterns were found as (136, 247) bp repeat units in two

subsets of *T. congolense* MBCs (MBC1, MBC3, and MBC6) and the other set (MBC4 and MBC11), respectively. Furthermore, a detailed description of the strand switch regions was presented, and these genomic regions have particular importance in trypanosomes as they serve as transcriptional initiation and termination sites of directional gene clusters in the core regions of the MBCs. In this context, the analysis suggested higher GC content in convergent SSRs in comparison to the divergent SSRs; additionally, the former sites are more likely to possess *tRNA* genes, which might suggest polycistronic transcription termination roles, as in *T. brucei*. However, divergent SSRs contain transposable elements *ingi* and conserved AT-rich DNA motifs that could serve as histone or polymerase II deposition sites, as previously noticed in *T. brucei* (Respuela *et al.*, 2008; Siegel *et al.*, 2009).

Protein clustering approach of the product of the annotated genes suggested 519 new gene families shared with *T. brucei* containing 604 proposed genes of *T. congolense*, and the functional enrichment analysis of these genes revealed their roles in different cellular and pathogenic aspects. Interestingly, the genomic localization of these genes was mostly in the internal regions of MBCs. Additionally, regarding other set of genes, 291 unique single genes with known functions were analysed and their localization is more likely to happen in the subtelomeric regions of the chromosomes, suggesting their *T. congolense* specific roles. They might involve in different cell functions like transportation of metabolites, DNA and RNA replication, protein folding, structural proteins, microtubule movement and host immune evasion mechanism; proposing an important subset of genes that could be potential drug or vaccine targets.

In Chapter four, a *de novo* genome assembly and annotation was carried out on *T. vivax* IL1392 genomic DNA in order to achieve more physical completeness of the genomic territories of this organism, as the current Sanger based draft genome is highly fragmented and interrupted with many gaps, that could affect the downstream analyses such as genomic structural and genome population studies. The adopted approach resulted in relatively super-long contigs that enabled us to discover different novel aspects in *T. vivax* genome:

Large chromosomal rearrangements on a large scale were recovered. These large-scale translocations hampered the efforts to generate chromosomal level assembly based on synteny to the *T. brucei* MBCs. Detailed description of the putative genes affected by these proposed translocations were presented in detail. Protein clustering analysis with *T. brucei* and the current Sanger assembly proteomic database revealed the presence of 7,292 putative new unique genes to the *T. vivax* IL1392 PacBio assembly; 6,444 with known functional assignment. The functional annotation analysis of these putative genes showed that their protein products could be involved in various metabolic pathways, transportation, cell structure, and evasion from the host immune factors.

The strand switch regions analysis revealed similar trends to those found in *T. congolense* in chapter two in terms of GC content and features in divergent or convergent SSRs. These findings are in consistent with other trypanosomatid genomes, suggested conserved roles for these regions throughout the trypanosomatids.

Noteworthy, long read SMRT sequencing and structural genomic approaches have been recently used in improving available genome assemblies in a number of protozoan species and analysis of chromosomal rearrangements. In accordance with our findings of chromosomal rearrangement among the three African trypanosome species, large-scale chromosomal rearrangement and formation of mosaic chromosomes were also found between different pathogenic groups of *Leishmania* (Britto *et al.*, 1998). Moreover, resequencing of the *L. infantum* genome using PacBio SMRT sequencing was used to improve the genome reference sequence of this parasite and was able to reveal more physical linkage of complete sequences of 36 chromosomes, showing more tandemly repeated genes and resolving ambiguous nucleotides in the current reference genome (González-De La Fuente *et al.*, 2017).

Genomic structural rearrangement in some genomic sites was also noticed between two species of *Crithidia*. (*C. bombi* and *C. expoeki*) at a scaffold level of PacBio sequence, while they shared high genomic synteny (Schmid-Hempel *et al.*, 2018).

Other studies showed that within-species chromosomal heterogeneity in size and content between different strains of *G. intestinalis* were observed in this human pathogen (Le Blancq, Korman and van der Ploeg, 1992; Ankarklev *et al.*, 2015).

The ability of PacBio SMRT in combination with Hi-C data in improving genome assembly in comparison to the available Sanger-based reference genome of *Plasmodium knowlesi* was recently published; showed noticeable improvement in structural contiguity and revealed incorrect assignments of some genomic regions by the Sanger assembly (Lapp *et al.*, 2018), with more representation of genes and pseudogenes in subtelomeric regions (Pasini *et al.*, 2017).

Remarkably, inter and intra-chromosomal rearrangements were also retrieved from the medically and important zoonotic species of genus *Plasmodium* based on genomic structural analysis of PacBio SMRT genome assemblies. This showed predicted genome synteny and synteny breakage across different species of this parasite (*P. malariae*, *P. vivax*, *P. falciparum*, *P. ovale curtisi* and *P. o. wallikeri*) (Rutledge *et al.*, 2017).

### **6.1.2 *T. congolense* minichromosomes and potential expression sites**

A detailed description of protein-coding and conserved non-coding DNA sequences for a novel African trypanosomes subset of chromosomes was achieved for the first time in *T. congolense* genome (Chapter three). Minichromosomes are characterized by their high repeated DNA motifs and they involve mainly *VSG* genes that are used in the parasite evasion from the mammalian host immune system. General conserved trend was noticed in the structure of these chromosomes. They characterized by central repeated DNA sequence of 369 bp motifs of varying number that delineate the length of each mini-chromosome, flanked by two conserved non-coding DNA sequences followed by a subtelomeric region of length about 5 kb hosting variable features; however, it was characterized mainly by distal *VSG* gene/pseudogenes that end before the start of the distinctive repeats at the two ends of the linear eukaryotic chromosomes (telomeres) of a sequence unit

consist of TTAGGG. Furthermore, expression sites of *T. congolense* has been described in this work most likely to have similar structure to the canonical structure of minichromosomes (Appendix B).

### **6.1.3 Phylogenomic analysis of Kinetoplastids and African Trypanosomes**

A comparative phylogenomic study was carried out on seven kinetoplastid organisms that live in different environmental niches such as free living, parasitic on vertebrate and non-vertebrate hosts, and extra and intracellular pathogens (Chapter five). The analysis suggested 36,411 core genes within 4,163 gene families representing shared genes across all analysed kinetoplastids. Possible gene expansion and contraction between the free living and the parasitic ones suggested high variation devoted mainly to the surface proteins. A similar analysis to genes shared among the parasitic species revealed 673 gene families shared amongst the parasitic kinetoplastids, mainly consisting of 1:1 orthologous genes (449/673). The functional enrichment analysis of this set of genes revealed important information as they are involved in cell growth, metabolic processes, transport and drug metabolism suggesting a particular importance as they could serve as candidates for drug and/or vaccine targeting.

The African trypanosomes showed specific gene families shared between the three African trypanosomes, with the analysis focused on Haptoglobin-Hemoglobin receptor variants with apparent extreme gene contraction in *T. brucei* (one gene on MBC six) and the highest expansion noticed in *T. vivax* genome (20) copies, while the vast majority of *T. congolense* members (5/6 copies) showed tandem expansion on MBC 10. Moreover, the analysis of genes shared between *T. congolense* and *T. brucei* recovered genes that have particular importance during the insect midgut stages (a shared element of the life cycle between the two trypanosomes). Meanwhile, the *T. vivax* genome exhibited more species-specific gene families in comparison to the other African trypanosomes presented in this study.

To sum up, the analysis done in this project shed light on a number of possible novel findings in the African trypanosomes genomes. The proposed chromosomal rearrangements might have shaped the speciation and evolution of these trypanosomes. The detailed description of the *T. congolense* mini-chromosomes presented here could help for more understanding of VSG gene expression from this repertoire, and provided the foundation and, a useful database for future work on this enigmatic set of chromosomes. Finally, the analysis of gene sets among selected kinetoplastids living in different niches presented by whole proteomics-phylogenomic analysis revealed genes that were involved mainly in parasitism and those which evolved perhaps to cope with specific environmental stresses.

## 6.2 Future perspective

This project has opened other questions need to be answered and more work to be done, such as:

- 1- What are the possible structural differences on genomic level among different strains or geographically separated trypanosomes (i.e. population genomic studies)?
- 2- What are the variants of *T. brucei* mini-chromosomes and how do they differ from those of *T. congolense* or even between different strains?
- 3- Achieving chromosomal-level assembly of *T. vivax* using techniques other than synteny to a reference assembly like FISH technique, optical mapping or Hi-C technique, or a combination of these techniques.
- 4- More in depth investigation based on PCR amplification and verification is needed on singleton genes of *T. congolense* strains and *T. vivax* predicted in this project followed by testing of selective pressure on this subset of genes using  $d_N/d_S$  analysis to study the possible evolutionary history and importance of these genes to the parasite.
- 5- Further investigation of phylogenomics to show events of possible genomic duplication or gene loss and gain events on phylogenetic trees by using an appropriate statistical tool like CAFE to test gene expansion or contraction of certain gene families.

## References

- Aas, F. E. *et al.* (2002) 'Competence for natural transformation in *Neisseria gonorrhoeae*: Components of DNA binding and uptake linked to type IV pilus expression', *Molecular Microbiology*, 46(3), pp. 749–760. doi: 10.1046/j.1365-2958.2002.03193.x.
- Van Den Abbeele, J. *et al.* (1999) 'Trypanosoma brucei spp. development in the tsetse fly: characterization of the post-mesocyclic stages in the foregut and proboscis.', *Parasitology*, 118 ( Pt 5(5), pp. 469–478. doi: 10.1017/S0031182099004217.
- Akiyoshi, B. and Gull, K. (2013) 'Evolutionary cell biology of chromosome segregation: insights from trypanosomes', *Open Biology*, 3(5), pp. 130023–130023. doi: 10.1098/rsob.130023.
- Aksoy, S. (1991) 'Site-specific retrotransposons of the trypanosomatid protozoa', *Parasitology Today*, 7(10), pp. 281–285. doi: 10.1016/0169-4758(91)90097-8.
- Al-Khedery, B. and Allred, D. R. (2006) 'Antigenic variation in *Babesia bovis* occurs through segmental gene conversion of the ves multigene family, within a bidirectional locus of active transcription', *Molecular Microbiology*, 59(2), pp. 402–414. doi: 10.1111/j.1365-2958.2005.04993.x.
- Alemayehu, B. and Alemayehu, M. (2017) 'Leishmaniasis: A Review on Parasite, Vector and Reservoir Host', *Health Science Journal*, 11(4). doi: 10.21767/1791-809X.1000519.
- Alexandre, S. *et al.* (1988) 'Putative genes of a variant-specific antigen gene transcription unit in *Trypanosoma brucei*.' , *Molecular and cellular biology*, 8(6), pp. 2367–78. doi: 10.1128/MCB.8.6.2367.Updated.
- Alexandre, S. *et al.* (1990) 'Differential expression of a family of putative adenylate/guanylate cyclase genes in *Trypanosoma brucei*', *Molecular and Biochemical Parasitology*, 43(2), pp. 279–288. doi: 10.1016/0166-6851(90)90152-C.

- Allred, D. R. *et al.* (2000) 'The ves multigene family of *B. bovis* encodes components of rapid antigenic variation at the infected erythrocyte surface', *Molecular Cell*, 5(1), pp. 153–162. doi: 10.1016/S1097-2765(00)80411-6.
- Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool.', *Journal of molecular biology*, 215(3), pp. 403–10. doi: 10.1016/S0022-2836(05)80360-2.
- Amin, M. L. (2013) 'P-glycoprotein inhibition for optimal drug delivery', *Drug Target Insights*, pp. 27–34. doi: 10.4137/DTI.S12519.
- Ankarklev, J. *et al.* (2015) 'Comparative genomic analyses of freshly isolated *Giardia intestinalis* assemblage A isolates', *BMC Genomics*, 16(1). doi: 10.1186/s12864-015-1893-6.
- Armour, J. A. L. (2006) 'Tandemly repeated DNA: why should anyone care?', *Mutation research*, 598(1–2), pp. 6–14. doi: 10.1016/j.mrfmmm.2006.01.013.
- Assefa, S. *et al.* (2009) 'ABACAS: algorithm-based automatic contiguation of assembled sequences.', *Bioinformatics (Oxford, England)*, 25(15), pp. 1968–9. doi: 10.1093/bioinformatics/btp347.
- Athman, R. and Philpott, D. (2004) 'Innate immunity via Toll-like receptors and Nod proteins', *Current Opinion in Microbiology*, pp. 25–32. doi: 10.1016/j.mib.2003.12.013.
- Ayala, D. *et al.* (2011) 'Chromosomal inversions, natural selection and adaptation in the malaria vector *Anopheles funestus*', *Molecular Biology and Evolution*, 28(1), pp. 745–758. doi: 10.1093/molbev/msq248.
- Ayala, D., Ullastres, A. and González, J. (2014) 'Adaptation through chromosomal inversions in *Anopheles*', *Frontiers in Genetics*. doi: 10.3389/fgene.2014.00129.
- Bacolla, A. *et al.* (2008) 'Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties', *Genome Research*, 18(10), pp. 1545–1553. doi: 10.1101/gr.078303.108.
- Bailey, J. A. and Eichler, E. E. (2006) 'Primate segmental duplications:

Crucibles of evolution, diversity and disease', *Nature Reviews Genetics*, pp. 552–564. doi: 10.1038/nrg1895.

Baldi, S. and Becker, P. B. (2013) 'The variant histone H2A.V of *Drosophila* - Three roles, two guises', *Chromosoma*, pp. 245–258. doi: 10.1007/s00412-013-0409-x.

Barry, J. D. *et al.* (1998) 'VSG gene control and infectivity strategy of metacyclic stage *Trypanosoma brucei*', *Molecular and Biochemical Parasitology*, pp. 93–105. doi: 10.1016/S0166-6851(97)00193-X.

Barry, J. D. and McCulloch, R. (2001) 'Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite.', *Advances in parasitology*, 49, pp. 1–70. doi: 10.1016/S0065-308X(01)49037-3.

Baruch, D. I. *et al.* (1995) 'Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes', *Cell*, 82(1), pp. 77–87. doi: 10.1016/0092-8674(95)90054-3.

Bastin, P. *et al.* (1999) 'Flagellar morphogenesis: protein targeting and assembly in the paraflagellar rod of trypanosomes.', *Molecular and cellular biology*, 19(12), pp. 8191–8200. doi: 10.1128/MCB.19.12.8191.

Batram, C. *et al.* (2014) 'Expression site attenuation mechanistically links antigenic variation and development in *Trypanosoma brucei*', *eLife*, 3, p. e02324. doi: 10.7554/eLife.02324.

Bayne, R. A. L. *et al.* (1993) 'A major surface antigen of procyclic stage *Trypanosoma congolense*', *Molecular and Biochemical Parasitology*, 61(2), pp. 295–310. doi: 10.1016/0166-6851(93)90075-9.

Bayoh, M. N., Thomas, C. J. and Lindsay, S. W. (2001) 'Mapping distributions of chromosomal forms of *Anopheles gambiae* in West Africa using climate data', *Medical and Veterinary Entomology*, 15(3), pp. 267–274. doi: 10.1046/j.0269-283X.2001.00298.x.

Becker, M. *et al.* (2004) 'Isolation of the repertoire of VSG expression site containing telomeres of *Trypanosoma brucei* 427 using transformation-associated recombination in yeast', *Genome Research*, 14, pp. 2319–2329. doi: 10.1101/gr.2955304.

Bednar, J. *et al.* (1998) 'Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin.', *Proceedings of the National Academy of Sciences of the United States of America*, 95(24), pp. 14173–8. doi: 10.1073/pnas.95.24.14173.

Beecroft, R. P., Roditi, I. and Pearson, T. W. (1993) 'Identification and characterization of an acidic major surface glycoprotein from procyclic stage *Trypanosoma congolense*', *Molecular and Biochemical Parasitology*, 61(2), pp. 285–294. doi: 10.1016/0166-6851(93)90074-8.

Belli, S. I. (2000) 'Chromatin remodelling during the life cycle of trypanosomatids', *International Journal for Parasitology*, pp. 679–687. doi: 10.1016/S0020-7519(00)00052-7.

Bengaly, Z. *et al.* (2002) 'Comparative pathogenicity of three genetically distinct *Trypanosoma congolense*-types in inbred Balb/c mice', *Veterinary Parasitology*, 105(2), pp. 111–118. doi: 10.1016/S0304-4017(01)00609-4.

Benson, G. (1999) 'Tandem Repeats Finder: a program to analyse DNA sequences', *Nucleic Acids Res.*, 27(2), pp. 573–578.

Benson, G. (1999) 'Tandem repeats finder: A program to analyze DNA sequences', *Nucleic Acids Research*, 27(2), pp. 573–580. doi: 10.1093/nar/27.2.573.

Bentkowski, P., Van Oosterhout, C. and Mock, T. (2015) 'A model of genome size evolution for prokaryotes in stable and fluctuating environments', *Genome Biology and Evolution*, 7(8), pp. 2344–2351. doi: 10.1093/gbe/evv148.

Berg, P. R. *et al.* (2016) 'Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod', *Scientific Reports*, 6. doi: 10.1038/srep23246.

Berg, P. R. *et al.* (2017) 'Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions', *Heredity*, 119(6), pp. 418–428. doi: 10.1038/hdy.2017.54.

Berlin, K. *et al.* (2015) 'Assembling large genomes with single-molecule sequencing and locality-sensitive hashing.', *Nature biotechnology*, 33(6), pp. 623–630. doi: 10.1038/nbt.3238.

Bermudez-Santana, C. *et al.* (2010) 'Genomic organization of eukaryotic tRNAs', *BMC Genomics*, 11(1). doi: 10.1186/1471-2164-11-270.

Berriman, M. *et al.* (2002) 'The architecture of variant surface glycoprotein gene expression sites in *Trypanosoma brucei*', *Mol Biochem Parasitol*, 122, pp. 131–140. doi: 10.1016/S0166-6851(02)00092-0.

Berriman, M. (2005) 'The Genome of the African Trypanosome *Trypanosoma brucei*', *Science*, 309(5733), pp. 416–422. doi: 10.1126/science.1112642.

Bhattacharya, S., Bakre, A. and Bhattacharya, A. (2002) 'Mobile genetic elements in protozoan parasites.', *Journal of genetics*, 81(2), pp. 73–86. doi: 10.1007/BF02715903.

De Bie, T. *et al.* (2006) 'CAFE: a computational tool for the study of gene family evolution.', *Bioinformatics (Oxford, England)*, 22(10), pp. 1269–71. doi: 10.1093/bioinformatics/btl097.

Le Blancq, S. M., Korman, S. H. and van der Ploeg, L. H. t (1992) 'Spontaneous chromosome rearrangements in the protozoan *Giardia lamblia*: Estimation of mutation rates', *Nucleic Acids Research*, 20(17), pp. 4539–4545. doi: 10.1093/nar/20.17.4539.

Boetzer, M. and Pirovano, W. (2014) 'SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information.', *BMC bioinformatics*, 15(1), p. 211. doi: 10.1186/1471-2105-15-211.

Van Den Bossche, P. *et al.* (2011) 'Virulence in *Trypanosoma congolense* Savannah subgroup. A comparison between strains and transmission cycles',

*Parasite Immunology*, pp. 456–460. doi: 10.1111/j.1365-3024.2010.01277.x.

Brandenburg, J. *et al.* (2007) 'Multifunctional class I transcription in *Trypanosoma brucei* depends on a novel protein complex.', *The EMBO journal*, 26(23), pp. 4856–66. doi: 10.1038/sj.emboj.7601905.

Branquinha, M. H. *et al.* (2013) 'Calpains: potential targets for alternative chemotherapeutic intervention against human pathogenic trypanosomatids.', *Current medicinal chemistry*, 20(25), pp. 3174–85. doi: 10.2174/0929867311320250010.

Bringaud, F. *et al.* (2004) 'The ingi and RIME non-LTR Retrotransposons are Not Randomly Distributed in the Genome of *Trypanosoma brucei*', *Molecular Biology and Evolution*, 21(3), pp. 520–528. doi: 10.1093/molbev/msh045.

Britto, C. *et al.* (1998) 'Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World *Leishmania* genomes', *Gene*, 222(1), pp. 107–117. doi: 10.1016/S0378-1119(98)00472-7.

Brun, R. *et al.* (2010) 'Human African trypanosomiasis.', *The Lancet*, 375(9709), pp. 148–159. doi: 10.1016/S0140-6736(09)60829-1.

Bryant, J. M. *et al.* (2017) 'CRISPR/Cas9 Genome Editing Reveals That the Intron Is Not Essential for *var2csa* Gene Activation or Silencing in *Plasmodium falciparum*', *mBio*, 8(4), pp. e00729-17. doi: 10.1128/mBio.00729-17.

Bubliy, O. A. and Loeschcke, V. (2002) 'Effect of low stressful temperature on genetic variation of five quantitative traits in *Drosophila melanogaster*', *Heredity*, 89(1), pp. 70–75. doi: 10.1038/sj.hdy.6800104.

Burns, K. H. (2017) 'Transposable elements in cancer', *Nature Reviews Cancer*, pp. 415–424. doi: 10.1038/nrc.2017.35.

Burrows, L. L. (2012) 'Prime time for minor subunits of the type II secretion and type IV pilus systems', *Molecular Microbiology*, 86(4), pp. 765–769. doi: 10.1111/mmi.12034.

Bütikofer, P. *et al.* (2002) 'Glycosylphosphatidylinositol-anchored surface

molecules of *Trypanosoma congolense* insect forms are developmentally regulated in the tsetse fly', *Molecular and Biochemical Parasitology*, 119(1), pp. 7–16. doi: 10.1016/S0166-6851(01)00382-6.

Camacho, C. *et al.* (2009) 'BLAST plus: architecture and applications', *BMC Bioinformatics*, 10(421), p. 1. doi: Artn 421\nDoi 10.1186/1471-2105-10-421.

Camargo, E. P. (1999) 'Phytomonas and other trypanosomatid parasites of plants and fruit.', *Advances in parasitology*, 42, pp. 29–112. doi: [http://dx.doi.org/10.1016/S0065-308X\(08\)60148-7](http://dx.doi.org/10.1016/S0065-308X(08)60148-7).

Capy, P. (1998) 'A plastic genome', *Nature*, pp. 522–523. doi: 10.1038/25007.

Capy, P. *et al.* (2000) 'Stress and transposable elements: Co-evolution or useful parasites?', *Heredity*, pp. 101–106. doi: 10.1046/j.1365-2540.2000.00751.x.

Carruthers, L. M. *et al.* (1998) 'Linker histones stabilize the intrinsic salt-dependent folding of nucleosomal arrays: Mechanistic ramifications for higher-order chromatin folding', *Biochemistry*, 37(42), pp. 14776–14787. doi: 10.1021/bi981684e.

de Carvalho, E. F. *et al.* (1990) 'HSP 70 gene expression in *Trypanosoma cruzi* is regulated at different levels', *Journal of Cellular Physiology*, 143(3), pp. 439–444. doi: 10.1002/jcp.1041430306.

Carver, T. *et al.* (2012) 'Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data', *Bioinformatics*, 28(4), pp. 464–469. doi: 10.1093/bioinformatics/btr703.

Carver, T. J. *et al.* (2005) 'ACT: The Artemis comparison tool', *Bioinformatics*, 21(16), pp. 3422–3423. doi: 10.1093/bioinformatics/bti553.

Casacuberta, E. and González, J. (2013) 'The impact of transposable elements in environmental adaptation', *Molecular Ecology*, pp. 1503–1517. doi: 10.1111/mec.12170.

Chaisson, M. J. P., Wilson, R. K. and Eichler, E. E. (2015) 'Genetic variation

and the de novo assembly of human genomes', *Nature Reviews Genetics*, 16(11), pp. 627–640. doi: 10.1038/nrg3933.

Chaisson, M. J. and Tesler, G. (2012) 'Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory', *BMC Bioinformatics*, 13(1), p. 238. doi: 10.1186/1471-2105-13-238.

Chang, D. and Duda, T. F. (2012) 'Extensive and continuous duplication facilitates rapid evolution and diversification of gene families', *Molecular Biology and Evolution*, 29(8), pp. 2019–2029. doi: 10.1093/molbev/mss068.

Chaves, I. *et al.* (1999) 'Control of variant surface glycoprotein gene-expression sites in *Trypanosoma brucei*', *EMBO Journal*, 18(17), pp. 4846–4855. doi: 10.1093/emboj/18.17.4846.

Chen, H. and Boutros, P. C. (2011) 'VennDiagram: A package for the generation of highly-customizable Venn and Euler diagrams in R', *BMC Bioinformatics*, 12. doi: 10.1186/1471-2105-12-35.

Chénais, B. (2013) 'Transposable elements and human cancer: A causal relationship?', *Biochimica et Biophysica Acta*, 1835(1), pp. 28–35. doi: 10.1016/j.bbcan.2012.09.001.

Cherenet, T. *et al.* (2004) 'Seasonal prevalence of bovine trypanosomosis in a tsetse-infested zone and a tsetse-free zone of the Amhara Region, north-west Ethiopia.', *The Onderstepoort journal of veterinary research*, 71, pp. 307–312. doi: 10.4102/ojvr.v71i4.250.

Chin *et al.* (2013) *Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data*, *Nature methods*. Available at: <http://www.nature.com/nmeth/journal/v10/n6/pdf/nmeth.2474.pdf> (Accessed: 13 July 2015).

Chin *et al. et al.* (2013) 'Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data', *Nature methods*, 10(6), pp. 563–571. doi: 10.1038/nmeth.2474.

Christiane Hertz-Fowler, H. R. and M. B. (2007) *Trypanosomes: After The Genome*. Edited by U. David Barry, Richard McCulloch, Jeremy Mottram and Alvaro Acosta-Serrano Wellcome Centre for Molecular Parasitology, University of Glasgow, Glasgow G12 8QQ. Glasgow G12 8QQ, UK: Horizon Bioscience. Available at: <http://www.caister.com/backlist/horizonbioscience/try.html>.

Chung, H. M. *et al.* (1990) 'Architectural organization in the interphase nucleus of the protozoan *Trypanosoma brucei*: location of telomeres and mini-chromosomes.', *The EMBO journal*, 9(8), pp. 2611–9. doi: 10.1016/0168-9525(90)90290-M.

Clancy, S. and Shaw, K. M. (2008) 'DNA Deletion and Duplication and the Associated Genetic Disorders', *Nature education*, 1(23), pp. 5–9. Available at: <http://www.nature.com/scitable/topicpage/dna-deletion-and-duplication-and-the-associated-331>.

Cortez, A. P. *et al.* (2006) 'The taxonomic and phylogenetic relationships of *Trypanosoma vivax* from South America and Africa', *Parasitology*, 133(2), pp. 159–169. doi: 10.1017/S0031182006000254.

Coustou, V. *et al.* (2010) 'Complete in vitro life cycle of *Trypanosoma congolense*: Development of genetic tools', *PLoS Neglected Tropical Diseases*, 4(3). doi: 10.1371/journal.pntd.0000618.

Craig, N. L. (1997) 'Target site selection in transposition', *Annu.Rev.Biochem.*, 66, pp. 437–474.

Cremer, T. and Cremer, M. (2010) 'Chromosome territories.', *Cold Spring Harbor perspectives in biology*. doi: 10.1101/cshperspect.a003889.

Croft, J. A. *et al.* (1999) 'Differences in the localization and morphology of chromosomes in the human nucleus', *Journal of Cell Biology*, 145(6), pp. 1119–1131. doi: 10.1083/jcb.145.6.1119.

Cross, G. A. M. M., Kim, H.-S. S. and Wickstead, B. (2014) 'Capturing the variant surface glycoprotein repertoire (the VSGnome) of *Trypanosoma brucei*

Lister 427', *Molecular and Biochemical Parasitology*. Elsevier B.V., 195(1), pp. 59–73. doi: 10.1016/j.molbiopara.2014.06.004.

D'Archivio, S. *et al.* (2011) 'Genetic engineering of trypanosoma (duttonella) vivax and in vitro differentiation under axenic conditions', *PLoS Neglected Tropical Diseases*, 5(12). doi: 10.1371/journal.pntd.0001461.

Daher, W. *et al.* (2007) 'A Toxoplasma gondii leucine-rich repeat protein binds phosphatase type 1 protein and negatively regulates its activity', *Eukaryotic Cell*, 6(9), pp. 1606–1617. doi: 10.1128/EC.00260-07.

Dalmaso, M. C., Sullivan, W. J. and Angel, S. O. (2011) 'Canonical and variant histones of protozoan parasites.', *Frontiers in bioscience : a journal and virtual library*, 16, pp. 2086–2105. doi: 10.2741/3841.

Daniels, J. P., Gull, K. and Wickstead, B. (2010) 'Cell Biology of the Trypanosome Genome', *Microbiology and Molecular Biology Reviews*, 74(4), pp. 552–569. doi: 10.1128/mnbr.00024-10.

Dayo, G. K. *et al.* (2010) 'Prevalence and incidence of bovine trypanosomosis in an agro-pastoral area of southwestern Burkina Faso', *Research in Veterinary Science*, 88(3), pp. 470–477. doi: 10.1016/j.rvsc.2009.10.010.

Dean, P. *et al.* (2014) 'Transport proteins of parasitic protists and their role in nutrient salvage', *Frontiers in Plant Science*, 5. doi: 10.3389/fpls.2014.00153.

DeBarry, J. D. and Kissinger, J. C. (2014) 'A survey of innovation through duplication in the reduced genomes of twelve parasites', *PLoS ONE*, 9(6). doi: 10.1371/journal.pone.0099213.

DeJesus, E. *et al.* (2013) 'A Single Amino Acid Substitution in the Group 1 Trypanosoma brucei gambiense Haptoglobin-Hemoglobin Receptor Abolishes TLF-1 Binding', *PLoS Pathogens*, 9(4). doi: 10.1371/journal.ppat.1003317.

Delcher, A. L. (2002) 'Fast algorithms for large-scale genome alignment and comparison', *Nucleic Acids Research*, 30(11), pp. 2478–2483. doi: 10.1093/nar/30.11.2478.

DeLotto, R. and Schedl, P. (1984) 'A *Drosophila melanogaster* transfer RNA gene cluster at the cytogenetic locus 90BC', *Journal of Molecular Biology*, 179(4), pp. 587–605. doi: 10.1016/0022-2836(84)90157-8.

Demura, M. *et al.* (2007) 'Regional rearrangements in chromosome 15q21 cause formation of cryptic promoters for the CYP19 (aromatase) gene', *Human Molecular Genetics*, 16(21), pp. 2529–2541. doi: 10.1093/hmg/ddm145.

Deschamps, P. *et al.* (2011) 'Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids', *Molecular Biology and Evolution*, 28(1), pp. 53–58. doi: 10.1093/molbev/msq289.

Desquesnes, M. and Dia, M. L. (2003) 'Trypanosoma vivax: Mechanical transmission in cattle by one of the most common African tabanids, *Atylotus agrestis*', *Experimental Parasitology*, 103(1–2), pp. 35–43. doi: 10.1016/S0014-4894(03)00067-5.

Dickin, S. K. and Gibson, W. C. (1989) 'Hybridisation with a repetitive DNA probe reveals the presence of small chromosomes in *Trypanosoma vivax*', *Molecular and Biochemical Parasitology*, 33(2), pp. 135–142. doi: 10.1016/0166-6851(89)90027-3.

Dirie, M. F. *et al.* (1993) 'Comparative studies of *Trypanosoma* (*Duttonella*) *vivax* isolates from Colombia', *Parasitology*, 106(1), pp. 21–29. doi: 10.1017/S0031182000074771.

Doležel, D. *et al.* (2000) 'Phylogeny of the bodonid flagellates (Kinetoplastida) based on small-subunit rRNA gene sequences', *International Journal of Systematic and Evolutionary Microbiology*, 50(5), pp. 1943–1951. doi: 10.1099/00207713-50-5-1943.

Dolinski, K. and Botstein, D. (2007) 'Orthology and Functional Conservation in Eukaryotes', *Annual Review of Genetics*, 41(1), pp. 465–507. doi: 10.1146/annurev.genet.40.110405.090439.

Donelson, J. E. (2003) 'Antigenic variation and the African trypanosome genome', in *Acta Tropica*, pp. 391–404. doi: 10.1016/S0001-706X(02)00237-

1.

Dongen, S. Van (2000) 'A cluster algorithm for graphs', *Information Systems [INS]*, (R 0010), pp. 1–40. doi: INS-R0010.

Downing, T. *et al.* (2012) 'Genome-wide SNP and microsatellite variation illuminate population-level epidemiology in the *Leishmania donovani* species complex', *Infection, Genetics and Evolution*. Elsevier B.V., 12(1), pp. 149–159. doi: 10.1016/j.meegid.2011.11.005.

Dreesen, O. and Cross, G. A. M. (2006) 'Consequences of telomere shortening at an active VSG expression site in telomerase-deficient *Trypanosoma brucei*', *Eukaryotic Cell*, 5(12), pp. 2114–2119. doi: 10.1128/EC.00059-06.

Dreesen, O. and Cross, G. A. M. (2008) 'Telomere length in *Trypanosoma brucei*.', *Experimental parasitology*, 118(1), pp. 103–110. doi: 10.1016/j.exppara.2007.07.016.

Dreesen, O., Li, B. and Cross, G. A. M. (2007) 'Telomere structure and function in trypanosomes: a proposal', *Nature*, 5(January), pp. 70–75.

Dubcovsky, J. and Dvorak, J. (1995) 'Ribosomal RNA multigene loci: Nomads of the triticeae genomes', *Genetics*, 140(4), pp. 1367–1377.

Duda, T. F. and Palumbi, S. R. (1999) 'Molecular genetics of ecological diversification: Duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*', *Proceedings of the National Academy of Sciences*, 96(12), pp. 6820–6823. doi: 10.1073/pnas.96.12.6820.

Duffy, C. W. *et al.* (2017) 'Population genetic structure and adaptation of malaria parasites on the edge of endemic distribution', *Molecular Ecology*, 26(11), pp. 2880–2894. doi: 10.1111/mec.14066.

Dunbar, D. A. *et al.* (2000) 'The genes for small nucleolar RNAs in *Trypanosoma brucei* are organized in clusters and are transcribed as a polycistronic RNA.', *Nucleic acids research*, 28(15), pp. 2855–61. doi: Doi

10.1093/Nar/28.15.2855.

Edgar, R. C. and Edgar, R. C. (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput.', *Nucleic acids research*, 32(5), pp. 1792–7. doi: 10.1093/nar/gkh340.

Eichler, E. E. and Sankoff, D. (2003) 'Structural dynamics of eukaryotic chromosome evolution.', *Science (New York, N.Y.)*, 301(5634), pp. 793–7. doi: 10.1126/science.1086132.

Eid, J. *et al.* (2009) 'Real-Time DNA Sequencing from Single Polymerase Molecules', *Science*, 323(5910), pp. 133–138. doi: 10.1126/science.1162986.

Eisen, J. A. (1998) 'Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis', *Genome Research*, 8(3), pp. 163–167. doi: 10.1101/gr.8.3.163.

Ekblom, R. and Wolf, J. B. W. W. (2014) 'A field guide to whole-genome sequencing, assembly and annotation', *Evolutionary Applications*, 7(9), pp. 1026–1042. doi: 10.1111/eva.12178.

El-Sayed, N. M. *et al.* (2005) 'Comparative genomics of trypanosomatid parasitic protozoa.', *Science (New York, N.Y.)*, 309, pp. 404–409. doi: 10.1126/science.1112181.

Elias, M. C. Q. B. *et al.* (2002) 'Chromosome localization changes in the *Trypanosoma cruzi* nucleus', *Eukaryotic Cell*, 1(6), pp. 944–953. doi: 10.1128/EC.1.6.944-953.2002.

Elsik, C. G., Tellam, R. L. and Worley, K. C. (2009) 'The Genome Sequence of Taurine Cattle: A window to ruminant biology and evolution', *Science (New York, N.Y.)*, 324(5926), pp. 522–528. doi: 10.1126/science.1169588.

Emms, D. M. and Kelly, S. (2015) 'OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy', *Genome Biology*, 16(1), p. 157. doi: 10.1186/s13059-015-0721-2.

Engstler, M. (2004) 'Kinetics of endocytosis and recycling of the GPI-anchored

variant surface glycoprotein in *Trypanosoma brucei*', *Journal of Cell Science*, 117(7), pp. 1105–1115. doi: 10.1242/jcs.00938.

Ersfeld, K. and Gull, K. (1997) 'Partitioning of large and minichromosomes in *Trypanosoma brucei*.', *Science*, 276(5312), pp. 611–614. doi: 10.1126/science.276.5312.611.

Ersfeld, K., Melville, S. E. and Gull, K. (1999) 'Nuclear and genome organization of *Trypanosoma brucei*.', *Parasitology today (Personal ed.)*, 15(2), pp. 58–63. doi: 10.1016/S0169-4758(98)01378-7.

Eshita, Y. *et al.* (1992) 'Metacyclic form-specific variable surface glycoprotein-encoding genes of *Trypanosoma (Nannomonas) congolense*', *Gene*, 113(2), pp. 139–148. doi: 10.1016/0378-1119(92)90389-7.

Eyford, B. A. *et al.* (2011) 'Differential protein expression throughout the life cycle of *Trypanosoma congolense*, a major parasite of cattle in Africa', *Molecular and Biochemical Parasitology*, 177(2), pp. 116–125. doi: 10.1016/j.molbiopara.2011.02.009.

Fagone, P. and Jackowski, S. (2013) 'Phosphatidylcholine and the CDP-choline cycle', *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids*, pp. 523–532. doi: 10.1016/j.bbalip.2012.09.009.

Farine, L. *et al.* (2015) 'Phosphatidylethanolamine and phosphatidylcholine biosynthesis by the Kennedy pathway occurs at different sites in *Trypanosoma brucei*', *Scientific Reports*, 5(1), p. 16787. doi: 10.1038/srep16787.

Fasogbon, A. I., Knowles, G. and Gardiner, P. R. (1990) 'A comparison of the isoenzymes of *Trypanosoma (Duttonella) vivax* isolates from East and West Africa', *International Journal for Parasitology*, 20(3), pp. 389–394. doi: 10.1016/0020-7519(90)90156-H.

Feder, J. L., Nosil, P. and Flaxman, S. M. (2014) 'Assessing when chromosomal rearrangements affect the dynamics of speciation: Implications from computer simulations', *Frontiers in Genetics*, 5(AUG). doi: 10.3389/fgene.2014.00295.

- Fenn, K. and Matthews, K. R. (2007) 'The cell biology of *Trypanosoma brucei* differentiation', *Current Opinion in Microbiology*, pp. 539–546. doi: 10.1016/j.mib.2007.09.014.
- Ferrarini, M. *et al.* (2013) 'An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome', *BMC Genomics*, 14(1), p. 670. doi: 10.1186/1471-2164-14-670.
- Field, M. C. *et al.* (2004) 'New approaches to the microscopic imaging of *Trypanosoma brucei*.', *Microscopy and microanalysis: the official journal of Microscopy Society of America, Microbeam Analysis Society, Microscopical Society of Canada*, 10(5), pp. 621–636. doi: 10.1017/S1431927604040942.
- Field, M. C. and Carrington, M. (2009) 'The trypanosome flagellar pocket', *Nature Reviews Microbiology*, pp. 775–786. doi: 10.1038/nrmicro2221.
- Finn, R. D. *et al.* (2016) 'The Pfam protein families database: Towards a more sustainable future', *Nucleic Acids Research*, 44(D1), pp. D279–D285. doi: 10.1093/nar/gkv1344.
- Folgueira, C. and Requena, J. M. (2007) 'A postgenomic view of the heat shock proteins in kinetoplastids', *FEMS Microbiology Reviews*, pp. 359–377. doi: 10.1111/j.1574-6976.2007.00069.x.
- Frech, C. and Chen, N. (2013) 'Variant surface antigens of malaria parasites: Functional and evolutionary insights from comparative gene family classification and analysis', *BMC Genomics*, 14(1). doi: 10.1186/1471-2164-14-427.
- Freeman, J. L. *et al.* (2006) 'Copy number variation: New insights in genome diversity', *Genome Research*, pp. 949–961. doi: 10.1101/gr.3677206.
- Gabaldón, T. and Koonin, E. V. (2013) 'Functional and evolutionary implications of gene orthology', *Nature Reviews Genetics*, pp. 360–366. doi: 10.1038/nrg3456.
- Galanti, N. *et al.* (1998) 'Histone genes in Trypanosomatids', *Parasitology*

*Today*, pp. 64–70. doi: 10.1016/S0169-4758(97)01162-9.

Ganley, A. R. D. and Kobayashi, T. (2007) 'Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data', *Genome Research*, 17(2), pp. 184–191. doi: 10.1101/gr.5457707.

Gardiner, P. R. *et al.* (1987) 'Identification and isolation of a variant surface glycoprotein from *Trypanosoma vivax*.' , *Science*, 235(4790), pp. 774–777.

Gardiner, P. R. (1989) 'Recent Studies of the Biology of *Trypanosoma vivax*' , *Advances in Parasitology*, 28(C), pp. 229–317. doi: 10.1016/S0065-308X(08)60334-6.

Gardiner, P. R. and Wilson, A. J. (1987) '*Trypanosoma (Duttonella) vivax*' , *Parasitology Today*, pp. 49–52. doi: 10.1016/0169-4758(87)90213-4.

Garside, L., Bailey, M. and Gibson, W. (1994) 'DNA content and molecular karyotype of trypanosomes of the subgenus *Nannomonas*' , *Acta Tropica*, 57(1), pp. 21–28. doi: 10.1016/0001-706X(94)90089-2.

Garza, M. *et al.* (2014) 'Projected Future Distributions of Vectors of *Trypanosoma cruzi* in North America under Climate Change Scenarios' , *PLoS Neglected Tropical Diseases*, 8(5). doi: 10.1371/journal.pntd.0002818.

Geerts, S. *et al.* (2001) 'African bovine trypanosomiasis: The problem of drug resistance' , *Parasitology Today*, 17(1), pp. 25–28. doi: 10.1016/S0169-4758(00)01827-5.

Giambiagi-deMarval, M., Souto-Padrón, T. and Rondinelli, E. (1996) 'Characterization and cellular distribution of heat-shock proteins HSP70 and HSP60 in *Trypanosoma cruzi*.' , *Experimental parasitology*, 83(3), pp. 335–45. doi: 10.1006/expr.1996.0081.

Gibellini, F., Hunter, W. N. and Smith, T. K. (2008) 'Biochemical characterization of the initial steps of the Kennedy pathway in *Trypanosoma brucei*: the ethanolamine and choline kinases' , *Biochemical Journal*, 415(1),

pp. 135–144. doi: 10.1042/BJ20080435.

Gibson, W. *et al.* (2008) 'The use of yellow fluorescent hybrids to indicate mating in *Trypanosoma brucei*', *Parasites and Vectors*, 1(1). doi: 10.1186/1756-3305-1-4.

Gibson, W. (2012) 'The origins of the trypanosome genome strains *Trypanosoma brucei brucei* TREU 927, *T. b. gambiense* DAL 972, *T. vivax* Y486 and *T. congolense* IL3000', *Parasites and Vectors*. doi: 10.1186/1756-3305-5-71.

Gibson, W. *et al.* (2015) 'Genetic Recombination between Human and Animal Parasites Creates Novel Strains of Human Pathogen', *PLoS Neglected Tropical Diseases*, 9(3). doi: 10.1371/journal.pntd.0003665.

Gibson, W. and Bailey, M. (1994) 'Genetic exchange in *Trypanosoma brucei*: evidence for meiosis from analysis of a cross between drug-resistant transformants', *Molecular and Biochemical Parasitology*, 64(2), pp. 241–252. doi: 10.1016/0166-6851(94)00017-4.

Gibson, W. C., Dukes, P. and Gashumba, J. K. (1988) 'Species-specific DNA probes for the identification of African trypanosomes in tsetse flies.', *Parasitology*, 97 ( Pt 1), pp. 63–73. doi: 10.1017/S0031182000066749.

Gibson, W. C. and Garside, L. H. (1991) 'Genetic exchange in *Trypanosoma brucei brucei*: variable chromosomal location of housekeeping genes in different trypanosome stocks', *Molecular and Biochemical Parasitology*, 45(1), pp. 77–90. doi: 10.1016/0166-6851(91)90029-6.

Gillingwater, K., Mamabolo, M. V and Majiwa, P. A. O. (2010) 'Prevalence of mixed *Trypanosoma congolense* infections in livestock and tsetse in KwaZulu-Natal, South Africa.', *Journal of the South African Veterinary Association*, 81, pp. 219–223. doi: 10.4102/jsava.v81i4.151.

Girgis, H. Z. (2015) 'Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale.', *BMC bioinformatics*. BMC Bioinformatics, 16(1), p. 227. doi: 10.1186/s12859-015-0654-5.

Giroud, C. *et al.* (2009) 'Murine models for *Trypanosoma brucei gambiense* disease progression - From silent to chronic infections and early brain tropism', *PLoS Neglected Tropical Diseases*, 3(9). doi: 10.1371/journal.pntd.0000509.

Godiska, R. *et al.* (2009) 'Linear plasmid vector for cloning of repetitive or unstable sequences in *Escherichia coli*', *Nucleic Acids Research*, 38(6). doi: 10.1093/nar/gkp1181.

Gonçalves dos Santos Silva, A. *et al.* (2010) 'Telomere-Centromere-Driven Genomic Instability Contributes to Karyotype Evolution in a Mouse Model of Melanoma', *Neoplasia*, 12(1), pp. 11-14. doi: 10.1593/neo.91004.

Gonzales-Perdomo, M., Romero, P. and Goldenberg, S. (1988) 'Cyclic AMP and adenylate cyclase activators stimulate *Trypanosoma cruzi* differentiation', *Experimental Parasitology*, 66(2), pp. 205–212. doi: 10.1016/0014-4894(88)90092-6.

González-De La Fuente, S. *et al.* (2017) 'Resequencing of the *Leishmania infantum* (strain JPCM5) genome and de novo assembly into 36 contigs', *Scientific Reports*, 7(1). doi: 10.1038/s41598-017-18374-y.

Gonzalez, A. *et al.* (1984) 'Minichromosomal repetitive DNA in *Trypanosoma cruzi*: its use in a high-sensitivity parasite detection assay.', *Proceedings of the National Academy of Sciences of the United States of America*, 81(June), pp. 3356–3360. doi: 10.1073/pnas.81.11.3356.

Gouy, M., Guindon, S. and Gascuel, O. (2010) 'SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building.', *Molecular biology and evolution*, 27(2), pp. 221–224. doi: 10.1093/molbev/msp259.

de Graaf, C. A. and van Steensel, B. (2013) 'Chromatin organization: form to function.', *Current opinion in genetics & development*. Elsevier Ltd, 23(2), pp. 185–90. doi: 10.1016/j.gde.2012.11.011.

Gray, M. A. *et al.* (1984) 'In vitro cultivation of *Trypanosoma congolense*: the production of infective metacyclic trypanosomes in cultures initiated from

cloned stocks', *Acta Trop.*, 41(0001–706X (Print)), pp. 343–353.

Gremme, G., Steinbiss, S. and Kurtz, S. (2013) 'Genome tools: A comprehensive software library for efficient processing of structured genome annotations', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(3), pp. 645–656. doi: 10.1109/TCBB.2013.68.

Guindon, S. *et al.* (2010) 'New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0', *Systematic Biology*, 59(3), pp. 307–321. doi: 10.1093/sysbio/syq010.

Guindon, S. and Gascuel, O. (2003) 'A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood', *Systematic Biology*, 52(5), pp. 696–704. doi: 10.1080/10635150390235520.

Guindon, S. and Gascuel, O. (2003) 'A Simple, Fast, and Accurate Method to Estimate Large Phylogenies by Maximum Likelihood', *Systematic Biology*, 52(5), pp. 696–704. doi: 10.1080/10635150390235520.

Haag, J., O'hUigin, C. and Overath, P. (1998) 'The molecular phylogeny of trypanosomes: evidence for an early divergence of the Salivaria', *Molecular and Biochemical Parasitology*. doi: 10.1016/S0166-6851(97)00185-0.

Haber, J. E. and Leung, W. Y. (1996) 'Lack of chromosome territoriality in yeast: promiscuous rejoining of broken chromosome ends.', *Proceedings of the National Academy of Sciences of the United States of America*, 93(24), pp. 13949–54. doi: 10.1073/pnas.93.24.13949.

Haenni, S. *et al.* (2006) 'The procyclin-associated genes of *Trypanosoma brucei* are not essential for cyclical transmission by tsetse', *Molecular and Biochemical Parasitology*, 150(2), pp. 144–156. doi: 10.1016/j.molbiopara.2006.07.005.

HaileMeskel, T. M. (2016) *Trypanosomiasis costs 37 African countries USD 4.5 Billion yearly, FAO, Food and Agriculture Organization of the United Nations.*

Hajduk, S. L., Siqueira, A. M. and Vickerman, K. (1986) 'Kinetoplast DNA of *Bodo caudatus*: a noncatenated structure.', *Molecular and Cellular Biology*, 6(12), pp. 4372–4378. doi: 10.1128/MCB.6.12.4372.

Hall, J. P. J., Wang, H. and David Barry, J. (2013) 'Mosaic VSGs and the Scale of *Trypanosoma brucei* Antigenic Variation', *PLoS Pathogens*, 9(7). doi: 10.1371/journal.ppat.1003502.

Hammond, J. W., Cai, D. and Verhey, K. J. (2008) 'Tubulin modifications and their cellular functions', *Current Opinion in Cell Biology*, pp. 71–76. doi: 10.1016/j.ceb.2007.11.010.

Han, X. *et al.* (2007) 'Type IV fimbrial biogenesis is required for protease secretion and natural transformation in *Dichelobacter nodosus*', *Journal of Bacteriology*, 189(14), pp. 5022–5033. doi: 10.1128/JB.00138-07.

Han, Y. *et al.* (2009) 'Centromere repositioning in cucurbit species: Implication of the genomic impact from centromere activation and inactivation', *Proceedings of the National Academy of Sciences*, 106(35), pp. 14937–14941. doi: 10.1073/pnas.0904833106.

Hancock, A. M. *et al.* (2010) 'Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency', *Proceedings of the National Academy of Sciences*, 107(Supplement\_2), pp. 8924–8930. doi: 10.1073/pnas.0914625107.

Happel, N. and Doenecke, D. (2009) 'Histone H1 and its isoforms: Contribution to chromatin structure and function', *Gene*, pp. 1–12. doi: 10.1016/j.gene.2008.11.003.

Harewood, L. and Fraser, P. (2014) 'The impact of chromosomal rearrangements on regulation of gene expression', *Human Molecular Genetics*, 23(R1). doi: 10.1093/hmg/ddu278.

Hayes, P. *et al.* (2014) 'Modulation of a cytoskeletal calpain-like protein induces major transitions in trypanosome morphology', *Journal of Cell Biology*, 206(3), pp. 377–384. doi: 10.1083/jcb.201312067.

Hecker, H. *et al.* (1994) 'The chromatin of trypanosomes', *International Journal for Parasitology*, 24(6), pp. 809–819. doi: 10.1016/0020-7519(94)90007-8.

Helm, J. R. *et al.* (2009) 'Analysis of expressed sequence tags from the four main developmental stages of *Trypanosoma congolense*.', *Molecular and biochemical parasitology*, 168(1), pp. 34–42. doi: 10.1016/j.molbiopara.2009.06.004.

Hertz-Fowler, C. *et al.* (2008) 'Telomeric expression sites are highly conserved in *Trypanosoma brucei*', *PLoS ONE*, 3(10). doi: 10.1371/journal.pone.0003527.

Higgins, M. K. *et al.* (2013) 'Structure of the trypanosome haptoglobin-hemoglobin receptor and implications for nutrient uptake and innate immunity', *Proceedings of the National Academy of Sciences*, 110(5), pp. 1905–1910. doi: 10.1073/pnas.1214943110.

Hirumi, H. and Hirumi, K. (1991) 'In vitro cultivation of *Trypanosoma congolense* bloodstream forms in the absence of feeder cell layers.', *Parasitology*, 102 Pt 2, pp. 225–36. doi: 10.1017/S0031182000062533.

Hoare, C. A. (1972) *The trypanosomes of mammals: a zoological monograph*, *The trypanosomes of mammals A zoological monograph*. doi: 10.3109/09638237.2012.705929.

Hoek, M., Engstler, M. and Cross, G. A. M. (2000) 'Expression-site-associated gene 8 (ESAG8) of *Trypanosoma brucei* is apparently essential and accumulates in the nucleolus', *Journal of Cell Science*, 113, pp. 3959–3968.

Horn, D. (2001) 'Nuclear gene transcription and chromatin in *Trypanosoma brucei*', *International Journal for Parasitology*, pp. 1157–1165. doi: 10.1016/S0020-7519(01)00264-8.

Hotz, H. R. *et al.* (1997) 'Mechanisms of developmental regulation in *Trypanosoma brucei*: A polypyrimidine tract in the 3'-untranslated region of a surface protein mRNA affects RNA abundance and translation', *Nucleic Acids Research*, 25(15), pp. 3017–3025. doi: 10.1093/nar/25.15.3017.

Hovel-Miner, G. A. *et al.* (2012) 'Telomere Length Affects the Frequency and Mechanism of Antigenic Variation in *Trypanosoma brucei*', *PLoS Pathogens*, 8(8). doi: 10.1371/journal.ppat.1002900.

Hu, Z.-L., Bao, J. and Reecy, J. (2008) 'CateGORizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories', *Online Journal of Bioinformatics*, pp. 108–112. Available at: <http://users.comcen.com.au/~journals/geneontologyabs2008.htm>.

Hughes, L. *et al.* (2017) 'Patterns of organelle ontogeny through a cell cycle revealed by whole-cell reconstructions using 3D electron microscopy', *Journal of Cell Science*, 130(3), pp. 637–647. doi: 10.1242/jcs.198887.

Hull, M. W. *et al.* (1994) 'tRNA genes as transcriptional repressor elements.', *Molecular and cellular biology*, 14(2), pp. 1266–77. doi: 10.1128/MCB.14.2.1266.Updated.

Hunter, S. *et al.* (2009) 'InterPro: The integrative protein signature database', *Nucleic Acids Research*, 37(SUPPL. 1). doi: 10.1093/nar/gkn785.

Imboden, M. A. *et al.* (1987) 'Transcription of the intergenic regions of the tubulin gene cluster of *Trypanosoma brucei*: Evidence for a polyclstronic transcription unit in a eukaryote', *Nucleic Acids Research*, 15(18), pp. 7357–7368. doi: 10.1093/nar/15.18.7357.

Initiative, I. G. G. (2014) 'Genome sequence of the tsetse fly (*Glossina morsitans*): vector of African trypanosomiasis.', *Science*, 344(6182), pp. 380–386. doi: 10.1126/science.1249656.

InoKuchi, H. O. and H. (1986) 'O R G A N I Z A T I O N OF TRANSFER RNA GENES IN PROKARYOTES Department of Biophysics , Faculty of Science , Kyoto Unh , ersiv ', *Advances in Biophysics*, 21, pp. 35–47.

Iskow, R. C., Gokcumen, O. and Lee, C. (2012) 'Exploring the role of copy number variants in human adaptation', *Trends in Genetics*, pp. 245–257. doi: 10.1016/j.tig.2012.03.002.

Ivens, A. C. *et al.* (2005) 'The genome of the kinetoplastid parasite, *Leishmania major*', *Science*, 309(5733), pp. 436–442. doi: 10.1126/science.1112680.

Jackson, A. P. (2007) 'Evolutionary consequences of a large duplication event in *Trypanosoma brucei*: Chromosomes 4 and 8 are partial duplicons', *BMC Genomics*, 8. doi: 10.1186/1471-2164-8-432.

Jackson, A. P. *et al.* (2010) 'The genome sequence of *Trypanosoma brucei gambiense*, causative agent of chronic human African Trypanosomiasis', *PLoS Neglected Tropical Diseases*, 4(4). doi: 10.1371/journal.pntd.0000658.

Jackson, A. P. *et al.* (2012) 'Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species', *Pnas*, 109(9), pp. 3416–3421. doi: 10.1073/pnas.1117313109.

Jackson, A. P. *et al.* (2013) 'A Cell-surface Phylome for African Trypanosomes', *PLoS Neglected Tropical Diseases*, 7(3), p. e2121. doi: 10.1371/journal.pntd.0002121.

Jackson, A. P. *et al.* (2014) 'The evolutionary dynamics of variant antigen genes in *Babesia* reveal a history of genomic innovation underlying host-parasite interaction', *Nucleic Acids Research*, 42(11), pp. 7113–7131. doi: 10.1093/nar/gku322.

Jackson, A. P. *et al.* (2015) 'Global gene expression profiling through the complete life cycle of *Trypanosoma vivax*', *PLoS Neglected Tropical Diseases*, 9(8). doi: 10.1371/journal.pntd.0003975.

Jackson, A. P. (2016) 'Gene family phylogeny and the evolution of parasite cell surfaces', *Molecular and Biochemical Parasitology*. Elsevier B.V., 209(1–2), pp. 64–75. doi: 10.1016/j.molbiopara.2016.03.007.

Jackson, A. P. *et al.* (2016) 'Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism', *Current Biology*, 26, pp. 161–172. doi: 10.1016/j.cub.2015.11.055.

JACKSON, A. P. (2015) 'Genome evolution in trypanosomatid parasites',

*Parasitology*, 142(S1), pp. S40–S56. doi: 10.1017/S0031182014000894.

Jackson, A. P. A. *et al.* (2007) 'Tandem gene arrays in *Trypanosoma brucei*: Comparative phylogenomic analysis of duplicate sequence variation', *BMC Evolutionary Biology*, 7(1), p. 54. doi: 10.1186/1471-2148-7-54.

Jackson, A. P. and Barry, J. D. (2012) 'The Evolution of Antigenic Variation in African Trypanosomes', in Sibley, L. D., Howlett, B. J., and Heitman, J. (eds) *Evolution of Virulence in Eukaryotic Microbes*. Wiley-Blackwell, pp. 324–337.

Jackson, D. G., Windle, H. J. and Voorheis, H. P. (1993) 'The identification, purification, and characterization of two invariant surface glycoproteins located beneath the surface coat barrier of bloodstream forms of *Trypanosoma brucei*', *Journal of Biological Chemistry*, 268(11), pp. 8085–8095.

Jaskowska, E. *et al.* (2015) 'Phytomonas: trypanosomatids adapted to plant environments', *PLoS pathogens*, p. e1004484. doi: 10.1371/journal.ppat.1004484.

Jefferies, D., Helfrich, M. P. and Molyneux, D. H. (1987) 'Cibarial infections of *Trypanosoma vivax* and *T. congolense* in *Glossina*', *Parasitology Research*, 73(4), pp. 289–292. doi: 10.1007/BF00531079.

Jenni, L. *et al.* (1986) 'Hybrid formation between African trypanosomes during cyclical transmission', *Nature*, 322(6075), pp. 173–175. doi: 10.1038/322173a0.

Jiang, L. *et al.* (2013) 'PfSETvs methylation of histone H3K36 represses virulence genes in *Plasmodium falciparum*', *Nature*, 499(7457), pp. 223–227. doi: 10.1038/nature12361.

Jimenez, V. (2014) 'Dealing with environmental challenges: Mechanisms of adaptation in *Trypanosoma cruzi*', *Research in Microbiology*, 165(3), pp. 155–165. doi: 10.1016/j.resmic.2014.01.006.

Johnson, L. S., Eddy, S. R. and Portugaly, E. (2010) 'Hidden Markov model speed heuristic and iterative HMM search procedure', *BMC Bioinformatics*, 11.

doi: 10.1186/1471-2105-11-431.

Jones, T. W. and Dávila, A. M. R. (2001) 'Trypanosoma vivax - Out of Africa', *Trends in Parasitology*, pp. 99–101. doi: 10.1016/S1471-4922(00)01777-3.

Joshi, P. B. *et al.* (2002) 'Targeted gene deletion in Leishmania major identifies leishmanolysin (GP63) as a virulence factor.', *Molecular and biochemical parasitology*, 120, pp. 33–40. doi: 10.1016/S0166-6851(01)00432-7.

Katoh, K. and Standley, D. M. (2013) 'MAFFT multiple sequence alignment software version 7: Improvements in performance and usability', *Molecular Biology and Evolution*, 30(4), pp. 772–780. doi: 10.1093/molbev/mst010.

Kearse, M. *et al.* (2012) 'Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data', *Bioinformatics*, 28(12), pp. 1647–1649. doi: 10.1093/bioinformatics/bts199.

Kennedy, P. G. E. (2004) 'Human African trypanosomiasis of the CNS: current issues and challenges.', *The Journal of clinical investigation*, 113(4), pp. 496–504. doi: 10.1172/JCI21052.

Kennedy, P. G. E. (2006) 'Diagnostic and neuropathogenesis issues in human African trypanosomiasis', *International Journal for Parasitology*, 36(5), pp. 505–512. doi: 10.1016/j.ijpara.2006.01.012.

Kennedy, P. G. E. (2008) 'The continuing problem of human African trypanosomiasis (sleeping sickness)', *Annals of Neurology*, 64(2), pp. 116–126. doi: 10.1002/ana.21429.

Khan, M. F. *et al.* (2015) 'Dataset for distribution of INGI/RIME and SLACS CRE transposable elements in Trypanosoma brucei genome', *Data in Brief*, 5, pp. 818–821. doi: 10.1016/j.dib.2015.10.040.

Kinabo, L. D. B. (1993) 'Pharmacology of existing drugs for animal trypanosomiasis', *Acta Tropica*, pp. 169–183. doi: 10.1016/0001-706X(93)90091-O.

Kirn, T. J., Bose, N. and Taylor, R. K. (2003) 'Secretion of a soluble

colonization factor by the TCP type 4 pilus biogenesis pathway in *Vibrio cholerae*', *Molecular Microbiology*, 49(1), pp. 81–92. doi: 10.1046/j.1365-2958.2003.03546.x.

Knippschild, U. *et al.* (2005) 'The casein kinase 1 family: Participation in multiple cellular processes in eukaryotes', *Cellular Signalling*, pp. 675–689. doi: 10.1016/j.cellsig.2004.12.011.

Kohl, L., Sherwin, T. and Gull, K. (1999) 'Assembly of the paraflagellar rod and the flagellum attachment zone complex during the *Trypanosoma brucei* cell cycle.', *The Journal of eukaryotic microbiology*, 46(2), pp. 105–109. doi: 10.1111/j.1550-7408.1999.tb04592.x.

Kolev, N. G. *et al.* (2010) 'The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution', *PLoS Pathogens*, 6(9). doi: 10.1371/journal.ppat.1001090.

Kolev, N. G., Günzl, A. and Tschudi, C. (2017) 'Metacyclic VSG expression site promoters are recognized by the same general transcription factor that is required for RNA polymerase I transcription of bloodstream expression sites', *Molecular and Biochemical Parasitology*, 216, pp. 52–55. doi: 10.1016/j.molbiopara.2017.07.002.

Koonin, E. V. (2005) 'Orthologs, Paralogs, and Evolutionary Genomics', *Annual Review of Genetics*, 39(1), pp. 309–338. doi: 10.1146/annurev.genet.39.073003.114725.

Koren, S. *et al.* (2016) 'Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation', *bioRxiv*. Available at: <http://biorxiv.org/content/early/2016/08/24/071282> (Accessed: 26 April 2017).

Kosakovsky Pond, S. L. *et al.* (2006) 'GARD: A genetic algorithm for recombination detection', *Bioinformatics*, 22(24), pp. 3096–3098. doi: 10.1093/bioinformatics/btl474.

Kosakovsky Pond, S. L. *et al.* (2011) 'A random effects branch-site model for detecting episodic diversifying selection', *Molecular Biology and Evolution*,

28(11), pp. 3033–3043. doi: 10.1093/molbev/msr125.

Kraemer, S. M. and Smith, J. D. (2003) 'Evidence for the importance of genetic structuring to the structural and functional specialization of the *Plasmodium falciparum* var gene family', *Molecular Microbiology*, 50(5), pp. 1527–1538. doi: 10.1046/j.1365-2958.2003.03814.x.

Krude, T. (1995) 'Chromatin: Nucleosome assembly during DNA replication', *Current Biology*, 5(11), pp. 1232–1234. doi: 10.1016/S0960-9822(95)00245-4.

Krutilina, R. I. *et al.* (2003) '[Recognition of internal (TTAGGG)<sub>n</sub> repeats by telomeric protein TRF1 and its role in maintenance of chromosomal stability in Chinese hamster cells]', *Tsitologiya*, 45(12), pp. 1211–1220.

Kuhn, R. M., Clarke, L. and Carbon, J. (1991) 'Clustered tRNA genes in *Schizosaccharomyces pombe* centromeric DNA sequence repeats', *Proceedings of the National Academy of Sciences of the United States of America*, 88(4), pp. 1306–1310. doi: 10.1073/pnas.88.4.1306.

Kukla, B. A. *et al.* (1987) 'Use of species-specific DNA probes for detection and identification of trypanosome infection in tsetse flies', *Parasitology*, 95(01), p. 1. doi: 10.1017/S0031182000057498.

Kurtz, S. *et al.* (2004) 'Versatile and open software for comparing large genomes.', *Genome biology*, 5(2), p. R12. doi: 10.1186/gb-2004-5-2-r12.

Kvikstad, E. M. and Makova, K. D. (2010) 'The (r)evolution of SINE versus LINE distributions in primate genomes: Sex chromosomes are important', *Genome Research*, 20(5), pp. 600–613. doi: 10.1101/gr.099044.109.

Kyes, S. A., Kraemer, S. M. and Smith, J. D. (2007) 'Antigenic variation in *Plasmodium falciparum*: Gene organization and regulation of the var multigene family', *Eukaryotic Cell*, 6(9), pp. 1511–1520. doi: 10.1128/EC.00173-07.

LaCount, D. J. *et al.* (2003) 'Expression and Function of the *Trypanosoma brucei* Major Surface Protease (GP63) Genes', *Journal of Biological*

- Chemistry*, 278(27), pp. 24658–24664. doi: 10.1074/jbc.M301451200.
- Lalmanach, G. *et al.* (2002) 'Congopain from *Trypanosoma congolense*: Drug target and vaccine candidate', *Biological Chemistry*, pp. 739–749. doi: 10.1515/BC.2002.077.
- Lamb, J. C. and Birchler, J. A. (2003) 'The role of DNA sequence in centromere formation', *Genome Biology*. doi: 10.1186/gb-2003-4-5-214.
- Lander, E. S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921. doi: 10.1038/35057062.
- Lane-Serff, H. *et al.* (2016) 'Evolutionary diversification of the trypanosome haptoglobin-haemoglobin receptor from an ancestral haemoglobin receptor', *eLife*, 5(APRIL2016). doi: 10.7554/eLife.13044.
- De Lange, T. (2005) 'Shelterin: The protein complex that shapes and safeguards human telomeres', *Genes and Development*, pp. 2100–2110. doi: 10.1101/gad.1346005.
- Lapp, S. A. *et al.* (2018) 'PacBio assembly of a *Plasmodium knowlesi* genome sequence with Hi-C correction and manual annotation of the SICAvax gene family', *Parasitology*, pp. 71–84. doi: 10.1017/S0031182017001329.
- Larkin, M. A. *et al.* (2007) 'Clustal W and Clustal X version 2.0', *Bioinformatics*, 23(21), pp. 2947–2948. doi: 10.1093/bioinformatics/btm404.
- Laslett, D. and Canback, B. (2004) 'ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences', *Nucleic Acids Research*, 32(1), pp. 11–16. doi: 10.1093/nar/gkh152.
- Lee, H., Gurtowski, J. and Yoo, S. (2014) 'Error correction and assembly complexity of single molecule sequencing reads', *bioRxiv*, pp. 1–17. doi: 10.1101/006395.
- Leeflang, P., Buys, J. and Blotkamp, C. (1976) 'Studies on *Trypanosoma vivax*: Infectivity and serial maintenance of natural bovine isolates in mice', *International Journal for Parasitology*, 6(5), pp. 413–417. doi: 10.1016/0020-

7519(76)90027-8.

Lefort, V., Longueville, J.-E. and Gascuel, O. (2017) 'SMS: Smart Model Selection in PhyML', *Molecular Biology and Evolution*, pp. 4–6. doi: 10.1093/molbev/msx149.

Lehane, M. J., Aksoy, S. and Levashina, E. (2004) 'Immune responses and parasite transmission in blood-feeding insects', *Trends in Parasitology*, pp. 433–439. doi: 10.1016/j.pt.2004.07.002.

Li, H. (2011) 'Tabix: Fast retrieval of sequence features from generic TAB-delimited files', *Bioinformatics*, 27(5), pp. 718–719. doi: 10.1093/bioinformatics/btq671.

Li, H. (2013) 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM', 00(00), pp. 1–3. doi: arXiv:1303.3997 [q-bio.GN].

Li, L. *et al.* (1998) 'In vitro and in vivo reconstitution and stability of vertebrate chromosome ends', *Nucleic Acids Research*, 26(12), pp. 2908–2916. doi: 10.1093/nar/26.12.2908.

Li, L., Stoeckert, C. J. J. and Roos, D. S. (2003) 'OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes -- Li et al. 13 (9): 2178 -- Genome Research', *Genome Research*, 13(9), pp. 2178–2189. doi: 10.1101/gr.1224503.candidates.

Li, Z. *et al.* (2012) 'Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph', *Briefings in Functional Genomics*, 11(1), pp. 25–37. doi: 10.1093/bfgp/elr035.

Lian, S. *et al.* (2014) 'A de novo genome assembly algorithm for repeats and nonrepeats', *BioMed Research International*, 2014. doi: 10.1155/2014/736473.

Liang, X.-H. *et al.* (2005) 'A genome-wide analysis of C/D and H/ACA-like small nucleolar RNAs in *Trypanosoma brucei* reveals a trypanosome-specific pattern of rRNA modification.', *Rna*, 11, pp. 619–645. doi:

10.1261/rna.7174805.

Lin, J. H. and Yamazaki, M. (2003) 'Role of P-glycoprotein in pharmacokinetics: Clinical implications', *Clinical Pharmacokinetics*, pp. 59–98. doi: 10.2165/00003088-200342010-00003.

Liniger, M. *et al.* (2001) 'Overlapping sense and antisense transcription units in *Trypanosoma brucei*', *Molecular Microbiology*, 40(4), pp. 869–878. doi: 10.1046/j.1365-2958.2001.02426.x.

Liniger, M. *et al.* (2003) 'Cleavage of trypanosome surface glycoproteins by alkaline trypsin-like enzyme(s) in the midgut of *Glossina morsitans*', *International Journal for Parasitology*, 33(12), pp. 1319–1328. doi: 10.1016/S0020-7519(03)00182-6.

Loomis, E. W. *et al.* (2013) 'Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene', *Genome Research*, 23(1), pp. 121–128. doi: 10.1101/gr.141705.112.

Lopez, M. A., Saada, E. A. and Hill, K. L. (2015) 'Insect stage-specific adenylate cyclases regulate social motility in African trypanosomes', *Eukaryotic Cell*, 14(1), pp. 104–112. doi: 10.1128/EC.00217-14.

Loveless, B. C. *et al.* (2011) 'Structural characterization and epitope mapping of the glutamic acid/alanine-rich protein from *Trypanosoma congolense*: Defining assembly on the parasite cell surface', *Journal of Biological Chemistry*, 286(23), pp. 20658–20665. doi: 10.1074/jbc.M111.218941.

Lowell, J. E. *et al.* (2005) 'Histone H2AZ dimerizes with a novel variant H2B and is enriched at repetitive DNA in *Trypanosoma brucei*.' , *Journal of cell science*, 118(Pt 24), pp. 5721–5730. doi: 10.1242/jcs.02688.

Lowell, J. E. and Cross, G. A. M. (2004) 'A variant histone H3 is enriched at telomeres in *Trypanosoma brucei*.' , *Journal of Cell Science*, 117(Pt 24), pp. 5937–5947. doi: 10.1242/jcs.01515.

Lowry, D. B. and Willis, J. H. (2010) 'A widespread chromosomal inversion

polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation', *PLoS Biology*, 8(9). doi: 10.1371/journal.pbio.1000500.

Luger, K. *et al.* (1997) 'Crystal structure of the nucleosome core particle at 2.8 Å resolution.', *Nature*, 389(6648), pp. 251–60. doi: 10.1038/38444.

MacGregor, P. and Matthews, K. R. (2010) 'New discoveries in the transmission biology of sleeping sickness parasites: Applying the basics', *Journal of Molecular Medicine*, pp. 865–871. doi: 10.1007/s00109-010-0637-y.

Macías, F., López, M. C. and Thomas, M. C. (2016) 'The Trypanosomatid Pr77-hallmark contains a downstream core promoter element essential for transcription activity of the Trypanosoma cruzi L1Tc retrotransposon', *BMC Genomics*, 17(1). doi: 10.1186/s12864-016-2427-6.

Magona, J. W., Walubengo, J. and Odimin, J. T. (2008) 'Acute haemorrhagic syndrome of bovine trypanosomosis in Uganda', *Acta Tropica*, 107(2), pp. 186–191. doi: 10.1016/j.actatropica.2008.05.019.

Mair, G. *et al.* (2000) 'A new twist in trypanosome RNA metabolism: Cis-splicing of pre-mRNA', *RNA*, 6(2), pp. 163–169. doi: 10.1017/S135583820099229X.

Majekodunmi, A. O. *et al.* (2013) 'A longitudinal survey of African animal trypanosomiasis in domestic cattle on the Jos Plateau, Nigeria: Prevalence, distribution and risk factors', *Parasites and Vectors*, 6(1). doi: 10.1186/1756-3305-6-239.

Majiwa, P. A. O. O. and Webster, P. (1987) 'A repetitive deoxyribonucleic acid sequence distinguishes Trypanosoma simiae from T. congolense', *Parasitology*, 95(3), pp. 543–558. doi: 10.1017/S0031182000057978.

Marcello, L. and Barry, J. D. (2007a) 'Analysis of the VSG gene silent archive in Trypanosoma brucei reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure', *Genome Research*,

17(9), pp. 1344–1352. doi: 10.1101/gr.6421207.

Marcello, L. and Barry, J. D. (2007b) 'From silent genes to noisy populations - Dialogue between the genotype and phenotypes of antigenic variation', in *Journal of Eukaryotic Microbiology*, pp. 14–17. doi: 10.1111/j.1550-7408.2006.00227.x.

Marchat, L. A. *et al.* (2015) 'DEAD/DEXH-Box RNA helicases in selected human parasites', *Korean Journal of Parasitology*, pp. 583–595. doi: 10.3347/kjp.2015.53.5.583.

Maree, J. P. and Patterson, H. G. (2014) 'The epigenome of *Trypanosoma brucei*: A regulatory interface to an unconventional transcriptional machine', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, pp. 743–750. doi: 10.1016/j.bbagr.2014.05.028.

Martínez-Calvillo, S. *et al.* (2003) 'Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region', *Molecular Cell*, 11(5), pp. 1291–1299. doi: 10.1016/S1097-2765(03)00143-6.

Martínez-Calvillo, S. *et al.* (2004) 'Transcription initiation and termination on *Leishmania major* chromosome 3', *Eukaryotic Cell*, 3(2), pp. 506–517. doi: 10.1128/EC.3.2.506-517.2004.

Maslov, D. A. *et al.* (2013) 'Diversity and phylogeny of insect trypanosomatids: All that is hidden shall be revealed', *Trends in Parasitology*, pp. 43–52. doi: 10.1016/j.pt.2012.11.001.

Masumu, J. *et al.* (2006) 'Comparison of the virulence of *Trypanosoma congolense* strains isolated from cattle in a trypanosomiasis endemic area of eastern Zambia', *International Journal for Parasitology*, 36(4), pp. 497–501. doi: 10.1016/j.ijpara.2006.01.003.

Masumu, J. *et al.* (2009) 'Cross-protection between *Trypanosoma congolense* strains of low and high virulence', *Veterinary Parasitology*, 163(1–2), pp. 127–131. doi: 10.1016/j.vetpar.2009.04.006.

- Matthews, K. R. (2005) 'The developmental cell biology of *Trypanosoma brucei*', *Journal of Cell Science*, 118(2), pp. 283–290. doi: 10.1242/jcs.01649.
- Mattick, J. S. (2002) 'Type IV Pili and Twitching Motility', *Annual Review of Microbiology*, 56(1), pp. 289–314. doi: 10.1146/annurev.micro.56.012302.160938.
- Mattioli, R. C. *et al.* (2004) 'Tsetse and trypanosomiasis intervention policies supporting sustainable animal-agricultural development', *Agriculture & Environment*, 22(22), pp. 310–314.
- McClintock, B. (1984) 'The significance of responses of the genome to challenge', *Science*, 226(4676), pp. 792–801. doi: 10.1126/science.15739260.
- McHale, L. *et al.* (2006) 'Plant NBS-LRR proteins: adaptable guards.', *Genome biology*, 7, p. 212. doi: 10.1186/gb-2006-7-4-212.
- Melville, S. E. *et al.* (1998) 'The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* and the assignment of chromosome markers', *Molecular and Biochemical Parasitology*, 94(2), pp. 155–173. doi: 10.1016/S0166-6851(00)00316-9.
- Melville, S. E. *et al.* (2000) 'The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* stock 427', *Molecular and Biochemical Parasitology*, 111(2), pp. 261–273. doi: 10.1016/S0166-6851(00)00316-9.
- Melville, S. E., Gerrard, C. S. and Blackwell, J. M. (1999) 'Multiple causes of size variation in the diploid megabase chromosomes of African trypanosomes', *Chromosome Research*, 7(3), pp. 191–203. doi: 10.1023/A:1009247315947.
- Mendoza-Palomares, C. *et al.* (2008) 'Molecular and biochemical characterization of a cathepsin B-like protease family unique to *Trypanosoma congolense*', *Eukaryotic Cell*, 7(4), pp. 684–697. doi: 10.1128/EC.00405-07.
- Meyne, J., Ratliff, R. L. and Moyzis, R. K. (1989) 'Conservation of the human telomere sequence (TTAGGG)<sub>n</sub> among vertebrates.', *Proceedings of the National Academy of Sciences of the United States of America*, 86(18), pp.

7049–53. doi: 10.1073/pnas.86.18.7049.

Mihok, S. *et al.* (1995) 'Mechanical transmission of *Trypanosoma* spp. by African Stomoxiinae (Diptera: Muscidae)', *Tropical medicine and parasitology*, 46, pp. 103–105.

Milligan, P. J. and Baker, R. D. (1988) 'A model of tsetse-transmitted animal trypanosomiasis.', *Parasitology*, 96 ( Pt 1)(January), pp. 211–239. doi: 10.1017/S0031182000081774.

Misteli, T. (2001) 'Protein dynamics: implications for nuclear architecture and gene expression.', *Science (New York, N.Y.)*, 291(5505), pp. 843–847. doi: 10.1126/science.291.5505.843.

Misteli, T. (2005) 'Concepts in nuclear architecture', *BioEssays*, pp. 477–487. doi: 10.1002/bies.20226.

Mogk, S. *et al.* (2014) 'The lane to the brain: How African trypanosomes invade the CNS', *Trends in Parasitology*, pp. 470–477. doi: 10.1016/j.pt.2014.08.002.

Mok, B. W. *et al.* (2008) 'A highly conserved segmental duplication in the subtelomeres of *Plasmodium falciparum* chromosomes varies in copy number', *Malaria Journal*, 7. doi: 10.1186/1475-2875-7-46.

Moloo, S. K., Kutuza, S. B. and Desai, J. (1987) 'Comparative study on the infection rates of different *Glossina* species for East and West African *Trypanosoma vivax* stocks', *Parasitology*, 95(3), pp. 537–542. doi: 10.1017/S0031182000057966.

Morrison, L. J., Marcello, L. and McCulloch, R. (2009) 'Antigenic variation in the African trypanosome: Molecular mechanisms and phenotypic complexity', *Cellular Microbiology*, 11(12), pp. 1724–1734. doi: 10.1111/j.1462-5822.2009.01383.x.

Moser, D. R. *et al.* (1989) 'Detection of *Trypanosoma congolense* and *Trypanosoma brucei* subspecies by DNA amplification using the polymerase chain reaction', *Parasitology*. University of Liverpool Library, 99(01), p. 57. doi:

10.1017/S0031182000061023.

Mossaad, E. *et al.* (2017) 'Trypanosoma vivax is the second leading cause of camel trypanosomosis in Sudan after Trypanosoma evansi', *Parasites and Vectors*, 10(1). doi: 10.1186/s13071-017-2117-5.

Mottram, J. C., Brooks, D. R. and Coombs, G. H. (1998) 'Roles of cysteine proteinases of trypanosomes and Leishmania in host-parasite interactions', *Current Opinion in Microbiology*, 1(4), pp. 455–460. doi: 10.1016/S1369-5274(98)80065-9.

Mottram, J. C., Coombs, G. H. and Alexander, J. (2004) 'Cysteine peptidases as virulence factors of Leishmania', *Current Opinion in Microbiology*, pp. 375–381. doi: 10.1016/j.mib.2004.06.010.

Murphy, W. J. *et al.* (2005) 'Evolution: Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps', *Science*, 309(5734), pp. 613–617. doi: 10.1126/science.1111387.

Murrell, B. *et al.* (2013) 'FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection', *Molecular Biology and Evolution*, 30(5), pp. 1196–1205. doi: 10.1093/molbev/mst030.

Nagarajan, N. and Pop, M. (2009) 'Parametric Complexity of Sequence Assembly: Theory and Applications to Next Generation Sequencing', *Journal of Computational Biology*, 16(7), pp. 897–908. doi: 10.1089/cmb.2009.0005.

Nattestad, M., Chin, C.-S. and Schatz, M. C. (2016) 'Ribbon: Visualizing complex genome alignments and structural variation', *bioRxiv*, 0344, p. 82123. doi: 10.1101/082123.

Navarro, M. (1999) 'Trypanosoma brucei variant surface glycoprotein regulation involves coupled activation/inactivation and chromatin remodeling of expression sites', *The EMBO Journal*, 18(8), pp. 2265–2272. doi: 10.1093/emboj/18.8.2265.

Navarro, M. and Cross, G. a (1996) 'DNA rearrangements associated with

multiple consecutive directed antigenic switches in *Trypanosoma brucei*.', *Molecular and cellular biology*, 16(7), pp. 3615–3625. doi: 10.1128/mcb.16.7.3615.

Navarro, M. and Gull, K. (2001) 'A pol I transcriptional body associated with VSG mono-allelic expression in *Trypanosoma brucei*.', *Nature*, 414(6865), pp. 759–763. doi: 10.1038/414759a.

Nawrocki, E. P. and Eddy, S. R. (2013) 'Infernal 1.1: 100-fold faster RNA homology searches', *Bioinformatics*, 29(22), pp. 2933–2935. doi: 10.1093/bioinformatics/btt509.

Neafsey, D. E. *et al.* (2012) 'The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*', *Nature Genetics*, 44(9), pp. 1046–1050. doi: 10.1038/ng.2373.

Newman, T. L. *et al.* (2005) 'A genome-wide survey of structural variation between human and chimpanzee', *Genome Research*, 15(10), pp. 1344–1356. doi: 10.1101/gr.4338005.

Nilsson, D. *et al.* (2010) 'Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of *Trypanosoma brucei*', *PLoS Pathogens*, 6(8), pp. 21–22. doi: 10.1371/journal.ppat.1001037.

Noma, K. *et al.* (2006) 'A Role for TFIIIC Transcription Factor Complex in Genome Organization', *Cell*, 125(5), pp. 859–872. doi: 10.1016/j.cell.2006.04.028.

Noor, M. A. F. *et al.* (2001) 'Chromosomal inversions and the reproductive isolation of species', *Proceedings of the National Academy of Sciences*, 98(21), pp. 12084–12088. doi: 10.1073/pnas.221274498.

Norris, S. J. (2014) 'vls Antigenic Variation Systems of Lyme Disease *Borrelia*: Eluding Host Immunity through both Random, Segmental Gene Conversion and Framework Heterogeneity', *Microbiology Spectrum*, 2(6), pp. 1–18. doi: 10.1128/microbiolspec.MDNA3-0038-2014.

O'Sullivan, R. J. and Karlseder, J. (2010) 'Telomeres: Protecting chromosomes against genome instability', *Nature Reviews Molecular Cell Biology*, pp. 171–181. doi: 10.1038/nrm2848.

Obado, S. O. *et al.* (2005) 'Functional mapping of a trypanosome centromere by chromosome fragmentation identifies a 16-kb GC-rich transcriptional "strand-switch" domain as a major feature', *Genome Research*, 15(1), pp. 36–43. doi: 10.1101/gr.2895105.

Obado, S. O. *et al.* (2007) 'Repetitive DNA is associated with centromeric domains in *Trypanosoma brucei* but not *Trypanosoma cruzi*', *Genome biology*, 8(3), p. R37. doi: 10.1186/gb-2007-8-3-r37.

Office, R. *et al.* (2015) 'Human African trypanosomiasis Cases of sleeping sickness drop to lowest level in 75 years', pp. 1–2.

Ogbadoyi, E. *et al.* (2000) 'Architecture of the *Trypanosoma brucei* nucleus during interphase and mitosis.', *Chromosoma*, 108(8), pp. 501–513. doi: 10.1007/s004120050402.

Ooi, C.-P. *et al.* (2016) 'The Cyclical Development of *Trypanosoma vivax* in the Tsetse Fly Involves an Asymmetric Division', *Frontiers in Cellular and Infection Microbiology*, 6. doi: 10.3389/fcimb.2016.00115.

Ooi, C.-P. and Bastin, P. (2013) 'More than meets the eye: understanding *Trypanosoma brucei* morphology in the tsetse', *Frontiers in Cellular and Infection Microbiology*, 3. doi: 10.3389/fcimb.2013.00071.

Osório, A. L. A. R. *et al.* (2008) '*Trypanosoma* (*Duttonella*) *vivax*: its biology, epidemiology, pathogenesis, and introduction in the New World - a review', *Memórias do Instituto Oswaldo Cruz*, 103(1), pp. 1–13. doi: 10.1590/S0074-02762008000100001.

Otto, T. D. *et al.* (2011) 'RATT: Rapid Annotation Transfer Tool.', *Nucleic acids research*, 39(9), pp. 1–7. doi: 10.1093/nar/gkq1268.

Parsons, M., Valentine, M. and Carter, V. (1993) 'Protein kinases in divergent

eukaryotes: identification of protein kinase activities regulated during trypanosome development', *Proceedings of the National Academy of Sciences of the United States of America*, 90(7), pp. 2656–2660. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC46154/>.

Pasini, E. M. *et al.* (2017) 'An improved genome assembly reveals *Plasmodium cynomolgi* an unexpected methyltransferase gene expansion [version 1; referees: 2 approved]', *Wellcome Open Research*, 11(2), pp. 424242–30. doi: 10.12688/wellcomeopenres.11864.1.

Pays, E. *et al.* (1985) 'Telomeric reciprocal recombination as a possible mechanism for antigenic variation in trypanosomes', *Nature*, 316(6028), pp. 562–564. doi: 10.1038/316562a0.

Pays, E. *et al.* (1989) 'The genes and transcripts of an antigen gene expression site from *T. brucei*', *Cell*, 57(5), pp. 835–845. doi: 10.1016/0092-8674(89)90798-8.

Pays, E. *et al.* (2001) 'The VSG expression sites of *Trypanosoma brucei*: Multipurpose tools for the adaptation of the parasite to mammalian hosts', *Molecular and Biochemical Parasitology*, 114(1), pp. 1–16. doi: 10.1016/S0166-6851(01)00242-0.

Pays, E., Vanhamme, L. and Pérez-Morga, D. (2004) 'Antigenic variation in *Trypanosoma brucei*: Facts, challenges and mysteries', *Current Opinion in Microbiology*, pp. 369–374. doi: 10.1016/j.mib.2004.05.001.

Peacock, L. *et al.* (2012) 'The life cycle of *Trypanosoma* (*Nannomonas*) *congolense* in the tsetse fly', *Parasites & Vectors*, 5(1), p. 109. doi: 10.1186/1756-3305-5-109.

Peacock, L. *et al.* (2014) 'Meiosis and haploid gametes in the pathogen *Trypanosoma brucei*', *Current Biology*, 24(2), pp. 181–186. doi: 10.1016/j.cub.2013.11.044.

Peitl, P. *et al.* (2002) 'Chromosomal rearrangements involving telomeric DNA sequences in Balb/3T3 cells transfected with the Ha-ras oncogene.',

*Mutagenesis*, 17(1), pp. 67–72. doi: 10.1093/mutage/17.1.67.

Peña, I. *et al.* (2015) 'New Compound Sets Identified from High Throughput Phenotypic Screening Against Three Kinetoplastid Parasites: An Open Resource', *Scientific Reports*, 5(1), p. 8771. doi: 10.1038/srep08771.

Pérez-Morga, D. . *et al.* (2001) 'Organization of telomeres during the cell and life cycles of trypanosoma brucei', *Journal of Eukaryotic Microbiology*, 48(2), pp. 221–226. Available at: <http://www.scopus.com/inward/record.url?eid=2-s2.0-0034913709&partnerID=40&md5=226782f2e493ab2865109982cb114f9e>.

Phillippy, A. M., Schatz, M. C. and Pop, M. (2008) 'Genome assembly forensics: Finding the elusive mis-assembly', *Genome Biology*, 9(3). doi: 10.1186/gb-2008-9-3-r55.

Pinchbeck, G. L. *et al.* (2008) 'Trypanosomosis in The Gambia: prevalence in working horses and donkeys detected by whole genome amplification and PCR, and evidence for interactions between trypanosome species.', *BMC veterinary research*, 4, p. 7. doi: 10.1186/1746-6148-4-7.

Van der Ploeg, L. H. *et al.* (1984) 'Chromosomes of kinetoplastida.', *The EMBO journal*, 3(13), pp. 3109–3115. doi: 10.1002/j.1460-2075.1984.tb02266.x.

Pond, S. L. K. 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. doi: 10.1093/molbev/msi105.

Price, M. N., Dehal, P. S. and Arkin, A. P. (2010) 'FastTree 2 - Approximately maximum-likelihood trees for large alignments', *PLoS ONE*, 5(3). doi: 10.1371/journal.pone.0009490.

Prokopowich, C. D., Gregory, T. R. and Crease, T. J. (2003) 'The correlation between rDNA copy number and genome size in eukaryotes', *Genome*, 46(1), pp. 48–50. doi: 10.1139/g02-103.

Puechberty, J. *et al.* (2007) 'Compared genomics of the strand switch region of Leishmania chromosome 1 reveal a novel genus-specific gene and conserved structural features and sequence motifs', *BMC Genomics*, 8. doi: 10.1186/1471-2164-8-57.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: A flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.

Qvit, N. *et al.* (2016) 'Scaffold proteins LACK and TRACK as potential drug targets in kinetoplastid parasites: Development of inhibitors', *International Journal for Parasitology: Drugs and Drug Resistance*, 6(1), pp. 74–84. doi: 10.1016/j.ijpddr.2016.02.003.

R Core Team (2016) 'R: A Language and Environment for Statistical Computing', *R Foundation for Statistical Computing*, p. 3503. doi: 10.1007/978-3-540-74686-7.

Ramsey, J. M. *et al.* (2015) 'Atlas of Mexican Triatominae (Reduviidae: Hemiptera) and vector transmission of Chagas disease', *Memorias do Instituto Oswaldo Cruz*, 110(3), pp. 339–352. doi: 10.1590/0074-02760140404.

Rassi, A., Rassi, A. and Marin-Neto, J. A. (2010) 'Chagas disease', *The Lancet*, pp. 1388–1402. doi: 10.1016/S0140-6736(10)60061-X.

Raynaud, C. M. *et al.* (2008) 'Telomere length, telomeric proteins and genomic instability during the multistep carcinogenic process', *Critical Reviews in Oncology/Hematology*, pp. 99–117. doi: 10.1016/j.critrevonc.2007.11.006.

Redpath, M. B. *et al.* (2000) 'ESAG11, a new VSG expression site-associated gene from *Trypanosoma brucei*', *Molecular and Biochemical Parasitology*, 111(1), pp. 223–228. doi: 10.1016/S0166-6851(00)00305-4.

Respuela, P. *et al.* (2008) 'Histone acetylation and methylation at sites initiating divergent polycistronic transcription in *Trypanosoma cruzi*', *Journal of Biological Chemistry*, 283(23), pp. 15884–15892. doi: 10.1074/jbc.M802081200.

Rhoads, A. and Au, K. F. (2015) 'PacBio Sequencing and Its Applications', *Genomics, Proteomics and Bioinformatics*, pp. 278–289. doi: 10.1016/j.gpb.2015.08.002.

Richard, G.-F., Kerrest, A. and Dujon, B. (2008) 'Comparative genomics and molecular dynamics of DNA repeats in eukaryotes.', *Microbiology and molecular biology reviews: MMBR*, 72(4), pp. 686–727. doi: 10.1128/MMBR.00011-08.

Robinson, D. R. *et al.* (1995) 'Microtubule polarity and dynamics in the control of organelle positioning, segregation, and cytokinesis in the trypanosome cell cycle', *Journal of Cell Biology*, 128(6), pp. 1163–1172. doi: 10.1083/jcb.128.6.1163.

Robinson, N. P. *et al.* (1999) 'Predominance of duplicative VSG gene conversion in antigenic variation in African trypanosomes.', *Molecular and cellular biology*, 19(9), pp. 5839–46.

Rodgers, J. (2010) 'Trypanosomiasis and the brain', *Parasitology*, 137(14), pp. 1995–2006. doi: 10.1017/S0031182009991806.

Roditi, I. and Clayton, C. (1999) 'An unambiguous nomenclature for the major surface glycoproteins of the procyclic form of *Trypanosoma brucei*', *Molecular and Biochemical Parasitology*, 103(1), pp. 99–100. doi: 10.1016/S0166-6851(99)00124-3.

Rodrigues, A. C. *et al.* (2008) 'Phylogenetic analysis of *Trypanosoma vivax* supports the separation of South American/West African from East African isolates and a new *T. vivax*-like genotype infecting a nyala antelope from Mozambique', *Parasitology*, 135(11), pp. 1317–1328. doi: 10.1017/S0031182008004848.

Rodrigues, A. C. *et al.* (2014) 'Congopain genes diverged to become specific to Savannah, Forest and Kilifi subgroups of *Trypanosoma congolense*, and are valuable for diagnosis, genotyping and phylogenetic inferences', *Infection, Genetics and Evolution*, 23, pp. 20–31. doi: 10.1016/j.meegid.2014.01.012.

Rooney, A. P. and Ward, T. J. (2005) 'Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm', *Proceedings of the National Academy of Sciences of the United States of America*, 102(14), pp. 5084–5089. doi: 10.1073/pnas.0409689102.

Rotureau, B. *et al.* (2012) 'A new asymmetric division contributes to the continuous production of infective trypanosomes in the tsetse fly', *Development*, 139(10), pp. 1842–1850. doi: 10.1242/dev.072611.

Rotureau, B. and Van Den Abbeele, J. (2013) 'Through the dark continent: African trypanosome development in the tsetse fly', *Frontiers in Cellular and Infection Microbiology*, 3. doi: 10.3389/fcimb.2013.00053.

Rotureau, B., Subota, I. and Bastin, P. (2011) 'Molecular bases of cytoskeleton plasticity during the *Trypanosoma brucei* parasite cycle', *Cellular Microbiology*, 13(5), pp. 705–716. doi: 10.1111/j.1462-5822.2010.01566.x.

Rovira, C., Beermann, W. and Edström, J. erik (1993) 'A repetitive DNA sequence associated with the centromeres of *chironomus pallidivittatus*', *Nucleic Acids Research*, 21(8), pp. 1775–1781. doi: 10.1093/nar/21.8.1775.

RStudio, T. (2016) 'RStudio: Integrated Development for R', [Online] RStudio, Inc., Boston, MA URL <http://www.rstudio.com>, p. RStudio, Inc., Boston, MA. doi: 10.1007/978-81-322-2340-5.

Rutherford, K. *et al.* (2000) 'Artemis : sequence visualization and annotation', 16(10), pp. 944–945.

Rutherford, K. *et al.* (2000) 'Artemis: sequence visualization and annotation', *Bioinformatics*, 16(10), pp. 944–945. doi: 10.1093/bioinformatics/16.10.944.

Rutledge, G. G. *et al.* (2017) '*Plasmodium malariae* and *P. ovale* genomes provide insights into malaria parasite evolution', *Nature*, 542(7639), pp. 101–104. doi: 10.1038/nature21038.

Ruwende, C. and Hill, A. (1998) 'Glucose-6-phosphate dehydrogenase deficiency and malaria', *J Mol Med*, 76, pp. 581–588. doi:

10.1007/s001090050253.

Sakurai, T., Sugimoto, C. and Inoue, N. (2008) 'Identification and molecular characterization of a novel stage-specific surface protein of *Trypanosoma congolense* epimastigotes', *Molecular and Biochemical Parasitology*, 161(1), pp. 1–11. doi: 10.1016/j.molbiopara.2008.05.003.

Salmela, L. *et al.* (2017) 'Accurate self-correction of errors in long reads using de Bruijn graphs', *Bioinformatics*, 33(6), pp. 799–806. doi: 10.1093/bioinformatics/btw321.

Salmon, D. *et al.* (1994) 'A novel heterodimeric transferrin receptor encoded by a pair of VSG expression site-associated genes in *T. brucei*', *Cell*, 78(1), pp. 75–86. doi: 10.1016/0092-8674(94)90574-6.

Santrich, C. *et al.* (1997) 'A motility function for the paraflagellar rod of *Leishmania* parasites revealed by PFR-2 gene knockouts', *Molecular and Biochemical Parasitology*, 90(1), pp. 95–109. doi: 10.1016/S0166-6851(97)00149-7.

Sbicego, S. *et al.* (1999) 'The use of transgenic *Trypanosoma brucei* to identify compounds inducing the differentiation of bloodstream forms to procyclic forms', *Molecular and Biochemical Parasitology*, 104(2), pp. 311–322. doi: 10.1016/S0166-6851(99)00157-7.

Schell, D. *et al.* (1991) 'A transferrin-binding protein of *Trypanosoma brucei* is encoded by one of the genes in the variant surface glycoprotein gene expression site', *EMBO Journal*, 10(5), pp. 1061–1066.

Schlagenhauf, E., Etges, R. and Metcalf, P. (1998) 'The crystal structure of the *Leishmania major* surface proteinase leishmanolysin (gp63)', *Structure*, 6(8), pp. 1035–1046. doi: 10.1016/S0969-2126(98)00104-X.

Schmid-Hempel, P. *et al.* (2018) 'The genomes of *Crithidia bombi* and *C. expoeki*, common parasites of bumblebees', *PLoS ONE*, 13(1). doi: 10.1371/journal.pone.0189738.

Schmuñis, G. (2013) 'Status of and cost of Chagas disease worldwide', *The Lancet Infectious Diseases*, pp. 283–284. doi: 10.1016/S1473-3099(13)70032-X.

Schubert, I. and Rieger, R. (1985) 'A new mechanism for altering chromosome number during karyotype evolution', *Theoretical and Applied Genetics*, 70(2), pp. 213–221. doi: 10.1007/BF00275324.

Schueler, M. G. *et al.* (2001) 'Genomic and genetic definition of a functional human centromere', *Science*, 294(5540), pp. 109–115. doi: 10.1126/science.1065042.

Schwede, A., Kramer, S. and Carrington, M. (2012) 'How do trypanosomes change gene expression in response to the environment?', *Protoplasma*, 249(2), pp. 223–238. doi: 10.1007/s00709-011-0282-5.

Seong, H. A. and Ha, H. (2012) 'Murine protein serine-threonine kinase 38 activates p53 function through Ser 15 phosphorylation', *Journal of Biological Chemistry*, 287(25), pp. 20797–20810. doi: 10.1074/jbc.M112.347757.

Sexton, T. *et al.* (2012) 'Three-dimensional folding and functional organization principles of the *Drosophila* genome', *Cell*, 148(3), pp. 458–472. doi: 10.1016/j.cell.2012.01.010.

Shahada, F. *et al.* (2007) 'Absence of correlation between karyotype profiles of *Trypanosoma congolense* and resistance to isometamidium chloride.', *Veterinary parasitology*, 147(3–4), pp. 311–4. doi: 10.1016/j.vetpar.2007.04.016.

Sharma, R. *et al.* (2008) 'Asymmetric Cell Division as a Route to Reduction in Cell Length and Change in Cell Morphology in Trypanosomes', *Protist*, 159(1), pp. 137–151. doi: 10.1016/j.protis.2007.07.004.

Sharma, R. *et al.* (2009) 'The heart of darkness: growth and form of *Trypanosoma brucei* in the tsetse fly', *Trends in Parasitology*, pp. 517–524. doi: 10.1016/j.pt.2009.08.001.

- Sharp, A. J. *et al.* (2005) 'Segmental duplications and copy-number variation in the human genome.', *American journal of human genetics*, 77(1), pp. 78–88. doi: 10.1086/431652.
- Shaw, J. J. and Lainson, R. (1972) 'Trypanosoma vivax in Brazil', *Annals of Tropical Medicine and Parasitology*, 66(1), pp. 25–32. doi: 10.1080/00034983.1972.11686794.
- She, X. *et al.* (2008) 'Mouse segmental duplication and copy number variation', *Nature Genetics*, 40(7), pp. 909–914. doi: 10.1038/ng.172.
- Shea, C., Lee, M. G. S. and Van der Ploeg, L. H. T. (1987) 'VSG gene 118 is transcribed from a cotransposed pol I-like promoter', *Cell*, 50(4), pp. 603–612. doi: 10.1016/0092-8674(87)90033-X.
- Shea, C. and Van der Ploeg, L. H. (1988) 'Stable variant-specific transcripts of the variant cell surface glycoprotein gene 1.8 expression site in Trypanosoma brucei', *Mol Cell Biol*, 8(2), pp. 854–859.
- Shimodaira, H. and Hasegawa, M. (1999) 'Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference', *Molecular Biology and Evolution*, 16(8), pp. 1114–1116. doi: 10.1093/oxfordjournals.molbev.a026201.
- Shozu, M. *et al.* (2003) 'Estrogen excess associated with novel gain-of-function mutations affecting the aromatase gene.', *The New England journal of medicine*, 348(19), pp. 1855–1865. doi: 10.1097/01.OGX.0000109265.21876.17.
- Siegel, T. N. *et al.* (2009) 'Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei', *Genes and Development*, 23(9), pp. 1063–1076. doi: 10.1101/gad.1790409.
- Simão, F. A. *et al.* (2015) 'BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics*, 31(19), pp. 3210–3212. doi: 10.1093/bioinformatics/btv351.

Simo, G. *et al.* (2010) 'Identification of subspecies specific genes differentially expressed in procyclic forms of *Trypanosoma brucei* subspecies', *Infection, Genetics and Evolution*, 10(2), pp. 229–237. doi: 10.1016/j.meegid.2009.11.003.

Sinyangwe, L. *et al.* (2004) 'Trypanocidal drug resistance in eastern province of Zambia.', *Veterinary parasitology*, 119, pp. 125–135. doi: 10.1016/j.vetpar.2003.11.007.

Sloof, P. *et al.* (1983) 'Characterization of satellite DNA in *Trypanosoma brucei* and *Trypanosoma cruzi*', *Journal of Molecular Biology*, 167(1), pp. 1–21. doi: 10.1016/S0022-2836(83)80031-X.

Smith, J. D. *et al.* (1995) 'Switches in expression of plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes', *Cell*, 82(1), pp. 101–110. doi: 10.1016/0092-8674(95)90056-X.

Smogorzewska, A. and de Lange, T. (2004) 'Regulation of Telomerase by Telomeric Proteins', *Annual Review of Biochemistry*, 73(1), pp. 177–208. doi: 10.1146/annurev.biochem.73.071403.160049.

Söding, J. (2005) 'Protein homology detection by HMM-HMM comparison', *Bioinformatics*, 21(7), pp. 951–960. doi: 10.1093/bioinformatics/bti125.

Sonnhammer, E. L. L. and Koonin, E. V. (2002) 'Orthology, paralogy and proposed classification for paralog subtypes', *Trends in Genetics*, pp. 619–620. doi: 10.1016/S0168-9525(02)02793-2.

Spencer-Smith, R. *et al.* (2012) 'Sequence features contributing to chromosomal rearrangements in *Neisseria gonorrhoeae*', *PLoS ONE*, 7(9). doi: 10.1371/journal.pone.0046023.

Stamatakis, A. (2014) 'RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, 30(9), pp. 1312–1313. doi: 10.1093/bioinformatics/btu033.

- Stanke, M. *et al.* (2004) 'AUGUSTUS: A web server for gene finding in eukaryotes', *Nucleic Acids Research*, 32(WEB SERVER ISS.). doi: 10.1093/nar/gkh379.
- Steinbiss, S. *et al.* (2016) 'Companion: a web server for annotation and analysis of parasite genomes.', *Nucleic acids research*. doi: 10.1093/nar/gkw292.
- Stevens, J. and Gibson, W. (1999) 'The evolution of salivarian trypanosomes.', *Memórias do Instituto Oswaldo Cruz*, 94(2), pp. 225–228. doi: 10.1590/S0074-02761999000200019.
- Stevens, J. R. *et al.* (1999) 'The ancient and divergent origins of the human pathogenic trypanosomes, *Trypanosoma brucei* and *T. cruzi*', *Parasitology*, 118(1), pp. 107–116. doi: 10.1017/S0031182098003473.
- Stevens, J. R. *et al.* (2001) 'The molecular evolution of trypanosomatidae', *Advances in Parasitology*, pp. 1–56. doi: 10.1016/S0065-308X(01)48003-1.
- Stevens, J. and Rambaut, A. (2001) 'Evolutionary rate differences in trypanosomes', *Infection, Genetics and Evolution*, 1(2), pp. 143–150. doi: 10.1016/S1567-1348(01)00018-1.
- Steverding, D. (2008) 'The history of African trypanosomiasis', *Parasites & Vectors*, 1(1), p. 3. doi: 10.1186/1756-3305-1-3.
- Stockdale, C. *et al.* (2008) 'Antigenic variation in *Trypanosoma brucei*: Joining the DOTs', *PLoS Biology*, pp. 1386–1391. doi: 10.1371/journal.pbio.0060185.
- Stuart, K. *et al.* (2008) 'Kinetoplastids: Related protozoan pathogens, different diseases', *Journal of Clinical Investigation*, pp. 1301–1310. doi: 10.1172/JCI33945.
- Su, X. zhuan *et al.* (1995) 'The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of plasmodium falciparum-infected erythrocytes', *Cell*, 82(1), pp. 89–100. doi: 10.1016/0092-8674(95)90055-1.

- Sumba, A. L., Mihok, S. and Oyieke, F. A. (1998) 'Mechanical transmission of *Trypanosoma evansi* and *T. congolense* by *Stomoxys niger* and *S. taeniatus* in a laboratory mouse model', *Medical and Veterinary Entomology*, 12(4), pp. 417–422. doi: 10.1046/j.1365-2915.1998.00131.x.
- Sun, J. *et al.* (2010) 'Gene duplication in the genome of parasitic *Giardia lamblia*', *BMC Evolutionary Biology*, 10(1). doi: 10.1186/1471-2148-10-49.
- Sun, X. *et al.* (2003) 'Sequence analysis of a functional *Drosophila* centromere', *Genome Research*, 13(2), pp. 182–194. doi: 10.1101/gr.681703.
- Sundberg, L.-R. and Pulkkinen, K. (2015) 'Genome size evolution in macroparasites', *International Journal for Parasitology*, 45(5), pp. 285–288. doi: 10.1016/j.ijpara.2014.12.007.
- Sunter, J. D. and Gull, K. (2016) 'The Flagellum Attachment Zone: "The Cellular Ruler" of Trypanosome Morphology', *Trends in Parasitology*, pp. 309–324. doi: 10.1016/j.pt.2015.12.010.
- Supek, F. *et al.* (2011) 'Revigo summarizes and visualizes long lists of gene ontology terms', *PLoS ONE*, 6(7). doi: 10.1371/journal.pone.0021800.
- Swain, M. T. *et al.* (2012) 'A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs', *Nature Protocols*, 7(7), pp. 1260–1284. doi: 10.1038/nprot.2012.068.
- Swallow, B. (1999) 'Impacts of trypanosomiasis on African agriculture.', *International Livestock Research Institute, Nairobi, Kenya.*, pp. 1–46. Available at: <http://www.cabdirect.org/abstracts/20003010139.html>.
- Symula, R. E. *et al.* (2012) '*Trypanosoma brucei gambiense* group 1 is distinguished by a unique amino acid substitution in the HpHb receptor implicated in human serum resistance', *PLoS Neglected Tropical Diseases*, 6(7). doi: 10.1371/journal.pntd.0001728.
- Talbert, P. B., Bayes, J. J. and Henikoff, S. (2009) 'Evolution of centromeres and kinetochores: A two-part fugue', in *The Kinetochores: From Molecular*

*Discoveries to Cancer Therapy*, pp. 193–229. doi: 10.1007/978-0-387-69076-6\_7.

Tetley, L. and Vickerman, K. (1985) 'Differentiation in *Trypanosoma brucei*: host-parasite cell junctions and their persistence during acquisition of the variable antigen coat.', *Journal of cell science*, 74, pp. 1–19.

Tham, W. H. and Zakian, V. A. (2002) 'Transcriptional silencing at *Saccharomyces* telomeres: Implications for other organisms', *Oncogene*, pp. 512–521. doi: 10.1038/sj/onc/1205078.

Thévenaz, P. and Hecker, H. (1980) 'Distribution and attachment of *Trypanosoma* (*Nannomonas*) *congolense* in the proximal part of the proboscis of *Glossina morsitans morsitans*.' , *Acta tropica*, 37(2), pp. 163–175.

Thomashow, L. S. *et al.* (1983) 'Tubulin genes are tandemly linked and clustered in the genome of *trypanosoma brucei*', *Cell*, 32(1), pp. 35–43. doi: 10.1016/0092-8674(83)90494-4.

Thompson, M. *et al.* (2003) 'Nucleolar Clustering of Dispersed tRNA Genes', *Science*, 302(5649), pp. 1399–1401. doi: 10.1126/science.1089814.

Tjong, H. *et al.* (2012) 'Physical tethering and volume exclusion determine higher-order genome organization in budding yeast', *Genome Research*, 22(7), pp. 1295–1305. doi: 10.1101/gr.129437.111.

Tosato, V. *et al.* (2001) 'Secondary DNA structure analysis of the coding strand switch regions of five *Leishmania major* Friedlin chromosomes', *Current Genetics*, 40(3), pp. 186–194. doi: 10.1007/s002940100246.

Treangen, T. J. and Salzberg, S. L. (2012) 'Repetitive DNA and next-generation sequencing: Computational challenges and solutions', *Nature Reviews Genetics*, 13(1), pp. 36–46. doi: 10.1038/nrg3117.

Truc, P. *et al.* (2013) 'Atypical Human Infections by Animal Trypanosomes', *PLoS Neglected Tropical Diseases*, 7(9). doi: 10.1371/journal.pntd.0002256.

Tsai, A. G. and Lieber, M. R. (2010) 'Mechanisms of chromosomal

rearrangement in the human genome.', *BMC genomics*, 11 Suppl 1, p. S1. doi: 10.1186/1471-2164-11-S1-S1.

Tůmová, P. *et al.* (2016) 'Constitutive aneuploidy and genomic instability in the single-celled eukaryote *Giardia intestinalis*', *MicrobiologyOpen*, 5(4), pp. 560–574. doi: 10.1002/mbo3.351.

Turner, C. M. R. *et al.* (1988) 'An estimate of the size of the metacyclic variable antigen repertoire of *Trypanosoma brucei rhodesiense*', *Parasitology*, 97(2), pp. 269–276. doi: 10.1017/S0031182000058479.

Ukaegbu, U. E. *et al.* (2015) 'A Unique Virulence Gene Occupies a Principal Position in Immune Evasion by the Malaria Parasite *Plasmodium falciparum*', *PLoS Genetics*, 11(5), pp. 1–26. doi: 10.1371/journal.pgen.1005234.

Ullastres, A. *et al.* (2014) 'Unraveling the effect of genomic structural changes in the rhesus macaque - implications for the adaptive role of inversions', *BMC Genomics*, 15(1). doi: 10.1186/1471-2164-15-530.

Urwyler, S. *et al.* (2007) 'A family of stage-specific alanine-rich proteins on the surface of epimastigote forms of *Trypanosoma brucei*', *Molecular Microbiology*, 63(1), pp. 218–228. doi: 10.1111/j.1365-2958.2006.05492.x.

Usdin, K. and Grabczyk, E. (2000) 'DNA repeat expansions and human disease', *Cellular and Molecular Life Sciences*, pp. 914–931. doi: 10.1007/PL00000734.

Utz, S. *et al.* (2006) 'Trypanosoma congolense procyclins: Unmasking cryptic major surface glycoproteins in procyclic forms', *Eukaryotic Cell*, 5(8), pp. 1430–1440. doi: 10.1128/EC.00067-06.

Vanderelst, D. and Speybroeck, N. (2010) 'Quantifying the lack of scientific interest in neglected tropical diseases', *PLoS Neglected Tropical Diseases*. doi: 10.1371/journal.pntd.0000576.

Vanhollebeke, B. *et al.* (2008) 'A Haptoglobin-Hemoglobin Receptor Conveys Innate Immunity to *Trypanosoma brucei* in Humans', *Science*, 320(5876), pp.

677–681. doi: 10.1126/science.1156296.

Vaughan, S. (2003) 'The trypanosome flagellum', *Journal of Cell Science*, 116(5), pp. 757–759. doi: 10.1242/jcs.00287.

Vickerman, K. (1985) 'Developmental cycles and biology of pathogenic trypanosomes.', *British medical bulletin*, 41(2), pp. 105–114.

Vickerman, K. *et al.* (1988) 'Biology of African trypanosomes in the tsetse fly', *Biology of the Cell*, 64(2), pp. 109–119. doi: 10.1016/0248-4900(88)90070-6.

Walker, B. J. *et al.* (2014) 'Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement', 9(11). doi: 10.1371/journal.pone.0112963.

Wang, X. *et al.* (2012) 'Characterization of the unusual bidirectional ves promoters driving vesa1 expression and associated with antigenic variation in', *Eukaryotic Cell*, 11(3), pp. 260–269. doi: 10.1128/EC.05318-11.

Wastling, S. L. *et al.* (2011) 'Latent *Trypanosoma brucei gambiense* foci in Uganda: A silent epidemic in children and adults?', *Parasitology*, 138(12), pp. 1480–1487. doi: 10.1017/S0031182011000230.

Waterston, R. H. *et al.* (2002) 'Initial sequencing and comparative analysis of the mouse genome', *Nature*, 420(6915), pp. 520–562. doi: 10.1038/nature01262.

Weckselblatt, B. and Rudd, M. K. (2015) 'Human Structural Variation: Mechanisms of Chromosome Rearrangements', *Trends in Genetics*, pp. 587–599. doi: 10.1016/j.tig.2015.05.010.

Weedall, G. D. and Hall, N. (2015) 'Sexual reproduction and genetic exchange in parasitic protists', *Parasitology*, pp. S120–S127. doi: 10.1017/S0031182014001693.

Weil, C. F. (2009) 'Too many ends: Aberrant transposition', *Genes and Development*, pp. 1032–1036. doi: 10.1101/gad.1801309.

- Weischenfeldt, J. *et al.* (2013) 'Phenotypic impact of genomic structural variation: Insights from and for human disease', *Nature Reviews Genetics*, pp. 125–138. doi: 10.1038/nrg3373.
- Wellde, B. *et al.* (1974) 'Trypanosoma congolense: I. Clinical observations of experimentally infected cattle', *Experimental Parasitology*, 36(1), pp. 6–19.
- Wells, J. M. *et al.* (1987) 'DNA contents and molecular karyotypes of hybrid Trypanosoma brucei', *Molecular and Biochemical Parasitology*, 24(1), pp. 103–116. doi: 10.1016/0166-6851(87)90121-6.
- White, M. J. D. (1969) 'Chromosomal rearrangements and speciation in animals', *Annual Review of Genetics*, 3(7), pp. 75–98. doi: 10.1146/annurev.ge.03.120169.000451.
- Wicker, T. *et al.* (2007) 'A unified classification system for eukaryotic transposable elements', *Nature Reviews Genetics*, pp. 973–982. doi: 10.1038/nrg2165.
- Wickstead, B., Ersfeld, K. and Gull, K. (2004) 'The small chromosomes of Trypanosoma brucei involved in antigenic variation are constructed around repetitive palindromes.', *Genome research*, 14(6), pp. 1014–24. doi: 10.1101/gr.2227704.
- Wickstead, B. and Gull, K. (2007) 'Dyneins across eukaryotes: A comparative genomic analysis', *Traffic*, 8(12), pp. 1708–1721. doi: 10.1111/j.1600-0854.2007.00646.x.
- Williams, R. O., Young, J. R. and Majiwa, P. A. O. (1982) 'Genomic environment of T. brucei VSG genes: presence of a minichromosome', *Nature*, 299, pp. 417–421.
- Wilson, G. M. *et al.* (2006) 'Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla', *Genome Research*, 16(2), pp. 173–181. doi: 10.1101/gr.4456006.
- World Health Organization (2013) *Trypanosomiasis, Human African (sleeping*

sickness)., *World Health Organization*.

Wu, L. *et al.* (2017) 'Direct comparison of performance of single nucleotide variant calling in human genome with alignment-based and assembly-based approaches', *Scientific Reports*, 7(1). doi: 10.1038/s41598-017-10826-9.

Xiao, Y. P., Al-Khedery, B. and Allred, D. R. (2010) 'The Babesia bovis VESA1 virulence factor subunit 1b is encoded by the 1?? branch of the ves multigene family', *Molecular and Biochemical Parasitology*, 171(2), pp. 81–88. doi: 10.1016/j.molbiopara.2010.03.001.

Yeaman, S. (2013) 'Genomic rearrangements and the evolution of clusters of locally adaptive loci', *Proceedings of the National Academy of Sciences*, 110(19), pp. E1743–E1751. doi: 10.1073/pnas.1219381110.

Young, C. J. and Godfrey, D. G. (1983) 'Enzyme polymorphism and the distribution of Trypanosoma congolense isolates.', *Annals of tropical medicine and parasitology*, 77(5), pp. 467–81. doi: 10.1080/00034983.1983.11811740.

Zafra, G. *et al.* (2011) 'Direct analysis of genetic variability in Trypanosoma cruzi populations from tissues of Colombian chagasic patients', *Human Pathology*, 42(8), pp. 1159–1168. doi: 10.1016/j.humpath.2010.11.012.

Zakian, V. A. (1997) 'Life and cancer without telomerase', *Cell*, pp. 1–3. doi: 10.1016/S0092-8674(01)80001-5.

Zapata, L. *et al.* (2016) 'Chromosome-level assembly of *Arabidopsis thaliana* L *er* reveals the extent of translocation and inversion polymorphisms', *Proceedings of the National Academy of Sciences*, 113(28), pp. E4052–E4060. doi: 10.1073/pnas.1607532113.

Zhao, P. *et al.* (2016) 'Structural Variant Detection by Large-scale Sequencing Reveals New Evolutionary Evidence on Breed Divergence between Chinese and European Pigs', *Scientific Reports*, 6. doi: 10.1038/srep18501.

Zlatanova, J. *et al.* (2009) 'The Nucleosome Family: Dynamic and Growing', *Structure*, pp. 160–171. doi: 10.1016/j.str.2008.12.016.

Zmasek, C. M. and Eddy, S. R. (2002) 'RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs.', *BMC bioinformatics*, 3, p. 14. doi: 10.1186/1471-2105-3-14.

Zomerdijk, J. C. B. m, Kieft, R. and Borst, P. (1992) 'A ribosomal RNA gene promoter at the telomere of a mini-chromosome in *Trypanosoma brucei*', *Nucleic Acids Research*, 20(11), pp. 2725–2734. doi: 10.1093/nar/20.11.2725.

## Appendix A Perl script and command lines

### 1. Perl script used to convert a BED file format to a GFF file format:

```
#!/usr/bin/perl
# The bed file is described at http://genome.ucsc.edu/FAQ/FAQformat#format1
# the contents include the following in this order:
# chromosome, start (0-based), stop (exclusive), name, score, strand, etc.
print "\n\t This program will convert *.bed files to a *.gff v2 file\n";

# Input
if ($ARGV[0]) {
    $filename = $ARGV[0];
} else {
    print "\n\t Please type in the bed file name ";
    chomp($filename = <STDIN>);
}
unless ($filename =~ /\.bed$/) {die "please enter a *.bed file\n"}

# Ask for specific GFF information
print "What is the name for this data? ";
my $type = <STDIN>;
chomp $type;
if ($type =~ /\s/) {die("Can't have whitespace in $type\n")}
print "Enter new source name [default: data] ";
my $source = <STDIN>;
chomp $source;
if ($source eq "") {$source = 'data'}

# Do the conversion
open INFILE, $filename;
my @output;
while (my $line = <INFILE>) {
    if ($line =~ /^track/i) {next} # skip the track definition line
    chomp $line;
    my @data = split /\t/, $line;
    my $refseq = $data[0];
```

```

my $start = $data[1] + 1; # need to shift from 0-based indexing
my $end = $data[2] - 1; # need to shift from exclusive number to an inclusive
number
my $score;
# score is optional in the bed format
if ($data[4]) { $score = $data[4] } else { $score = '.' }
my $strand;
# strand is optional, but if present is either + or -
if ($data[5]) { $strand = $data[5] } else { $strand = '.' }
my $phase = '.';
my ($name, $group);
# name is optional
if ($data[3]) {
    $name = $data[3];
    $group = "$type \"$name\"";
} else {
    $group = "Experiment \"$type\"";
}
push @output,
"$refseq\t$source\t$type\t$start\t$end\t$score\t$strand\t$phase\t$group\n";
}
close INFILE;

# Output
$filename =~ s/\.bed$//;
open OUTFILE, ">$filename.gff";
print OUTFILE "##gff-version 2\n";
print OUTFILE "# generated using program $0\n";
print OUTFILE "# from source file $filename.bed\n";
print OUTFILE @output;
close OUTFILE;

```

## 2. **Extraction of Gene Ontology enrichment terms GO IDs from annotation file:**

A list of gene IDs were obtained from OrthoFinder output file. Then the GO terms were extracted using following command line:

```
grep -wFf GeneID.list Annot.gff3 |grep 'polypeptide'|grep 'Ontology term'| awk -F"Ontology term=" '{sub("-[^-];", "", $NF); print $NF}'| sed 's/;/\n/g'|sed 's/;/\n/g'|grep 'GO' > GO_ids .list
```

## 3. Extract genes/gene families Shared between all Kinetoplastids:

```
sed '/Tb/!d; /TcIL3000/!d; /TcCL/!d; /Lm/!d; /Tv/!d; /CFAC/!d; /BS/!d' OrthologousGroups.csv| grep ',
```

- To extract the gene families contain one gene only and shared across all kinetoplastids, “grep -v” instead “grep” was added to the above command line.
- For the extraction of gene families for all other combinations of kinetoplastids like the parasitic and African trypanosomes I’ve used the same above syntax with the corresponding search of species identifiers for each combination.
- All the command lines and bioinformatics tools in this analysis were used on LINUX system of CGR HPC clusters.

Appendix B *T. congolense* expression sites draft  
paper

***Trypanosoma congolense* has VSG-  
containing canonical telomeric  
structures**

Ali Hadi Abbas<sup>1,2†</sup>, Sara Silva Pereira<sup>3†\*</sup>, Simon D'Archivio<sup>4</sup>, Bill Wickstead<sup>4</sup>, Liam J. Morrison<sup>5</sup>, Neil Hall<sup>6</sup>, Christiane Hertz-Fowler<sup>1</sup>, Alistair C. Darby<sup>1</sup>, Andrew Jackson<sup>3</sup>

Ali Hadi Abbas - [A.H.Abbas@liverpool.ac.uk](mailto:A.H.Abbas@liverpool.ac.uk); alih.abbas@uokufa.edu.iq

Sara Silva Pereira\* - sara.silva-pereira@liverpool.ac.uk

Simon D'Archivio - simon.d'archivio@nottingham.ac.uk

Bill Wickstead - [bill.wickstead@nottingham.ac.uk](mailto:bill.wickstead@nottingham.ac.uk)

[Liam Morrison – liam.morrison@roslin.ed.ac.uk](mailto:Liam.Morrison@roslin.ed.ac.uk)

Neil Hall - neil.hall@earlham.ac.uk

Christiane Hertz-Fowler - chf@liverpool.ac.uk

Alistair Darby - acdarby@liverpool.ac.uk

Andrew Jackson - A.P.Jackson@liverpool.ac.uk

1 Centre for Genomic Research, Biosciences Building, Crown Street, Liverpool L69 7ZB, United Kingdom, 2 Department of Pathology, Faculty of Veterinary Medicine, University of Kufa, Najaf, Iraq, 3 Department of Infection Biology, Institute of Infection and Global Health, University of Liverpool, Liverpool Science Park Ic2, 146 Brownlow Hill, Liverpool L3 5RF, United Kingdom, 4 Centre for Genetics and Genomics, School of Life Sciences, Queen's Medical Centre, Nottingham NG7 2UH, 5 Division of infection and Immunity, Roslin Institute, Easter Bush, Midlothian, Edinburgh EH25 9RG, UK, 6 Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, UK

† Contributed equally

\*Corresponding author

## Abstract (250w)

African trypanosomiasis is a vector-borne disease of humans and animals caused by African trypanosomes (*Trypanosoma* spp.). Parasite survival in the bloodstream of their vertebrate hosts depends on antigenic variation of the variant surface glycoproteins (VSG) covering their cell surface. In *T. brucei*, a model for antigenic variation, VSG expression occurs from specialized telomeric structures known as VSG expression sites (VES). The *T. brucei* VES has a canonical topology, consisting of repeat sequences, a promoter, a VSG and diverse 'Expression Site Associated Genes (ESAGs)', mostly implicated in cell surface function. In other species, such as the major veterinary pathogen *T. congolense*, conserved telomeric structures have never been described, even though VSG expression is also telomeric. Here, we used long-read genome sequencing on the PacBio platform to establish that *T. congolense* has VSG-containing canonical telomeric structures that are analogous rather than orthologous to the *T. brucei* VES. The *T. congolense* canonical telomeric structures are mostly present in the minichromosomes and contain a complex 369bp repeat, the VSG, and four conserved non-coding regions. Non-VSG families, such as ESAG3-like, cathepsin-B and the DEAH-box RNA helicase, are found sporadically along the telomeric structures, but are not canonical features. Trypanosome antigenic variation is a model system for disease genetics. Through comparison of the VSG-containing canonical telomeric structures in *T. brucei* and *T. congolense*, we are able to reconstruct the evolution of antigenic variation in these enigmatic pathogens, and so illuminate the long process of host-parasite coevolution.

**Keywords:** *Trypanosoma congolense*, telomere biology, antigenic variation, long-read genome sequencing, variant surface glycoproteins, Expression Site Associated Genes

## Background

Telomeres are specialized nucleoprotein structures found at the chromosome ends whose main function is sequence stabilization (De Lange, 2005). In most organisms, telomeres are involved in heterochromatic structures to protect the chromosome against nucleolytic degradation and spurious recombination (Tham and Zakian, 2002; Smogorzewska and de Lange, 2004). In African trypanosomes, hemoparasites of humans and livestock, telomeres are likely to perform additional functions, especially in regulating antigenic variation (Dreesen, Li and Cross, 2007). In *Trypanosoma brucei*, the causative agent of Human sleeping sickness, the regions immediately upstream the telomeres harbor the well-studied variant surface glycoprotein (VSG) expression sites (Barry and McCulloch, 2001). Most antigenically variable pathogens have dedicated machineries for the expression of variant surface antigens (Becker *et al.*, 2004; Al-Khedery and Allred, 2006; Kyes, Kraemer and Smith, 2007; Hertz-Fowler *et al.*, 2008; Wang *et al.*, 2012; Norris, 2014). In *T. brucei*, bloodstream VSG genes are transcribed by RNA polymerase I from multiple polycistronic bloodstream form VSG expression sites (BES) (Shea, Lee and Van der Ploeg, 1987; Alexandre *et al.*, 1988; Shea and Van der Ploeg, 1988; Pays *et al.*, 1989; Navarro and Gull, 2001). These are located immediately upstream the telomeres of megabase chromosomes and consist of a promoter, sporadic transposable elements (i.e. RIME/ingi), 12 different 'expression site associated genes (ESAGs)', a 70-bp repeat, and the VSG (Berriman *et al.*, 2002; Becker *et al.*, 2004; Hertz-Fowler *et al.*, 2008). With the exception of ESAG8, which is a putative nuclear DNA-binding protein (Hoek, Engstler and Cross, 2000), all ESAGs have cell-surface roles: ESAG1 is a *T. brucei* specific gene; ESAG2 is an ancestral, invariant b-type VSG (Jackson *et al.*, 2012); ESAG3 and ESAG5 are membrane-associated proteins (Hertz-Fowler *et al.*, 2008); ESAG4 is an adenylate cyclase (Pays *et al.*, 1989; Jackson *et al.*, 2013); ESAG6 and ESAG7 are transferrin receptors (Salmon *et al.*, 1994); ESAG9 has an

unknown function; ESAG10 is a folate transporter (Hertz-Fowler *et al.*, 2008); ESAG11 is a modified invariant surface glycoprotein (Jackson, Windle and Voorheis, 1993); and ESAG12 is a *T. brucei*-specific gene of unknown function (Hertz-Fowler *et al.*, 2008). Furthermore, with the exception of ESAG10 that is identical to its core homologs, all ESAGs have derived from multi-copy gene families distributed throughout the subtelomere (Jackson *et al.*, 2013). Yet, they have been independently recruited to the ES, where they adapted and evolved concertedly (Jackson *et al.*, 2013). Of the 12 ESAGs, only three (ESAG 9, 10, 12) are not essential for BES activation.

Despite wide knowledge of *T. brucei* BES, in the veterinary parasite *Trypanosoma congolense*, responsible for the majority of animal African trypanosomiasis cases in Africa and estimated losses of 4.5 billion USD per year (HaileMeskel, 2016), such structures have not been described and telomeres have not been studied. Telomeres are complex regions to resolve with conventional sequencing techniques. The *T. brucei* telomeric-associated structures have been described using laborious and inefficient cloning-derived techniques, such as bacterial-associated cloning (BAC) and transformation-associated recombination (TAR) cloning (Berriman *et al.*, 2002; Becker *et al.*, 2004; Hertz-Fowler *et al.*, 2008). With the emergence of long-read genome sequencing technologies, such as the SMRT cell sequencing from Pacific Biosciences ('PacBio sequencing') (Rhoads and Au, 2015), these challenges can be overcome. Telomeres can be sequenced in reads as long as 60Kb, which can often contain telomeric-associated structures of the genome in single reads bypassing genome assembly. Here, we have used PacBio technology to sequence the genomes of two strains of *T. congolense* savannah, IL3000 and Tc1/148. We reveal a comprehensive set of VSG-rich telomere-associated canonical structures in *T. congolense* not orthologous to those in *T. brucei*.

Furthermore, we investigate the nature and role of telomere-associated genes in *T. congolense*.

## Methods

### 6.3 Parasite stocks and culture

***T. congolense* savannah 1/148 (MBOI/NG/60/1-148)** (Young and Godfrey, 1983): Tc1/148 procyclic forms were cultured at the Liverpool School of Tropical Medicine in modified Eagle's medium (MEM)-based modified differentiating trypanosome medium (DTM) (10% fetal bovine serum, 2mM l-glutamine, 10mM l-proline) in 25cm<sup>2</sup> flasks and incubated at 27°C, 5% CO<sub>2</sub>.

***T. congolense* savannah IL3000** (Gibson, 2012): TcIL3000 blood stage forms were cultured at the University of Glasgow in TcBSF-3 media (<http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0000618>) in 24 well plates and incubated at 34°C in humid incubator of 5% CO<sub>2</sub> atmosphere to the mid log phase, as described by Coustou et al. (2010).

Cultures were harvested when approximately 1.2x10<sup>9</sup> cells were obtained by centrifugation at 1000 g and DNA was extracted from the cell pellet using a Qiagen DNeasy Blood and Tissue Kit, following the manufacturer's instructions.

### 6.4 DNA extraction and sequencing

High Molecular Weight DNA was extracted from 1.2x10<sup>9</sup> cells using a double phenol: chloroform protocol. High Molecular Weight DNA was extracted from 1.2x10<sup>9</sup> cells by phenol: chloroform protocol. Cells were centrifuged at 1500g for 10 minutes

and washed in 10ml cold PBS. Cells were centrifuged at 1500g for 10 minutes and supernatant was discarded. Pellet was resuspended in 500ul PBS and incubated with 6ml TELT buffer (1.5M LiCl anhyd, 50mM Tris-HCl pH8.0, 62.5mM EDTA pH8.0, 4% Triton-X) at room temperature for 5minutes. In a fume hood, 7ml of 1:1 phenol: chloroform was added and mixed by inversion for 5 minutes or until emulsion was formed. Solution was centrifuged at >3000g for 5 minutes and aqueous solution collected in a 50ml tube. The remaining phases were discarded. Two volumes of ethanol were added to the aqueous solution, mixed by inversion, incubated on ice for 10 minutes and centrifuged at 4000g for 20 minutes. Supernatant was discarded by gentle decantation and pellet was washed in 2 volumes of freeze-cold 70% ethanol. Solution was centrifuged at 3000g for 5min and supernatant decanted. Pellet was left to air dry at 70°C for 5 minutes and then re-dissolved in 600ul TE50 (10mM Tris-HCl pH8.0, 50mM EDTA pH8.0). 150um/ml of RNase A was added to the resuspended pellet and incubated for 1 hour at 37C. Subsequently, 300ug/ul of Proteinase K was added to the solution and incubated for 2 hours at 50°C. After the incubation period, 600ul 1:1 phenol: chloroform was added in a fume hood and mixed by inversion for 5 minutes. Solution was centrifuged at > 3000g for 5 minutes and aqueous fraction collected in a 1.5ml tube. To the aqueous solution, 1 volume of isopropanol and 0.1 volumes of 3M sodium acetate (NaOAc) were added. The solution was centrifuged at 1500g for 15 minutes at 4C. Supernatant was discarded and pellet was washed in 1ml freeze-cold 70% ethanol. Pellet was left to air dry until no ethanol was visible. Finally, the pellet was left to re-dissolve in TE50 (2ul/10<sup>7</sup> cells) at 4°C overnight, without pipetting or mechanically disturbance. DNA was used to prepare 20Kb genomic libraries, sequenced on the PacBio® SMRT sequencer RSII (Pacific BioSciences, USA), and assembled using the Hierarchical Genome Assembly Process 3 (HGAP3) (Chin et al. *et al.*, 2013), under default conditions and a predicted genome size of 34Mb.

## 6.5 Assembly and annotation

The Tc1/148 assembly contained 536 contigs (n50 = 421,740bp), whilst the IL3000 assembly resulted in 1,541 contigs (n50 = 156,211bp). Tc1/148 assembled contigs were annotated using the web server Companion (Steinbiss *et al.*, 2016), using RATT (Otto *et al.*, 2011) on species mode to transfer relevant annotation from *T. brucei* 927 and doing *ab initio* gene finding in AUGUSTUS (Stanke *et al.*, 2004) with a score threshold of 0.7 to make gene prediction more sensitive. TcIL3000 assembled contigs were annotated using the same procedures, followed by manual curation based on BLASTx protein evidence.

## 6.6 Telomere assembly and annotation

Assemblies were screened for the telomeric repeat using Repeat Masker (<http://repeatmasker.org>) and manually curated. Telomere annotation was performed by sequence similarity search using BLASTn, tBLASTx (Altschul *et al.*, 1990) with a significance threshold (E-value) of  $10^{-4}$ , ACT (Carver *et al.*, 2005), and multiple sequence alignment (ClustalW (Larkin *et al.*, 2007)). The features identified were annotated into the respective contigs using Artemis (K Rutherford *et al.*, 2000). The *T. congolense* 1/148 genome has been deposited at DDBJ/ENA/GenBank under the accession NHOR000000000. The *T. congolense* IL3000 genome has been deposited at DDBJ/ENA/GenBank under the accession PQVL000000000.

## 6.7 Multiple Sequence Alignment

Nucleotide sequences of conserved non-coding regions of *T. congolense* 1/148 were aligned with ClustalW (Larkin *et al.*, 2007) and manually curated. The 126 conserved non-coding Region 1 (CNR1) sequences produced an alignment of 3584 nucleotides. The conserved Non-Coding Region 2 (CNR2) produced an alignment of 37 sequences and 176 nucleotides. The conserved Non-Coding Region 3 (CNR3) produced an alignment of 15 sequences and 199 nucleotides. The conserved Non-Coding Region 4 (CNR4) produced an alignment of 21 sequences and 148 nucleotides. Amino acid sequences of conserved coding regions found in telomere-containing contigs (Fam15, Fam53, DEAH-box RNA helicase, cathepsin B) were aligned with ClustalW (Larkin *et al.*, 2007). The Fam15 alignment was comprised of 217 sequences from *T. brucei* and *T. congolense* IL3000 and 1/148 of 435 amino acids. The Fam53 protein alignment consists of 173 sequences from *T. brucei*, *T. congolense* IL3000 and 1/148, *T. vivax* and *T. cruzi* and has 294 amino acids. The DEAH-box RNA helicase alignment contains 50 sequences from *T. brucei* and *T. congolense* IL3000 and 1/148 of 1089 amino acids. The cathepsin B alignment has 23 sequences from *T. brucei* and *T. congolense* IL3000 and 1/148 of 347 amino acids.

## 6.8 Phylogenetic Analysis

Phylogenies were estimated from nucleotide sequence alignments with maximum likelihood (ML) following automatic model selection (Lefort, Longueville and Gascuel, 2017) using PHYML v3.0 (S Guindon and Gascuel, 2003). Robustness was assessed with 100 bootstrap replicates.

The differences in phylogenetic signal between different loci of the expression site were calculated by evaluating the significance of the differences in likelihood

values between the optimal tree and the constrained tree. Phylogenetic topologies were constrained to the topology of CNR1 or CNR3 trees. Likelihood values recorded for the phylogenies of CNR2-4 and the differences between constrained and unconstrained topologies were calculated and statistical significance evaluated using the Shimodaira-Hasegawa test (Shimodaira and Hasegawa, 1999) in RaxML (Stamatakis, 2014). Phylogenetic signal was not measured between CNR2-CNR3/3 because the number of contigs containing both features was too low.

### **Recombination Tests**

The role of recombination in DEAH-box RNA helicase expansion was investigated by predicting breakpoints with the Genetic Algorithm for Recombination Detection (GARD) (Pond et al. 2006). GARD was run using the REV model, under the AICc information criterion. The KH test was applied to test for rate heterogeneity to account for the effect of significant topological incongruences. The role of selection in sequence evolution was assessed with three site-level selection tests (Fixed Effects Likelihood (FEL) to directly estimate  $d_n/d_s$  ratios (Pond and Frost, 2005); Random Effects Likelihood (REL) to infer selection pressures using an empirical Bayes approach and model  $d_n/d_s$  ratios at individual sites based on a pre-defined distribution; and Fast Unbiased Bayesian Approximation (FUBAR) to estimate the  $d_n/d_s$  ratio based on Bayesian Inference using a MCMC routine (Murrell *et al.*, 2013)) and one branch-level test (Branch-Site REL, which is an empirical Bayes approach to infer selection pressures in individual phylogenetic lineages (Kosakovsky Pond *et al.*, 2011)).

## **Results**

The Tc1/148 genome assembly generated 153 telomere-containing contigs. Three contigs represent portions of the megabase chromosomes 6, 10, and 11 (1.16Mb, 1.78Mb, and 1.57Mb) (Figure 1), whilst 25 represent full minichromosomes of 20,914 to 37,974bp in length (4). These were often derived from single sequencing reads and therefore unaffected by the assembly process, which can be problematic in complex repetitive regions. The remaining 125 contigs could not be indisputably allocated to either structure, although the majority is likely to represent partial minichromosomes. The IL3000 genome assembly produced 128 telomere-containing contigs, of which 2 belong to chromosome 6 and 11; 7 are full minichromosomes of 21,075 to 37,994bp in length (4) and the remaining putative partial minichromosomes. Our results suggest that, like the telomeric VSG expression sites in *T. brucei*, the telomeres of *T. congolense* have a canonical structure, conserved across chromosomes and strains. However, unlike *T. brucei*, they are mostly minichromosomes. Therefore, the features identified will be hereafter analyzed collectively in the format (Tc1/148; IL3000).

### **The canonical structure of *T. congolense* telomeres**

Early studies of the *T. congolense* karyotype have described the minichromosomes to contain a long complex repeat of 369bp, analogous to the 177bp repeat of *T. brucei* (Gibson, Dukes and Gashumba, 1988; Moser *et al.*, 1989). We found it in most of the telomeric contigs (80%; 99%) (Figure 2), including in the contig spanning chromosomes 10 (Figure 1). In strain Tc1/148, the repeat is 359bp rather than 369bp, due to a deletion between nucleotides 241 and 250. Despite that, it is broadly conserved in both strains and between contigs. A Southern blot analysis confirms that this repeat mostly represents minichromosomes (Figure 3). Conservative estimates from integration of the minichromosome ethidium stain

against single-copy chromosomes (as previously performed for *T. brucei* (Cross, Kim and Wickstead, 2014)) indicate that minichromosomes represent ~12Mb of DNA, which for an average minichromosome size of ~30kb, suggests approximately 400 MCs per nucleus. When found in megabase or intermediate chromosomes, the 369bp repeat is always the most telomeric-distal feature, separating the canonical telomeric structures from the subtelomeres, whilst in the minichromosomes it locates at the center. Its size directly influences the size of the minichromosome. Downstream the repeat, we find a GC-rich conserved non-coding region (CNR1). Despite the variable size of up to 3584bp, there is a stable region of 1368 nucleotides with 65% sequence identity, which is common to all contigs, and an ultra-conserved core with 87% sequence identity across 302 nucleotides between positions 1893 and 2195 (Figure 2a). In complete minichromosomes, the presence of CNR1 adjacent to both ends of the 369bp repeat is evidence for two distinct telomeric structures, one at each end, thus corroborating the view that the repeat represents the upstream boundary of the telomeric context (Figure).

We have also identified at least 3 conserved non-coding regions (CNR2-4) that are specific to the telomeric context and, whilst not constitutively present, maintain their position relative to each other and to the remaining elements (Figure 2). CNR2 is the most abundant, found in 31% and 60% of the contigs, usually downstream CNR1 and preceding a VSG gene. This structure is composed of 180bp repeats, and often found 2.5Kb upstream the VSG and 4kb upstream the telomere. The remaining conserved non-coding regions always locate downstream the VSG. CNR3 is a highly conserved, 200-nucleotide sequence found in 8% and 20% of the telomeric contigs. It is usually placed 1Kb downstream the VSG and 0.5Kb upstream the telomere. CNR4 is a 150bp region present in 12% and 23% of the telomeric contigs. It contains a 46bp AT-rich motif and locates 1.5Kb downstream the VSG and 0.5Kb upstream the telomere. Both CNR3 and CNR4 are annotated in the original

IL3000 genome as coding sequences (TcIL3000\_04880 and TcIL3000\_0\_12610); yet there are strong reasons to consider them non-coding: there is no evidence for their expression in the available datasets (EST library data (Helm *et al.*, 2009), epimastigote and metacyclic transcriptomes (Silva Pereira *et al.*, submitted), and bloodstream form IL3000 transcriptome (The Wellcome Trust Sanger Institute (WTSI),

<https://www.ebi.ac.uk/arrayexpress/experiments/E-ERAD-440/>)), they have multiple internal stop codons, and CNR3 has a higher GC content than the coding sequence average (60 vs. 51.7%). Both of these structures are related to the presence of pseudogenic VSGs. In fact, CNR3 and CNR4 exist more often associated with a pseudogenic VSG than an intact telomeric VSG, for instance in Tc1/148 an intact telomeric VSG only associates with CNR3 in 10% of the one occasion, and with CNR4 in three.

The most abundant gene in the telomeric region is the VSG. VSG genes are found in 59% and 40% of the telomeric-associated structures. In the vast majority, they are the most telomere-proximal coding sequence (Figure 2). As the *T. congolense* VSG repertoire is composed of 15 clades or ‘phylotypes’ (Jackson and Barry, 2012; Jackson *et al.*, 2012), we have profiled the telomeric VSGs to understand their distribution. Our results show that the telomeric VSGs represent most, but not all phylotypes (Figure 2b). Phylotype 6 and 9 were not observed at the telomeres in either Tc1/148 or IL3000. Furthermore, the size of each phylotype in the whole genome does not correlate with their abundance in the telomeric structure.

***T. congolense* telomeric-associated genes lack evidence for sequence adaptation to the telomere**

Other coding regions found within the telomeric structures include the transposable element *ingi*, *Fam67* (Figure 2c), DEAH-box RNA helicases (Figure 2d), *Fam15* (Figure 2e), *Fam53* (Figure 2d), and retrotransposon hot spot (RHS) protein. They are sporadic and variable in position.

*Fam67* encodes cathepsin B, a family of cysteine proteases that is single-copy in *T. brucei* and *T. vivax*, but has expanded in *T. congolense*. It is essential for *T. congolense* survival, being implicated in lysosomal protein degradation and immunogenicity (Mendoza-Palomares *et al.*, 2008). We found two contigs in Tc1/148 and one contig in IL3000 with similar copies of cathepsin B. Although they are homologous to the cathepsin-B genes found in the subtelomeres, showing no evidence for a *T. congolense*-specific adaptation of cathepsin-B to the telomeres, they belong to the subgroup of proteins where the catalytic cysteine has been replaced by a serine residue (Figure 2c).

DEAH-box RNA helicases were found in 11% and 5% of the telomeric regions. They are occasionally arranged in tandem pairs and the telomeric copies are similar to those in the subtelomeres. However, they all derive from a *T. congolense*-specific expansion of TcIL3000.6.290, a single-copy gene common to all trypanosomatids, located at a strand-switch region of chromosome 6, flanked at the 5' end by a conserved serine/threonine protein phosphatase that has been lost in *T. congolense*, and a dephospho-CoA kinase at the 3' end (TcIL3000\_6\_260). To test the contribution of selection to the expansion of this gene family, we searched for evidence of recombination using GARD (Kosakovsky Pond *et al.*, 2006) and subsequently performed three tests of site-level selection (FEL, FUBAR, and REL (Pond and Frost, 2005; Murrell *et al.*, 2013)) and one test of branch-level selection (BSR). Two significant breakpoints were found at nucleotide 873 and 1522, but inspection of the sequence alignment did not reveal an obvious recombination point. Regarding the role of selection, all tests agreed on one site being under diversifying selection, whilst

the remaining sites showed evidence for purifying selection, suggesting that the DEAH/box RNA helicase expansion in *T. congolense* is not driven by positive selection or gene conversion.

Fam15 is the ESAG6/7-like transferrin-receptor family, which has evolved from VSGs in *T. congolense* and *T. brucei* (Jackson *et al.*, 2012). In *T. brucei*, Fam15 contains the GPI-positive ESAG6, the GPI-negative ESAG7, and a single GPI-positive/GPI-negative tandem pair (Tb927.7.3250/3260) at a strand-switch region on chromosome 7. In this species, Fam15 is almost exclusive to the expression site and exists in tandem pairs. However, in *T. congolense*, Fam15 has expanded (45 genes compared to 23 in *T. brucei*), and both GPI-positive and GPI-negative versions of the gene are abundant throughout the subtelomeres. In the telomeres, we found Fam15 members in 5% and 4% of the contigs (Figure 2e), never as tandem pairs. Phylogenetic analysis further verifies that the *T. congolense* transferrin receptors found in the telomeres are undistinguishable from those in the subtelomeres, being paraphyletic to ESAG6/7 and therefore not orthologous (Figure 2e).

Fam53 encodes ESAG3 and ESAG3-like proteins in all trypanosomatids, in *T. brucei* Fam53 has considerably expanded and transposed to the BES to become involved in VSG expression. Despite this expansion, *T. brucei* retains subtelomeric copies of ESAG3-like genes, i.e. GRESAG3, which represent the ancestral form of the expression-site isoforms. In the *T. congolense* telomeres, Fam53 genes were found in 3% and 15% of the contigs. These sequences are homologous to the genes located in the subtelomeres and not orthologous to *T. brucei* ESAG3 genes. However, they too represent a gene expansion, as they form a clade distinct from two genes copies closer to GRESAG3, mimicking the situation in *T. brucei*.

RHS pseudogenes were found in the telomeres in 10% and 3% of the contigs (Figure 2). The lack of intact genes in these regions suggests they may result from

sporadic, arbitrary gene movements. Six additional coding sequences representing four hypothetical proteins were found in the context of the Tc1/148 telomeres, corresponding to genes TcIL3000\_0\_16860, TcIL3000\_0\_59950, TcIL3000\_0\_02720, and TcIL3000\_0\_51940 from the original IL3000 assembly.

Together, these results indicate that whilst *T. congolense* has non-VSG telomeric-associated genes, they are either paralogous or analogous, but not orthologous to *T. brucei* ESAGs and none approach their exclusivity, as they are not functionally distinct isoforms found exclusively in a telomeric context.

### **Recombination is a driver of telomeric sequence evolution**

To assess whether *T. congolense* telomeric regions are subject to homologous recombination and sequence reorganization, we have analyzed the differences in phylogenetic signal across conserved features. In the absence of recombination, the phylogenetic relationships between features along the canonical structure should be constant, only reflecting the pattern of telomere duplication. However, if homologous recombination plays a role in sequence evolution, the optimal topology for each feature will be distinct to reflect the pattern of sequence exchange and telomeric reorganization. The comparison of phylogenetic signal between CNR1 and CNR2-4 and between CNR3 and CNR4 reveals significant differences, suggesting that homologous recombination is driving sequence evolution in *T. congolense* telomeres (Figure 4). However, further analysis of recombination using GARD did not reveal any significant breakpoints within the conserved sequences.

## **Discussion**

The recent improvements in PacBio sequencing (Rhoads and Au, 2015) have allowed us to recover intact telomeric regions in unprecedented numbers. The analysis of telomeric and subtelomeric regions of two *T. congolense* strains shows that *T. congolense* has over one hundred canonical telomeric structures present mainly, but not exclusively, in minichromosomes. These consist of a complex repeat region; the VSG; and four conserved non-coding regions, here named CNR1-4, which although not present in all telomeres, have a conserved relative position. Moreover, we found sporadic non-VSG genes.

When the 369bp repeat was discovered, it was proposed as a feature of the *T. congolense* minichromosomes (Kukla *et al.*, 1987; Gibson, Dukes and Gashumba, 1988; Moser *et al.*, 1989), analogous to the 177bp repeat in *T. brucei* (Sloof *et al.*, 1983; Gibson, Dukes and Gashumba, 1988), the 170bp repeat in *T. vivax* (Dickin and Gibson, 1989), the 550bp repeat in *T. simiae* (Majiwa and Webster, 1987) and the 195bp repeat satellite DNA in *T. cruzi* (Gonzalez *et al.*, 1984). All these repeats correspond to 5-10% of the total nuclear genome and, with the exception of *T. vivax*, are AT-rich (29-35% GC) (Gonzalez *et al.*, 1984; Dickin and Gibson, 1989; Moser *et al.*, 1989). *T. brucei* Lister 427 has been estimated to contain 96 minichromosomes per cell, of a mean size of 75kb (Cross, Kim and Wickstead, 2014). Our experimental data suggests that minichromosomes in *T. congolense* are smaller (mean size of ~30kb) and more numerous. Although the majority of the canonical telomeric regions described here belong to minichromosomes and intermediate chromosomes, we surprisingly found the 369bp repeat to be associated with chromosome 10, suggesting it may sporadically be found in larger chromosomes. We find the structure of *T. congolense* minichromosomes to contain the 369bp tandem repeat in the center, flanked by conserved non-coding regions, VSG genes, sporadic non-VSG genes, such as ESAG3 and DEAH-box RNA helicases, and the eukaryotic telomeres. This canonical structure is consistent with *T. brucei* minichromosomes. They too are

defined by a large repetitive region of 177bp tandem repeats (Wickstead, Ersfeld and Gull, 2004), VSG genes (Williams, Young and Majiwa, 1982; Robinson *et al.*, 1999), other repetitive non-coding regions, such as the 70bp repeats (Sloof *et al.*, 1983), and the eukaryotic telomeres (Zomerdijk, Kieft and Borst, 1992). The minichromosomes in *T. brucei* are transcriptionally silent. VSGs harbored in the minichromosomes can only become functional when transposed to the BES present at the end of the megabase chromosomes. However, in *T. congolense* such structures remain to be defined, therefore it is possible that VSG expression is occurring from the minichromosomes. The abundance of highly conserved non-coding regions in conserved relative positioning certainly suggest that the telomeric structures described in this study may represent *T. congolense* VSG expression sites. If that is the case, then unlike *T. brucei*, minichromosomes in *T. congolense* are more than a VSG archive, harboring the majority of BES, which suggest important consequences in the mechanism of VSG expression.

Structurally, the main differences between *T. congolense* and *T. brucei* telomeric structures are the nature and abundance of coding regions in the polycistronic transcription unit, as well as the existence of defined anchor points for sequence exchange. *T. congolense* has few and non-VSG coding regions and none showing the exclusivity and sequence adaptation of *T. brucei* ESAGs. In contrast, *T. brucei* ESAGs are a necessary and constitutive feature of the BES (Becker *et al.*, 2004; Hertz-Fowler *et al.*, 2008). This species has thirteen characterized ESAGs, most of which are essential for activation and retain their order throughout the generic BES structure (Hertz-Fowler *et al.*, 2008). With the exception of ESAG3 and ESAG10, all *T. brucei* ESAGs were specifically recruited from the core and subtelomeres and independently diversified in the BES (Jackson *et al.*, 2012). This is evidenced by their monophyly and explained by rare sequence transposition between telomeric and non-telomeric loci; the absence of their orthologs from the *T. congolense* telomeres

corroborates such findings. However, *T. brucei* telomeric ESAG3 are not monophyletic; instead, they form an expanded clade of both subtelomeric and telomeric genes, distinct from the group of basal Fam53 branches. The basal group is the ancestral lineage within the family, as they cluster with *T. congolense*, *T. vivax*, and *T. cruzi* orthologs. The original IL3000 genome contained a single copy of ESAG3-like genes, similar to the ancestral lineage. However, when we add the Fam53 sequences from the new IL3000 assembly and Tc1/148, we see that most *T. congolense* Fam53 genes cluster together away from the ancestral lineage. This suggests that as in *T. brucei*, ESAG3-like genes have expanded in *T. congolense*. As such, the most parsimonious explanation is common ancestry for both *T. congolense* and *T. brucei* Fam53 genes, as previously proposed for the transferrin receptors (Jackson *et al.*, 2012). Furthermore, as the telomeric and sub-telomeric genes are structurally indistinguishable, we propose that transposition of ESAG3-like genes between the subtelomeres and the telomeres is frequent. The DEAH-box RNA helicases show a similar pattern of expansion. Sub-telomeric and telomeric gene copies derive from a single-copy core chromosomal lineage with orthologs in *T. brucei* and *T. vivax*. The telomeric gene copies are also undistinguishable from the subtelomeric copies and only present in 11% and 5% of the described canonical structures.

The remaining genes found in the telomeric canonical structure (i.e. Fam15, Fam67 and RHS) have been found in a small percentage of telomeres and they lack the intrinsic features of an ESAG: they are undistinguishable from subtelomeric homologous; therefore they have not independently adapted to the BES; and they do not show a conserved relative position in the telomeres. Therefore, we propose that they are not part of the *T. congolense* telomeric canonical structure.

Although the absence of *sensu stricto* ESAGs may be surprising given their importance for VSG expression in *T. brucei*, they have adapted independently in *T.*

*brucei* (Jackson *et al.*, 2013), hence it is not unrealistic to consider that *T. congolense* has evolved different survival mechanisms that do not require constitutive telomeric-associated genes. For example, whilst the large expression levels of transferrin receptors in *T. brucei* must be achieved by the expression of ESAG6 and 7 due to their scarcity in the other loci (Schell *et al.*, 1991; Salmon *et al.*, 1994; Jackson *et al.*, 2013), *T. congolense* has evolved a large repertoire of subtelomeric transferrin receptors whose expression would effortlessly suffice the cell requirement (Jackson *et al.*, 2013).

Despite the scarcity of genes, the *T. congolense* telomeres have at least five conserved non-coding regions, i.e. the 369bp repeat and CNR1-4. Even though they do not exist in all telomeric structures, they retain their relative positioning to each other and to the telomere; therefore, they are considered part of the telomere canonical structure. It is interesting to speculate about their role, if any, in antigenic variation. The most straightforward hypotheses would be either a role in sequence reorganization and/or VSG exchange, or in ES regulation. The phylogenetic signal along the ES of *T. congolense* is not constant; in fact, phylogenetic comparison of the cohort of CNR1 and CNR2-4 sequences shows distinct trees, suggesting frequent sequence exchange between ES. In *T. brucei*, VSG switching is facilitated by the 70bp repeat upstream the VSG and the VSG C-terminal domain, allowing sequence exchange between the subtelomeres and the BES. Yet, all of the conserved coding regions found in *T. congolense* are specific to the expression site; they could serve as annealing regions to facilitate telomeric exchange and gene conversion between telomeres, but not to recruit subtelomeric VSG. Whilst the higher number of BES in *T. congolense* may expose a reduced need for VSG gene conversion from the subtelomeres to the ES, the pool subtelomeric VSG may be too large to justify an intrinsic inability to transpose them to the telomeres. It is possible that the VSG CTDs are 3' anchor points for homologous recombination, but the correspondent 5' structure

remains unknown. If CNR1-4 are not involved in sequence reorganization, they may be regulators of ES activation. Evidence of regulatory non-coding regions is ample in parasitic genomes. Expression of *var* genes encoding the PfEMP1 proteins of *Plasmodium falciparum* (Baruch *et al.*, 1995; Smith *et al.*, 1995; Su *et al.*, 1995) is driven by their 5' noncoding sequences (Kyes, Kraemer and Smith, 2007); whilst monoallelic exclusion is controlled by the variant silencing gene PfSETvs through histone methylation and long-coding RNA binding (Jiang *et al.*, 2013). In *Babesia bovis*, a bovine haemoparasite, expression of variant erythrocyte surface antigen-1 (VESA1) is controlled by a bidirectional promoter and multiple non-coding regulatory regions flanking the variant antigens (Al-Khedery and Allred, 2006; Wang *et al.*, 2012). These non-coding regions can not only regulate promoter activation, but also drive *in situ* transcriptional switching (Wang *et al.*, 2012).

Finally, the VSGs found in the ES of both strains were diverse. The *T. congolense* VSG repertoire is divided in 15 self-contained phylotypes, of distinct genomic proportions but present in all *T. congolense* genomes analyzed to date (Silva Pereira *et al.*, submitted; Jackson *et al.* 2012). In the telomere-associated structures, we found VSGs from most, but not all phylotypes, despite these phylotypes being abundant in other parts of the genome. In particular, phylotype 6 and 9 were not observed in the telomeres of either Tc1/148 or IL3000. Such absence, which is hardly explained by chance, might suggest that these phylotypes encode non-variant genes expressed from the subtelomeres. Neofunctionalization of variant antigens is a recurrent phenomenon in African trypanosomes, illustrated by the examples of ESAG2, the transferrin receptor, and the VSG-related (VR) genes (Schell *et al.*, 1991; Salmon *et al.*, 1994; Marcello and Barry, 2007a; Jackson *et al.*, 2012; Jackson, 2016), as well as in other parasites, such as *ves2* in *Babesia bovis* (Allred *et al.*, 2000; Xiao, Al-Khedery and Allred, 2010; Jackson *et al.*, 2014), *vir-D* in *Plasmodium vivax* (Neafsey *et al.*, 2012; Frech and Chen, 2013; Jackson, 2016), and *var2csa* in

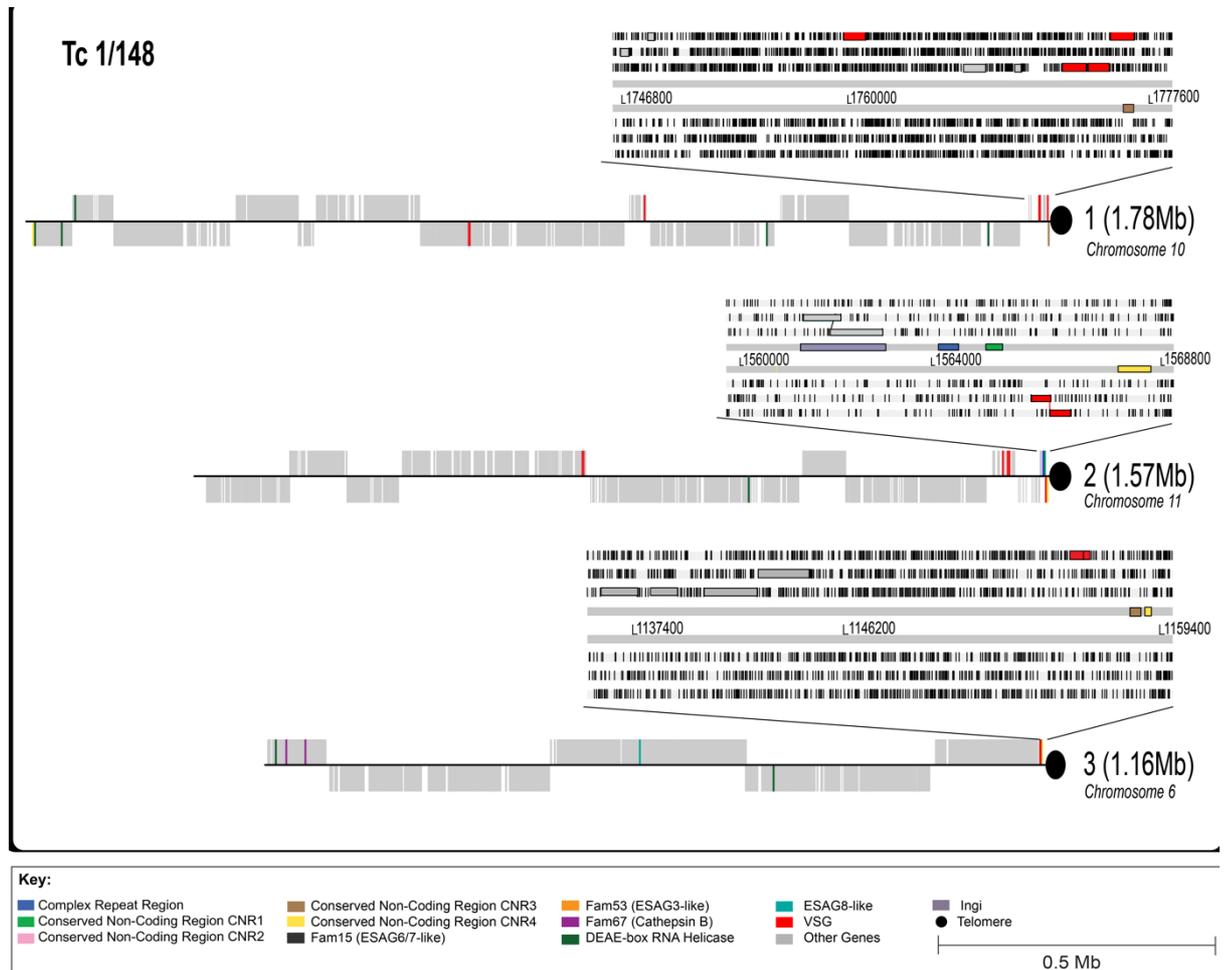
*Plasmodium falciparum* (Kraemer and Smith, 2003; Ukaegbu *et al.*, 2015; Bryant *et al.*, 2017). Whilst functional variant antigens are the fastest evolving genes in antigenically variable parasitic genomes, specific neofunctionalization events are usually associated with orthology conservation and purifying selection to support important roles in pathogenesis, virulence or parasite survival, which may well apply to *T. congolense* VSG phlotypes 6 and 9.

## Conclusions

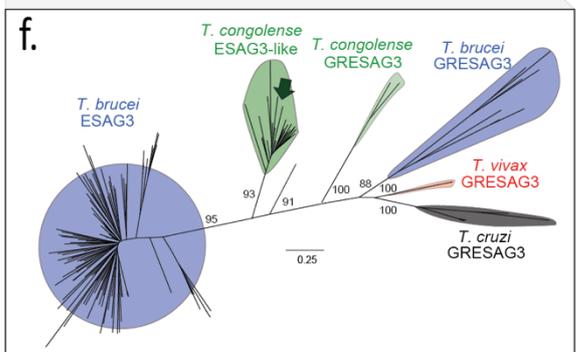
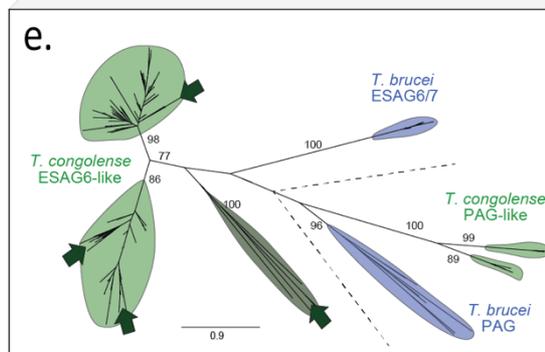
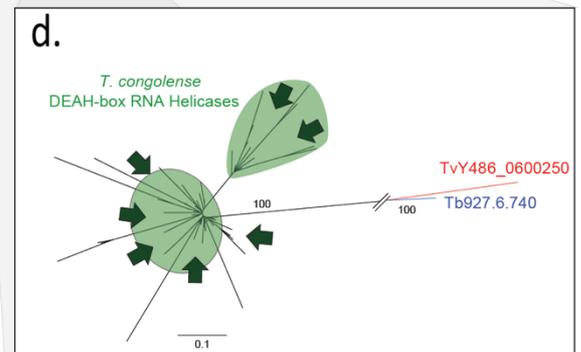
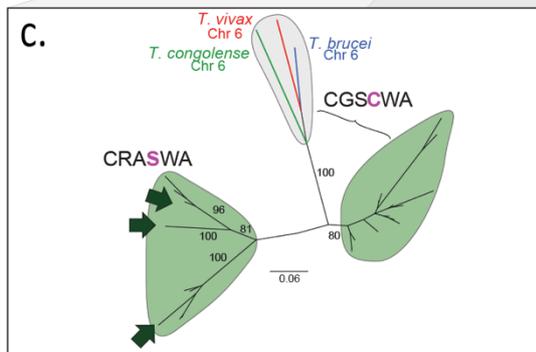
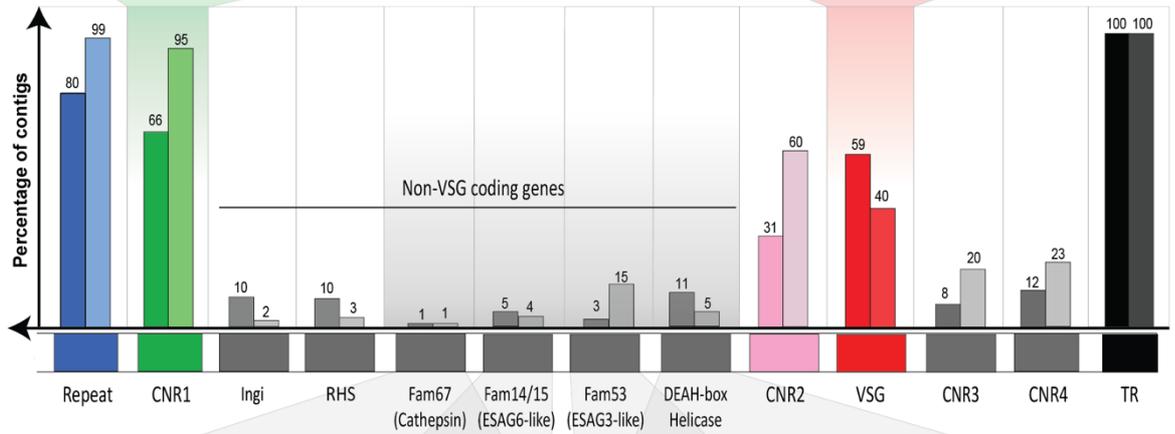
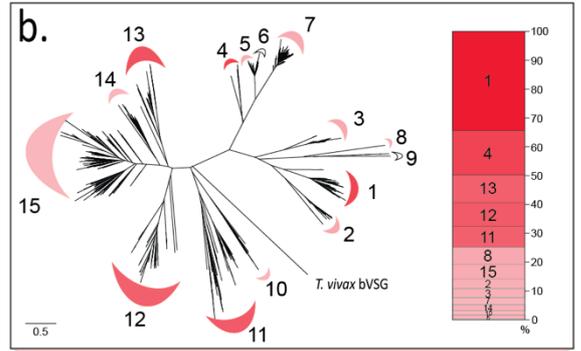
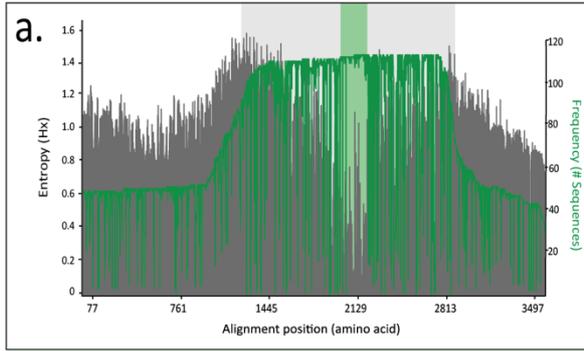
We described 278 telomeric structures located in minichromosomes of variable lengths and in one megabase chromosome. These structures are canonical, conserved across telomeres and strains, and composed of one complex repeat, VSG gene(s), and at least four conserved non-coding regions, specific to the telomeric context. The presence of non-VSG genes, including cathepsin B, DEAH-box RNA helicase, and ESAG3-like genes, was sporadic and they lack evidence of sequence adaptation to the telomeric environment, revealing a major dichotomy with the telomeric BES structure in *T. brucei*. Furthermore, we showed that homologous recombination and sequence reorganization is happening, although the specific recombination breakpoints are yet to be determined. We suggest that the structures described here are VSG expression sites, analogous to *T. brucei* BES. As such, we propose that their lowest common ancestor also expressed VSGs from telomeric ES, although extensive sequence diversification has erased their signature. The substantial differences in structure found in this study may bring large implications to the molecular mechanism of antigenic variation in both species, and be linked to the ability to survive in varying host environments. This work raises a number of important

questions about the species-specific mechanisms of antigenic variation, such as the functional role of the conserved non-coding regions found in *T. congolense* BES, the reason behind the ESAG scarcity in expression sites, and the *T. congolense*-specific fitness advantage of evolving canonical expression sites in the minichromosomes.

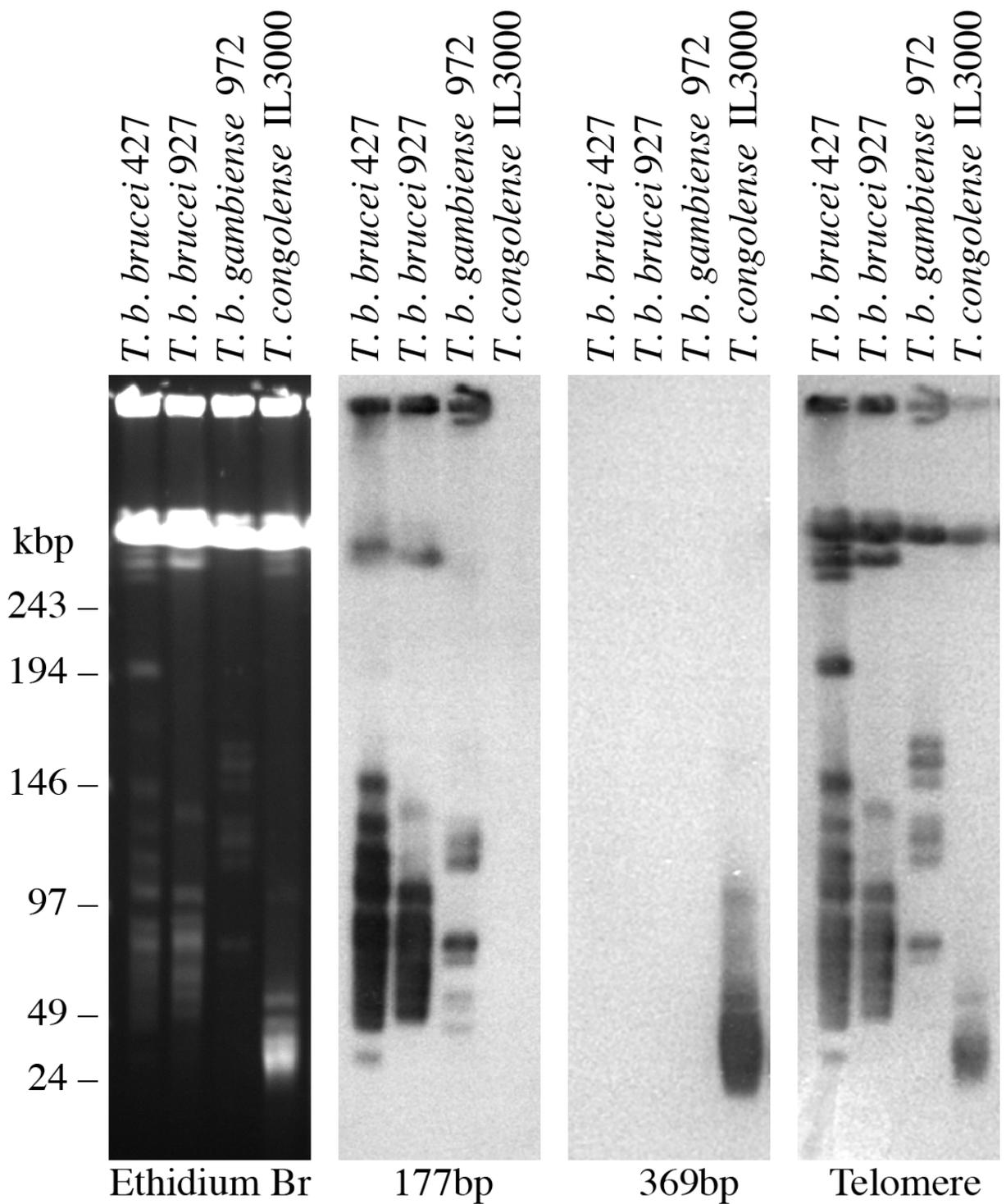
# Figures



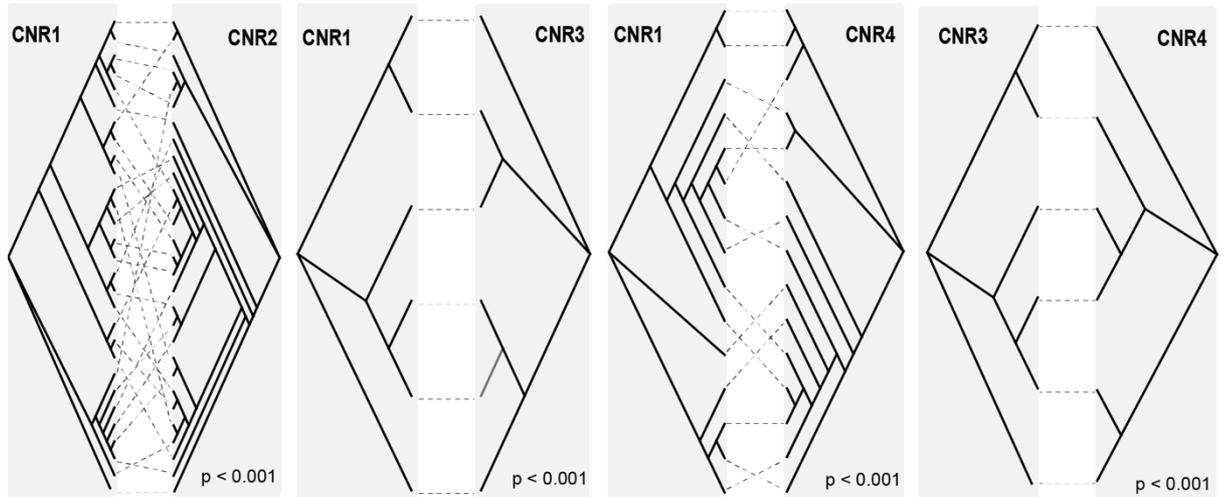
**Fig.1 Structure of the telomere-containing contigs from the Tc1/148 megabase chromosomes.** Contigs 1-3 belong to chromosome 10, 5, 6, respectively, according to sequence similarity searches. Contigs are drawn to scale and have been aligned at their telomeric end. Conserved features are shown according to key.



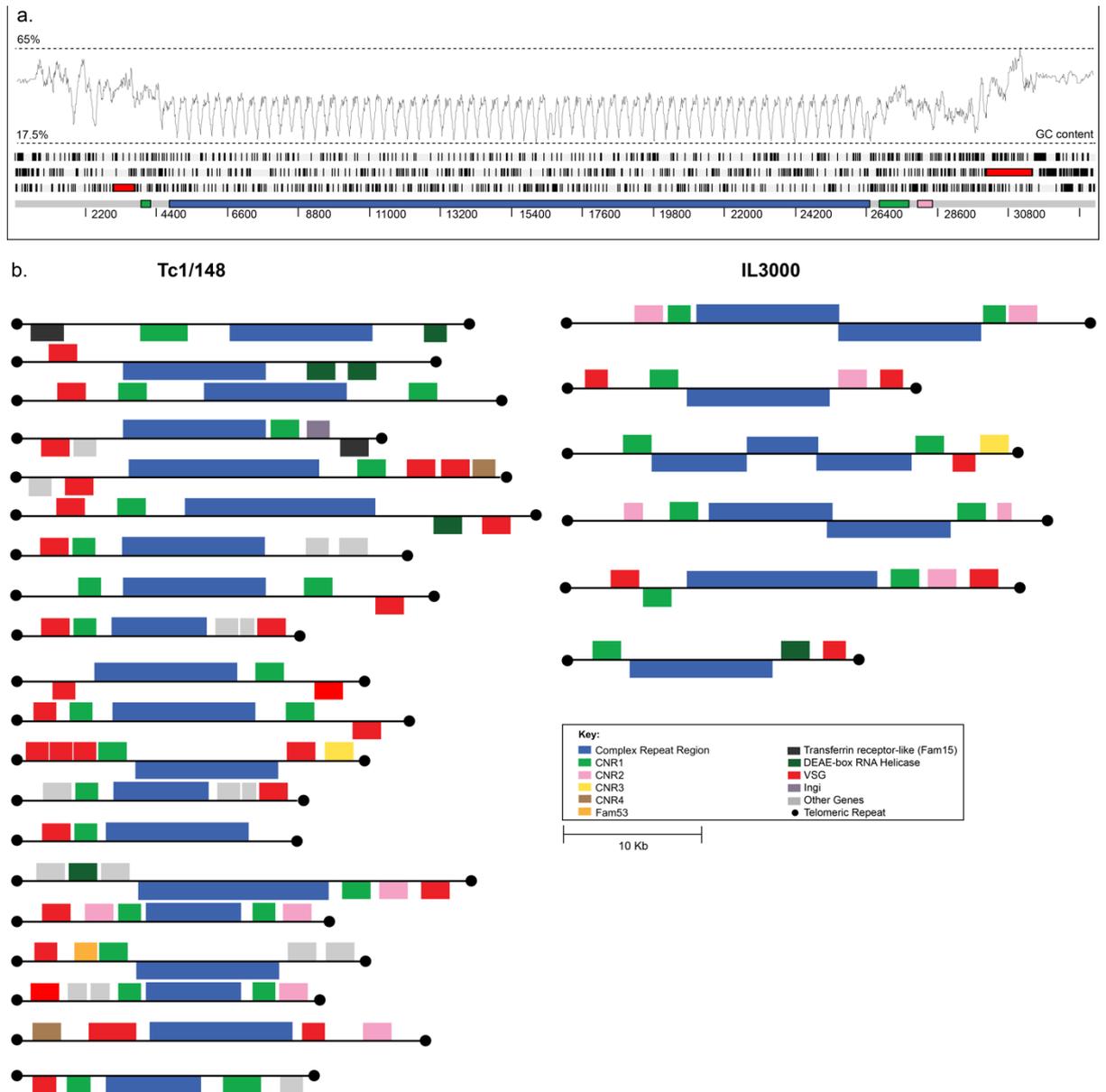
**Fig.2 The structure and gene content of the *T. congolense* canonical telomeric structures.** The central panel shows a cartoon (not to scale) of the canonical expression, indicating the frequency with which features were observed. **(a)** Sequence conservation around conserved non-coding region (CNR) 1, expressed as the entropy in the multiple sequence alignment (left axis) and the residue conservation across contigs (right axis). **(b)** Maximum likelihood phylogeny of VSG amino acid sequences and the contribution of phylotypes 1-15 to total VSG repertoire observed within expression sites. **(c)** Maximum likelihood phylogeny of cathepsin B amino acid sequences (Fam67) estimated with PHYML (Guindon *et al.*, 2010) with a WAG+ model and 100 bootstrap replicates, showing the position of telomeric-associated genes. **(d)** Maximum likelihood phylogeny of ATP-dependent DEAH RNA helicase amino acid sequences estimated with PHYML (Guindon *et al.*, 2010) with a VT++F model and 100 bootstrap replicates, showing the position of telomeric-associated genes. **(e)** Maximum likelihood phylogeny of transferrin receptor-like amino acid sequences (Fam14/15) estimated with PHYML (Guindon *et al.*, 2010) with a JTT+ model and 100 bootstrap replicates, showing the position of telomeric-associated genes. **(f)** Maximum likelihood phylogeny of ESAG3-like amino acid sequences (Fam53) estimated with PHYML (Guindon *et al.*, 2010) with a WAG+ model and 100 bootstrap replicates, showing the position of telomeric-associated genes. Terminal nodes are coloured by species according to key; bootstraps higher than 70% are shown in the internal nodes. Green arrows show the position of *T. congolense* telomeric genes.



**Fig.3** The karyotype of *T. brucei* spp. and *T. congolense* as shown by staining with ethidium bromide, hybridization to labelled minichromosome satellite repeats (177bp in *T. brucei* spp. and 369bp in *T. congolense*, and hybridization to labelled telomeres.



**Fig.4 The difference in phylogenetic signal along the conserved non-coding regions of the telomeric structures.** The first three tanglegrams relate the CNR1 phylogeny with those of CNR2-3 and the last relates CNR3 with CNR4. Grey lines link corresponding contigs. Incongruence is highest in CNR2 and lowest in CNR3, although still significant.



**Fig.5** The structure of the *T. congolense* minichromosomes in Tc1/148 and IL3000. A. Artemis plot of a complete minichromosome with a GC content graph. B. Minichromosomes are drawn to scale and have been aligned at their 5' end. Conserved features are shown according to key.

## Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The datasets analysed during the current study are available from the corresponding author on reasonable request. The genomes of Tc1/148 and TcIL3000 have been deposited in Genbank (accession NHOR00000000 and PQVL00000000).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by a Grand Challenges (Round 11) award from the Bill and Melinda Gates Foundation and a BBSRC New investigator Award (BB/M022811/1) to APJ; by Iraqi ministry of higher education and scientific research/Iraqi cultural Attache' award (977) awarded to Alistair C. Darby and Ali H. Abbas.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Acknowledgements

I would like to acknowledge Iraqi ministry of Higher Education and Scientific research/University of Kufa/Faculty of veterinary Medicine for providing the funding opportunity of part of this research.

## References

Al-Khedery B, Allred DR. 2006. Antigenic variation in *Babesia bovis* occurs through segmental gene conversion of the ves multigene family, within a bidirectional locus of active transcription. *Mol. Microbiol.* 59:402–414. doi: 10.1111/j.1365-2958.2005.04993.x.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–10. doi: 10.1016/S0022-2836(05)80360-2.

Barry JD, Ginger ML, Burton P, McCulloch R. 2003. Why are parasite contingency genes often associated with telomeres? *Int. J. Parasitol.* 33:29–45. doi: 10.1016/S0.

Berriman M et al. 2002. The architecture of variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. *Mol Biochem Parasitol.* 122:131–140. doi: 10.1016/S0166-6851(02)00092-0.

Berriman M et al. 2005. The genome of the African trypanosome *Trypanosoma brucei*. *Science.* 309:416–422. doi: 10.1126/science.1112642.

Brown C a, Murray AW, Verstrepen KJ. 2010. Rapid Expansion and Functional Divergence of Sub-telomeric Gene Families in Yeasts. *Curr. Biol.* 20:895–903. doi: 10.1016/j.cub.2010.04.027.Rapid.

Bruen TC, Philippe H, Bryant D. 2006. A Simple and Robust Statistical Test for Detecting the Presence of Recombination. *Genetics.* 172:2665–2681. doi: 10.1534/genetics.105.048975.

Carver TJ et al. 2005. ACT: The Artemis comparison tool. *Bioinformatics.* 21:3422–3423. doi: 10.1093/bioinformatics/bti553.

Coustou V, Guegan F, Plazolles N, Baltz T. 2010. Complete in vitro life cycle of

*Trypanosoma congolense*: Development of genetic tools. *PLoS Negl. Trop. Dis.* 4. doi: 10.1371/journal.pntd.0000618.

Cross G a M, Kim HS, Wickstead B. 2014. Capturing the variant surface glycoprotein repertoire (the VSGnome) of *Trypanosoma brucei* Lister 427. *Mol. Biochem. Parasitol.* 195:59–73. doi: 10.1016/j.molbiopara.2014.06.004.

Das S, Nozawa M, Klein J, Nei M. 2008. Evolutionary dynamics of the immunoglobulin heavy chain variable region genes in vertebrates. *Immunogenetics.* 60:47–55. doi: 10.1007/s00251-007-0270-2.

Deitsch KW, Lukehart SA, Stringer JR. 2009. Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nat. Rev. Microbiol.* 7:493–503. doi: 10.1038/nrmicro2145.

Dickin SK, Gibson WC. 1989. Hybridisation with a repetitive DNA probe reveals the presence of small chromosomes in *Trypanosoma vivax*. *Mol. Biochem. Parasitol.* 33:135–142. doi: 10.1016/0166-6851(89)90027-3.

Duraisingh MT, Horn D. 2016. Epigenetic Regulation of Virulence Gene Expression in Parasitic Protozoa. *Cell Host Microbe.* 19:629–640. doi: 10.1016/j.chom.2016.04.020.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797. doi: 10.1093/nar/gkh340.

Frank M, Deitsch K. 2006. Activation, silencing and mutually exclusive expression within the var gene family of *Plasmodium falciparum*. *Int. J. Parasitol.* 36:975–985. doi: 10.1016/j.ijpara.2006.05.007.

Gibson W. 2012. The origins of the trypanosome genome strains *Trypanosoma brucei* brucei TREU 927, *T. b. gambiense* DAL 972, *T. vivax* Y486 and *T. congolense* IL3000.

Parasit. Vectors. 5:71. doi: 10.1186/1756-3305-5-71.

Gibson WC, Dukes P, Gashumba JK. 1988. Species-specific DNA probes for the identification of African trypanosomes in tsetse flies. *Parasitology*. 97:63–73. doi: 10.1017/S0031182000066749.

Glover L, Alsford S, Horn D. 2013. DNA Break Site at Fragile Sub-telomeres Determines Probability and Mechanism of Antigenic Variation in African Trypanosomes. *PLoS Pathog*. 9. doi: 10.1371/journal.ppat.1003260.

Guindon S et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* 59:307–321. doi: 10.1093/sysbio/syq010.

Guindon S, Gascuel O. 2003. A Simple, Fast, and Accurate Method to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.* 52:696–704. doi: 10.1080/10635150390235520.

Hayashida K et al. 2013. MDM2 regulates a novel form of incomplete neoplastic transformation of *Theileria parva* infected lymphocytes. *Exp. Mol. Pathol.* 94:228–238. doi: 10.1016/j.yexmp.2012.08.008.

Helm JR et al. 2009. Analysis of expressed sequence tags from the four main developmental stages of *Trypanosoma congolense*. *Mol. Biochem. Parasitol.* 168:34–42. doi: 10.1016/j.molbiopara.2009.06.004.

Hertz-Fowler C et al. 2008. Telomeric expression sites are highly conserved in *Trypanosoma brucei*. *PLoS One*. 3:e3527. doi: 10.1371/journal.pone.0003527.

Horn D. 2014. Antigenic variation in African trypanosomes. *Mol. Biochem. Parasitol.* 195:123–129. doi: 10.1016/j.molbiopara.2014.05.001.

Hovel-Miner G, Mugnier MR, Goldwater B, George A. 2016. A Conserved DNA Repeat Promotes Selection of a Diverse Repertoire of *Trypanosoma brucei* Surface Antigens from the Genomic Archive. *PLoS Genet.* 1–19. doi: 10.1371/journal.pgen.1005994.

Jackson AP et al. 2013. A Cell-surface Phylome for African Trypanosomes. *PLoS Negl. Trop. Dis.* 7:e2121. doi: 10.1371/journal.pntd.0002121.

Jackson AP et al. 2012. Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. *Proc. Natl. Acad. Sci. U. S. A.* 109:3416–21. doi: 10.1073/pnas.1117313109.

Jackson AP. 2016. Gene family phylogeny and the evolution of parasite cell surfaces. *Mol. Biochem. Parasitol.* 209:64–75. doi: 10.1016/j.molbiopara.2016.03.007.

Jackson AP et al. 2010. The genome sequence of *Trypanosoma brucei gambiense*, causative agent of chronic human African Trypanosomiasis. *PLoS Negl. Trop. Dis.* 4. doi: 10.1371/journal.pntd.0000658.

Jiang L et al. 2013. PfSETvs methylation of histone H3K36 represses virulence genes in *Plasmodium falciparum*. *Nature.* 499:223–227. doi: 10.1038/nature12361.

Kissinger JC, DeBarry J. 2011. Genome cartography: Charting the apicomplexan genome. *Trends Parasitol.* 27:345–354. doi: 10.1016/j.pt.2011.03.006.

Kosakovsky Pond SL, Frost SDW, Muse S V. 2005. HyPhy: Hypothesis testing using phylogenies. *Bioinformatics.* 21:676–679. doi: 10.1093/bioinformatics/bti079.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. GARD: A genetic algorithm for recombination detection. *Bioinformatics.* 22:3096–3098. doi: 10.1093/bioinformatics/btl474.

Kukla BA, Majiwa PAO, Young JR, Moloo SK, ole-Moiyoi O. 1987. Use of species-specific DNA probes for detection and identification of trypanosome infection in tsetse flies. *Parasitology*. 95:1. doi: 10.1017/S0031182000057498.

Larkin MA et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*. 23:2947–2948. doi: 10.1093/bioinformatics/btm404.

Lefort V, Longueville J-E, Gascuel O. 2017. SMS: Smart Model Selection in PhyML. *Mol. Biol. Evol.* 4–6. doi: 10.1093/molbev/msx149.

Majiwa PAO, Matthyssens G, Williams RO, Hamers R. 1985. Cloning and analysis of *Trypanosoma* (*Nannomonas*) *congolense* ILNat 2.1 VSG gene. *Mol. Biochem. Parasitol.* 16:97–108. doi: 10.1016/0166-6851(85)90052-0.

Majiwa PAO, Webster P. 1987. A repetitive deoxyribonucleic acid sequence distinguishes *Trypanosoma simiae* from *T. congolense*. *Parasitology*. 95:543. doi: 10.1017/S0031182000057978.

Mamoudou A, Njanloga A, Hayatou A, Suh PF, Achukwi MD. 2016. Animal trypanosomosis in clinically healthy cattle of north Cameroon: Epidemiological implications. *Parasites and Vectors*. 9:1–8. doi: 10.1186/s13071-016-1498-1.

Mendoza-Palomares C et al. 2008. Molecular and biochemical characterization of a cathepsin B-like protease family unique to *Trypanosoma congolense*. *Eukaryot. Cell*. 7:684–697. doi: 10.1128/EC.00405-07.

Milne I et al. 2009. TOPALi v2: A rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics*. 25:126–127. doi: 10.1093/bioinformatics/btn575.

Moser DR et al. 1989. Detection of *Trypanosoma congolense* and *Trypanosoma brucei* subspecies by DNA amplification using the polymerase chain reaction.

Parasitology. 99:57. doi: 10.1017/S0031182000061023.

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32:268–274. doi: 10.1093/molbev/msu300.

Otto TD, Dillon GP, Degraeve WS, Berriman M. 2011. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.* 39. doi: 10.1093/nar/gkq1268.

Rehmeyer C et al. 2006. Organization of chromosome ends in the rice blast fungus, *Magnaporthe oryzae*. *Nucleic Acids Res.* 34:4685–4701. doi: 10.1093/nar/gkl588.

Reid AJ. 2015. Large, rapidly evolving gene families are at the forefront of host-parasite interactions in Apicomplexa. *Parasitology.* 142:857–870. doi: 10.1017/S0031182014001528.

Robinson NP, Burman N, Melville SE, Barry JD. 1999. Predominance of duplicative VSG gene conversion in antigenic variation in African trypanosomes. *Mol. Cell. Biol.* 19:5839–46.

Rudenko G. 2011. African trypanosomes: the genome and adaptations for immune evasion. *Essays Biochem.* 51:47–62. doi: 10.1042/bse0510047.

Rutherford K et al. 2000. Artemis: sequence visualization and annotation. *Bioinformatics.* 16:944–945. doi: 10.1093/bioinformatics/16.10.944.

Salmon D et al. 1994. A novel heterodimeric transferrin receptor encoded by a pair of VSG expression site-associated genes in *T. brucei*. *Cell.* 78:75–86. doi: 10.1016/0092-8674(94)90574-6.

Sargeant TJ et al. 2006. Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biol.* 7. doi: 10.1186/gb-2006-7-2-r12.

Shimodaira H, Hasegawa M. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol. Biol. Evol.* 16:1114–1116. doi: 10.1093/oxfordjournals.molbev.a026201.

Sloof P, Bos JL, et al. 1983. Characterization of satellite DNA in *Trypanosoma brucei* and *Trypanosoma cruzi*. *J. Mol. Biol.* 167:1–21. doi: 10.1016/S0022-2836(83)80031-X.

Sloof P, Menke HH, Caspers MPM, Borst P. 1983. Size fractionation of *Trypanosoma brucei* DNA: Localization of the 177-bp repeat satellite DNA and a variant surface glycoprotein gene in a minichromosomal DNA fraction. *Nucleic Acids Res.* 11:3889–3901. doi: 10.1093/nar/11.12.3889.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 30:1312–1313. doi: 10.1093/bioinformatics/btu033.

Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Res.* 32. doi: 10.1093/nar/gkh379.

Steinbiss S et al. 2016. *Companion*: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res.* 44:gkw292. doi: 10.1093/nar/gkw292.

Takeet MI et al. 2013. Molecular survey of pathogenic trypanosomes in naturally infected Nigerian cattle. *Res. Vet. Sci.* 94:555–561. doi: 10.1016/j.rvsc.2012.10.018.

Vink C, Rudenko G, Seifert HS. 2012. Microbial antigenic variation mediated by homologous DNA recombination. *FEMS Microbiol. Rev.* 36:917–948. doi: 10.1111/j.1574-6976.2011.00321.x.Microbial.

Wang X et al. 2012. Characterization of the unusual bidirectional ves promoters driving vesa1 expression and associated with antigenic variation in. *Eukaryot. Cell.*

11:260–269. doi: 10.1128/EC.05318-11.

Wickstead B, Ersfeld K, Gull K. 2004. The Small Chromosomes of *Trypanosoma brucei* Involved in Antigenic Variation Are Constructed Around Repetitive Palindromes. *Genome Res.* 1014–1024. doi: 10.1101/gr.2227704.).

Williams RO, Young JR, Majiwa PAO. 1982. Genomic environment of *T. brucei* VSG genes: presence of a minichromosome. *Nature.* 299:417–421.

Young CJ, Godfrey DG. 1983. Enzyme polymorphism and the distribution of *Trypanosoma congolense* isolates. *Ann. Trop. Med. Parasitol.* 77:467–81. doi: 10.1080/00034983.1983.11811740.

Zomerdijk JCB m, Kieft R, Borst P. 1992. A ribosomal RNA gene promoter at the telomere of a minichromosome in *Trypanosoma brucei*. *Nucleic Acids Res.* 20:2725–2734. doi: 10.1093/nar/20.11.2725.

## **Appendix C Additional materials of Chapter 4 and 5**

See supplementary data.