THE UNIVERSITY *of* EDINBURGH

THE UNIVERSITY of EDINBURGH

Kinetoplast DNA Dynamics in Trypanosoma Species: The Impact of Life Cycle Variations and Reproduction Strategies

Zihao Chen

School of Biological Sciences

Thesis submitted for a degree of

Doctor of Philosophy

University of Edinburgh

2024

I. Preface

i. Abstract

Trypanosomatids are unicellular, flagellated obligatory protozoa parasites. Many dixenous trypanosomatids, such as trypanosome parasites in the genus *Trypanosoma*, cause diseases in humans and livestock. Human diseases due to trypanosome parasites mainly occur in developing or undeveloped countries, including Chagas in South and Central America (*Trypanosoma cruzi*), chronic Human African Trypanosomiasis (HAT) in Central and West Africa (*Trypanosoma. brucei. gambiense* type 1), and acute HAT in East Africa (*Trypanosoma brucei rhodesiense*). Meanwhile, *Trypanosoma brucei brucei*, *Trypanosoma congolense*, *Trypanosoma brucei equiperdum*, and *Trypanosoma evansi* afflict animals and cause Animal African Trypanosomiasis (AAT), nagana, dourine, and surra respectively.

The single mitochondrion of trypanosomatids contains a massive genome, the kinetoplast. Within an individual parasite, the kinetoplast DNA (kDNA) forms a chainmail-like network with two types of catenated DNA molecules: 20 to 50 copies of essentially identical maxicircles and thousands of highly heterogeneous minicircles. Maxicircle encodes ribosomal and electron transport chain subunits. The pre-mRNAs of 12 genes require post-transcriptional editing directed by short “guide RNAs” (gRNAs) encoded on minicircles. To produce translatable mRNAs, trypanosomatids must cover all editing sites with at least one gRNA. In species with extensive editing such as *T. brucei*, the kDNA network contains a highly diverse population of minicircles and encodes hundreds of distinct gRNAs.

The lifecycle of dixenous trypanosomatids involves insect vectors and mammalian hosts. During clonal reproduction, imperfect replication and segregation of kDNA may cause some minicircles encoding essential genes to drift towards a dangerously low abundance. In trypanosome parasites, sexual reproduction occurs exclusively in the insect vectors and results in mixing of the mitochondrial genome in the progeny. Circulating minicircles among tsetse-transmissible isolates, sexual reproduction potentially rescues low-abundance gRNA genes in the progeny by replenishing it with copies from the other parent. In addition, the different metabolisms at the mammalian and insect stages entail a different set of essential maxicircle genes and a lower demand for editing capacity in the mammalian stage. Consequently, deviations from the typical lifecycle can present a unique challenge in maintaining the kDNA integrity.

We propose that sexual reproduction is key in combating genetic diversity loss in kDNA. Conversely, the absence of sexual recombination reduces kDNA complexity. Using next-generation sequencing data, we have assembled and examined the kDNA from trypanosome species and subspecies with different life histories. Also transmitted by tsetse, the human pathogen *T. b. gambiense* type 1 reproduces strictly clonally. The clonal *T. b. equiperdum* and *T. b. evansi* no longer rely on tsetse but are transmitted directly between mammals. The dixenous *T. congolense* is colocalized with *T. brucei* and reproduces sexually in the proboscis of tsetse flies.

We report a highly conserved minicircle population characteristic of *T. b. gambiense* type 1 in 117 isolates. Comparing *T. b. gambiense* type 1 to the sexual *T. brucei* subspecies, we observed substantial kDNA streamlining in the asexual isolates with evidence of approaching tsetse transmissibility loss.

We confirm that in three groups of asexual kDNA-independent *T. b. evansi* and *T. b. equiperdum*, the minicircle genomes consist of thousands of a single minicircle class specific to each group. A putatively kDNA-independent ecotype, *T. b. equiperdum* group OVI retains moderate kDNA complexity and can probably produce fully edited mRNAs of A6 and RPS12, the only edited maxicircle genes required in mammalian bloodstream.

Comparison between *T. brucei* and *T. congolense* gRNA annotation revealed highly conserved editing blocks that cover the edited mRNAs with minimal overlaps. The results shed light on the evolution of the editing cascade.

ii. Lay summary

Causing human and animal diseases, many trypanosome parasites are transmitted by blood-sucking insects. The parasites adapt to the different environments in the insect vectors and mammalian hosts and must maintain enough DNA information to support the necessary cell activities. The single-cellular trypanosome parasite has a single mitochondrion. The massive mitochondrial genome consists of two types of circular DNAs: 20 to 50 copies of practically identical maxicircles and thousands of highly diverse minicircles. The maxicircles and minicircles contain complementary information. Maxicircles provide instructions for making essential proteins, but the instructions usually require modifications guided by the information on minicircles. Hence, maxicircles and minicircles act like two sets of interdependent keys for deciphering an encoded message. For each protein encoded by maxicircles, multiple keys from minicircles are needed to produce the complete message. Some keys are present in multiple different minicircles, and the copy number of minicircles varies.

When the parasites clone themselves, the clones do not always receive a perfect copy of the minicircles, so their abundance fluctuates between generations. However, while the trypanosome parasites only produce clones in the mammalian hosts, in the insect vectors, under the right conditions, they may have sex. In humans and most mammals, the mitochondrial genome is always inherited from the egg, or the 'mother'. However, the trypanosome parasites inherit maxicircles from one parent but have minicircles from both parents. This way the hybrids can probably regain the minicircles lost in one of the parents and even acquire new minicircles. However, a few groups of trypanosome parasites have given up sex and only replicate clonally. We were interested in the impact of this atypical lifestyle on the mitochondrial genome, especially on the minicircle population.

First, we looked at the mitochondrial genome of the sub-Saharan trypanosome parasites. A human-infective subspecies no longer performs sexual reproduction. We compared the minicircle population among sub-Saharan trypanosomes and realized that all the groups within this clonal subspecies shared a similar minicircle composition. In addition, the clonal group does not have as many different types of minicircles as the other groups. We also check the parasites' ability to decipher the maxicircle-encoded information. We concluded that the clonal subspecies is not as competent as the subspecies capable of sexual reproduction in decoding the information. We proposed that most cells of the clonal subspecies cannot produce some vital proteins for the insect stage. The observation suggested that sexual reproduction is critical for maintaining a healthy and fully functional mitochondrial genome in trypanosome parasites.

Secondly, we checked the mitochondrial genome of trypanosome parasites believed to be able to survive without the mitochondrial genome. These parasites no longer proliferate in insects. We found no mitochondrial DNA or a homogeneous minicircle population in most groups. Neither was enough for deciphering maxicircle-encoded information as expected. However, we detected in one group a small set of minicircles that provide enough information for making two proteins. The two proteins are the only essential ones from the mitochondrial genome that require additional corrections when the parasites live in

mammals. In addition, a drug that disrupted mitochondrial DNA production managed to kill the parasite. We concluded that this group represented a unique case of trypanosome parasites that proliferate solely in mammals, do not have sex, but still need the mitochondrial genome. They will probably live with the minimal minicircle set until a vital key is lost or until they evolve to live without the mitochondrial genomes.

We were also curious about the process of decoding information on maxicircles to make the proteins. Each key targets and fixes a certain region on the maxicircle-encoded instruction. Comparing the keys in different trypanosome parasites showed that groups of keys are located in the same areas. The groups of keys are spaced at semi-regular intervals, so the decoding proceeds in a cascade with fixed steps. We proposed that the size of the keys and the intervals were determined by other proteins participating in the decoding procedure.

iii. Acknowledgments

It would be unfair to say that pursuing a doctorate is an unexpected journey, but it does seem to end more abruptly compared to how it started. Much do I love Edinburgh now; the transition was not necessarily a smooth one. I remember the first day stepping down the tram into the hustle and bustle of Prince's Street in 2019 to start my Master's, after four years in the quiet Maine countryside. I wondered for a second or two if I should simply catch the next flight away.

Striving through the master's course, I spent just as many hours in the library writing post-grad applications to work my way back to Maine and my old friends. Needless to say, a great part of that came from a coping mechanism for the departure for which I was not mentally prepared. My parents, being gentle and understanding, would indulge me in reflecting on my college years as I revisited memories of pier jumping, bird watching, and the many stimulating conversations I had with friends by the fireplace. However, they also reminded me that nostalgia would not buy me a ticket to the past. Knowing they were right, I tried to open up to the new surroundings and the opportunities it offered.

In February 2020, when I chose Edinburgh over the US and decided to start the PhD at Schnaufer lab, I eventually drew a proper period over the old chapters to embrace changes and a new era of my life. Time flies, and I have made Edinburgh yet another 'second home'. Besides the degree itself, the city has introduced me to myriad adventures, from spinning fire at Samhain to singing acapella at folk music pubs. I have also met people from various backgrounds, and I will start by addressing a few key figures through this time.

Parasitology was a new topic to me. I would like to express my heartfelt gratitude to my supervisors, Achim and Nick, for their unwavering support and guidance throughout my PhD journey. Your expertise, encouragement, and constructive feedback have been invaluable in shaping my research and academic growth. Thank you for your patience and for providing a nurturing environment that allowed me to explore new ideas and develop my skills. I am deeply appreciative of the time and effort you invested, and this work would not have been possible without your mentorship.

Before I joined Schnaufer lab, one of the old members Lizzie gave me a tour of the building and a comprehensive overview of the work involved. The hard drives and codes she left have been a lifesaver for many occasions in the first two years. Unexpectedly, the PhD started quietly during COVID. After the Lockdown, I finally met other members of the lab in person and learned about the tradition of Secret Santa, Christmas dinner, and the birthday cake rota. I would like to say thank you to Laurine, who has been supportive and kind and always got the answer to the confusing paperwork and online systems such as People and Money and Diversity Travel. Also thank you for smoothening my adjustment to working at Ashworth and introducing me to the wonderful people sharing the building.

I also received other help and guidance from people outside the lab group. I would like to thank EASTBIO for funding my study. I am equally grateful to our collaborator Fre from ITM for providing the data essential to all the analysis that you will soon read about. I would like to thank Dr Alasdair Ivens. Besides providing the job opportunity as a demonstrator for his

course, Al also helped me out with quite a few technical issues when I did not know where to start. Another person I would like to thank is Dr Daniel Barker, for discussing the idea of building a phylogeny using discrete traits like the minicircle population.

Through the pursuit of my degree, my mom and dad have always been patient and sensible, besides looking after the plants I left at home. I am grateful that they have faith in me but also know my weaknesses in handling bureaucracy and paperwork (hence the need for a bit of push when it comes to that). As in 2019, they listened to my thoughts, especially when I had doubts. When we moved on to less heavy topics, they seemed to enjoy teasing me by sending pictures of the food they made, the sunset from the coast, and the dozens of different species of palm trees around the reservoir near home. Sometimes they would send photos of my paintings in the living room. We were not great at saying we missed each other, but something was obvious enough and went without saying.

I would not have enjoyed the four years without the many new friends I made here, and they were key to my sanity as I approached the finish line. When I literally became a recluse in the library basement in the last few months, Douglas, who also helped me rescue a red currant tree, and Errikos dragged me to a lovely concert at St. Giles and reminded me of the nice things in life. Many friends from the squash club had passed their PhD before me, and their experiences and advice helped me to steer through the treacherous water. I am particularly grateful for the encouragement from Dr Fiona, who also checked on me from time to time in case I got into a mental crash. Thankfully, since we went through the timeline together a few times in advance, the hypothetical crisis never happened. Instead, she received photos of my orchid blossoms when I genuinely needed to think about something else.

Before I tied myself to the library desktops, I walked the West Highland Way with Rhys in September. The weather was blessed, and it was a much-needed break to prepare myself for the final challenge. I am thankful for the bantering and debates, which reinvigorated me along with the majestic view, and Rhys who managed to stand ‘Z for being Z and talking about moss and sundew and fungi and bugs every other hour’. I was also glad he taught me how to make a fire, although he said he was disappointed in me for refusing to blow into the fire but building a simple stone stove to fan into instead.

It is the beginning of winter when I submit this piece, so forgive me if I did not write the acknowledgment with a more cheerful tone. I know it is not the end but the beginning of many new adventures. Hopefully, I will continue to be amazed by the views on the way and find friendship in the people I meet.

------10/31/2024 Main Library

iv. Key abbreviations

AAT: Animal African Trypanosomiasis

BSF: bloodstream form

CSB: Conserved Sequence Blocks

EMF: epimastigote form

ETC: electron transport chain

HAT: Human African Trypanosomiasis

ISSP: initiation sequence starting position

kDNA: kinetoplast DNA

KFZ: kinetoflagellar zone

PF: procyclic form

PMC: percentage of mapped CSB-3-containing reads

KREN: kinetoplastid RNA editing endonuclease

KREP: kinetoplastid RNA editing protein

KREX: kinetoplastid RNA editing exonuclease

MCN: minicircle copy number

RECC: RNA editing core complex

RESC: RNA editing substrate binding complex

SID: sequence identity

SRA gene: serum resistance-associated gene

UMS: universal minicircle sequence, or CSB-3

UMSBP: universal minicircle sequence binding protein

II. Content

i. Table of content

ii. Table of figures

iii. Table of tables

1 Introduction

1.1 The classification of kinetoplastids and trypanosomatids

Euglenozoa contains members of four classes: Diplonemea, Euglenida, Kinetoplastea, and Symbiontida [1]. A sister clade to Diplonemea, Kinetoplastea includes free-living and parasitic protozoa known for their extraordinarily massive and complex mitochondrial genome, the kinetoplast. Kinetoplast DNA is referred to as kDNA. Unlike the other three more obscure clades, Kinetoplastea contains parasitic taxa that have drawn researchers' attention because of the public health and economic challenges they imposed globally, particularly in developing countries [2-4].

Assuming no reversions to a free-living state, parasitism has evolved independently on at least four occasions within kinetoplastids, one of which gave rise to trypanosomatids of the order Trypanosomatida [5]. Trypanosomatida includes monoxenous or dixenous obligate endoparasites of arthropods, plants, leeches, vertebrates, and ciliates [6-12]. At the time of writing, evidence suggests that trypanosomatids have emerged from bodonids as a sister clade of free-living eubodonids [13, 14]. The most well-studied eubodonid, *Bodo saltans*, is shown to be closely related to trypanosomatids [13, 14].

Trypanosomatida contains a single family Trypanosomatidae, which includes three dixenous genera: *Leishmania*, *Phytomonas*, and *Trypanosoma*, eleven monoxenous genera including *Crithidia*, and three genera (that form the subfamily Strigomonadinae) characterized by the presence of endosymbiotic bacteria [15, 16]. Dixenous mammalian parasites are now thought to have evolved from ancestral insect parasites after the vertebrate land invasion no sooner than 380 million years ago (mya) [17]. The evolution of dixenous lifecycles that include development in mammalian hosts has occurred multiple times, and the nowadays successful mammalian parasites, *Trypanosoma* and *Leishmania*, probably represent the survivors of a series of bottlenecks [5].

The insect-first hypothesis also suggests new mammalian pathogens may emerge from the extant insect parasites [5]. The experimental infestation of mouse epidermal fibroblast by *Crithidia* and *Herpetomonas* species exemplifies the potential of an insect parasite to adapt to a foreign biochemical environment, although no evidence has suggested their ability to evade the mammalian immune system [5, 18].

Most studies on Trypanosomatids are centered around the dixenous genus *Trypanosoma* and *Leishmania* due to the medical and veterinary challenges associated with their endemics. Despite their vast diversity and distribution, the monoxenous trypanosomatids receive less attention [15]. Nevertheless, the monoxenous *C. fasciculata* has been a model species in studying kinetoplast replication and segregation, especially in the 20th century [19-25]. The kinetoplast of *C. fasciculata* is now widely used in physics research to study polymers with planar structures, for example, polycatenanes and Olympic gels [26-28]. Fragments of the *C. fasciculata* kinetoplast have been assembled and annotated [29-31]. Our collaboration with the Michieletto lab contributed to the deep sequencing and annotation of the *C. fasciculata* mitochondrial genome [32].

1.2 Life history of trypanosomatids

Dixenous trypanosomatids complete cyclical developments through insect vectors and mammalian hosts and undergo morphological and metabolic changes to adapt to distinct biochemical environments [33]. Subspecies of *T. brucei* and *T. congolense* navigate the host-vector transitions through distinct paths [34].

1.2.1 Life history of *T. brucei*

The development of the exclusively extracellular *T. brucei* has been well-charted (Figure 1-1). *T. brucei* proliferates asexually in a highly mobile slender form by mitosis in the mammalian host. The slender form *T. brucei* satisfies the energy need via glycolysis of bloodstream glucose and does not maintain a functional electron transport chain. The parasites lack oxidative phosphorylation activities and have highly reduced tubular mitochondria without cristae [33]. Without the electron transport chain, to maintain the vital mitochondrial membrane potential, bloodstream form (BSF) *T. brucei* employs the ATP synthase complex in a manner opposite to that in the insect stage and in most other eukaryotes, namely conducting ATP hydrolysis to maintain the mitochondrial membrane potential by proton pumping, instead of ATP production [35].

Each wave of parasitaemia stimulates the differentiation of the slender form *T. brucei* into a stumpy form, cued by quorum sensing [34]. Arrested in G1/G0 phase, the less mobile stumpy form parasites prepare for infecting the insect vector through a series of adaptations, including the remodeling of a cristate and more activated mitochondrion capable of α-ketoglutarate metabolism. The stumpy form parasites enter the tsetse fly midgut during the bloodmeal and differentiate into the proliferative procyclic forms to establish an infective reservoir [36].

After 1-2 weeks, the parasites migrate to the anterior of the fly and differentiate into asymmetrically dividing epimastigotes in proventriculus [36]. Colonization of the salivary gland, where sexual reproduction of *T. brucei* occurs, involves the attachment to the salivary gland epithelium by epimastigotes and proliferation, leading to the infective metacyclic trypomastigotes [36]. As the fly feeds and injects saliva, the metacyclic trypomastigotes enter the biting sites to initiate a new round of infection in the mammalian host.

Unlike in the mammalian bloodstream, in the tsetse vector *T. brucei* employs a fully functional electron transport chain (ETC) supported by a more branched and cristate mitochondrion [33, 37, 38]. Without the readily available bloodstream glucose, *T. brucei* instead relies on proline as the major source of energy, downregulating the high-affinity hexose transporters and the glycolytic enzymes and upregulating enzymes that partake in proline uptake and catabolism [39, 40]. Despite the presence of all citric acid cycle enzymes, *T. brucei* does not complete the citric cycle but uses part of the pathway for purposes other than energy production, such as fatty acid biosynthesis, degradation of proline and glutamate to succinate, and the generation of malate [37]. ETC activities lead to the elevation of reactive oxygen species levels which appear to act as signalling molecules driving development progression [40].

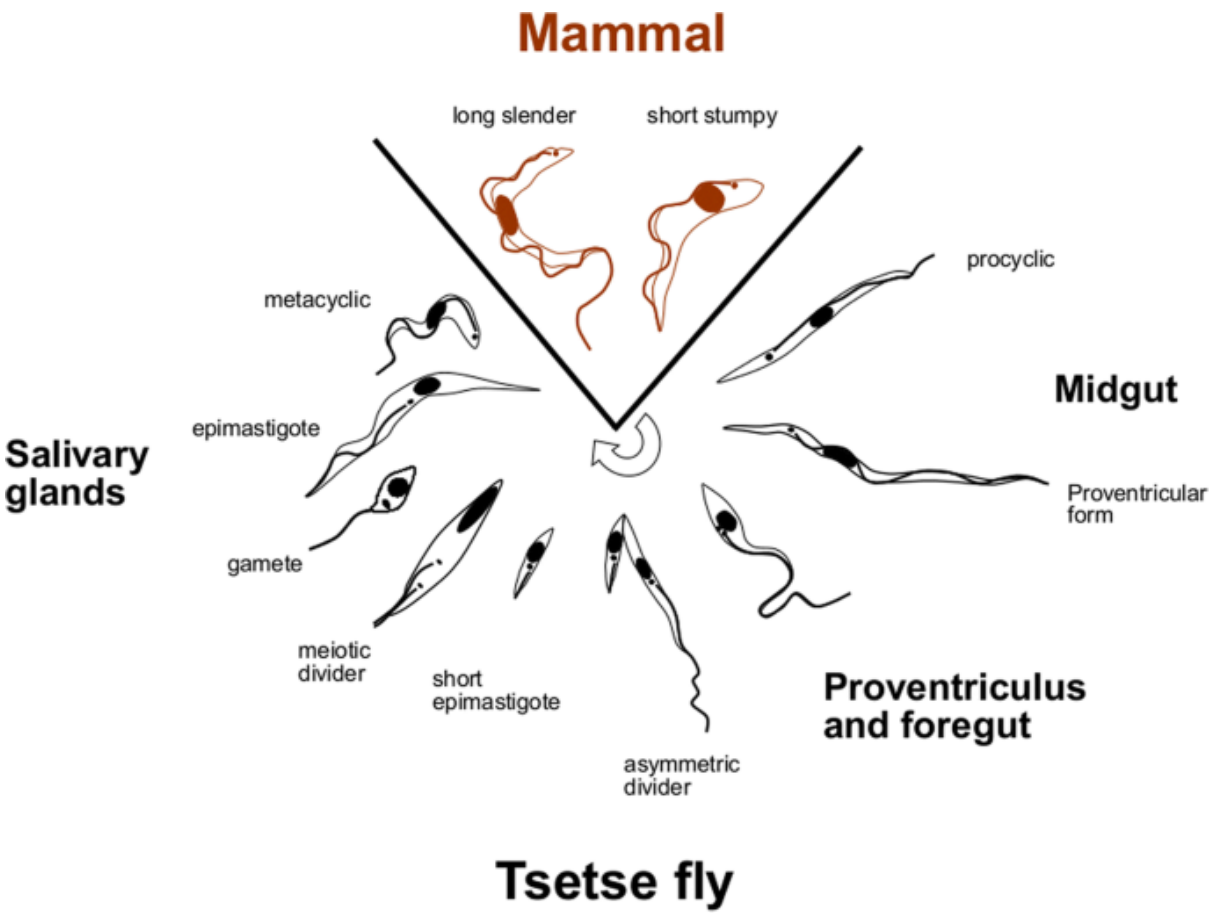


Figure 1-1. **Diagram of the life-cycle of *T. brucei*, including sexual stages (taken from [41])**

1.2.2 Life history of *T. congolense*

Compared to *T. brucei*, the lifecycle of *T. congolense* remains less well-documented, as it mainly afflicts animals and historically did not enjoy priority in a field focusing on human diseases. Although *T. congolense* colocalizes with *T. brucei* in the tsetse belt and is transmitted by tsetse fly, it exhibits substantial differences from *T. brucei* in its cyclical development. Without activating the ETC, BSF *T. congolense* also relies on glycolysis of bloodstream glucose yet generates mainly succinate, acetate, and malate instead of pyruvate as the end product [42]. *T. congolense* also experiences density-dependent cell cycle arrest in the mammalian bloodstream. Without observable morphological differentiations, *T. congolense* preadapted to the insect vectors exhibits a unique gene expression profile that suggests regulations distinct from *T. brucei* [34, 42, 43].

After infecting tsetse flies through a blood meal, the BSF *T. congolense* develops into procyclic form within the midgut, proliferates, and elongates at the anterior end of the cell [44]. Colonization of proventriculus by the slender trypomastigotes occurs as early as six days after infection [44]. The proventricular trypomastigotes migrate through the peritrophic matrix to the foregut lumen and produce epimastigotes without going through the asymmetrical division recorded in *T. brucei* [44].

Unlike the uniform epimastigotes attached to the salivary gland in *T. brucei*, the *T. congolense* epimastigotes attached to the proboscis lining are morphologically variable. The elongated epimastigotes proliferate and subsequently develop into infective metacyclics

with considerable reduction in length. *T. congolense* sexual reproduction occurs exclusively in the proboscis [45] [46].

The enlargement of mitochondrion from BSF to procyclic and metacyclic manifests the increase in mitochondrial activity in insect-stage parasites [47]. Meanwhile, the metacyclic trypomastigotes preadapted to mammalian bloodstream fulfill energy requirements primarily via glycolysis and produce mainly trypanosome alternative oxidase (TAO) at the endpoint [47].

1.2.3 Sexual reproduction

Evidence of sexual recombination in *T. brucei* dates back to 1980 [48]. As a non-obligatory feature of cyclical development, sexual reproduction of *T. brucei* occurs in the tsetse salivary gland [49, 50]. In contrast, evidence of hybrids suggests that *T. congolense* and *T. simiae* reproduce sexually exclusively in the proboscis [45] [46]. Analysis of naturally occurring hybrids among field isolates and experimental genetic exchange in the laboratory have shed light on the mechanism and epidemiology of sexual recombination [51].

The strong immune response to the first strain in the tsetse fly hinders the establishment of the second strain via a later blood meal [52]. Consequently, although sequential bloodmeals from animals infested with different strains potentially lead to coinfection in the vector and provide an opportunity for genetic exchange by sexual reproduction, sexual recombination between strains mostly likely occurs when the fly bites a coinfected mammal [52]. Another requirement is that both strains reach and colonize the same salivary gland, which does not always occur when the fly takes a blood meal that contains two strains [53, 54]. However, contrary to previous findings that recombination requires two different strains, evidence of intraclonal hybrids has also been reported for *T. brucei* and suggests genome reshuffling via tsetse transmission without coinfection [55].

Overall, the inheritance of genetic markers in hybrid progeny from crosses of *T. brucei* is consistent with Mendelian genetics [51, 56]. After the observation of hybrids, the production of gametes had been speculated and the expression of meiosis-specific protein had been detected [52, 57]. Gametes have not been observed until recently. The fusion of promastigote-like cells that interact with each other via their flagella has been visualized with cytoplasmic fluorescent proteins [58]. The observation of intermediate cells with unique nuclei and kinetoplast count led to the proposal of an asymmetrical meiotic division model [59].

A non-obligatory part of the *Leishmania* lifecycle as well, sexual reproduction occurs in the midgut of sandflies [60-62]. Despite hybrids having been generated experimentally and detected in the field, no haploid gametes have been identified. Evidence of inter- [63, 64] [65, 66] and intra-specific [67, 68] hybridization within *Leishmania* has been detected over the years, raising concerns about the consequences of these transfers on phenotypes such as drug resistance or pathogenicity.

Besides the recombination of the nuclear genome, sexual reproduction also results in the mixing of the mitochondrial genome, suggesting that both mitochondrial and cell fusion

occur. Experimental crosses of *T. brucei* and *Leishmania spp.* result in heterogeneous networks in hybrids [25, 69-73]. Early-stage hybrids exhibit a mixed maxicircle population [71]. It has been proposed that the initially heterogeneous maxicircles become homogeneous after several generations of random partition between daughter cells during mitotic division, giving the impression of a superficially uniparental inheritance [70, 71]. In contrast, the kDNA of the progeny contains minicircles from both parents and thus maintains a biparental inheritance of minicircles [69, 70, 73].

Sexual reproduction facilitates the circulation and expansion of harmful traits such as drug and human serum resistance and generates novel human infective strains [74]. Hybrids with recombined VSG and minicircle repertoires have been detected in *T. brucei* in the field, including hybrids between human infective and non-human infective trypanosomes in the same geographic areas [69]. Animal reservoirs allow such hybridizations to occur [4]. In Uganda, the human pathogen *T. b. rhodesiense* and the animal pathogen *T. b. brucei* form a single breeding population, resulting in the emergence of novel *T. b. rhodesiense* from *T. b. brucei* with different genetic backgrounds, probably via spreading the VSG-derived SRA gene into previously non-human-infective lineages [4].

1.2.4 Clonal groups of *T. brucei*

1.2.4.1 *T. b. gambiense type 1*

The underlying reasons for the loss of sexual reproduction in certain groups of *T. brucei* remain unclear. Sub-Saharan *T. brucei* subspecies capable of sexual reproduction are genetically interconnected due to hybridizations, raising concerns about the emergence of novel HAT strains from the broader reservoir of animal pathogens [74-76]. Conversely, microsatellite and multilocus genotype analyses indicate that Central African *T. b. gambiense* type 1 is monophyletic with minimal gene flow between populations and being strictly clonal reproduction [77-79]. Nevertheless, the expression of meiosis-specific genes has been detected in *T. b. gambiense* type 1, which raises the question what factors are involved in the genetic isolation of the subspecies [58].

The asexual reproduction isolates the nuclear and mitochondrial genomes of *T. b. gambiense* type 1, resulting in a highly unique kDNA composition. A recent study has reported minicircle diversity of 89 to 122 classes per network, while the same minicircle classes were not detected in other *T. brucei* subspecies [80]. The minicircle diversity is much lower than the 399 unique classes detected in *T. b. brucei* EATRO1125, capable of sexual reproduction [81].

1.2.4.2 *T. b. equiperdum and T. b. evansi*

Both *T. b. equiperdum* and *T. b. evansi* have generally lost the ability to transform from slender to stumpy form, hence termed monomorphic. However, there are scarce reports of stumpy forms in *T. b. evansi* [82]. The monomorphic parasites may have evolved from tsetse-transmissible *T. b. brucei* by switching to direct transmission between mammals [83-85]. *T. b. equiperdum* is transmitted between horses during copulation, while *T. b. evansi* is transmitted mechanically via biting insects and vampire bats [83, 86]. Freedom from tsetse

transmission allows the monomorphic strains to expand their territory beyond the tsetse belt.

The kDNA of three monophyletic groups, *T. b. equiperdum* type C, *T. b. evansi* type A, and *T. b. evansi* type B, contains a homogeneous minicircle population incapable of editing mRNAs required for either BSF or PCF parasites [87, 88]. Surprisingly, functional editing complexes have also been detected in *T. brucei* presumably without editing activities [89].

Genetic studies suggest at least four independent events that have led to the *T. b. equiperdum* and *T. b. evansi* groups [90-93]. Old and New World *T. b. equiperdum* (type C/BoTat, type OVI) is closely related to the Kiboko/Sindo *T. b. brucei* in Eastern Africa, whereas *T. b. evansi* type A and type B show an ancestry within Western African [90, 92]. The multiple origins of monomorphic strains imply that the tsetse-transmissible pleomorphic *T. brucei* may serve as a reservoir for the emergence of novel tsetse-independent cell lines.

It has been reported that *T. b. equiperdum* and *T. b. evansi* have evolved kDNA independence and undergone partial or complete loss of the organellar genome, hence becoming dyskinetoplastic or akinetoplastic respectively [94-96]. A single amino acid mutation of the nuclearly-encoded F_1F_O-ATP synthase subunit γ compensates for the kDNA loss with increased reliance on the ADP/ATP carrier to maintain the mitochondrial membrane potential [35, 95, 96]. The mutation is not detected in *T. b. equiperdum* type OVI, whose minicircle diversity remains uncharacterized [91].

Historically, the classification of the monomorphic asexual trypanosomes prioritizes their pathological and epidemiological features with little emphasis on evolutionary genetics [83]. Based on emerging evidence, a more updated classification would view each *T. b. evansi* or *T. b .equiperdum* group as a distinct lineage of *T. brucei*. In this thesis, as an expediency, we adopt the view that both are subspecies of *T. brucei* and address them as *T. b. equiperdum* and *T. b. evansi* [90, 91]. Admittedly, the polyphyletic origins are inconsistent with a subspecies classification and make both groups polyphyletic and sometimes paraphyletic [90, 91]. Nevertheless, we investigated all groups, i.e. type A, B, C, and OVI, without assuming their relatedness with each other, so retaining the conventional nomenclature did not compromise the integrity of this work.

1.3 Diseases caused by trypanosomatids, and their treatments

1.3.1 Human diseases caused by trypanosomatids

Some members of the dixenous genera *Trypanosoma* and *Leishmania* are causative agents of human diseases classified as Neglected Tropical Diseases (NTDs) by the World Health Organization (WHO), including African sleeping sickness or Human African Trypanosomiasis (HAT), Chagas disease, and leishmaniasis. NTDs are prevalent in impoverished tropical areas and cause substantial social, economic, and public health burdens in developing countries [97]. Despite the lack of attention to NTD distributions among different socioeconomic groups, multiple studies suggest that NTD prevalence is biased toward disadvantaged populations [98].

While the living conditions associated with poverty facilitate transmission, the cost and unavailability of preventive and curative schemes increase the infection rate in the poor population. The decreased working time and increased medical expenses exacerbate the economic burden on the disadvantaged groups, creating greater inequality of NTD distributions across the socioeconomic gradients [98]. On the other hand, horizontal control strategies reduce vector prevalence, raise NTD awareness in all strata, and reduce inequalities [98, 99]. In the following sections, we will summarize the three main trypanosomatid-associated NTDs, caused by *T. brucei*, *T. cruzi*, and *Leishmania*.

1.3.1.1 *Trypanosoma brucei*

Subspecies of *T. brucei* cause HAT in sub-Saharan Africa with different symptoms. Chronic HAT in Central and West Africa is caused by *T. b. gambiense* type 1, and acute HAT in East Africa is caused by *T. b. rhodesiense* [100, 101]. Both subspecies have evolved to expand the host range by acquiring resistance to the trypanosome lytic factors (TLFs) in human serum [102, 103]. Nevertheless, some Western African human populations have regained resistance to *T. b. rhodesiense* at the cost of a higher risk of kidney sclerosis [102, 104].

As an extracellular blood parasite, *T. brucei* invades intravascular and extravascular spaces. Nevertheless, the parasite also propagates in the adipose tissue, which probably contributes to the characteristic progressive wasting of its hosts and infections are typically lethal if left untreated [105]. Although *T. brucei* mainly relies on tsetse flies (*Glossina. spp*) for transmission, congenital and sexual transmission of *T. b. gambiense* type 1 has also been reported [106].

The disease initiates in the cutaneous stage as the parasites invade the tsetse fly bite site and form ulcerating nodules known as trypanosomal chancre. Complex parasite-host interactions occur between the parasites and the immune cells in the skin as the infection establishes [107]. The infection may develop over months (*T. b. gambiense*) or weeks (*T. b. rhodesiense*) in the hemo-lymphatic stage manifested by intermittent fever, headaches, rigors, muscle and joint pain, and transient facial swelling as the parasites proliferate in the patient's body fluid [108]. The central nervous system (CNS) stage occurs as the parasites penetrate the blood-brain barrier and proliferate in the brain tissue [108]. In the CNS stage,

T. brucei alters neuronal function, leading to the characteristic disruption of the 24-hour sleep-wake cycle that gives HAT its common name African Sleeping Sickness [108].

The history of HAT in 19th and 20th century Africa mirrors the complicated colonial history and political turbulence [109]. Exacerbated by colonial exploitation and population movements, several severe HAT epidemics swept through sub-Saharan Africa, particularly in the Congo Basin and Uganda [110]. By the mid-20th century, coordinated campaigns by colonial governments and later independent African states successfully reduced the incidence of the disease in many regions. However, political instability, economic decline, and the breakdown of health services in many African countries during the 1970s and 1980s led to a resurgence of trypanosomiasis, particularly in war-torn regions where control measures had collapsed [109].

In response to the resurgence, renewed efforts by organizations including WHO and African governments have focused on enhanced surveillance, improved treatment options, and vector control [101]. These efforts have led to a significant decline in reported cases, from over 35000 cases annually in the late 90s to 663 in 2022, with Togo and Cote d'Ivoire being the first countries to be validated for achieving elimination of HAT as a public health problem at the national level [101]. The powerful advancements in public health motivate the WHO to chart the new ambitious targets for NTDs in 2021-2030. However, little is known about the current state of asymptomatic human carriers and animal reservoirs, which may present unexpected challenges to HAT elimination [100, 106, 111].

1.3.1.2 *Trypanosoma cruzi*

The intracellular parasite *Trypanosoma cruzi* causes endemics of Chagas disease in 21 Latin American countries. Commonly transmitted by members of the subfamily Triatominae known as kissing bugs, Chagas disease can also be transmitted orally through contaminated food, drinks, or meat [112, 113], sexually [114], or congenitally [115]. The symptoms of Chagas may be deceptive during the onset acute phase, including flu-like symptoms, mild inflammatory responses, diarrhea, and vomiting. The symptoms may disappear for years before the parasites invade the cardiac tissue in the chronic phase manifested in dilated cardiomyopathy, thromboembolic phenomena, and arrhythmias that may culminate in heart failure [116]. Nowadays, treatment for parasitic infection remains limited to benznidazole and nifurtimox [117]. Outbreaks of Chagas are frequently associated with low socioeconomic positions, which necessitates the enforcement of the One Health approach.

1.3.1.3 *Leishmania*

At least 20 species in the genus *Leishmania* are transmitted by the bite of infected female Phlebotominae sandflies and cause the vector-borne diseases leishmaniases [118]. Leishmaniases are endemic in large areas of the tropics, subtropics, and the Mediterranean basin, including more than 98 countries globally. The disease may manifest in four forms according to the different species and the location of the parasite in the mammalian tissues: visceral, cutaneous, diffuse cutaneous, and mucocutaneous. The most widespread and

generally self-healing cutaneous forms (CL) cause skin lesions and ulcers on exposed parts of the body [119]. The visceral form (VL), also known as kala-azar, results from the infection of phagocytes within the reticuloendothelial system [119]. Manifested in fever, weight loss, anemia, and spleen and liver enlargement, VL can be lethal if left untreated and is less common than CL [120].

VL can be caused by different species. *Leishmania donovani* (in regions of India, Pakistan, China, and Africa) and *Leishmania infantum* (in the Mediterranean region) cause VL in the Old World. Considered an agent of CL, *L. tropica* has caused cases of VL in the Middle East. In the New World, *L. infantum* primarily causes VL in Brazil [119].

Furthermore, coinfection with Leishmania and HIV imposes severe public health hazards in co-endemic regions. Both HIV and *Leishmania* proliferate within immune cells, which leads to reciprocal modulation of disease pathogenesis and reactivation of leishmaniasis in immunocompromised patients [121]. Assuming 10% fatality, it is estimated that 20,000 to 40,000 leishmaniasis deaths occur per year [120].

1.3.1.4 *Atypical trypanosome infections*

Humans are resistant to most trypanosome parasites due to the TLFs in serum. However, rare cases of atypical trypanosome infection have been reported [122-124]. Most infections are transient or treated, although at least two have been fatal [123]. In 2022, among the 21 human cases from the published information, 10 (47%) were due to *T. lewisi*, followed by 5 (24%) cases of *T. b. evansi*, 4 (19%) cases of *T. b. brucei* and 1 (5%) case each of *T. vivax* and *T. congolense* [125]. The authors suggest that the increasing reports of animal trypanosome species in humans present an alarming perspective of host range expansion and probable acquisition of novel TLF resistance. Alternatively, improved medical facilities and surveillance may also explain the increased reporting from developing countries, while some cases could be attributed to TLF deficiency in compromised immune systems.

1.3.2 Available treatments for human diseases

1.3.2.1 *Human African Trypanosomiasis*

The treatment of human trypanosomiasis depends on the type of parasite and the stage of infection. Haemo-lymphatic stage (first-stage) is defined as ≤ 5 white blood cell (WBC)/µL and no trypanosomes in cerebrospinal fluid (CSF), and meningo-encephalitic stage (second-stage) is defined as > 5 WBC/µL or trypanosomes in CSF [126]. In chronic *T. b. gambiense* infection, severe meningo-encephalitic stage (severe second-stage) occurs with CSF WBC ≥ 100/µL with or without trypanosomes in CSF [126].

Treatment for first-stage *T. b. gambiense* infection often involves pentamidine administered by intramuscular injection [126] [127]. An aromatic diamidine first introduced in the 1940s, pentamidine exerts its trypanocidal effects via complex and poorly understood mechanisms. The trypanocidal effects of pentamidine are probably associated with kDNA structure disruption, group I intron catalytic activity inhibition via RNA binding, inhibition of a plasma-

membrane Ca^{2+} ATPase, and disruption of mitochondrial membrane potential [127]. Despite the unexplained failure of *T. b. rhodesiense* treatment, no field resistance to pentamidine has been reported for *T. b. gambiense.*

Introduced in 1920, suramin (Germanin) remains the first-choice treatment for early-stage rhodesiense HAT in children < 6 years or body weight < 20 kg [126, 128]. Highly soluble in water and given intravenously by injection, suramin probably acts as a promiscuous inhibitor for various enzymes and receptors, which may explain the scarcity of high-level resistance [127]. Nevertheless, the severe and sometimes fatal side effects of suramin made fexinidazole a better alternative when introduced in 2010 [129].

Eflornithine in combination with nifurtimox (NECT) is commonly used to treat late-stage *T. b. gambiense* infection, with documented cure rates of 95–98% and fatality rates of < 1% [126]. Eflornithine's mechanism of action is based on irreversible inhibition of ornithine decarboxylase (ODC), while the mechanism of nifurtimox remains to be elucidated [127]. As a drug combination, NECT also discourages the parasite from developing drug resistance.

Introduced in 1949, melarsoprol remained the sole effective treatment of second-stage rhodesiense HAT until approval of fexinidazole [126] [128]. Despite the efficacy of the arsenical in killing the parasites via inhibition of several important metabolic and transport functions, melarsoprol is notorious for its side effects, including serious and often fatal reactive encephalopathy in 5–10% of cases [127].

Being the first oral trypanosomiasis treatment and a breakthrough in combating these NTDs, fexinidazole frees medical staff from intravenous or intramuscular injections, so lowers the risk of catheter or needle-related infection. It exerts trypanocidal effects via the bioreductive activation by the parasite's nitroreductase enzymes to produce sulfoxide and sulfone metabolites [126]. Taken as prescribed for gambiense HAT, fexinidazole shows equivalent efficacy to pentamidine in first-stage and to NECT in second-stage, yet an inferior efficacy to NECT in severe second-stage (CSF WBC ≥ 100/µL) [130, 131]. For rhodesiense, fexinidazole performs as well as suramin in first-stage and better than melarsoprol in second-stage. Validated by phase I-III trials in 2018, fexinidazole is now available to the National Control Programmes of HAT-endemic countries as a free treatment [132] [133]. Nevertheless, the selectable reciprocal cross-resistance to fexinidazole and nifurtimox indicates a need for close resistance monitoring and continuous investment in drug discovery and development [134].

1.3.2.2 *Chagas disease*

The current drugs for Chagas disease are benznidazole and nifurtimox, both carrying severe side effects [117]. The drugs are effective in the early stage, but the efficacy declines as the disease develops into a later stage, resulting in difficulty curing chronic Chagas [135]. The ongoing trials for alternative treatments have not yet yielded conclusive positive results [117]. Hence, improving social participation and awareness to achieve a more prompt diagnosis is key for early intervention and treatment of Chagas [136].

1.3.2.3 *Leishmaniasis*

Although CL is generally self-healing, medications, including pentamidine, may help to prevent permanent cosmetic damages [137]. Treating the lethal VL is complicated by accessibility, adverse effects, and cost [138]. Most treatments do not support oral preparation. Patients prescribed the lengthy treatment of sodium stibogluconate (SSG) also suffer from liver and cardiac toxicities. The lengthy renal toxic Amphotericin B treatment incurs infusion reactions, while the quicker liposomal amphotericin B is expensive. The cheap alternative paromomycin exhibits toxicity to the auditory system and liver. The oral treatments miltefosine and pentamidine have lower efficacy against HIV co-infection and various toxicities [138].

Clinical resistance to mainstream drugs such as SSG and paromomycin presents further challenges. Despite recent advances in antileishmanial drug discovery and the promising aspects of vaccination, many issues associated with the available treatments remain unresolved [139] [140]. The high mortality and the advent of resistance to current agents necessitate continuous research to combat leishmaniases [141].

1.3.3 Human serum resistance and VSG in the African trypanosome

Presumably similar to the ancestral trypanosomatids, *Leishmania* and Stercorarian trypanosomes have heterogeneous cell surfaces covered with various antigens encoded on multicopy gene families [142]. On the contrary, African trypanosome parasites, including *T. brucei*, *T. congolense*, and *T. vivax*, have evolved a system of monoallelic expression and coat the cell surfaces predominantly with one type of the highly diverse variant surface glycoprotein (VSG) [142]. Trypanosomes express distinct VSGs at different times to 'switch' their antigenic signatures [143]. The mosaic antigenic demography leads to the characteristic cycle of parasitemia, where the clearance of the parasites coated with dominant VSG by the host's immune response is followed by the growth of immune-escaped parasites that express different VSGs. This system makes complete clearance without intervention almost impossible.

The strong selection in favor of novel VSGs drives its diversification and gene conversion, culminating in the immense repertoire of thousands of VSG genes and pseudogenes [144]. For thirty years since the first description of VSG structure, all VSGs have been assumed to be homodimers [145, 146]. However, some VSG monomers are now shown to self-oligomerize when constrained at high density to the cell surface, hence assuming distinct states during trafficking and anchoring [147]. To further complicate the story, post-translational O-linked glycosylation at the top of the VSGs enhances the parasites' ability to evade the immune response [148]. Consequently, designing a structure-based classification scheme for the VSG proteins is highly challenging [149].

T. b. rhodesiense and *T. b. gambiense* have overcome the effect of TLFs and become human infective via repurposed VSGs. Co-transcribed with the expressed VSGs, the serum resistance-associated (SRA) gene is a truncated VSG that confers resistance to human serum

in *T. b. rhodesiense* [150, 151]. Localizing to the endosomal network [152], SRA binds to the trypanolytic toxin in TLFs, APOL-1, and prevents it from lysing trypanosomes [153, 154]. The TLF neutralizing mechanism in *T. b. gambiense* involves a GPI-anchored b-VSG protein TgsGP [155, 156]. TgsGP hinders APOL-1-mediated cell lysis by inducing membrane stiffening, buying time for APOL-1 degradation. Besides, VSGs also play roles in drug resistance, immune modulation, iron transport, and transmission [144]. The seemingly unbounded evolution potential of VSGs and their importance to host-parasite interaction have kindled research interest since its discovery.

1.3.4 Animal diseases caused by trypanosomatids

Members of the genus *Trypanosoma* infect many animals of economic importance, such as insect pollinators, rabbits, dogs, horses, camels, cattle, and small ruminants [3, 8, 157-160]. Leishmaniases seriously affect dogs, while in cattle and horses, the impact is limited [161, 162].

Trypanosoma parasites have also been detected in endangered marsupial and petropine species [7, 163]. The zoonotic New World *Trypanosoma cruzi* infests humans and domestic and wild mammals [84, 163], indicating the risk of animal trypanosomiasis as a repertoire for human diseases [46, 74, 111].

1.3.4.1 *Animal African Trypanosomiasis*

A devastating disease for domestic animals within the tsetse belt in sub-Saharan Africa, Animal African Trypanosomiasis (AAT), or nagana, is mainly caused by *Trypanosoma vivax, T. b. brucei*, *T. congolense*, and *Trypanosoma simiae*. Endemic to 36 countries, AAT infests about 10 million km^2 of African landmass [164, 165]. In Nigeria alone, over 78% of cattle are located in areas overlapping the territories of various *Glossina* species [166]. Overall, AAT threatens 48%, 76%, 28%, and 8% of the total cattle population in Western, Central, Eastern, and Southern Africa and results in up to 4.75 billion USD loss of GDP per year [167, 168].

Nearly 309 million livestock keepers live below $2 per day in sub-Saharan Africa, where livestock production accounts for 40% of total household income [169]. The areas with high cattle density along the margins of the tsetse belt have suffered substantial economic loss to AAT and will obtain the most rewards from effective control [168, 170]. The high mortality of livestock and the expense of AAT prevention and treatments cast a heavy burden on households in impoverished rural areas. The dependency of farmers on livestock makes AAT an urgent issue. Despite decades of attempts to control AAT, the lack of organized national control in developing countries hampers combating AAT [171, 172].

Generally chronic and fatal if left untreated, AAT results in considerable weight loss and anemia in livestock, along with other symptoms including fever, oedema, adenitis, dermatitis, and nervous disorders [173-176]. In dairy animals, trypanosomiasis significantly reduces milk yield, affecting both the quantity and quality of milk [177].

The evolution of AAT varies according to the infesting species, and susceptibility also varies among livestock [46, 164]. *T. simiae* causes acute AAT and rapid death in improved pig strains [84, 91]. Albeit highly pathogenic for horses and dogs, *T. brucei* infection in cattle is usually asymptomatic. While the West African humpless cattle and the Guinean strain of goats are resistant to *T. congolense* and *T. vivax*, the parasites are highly pathogenic to Zebu cattle[177].

The prevalence of different trypanosome species varies across sub-Saharan Africa. While *T. congolense* predominantly circulates in Eastern and Central Africa, *T. vivax* mainly occurs in Western Africa, and *T. b. brucei* is found through the tsetse belt [46, 178]. Where the species distributions overlap, the cattle frequently suffer from coinfection of mainly *T. congolense*, *T. vivax*, and *T. brucei [178]*. Hence, the distribution of tsetse flies and trypanosome parasites strongly influences the structure of the animal industry in sub-Saharan Africa [3, 177].

1.3.4.2 *Surra and dourine*

Some *T. brucei* subspecies has expanded beyond the tsetse belt, imposing global veterinary challenges. [82]Veterinarians have treated surra and dourine strictly separately, yet this decision has become increasingly questionable with emerging evidence [93].

T. b. evansi is transmitted mechanically via biting insects and vampire bats and causes the disease surra [83, 179]. As mechanical transmission relies on the parasite's survival within the oral cavity of the vector, the transmission rate increases with a shorter interval between blood meals [180]. Between 1906-2017, surra decimated thousands of livestock and caused great economic loss in Africa, Asia, and South America, and rare cases occurred in Europe due to imported animals [181]. A generally mild or asymptomatic infection in cattle, buffalo, sheep, goats, and pigs, surra can cause acute symptoms in camel, horses, and dogs, resulting in fever, weakness, lethargy, anemia, and severe weight loss that culminate in death [182]. Several curative/preventive trypanocides are available for surra, each with advantages and drawbacks depending on the host and evolution of the disease, while other preventive measures, such as vector control and preventive injections, also help to reduce surra instances [179].

Transmitted between horses during copulation, *T. b. equiperdum* no longer requires insect vectors and causes a worldwide equid disease called dourine [83]. Once widespread when horses played important socioeconomic and military roles, dourine is now mostly absent from Western Europe, Australia, and the USA, although sporadic outbreaks have been reported in Italy [84]. Dourine remains endemic in areas of Asia, Africa, Russia, the Middle East, and Eastern Europe [84, 183, 184].

Mainly a tissue instead of a blood parasite, *T. b. equiperdum* causes a characteristic symptom of periodic exacerbation and relapse and usually kills the host. The unavailability of vaccines dictates that the prevention of dourine relies on avoiding natural mating or

artificial fertilization with infected horses or semen from infected stallions, stressing the need for effective screening and diagnosis [184, 185].

Despite often indicative clinical signs of dourine, most serological markers, including ones in the most widely used CFT test, cannot distinguish the morphologically identical *T. b. equiperdum* from *T. b. evansi* and *T. b. brucei* and fail to confirm the diagnosis in co-endemic regions [183]. Before knowledge of their close evolutionary relationship to *T. brucei* and understanding of the kDNA independence adaptations, a prominent distinction between the highly conserved *T. b. equiperdum* and *T. b. evansi* was the presence of a mitochondrial genome component, the maxicircle, in *T. b. equiperdum* [83]. A PCR test for maxicircle-encoded genes has identified two highly virulent yet genetically distinct horse isolates from Venezuela as *T. b. equiperdum* strains [186], while the isolates were initially classified as *T. b. evansi [187]*. Furthermore, 354 homozygous SNPs identified from genome-wide SNP analysis for three monophyletic *T. b. equiperdum* strains may become promising markers for diagnosis [90].

1.4 Cell biology of trypanosomatids

1.4.1 Morphology of trypanosomatids

The corkscrew-like motion of some trypanosomatids gives the obligate parasitic family its name: *trypano* (borer) and *soma* (body). Assuming distinct morphology specific to each developmental stage of their lifecycles, the single-flagellated trypanosomatids exhibit four main forms with various positions of the flagellum relative to the nucleus (Figure 1-2) [188, 189]. In amastigotes, promastigotes, epimastigotes, and the less common choanomastigotes, the flagellum is anterior to the nucleus, while the flagellum is posterior to the nucleus in trypomastigote and the rarer ophisthomastigote. An undulating membrane attaches the flagellum to the cell body in epimastigotes and trypomastigotes. The transformation associated with transitions between developmental stages requires substantial cell remodeling achieved by rearranging the microtubule array [189].

Essential for viability and mobility, the flagellum exits the cytoplasm through a plasma membrane invagination known as the flagellar pocket. Acting as the sole site of endocytosis and secretion, the flagellar pocket regulates protein trafficking, immune evasion, and other aspects of host-parasite interactions [190]. The basal body anchors the flagellum in the cytoplasm. Unlike most eukaryotes, trypanosomatids have a single mitochondrion, whose genome is known as the kinetoplast. The tripartite attachment complex (TAC) connects the basal body to the kinetoplast (Figure 1-3). Spanning two mitochondrial membranes, TAC is crucial for the segregation of replicated kinetoplast, while more components of TAC are still being identified [191, 192]. The relocation of the base of the flagellum and the kinetoplast is one of the most noticeable changes in cell morphology during cyclical developments [57, 193].

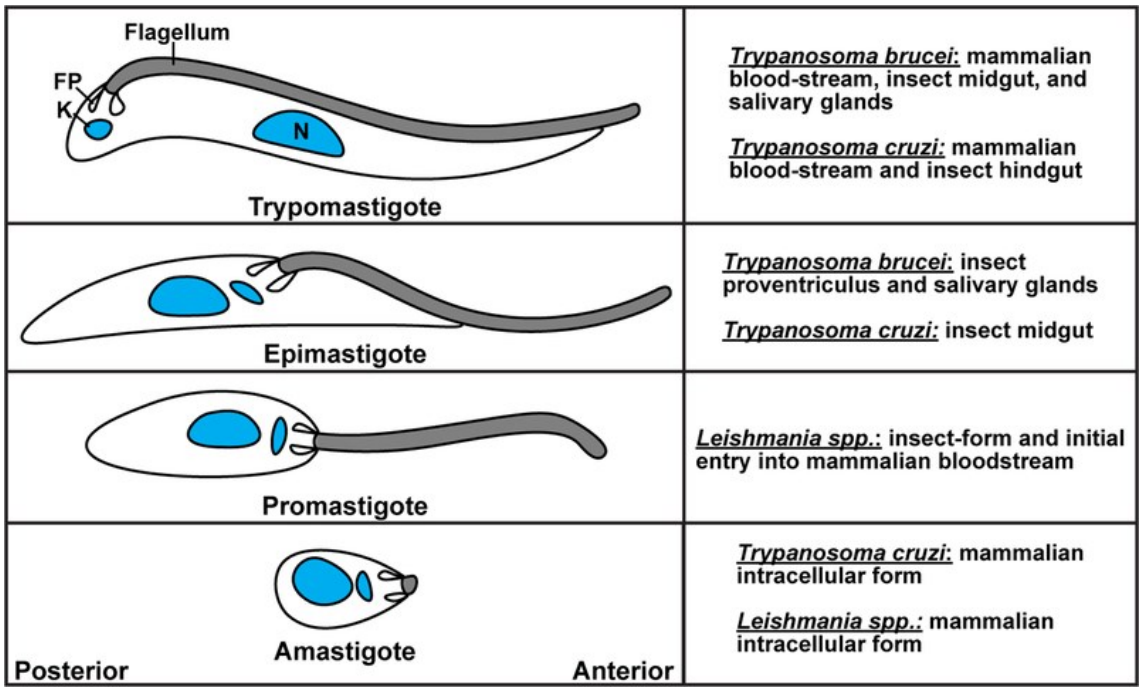


Figure 1-2. **The primary morphologies of human-infective trypanosomatids (taken from [189]).**

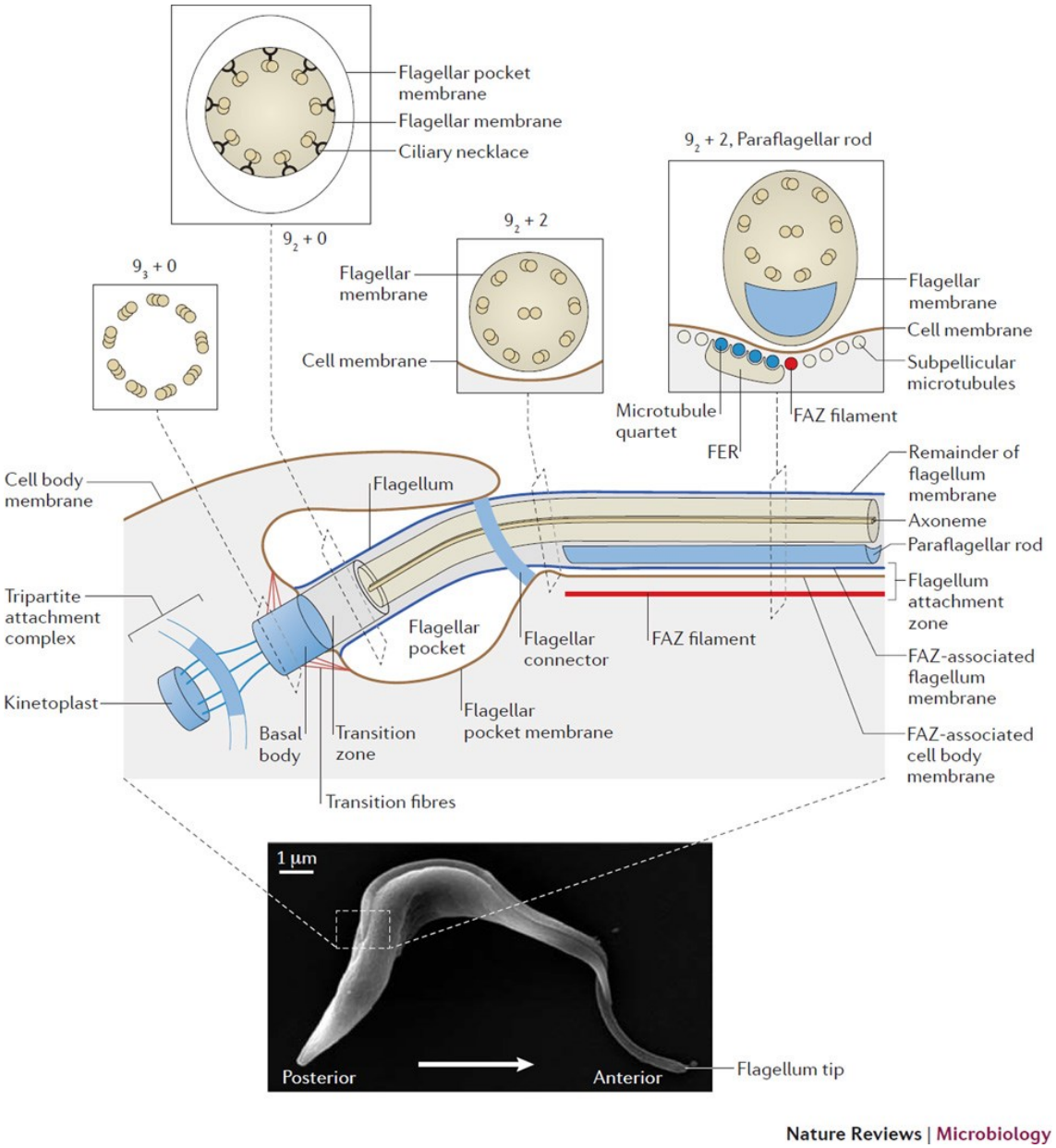


Figure 1-3. ***T. brucei* flagellum overview (taken from [194]).**

Top shows a cartoon diagram of the flagellum emerging from the flagellar pocket at the posterior end of the cell (boxed region in bottom panel). Bottom shows cartoon diagram of a *T. brucei* cell with flagellum in black. Cell movement (arrow) is with the flagellum tip leading.

1.4.2 The structure of the kinetoplast

The kDNA network in trypanosomatids consists of a chain-mail-like network of two types of circular DNAs: dozens of practically identical maxicircles of 20 to 40 kb and thousands of smaller (0.5 to 10 kb, depending on the species), typically heterogeneous, minicircles [195]. For instance, the *T. b. brucei* kDNA harbors 20-50 copies of ~23-kb maxicircle and 5 to 10k 1-kb minicircles. Maxicircles and minicircles encode complementary information. The kDNA structure of the model species *C. fasciculata* has been thoroughly investigated.

The TAC is essential for positioning kDNA within the mitochondrion and its segregation during cytokinesis [196]. The kDNA is condensed into a compact disk structure via protein-protein and protein-DNA interactions through stepwise formation and expansion of multiple condensation foci [197]. The disk's thickness is approximately the diameter of a minicircle, while on average each minicircle is connected to around three neighbors (minicircle valence=3) within a topologically controlled network [28, 198]. Minicircles at the center of the network are probably more connected than the minicircles at the periphery, and the rim of the network sequestrates most maxicircles and has significantly larger minicircle density [28]. Other trypanosomatids probably share a similar network structure with different minicircle sizes and valence. The localization of minicircle classes and maxicircle also seems to differ among species, which remains to be elucidated by further investigations [199]. [200]

1.4.3 Maxicircles

Homologous to other eukaryotic mitochondrial genomes, maxicircles encode genes essential to oxidative phosphorylation and mitochondrial translation. Trypanosomatid maxicircles are syntenic [92, 201].

In *T. brucei*, the maxicircles encode 2 rRNA genes [202] and 18 protein-coding genes including 12 cryptogenes [203]. The pre-mRNA products of cryptogenes require post-transcriptional editing by insertion or deletion of uridines to become translatable mature mRNAs. The maxicircle genes encode subunits of respiratory complex I, III, IV, V, and ribosomal subunits. Among the eight respiratory complex I, or NADH dehydrogenase (ND): ubiquinone oxidoreductase, subunit mRNAs (ND1, ND2, ND3, ND4, ND5, ND7, ND8, ND9), four (ND3, ND7, ND8, ND9) require post-transcriptional editing [204-208]. MURF1 is now known as ND2 [209]. The mRNAs of the respiratory complex III subunit apocytochrome *b* (CYB) [210], two (COX2, COX3) out of three (COX1-3) respiratory complex IV (cytochrome *c* oxidase) subunits [211-214], respiratory complex V, F_1F_O-ATP synthase subunit 6 (A6) [215] are edited. Among the two ribosomal protein subunit mRNAs (us3m, formerly known as MURF5, and us12m), us12m is extensively edited [216]. Finally, the maxicircle also contains three edited open reading frames (ORFs) with a suspected role as respiratory complex I subunits: C-rich region (CR) 3, CR4 [217], and MURF2 [217]. Beyond the largely conserved coding region is the variable region that drastically differs among species, subspecies, and even strains, whose function remains understudied yet is believed to participate in maxicircle replication [218-220].

It has been proposed that in *T. brucei*, maxicircle genes are transcribed as large polycistronic transcripts, and the adjacent genes occasionally overlap due to the compact gene organization [221-223]. However, more recently, analysis of the RNA polymerase occupancy of maxicircle DNA and the positions of mature 5' termini of procyclic stage mRNAs suggested that kinetoplast genes are transcribed as somewhat independent units in *T. brucei* [224]. Although the formation of the 3' ends may rely on the cleavage of a polycistronic transcript, gene-specific transcription initiation sites define the 5' ends of mRNAs [224].

1.4.4 Minicircles

Minicircles are highly abundant and heterogeneous in kDNA-dependent trypanosomatids. In *T. brucei*, each cell may contain 5-10k minicircles with ~400 distinct minicircle classes with ≤ 95% SID [198, 225]. The only known function of minicircles is to encode gRNAs [87]. Both strands of minicircles are transcribed into approximately 800-nt gRNA precursors – initiated at gRNA genes - which are processed into mature gRNAs by uridylation-induced, antisense transcription-controlled 3'-5' exonucleolytic degradation [226].

T. brucei and *T. congolense* gRNA genes almost always reside in cassettes flanked by 18-bp imperfectly conserved inverted repeats, whereas such cassettes are not found in *Leishmania* or *C. fasciculata* [30, 73, 225, 227, 228]. *T. brucei* minicircles have two to four cassettes, while *T. congolense* minicircles assembled up-to-date contain three cassettes [225, 228]. However, not all cassettes contain gRNAs aligned to the known edited mRNAs, called canonical gRNAs. Some cassettes may contain 'non-canonical' gRNAs identified based on nucleotide bias and not complementary to known edited mRNAs [81].

Comparing the minicircle conserved regions from eight trypanosome species reveals a common sequence motif (or conserved region) consisting of three conserved sequence blocks (CSBs), whose conserved order and spacing indicate their ancestral and vital roles [19, 197, 229]. The motif includes a 10-bp CSB-1, a more variable 8-bp sequence CSB-2, and a previously considered invariant 12-bp CSB-3, also known as universal minicircle sequence (UMS), that acts as a protein binding site and the origin of minicircle replication [19, 230]. Despite conserved motifs, the structures of minicircles vary drastically among trypanosomatids. The number of conserved regions ranges from four in *T. cruzi* to one in *Leishmania* and *T. brucei*, while each minicircle may encode one guide RNA (gRNA) as in *Leishmania* or up to four as in *T. brucei* [231].

1.4.5 The evolution of the kinetoplast

The focus on a few restricted branches, mainly the parasitic trypanosomatids, had resulted in an inaccurate impression that a kinetoplast with a condensed network of catenated circular molecules is established and widely shared by members of Kinetoplastida [198]. However, extending the investigation into previously ignored lineages reveals substantial variations in kinetoplast morphology that shed light on the emergence of the 'classic' kinetoplast with a chain-mail-like network structure [200].

The emergence of minicircles and the formation of kDNA networks are most likely separate evolutionary events [232]. The ancestral state of the kDNA structure is probably similar to the 'pan-kDNA' structure observed in the snail parasite *Cryptobia helicis*, where the supercoiled minicircles mostly remain monomeric and dispersed throughout the mitochondrial matrix [200] (Figure *1-4*). The resemblance to plasmids suggests that the minicircles may have evolved from ancient mitochondrial plasmids. Subsequently, the loss of supercoiling allows the clustering of minicircles into larger aggregates with multiple foci (poly-kDNA), as in *Dimastigella trypaniformis*, or a single focus (pro-kDNA), as in *Bodo saltans*, eventually leading to catenation of circular DNAs and the highly condensed disc-shaped kDNA in the probably most derived group of kinetoplastids: the obligate parasitic

trypanosomatids [200, 233]. The kDNA network may have emerged to reduce the risk of losing essential minicircles during segregation of kDNA and maintain the integrity and diversity of the minicircle population [234].

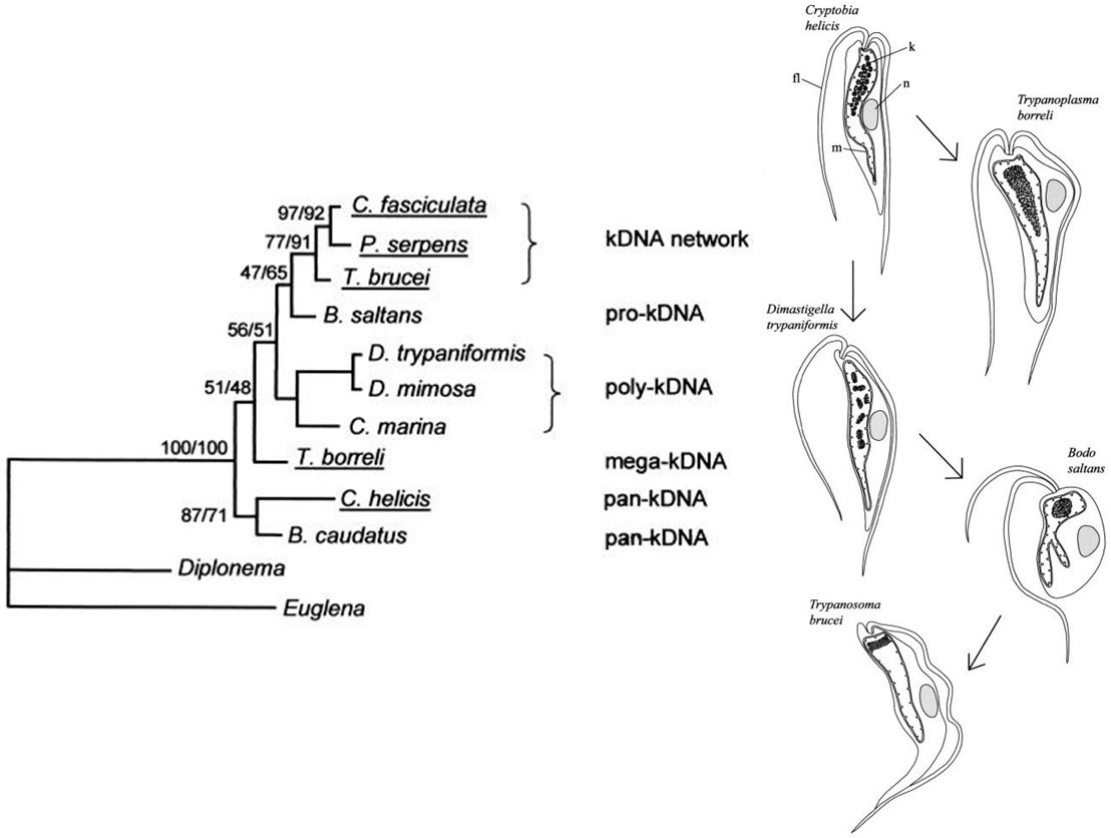


Figure 1-4. **Proposed evolution of kinetoplastids, emphasizing differences in kDNA organization and compaction (taken from [200]).**

Majority-consensus maximum-likelihood tree was constructed by using a small-subunit rRNA alignment. kDNA (k) is the structure within the mitochondrial matrix. fl, flagellum; m, mitochondrion; n, nucleus.

1.4.6 Replication and segregation of the kinetoplast

Trypanosomatids replicate via mitosis, while cell lines capable of sexual reproduction also produce gametes via meiosis [58]. Prior to cell division, the cell also completes replication and segregation of its organelle genomes. Although the mitochondrial and nuclear S phases are fairly synchronized, the replicated kinetoplast segregation precedes the nuclear division [235].

Little is known about the replication of maxicircles except that they remain attached to the kDNA disc during the unidirectional replication [236]. UMS binding protein (UMSBP) binds to UMS and initiates the assembly of the replication apparatus for minicircles [19]. The UMSBP-mediated decondensation results in a more accessible kDNA network and facilitates the release of catenated minicircles by topoisomerase II [237]. It has been widely accepted that during kDNA replication in *C. fasciculata* and *T. brucei*, the free minicircles enter the zone between the flagellar face of the disk and the mitochondrial membrane, the

kinetoflagellar zone, or KFZ, where UMSBP is localized [20, 238]. The minicircles are replicated in KFZ. The original minicircles remain circularized covalently, whereas the duplicates are gapped [29]. The difference probably ensures that each circularized DNA is copied once only [29]. Furthermore, two protein assemblies opposite each other on the kinetoplast periphery, known as the antipodal sites, mediate the distribution of duplicated minicircles into the daughter networks [21, 23].

Two mechanisms of kinetoplast replication have been described in different trypanosomatids [196, 198] (Figure 1-5). Both are believed to reduce the risk of losing essential minicircle classes during random segregation and increase the accuracy of assigning daughter minicircles into different daughter networks. *T. brucei* resides on presumably the most basal branch of the trypanosomatid phylogeny. It uses a unique, presumably more primitive, polar mechanism no longer employed by the more derived lineages [13, 22, 239]. Instead, these other trypanosomatids, including *L. tarentolae*, *L. donovani*, *P. serpens*, and *T. cruzi*, rely on an annular mechanism that involves the rotation of the kDNA disk.

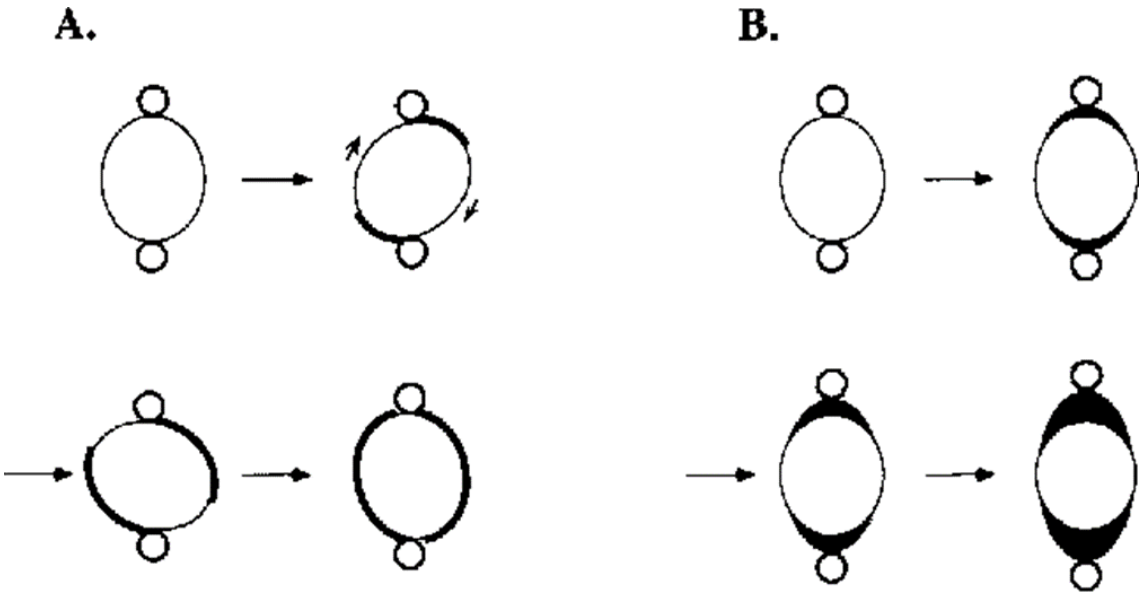


Figure 1-5 **Schematic representation of ring (A) and polar (B) mechanism of kinetoplast replication (taken from [198]).**

(A) The kDNA disk rotates relative to the antipodal sites, resulting in a thick ring-shaped network that is probably cut in the middle to release the duplicated kDNAs. **(B)** A zone of gapped minicircles forms at each network pole during replication and enlarges as the central zone shrinks, resulting in a dumbbell-shaped network with maxicircles in the centre, which are cleaved to unlink the sister networks upon segregation.

The polar mechanism in *T. brucei* does not involve the rotation of the kinetoplast [240]. Instead, minicircles are released into KFZ for replication and the daughter minicircles are distributed to the opposite ends of the network where the antipodal sites are found [22, 240]. In other words, the replicated minicircles are transported via an unknown mechanism to accumulate at the opposite ends of the kDNA network, or the ‘poles’. The central zone where the maxicircles are localized shrinks as the poles enlarge, forming a dumbbell-shaped network. The lateral elongation and a central constriction of the network probably drive the

segregation of kDNA, while ultimately the maxicircles are cleaved to unlink the sister networks [240].

More recently, an alternative loose-diploid model has been proposed for *T. brucei* kDNA replication, in which the minicircles are released, replicated, and reattached at the same lobe of the disc at the antipodal site [236]. In addition, the model assumes that the replication and reattachment of the leading strand is faster than the lagging strand, and the difference in reattachment rate results in a loose spatial separation of the daughter minicircles. After the division of the disc, the segregation ensures that the new disc contains a nearly, if not complete set of essential minicircles in each lobe, hence giving a loose diploidy to the kDNA network. This model explains the localization of some replication-associated proteins at the antipodal sites instead of KFZ and does not assume an unknown mechanism that distributes the duplicated minicircles evenly into the daughter discs. The 'sloppiness' of the minicircle localization into each lobe and the division mechanism account for the fluctuation of the minicircle population [225, 241].

Other kinetoplastids use the annular mechanism, in which the distribution of minicircles into daughter networks is achieved by attaching replicated minicircles to a rotating network. The kDNA disk rotates relative to the antipodal sites [24, 25, 242]. As the freed and replicated minicircles adhere adjacent to the antipodal sites, the movement results in the reattachment of minicircles over the entire network periphery and hence a thick ring-shaped network surrounding the covalently linked minicircles at the center yet-to-be replicated [198]. After the completion of minicircle replications, the duplicated network undergoes dramatic remodeling that reduces the minicircle valance back to three and doubles the network size. Meanwhile, the gaps in newly assembled minicircles are repaired. The flat edge observed under electron microscopy suggests that topoisomerase probably splits the doubled elliptical network with a straight cut along its short axis [198]. The daughter networks are subsequently rearranged covalently to restore the typical disk shape.

Despite the complicated replication and segregation mechanisms, the fluctuation in minicircle class copy number over time suggests that a 100% accurate distribution of daughter minicircles into daughter cells is unlikely [225]. Other evidence includes the loss of minicircle classes observed in lab cultures of *L. tarentolae* [243] and the minicircle homogenization in kDNA-independent *T. brucei* [87]. Computational simulations support the hypothesis that the minicircles are randomly segregated between daughter cells when the parent cell divides [241, 244]. The degree of stochasticity in minicircle segregation and the mechanisms in stabilizing minicircle population remain to be determined experimentally.

1.5 Trypanosomatid mRNA editing

1.5.1 RNA editing in prokaryotes and eukaryotes

RNA editing occurs widely among eukaryotes [245]. Over 100 million A-to-I editing sites are located in human primate-specific Alu sequences but barely in coding sequences [246]. In coleoid cephalopods, neural transcriptomes show A-to-I RNA editing enrichment, which results in the diversification of proteomes that presumably bear a selective advantage [247]. In bacteria, besides the A-to-I editing on the wobble position of the ACG anticodon in the tRNA, A-to-I editing in coding regions of mRNA is also detected [248].

However, the unique editing mechanism by insertion and deletion of uridylates (U-indel) - the first type of RNA editing discovered (Benne et al 1986) - has only been described in kinetoplastids. Most studies on U-indel focus on pathogenic trypanosomatids of humans and animals [249, 250]. Nevertheless, free-living *B. saltans* in the sister taxon Eubodonida [251] and the extracellular fish parasite *Trypanoplasma borreli* in Parabodonida also exhibit U-indel editing [252-254]. Despite the lack of direct evidence of mRNA editing in basal Bodonids, uridine enrichment and variation in COXII and ND5 mRNAs and the detection of homologs of the trypanosomatid RNA editing machinery indicate the presence of editing-like mechanisms in Neobodonida [14]. The evidence suggests RNA editing is an ancestral trait that emerged before the divergence of Metakinetoplastina [14].

1.5.2 Structure and evolution of gRNAs

Guide RNAs are short (40-60 nt) untranslated transcripts, whose only known function is to direct post-transcriptional editing of maxicircle-encoded mRNAs. A *T. brucei* gRNA contains an AT-rich initiation sequence at its 5’ end, an anchor region for target mRNA recognition, and a guiding sequence that directs editing by its complementarity to the edited sequence, followed by a 3’ untemplated oligo(U) tail from posttranscriptional uridylation [225, 226]. Although the function of the U-tail is not fully understood, RNA-RNA cross-linking has shown that the U-tail preferably interacts with the purine-rich region on the unedited ‘pre-mRNA’ [255] and is protected by the cognate mRNA [256]. Hence, the U-tail probably plays a role in stabilization of gRNA-mRNA interaction during the editing.

A key unanswered question concerns the evolutionary origin of gRNA genes. Novel gRNAs may have originated from reverse-transcribed, partially duplicated, and subsequently inverted (or transcribed in the opposite direction) fully edited mRNAs [250]. Alternatively, gRNA genes could also predate the emergence of cryptogenes in a constructive neutral evolution model [257]. The original gene (or fragment thereof) was duplicated, and anti-sense transcription produced a gRNA, the existence of which, together with a piece of primordial editing machinery, subsequently allowed gene mutation by T deletion [257].

T. brucei and *T. congolense* encode almost all gRNAs on the minicircles [225, 230]. *Leishmania* also houses a complex kDNA network, whose maxicircles encode the same set of rRNAs and protein-coding genes [25, 203]. However, unlike *T. brucei* maxicircles that encode only two gRNAs, including the *cis*-gRNA within the COX2 mRNA 3’ UTR, a greater number of maxicircle-encoded gRNAs have been described in *C. fasciculata* and *L. tarentolae* and later

in *L. peruviana* and *L.braziliensis* [66, 258, 259]. The minicircle structure, editing mechanisms, and mRNA editing patterns of *Leishmania* are distinct from *Trypanosoma*, providing an intriguing subject for investigating the evolution of the editing apparatus [66, 260].

Several questions emerge when comparing the maxicircle gRNAs in the presumably ancestral *T. brucei* to the more derived *Leishmania*. Why does *T. brucei* encode only two gRNAs on maxicircles, while *L. peruviana* and *L. braziliensis* have as many as 22 maxicircle gRNAs [66, 225]? Did *Leishmania* and *Crithidia* acquire additional maxicircle gRNAs after diverging from *T. brucei*, or did the latter lose the maxicircle gRNAs after switching to minicircles for encoding most gRNAs? Finally, is the lack of maxicircle gRNAs associated with the extensive editing in *T. brucei*?

1.5.3 mRNA editing in trypanosome parasites at different life stages

In *T. brucei* and *T. congolense*, the two rRNAs and mRNAs of six genes do not require posttranscriptional editing. The mRNAs of three (COX2, CYb, MURF2) of the 12 cryptogenes are 'minimally' edited, i.e. edited over a limited region, and the mRNAs of the remaining nine cryptogenes are edited extensively and thus termed 'pan-edited' (Figure *1-6*A). As mentioned above, the COX2 gRNA responsible for the four uridine insertions to restore the ORF is present *in cis* at the mRNA's 3' UTR, while other mRNAs are *trans*-edited [211, 213].

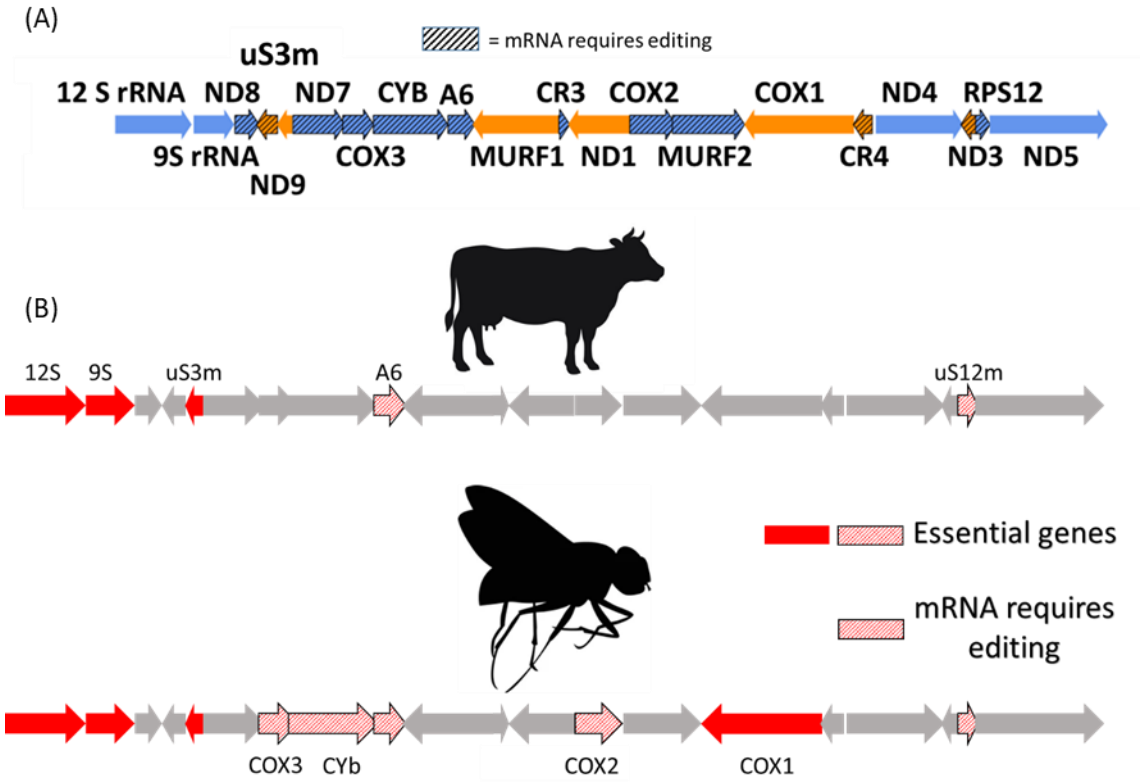


Figure 1-6. **Schematic representation of maxicircle coding region and essential maxicircle genes at mammalian and insect stage for *T. brucei*.**

(A) Schematic representation of the maxicircle coding region, including 2 rRNA genes and 18 protein coding genes. Cryptogenes that require posttranscriptional editing are hashed. **(B)** known essential maxicircle genes at two key stages of development. Solid red: essential genes where the RNA products don't require editing. Hashed red: essential cryptogenes.

To produce translatable mRNAs, trypanosomatids must cover all editing sites with at least one gRNA, demanding at least 200 different gRNAs to produce all fully edited mRNAs in *T. brucei*, assuming an average gRNA length of 40 bp and minimal overlaps (arbitrarily set at 6 nt for this calculation) necessary for the anchor regions [225, 241]. The large set of gRNAs required in *T. brucei* leads to a diverse minicircle population. Nevertheless, not all gRNA genes, and hence not all minicircles, are essential for the parasite at all life stages.

Dixenous trypanosomatids that complete cyclical development between mammalian hosts and insect vectors undergo drastic remodelling of mitochondrial structure and function to accommodate different metabolisms. The shift to proline metabolism and oxidative phosphorylation (OXPHOS) in insect-stage *T. brucei* entails that most of the maxicircle encoded ETC subunits are required. Nevertheless, complex I is active but not essential in procyclic *T. brucei* [261], which also uses alternative NADH dehydrogenase (NDH2) for cytosolic NAD+ generation. Although the involvement of respiratory complex I remains elusive in any lifecycle stage, respiratory complexes III, IV, and V, besides both rRNAs and mitoribosomal subunits, play an essential role in insect stage *T. brucei*, and the corresponding maxicircle genes must be transcribed and edited accordingly [262, 263] (Figure *1-6*B).

Conversely, the bounty of bloodstream glucose in mammalian hosts enables trypanosomatids to employ less efficient glycolysis to satisfy the energy requirement and maintain a highly reduced tubular mitochondrion without OXPHOS [264]. Consequently, besides the two rRNAs and two ribosomal subunits, the only confirmed essential maxicircle gene encodes subunit A6 of the F_1F_O-ATP synthase, which maintains the mitochondrial membrane potential via ATP-hydrolysis mediated proton pumping [35, 264, 265] (Figure *1-6*B).

1.5.4 Proteins involved in mRNA editing

The holo-editosome includes proteins of two specialized complexes that appear to interact dynamically: a modular RNA editing substrate binding complex (RESC) and the catalytic ~20S RNA editing core complex (RECC) [266].

RESC binds RNA editing substrates (pre-edited mRNA and gRNA), intermediates (partially edited mRNAs), and products (fully edited mRNAs). RESC loads a single gRNA each time [267], which is responsible for the transient interaction with the catalytic RECC [268, 269]. Three versions of RESC have been described: gRNA-stabilizing RESC-A (consisting of RESC1-6), gRNA-mRNA-binding RESC-B (consisting of RESC5-11, RESC13, and RESC14), and RESC-C (consisting of RESC5, RESC8, RESC10, and RESC14) [269]. The function of RESC-C remains uncertain, although it also binds mRNA and gRNA. It may represent an intermediate of RESC-B assembly or a step in RESC-B remodelling during editing [269].

In RESC-A, the RESC1-RESC2 heterodimer binds exclusively to 5'-triphosphate nucleosides, hence binding gRNA but no other mitochondrial RNAs [270]. The gRNA is protected in RESC-A from degradation by forming a hairpin structure that locks its anchor and guiding region and prevents the access of mRNA [269]. The gRNA 5' end is wrapped in the triphosphate binding tunnel of RESC2 while the 3′ end docks into the crevice formed by RESC5 and RESC6 [269]. RESC5 also selects gRNAs with longer U-tails.

Remodelling RESC-A into RESC-B unfolds the gRNA and exposes the 5' region, while the 3' end is kept in close interaction with RESC5, RESC6, and RESC10 [269]. The mRNA traverses the entire RESC-B in the opposite direction as the gRNA, which allows hybridization beyond the RESC surface [269]. Subsequently, RESC-B acts as the editing-competent substrate of the RECC [269], extending the gRNA-mRNA duplex beyond the RESC surface as U-indel restores the complementarity [269]. The ratio of fully edited mRNA associated with each subunit increases from RESC12 and RESC13 preferably associated with unedited mRNAs to RESC5, RESC6, and RESC10 in close contact with the gRNA-mRNA duplex [269].

The RECCs are also known as ~20S editosomes. Three types of RECCs have been described, each containing a common set of 12 proteins and one of the three mutually exclusive RNase III endonucleases with associated specific partner proteins that stabilize RECCs via dimerization [271] (Figure *1-7*). The RECC that contains the kinetoplastid RNA editing endonuclease (KREN)1/kinetoplastid RNA editing protein binding (KREPB)8 protein pair and kinetoplastid RNA editing exonuclease (KREX)1 cleaves sites for uridine deletion. The other

two versions of RECC contain KREN2/KREPB7 or KREN3/KREPB6 protein pairs and catalyze uridine insertion with different substrate preferences [271].

RECCs also consist of proteins without known catalytic properties that influence the parasite's viability [272, 273]. Some of the non-catalytic proteins probably differentially regulate stage-specific editing activity, while others are structural components or important for substrate binding [263]. Mutations on ZnF, RNase III, and RAM domains of the KREPB proteins causes growth defects in both BSF and PF *T. brucei* have been reported [274]. The mutations also affect the abundance of their cognate KRENs and edited mRNAs differentially [274]. Although repression of some noncatalytic proteins is lethal in BSF and procyclic *T. brucei*, the repression impacts RECC structure differently [272]. Meanwhile, parasites' response to point mutations in, or loss of, some of the non-catalytic proteins regarding cell growth, editosome integrity, and RNA editing differs between BSF and procyclic parasites [263, 272, 273, 275].

It has been proposed that RECC catalyzes a single round of mRNA editing in the following steps. After endonucleases cleave the mRNA substrate at the first mismatch 5' to the anchor (with respect to the mRNA) to expose the editing sites, 3' terminal uridylyltransferase (TUTase) and U-specific 3'exonuclease (exoUase) catalyze uridine insertion and deletion (U-indel), respectively, using gRNA as a template [276]. Upon completion of U-indel, RNA ligases reconnect the mRNA fragments [271, 277], extending the duplex between gRNA and mRNA. RECC can then move on to the next editing site.

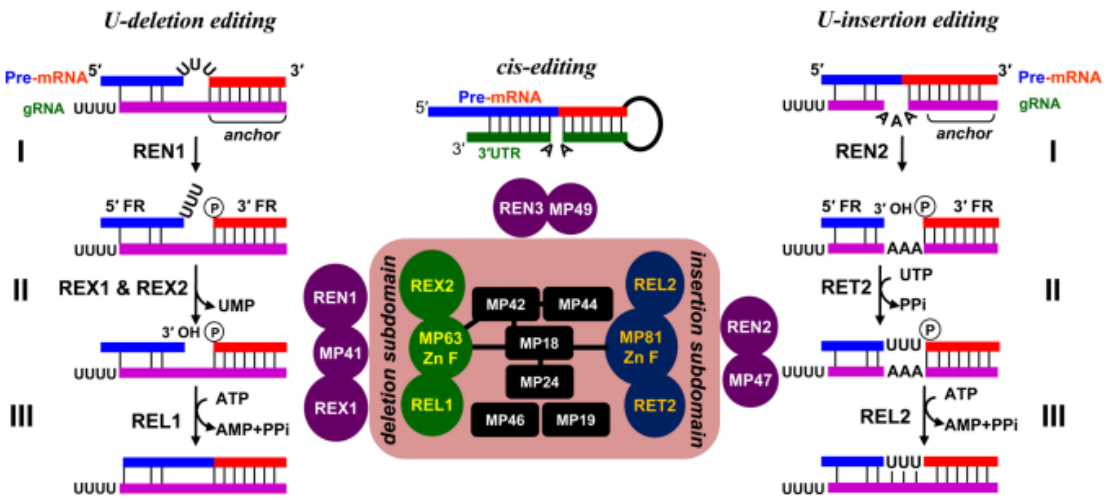


Figure 1-7 **The RNA editing core complex (RECC) catalyzes elementary RNA editing reactions (taken from [268]).**

Direct protein-protein interactions within the core complex are depicted by black bars. Roman numerals signify three elementary steps of RNA editing: mRNA cleavage, U-deletion or insertion and mRNA ligation. MP: mitochondrial protein (structural and/or RNA binding components); REX: RNA editing exonuclease; REN: RNA editing endonuclease; REL: RNA editing ligase; RET: RNA editing TUTase; anchor: 5-15-nt long double-stranded region formed by the 5' portion of the gRNA and pre-edited mRNA.

1.5.5 The polarity of mRNA editing

The interaction of pre-mRNA and successive gRNAs dictates an overall 3' to 5' directionality for editing [259, 278]. The gRNAs recognize the target area on pre-mRNA via Watson-Crick base-pairing of the anchor regions. The anchor of the first gRNA in a gRNA cascade aligns to the never-edited region downstream of the region on the mRNA to be edited (the 'editing domain'). The 3' most gRNAs are termed initiation gRNAs as they start the editing cascade by covering the first editing sites. Most mRNAs have only one site for editing initiation. The pan-edited ND7 presents an exception that contains two editing domains and practically two sets of initiation gRNAs each for one domain [206].

Most subsequent gRNAs rely on the correct modifications directed by the preceding gRNA to generate sequences recognizable by their anchors [259]. The prerequisite of accurately edited upstream mRNA for gRNA anchoring means that, overall, editing proceeds from 3' to 5' on mRNAs [258, 259]. Cryptogene mRNAs with unedited 5' regions and progressively edited 3' regions validate the polarity of mRNA editing [279].

The polarity of editing results in unique transcriptomic patterns, including intermediates and dead ends. Fully edited mRNAs only account for a small fraction of transcripts from pan-edited mRNAs [280]. In reality, within an editing block (the region of editing directed by a single gRNA), the directionality is not strictly 3'-5'. Instead, editing proceeds through dynamic and progressive realignment of gRNA and partially edited mRNA, creating 'junction' regions as intermediates that neither resemble unedited nor fully edited sequences [223]. Editing intermediates that contain junctions are the most abundant among transcriptomes [249, 279]. The biased distribution of junction ends along mRNAs (termed 'intrinsic pause sites') suggests that canonical editing tends to end at particular editing sites of the mRNAs. The difficulty in making canonical modifications of certain editing sites is probably associated with mRNA structures, mRNA-gRNA interaction, or gRNA structure affect editosome activities [249].

Besides intrinsic pauses, alternative or incorrect gRNAs can also alter the canonical editing patterns and result in junctions [249, 280]. Nevertheless, the flexibility of gRNA utilization opens up the possibility of producing alternative ORFs from the same pre-mRNA, although whether any such non-canonical ORFs are translated remains unknown [249]. So far, ND7-9, COX3, CR3, and A6 mRNAs show evidence of alternative ORFs, and the corresponding gRNAs are identified [225, 281-284]. Although the function of the putative novel protein sequences is unknown for most alternative mRNAs, the alternative COX3 mRNA was suggested to be translated into a putative mitochondrial membrane protein in *T. brucei* [281].

1.5.6 The evolution of mRNA editing domains in kinetoplastids

Kinetoplastid species vary in the editing domain sizes probably due to a history of editing site loss via retroposition (Figure *1-8*). For a given cryptogene, the early diverging lineages including *T. brucei* and *T. congolense* perform more extensive mRNA editing, and the versions with more restricted editing domains in the more derived lineages may represent editing site loss via retroposition [250]. An alternative explanation is that lineages on the

basal branches accumulated more editing sites after the divergence of the more derived clades.

A detailed comparison of editing domains in kinetoplastids shows that the extent of editing differs substantially in different lineages (Figure *1-8*). *T. borreli* [252], *C. fasculata* [31], *L. tarentolae* [216], and *T. brucei* [225] all show extensive editing over RPS12 mRNA. Heavily edited COX3 is detected in *T. brucei* and *Herpetomonas muscarum* from separate clades, while *T. borreli*, *Leishmania*, and *Blastocrithidia cullicis* have unedited COX3. A6 and ND7 are pan-edited in *T. brucei*, but the editing domains shrink towards the 5' end in more derived trypanosomatids [250]. The novel editing patterns unseen in related lineages suggest that some editing sites may emerge independently. For instance, only *T. borreli* but no other species examined so far have edited COXI and a 3' editing domain on CYB [250, 254]. Meanwhile, editing of ND5 has only been recorded in *B. saltans* [251].

Some genes including ND4 probably never undergo U-indel among known kinetoplastids. Unedited in the more basal taxa *T. borreli* and *C. helicis*, the highly conserved cis-editing in COX2 is shared by bodonid species (2 insertions) and trypanosomatids (4 insertions) [251-253]. Compared to pan-edited genes, non-edited and minimally edited genes have significantly lower CG% approaching the limit of CG loss, where further reduction will introduce nonsynonymous substitutions [92].

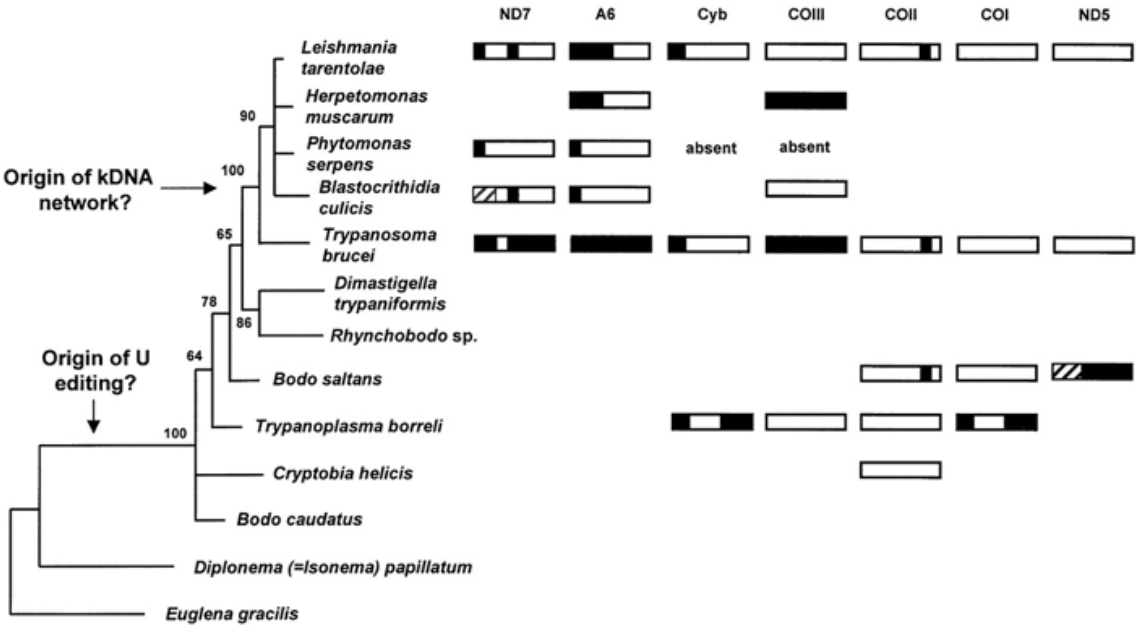


Figure 1-8. **Phylogenetic analysis of kinetoplastid RNA editing (taken from [250]).**

A maximum likelihood ribosomal RNA phylogenetic tree representing only the main trypanosomatid and some bodonid lineages (J. Lukeš and D. Maslov, unpublished results) with the corresponding bootstrap values is shown on the left. A representation of the known cryptogenes is shown on the right: open boxes correspond to unedited sequences, black boxes to pre-edited regions or edited cryptogenes, and cross-hatched boxes indicate a lack of information.

1.5.7 The driving force behind mRNA editing in kinetoplastids

Unlike nuclear DNA, mitochondrial DNA (mtDNA) is unprotected by histone. Meanwhile, oxidative phosphorylation in mitochondria entails that mtDNA is more exposed to reactive oxygen species, which leads to DNA damage. It was believed that mtDNA reparation occurs minimally if not at all, and the aged mitochondria with damaged mtDNA were degraded and replaced by newly generated successors with intact mtDNA. Now it is known that mitochondrion is capable of limited types DNA reparation [285-287], but similar reparation pathways remain undescribed in kinetoplastids. Furthermore, containing a single kinetoplast instead of multiple mitochondria per cell imposes a unique challenge to kinetoplastids. It probably necessitates compensatory mechanisms for mtDNA damage, as the cells cannot renew the kinetoplast via degradation and regeneration. It is likely that RNA editing initially serves as a curation system to ensure the expression of mitochondrial genes by restoring the open reading frame [288].

An editing mechanism probably allows additional changes to mtDNA and the accumulation of editing sites to be tolerated. Attempts to identify the evolutionary advantage of editing site expansion in kinetoplastids have proposed potential benefits including (1) additional regulation for mitochondrial gene expression [289-293]; (2) more genetic variations that precipitate evolution [294]; (3) the ability to generate multiple protein products via alternative editing [279, 281, 282]; (4) tolerance to mutations accumulating in the mitochondrial genome due to production of reactive oxygen species [288]. Admittedly, these benefits may not be decisive enough to justify the accumulation of editing sites, and no evidence has been uncovered to validate or refute these theories.

There is no consensus on whether U-indel is ‘on the way in’ or ‘on the way out’. As discussed above, a retroposition model has been proposed for the loss of editing in evolution, where the cells replace the original cryptogenes with cDNAs from partially edited mRNAs produced by unidentified mitochondrial reverse transcriptase activity [295]. Editing site removal substantially reduces the set of essential gRNAs and allows the loss of minicircle classes. Given the polarity of mRNA editing, this model would predict a 3’ to 5’ directionality in editing site loss. If *T. brucei* with extensively edited mRNAs represents the ancestral state, the retreating of A6 and ND7 editing domains to the 5’ end in postulated more derived lineages agrees with the prediction [250]. The remaining 5’ editing sites may allow post-transcriptional gene regulation, although direct evidence for this is lacking. However, suggesting that U-indel is ‘on the way out’ due to its apparent lack of benefits does not explain its origin and expansion in the first place, while *de facto* posing a further question on what evolutionary factors have changed to arrest the accumulation of editing sites and instigate the streamlining a seemingly unnecessarily complex and costly mechanism.

Dated to the divergence of euglenoid and kinetoplastids, the mRNA editing mechanism may have initially evolved in the facultative anaerobic ancestors to repair damages from reactive oxygen species in aerobic environments [250]. Acquiring strong selective advantage subsequently in permanent aerobic environments, the mechanism may also drive the fixation of mutations at editable sites on mitochondrial genes and the complementary

templates. Constructive neutral evolution (CNE) may lead to the expansion of mRNA editing without invoking immediate benefits [296]. A bias for uridine deletions instead of insertions in genes inevitably leads to the accumulation of editing sites tolerated by existent editing machinery [296].

1.6 Gaps in knowledge and key questions

The advancement in high throughput sequencing technologies enables the investigation of the kDNA genome functions and the mRNA editing mechanisms at higher resolution [279, 297]. Although the minicircle population in *T. brucei* isolates capable of sexual reproduction is known to be diverse [80, 225], a comprehensive study comparing the complexity of kDNA and the editing capacity of *T. brucei* subspecies adapted to different life cycles is lacking.

In addition, besides the investigation of the editing apparatus (1.5.4) and the polarity of editing (1.5.5), little is known about how the progress of editing, as exemplified by the alignment of gRNAs, compares between groups of trypanosomes. To address these questions, we conducted a detailed investigation of 224 African *T. brucei* isolates.

In addition, despite the economic relevance of *T. congolense* and emerging evidence for important differences in its bioenergetics to *T. brucei*, little is known about the minicircle population and the editing activities in *T. congolense* [42]. Guide RNA containing cassettes flanked by inverted repeats have been reported based on two *T. congolense* minicircles. The repeats share substantial similarities with repeats described in *T. brucei* [228, 230]. Three edited mRNA sequences and their stage-specific relative abundance are conserved between two species [204, 289]. Nevertheless, these studies are insufficient in fully capturing the kDNA dynamics of *T. congolense*, which is further complicated by hybridization between distinct lineages [298, 299]. We conducted a detailed analysis of three isolates, including the reference IL3000, to provide a more comprehensive image of the editing dynamics in *T. congolense*.

Finally, the historical diagnosis and classification of *T. b. equiperdum* and *T. b. evansi* have been controversial [83, 179]. The emerging molecular evidence suggests multiple independent origins of these parasites [90, 91] and necessitates closely examining their kDNA with higher resolution. In addition, preliminary investigations have revealed unexpected features in minicircle populations of certain *T. b. equiperdum* isolates [73, 186], which raises further questions on their editing capacity, gRNA features, and reliance on kDNA. To address these questions, we analysed 43 *T. b. equiperdum* and *T. b. evansi* samples from various geographical origins collected over decades.

2 Methods

2.1 General announcements

2.1.1 List of exceptions

All work was carried out by myself, except the following:

- *De novo assembly* of the nuclear genome of sub-Saharan *T. brucei* isolates 3.1)
- Detection of SRA and TbgsGP marker genes in sub-Saharan *T. brucei* field isolates 3.1)
- Nuclear genome phylogeny tree of sub-Saharan *T. brucei* isolates 6.4.1)
- Ethidium bromide challenge of type OVI *T. b. equiperdum* [300] (7.1.4)

Material and data provided by collaborators were specified in the following sections.

2.1.2 Supplementary data

All customed Python scripts are available on Github: https://github.com/Zedthedrifter/Thesis_2024.git.

The Supplementary Figures, Tables, and Datasets are available on Figshare (Table *2-1*).

The NGS data will be uploaded to publicly accessible archives before the publication of corresponding papers with our collaborators.

Table 2-1. **Availability of Supplementary information**

Item	**DOI link**
Supplementary Figures and Tables	https://doi.org/10.6084/m9.figshare.27186972
T. b. gambiense type 1 Mongo gRNA alignments	https://doi.org/10.6084/m9.figshare.27174039
T. b. gambiense type 1 collective gRNA alignments	https://doi.org/10.6084/m9.figshare.27146595
Sub-Saharan *T. brucei* minicircle annotation	https://doi.org/10.6084/m9.figshare.27174027
T. b. equiperdum type OVI gRNA alignments	https://doi.org/10.6084/m9.figshare.27063367
T. congolense IL3000 gRNA alignments	https://doi.org/10.6084/m9.figshare.27020074
T. congolense Kapeya gRNA alignments	https://doi.org/10.6084/m9.figshare.27020083
T. congolense UPKZN gRNA alignments	https://doi.org/10.6084/m9.figshare.27020089

2.1.3 Workflow of kDNA analysis

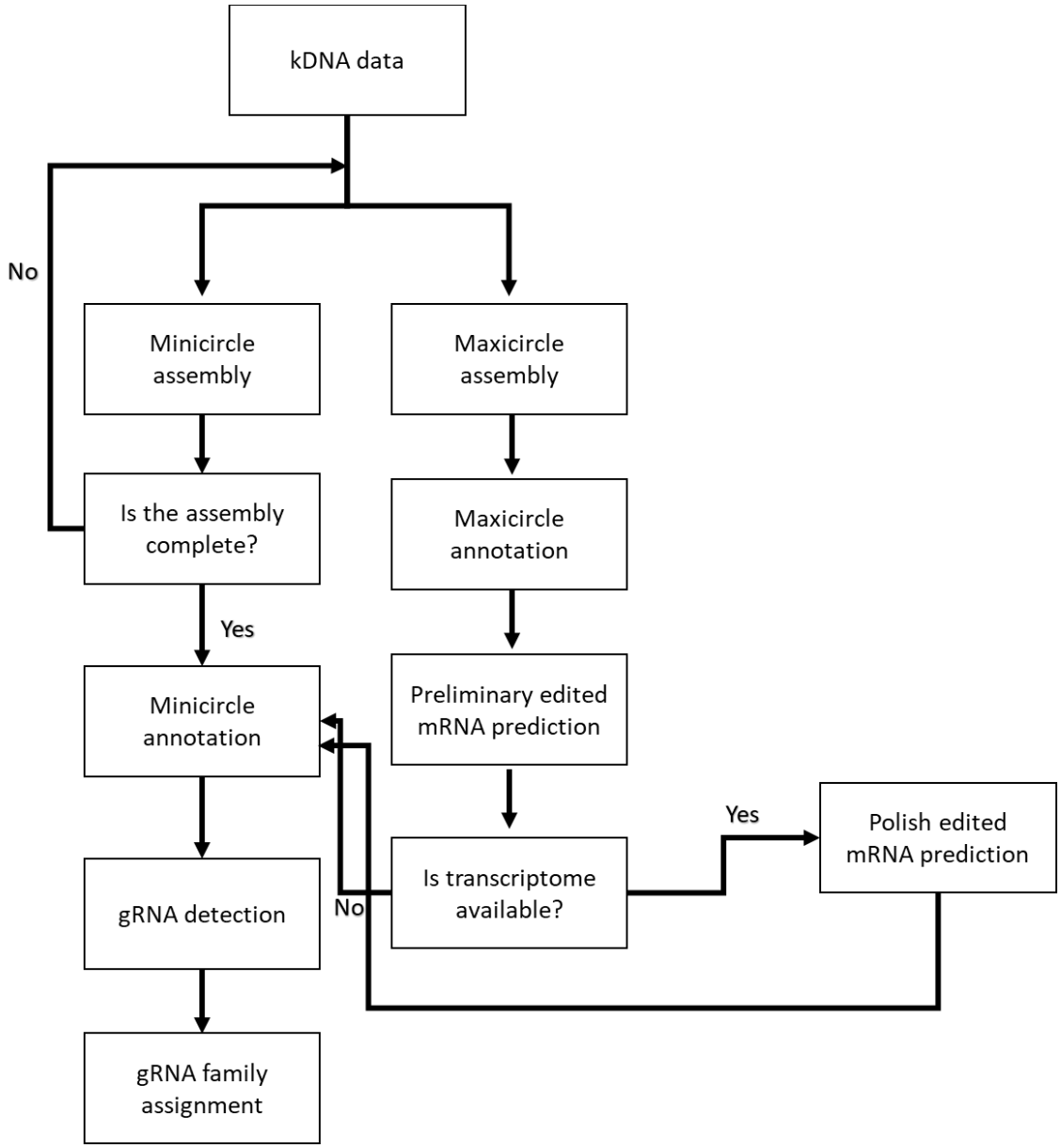


Figure 2-1 **Workflow diagram of the kDNA analysis procedure.**

The diagram summaries the general workflow of our kDAN analysis project. After acquisition of the kDNA data, maxicircles and minicircles were assembled. The maxicircles were annotated to extract the unedited mRNAs. We modified the published EATRO1125 edited mRNAs, with the strain-specific maxicircles from *de novo* assembly to obtain the preliminary edited mRNA predictions. When transcriptome data were available, we polished the prediction using the transcriptome reads. The unedited and edited mRNAs were used for minicircle annotation and gRNA identifications. The completeness of the minicircle assembly is assessed before proceeding to minicircle annotation. The gRNAs were assigned into gRNA families for further analysis.

2.2 Metadata and data availability

Table 2-2. ***T. brucei*** **isolates included in this study**

Isolate	*Taxon*	Year of Isolation	Continent	Host
AnTat-4-1	*T. b. equiperdum (mixed?)*	?		horse
AnTat-4-1-bis	*T. b. equiperdum (mixed?)*			horse
E28	*T. b. equiperdum (mixed?)*	1977		horse
Alfort	*T. b. equiperdum A*			
American-Strain	*T. b. equiperdum A*			
ATCC-30023	*T. b. equiperdum A*			
ATCC30019	*T. b. equiperdum A*			horse
ATCC-30019	*T. b. equiperdum A*			horse
Canadian-Strain	*T. b. equiperdum A*			horse
STIB818	*T. b. equiperdum A*	1979		horse
SVP	*T. b. equiperdum A*			
HAMBURG	*T. b. equiperdum A*			horse
BoTat-1-1	*T. b. equiperdum C*			horse
BoTat-1-1-bis	*T. b. equiperdum C*	1924	North Africa	horse
940	*T. b. equiperdum OVI*	2008	East Africa	horse
Te-Ap-N-D1	*T. b. equiperdum OVI*	1991		horse
Te-Ap-N-D1-bis	*T. b. equiperdum OVI*	1991	South America	horse
OVI	*T. b. equiperdum OVI*			horse
Kenya-a	*T. b. evansi (mixed?)*		East Africa	camel
Kenya-c	*T. b. evansi (mixed?)*		East Africa	camel
AnTat-3-1	*T. b. evansi A*	1969	South America	capybara
AnTat-3-3	*T. b. evansi A*	1969	South America	capybara
Colombia	*T. b. evansi A*	1973		horse
Kazakstan	*T. b. evansi A*	1995	Asia	camel
280104	*T. b. evansi A*	1982	Asia	water buffalo
STIB-816	*T. b. evansi A*	1978		bactrian camel
MCAM-ET-2013-MU-01	*T. b. evansi A*	2013		camel
MCAM-ET-2013-MU-02	*T. b. evansi A*	2013		camel
MCAM-ET-2013-MU-04	*T. b. evansi A*	2013		camel
MCAM-ET-2013-MU-05	*T. b. evansi A*	2013		camel
MCAM-ET-2013-MU-09	*T. b. evansi A*	2013		camel
MCAM-ET-2013-MU-17	*T. b. evansi A*	2013		camel
Kenya	*T. b. evansi A*		East Africa	camel
Merzouga-56	*T. b. evansi A*	1997		camel
Zagora-I-17	*T. b. evansi A*	1997		camel
Philippines	*T. b. evansi A*	1996		water buffalo
AnTar-7	*T. b. evansi A*			
MCAM-ET-2013-MU-14	*T. b. evansi B*	2013		camel
AGAL-CI-78-TCH312	*T.b. brucei*	1978	West Africa	chicken
AGAL-CI-78-TCH312-bis	*T.b. brucei*	1978	West Africa	chicken
GMOM-ZM-83-TRPZ-317	*T.b. brucei*	1983	Southern Africa	tsetse
GMOM-ZM-83-TRPZ323	*T.b. brucei*	1983	Southern Africa	tsetse
Etat-1-2R	*T.b. brucei*	1960	East Africa	tsetse

GPAL-ZM-80-TRPZ13	*T.b. brucei*	1980	Southern Africa	tsetse
GPAL-ZM-83TRPZ265	*T.b. brucei*	1983	Southern Africa	tsetse
GPAL-ZM-83-TRPZ320	*T.b. brucei*	1983	Southern Africa	tsetse
GPAP-CI-82-KP10-29	*T.b. brucei*	1982	West Africa	tsetse
GPAP-CI-82-KP10-29-bis	*T.b. brucei*	1982	West Africa	tsetse
AnTat-5-2	*T.b. brucei*	1975	West Africa	cattle
Lister-427-AT1-KO	*T.b. brucei*	1956	East Africa	cattle
MBO-NG-74-R10	*T.b. brucei*	1974	West Africa	cattle
MBO-NG-74-R10-bis	*T.b. brucei*	1974	West Africa	cattle
MBOT-GM-77-GB2	*T.b. brucei*	1977	West Africa	cattle
MCAP-CI-91-BALEA-2	*T.b. brucei*	1991	West Africa	goat
J10	*T.b. brucei*	1973	Southern Africa	hyena
AnTat-17-1	*T.b. brucei*	1978	Central Africa	sheep
AnTat-34-1-P10	*T.b. brucei*	1978	Central Africa	sheep
MSUS-CI-78-TSW129-cloneA	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-78-TSW129-cloneC	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-78-TSW-157	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-78-TSW168	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-78-TSW178	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-78-TSW185-cloneA	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-78-TSW185-cloneC	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-78-TSW38-021	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-78-TSW382	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-78-TSW382-bis	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-82-TSW31-BO1	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW31-KP1	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW32-B01	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW36	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW62	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW62-bis	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW65-KP1-exbiit	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW75	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW95	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-83-TSW-11	*T.b. brucei*	1983	West Africa	pig
MSUS-CI-85-PTAG-130	*T.b. brucei*	1985	West Africa	pig
MSUS-NG-62-B8-18-cloneB	*T.b. brucei*	1962	West Africa	pig
AnTat-1-1	*T.b. brucei*	1966	East Africa	bushbuck
AnTat-1-1E	*T.b. brucei*	1966	East Africa	bushbuck
LIZZARD	*T.b. brucei*	1996	East Africa	monitor lizard
AnTar-2	*T.b. brucei*	na	na	
AnTar-30	*T.b. brucei*	na	na	
AnTat-25-1S-bis	*T.b. brucei*	na	na	
GTAC-ET-70-GAMBELA4	*T.b. brucei*	1970	East Africa	G.TACHINOIDES
ITMAP1892	*T.b. brucei*	na	na	
MALC-BF-80-AB14	*T.b. brucei*	na	West Africa	hartebeest
MALC-BF-80-AB25	*T.b. brucei*	1980	West Africa	hartebeest

MBOT-CI-78-TC348	*T.b. brucei*	na	West Africa	cattle
MKOD-BF-80-KD3	*T.b. brucei*	180	West Africa	waterbuck
MKOK-BF-80-KK1	*T.b. brucei*	1980	West Africa	kob
MKOK-BF-80-KK25	*T.b. brucei*	1980	West Africa	kob
MKOK-BF-80-KK7	*T.b. brucei*	1980	West Africa	kob
MKOK-BF-90-KK17	*T.b. brucei*	1980	West Africa	kob
MSUS-CI-78-TSW176	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-78-TSW185-bis	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-78-TSW187-bis	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-78-TSW190	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-78-TSW390-bis	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-82-TSW171	*T.b. brucei*	1978	West Africa	pig
MSUS-CI-82-TSW23	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW26	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW35	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW38-B01	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW46	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW53	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW65-KP1-bis	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW8	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-82-TSW84	*T.b. brucei*	1982	West Africa	pig
MSUS-CI-83-TSW12	*T.b. brucei*	1983	West Africa	pig
MTRG-BF-80-TS2	*T.b. brucei*	1980	West Africa	bushbuck
P16F	*T.b. brucei*	na	Central Africa	pig
UTAT-1-1	*T.b. brucei*	na	na	
YAOUNDE	*T.b. brucei*	na	na	
NTUMA	*T.b. gambiense type I*	na	Central Africa	
MOS-bis	*T.b. gambiense type I*	1974	Central Africa	human
MOERBEKE-82	*T.b. gambiense type I*	na	Central Africa	
D12K-ITMAP1857	*T.b. gambiense type I*	na	Central Africa	
KOLO	*T.b. gambiense type I*	na	Central Africa	
MOERBEKE-103	*T.b. gambiense type I*	na	Central Africa	
MA	*T.b. gambiense type I*	na	Central Africa	
Frala	*T.b. gambiense type I*	na	West Africa	
LiTat-1-3	*T.b. gambiense type I*	na	West Africa	
KISOU-BOBO-80-MURAZ15	*T.b. gambiense type I*	1980	West Africa	man
148BT	*T.b. gambiense type I*	2005	Central Africa	man
146BT	*T.b. gambiense type I*	2005	Central Africa	man
45BT	*T.b. gambiense type I*	2006	Central Africa	man
146AT	*T.b. gambiense type I*	2006	Central Africa	man
163AT	*T.b. gambiense type I*	2006	Central Africa	man
163AT-relapse	*T.b. gambiense type I*	2006	Central Africa	man
40AT	*T.b. gambiense type I*	2006	Central Africa	man
93AT	*T.b. gambiense type I*	2006	Central Africa	man
223AT	*T.b. gambiense type I*	2006	Central Africa	man
40BT	*T.b. gambiense type I*	2006	Central Africa	man
148AT	*T.b. gambiense type I*	2006	Central Africa	man
349BT	*T.b. gambiense type I*	2006	Central Africa	man

340AT	*T.b. gambiense type I*	2006	Central Africa	man
348BT	*T.b. gambiense type I*	2006	Central Africa	man
346BT	*T.b. gambiense type I*	2007	Central Africa	man
346AT-relapse	*T.b. gambiense type I*	2007	Central Africa	man
108AT	*T.b. gambiense type I*	2007	Central Africa	man
108BT	*T.b. gambiense type I*	2007	Central Africa	man
57AT	*T.b. gambiense type I*	2007	Central Africa	man
174AT	*T.b. gambiense type I*	2007	Central Africa	man
378BT	*T.b. gambiense type I*	2007	Central Africa	man
167AT	*T.b. gambiense type I*	2008	Central Africa	man
113BT	*T.b. gambiense type I*	2008	Central Africa	man
19BT	*T.b. gambiense type I*	2008	Central Africa	man
167BT	*T.b. gambiense type I*	2008	Central Africa	man
104BT	*T.b. gambiense type I*	2008	Central Africa	man
105BT	*T.b. gambiense type I*	2008	Central Africa	man
48BT	*T.b. gambiense type I*	2008	Central Africa	man
104AT	*T.b. gambiense type I*	2008	Central Africa	man
15BT-relapse	*T.b. gambiense type I*	2008	Central Africa	man
174BT	*T.b. gambiense type I*	2008	Central Africa	man
113AT	*T.b. gambiense type I*	2008	Central Africa	man
85BT	*T.b. gambiense type I*	2008	Central Africa	man
95BT	*T.b. gambiense type I*	2008	Central Africa	man
141BT	*T.b. gambiense type I*	2008	Central Africa	man
29BT	*T.b. gambiense type I*	2008	Central Africa	man
130BT	*T.b. gambiense type I*	2006	Central Africa	man
BRAZAVILLE-BB	*T.b. gambiense type I*	1973	Central Africa	man
AMBOU	*T.b. gambiense type I*	1980	Central Africa	man
NDMI	*T.b. gambiense type I*	1980	Central Africa	man
NKOUA	*T.b. gambiense type I*	1980	Central Africa	man
PEYA	*T.b. gambiense type I*	1980	Central Africa	man
LiTat-1-5-P9	*T.b. gambiense type I*	1952	West Africa	man
MHOM-CI-78-DALOA69	*T.b. gambiense type I*	1978	West Africa	man
MHOM-CI-78-DALOA72-cloneA	*T.b. gambiense type I*	1978	West Africa	man
MHOM-CI-79-DALOA74	*T.b. gambiense type I*	1979	West Africa	man
KOUAM-DALOA	*T.b. gambiense type I*	na	West Africa	man
ROUPO-VAVOUA--80-MURAZ-14	*T.b. gambiense type I*	1980	West Africa	man
OUSOU	*T.b. gambiense type I*	1982	West Africa	man
MHOM-CI-91-SIQUE1623	*T.b. gambiense type I*	1991	West Africa	man
BIM-AnTat-8-1-P8	*T.b. gambiense type I*	1976	Central Africa	man
Jua	*T.b. gambiense type I*	1976	Central Africa	man
BIP24	*T.b. gambiense type I*	1999	Central Africa	human
INRB-2009-56A-15BT	*T.b. gambiense type I*	2009	Central Africa	man
Mbadi	*T.b. gambiense type I*	1998	Central Africa	human
BAT1	*T.b. gambiense type I*	1999	Central Africa	human
BAT31	*T.b. gambiense type I*	1999	Central Africa	human
BAT33	*T.b. gambiense type I*	1999	Central Africa	human
MHOM-SD-82-MUSIKIA-cloneA	*T.b. gambiense type I*	1982	East Africa	man
F39UG	*T.b. gambiense type I*	1999	East Africa	human

LOGRA	*T.b. gambiense type I*	1968	Central Africa	man
MHOM-ZR-71-C126	*T.b. gambiense type I*	1971	Central Africa	man
Bosendja	*T.b. gambiense type I*	1972	Central Africa	man
LOKO	*T.b. gambiense type I*	1973	Central Africa	man
MBA	*T.b. gambiense type I*	1974	Central Africa	man
AnTat-11-17	*T.b. gambiense type I*	1974	Central Africa	man
ALJO	*T.b. gambiense type I*	1975	Central Africa	man
Nabe	*T.b. gambiense type I*	1995	Central Africa	man
Seka	*T.b. gambiense type I*	1995	Central Africa	man
Bage	*T.b. gambiense type I*	1995	Central Africa	man
Pakwa	*T.b. gambiense type I*	1995	Central Africa	man
MSUS-CI-82-TSW125-KP1-cloneB	*T.b. gambiense type I*	1982	West Africa	pig
AnTat-21-1	*T.b. gambiense type I*	na	Central Africa	
AnTat-9-1	*T.b. gambiense type I*	na	Central Africa	
KIN-1	*T.b. gambiense type I*	na	na	
LOKO-BIIT	*T.b. gambiense type I*	1973	Central Africa	human
LOKO-BIIT-P2	*T.b. gambiense type I*	1973	Central Africa	human
MHOM-SD-82-SUZENA	*T.b. gambiense type I*	1982	North Africa	man
Mongo	*T.b. gambiense type I*	na	na	
13_97D	*T.b. gambiense type I*	na	Central Africa	human
14_97D	*T.b. gambiense type I*	na	Central Africa	human
15_97D	*T.b. gambiense type I*	na	Central Africa	human
A005	*T.b. gambiense type I*	1988	Central Africa	human
BAT35	*T.b. gambiense type I*	1999	Central Africa	human
BIP20	*T.b. gambiense type I*	1999	Central Africa	human
BIP28	*T.b. gambiense type I*	1999	Central Africa	human
BIP33	*T.b. gambiense type I*	1999	Central Africa	human
BIP42	*T.b. gambiense type I*	1999	Central Africa	human
C3359	*T.b. gambiense type I*	1998	Central Africa	human
C3392	*T.b. gambiense type I*	1998	Central Africa	human
Demba	*T.b. gambiense type I*	1989	Central Africa	human
Doume1	*T.b. gambiense type I*	2001	Central Africa	human
F33	*T.b. gambiense type I*	na	na	human
F39	*T.b. gambiense type I*	na	na	human
F41	*T.b. gambiense type I*	na	na	human
F42	*T.b. gambiense type I*	na	na	human
Leontio	*T.b. gambiense type I*	1997	Central Africa	human
Mibene	*T.b. gambiense type I*	1996	Central Africa	human
Moyox2	*T.b. gambiense type I*	1998	East Africa	human
P26F	*T.b. gambiense type I*	na	Central Africa	pig
R47	*T.b. gambiense type I*	1998	East Africa	human
FEO	*T.b. gambiense type II*	1938	West Africa	man
FEO-AnTat-16-1	*T.b. gambiense type II*	1938	West Africa	man
Feo-AnTat-16-1-bis	*T.b. gambiense type II*	1938	West Africa	man
MHOM-CI-78-TH113	*T.b. gambiense type II*	1978	West Africa	man
MHOM-CI-TH162	*T.b. gambiense type II*	1978	West Africa	man
ABBA	*T.b. gambiense type II*	1983	West Africa	man
LIGO	*T.b. gambiense type II*	1984	West Africa	man

AnTat-25-1S	*T.b. gambiense type II*	1971	Central Africa	man
MHOM-CI-78-TH1-037	*T.b. gambiense type II*	na	West Africa	man
MHOM-CI-78-TH2	*T.b. gambiense type II*	1978	West Africa	man
MHOM-CI-79-MURAZ3	*T.b. gambiense type II*	1979	West Africa	man
Muraz3	*T.b. gambiense type II*	1979	West Africa	human
P8F	*T.b. gambiense type II*	na	Central Africa	human
Lister-427	*T.b. rhodesiense*	1956	East Africa	cattle
MBOI-ZM-82-TRPZ260	*T.b. rhodesiense*	1982	Southern Africa	cattle
MCAP-ZM-83-TRPZ267	*T.b. rhodesiense*	1983	Southern Africa	goat
ITM001	*T.b. rhodesiense*	2015	East Africa	man
MHOM-KE-81-LVH122	*T.b. rhodesiense*	1981	East Africa	man
MHOM-KE-82-LVH61R	*T.b. rhodesiense*	1982	East Africa	man
Rumphi	*T.b. rhodesiense*	2007	Southern Africa	man
AnTat-12-1S	*T.b. rhodesiense*	1971	East Africa	man
MHOM-SD-82-BIYAMINA	*T.b. rhodesiense*	1982	North Africa	man
MHOM-SD-82-BIYAMINA-bis	*T.b. rhodesiense*	1982	North Africa	man
MHOM-UG-77-KETRI-2355	*T.b. rhodesiense*	1977	East Africa	man
STIB847	*T.b. rhodesiense*	1990	East Africa	man
STIB882	*T.b. rhodesiense*	1993	East Africa	man
STIB883	*T.b. rhodesiense*	1994	East Africa	man
MHOM-ZM-80-TRPZ-23	*T.b. rhodesiense*	1980	Southern Africa	man
MHOM-ZM-82-TRPZ186	*T.b. rhodesiense*	1982	Southern Africa	man
MHOM-ZM-83-TRPZ-349	*T.b. rhodesiense*	1983	Southern Africa	man
STIB-851	*T.b. rhodesiense*	1990	East Africa	man
LOPEZ	*T.b. rhodesiense*	na	na	
MHOM-ET-67-GAMBELA1	*T.b. rhodesiense*	1967	East Africa	man
MHOM-ET-69-GAMBELA3	*T.b. rhodesiense*	1969	East Africa	man
MHOM-KE-82-LVH61R-bis	*T.b. rhodesiense*	1982	East Africa	man
TRPZ26	*T.b. rhodesiense*	na	na	

2.2.1 RNA extraction from *T. b. gambiense*

Our collaborators at the Institute of Tropical Medicine (ITM; Philippe Büscher, Jan Van Den Abbeele, Nick Van Reet, Frederick Van den Broeck), Antwerp, supplied two biological replicates of procyclic form *T. b. gambiense* type 1 isolate Mongo isolated with the KIVI method [301] in 500 µl RNAlater (ThermoFisher). The total number of cells in each sample was not provided. We isolated total cellular RNA using a standard phenol-chloroform extraction and ethanol precipitation procedure described below:

The centrifuge temperature was kept at 4 °C. After thawing on ice for 6 min, the cell pellet was mixed with pre-chilled TE buffer and centrifuged at 7,000 x *g*-force for 10 minutes. The process was repeated on the supernatant to maximize cell pellet collection. TRIzol reagent (1 ml; ThermoFisher) was mixed with the cells at room temperature for 5 min, followed by mixing 200 µl chloroform isoamyl alcohol by vigorous shaking and a 3-minute incubation at room temperature. After centrifuging at 12,000 x *g*-force for 15 minutes, the 500 µl aqueous

phase was transferred to a new tube and mixed with 500 µl isopropanol, then incubated at room temperature for 10 minutes, followed by centrifugation at 12,000 x *g*-force for 10 minutes. After the supernatant was removed, the pellet was resuspended in 1 ml 70% ethanol, vortexed for 20 seconds, and centrifuged at 7,500 x *g*-force for 10 minutes. The process was repeated before the supernatant was removed for the pellet to air dry for 5 minutes.

2.2.2 Next-generation RNA sequencing data generation (RNA-seq)

T. b. gambiense type 1 Mongo total cellular RNA (2.2.1.) was dissolved in 50 µl DEPC-treated water and evaluated with a NanoDrop spectrophotometer (ThermoFisher). Both samples passed the NanoDrop examination and were evaluated using a BioAnalyser 2100 (Agilent) for quality. The validated samples were Illumina-sequenced at Novogene using a standard mRNA library preparation procedure with poly-A enrichment. We submitted the minimal required volume of 10 µl (≥2000ng RNA) for each sample. Pair-end reads (150 nt) were generated, and 43,965,840 and 49,710,510 reads passed the quality filter for each replicate, respectively. Reads from the two replicates were combined for subsequent analysis.

ITM also provided whole-cell transcriptomic data for IL3000. RNA extracted from isolated BSF and epimastigote IL3000 were separated with the MirVana small RNA kit (ThermoFisher) into the larger RNAs (> 200 nt) and the small RNAs (sRNAs; < 200 nt). The BSF larger RNAs were sequenced with DNBSEQ (MGI), while the BSF sRNAs were sequenced with Illumina (150 nt). Both BSF and epimastigotes larger RNAs were also used to generate Iso-Seq SMRT reads (PacBio), yielding 82,044 and 95,524 unique reads, respectively.

C. fasciculata (strain CfC1 [302]) kDNA, purchased from Inspiralis, at 100 ng/µl in TE buffer (10 mM TrisHCl pH7.5, 1 mM EDTA) was sent for Illumina sequencing at Novogene. After Microbial Whole Genome Library Preparation (350 nt fragmentation), libraries were pooled for sequencing. Pair-end reads (150 nt) were generated and about 4.7 million reads passed the quality filter.

2.2.3 Whole-genome DNA sequencing data from miscellaneous sources

2.2.3.1 *T. brucei ssp.*

ITM provided whole-genome sequencing (WGS) data (BGI platform, 150 nt) of 262 *T. brucei* field isolates and laboratory strains, including 224 sub-Saharan isolates of tsetse-transmissible *T. brucei* subspecies (*T. b. gambiense* type 1 or 2; *T. b. rhodesiense*; *T. b. brucei*) (Table 2-3). The 38 samples of non-tsetse transmitted subspecies included 27 and 6 samples that each contained a single *T. b. evansi* and *T. b. equiperdum* cell line, respectively. In comparison, five samples probably contained multiple cell lines and were removed for the analysis. Reads mapped to reference *T. brucei* nuclear genome were removed for tsetse-dependent isolates, leaving the reads presumably originating from kDNA (see Supplementary Table.1 for all metadata).

The same WGS data were used in other published studies, in which the DNA preparation procedure was described [85, 93]. Purified trypanosomes were sedimented by centrifugation (3,000 × *g*, 10 min at 4° C) and DNA was extracted by standard phenol-

chloroform [85]. The concentration of extracted DNA was determined using a Qubit 4 Fluorometer (Invitrogen by Thermo Fisher Scientific) [85]. Paired end 150 bp sequences were generated using the DNA nanoball sequencing technology (DNBSEQ™) at the Beijing Genomics Institute (BGI) [85]. The number of cells used in the extractions was not provided.

The sub-Saharan *T. brucei* isolates were classified after *de novo* assembly of the nuclear genome of each *T. brucei* strain using Megahit [303]. The contigs were searched using blastn for marker genes diagnostic of *T. b. gambiense* type 1 and *T. b. rhodesiense*. We used two SRA markers (SRAsouth (AJ345058.1), SRAnorth (AJ345057.1)) and two TbgsGP markers (TgsGP_1 (AJ277951.1), TgsGP_2 (FN555988.1)) for the classification. The five supplementary *T. b. gambiense* type 1 markers (LiTat1.3_1 (AJ304413.1), LiTat1.3_2 (KJ499460.1), LiTat1.5_1 (EU257624.1), LiTat1.5_2 (HQ662603.1), VSG117 (S62479.1)) had low specificity and did not provide convincing guidance on *T. b. gambiense* type 1 classification.

Blastn hits with complete marker gene coverage were recorded for each isolate. We considered isolates with hits on TgsGP or SRA gene markers to be *T. b. gambiense* type 1 and *T. b. rhodesiense*, respectively. Isolates without hits on any marker genes yet extracted from humans were considered unknown human infective *T. brucei* (*T. b. gambiense type 2*), which probably included *T. b. gambiense type 2*, while the remaining isolates were labeled as *T. b. brucei*. Four isolates did not contain TgsGP hits but were considered *T. b. gambiense* type 1 based on conservation of minicircle populations [80].

Table 2-3. **Classification of sub-Saharan *T. brucei* isolates**

Subspecies	Total	Marker
T. b. gambiense type 1	107	TgsGP+
T. b. gambiense type 1	4	TgsGP-, minicircle population conservation
T. b. brucei UHI	13	TgsGP-, SRA-, extracted from human
T. b. rhodesiense	23	SRA+
T. b. brucei	77	TgsGP-, SRA-, extracted from animals or unknown hosts
Total	224	

2.2.3.2 *T. congolense*

ITM provided annotated maxicircle of the reference *T. congolense* strain IL3000 and Illumina sequencing data (MiSeq) of isolated kDNA from three Savannah subgroup *T. congolense* strains (IL3000, Kapeya, UPKZN). IL3000 was isolated in 1966 from a cow in Kenya [304, 305] and cultured in the laboratory ever since, while strains Kapeya and UPKZN represent natural trypanosome isolates, collected from a cow in 2003 in Zambia [306] and from a buffalo in South-Africa [307], respectively.

2.2.3.3 *Leishmania*

The Sacks laboratory (NIH, Bethesda) provided WGS data of *L. major* parental strains and their intraspecific hybrids for kDNA assembly.

2.3 Assembly of kDNA genomes

2.3.1 Data processing and quality assessment

Prior to kDNA assembly, to remove sequencing reads corresponding to the nuclear genome from WGS data, reads were aligned against the reference nuclear chromosomes of respective (sub)species using Bowtie 2 v2.3.5.1 --very-sensitive-local option [308] (Table *2-4*). Reads not aligned to the nuclear genome were extracted and converted to FASTQ format using SAMtools v1.10 [309].

Read quality assessment and adaptor sequence detection were performed with FASTQC [310]. To prepare the reads for assembly, the unmapped reads were filtered for high quality with fastp v 0.21.0 [311] using the following parameters: allow for ≤ 10% bases per read to have a phred-scaled quality ≤ 20, keep reads with a minimum length of 100 nt, and trim 3' tail if the read length exceeds a maximal length of 150 nt. The requirement of read length was lifted when processing shorter reads. The resulting high-quality 'denucleated' reads were used for assembly and minicircle detection.

Table 2-4**. Reference genome used for removing reads aligned to the nuclear genomes**

Species	Reference strain	Publication
Trypanosoma brucei brucei	TREU927 v4.6	[312]
Trypanosoma congolense	IL3000	[313]
Leishmania major	Friedlin 2021	[314]
Leishmania infantum	JPCM5	[315]
Leishmania braziliensis	MHOM/BR/75/M2904 2019	[316, 317]
Leishmania tropica	L590	[318]
Leishmania donovani	NTK282A1	[319]

2.3.2 Maxicircle assembly

Candidate maxicircle contigs were generated with KOMICS v1.8 [66] using MEGAHIT NGS assembler [303] using a series of kmer sizes from 29 to 119 with an increment of 20 nt. Maxicircle sequences were selected from the contigs by megablast [320] using the published maxicircle fragment of *Leishmania braziliensis* (GenBank M2904) as a query. Attempts to recover the variable region and circularize the maxicircles were unsuccessful.

Contigs of *T. brucei* and *T. congolense* isolates were aligned to the *T. b. brucei* EATRO1125 reference (GenBank MK584625.1) [225] for concatenation and reorientation to start with the 12S rRNA gene, while *Leishmania braziliensis* (M2904) included in the KOMICS program was used for *Leishmania* and *Crithidia* maxicircles. We visually assessed the completeness of the maxicircle assemblies using sequence alignment to references and kept contigs with complete coding regions for further analysis. Illumina read coverage was checked for each selected contig and no gaps in read coverage were detected. The maxicircle alignments also identified deletions, which were confirmed by mapping the denucleated reads to the reference EATRO1125 maxicircle coding region (Genbank accession: MK584625) using Bowtie 2 v2.3.5.1 --very-sensitive option. Regions of over 10 nucleotides with read depth ≤ 5 were examined manually.

The coding regions were annotated by sequence alignment to the respective annotated reference genome using Blastn and custom Python scripts. Conservation of sequence and synteny over the coding region was observed for all maxicircle assemblies.

2.3.3 Minicircle genome assembly

We performed primary assembly using MEGAHIT via KOMICS v1.8 with various kmer sizes to maximize minicircle capture (Table *2-5*). Contigs that contained the universally conserved CSB-3 12-mer (GGGGTTGGTGTA) or its reverse complement, with one allowed mismatch at any position to capture potential sequence variations, were tested for circularity, by searching for overlapping sequences at either end of each contig.

Table 2-5 **Read sources for minicircle assembly and choice of kmer sizes**

Species	Read length	Kmer size
Trypanosoma brucei ssp	150	99, 119, 129, 149
Trypanosoma congolense	150	99, 119, 129, 149
Leishmania major	100	49, 79, 89, 99
Crithidia fasciculata	150	99, 119, 129, 149

The overlaps at the start of circularized contigs were removed upon confirmation of circularity. *Trypanosoma* and *Leishmania* have a single replication origin per minicircle. Their circular contigs were converted to template strain and reoriented to start at CSB-1. *Crithidia* minicircles contain two replication origins and were instead reoriented to start at a conserved motif in the singular bend region. The set of circular contigs for each isolate was clustered with ≥ 95% sequence identity (SID) usearch vsearch –usearch_global [321] (including the conserved region) to obtain the set of unique minicircle contigs with each contig being a unique minicircle classes. Hence, a minicircle class has ≤ 95% SID compared to any other classes.

To search for alternative CSB-3 sequences that differed by more than one mismatch from the sequences published for *Leishmania* and *Crithidia*, a consensus was derived from the CSB-containing conserved region via minicircle alignment. The ~100-nt conserved sequence was used as a query to detect probable minicircles with more divergent CSB motifs in the contigs that did not contain the canonical CSB-3.

To detect minicircles that were not captured during *de novo* assembly or present as non-circularized fragments (for example because they had been present at very low abundance in the samples), we pooled all assembled minicircles from the same species and clustered them at 95% SID by vsearch –usearch_global [321]. High-quality ‘denucleated’ reads of each isolate were mapped globally to the pooled minicircle set of the corresponding species using Bowtie 2 v2.3.5.1 --very-sensitive option. The mapped reads were filtered allowing two mismatches at most. Minicircles with over 99% read coverage were visually examined for homogenous SNPs in IGV [322], corrected as necessary, and sequences were added to the set of the respective isolate after.

2.3.4 Completeness assessment

The percentage of mapped reads and the percentage of mapped CSB-3-containing reads were used to estimate the completeness of each minicircle assembly. The conserved region containing replication origin spans around 100 nt from CSB-1 and extends for less than 20 nt after CSB-3 [19, 225, 230]. Given a 100 -or 150-nt read length, the majority of CSB3-containing reads will extend into the non-conserved region unique to each minicircle. In addition, the 120 nt conserved regions are not perfectly identical among minicircles, and the variations add to the specificity of the read mapping [225]. On the other hand, the 12-mer CSB-3 motif is absent from nuclear DNA [225]. Therefore, reads containing the CSB-3 motif are specific for each minicircle class and can be considered markers for minicircle DNA. Hence, we considered the percentage of mapped CSB-3 containing reads (PMC) as a direct measure of its completeness. We expected almost all CSB3-containing reads to be mapped if the assembly faithfully reflected the kDNA complexity.

2.3.5 Estimation of minicircle copy numbers per network

To estimate the coverage of each minicircle, minicircle sequences were first extended by adding the last 150 nt from the 3' end to the 5' end (i.e. immediately before CSB1), so reads that spanned both ends could be accurately mapped to the linearized sequences.

In case of WGS data, we mapped the reads to (1) a reference nuclear genome consisting of the central region (40% to 60% of total length) of the 11 nuclear mega-basepair chromosomes, (2) the coding region of the maxicircle from *de novo* assembly if it had been assembled for the isolate/strain in question, and (3) the minicircle sequences from the isolate-specific *de novo* assembly with 150 nt extensions. The average read depths of (1) core nuclear genome, (2) maxicircle coding region, and (3) nt 400 to nt 600 on each minicircle were calculated using samtools coverage and samtools depth [323]. Assuming that most cells in the original sample were diploid, the average read depths of maxicircles and minicircles were normalized against the read depth of the core nuclear chromosomes.

If the sequenced sample represented isolated kDNA, i.e. the sequencing dataset contained only kDNA reads, the maxicircle copy number was assumed to be 30 per kDNA network [225]. The average read depths of minicircles were then normalized against the read depth of maxicircles for copy number estimation. Note that the assumption of 30 maxicircles per network might under or overestimate the absolute copy minicircle copy number, but this would not affect the calculation of relative abundance within each network.

2.4 kDNA annotation

2.4.1 Identification of conserved sequence motifs

The minicircles from each assembly were aligned to compare the conserved regions that contained the replication origin and the bent region. Within the origin of replication, CSBs were characterized using WebLogo [324].

Minicircles of *T. brucei* and *T. congolense* contain cassettes framed and defined by the imperfectly conserved 18-nt inverted repeats [81, 228]. The forward and reverse repeats were visually identified and extracted from alignments to generate consensus sequences using Weblogo [324].

For *Leishmania* and *Crithidia* minicircles, which are reported to lack inverted repeats, 50 nt upstream and downstream of predicted gRNAs for all minicircles were input to MEME [325] with default parameters. No inverted repeats flanking the gRNAs were detected [325].

2.4.2 Edited mRNA predictions

Edited sequences for *T. congolense* and *T. brucei* subspecies under investigation in this study could deviate from published sequences. The mitochondrial mRNAs were represented poorly in the mRNA data, and the fully edited mRNAs only accounted for a small proportion of the steady-state mitochondrial mRNA populations. Consequently, attempts to reliably identify fully edited mRNA sequences using whole cell transcriptomic data, where available, and T-aligner [326] were unsuccessful for pan-edited mRNAs. Exceptions with good coverage occurred near the 3' end of edited regions, due to the 3' to 5' directionality of editing. Nonetheless, in combination with an alternative procedure that relied on sequence homology of closely related species (elaborated in the following sections), the availability of transcriptomic data did improve our ability to predict fully edited mRNA sequences. In cases where no transcriptomic data were available, such as *T. b. equiperdum* type OVI, we exclusively utilized this alternative procedure.

When transcriptomic data were available, we detected alternatively edited sequences when multiple consensuses of fully edited sequences were observed. We recorded all the editing patterns that allowed ORFs to be created and searched for the gRNAs subsequently. Admittedly, this method ignored alternative editing that resulted in drastically different sequences, which probably did not exist.

2.4.2.1 *Prediction of T. congolense edited mRNAs*

T. congolense fully edited sequences were predicted in two steps. First, by careful comparison of the *T. congolense* maxicircle-encoded pre-edited mRNA sequences with published pre-edited and edited mRNA sequences from *T. brucei*, and second by polishing of these sequences with available transcriptome data for *T. congolense* IL3000 (see section 2.2.2).

In the first step, *T. congolense* pre-edited mRNA sequences were extracted from the annotated IL3000 maxicircle provided by ITM, Antwerp. The transcriptomic reads were mapped to the reference nuclear genome (Table *2-4*) using Bowtie2 --very-sensitive-local and the unmapped reads were extracted and merged in a single file (unpaired). The pre-

edited mRNAs were validated by mapping denucleated Illumina reads using Bowtie2 --very-sensitive-local.

The non-U bases of EATRO1125 and IL3000 pre-edited mRNAs were aligned. The U-indels over the IL3000 edited mRNAs were inferred from the editing patterns over the EATRO1125 edited mRNAs and added manually.

SNPs were altered at their corresponding locations. The majority of indels involved U-indels. As frameshift mutations may render the protein inactive, indels were initially corrected by adding or removing uridines to maximize the conservation of ORFs and the protein sequences following the same rationale as described in [73].

- If a U insertion in the IL3000 pre-edited mRNA was in the same location as an otherwise post-transcriptional U insertion in EATRO1125, we assumed that the insertion negated the need for an additional post-transcriptional insertion.
- If not, it was assumed to be post-transcriptionally deleted.
- For U deletions, no modification to the editing pattern was made.
- For non-U indels, to avoid a frameshift, the closest U-insertion was removed or extended to minimize alteration of the resulting protein

In the second step, the preliminary predictions were polished with IL3000 transcriptomic data. Long (150 bp) Illumina reads were mapped with Bowtie 2 v2.3.5.1 --very-sensitive-local option. SNPs and U-indels were identified from the mapping consensus. Pairs of single U polymorphisms were due to incorrect insertion positions and could be resolved by relocating the U-insertions. U-indels indicated wrong numbers of uridines between non-U residues and often occurred in heavily edited regions. The U-indels were corrected by carefully rearranging the U-indels between non-U residues so they agreed with the mapping consensus while retaining the ORFs. Regions without transcriptome coverage were not modified. Reads were mapped to the corrected mRNAs and the same procedures were repeated until all discrepancies were accounted for.

To validate the editing patterns over regions without Illumina coverage and to detect alternative editing, the polished edited mRNAs and pre-edited mRNAs were used as queries to search against the condensed PacBio SMRT reads of BSF and epimastigotes IL3000 by blastn. Hits to the edited mRNAs were extracted for alignment using mafft and careful manual examination [327].

2.4.2.2 *Prediction of T. b. gambiense* type 1 *edited mRNAs*

T. b. gambiense type 1 edited mRNAs were modified from the *T. b. brucei* EATRO1125 edited mRNAs [225]. First, the EATRO1125 pre-edited mRNAs were aligned to the *T. b. gambiense* type 1 maxicircle of the isolate Mongo from the *de novo* assembly to extract pre-edited Mongo mRNAs. Preliminary edited mRNA predictions were performed as described for *T. congolense*. Next, transcriptome reads mapping to procyclic *T. b. gambiense* type 1 Mongo (see section 2.2.1) were used to curate the predicted mRNA sequences as described for *T. congolense*.

No differences in non-T residues between EATRO1125 and Mongo were detected. Consequently, the polishing only modified U-indel patterns.

2.4.2.3 *Prediction of T. b. rhodesiense edited mRNAs*

Mapping published *T. b. rhodesiense* transcriptome data (NCBI BioProject accession number PRJEB23278) to pre-edited and edited mRNAs of *T. b. brucei* EATRO1125 revealed no differences in non-T nucleotides or editing sites. Therefore, the EATRO1125 mRNAs were used for *T. b. rhodesiense* gRNA annotation.

2.4.2.4 *Prediction of T. b. equiperdum and T. b. evansi edited mRNAs*

Transcriptomic data were not available for the kDNA-independent subspecies *T. b. equiperdum* and *T. b. evansi*. The pre-edited mRNAs and the preliminary edited mRNA predictions for *T. b. equiperdum* type OVI were performed as described in *T. congolense* using *T. b. brucei* EATRO1125 mRNAs and the *de novo* assembled maxicircle of isolate OVI.

sometimes the methods described here were complemented by using minicircle homology to look for homologous gRNAs that were identified in some isolates/subspecies but potentially missed in others. If new gRNA alignment could be obtained from U-indel modifications while preserving the ORF and protein sequence, the edited mRNAs would be corrected accordingly using the gRNA as a template.

2.4.2.5 *Prediction of Leishmania major edited mRNAs*

The *de novo* assembled *L. major* maxicircle was annotated by sequence homology to the published pre-edited mRNAs of *L. peruviana* HR78 and *L. braziliensis* LC1412 [66]. The *L. major* pre-edited mRNAs were extracted and validated with published *L. major* strain 'Ryan' transcriptome data [291]. Preliminary prediction of edited mRNAs was performed as described for *T. congolense* using *L. peruviana* HR78 edited mRNAs as a template. The polishing used *L. major* strain 'Ryan' whole cell transcriptome data and was performed as described for *T. congolense.*

2.4.2.6 *Prediction of C. fasciculata edited A6 and RPS12 mRNAs*

The *de novo* assembled *C. fasciculata* strain CfC1 maxicircle was aligned to the annotated *L. major* maxicircle (see 2.4.2.6) for annotation. The annotation was validated by complete read coverage of publicly available transcriptome data from adherent and swimming forms of *C. fasciculata* from infected mosquito hindguts or *in vitro* culture [292].

T-masked alignment of transcriptome data against the CfC1 pre-edited mRNAs by T-Aligner [326] and mapping with Bowtie2 against the published A6 and uS12m (RPS12) edited mRNA sequences [31] confirmed these editing patterns.

2.4.3 Guide RNA predictions

2.4.3.1 *T. congolense*

Having assembled the minicircles and maxicircle, and generated pre-edited and edited mRNA sequences of IL3000, gRNAs were annotated on the minicircles with a published kDNA annotation pipeline [225]. Detailed instructions and examples for the pipeline are given at https://github.com/nicksavill/kDNA-annotation.

Allowing G-U wobble base-pairing, all alignments between the kDNA and the edited mRNA sequences ≥ 25nt were detected. High-confidence gRNAs are defined as alignments ≥ 40 nt with at most one mismatch and anchor length ≥ 8 nt with Watson-Crick base-pairing. Gaps in the alignment were not permitted.

The high-confidence cassettes that contained the high-confidence gRNAs plus flanking sequences 63 nt upstream and 100 nt downstream of the gRNA 5' end were extracted to characterize cassette structures with the MEME motif searcher [325]. The forward inverted repeat, the reverse inverted repeat, and the initiation sequence were detected. The 5' to 3' order of the three motifs was expected to be: forward inverted repeat, initiation sequence, and reverse inverted repeat. The relative positions of the motifs were visually examined, and sequences with motifs in the wrong order were removed before motif extraction.

The nucleotide frequencies of the initiation sequence motifs of HQ gRNAs found by Meme allowed us to predict the position of the initiation sequences relative to the 3' end of the forward repeat in the absence of sRNA data. The nucleotide bias of the repeats calculated from the high-confidence cassettes enabled us to identify the cassettes on all minicircles, with an expected cassette number of three. The cassettes were labeled based on the acceptable ranges of the forward repeat starting positions (Table *2-6*). Minicircles with conflicting cassette labels were checked for false positives. The ranges were adjusted for individual annotation until all accepted as confirmed cassettes received a label.

Table 2-6**. Cassette positions used in *T. brucei* and *T. congolense* minicircle annotation**

species	Cassette I		Cassette II		Cassette III		Cassette IV		Cassette V	
T. congolense	150	300	300	500	596	800				
T. b. gambiense type 1	100	280	230	480	400	720	596	1000	1001	2000
Other *T. brucei*	100	259	230	409	395	469	470	703	646	2000
T. b. equiperdum	100	259	260	409	410	469	470	703	652	2000

Following cassette assignments, we added to the set of high-confidence gRNAs by identifying all canonical gRNAs in cassettes and orphan gRNAs outside of cassettes, with length ≥ 25 nt, anchor ≥ 6 nt, and at most 3 mismatches. For cassettes that contained multiple candidate gRNAs, at least one gRNA was over 35 nt.

Kapeya and UPKZN minicircles were annotated with IL3000 mRNAs using the same parameters and procedures.

A SAM file of IL3000 sRNAs mapping on minicircles was generated with Bowtie 2 v2.3.5.1 --very-sensitive-local option. The true initiation site, gRNA gene end positions, and the expression levels of gRNAs were inferred from the SAM file using the same criteria as in a

previous study of *T. b. brucei* [225]. This was only performed with *T. congolense* IL3000, as sRNA transcriptome data were unavailable for other isolates.

2.4.3.2 *T. b. gambiense* type 1

Minicircles were pooled from all *T. b. gambiense* type 1 isolates and clustered at 95% SID to be annotated in bulk. The annotation was performed as described for *T. congolense* using the *de novo* assembled maxicircle and the pre-edited and edited mRNAs of the *T. b. gambiense* type 1 isolate Mongo. The high-confidence cassettes containing the high-confidence gRNAs, plus flanking sequences 70 nt upstream and 100 nt downstream of the gRNA 5' end, were extracted. The minimal minicircle length and the expected number of cassettes were set to 25 nt and four, respectively. The ranges of the forward repeat starting positions for labeling cassettes are listed in Table *2-6*.

2.4.3.3 *Other T. brucei isolates*

Minicircles were pooled from 224 *T. brucei* isolates, including *T. b. gambiense* type 1, and clustered at 95% SID to be annotated in bulk. The annotation was performed as described for *T. congolense* using the published maxicircle and pre-edited and edited mRNAs of *T. b. brucei* EATRO1125 [225]. With over 6000 unique minicircle classes, due to limitations of the annotation pipeline, the annotation had to be performed in batches of 500 minicircles.

The parameters for high-confidence cassettes and gRNAs were identical to *T. b. gambiense* type 1. The ranges of the forward repeat starting positions for labeling cassettes are listed in Table *2-6*.

For minicircles present in *T. b. gambiense* type 1, the annotations using EATRO1125 mRNAs were ignored. Only the annotation using the Mongo mRNAs specific to *T. b. gambiense* type 1 in section 2.4.2.2 was used in the subsequent analysis.

2.4.3.4 *T. b. equiperdum type OVI*

Annotation of minicircles of three type OVI isolates, OVI, Te-Ap-ND1, and Dodola940, were performed as described for *T. congolense*, using the *de novo* assembled OVI maxicircle and the edited mRNA sequences, predicted as described above. The parameters for high-confidence cassettes and gRNAs were identical to *T. b. gambiense* type 1. The ranges of the forward repeat starting positions for labeling cassettes are listed in Table *2-6*. In addition, gRNAs identified via minicircle homologs were included as described in mRNA editing.

2.4.3.5 *L. major*

Leishmania gRNA annotation was conducted with a modified version of the aforementioned kDNA annotation pipeline, due to the lack of cassettes flanked by inverted repeats and different gRNA aligning features. Both G-U wobble base-pairing and Watson-Crick base-pairing were allowed over the anchor regions.

The pipeline found alignments between the kDNA and the edited mRNA sequences ≥ 25 nt for minicircles and ≥ 31 nt for maxicircles. *Leishmania* minicircles encode one gRNA within a conserved distance from the replication origin, as described previously [216]. The region was identified to span from 380 to 600 nt downstream of CSB-1 using high-confidence gRNAs with length ≥ 31 nt and one mismatch. We then filtered for alignments ≥ 25 nt with ≤ 3 mismatches within the expected gRNA-encoding region to identify potential gRNA matches of lower confidence. The candidates aligned to multiple distinct editing sites were retained as potential examples of gRNAs with alternative editing activities.

2.4.3.6 *C. fasciculata*

C. fasciculata strain CfC1 gRNA prediction was performed with a modified version of the kDNA annotation pipeline, as described for *L. major. C. fasciculata* minicircles encode one gRNA within a conserved distance from the replication origins, as described previously [31]. After identifying high-confidence gRNAs (≥ 31 nt), the range of the gRNA-encoding region was set from 1583 to 2100 nt. We then searched for alignments ≥ 26 nt with ≤ 3 mismatches within the expected gRNA-encoding region.

The published gRNAs on annotated minicircle fragments [31] were also detected in our annotation pipeline. Restriction enzyme cutting sites on *de novo* assembled *C. fasciculata* minicircles and maxicircle were predicted for enzymes available from New England Biolabs using regular expression via a custom Python script to facilitate enzyme choice.

2.5 Phylogenetic analyses

2.5.1 Phylogeny based on specific minicircle families

We were interested in the evolutionary history of minicircles with identical or similar cassette families. Type A minicircles from *T. b. evansi* type A isolates and homologs with ≥ 85% SID detected from the sub-Saharan *T. brucei* isolates were pooled and aligned with mafft [327]. Minicircles from selected families were aligned similarly. Maximum likelihood phylogenies of homologous minicircles were constructed as described for maxicircles. ModelFinder identified K3Pu+F+G4 as the best-fitting nucleotide substitution models that minimize the Bayesian information criterion (BIC) score [328]. The certainty of branching was obtained with 1000x bootstrap replications.

2.5.2 Phylogeny of minicircle family populations

After identifying minicircle families, we constructed a phylogeny based on the entire *T. brucei* minicircle population (excluding *T. b. evansi* and *T. b. equiperdum* isolates). First, the complete set of minicircle families from the 224 isolates was compiled and a matrix of all families versus the 224 isolates was calculated. For an isolate, each *T. brucei* minicircle family was used as a query to compare to the minicircle family set of the population to populate the matrix by the following criteria:

- For each query, minicircle families with identical cassette families or empty/non-canonical cassettes across all positions were considered ‘related’.
- Among related minicircle families, if a minicircle family contained identical cassette families as the query, it received the label ‘present’.
- Minicircle families unrelated to any of the queries were considered absent.

A Discrete Morphological Model [329-331] was used to the phylogeny. The entry for each isolate was translated into a morphological sequence using the following dictionary:

- Present: P
- Related: R
- Absent: -

We constructed the phylogeny using IQtree v 1.6.12 [332] and the –st MORPH argument to specify the morphological substitution model. We found the best fitting substitution models that minimize the Bayesian information criterion (BIC) score to be MK+FQ+R3, using ModelFInder [328]. The certainty of branching was obtained with 1000x bootstrap replications.

2.5.3 Phylogeny based on the maxicircle

The denucleated reads from the 224 isolates from the tsetse-transmissible *T. brucei* subspecies and the 12 *T. b. evansi* and *T. b. equiperdum* isolates that contained maxicircles were mapped against the published EATRO1125 maxicircle (Genbank accession: MK584625) using Bowtie 2 v2.3.5.1 --very-sensitive. SNPs including indels were called from the BAM files using multi-way pileup using the -mpileup command from bcftools 1.10.2 [323]. The allele quality score (QUAL) and RMS mapping quality (MQ) were visualized, and QUAL ≥ 400

and MQ ≥ 40 were chosen as criteria for SNP filtering. The SNPs were visualized with custom Python scripts to facilitate understanding of the phylogeny tree.

The SNPs were extracted from the vcf file and converted to a fasta file using a custom Python script as input for IQtree. ModelFinder identified the best-fitting nucleotide substitution models (TPM3u+F) that minimize the BIC score [328]. The maximum likelihood phylogenies were constructed IQtree v 1.6.12 with 1000x bootstrap replications to obtain the certainty of branching.

3 kDNA streamlining in clonal tsetse-transmissible subspecies *T. b. gambiense* type 1

In this project, we investigated the editing capacity of the strictly clonal subspecies, *T. b. gambiense* type 1 and the conservation of editing patterns in *T. brucei*. We characterized the kDNA composition of 224 sub-Saharan *T. brucei* isolates by *de novo* assembly of the mitochondrial genome. We annotated the Mongo maxicircle via alignment to the annotated EATRO1125 maxicircle and extracted unedited mRNAs. To examine the clonal type 1 editing capacity in detail, we sequenced mRNAs from *T. b. gambiense* type 1 isolate Mongo to predict the fully edited versions of the edited mRNAs.

Using the predicted edited mRNAs, we annotated the Mongo minicircles, and then all the minicircles pooled from *T. b. gambiense* type 1. Isolates from other subspecies were annotated with published EATRO1125 edited mRNA [225]. We compared the percentage of covered editing sites of edited mRNAs among the *T. b. brucei, T. b. rhodesiense*, *T. b. gambiense* type 2, and *T. b. gambiense* type 1 to reveal differences in editing capacity probably associated with the lack of sexual reproduction.

To investigate the conservation of editing blocks, we introduced the concept of gRNA families. We group gRNAs into families based on the conserved initiation sequence starting positions (ISSPs). We inferred the range of editing blocks as the mean coverage of the guiding sequences of gRNAs from the same family. After characterizing the editing block patterns, we compared the end positions of the overlapping editing blocks to examine their conservation among the four groups.

3.1 Isolate annotation

Our collaborators at ITM queried SRA and TgsGP markers against the *de novo* assembled nuclear genome of each of the 224 field isolates from sub-Saharan Africa (Supplementary Table.2). Using the TgsGP and SRA markers, we identified 107 *T. b. gambiense* type 1 isolates and 23 *T. b. rhodesiense* isolates (Table *3-1*). For each isolate, all hits on the markers had at least 99% SID. The average read depths were between 38 and 236 on SRA markers of *T. b. rhodesiense* and between 25 and 214 on TgsGP markers of *T. b. gambiense*. Among the remaining isolates without TgsGP or SRA markers, 16 West and Central African samples from humans presumably exhibited human serum resistance via unknown mechanisms and were classified as *T. b. gambiense* type 2, Jamonneau et al. [333]. The remaining 78 isolates were labeled as *T. b. brucei.*

Table 3-1. **Classification of sub-Saharan *T. brucei* isolates**

Subspecies	Total	Marker
T. b. gambiense type 1	107	TgsGP+
T. b. gambiense type 2	16	TgsGP-, SRA-, extracted from human
T. b. rhodesiense	23	SRA+
T. b. brucei	78	TgsGP-, SRA-, extracted from animals or unknown hosts
Total	224	

T. b. gambiense type 1 included isolates collected between 1968 and 2009. As expected, most isolates came from Central (84) or West (12) Africa, while three came from Uganda (East Africa), one from South Sudan (East Africa), and one from Sudan (North Africa). Although both *T. b. gambiense* type 1 and *T. b. rhodesiense* have been reported in Uganda, they do not overlap. Furthermore, *T. b. gambiense* type 1 found mostly in West and Central Africa [334, 335], with an additional focus in Northern Ugada or South Sudan [336]. In addition, 51 isolates came from the Democratic Republic of the Congo (DRC). Most isolates with host records were extracted from humans (94) with one exception from a pig.

A closer examination of four isolates suggested different taxon annotations. LOKO, LOKO-BIIT, and LOKO-BIIT-P2 were isolated from humans in DRC , and P26F was isolated from a pig in Cameroon. The four isolates had either TgsGP or SRA markers. Despite the single-copy TgsGP gene being the gold standard for detecting *T. b. gambiense* type 1, the analytical sensitivity of tests based on TgsGP markers is limited because they target a hemizygous single-copy gene [337]. The minicircle compositions, which will be explained in later sections, of the four isolates were highly similar to *T. b. gambiense* type 1 and could be identified as *T. b. gambiense* type 1 based on the presence of conserved unique minicircles, as demonstrated by the novel diagnostic assay [80]. The geographical distributions did not contradict the annotations. Consequently, we reannotated the four isolates as *T. b. gambiense* type 1 for the subsequent analysis, which brought the number of *T. b. gambiense* type 1 isolates to 111.

T. b. rhodesiense included isolates collected between 1956 and 2015. For isolates with known origins, 13 came from East Africa, six from Southern Africa, and two from North Africa. No isolate came from West or Central Africa as expected. Most isolates were extracted from humans (18), while two were isolated from cattle and one from a goat.

The group we expediently named *T. b. gambiense* type 2 included isolates collected between 1938 and 1984. Most isolates in this study came from Côte d'Ivoire (8), and the rest came from DRC (3), Togo (3), Cameroon (1), and Rwanda (1). Some isolates from pigs were historically labeled as type 2 but are now considered *T. b. brucei* [333]. This group included reported type 2 isolates: FEO, ABBA, MHOM-CI-TH162, MHOM-CI-TH113, MHOM-CI-78-TH2, MHOM-CI-78-TH1-037, LIGO, MURAZ3 [333]. Reports of type 2 mainly occurred between 1978 and 1992, and the latest report was made in Ghana in 2013 [333, 338]. *T. b. gambiense type 2* always appears in geographical areas where type 1 is also found [333, 339]. Admittedly, this group also contained two unassigned human infective isolates: P8F and AnTat-25-1S. The minicircle compositions suggested that they were unlikely to be *T. b. gambiense* type I. Since we observed the limited sensitivity of the single marker gene TgsGP when annotating *T. b. gambiense* type 1, SRA might experience the same issue. Hence, some of the unclassified human infective *T. brucei* might be unidentified *T. b. rhodesiense* instead of novel human infective cell lines.

T. b. brucei included isolates collected between 1960 and 1996. The dataset was heavily biased towards West African isolates (56) but contained a few isolates from Central (4), East

(6), or Southern (6) Africa. Most samples were isolated in Côte d'Ivoire (43). As expected, the geographical distribution of *T. b. brucei* overlapped that of *T. b. gambiense* and *T. b. rhodesiense* [74, 76]. As this group was defined as the lack of TgsGP or SRA markers and the assay could sometimes be not sufficiently sensitive in detecting the markers, we admitted that some of the isolates could have been unidentified *T. b. gambiense* type 2 or *T. b. rhodesiense*.

3.2 Maxicircle assembly and annotation

3.2.1 Maxicircle assembly and maxicircle deletions

The tsetse-transmissible trypanosomes are expected to have functional maxicircles. We assembled the maxicircles with the bespoke pipeline KOMICS [225]. From the *de novo* assemblies of 219/224 isolates, we recovered maxicircles aligned to the EATRO1125 reference without gaps, while five isolates were aligned with extensive gaps. The deletion was confirmed by mapping Illumina reads to the respective maxicircles.

The five isolates shared a 148 nt deletion from nt 999 to nt 1147 on the EATRO1125 reference. The deletion occurred in the conserved region before the start of the 12S rRNA gene at nt 1364. One of the five isolates was *T. b. brucei* J10 closely related to *T. b. equiperdum* type C [90]. The same deletion was detected in *T. b. equiperdum* type C and type OVI, which will be discussed in chapter five.

3.2.2 Maxicircle annotation and preliminary edited mRNA predictions

To polish the *T. b. gambiense* type 1 maxicircle annotation and infer unedited mRNAs, we generated transcriptome data of *T. b. gambiense* type 1 Mongo. Alignments of the Mongo *de novo* assembled maxicircle to the published *T. b. brucei* EATRO1125 maxicircle [225] resulted in 99.81% overall SID over the coding region, with conserved synteny for maxicircle-encoded genes, as expected. Unedited mRNAs were extracted from the annotation. The Mongo transcriptome reads were mapped to the reference nuclear genome and mRNAs (Table *3-2*). All unedited versions of cryptogenes except MURF2 had complete transcript coverage.

Table 3-2. **Summary of transcriptome read mapping of Mongo**

Total reads	93676350
Mapped to *T. brucei* TREU927 nuclear genome	92560189
Not mapped to *T. brucei* TREU927 nuclear genome	1116161
Mapped to un/never edited Mongo mRNAs	135972
Mapped to edited Mongo mRNAs	119196

Table 3-3**. Summary of transcriptome read mapping to unedited Mongo mRNAs**

Unedited mRNA	start	end	Read count	Coverage%	Mean depth	Mean baseq
A6	1	402	2321	100	523.8	34.7
COX2	1	673	17046	99.7	3760.4	35.5
COX3	1	463	15337	100	3740.3	34.7
CR3	1	163	127	100	41.7	35.3
CR4	1	284	89	100	17	33.6
CYB	1	1117	39113	100	5165.8	35.4
MURF2	1	1086	65	43.7	0.7	35.1
ND3	1	270	333	100	127.3	35.1
ND7	1	777	50207	100	8092.3	34.9
ND8	1	361	8821	100	3074.6	33.2
ND9	1	318	488	100	99.3	34.7
RPS12	1	218	2025	100	701.3	35.4

The conservation of pre-edited mRNAs enabled preliminary edited mRNA prediction via sequence alignments. We aligned the pre-edited mRNAs of Mongo and EATRO1125, which exhibited SID between 98.4% and 100%, with MURF2, CR3, ND8, and RPS12 being completely identical. No indels of non-U residues between EATRO1125 and Mongo were detected. We then aligned both pre-edited mRNAs to the edited EATRO1125 mRNAs. As described in 2.4.2, we inferred the uridine modifications on the Mongo pre-edited mRNAs based on the editing patterns in EATRO1125, introducing the same U-indels at aligned editing sites. When the unedited mRNAs differed, SNPs and U-indels were resolved in edited mRNAs to minimize the disruptions to the ORFs, following the guidelines shown in the examples (Table *3-4*). Thus, we obtained the preliminary prediction of edited Mongo mRNAs.

Although published edited mRNA prediction pipelines, such as T-aligner [326] and TREAT [249], can infer edited mRNAs from transcriptome data, the prevalence of junctions and the scarcity of fully edited mRNAs hinder *de novo* edited-mRNA predictions [249, 279, 297]. The level of fully edited mRNAs also varies at different life stages and sometimes further reduces the available fully edited mRNAs and undermines the predictions [262]. Meanwhile, *de novo* mRNA predictions may generate multiple probable ORFs for each mRNA. It would be beneficial if we were interested in all probable editing events, including misediting and alternative editing for substantially different protein products, which was not part of this thesis due to limitations on time and resources. Without confirmation by amplicons, we decided it was not worthwhile to investigate the atypical editing patterns, which narrowed the focus of this project to the ORFs corresponding to known protein sequences. Hence, only the edited mRNAs translated to protein sequences homologous to those in EATRO1125 were selected for annotation. This choice offsets the benefit of *de novo* predictions. Consequently, we decided to take advantage of the readily available mRNA dataset of EATRO1125 instead of starting from scratch.

Table 3-4. **Resolving SNPs and U-indels between Mongo and EATRO1125 pre-edited mRNAs**

Situation	mRNA	Alignment examples			
SNPs were altered at their corresponding locations	A6	Mongo	279	TGG**G**ATTGGGAATTGCCTT	298
				||| |||||||||||||	
		EATRO1125	280	TGG**A**ATTGGGAATTGCCTTT	299
For u deletions, no modification on the editing pattern was made	ND7	Mongo	521	GGA-TAGCGAGAGGGAGAA	538
				||| |||||||||||||	
		EATRO1125	522	GGA**T**TAGCGAGAGGGAGAA	540
the u-insertion here negated the need for an additional post-transcriptional insertion.	COX3	Mongo	278	GGGG**T**TTTTGGGGAAACCA	297
				|||| ||||||||||||	
		EATRO1125	279	GGGG-TTTTGGGGAAACCA	297

3.2.3 Edited mRNA polishing with transcriptome

We polished the preliminary edited mRNA predictions and identified alternative editing patterns using whole-cell transcriptome reads of procyclic form *T. b. gambiense* type 1 Mongo. All polished edited mRNAs had over 88% transcriptome coverage except MURF2,

which had the lowest mapped read count (26) as expected due to the low transcription level (Table *3-5*). We observed complete coverage over edited ND7, ND8, CR3, CYB, COX3, COX2, and RPS12, although read coverage tended to decrease towards the 5' ends. Other cryptogenes (MURF2, ND3, ND9, CR4, A6) contained gaps in coverage in their 5' ends. This was probably due to the relatively low level of fully edited transcripts in the whole cell transcriptome data, stage-specific expression, and differentiated regulation of kDNA-encoded genes. Although mRNAs of complex I subunits had 100% read coverage, the reasons for ambiguities in their edited mRNA predictions are discussed below.

Table 3-5**. Summary of transcriptome read mapping to polished edited Mongo mRNAs**

Edited mRNA	start	end	Read count	Coverage%	Mean depth	Mean baseq
A6_v1	1	817	380	88.1	44.6	34.3
A6_v2	1	818	1637	91.8	187.5	35.1
COX2	1	677	17045	99.7	3773.2	35.5
COX3_v1	1	965	8202	95.2	962.6	35.1
COX3_v2	1	965	7989	95.3	942.4	35.1
CR3	1	295	140	97.3	22	34.7
CR4	1	568	467	95.2	70.8	28.1
CYB	1	1151	39136	100	5037.7	35.4
MURF2	1	1108	26	38.6	0.6	35.4
ND3	1	462	168	100	29.5	34.9
ND7	1	1241	34124	100	2883.7	35
ND8_v1	1	574	851	100	139.1	34.7
ND8_v2	1	574	1135	100	179.7	34.9
ND8_v3	1	574	1169	100	181.9	34.7
ND9	1	645	959	89.9	140.6	35.3

Despite 100% transcriptome coverage over edited ND3, the read coverage dropped to ≤ 5 from nt 39 to nt 255, and the mapping did not provide a consensus due to multiple mismatches on non-U residues.

3.2.4 Alternative editing

We identified alternative editing with the transcriptome data. Two versions of A6 and ND8 in EATRO1125 have been reported [225]. Mongo transcriptomics detected A6_v2 mRNA but not A6_v1 and neither version of ND8. Instead, we identified three alternative editing patterns for the Mongo ND8 mRNA that involved other bases between nt 509-517 and nt 536-553 within the 3' UTR (Figure *3-1*). The three Mongo ND8 mRNAs differed from EATRO1125 by the loss of uridine insertions at nt 513 and nt 515 and the additional insertions at nt 511 and nt 517. Editing patterns of both EATRO1125 mRNAs and Mongo ND8_v1 were identical over nt 536-553. Mongo ND8_v2 differed by relocating the two uridines before the cytosine after it. Mongo ND8_v3 shared the same changes as ND8_v2 and had four additional unique edits, including uridine insertion between two adenines and between the adenine and guanine, the deletion of an encoded uridine at nt 550, and the

insertion of a uridine at nt 553. The ratio of reads that contained the sequence unique to ND8_v1, ND8_v2, and ND8_v3 was 21:58:43, suggesting differential editing preferences.

```
                    1         2         3         4         5         6         7
                    90123456789012345678901234567890123456789012345678901234567890123456789012345678901234
>EATRO1125 ND8_v1   AuGGuGuGAuuuAuuGUGuuuAuGuAuuuAAAGAAAuuCuAUGGUGAAAUUAAAUUUUGACUAAAU
>EATRO1125 ND8_v2   AuGGuGuGAuuuAuuGUGuuuAuGuAuuuAAAGAAAuuCuAUGGUGAAAUUAAAUUUUGACUAAAU
>MONGO ND8_v1       AuuGGGGAuuuuAuuGUGuuuAuGuAuuuAAAGAAAuuCuAUGGUGAAAUUAAAUUUUGACUAAAU
>MONGO ND8_v3       AuuGGGGAuuuuAuuGUGuuuAuGuAuuuAAAGAAACuuuAUGGUGAAAUUAAAUUUUGACUAAAU
>MONGO ND8_v2       AuuGGGGAuuuuAuuGUGuuuAuGuAuAuAAuGAAACuuuAGGUuGAAAUUAAAUUUUGACUAAAU
```

Figure 3-1. **Aligned ND8 mRNAs (nt 509-574) showing the alternative editing over 3' UTR of ND8 in EATRO1125 and Mongo**.

Mongo ND8 mRNAs have three alternative editing patterns that involve bases between nt 509-517 and nt 536-553 within the 3' UTR. These editing patterns are not described in EATRO1125. Highlights: blue: identical to EATRO1125, yellow: unique to Mongo

Although COX3 did not exhibit alternative editing in EATRO1125, the Mongo transcriptome revealed two editing patterns over the 3'-most editing block, just upstream of the stop codon, that differed by one less uridine insertion at nt 942 and an additional uridine at nt 947, thus maintaining the position of the stop codon. All other editing sites were identical between the two versions. The editing resulted in protein sequences that differed in the two amino acids before the C-terminal tryptophan.

3.2.5 Maxicircle encoded gRNAs

We were interested in the maxicircle encoded gRNAs in the clonal *T. b. gambiense* type 1. The *T. b. brucei* EATRO1125 maxicircle encodes gRNAs for the minimal editing of mitochondrial unidentified open reading frame 2 (MURF2) and for cytochrome *c* oxidase subunit 2 (COX2), where four uridines are inserted via cis-editing directed by a gRNA encoded in its 3' UTR [213, 225].

Identification of *T. b. gambiense* type 1 Mongo maxicircle-encoded gRNAs by complementarity search and alignment with the *T. b. brucei* EATRO1125 maxicircle revealed synteny and sequence conservation of these two gRNAs. Despite low coverage, edited MURF2 reads were detected in the procyclic Mongo transcriptome (Table *3-5*). Transcriptomics similarly confirmed the expected U-insertions in COXII (Table *3-5*). We did not detect additional maxicircle gRNAs in Mongo.

3.3 Minicircle assembly and general features

3.3.1 Completeness of minicircle assembly

We characterized the minicircle population of the four *T. b. brucei* subspecies. Minicircle assembly for each isolate was performed using MEGAHIT [303] via KOMICS [66]. The percentages of mapped reads (mapped reads/all reads) ranged between 1.81% and 59.40% (Figure *3-2* A, Supplementary Table.3).

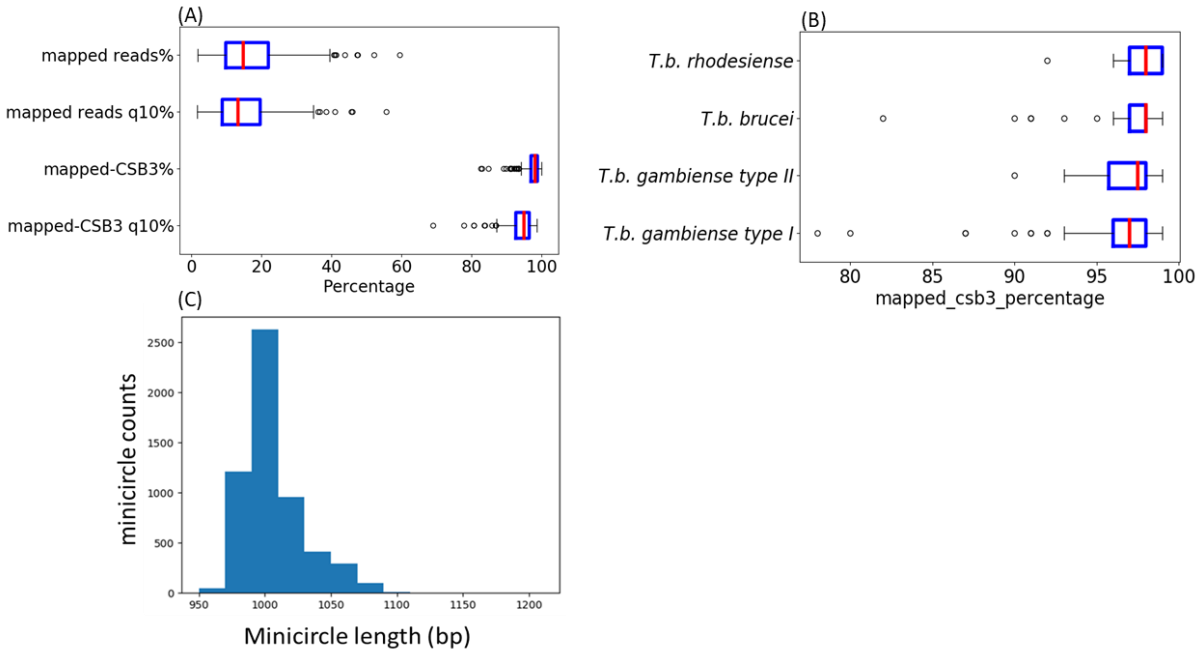


Figure 3-2**. Summary of completeness assessment and minicircle size distribution.**

(A) The percentage of mapped reads, mapped reads with mapping quality ≥ 10, mapped CSB3-containing reads, and mapped CSB3-containing reads with mapping quality ≥ 10 are shown. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point **(B)** Assemblies of 219/224 isolates exhibit over 90% PMC, with a mean PMC of 96.8%. **(C)** Size distribution of 5668 minicircle classes assembled from 224 sub-Saharan T. brucei isolates.

Conserved sequence block 3 (CSB-3), also known as the universal minicircle sequence (UMS), is present in one copy in all minicircles. To assess kDNA coverage and assembly completeness of a given isolate, we mapped the reads to the *de novo* assembled minicircles and maxicircle and calculated the percentage of mapped CSB-3-containing reads (PMC). We reasoned that almost all CSB-3 reads to be mapped if most of the more abundant minicircles were captured in the assembly. Our assemblies exhibited over 90% PMC in 219/224 isolates, with a mean PMC of 96.8% (Figure *3-2* B). only three *T. b. gambiense* type 1 and one *T. b. brucei* isolate had PMC ≤ 90%.

Full read coverage on all minicircle contigs was confirmed by visually examining the mapping in IGV [322]. Pooling minicircle classes from all isolates yielded 5668 distinct classes with less than 95% SID to any other minicircle in the assembly, including the 120 nt semi-conserved regions (the 95% cut-off was chosen for consistency with [225] and [81]). Circularized contigs ranged from 953 to 1206 bp in size, with a peak at around 1000 bp (Figure *3-2* C).

3.3.2 kDNA complexity

We compared the kDNA complexity of the four subspecies. The 107 *T. b. gambiense* type 1 isolates exhibited a unique and highly conserved minicircle composition profile. Out of the combined 5668 minicircle classes compiled from the 224 *T. brucei* isolates, only 195 were detected in *T. b. gambiense* type 1, with 132 being unique to this subspecies and not found in groups capable of sexual reproduction, namely *T. b. gambiense type 2*, *T. b. rhodesiense*, and *T. b. brucei*, as reported previously (Figure *3-3* A) [80]. We assessed the conservation of the 195 classes within *T. b. gambiense* type 1 and other subspecies. Although 61 classes were also detected outside *T. b. gambiense* type 1, only eight were shared by over 20% of other isolates (Figure *3-3* A). On the contrary, 50% of the 195 classes were shared by at least 65.7% of *T. b. gambiense* type 1 isolates (Figure *3-3* A). Moreover, apart from four strains that had undergone substantial kDNA reduction, all other *T. b. gambiense* type 1 isolates contained at least 90 out of the 195 classes. In contrast, other *T. brucei* subspecies typically contained no more than 13 of these minicircles in their kDNA network, except for four isolates (Figure *3-3* B).

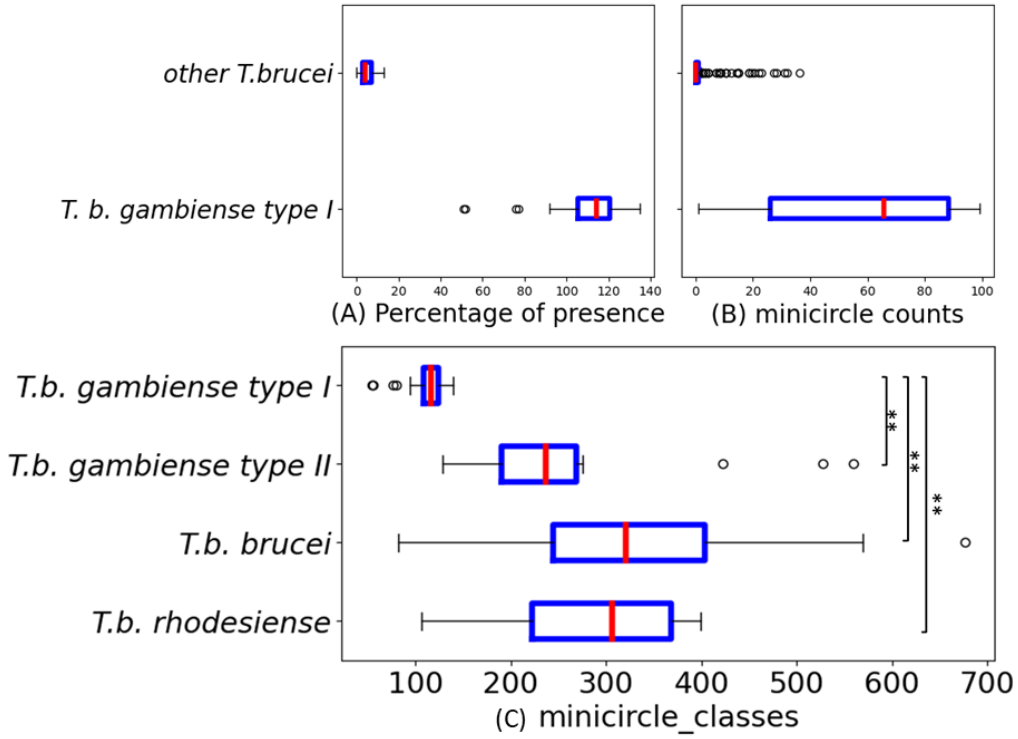


Figure 3-3. **Comparison of *T. b. gambiense* type 1 minicircle composition to the groups capable of sexual reproduction.**

The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point **(A)** For each of the 195 minicircle classes found in *T. b. gambiense* type 1, the graph plots what percentage of isolates the class is found in. 50% of classes were shared by at least 65.7% of *T. b. gambiense* type 1 isolates (red line = median). 61 classes were also detected in other *T. brucei ssp*. (for 132 classes the percentage is zero), but only 8 classes were present in at least 20% of other *T. b.* isolates. **(B)** For each isolate, the graph plots how many of the 195 *T. b. gambiense* type 1 minicircle classes are present. Four additional *T. b. gambiense* type 1 isolates were annotated despite the lack of TgsGP markers (note that these are apparent as two individual data points, one representing three isolates with nearly identical complexity). Besides four other isolates with unexpectedly low kDNA complexity, 107 *T. b. gambiense* type 1 isolates (bottom of the 80-percentile box) contained at least 90 of the 193 minicircle classes, while other subspecies contained no more than 13 of the 193 classes per network. **(C)** *T. b. gambiense* type 1 has a less complex kDNA network (unpaired t-test, $p<0.001$) and less intraspecific variation in kDNA complexity (Levene test for equal variance, $p<0.001$, $SD_{T.\ b.\ gambiense\ type\ 1} = 13.5$, $SD_{T.\ b.\ gambiense\ type\ 2} = 128.9$, $SD_{T.\ b.\ rhodesiense} = 84.0$, $SD_{T.\ b.\ brucei} = 119.3$) than other three subspecies.

Kinetoplast DNA complexity, here defined as the number of unique minicircle classes, was significantly higher in *T. b. gambiense type 2*, *T. b. rhodesiense*, and *T. b. brucei*, with mean values of 287, 291, and 329, respectively, compared to 115 classes per network observed in *T. b. gambiense* type 1 (Figure *3-3* C). Additionally, *T. b. gambiense* type 1 exhibited less variability in the level of kDNA complexity compared to taxa capable of sexual reproduction (Figure *3-3* C). Remarkably, two *T. b. gambiense* type 1 isolates (LiTat-1-3, LiTat-1-5-P9) contained 52 and 51 classes, respectively, and 2225.71 and 2871.58 copies of minicircles per network, respectively. Both isolates have been cultured for an extended period in the lab before DNA isolation, which probably resulted in the reduction in kDNA complexity that compromised the mRNA editing capacity.

3.3.3 Number of minicircles per network

T. brucei kDNA network is estimated to contain 20-50 maxicircles [195, 340]. A study using a subset of the 224 sub-Saharan isolates has calculated a slightly lower estimation of 17 maxicircles and around 2100 minicircles per network on average [80]. Here we assume 30 maxicircles per network as in the previous analysis of EATRO1125 [225] to estimate minicircle copy number (MCN). MCN was estimated as the ratio of minicircle read depth to maxicircle read depth (Supplementary Table 4).

On average, kDNA networks of *T. b. gambiense* type 1 and II, *T. b. brucei*, and *T. b. rhodesiense*, contained 3904, 3978, 4325, and 5098 minicircles per network, respectively (Figure *3-4*). Notably, the total MCN per kDNA network of *T. b. rhodesiense* significantly surpassed that of *T. b. gambiense* type 1 and II but did not differ from *T. b. brucei* (Figure *3-4*). It was uncertain if the higher kDNA complexity contributed to the larger network size in *T. b. rhodesiense*. However, the total MCN among the other three subspecies did not differ significantly. Furthermore, 19 *T. b. gambiense* type 1 and two *T. b. brucei* isolates had a total MCN < 2000 minicircles.

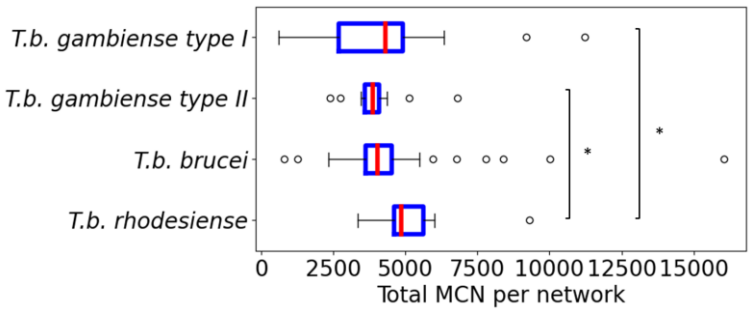


Figure 3-4**. Total minicircle copy number (MCN) in four *T. brucei* subspecies.**

T. b. rhodesiense had significantly higher total MCN than *T. b. gambiense* type 1 (p=0.002) and *T. b. gambiense type 2* (p=0.006) but not *T. b. brucei* (p=0.066) (unpaired t-test). The total MCN among the other three subspecies did not differ significantly (p>0.05). The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point.

3.3.4 The features of CSBs

We characterized the conserved features of the minicircles pooled from the four subspecies. Alignment of the 5668 minicircles revealed a 102 bp semi-conserved region (containing CSB-1, CSB-2, and CSB-3) (Figure *3-5*). CSB-1 was 100% conserved. CSB-2 exhibited more variability than CSB-1 and CSB-3. The top three most common CSB-2 motifs were TCCCGTGC, TCACGTGC, and TACCGTGC, accounting for 64.5%, 12.2%, and 11.5% of minicircle classes, respectively.

Table 3-6. **Alternative CSB-3 sequences were detected in three minicircle classes from three isolates**

Alternative CSB3 sequence (frequency for that class))	Minicircle class	Isolate	Average copy number per network
AGGGTTGGTGTA (100%)	Tb_mO_267	GPAL-ZM-83TRPZ265	2.8
GAGGTTGGTGTA (47%)	Tb_mO_1651	AnTat-25-1S	3.0
GGGGTTGGTATA (60%)	Tb_mO_4348	GPAL-ZM-83TRPZ265	10.9

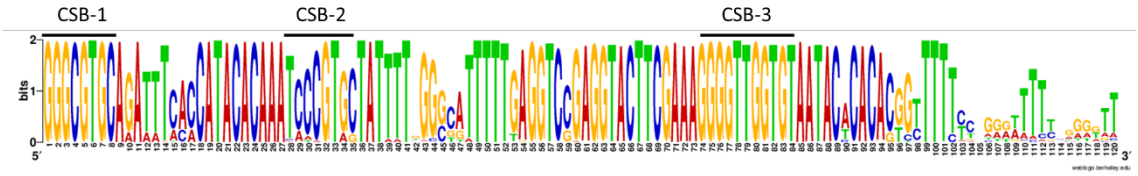


Figure 3-5**. Weblogo-generated sequence motif of the CSB-containing conserved regions from 5668 minicircles**.

The motif is generated with the first 120 nt of minicircles aligned at CSB1. The conserved region spans around 102 nt from CSB1.

While CSB-3 is conventionally considered universally conserved (GGGGTTGGTGTA), an alternative CSB-3 (GGGGTTGATGTA) has been reported in EATRO1125 [225]. We identified 11 minicircle classes with the reported alternative CSB-3 and 3 minicircle classes with novel single G-A substitutions (Table *3-6*). Inspection of read mappings revealed that each novel alternative CSB-3 was present in a single minicircle class from a single isolate. Two minicircle classes exhibited heterogeneity at the CSB-3 site: Tb_mO_1651 and Tb_mO_4348. The proportions of Tb_mO_1651 containing the canonical and alternative CSB-3 (GAGGTTGGTGTA) were 53% and 47%, respectively, while the proportions of Tb_mO_4348 containing the canonical and alternative CSB-3 (GGGGTTGGTATA) were 40% and 60%, respectively.

3.4 Minicircle annotation

To generate a database of gRNAs and for the assessment of editing capacity of editing block conservation, we annotated minicircles of the *T. brucei* subspecies. The complete set of sub-Saharan *T. brucei* minicircles, gRNAs, and cassette was deposited on Figshare (DOI: 10.6084/m9.figshare.27174027).

3.4.1 Minicircle annotation for groups capable of sexual reproduction

We *de novo* assembled and annotated minicircle genomes of the 113 isolates from groups capable of sexual reproduction using published *T. b. brucei* EATRO1125 edited mRNAs including two versions of A6 and ND8 and the published EATRO1125 maxicircle coding region [225]. Based on a cut-off of < 95% overall SID, 5534 distinct minicircles were pooled and annotated in batches of 500 as described.

To assess the completeness of the gRNA coverage in each subspecies, gRNAs encoded on minicircles present in each subspecies were pooled and mapped to the mRNA. We observed 100% gRNA coverage on all mRNAs except MURF2 and ND3 (Figure *3-6*). The missing gRNA for the two 3' most editing sites in MURF2 mRNA has been described in a previous study using EATRO1125 [225]. Because the set of minicircles represented the collective maximal kDNA complexity of each subspecies, all unaccounted-for editing sites in this annotation were shared by all isolates of the subspecies.

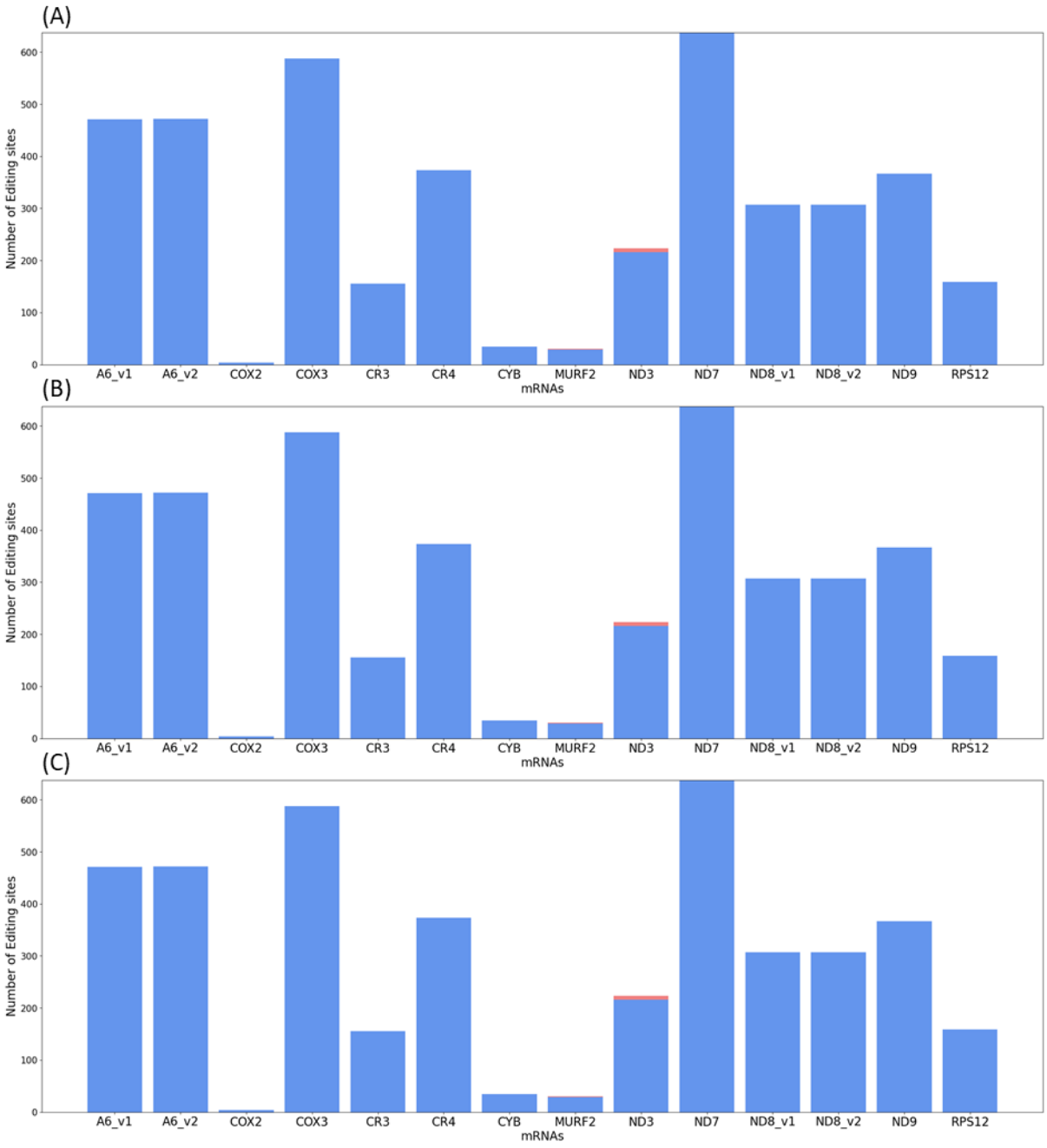


Figure 3-6. **gRNA coverage on published EATRO1125 mRNAs using gRNAs predicted from the collective set of minicircles from all (A) *T. b. gambiense* type 2, (B) *T. b. brucei*, and (C) *T. b. rhodesiense* isolates.**

(A-C) The annotation allows 100% gRNA coverage on all mRNAs except MURF2 and ND3. MURF2 has two 3' most editing sites not covered by gRNAs. The gap on ND3 spanned between nt 373 and nt 387, including eight editing sites.

The gaps in ND3 gRNA coverage were identical in the three subspecies. The gap spanned nt 373 to nt 387 in *T. b. gambiense* type 2 over eight editing sites, while in *T. b. rhodesiense* and *T. b. brucei* the gap spanned nt 373 to nt 385 over seven editing sites. In EATRO1125, the single gRNA responsible for editing sites from nt 373 to nt 396 was also absent [225]. The absence of the gRNA from all isolates capable of sexual reproduction suggested that the

editing pattern of ND3 over this region probably requires further analysis with RNAseq or PCR and modifications.

The gRNAs of COX2 and MURF2 were maxicircle-encoded, and each mRNA had one initiation gRNA. In *T. b. gambiense* type 2, all other mRNAs had multiple initiation gRNAs (Table 3-7). In *T. b. brucei*, ND3 and ND7 have single initiation gRNAs, and all other mRNAs have at least four initiation gRNAs (Table 3-8). In *T. b. rhodesiense*, ND9 also had a single initiation gRNA, while other mRNAs had between two (RPS12) and 20 (ND3) initiation gRNAs (Table 3-9).

Guide RNAs that were predicted to edit the shared regions of alternative versions of the same mRNA (such as the ND8_v1 and ND8_v3 initiation gRNA) were counted only once, which gave 4497, 11583, and 4784 unique gRNAs for *T. b. gambiense* type 2, *T. b. brucei*, and *T. b. rhodesiense,* respectively. The gRNAs and minicircle annotations were combined with annotations of *T. b. gambiense* type 1, which will be elaborated in the following sections, to make the complete dataset of sub-Saharan *T. brucei* gRNAs of the 224 isolates.

Table 3-7. ***T. b. gambiense* type 2 gRNA coverage on mRNAs of maxicircle-encoded cryptogenes. Published edited mRNAs of EATRO1125 were used.for minicircle annotation**

	total gRNAs	unique gRNAs	initiation gRNAs	missing gRNAs	all insertions	covered insertions	all deletions	covered deletions	% coverage
A6_v1	667	667	2	0	443	443	28	28	100
A6_v2	670	5	5	0	444	444	28	28	100
COX2	1	1	1	0	4	4	0	0	100
COX3	1027	1027	8	0	547	547	41	41	100
CR3	240	240	14	0	145	145	10	10	100
CR4	348	348	6	0	328	328	45	45	100
CYB	26	26	13	0	34	34	0	0	100
MURF2	2	2	0	1	26	25	4	4	96.7
ND3	196	196	17	1	210	202	13	13	96.4
ND7	901	901	9	0	551	551	87	87	100
ND8_v1	462	462	3	0	261	261	46	46	100
ND8_v2	465	6	4	0	260	260	47	47	100
ND9	375	375	4	0	346	346	21	21	100
RPS12	241	241	3	0	131	131	28	28	100

Note: unique gRNA does not double count gRNAs aligned to the regions with identical sequences on alternatively edited mRNAs

Table 3-8. ***T. b. brucei*** **gRNA coverage on mRNAs of maxicircle-encoded cryptogenes. Published edited mRNAs of EATRO1125 were used.for minicircle annotation**

	total gRNAs	unique gRNAs	initiation gRNAs	missing gRNAs	all insertions	covered insertions	all deletions	covered deletions	% coverage
A6_v1	1643	1643	4	0	443	443	28	28	100
A6_v2	1649	10	10	0	444	444	28	28	100
COX2	0	0	1	0	4	4	0	0	100
COX3	2700	2700	15	0	547	547	41	41	100
CR3	623	623	44	0	145	145	10	10	100
CR4	897	897	11	0	328	328	45	45	100
CYB	66	66	37	0	34	34	0	0	100
MURF2	1	1	0	1	26	25	4	4	96.7
ND3	457	457	1	1	210	203	13	13	96.9
ND7	2250	2250	1	0	551	551	87	87	100
ND8_v1	1264	1264	6	0	261	261	46	46	100
ND8_v2	1265	17	7	0	260	260	47	47	100
ND9	986	986	9	0	346	346	21	21	100
RPS12	666	666	8	0	131	131	28	28	100

Table 3-9. ***T. b. rhodesiense*** **gRNA coverage on mRNAs of maxicircle-encoded cryptogenes. Published edited mRNAs of EATRO1125 were used.for gRNA annotation**

	total gRNAs	unique gRNAs	initiation gRNAs	missing gRNAs	all insertions	covered insertions	all deletions	covered deletions	% coverage
A6_v1	676	676	3	0	443	443	28	28	100
A6_v2	677	4	4	0	444	444	28	28	100
COX2	1	1	1	0	4	4	0	0	100
COX3	1146	1146	7	0	547	547	41	41	100
CR3	247	247	19	0	145	145	10	10	100
CR4	388	388	6	0	328	328	45	45	100
CYB	26	26	14	0	34	34	0	0	100
MURF2	1	1	0	1	26	25	4	4	96.7
ND3	184	184	20	1	210	203	13	13	96.9
ND7	958	958	13	0	551	551	87	87	100
ND8_v1	468	468	3	0	261	261	46	46	100
ND8_v2	468	6	3	0	260	260	47	47	100
ND9	403	403	1	0	346	346	21	21	100
RPS12	275	275	2	0	131	131	28	28	100

3.4.2 kDNA annotation of *T. b. gambiense* type 1 Mongo isolate

The substantial difference between the minicircle compositions of *T. b. gambiense* type 1 and other sub-Saharan *T. brucei* groups raised the question of how the reduction in kDNA complexity affected editing capacity. To address this question, we report here the first detailed annotation of *T. b. gambiense* type 1 kDNA using transcriptome data of isolate Mongo. The gRNA alignments, gRNAs, and cassettes are uploaded to Figshare (DOI: https://doi.org/10.6084/m9.figshare.27174039).

We predicted the *T. b. gambiense* type 1 fully edited mRNAs using sequence and ORF homology to EATRO1125 edited mRNAs and polished the prediction with transcriptome data . We used the predicted fully edited mRNA sequences to identify gRNA genes in the

121 *T. b. gambiense* type 1 Mongo minicircle classes. The minicircle classes are unique circularized contigs clustered at 95% sequence identity (SID). Meanwhile, the same sequence may be aligned to alternatively edited mRNAs and recorded as distinct gRNA genes.

Prior identification of gRNA cassettes based on the semi-conserved, inverted 18-bp repeats detected three cassettes in 73 minicircles and four cassettes in 48 minicircles (a total of 411 cassettes) (Figure *3-7* A). Alignment allowing no gaps between gRNA and edited mRNA identified two maxicircle-encoded gRNA genes and 529 candidate minicircle-encoded gRNA genes. Excluding the gRNAs that edited the identical regions over alternatively edited mRNAs, the annotation yielded 343 unique gRNA genes, of which 338 (97.4%) were encoded in cassettes flanked by the 18-bp inverted repeats reported before in *T. brucei* and *T. congolense* [227, 228, 230, 341, 342] (Figure *3-7* B). Only 15 cassettes contained multiple gRNA genes. The four gRNAs whose genes were not bound within cassettes were termed 'orphan gRNAs' [225, 341, 343]. Not all cassettes contain gRNA genes, and a cassette may encode multiple gRNAs.

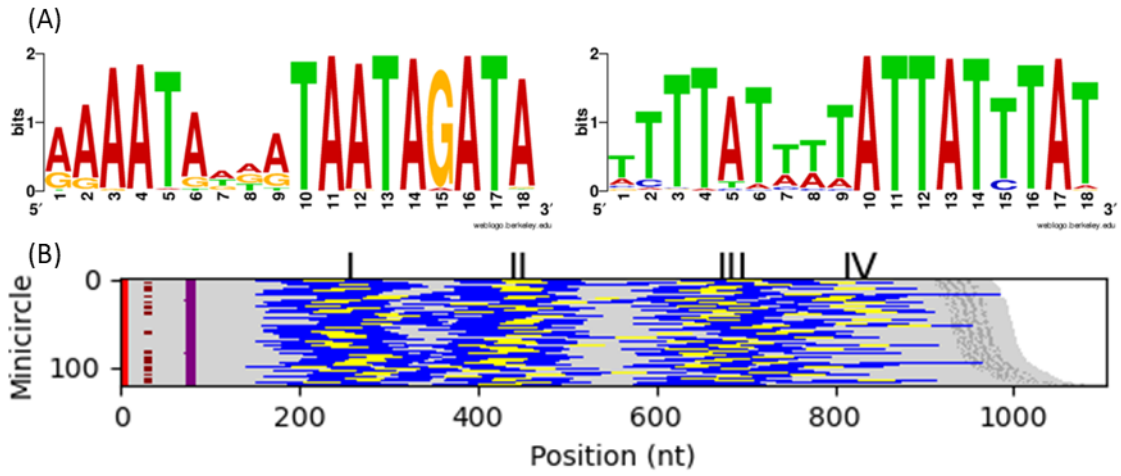


Figure 3-7. ***T. b. gambiense* type 1 minicircle annotation.**

(A) The imperfectly conserved forward and reverse repeats that frame gRNA gene cassettes. The repeat sequences were extracted from aligned Mongo minicircles. The consensus motifs were identical to the published sequences for *T. b. brucei* strain EATRO1125 [81]. **(B)** Structure of 121 *T. b. gambiense* type 1 Mongo minicircles ordered by length (rightmost brown line). Red, brown, and purple represent conserved sequence blocks CSB1, CSB-2, and CSB-3, respectively. The regions between the 5' of the forward 18 bp inverted repeats to the 3' of the inverted repeats are shown as dark blue. Cassette-associated and orphan canonical gRNA genes are shown in yellow. The labels for the four gRNA cassette positions, I–IV, are located at each cassette's median center position. Dark gray bars shows A-tracts of the bend region.

Almost all gRNAs were found on the coding strand as expected, although three minicircle classes were predicted to encode gRNAs on the template strand in cassette I (one ND8 gRNA and two COX3 gRNAs). Most of the EATRO1125 anti-sense gRNAs are also located in cassette I, suggesting that the cassette position was probably critical in anti-sense gRNA expression [81, 225]. Finally, we identified 88 non-canonical gRNA genes (i.e. putative gRNA genes not matching expected edited sequences) within the remaining cassettes using the nucleotide bias described in a previous study [225].

The gRNAs complementary region had an average length of 39.7 nt (sd = 5.62). The gRNAs covered 97.2% of editing sites and achieved over 93% editing site coverage over all mRNAs (Table 3-10, Figure *3-8* A). Individually, complete editing site coverage was observed on at least one version of most non-complex I mRNAs. Conversely, the complex I subunit gene mRNAs all contained unaccounted-for editing sites and had gRNA coverage ranging from 93.4% in ND8 to 97.3% in ND7. Given the average gRNA length of ~40 nt and a minimal anchor length of 6 nt [225], we predict that at least 15 additional gRNAs would be necessary to bridge the gaps in gRNA coverage. Some if not all of these may be hidden among the non-canonical gRNAs but were not identified by our pipeline because their quality scores didn't meet the cut-off. This could be because we chose the wrong (necessarily arbitrary) cut-off, or because our predicted edited mRNA sequences contain errors. We also identified the initiation gRNAs for all genes in at least one version of editing. Despite some level of gRNA redundancy in the downstream editing sites, all mRNAs had a single initiation gRNA including CR3, for which eight initiation gRNAs had been identified in *T. b. brucei* EATRO1125 [225].

Table 3-10. ***T. b. gambiense*** **type 1 Mongo gRNA coverage on mRNAs of maxicircle-encoded cryptogenes**

product	total gRNAs	unique gRNAs	total initiation gRNAs	missing gRNAs	insertions	insertions covered	deletions	deletions covered	% coverage
A6_v1	46	46	0	2	443	433	28	24	97.0
A6_v2	43	0	1	1	444	443	28	28	99.8
COX2	1	1	1	0	4	4	0	0	100
COX3_v1	76	76	1	0	544	544	42	42	100
COX3_v2	74	0	0	1	544	532	42	42	97.8
CR3	13	13	1	0	142	142	10	10	100
CR4	29	29	1	0	328	328	44	44	100
CYB	2	2	1	0	34	34	0	0	100
MURF2	1	1	0	1	26	24	4	4	93.3
ND3	18	18	1	2	211	201	19	19	95.7
ND7	70	70	1	3	550	541	86	78	97.3
ND8_v1	39	39	1	1	259	239	46	46	93.4
ND8_v2	36	1	1	1	260	240	47	47	93.5
ND8_v3	36	0	1	1	259	240	46	46	93.5
ND9	31	31	1	3	346	329	19	15	94.3
RPS12	16	16	1	0	130	130	28	28	100
Total	**531**	**343**	**13**	**16**	**4524**	**4404**	**489**	**473**	**97.2**

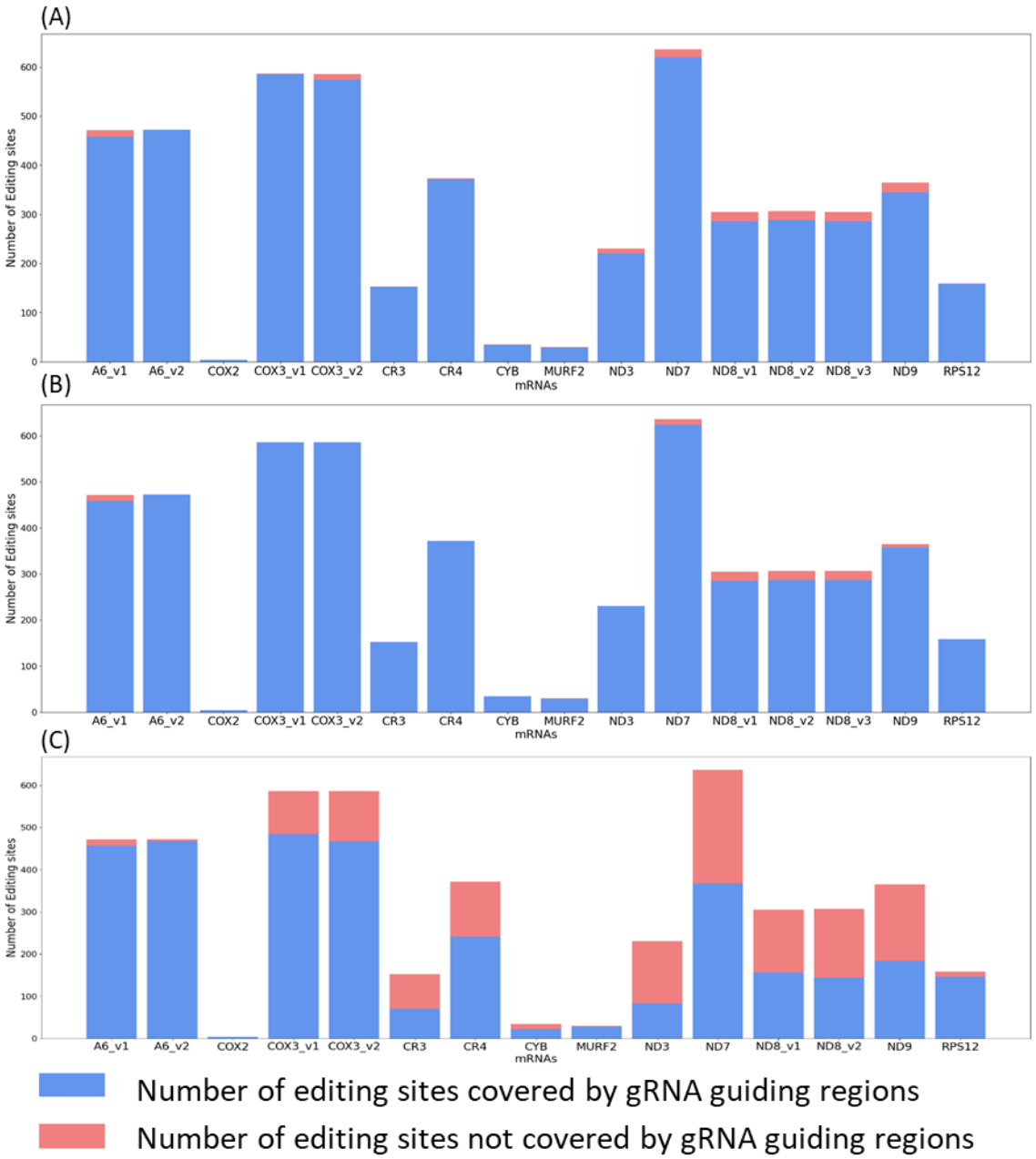


Figure 3-8. **gRNA coverage on *T. b. gambiense* type 1 Mongo mRNAs using gRNAs predicted from (A) *T. b. gambiense* type 1 Mongo, (B) collective *T. b. gambiense* type 1, and (C) *T. b. gambiense* type 1 LiTat-1-3 minicircles.**

(A) At least one version of non-complex I mRNAs has complete editing site coverage. Conversely, the complex I subunit gene mRNAs all contain unaccounted-for editing sites. **(B)** The annotation allows 100% gRNA coverage on nine mRNAs. No initiation gRNA for A6_v1 is detected. Gaps in gRNA coverage are observed on MURF2, ND7, ND8, and ND9. **(C)** LiTat-1-3 covers 36.1% to 82.6% editing site on most crypto genes, except for A6/RPS12, where over 90% editing sites were covered.

3.4.3 Minicircle annotation and gRNA coverage of *T. b. gambiense* type 1 isolates

We aimed to generate a complete set of gRNAs for the clonal *T. b. gambiense* type 1. Using the edited mRNAs inferred from isolate Mongo, we annotated minicircles pooled from all 111 *T. b. gambiense* type 1 isolates. Pooling minicircles and clustering the sequences at a 95% SID threshold yielded 195 unique minicircle classes, which we annotated using the

edited mRNAs predicted for the Mongo isolate, including the alternative versions of ND8, COX3, and A6 (Figure *3-8* B). Guide RNAs that were predicted to code for alternative versions of the same mRNA (such as the ND8_v1 and ND8_v3 initiation gRNA) were counted only once, which gave 545 unique gRNAs that provided ≥ 93% coverage on all mRNAs (Table *3-11*).

Table 3-11. ***T. b. gambiense*** **type 1 gRNA coverage on mRNAs of maxicircle-encoded cryptogenes**

product	total gRNAs	unique gRNAs	initiation gRNAs	missing gRNAs	insertions	insertions covered	deletions	deletions covered	% coverage
A6_v1	74	74	0	1	443	434	28	24	97.2
A6_v2	70	1	2	0	444	444	28	28	100
COX2	1	1	1	0	4	4	0	0	100
COX3_v1	108	108	1	0	544	544	42	42	100
COX3_v2	108	1	1	0	544	544	42	42	100
CR3	26	26	5	0	142	142	10	10	100
CR4	49	49	2	0	328	328	44	44	100
CYB	3	3	1	0	34	34	0	0	100
MURF2	1	1	0	1	26	24	4	4	93.3
ND3	26	26	1	0	211	211	19	19	100
ND7	113	113	2	2	550	543	86	81	98.1
ND8_v1	60	60	1	1	259	239	46	46	93.4
ND8_v2	56	1	1	1	260	240	47	47	93.5
ND8_v3	60	0	1	1	260	240	47	47	93.5
ND9	56	56	1	1	346	340	19	18	98.1
RPS12	25	25	1	0	130	130	28	28	100
Total	**836**	**545**	**21**	**8**	**4525**	**4441**	**490**	**480**	**97.9**

The annotation allowed 100% gRNA coverage on nine mRNAs. As the set of minicircles represented the collective maximal kDNA complexity of *T. b. gambiense* type 1, all unaccounted-for editing sites in this annotation were shared by all type 1 isolates. No initiation gRNA for A6_v1 was detected, suggesting that this alternative mRNA is probably not produced by *T. b. gambiense* type 1. Two of the 3'-most editing sites on MURF2 were not covered as reported previously in EATRO1125 [225]. The 5'-most editing site of the ND7 3' editing domain at nt 231 was not covered, besides the nine insertions and five deletions between nt 427 and nt 447. Three versions of ND8 did not have coverage over the 20 insertions in the 5' UTR. ND9 did not have coverage over the seven insertions and one deletion between nt 20 and nt 34. Ten mRNAs had a single initiation gRNA. Although we identified two A6_v2 initiation gRNAs, one was too short to provide the necessary coverage for the anchor sequence of the downstream gRNAs. The five CR3 initiation gRNAs and the two CR4 initiation gRNAs had identical sequences that strongly support homology. The two ND7 initiation gRNAs were slightly different. The complete set of gRNAs and alignments on edited mRNAs were deposited on Figshare (DOI: 10.6084/m9.figshare.27146595).

Some gRNA genes missing from Mongo were detected in other isolates. The collective minicircle set allowed complete gRNA coverage over ND3. The initiation gRNA

corresponding to COX3_v2 was absent from Mongo kDNA but annotated in the collective minicircle set. Both cases validated the mRNA prediction, suggesting that the apparent absence of certain gRNAs in Mongo was either due to (i) loss of non-essential gRNAs in most of the cells within the population and (ii) bulk sequencing's lack of sensitivity for low copy number minicircles.

3.4.4 Completeness of gRNA coverage in clonal isolates with reduced kDNA

We noticed that *T. b. gambiense* type 1 LiTat-1-3 and LiTat-1-5 isolates from Côte d'Ivoire had undergone a substantial reduction in kDNA complexity due to decades of lab culture in BSF.

The VSGs derived from *T. b. gambiense* type 1 isolates LiTat-1-3 and LiTat-1-5 have been used for diagnostic assay [344]. LiTat-1-5 was isolated in 1952, but the year of isolation of LiTat-1-3 was unrecorded, and we did not know for how long the two isolates were cultured separately. Nevertheless, after decades of culture as BSF, both isolates exhibited considerable loss of kDNA diversity compared to other *T. b. gambiense* type 1 isolates. We calculated $r=0.97$ ($P<0.001$) between the isolates for the correlation of MCN, which suggested the similarity of the kDNA networks.

The two isolates shared 51 minicircles, while LiTat-1-3 had an additional unique minicircle class. The annotation showed that the minicircle unique to LiTat-1-3, Tb_mO_4574, only encoded a gRNA that covered nt 417-462 on CR4 on cassette II, a gRNA that covered nt 270-311 on ND9 on cassette VI, and a gRNA that covered nt 814-853 on ND7 on cassette V. Hence, LiTat-1-3 and LiTat-1-5 had identical gRNA coverage on A6 and RPS12 (LiTat-1-3 shown in Figure *3-8* C). A6 and RPS12 had nearly complete gRNA coverage. In contrast, other cryptogenes had coverage between 36.1% to the maximum of 82.6% observed in COX3_v1. The A6_v2 ORF could form without the 5' most insertion, while three additional uridine insertions were unaccounted for downstream of the initiation gRNA (Figure *3-9*). The editing sites between nt 98 and nt 115 were unaccounted for in RPS12. However, since both isolates were unlikely to be kDNA-independent, the missed editing sites were mostly likely due to incomplete minicircle assembly and annotation.

Table 3-12. ***T. b. gambiense*** **type 1 LiTat-1-3 gRNA coverage on mRNAs of maxicircle-encoded cryptogenes**

product	total gRNAs	unique gRNA	total initiation gRNA	missing gRNAs	insertions	insertions covered	deletions	deletions covered	% coverage
A6_v1	37	37	0	2	443	433	28	24	97.0
A6_v2	38	1	1	2	444	440	28	28	99.2
COX2	1	1	1	0	4	4	0	0	100.0
COX3_v1	43	43	1	10	544	447	42	37	82.6
COX3_v2	38	0	0	12	544	431	42	36	79.7
CR3	4	4	0	5	142	66	10	4	46.1
CR4	14	14	0	7	328	217	44	24	64.8
CYB	1	1	1	1	34	22	0	0	64.7
MURF2	1	1	0	1	26	24	4	4	93.3
ND3	4	4	0	8	211	79	19	4	36.1
ND7	25	25	0	21	550	321	86	46	57.7
ND8_v1	14	14	0	9	259	138	46	18	51.1
ND8_v2	12	0	0	10	260	126	47	18	46.9
ND9	12	12	0	10	346	177	19	6	50.1
RPS12	9	9	1	2	130	120	28	26	92.4
total	253	166	5	100	4265	3045	443	275	70.8

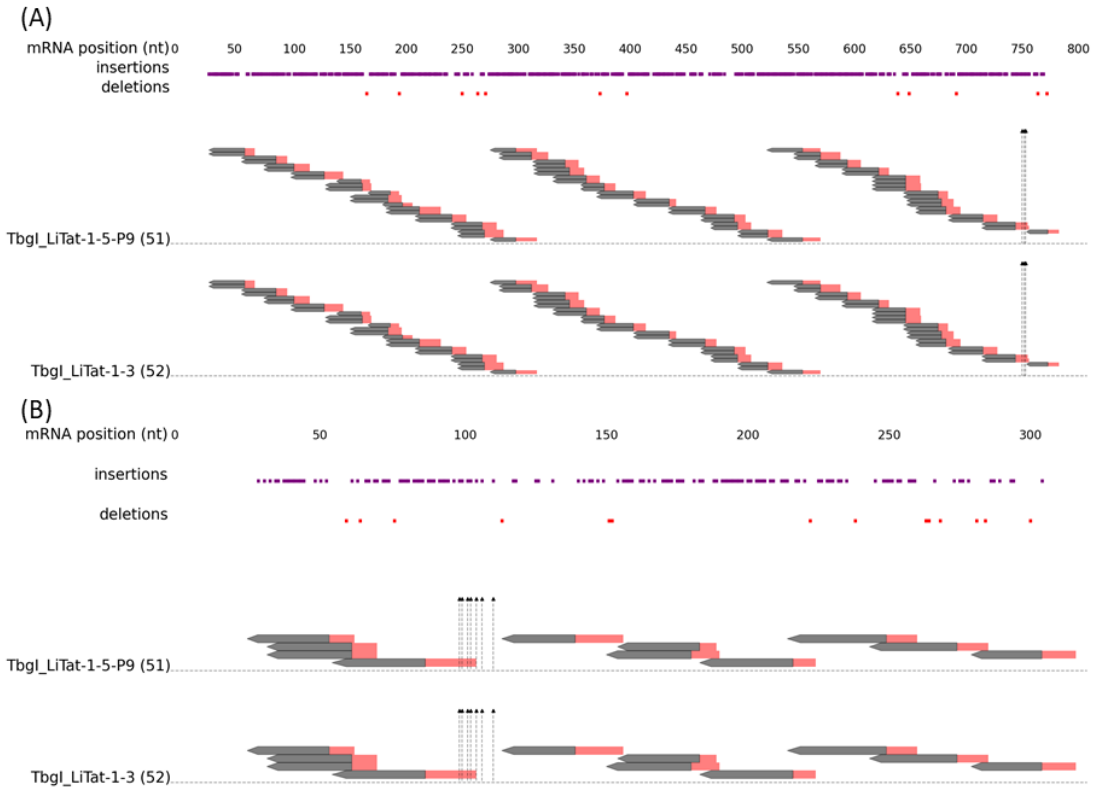


Figure 3-9. **LiTat-1-3 and LiTat-1-5 A6_v2 and RPS12 gRNA alignments.**

We only detected gRNA corresponding to A6_v2 in *T. b. gambiense* type 1. both isolates have nearly complete A6_v2 (A) and RPS12 (B) gRNA coverage. To show gRNA editing patterns, uridine Insertions are plotted in purple blocks, and uridine deletions are indicated below with red blocks. Each arrow represents a unique gRNA, with anchor region in red and guiding region in black. Dashed arrows indicate the editing sites not covered by gRNAs. Only one or two unique gRNAs are aligned to most editing sites in Mongo, whereas multiple unique gRNAs can be responsible for most editing sites in the isolates capable of sexual reproduction.

3.4.5 Summary of minicircle annotations of sub-Saharan *T. brucei*

To generate a dataset of gRNAs of sub-Saharan *T. brucei*, we combined minicircle annotations for the subspecies capable of sexual reproduction and the strictly clonal *T. b. gambiense* type 1. In total 5668 distinct minicircle classes were assembled, of which 5655 contained 19273 cassettes defined by semi-conserved 18-bp inverted repeats, while no cassette structure or gRNA genes was detected in the remaining 13 minicircles. We retrieved 18,808 gRNA genes within the cassette structures. Canonical gRNA genes were identified in >97% of minicircle classes (5510/5666) and 69.59% of cassettes (13,088/18,808), resulting in the annotation of 13,699 canonical gRNAs, including 297 orphan gRNAs found outside cassettes. The mean length of gRNA complementary sequence was 38 nt with a standard deviation of 6 nt.

Besides five and 43 minicircles that contain one or two cassettes, respectively, minicircles typically had three (3714) or four (1890) cassettes. Surprisingly, three minicircles contain five cassettes, a scenario not described in the EATRO1125 reference [225]. These minicircles encode different sets of canonical gRNAs in three cassettes (i.e. they appear to be unrelated) and each had two non-canonical cassettes. Cassettes I, II, and IV were major cassettes where most gRNAs were found, while fewer gRNAs were detected in cassettes III and V (Table *3-13*). In addition, 314 gRNAs were found in cassettes encoding multiple gRNAs, including cases of alternative editing where the same gRNA could be aligned to multiple versions of the same mRNA. Besides orphan gRNAs, minicircles encoded gRNAs on at most four canonical cassettes, with an average of 2.38 canonical positions per minicircle.

Table 3-13. **Counts of sub-Saharan *T. brucei* gRNA genes in each cassette**

Cassette	gRNA count
I	3521
II	3984
III	205
IV	4217
V	1475
Orphan	297
Total	13699

3.4.6 Completeness of gRNA coverage in individual *T. brucei* isolates

We investigated whether the distinct life history of *T. b. gambiense* type 1 (i.e., clonal reproduction, often chronic human infections) affected the maintenance of the ability to edit the various cryptogenes. We assessed the completeness of gRNA coverage for each isolate and cryptogene and compared the average percentage of covered editing sites (COX2 and MURF2, which lack minicircle-encoded gRNAs, were not assessed).

On average, the annotations achieved over 90% editing site coverage except for CYB and complex I subunits ND8 and ND9. While we detected no significant inter-specific differences for ND3, A6, and RPS12, *T. b. gambiense* type 1 showed slightly lower gRNA coverage for all other cryptogenes examined (Figure 3-10). Specifically, *T. b. gambiense* type 1 had less complete gRNA coverage on COX3, CYB, ND7, ND8, and ND9 compared to type 2, on COX3,

CR3, CR4, ND7, ND8, and ND9 compared to *T. b. rhodesiense*, and on CR3, CR4, CYB, ND7, ND8, and ND9 compared to *T. brucei* (Table *3-14*).

Table 3-14. **Mean coverage comparison between *T. b. gambiense* type 1 and other subspecies**.

Note: *: P < 0.05, **: P < 0.001

	***T. b. gambiense* type 1**	***T. b. rhodesiense***		***T. b. gambiense type 2***		***T. b. brucei***	
	Mean	**Mean**	**Pvalue**	**Mean**	**Pvalue**	**Mean**	**Pvalue**
A6	98.22	98.22	0.9874	98.32	0.8876	98.44	0.4891
COX3	96.35	97.99	0.0443*	98.82	0.0069*	97.60	0.0513
CR3	92.09	96.72	0.0059*	95.97	0.0664	95.40	0.0085*
CR4	94.89	98.38	0.0169*	96.76	0.3156	97.59	0.0097*
CYB	84.40	78.77	0.1722	94.96	0.0087*	94.85	0**
ND3	91.62	94.07	0.3036	93.67	0.4845	94.23	0.0682
ND7	92.94	97.85	0.0003**	97.94	0.0025*	97.73	0**
ND8	90.52	93.45	0.019*	95.37	0.0024*	94.66	0**
ND9	90.27	96.53	0.0026*	96.59	0.0108*	96.90	0**
RPS12	98.16	99.16	0.0869	98.68	0.4844	98.80	0.0666

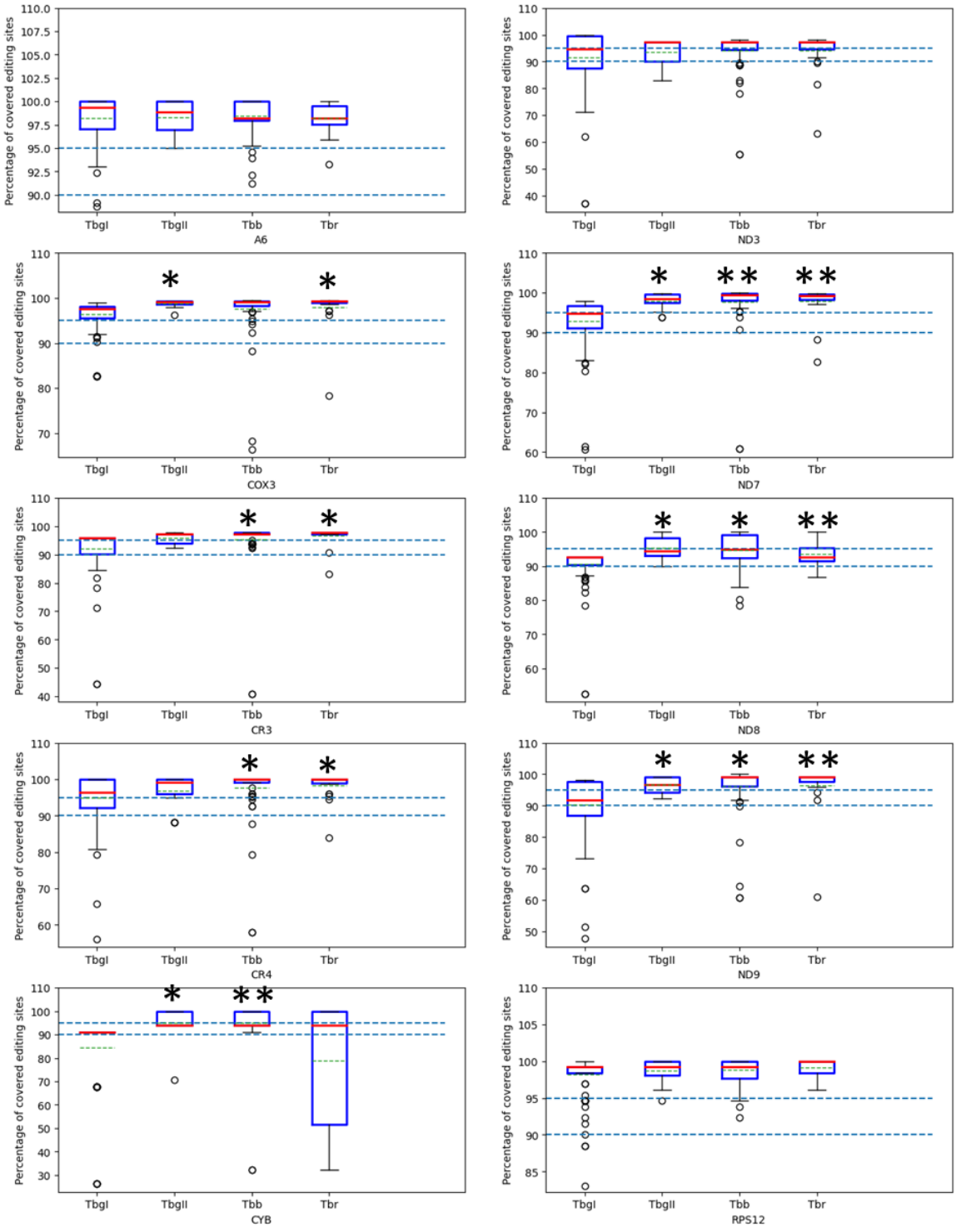


Figure 3-10**. Editing site coverage of edited maxicircle mRNAs by isolates from each *T. brucei* subspecies.**

90% and 95% coverage thresholds are indicated by dashed blue lines. The annotation achieved over 95% average gRNA coverage on A6, COX3, CR3, CR4, ND7, and RPS12. CYB and complex I subunits have slightly lower average coverage over at least 90% of editing sites. *T. b. gambiense* type 1 has significantly less complete editing site coverage than some or all subspecies capable of sexual reproduction on mRNAs except for ND3, A6, and RPS12. *: p<0.05, **: p<0.001. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point

The observation suggested the absence of gRNA genes or an abundance under the detection limit afforded by whole genome sequencing of cell populations. In other words, probably only a minor portion, if any, of the cells within each sample population retained the complete set of gRNA genes essential for tsetse transmission. We did not detect differences in gRNA coverage for cryptogenes required at the bloodstream stage, i.e. A6 and RPS12, suggesting that all the cells retained the ability to edit them as expected. The lack of significant difference over ND3 may have been caused by the gap in coverage shared by all isolates capable of sexual reproduction that masked the more subtle distinctions.

3.4.7 Editing capacity redundancy

Previous identifications of gRNAs in *T. brucei* and *Leishmania* typically found each editing site to be covered by multiple gRNAs encoded on distinct minicircle classes, suggesting there is considerable redundancy in editing capacity [225, 260]. To quantify editing redundancy among *T. brucei* isolates, we calculated the average depth of unique gRNAs over uridine insertions for each edited mRNA in all isolates, excluding the anchor regions and using only the guiding region to calculate the coverage (Figure *3-11*). We did not include the less abundant U-deletions because they required the same gRNAs and had identical gRNA depth as the adjacent nucleotides including insertions. *T. b. gambiense* type 1 had significantly lower editing redundancy than other subspecies, in concordance with the reduced kDNA network (unpaired t-test, $p<0.001$). The average depth difference over edited regions ranged from 0.68 compared to *T. b. gambiense type 2* over ND3 to 3.3 compared to *T. b. rhodesiense* over RPS12, with an average of 1.85 gRNAs less for a given editing site.

The gRNA complexity is lower across all mRNAs in *T. b. gambiense* type 1 compared to other subspecies (Figure *3-11*). Due to the various alternative editing discovered in *T. b. brucei* EATRO1125 and in *T. b. gambiense* type 1 in this study, to simplify the comparison of gRNA functionality, we mapped the gRNAs of different mRNA products of the same gene together (i.e. disregarding the version differences). As the alternative editing did not alter the sequence drastically, mapping the gRNAs to the same reference still preserved their position relative to the ORF.

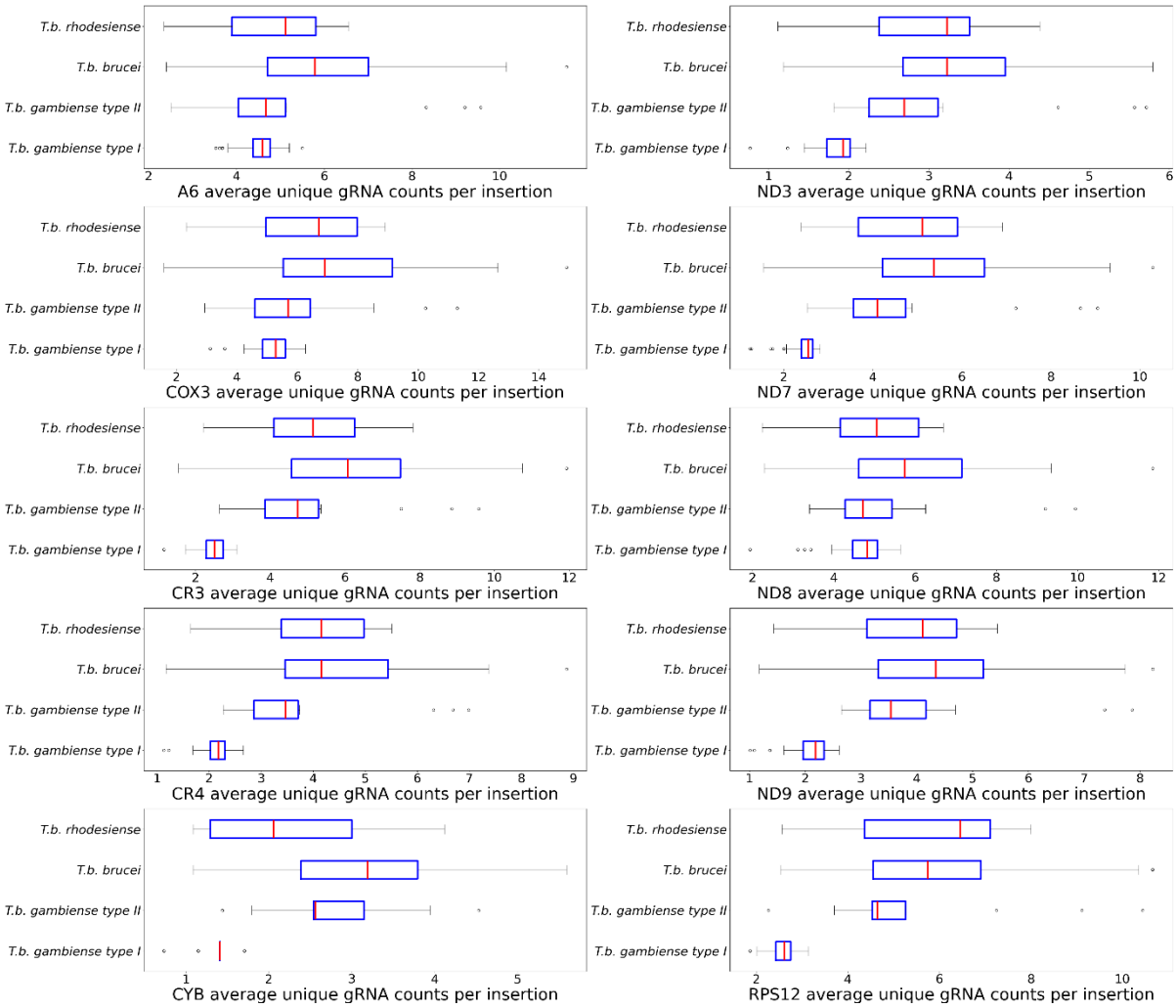


Figure 3-11. **Comparison of gRNA coverage over uridine insertions on edited mRNAs for four sub-Saharan *T. brucei* subspecies**.

T. b. gambiense type 1 had significantly lower gRNA coverage (unpaired t-test, $p<0.001$) than other subspecies on all edited mRNAs. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point

An example of unique gRNA count over RPS12 with four isolates each from one subspecies: *T. b. gambiense* type 1 Mongo, *T. b. gambiense type 2* AnTat-25-1S, *T. b. brucei* EATRO1125, and *T. b. rhodesiense* Rumphi is shown in Figure *3-12*. The isolates had 130, 431, 399, and 408 unique minicircle classes respectively. AnTat-25-1S and Rumphi were chosen as they had similar minicircle class counts to the reference EATRO1125. We observed highly redundant editing capacity on the isolates capable of sexual reproduction, as multiple unique gRNAs were aligned to the same editing sites. In contrast, most editing sites in Mongo were covered by one or two unique gRNAs only. We expected the more streamlined editing capacity in *T. b. gambiense* type 1 given the reduction in kDNA complexity as a result of asexual reproduction mode.

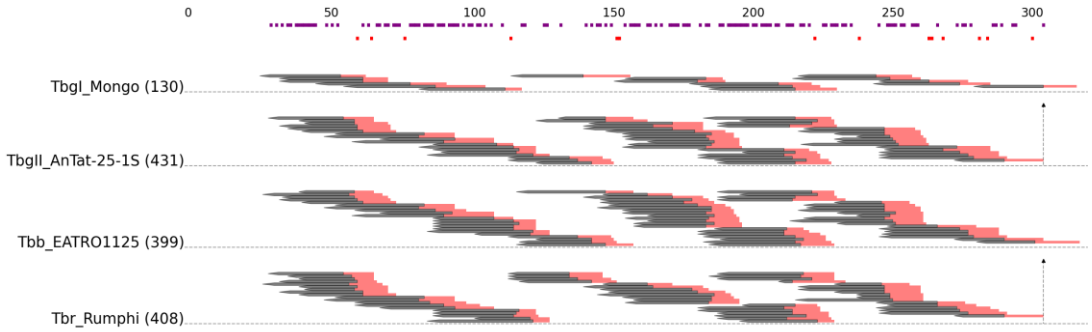


Figure 3-12. **Alignments of unique gRNAs of isolates each from one subspecies of sub-Saharan *T. brucei* over RPS12**

The clonal isolate has a more streamlined gRNA coverage than isolates capable of sexual reproduction. Insertions are plotted in purple blocks, and deletions are indicated below with red blocks. Each arrow represents a unique gRNA, with anchor region in red and guiding region in black.

3.5 Close examination of T. b. gambiense type 1 gRNA alignments

The concept of gRNA family allows us to describe interesting features of *T. b. gambiense* type 1 gRNA annotation. Here we report a detailed inspection of the gRNA alignment over edited mRNAs centered on *T. b. gambiense* type 1 isolate Mongo and other interesting observations made with all *T. b. gambiense* type 1 minicircles (https://doi.org/10.6084/m9.figshare.27174039).

3.5.1 Respiratory complex I / NADH:ubiquinone oxidoreductase

The non-T residues of ND3 differed by a single A-G substitution between EATRO1125 and Mongo. We observed nearly complete gRNA coverage on ND3 and estimated that two additional gRNAs were necessary to cover the nine missed editing sites from nt 114 to nt 126 and one at nt 185. Given the multitude of mismatches and the lack of consensus from nt 39 to nt 255, we could not rule out that the mistakes in mRNA prediction had resulted in the truncation of gRNAs. Nevertheless, the gRNA annotation using all *T. b. gambiense* type 1 minicircles provided complete coverage over these areas, which suggested that the prediction was valid. Notably, the ND3 initiation gRNA also directed editing on COX3 around 320 to 365 nt, a feature shared by the four ND3 initiation gRNAs in EATRO1125 [225].

ND7 contains two editing domains, from 5’ end to nt 180 and nt 230 to 3’ end. We observed nearly complete gRNA coverage for ND7, with just the 3’ editing domain requiring at least three additional gRNAs. In *T. brucei*, the 5’-most uridine insertion of the 3’ editing domain is directed by two orphan gRNA genes [225], which we did not detect in Mongo minicircles. The other two gaps in gRNA coverage spanned from nt 426 to nt 440 and nt 1022 to nt 1024, including eight insertions and deletions. For the 3’ editing domain, a single initiation gRNA encoded in cassette II of that minicircle was identified. The minicircle also encodes an ND8 gRNA on cassette I and an ND9 gRNA on cassette IV. The two *T. b. brucei* EATRO1125 ND7 initiation gRNAs are also encoded in cassette II of their respective minicircles, while both minicircles encode ND9 gRNAs on cassette IV similar to the one in *T. b. gambiense* type 1 Mongo. The conserved cassettes suggest a common origin for the Mongo and EATRO1125 minicircles.

We observed nearly complete gRNA coverage over ND8, with unaccounted-for editing sites restricted to the 5’ UTR. Whether these editing sites are biologically relevant and indeed conserved across *T. brucei* subspecies is uncertain. Two distinct initiation gRNAs as in EATRO1125 directed the three alternative editing patterns (Figure *3-13*). Most interestingly, the ND8_v1 and ND8_v3 appear to share the same initiation gRNA, which allows two ways of inserting three uridines around the cytosine from nt 545 to 548 (uuCu in ND8_v1 vs Cuuu in ND8_v3) due to G-U wobble base pairing. Both gRNA genes were located on cassette IV of their respective minicircles, but these minicircles shared no other conserved cassettes. We exclusively found the initiation gRNA for ND8_v1/3 in 33 isolates, four isolates only had the ND8_v2 initiation gRNA, and 68 isolates had the genes for both. However, we detected neither gRNA gene in six isolates, suggesting unaccounted-for editing patterns or degraded editing capacity.

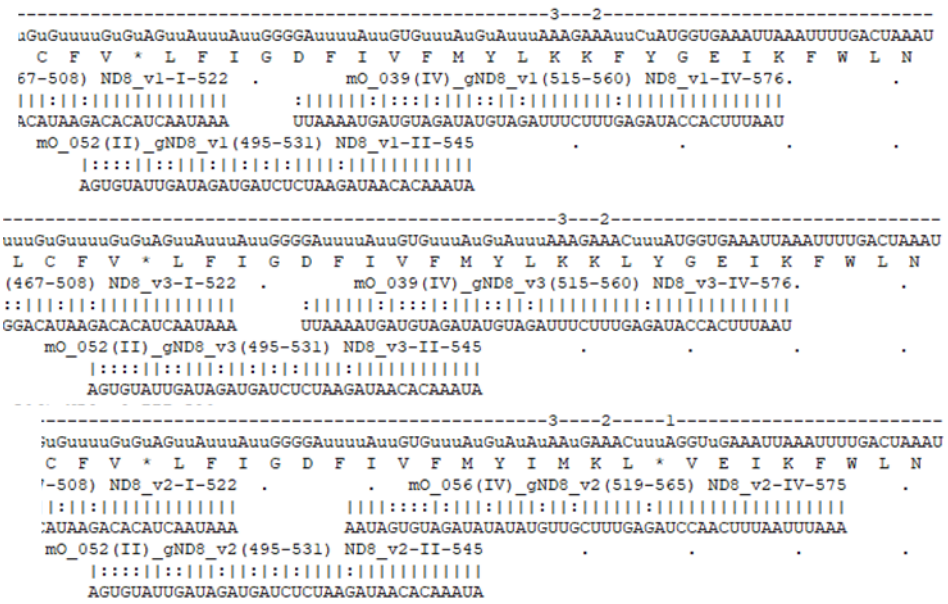


Figure 3-13. **The alternative editing over 3' UTR requires different ND8 initiation gRNAs.**

ND8_v1 and ND8_v2 share the same initiation gRNA, while ND8_v2 requires a different one. mRNAs are oriented in 5' to 3' direction, so gRNAs are aligned in 3' to 5' direction. Numbers in the top track indicate the number of uridine deletions that occur. Inserted uridines are in lowercase in the mRNA sequence. The protein sequence represents the longest ORF. For gRNA alignment, '|' indicates Watson-Crick base pairing, ':' indicates G-U base pairing, and '.' Indicates mismatch.

3.5.2 Respiratory complex III / cytochrome bc_1 complex

Only one subunit of complex III, apocytochrome *b* (CYB), is kDNA encoded in *T. brucei* [280]. CYB gRNAs have been reported to be orphan gRNAs that sit outside the typical gRNA gene cassettes [225, 341, 343]. We identified one orphan gRNA for each of the two editing blocks, compared to two and four, respectively, for *T. b. brucei* EATRO1125 [225]. The initiation CYB gRNAs of Mongo and EATRO1125 were found on minicircles with identical cassette families, and the Mongo minicircle for the upstream gRNA also shared the same cassette families with one of the four EATRO1125 minicircles.

3.5.3 Respiratory complex IV / cytochrome *c* oxidase

Among the three complex IV subunits encoded on the maxicircle, subunit 2 (COX2) and subunit 3 (COX3) pre-mRNAs require post-transcriptional editing. As in other trypanosomatids investigated, the gRNA that directs the insertion of four uridines into COX2 is encoded *in cis* in the mRNA's 3' UTR (see Chapter 4). In contrast to COX2, COX3 is edited extensively.

Transcriptomes revealed two alternative COX3 editing patterns (Figure *3-14*). The Mongo and EATRO1125 COX3 initiation gRNAs were found on different cassettes (III vs. IV), and the respective minicircle classes encoded different gRNA families except for cassette I. We only detected an initiation gRNA for COX3_v1 in Mongo, yet the COX3_v2 initiation gRNA was identified in the collective *T. b. gambiense* type 1 minicircles.

The two minicircle classes encoding the COX3_v1 and COX3_v2 initiation gRNAs had identical cassette families on other positions, indicating that accumulation of mutations in the ancestral gRNA gene was responsible for the divergence into two gRNAs that direct distinct editing patterns, similar to what has been described for A6 in EATRO1125 [225]. The initiation gRNAs for either version were found on a single minicircle class only. Among the 111 *T. b. gambiense* type I isolates, 25 only contained the initiation gRNA for COX3_v1, 41 only had the COX3_v2 counterpart, and 39 had both gRNAs. Six isolates lacked initiation gRNAs for either version of COX3.

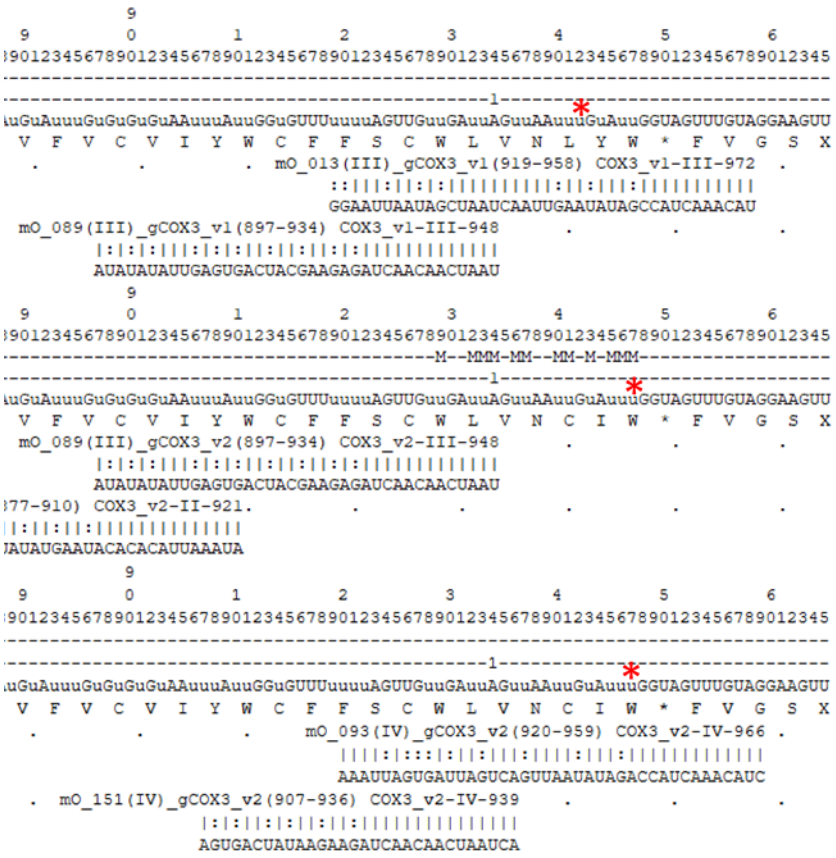


Figure 3-14. **An initiation gRNA was detected for COX3_v1 but not COX3_v2 in Mongo**.

The initiation gRNA for COX3_v2 is detected in other *T. b. gambiense* type 1 isolates (bottom alignment panel). The alternative U insertions are indicated by a red asterisk. The numbers in the first three rows show the coordinates of the bases. Insertions not covered by gRNAs are shown as 'M' on the track below. mRNAs are oriented in 5' to 3' direction, so gRNAs are aligned in 3' to 5' direction. Numbers in the top track indicate the number of uridine deletions that occur. Inserted uridines are in lowercase in the mRNA sequence. The protein sequence represents the longest ORF. For gRNA alignment, '|' indicates Watson-Crick base pairing, ':' indicates G-U base pairing, and '.' Indicates mismatch.

3.5.4 Respiratory complex V / F_1F_O-ATP synthase

The unedited A6 sequences differed between Mongo and EATRO1125 by a single A-G substitution at the 282nd nt. Two alternative versions of A6 editing have been described in *T. brucei* strains EATRO1125, Lister 427, and EATRO 164, which only differ in the 3' UTR, correspond to the use of distinct initiation gRNAs and may play a role in mRNA regulation

[225]. However, we only detected transcripts and gRNAs corresponding to the published A6_v2 mRNA in all *T. b. gambiense* type 1. The minicircles encoding A6_v2 initiation gRNA in Mongo and EATRO1125 contain homologous gRNA genes in the same cassettes while sharing 85% SID overall, probably representing a case where mutations in gRNA genes leads to a new functional gRNA and a novel editing pattern.

3.5.5 Mitoribosome

kDNA also encodes two mitoribosomal protein subunits, RPS12 (uS12m) and uS3m (formerly known as MURF5). The RPS12 mRNA requires extensive editing. Our annotation yielded 16 gRNAs that covered all RPS12 editing sites, including one initiation gRNA. The minicircles encoding the RPS12 initiation gRNAs showed complete cassette conservation between EATRO1125 and Mongo [225].

3.5.6 Unidentified open reading frames

The *T. brucei* maxicircle encodes three ORFs with uncertain function: CR3, CR4, and MURF2. As mentioned above, the gene encoding the second (upstream) gRNA for the minimally edited MURF2 was located on the maxicircle, as in other trypanosomatids. The initiation gRNA remained elusive, leaving two editing sites unaccounted for, as previously reported for *T. b. brucei* EATRO1125 [225]. The mRNAs for the putative Complex I subunits CR3 and CR4 are extensively edited. We identified gRNAs corresponding to all editing sites on both CR3 and CR4. Although EATRO1125 minicircles encode eight CR3 initiation gRNAs, only one was detected in *T. b. gambiense* type 1 Mongo.

3.6 Conservation of editing blocks within and between isolates and subspecies

Despite the difference in kDNA complexity and editing capacity, we wondered if the editing cascade was conserved among *T. brucei*. To answer this question, we grouped gRNAs into families based on their functional similarities. The gRNAs of *T. b. gambiense* type 1 and other *T. brucei* species were annotated with slightly different edited mRNAs (see above) and the mapping positions of gRNAs were adjusted by aligning the protein sequences. Meanwhile, different alternative mRNA products were detected in *T. b. gambiense* type 1 Mongo and *T. b. brucei* EATRO1125. Hence, for simplicity, we did not consider alternative editing when assigning gRNA families. The gRNAs for alternative mRNA products were mapped to the same references because for defining gRNA families we were only interested in the relative positions of gRNA mapping instead of the exact gRNA sequences. As identifying alternatively edited mRNAs was not a major goal of this study, we only recorded the alternatively edited mRNAs revealed by read mapping, which differed minimally among themselves. We then identified editing blocks corresponding to the average editing range of each family. Subsequently, we compared the editing block positions of the four *T. brucei* subspecies.

3.6.1 Assign gRNA families in Sub-Saharan T. brucei

When we compared alignments of gRNAs to their cognate mRNAs across isolates and subspecies it became apparent that, for any given mRNA, the editing blocks, i.e. the region of editing directed by a single gRNA, were remarkably conserved between redundant gRNAs for any one isolate, between isolates for any subspecies, and, indeed, between subspecies. This can be illustrated by mapping the position on mRNAs that corresponded to the 5’ ends of gRNAs, in the following defined as ‘initiation sequence starting positions’ (ISSPs). Note that these positions are distinct from the 5’ end of the anchor region, i.e. the nucleotide in the gRNA where the alignment with its cognate mRNA begins.

The 5’ ends of gRNA anchors were well-conserved among minicircles of different subspecies but not as tightly clustered compared to ISSPs (Figure *3-15* A, for all mRNAs, see Supplementary Figure 1). This is because the 5’ ends of anchor are “bioinformatically” defined at the start of Watson-Crick base pairing rather than biologically defined like the initiation sequence. The 3’ ends of anchor can also extend into the initiation sequence causing a little more variability in its position. The overall positions of the anchors, however, were relatively conserved, which would be shown later for gRNA families. With the arbitrary minimum anchor length of 6 nt, we calculated the average anchor length in *T. brucei* as 11.4 nt, concordant with the observed length of the anchor duplex between gRNA and mRNA [269].

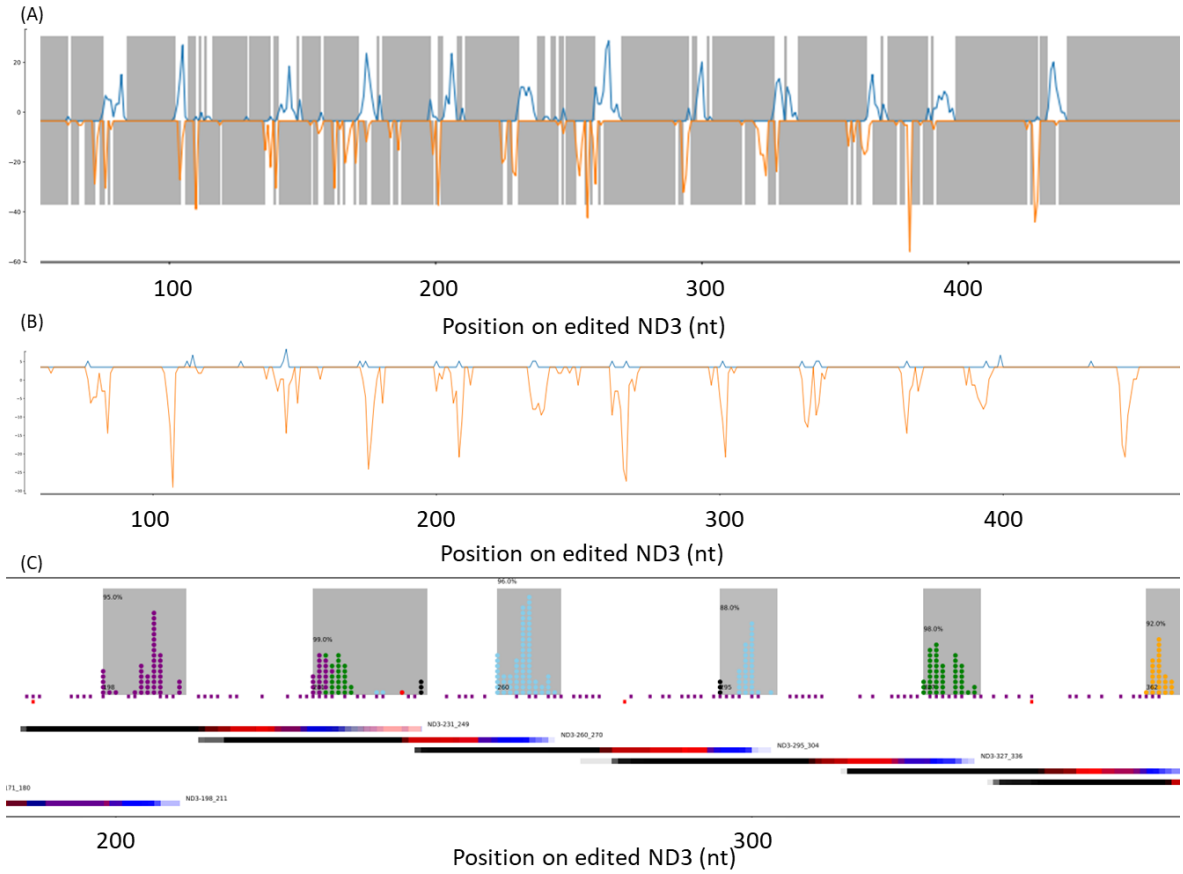


Figure 3-15**. Summary of gRNA families on edited ND3 mRNAs.**

(A) the initiation sequence starting positions (ISSPs) counts and anchor starting positions (ASPs) counts of gRNAs annotated from 224 *T. brucei* isolates on edited ND3 mRNA. Both ISSPs (top, blue) and ASPs (bottom, orange) cluster over limited regions over the mRNA instead of showing a uniform distribution, suggesting a conservation in gRNA mapping positions. Regions without ISSPs or ASPs mapping were shaded. Meanwhile, the clusters of ISSPs are more regular and compact than ASPs, which makes ISSPs a better marker for identifying gRNA families. **(B)** Initiation sequence starting positions (ISSPs) counts on edited ND3 mRNA. The ISSPs of *T. b. gambiense* type 1 (blue) coincide with those of other *T. brucei* subspecies (orange). The phase differences between the ORFs of Mongo and EATRO1125 reference sequences were resolved via protein sequence alignments, so that the relative positions of editing blocks and, most importantly initiation sequences starting positions (ISSPs), were comparable between the two sets. For simplicity, we did not consider the sequence variations of alternative editing and combined all gRNAs identified on different versions of the same pre-mRNA. **(C)** gRNA ISSPs and gRNA coverages of all *T. brucei* isolates on ND3edited mRNA. The positions of initiation sequences and anchors of gRNAs from the same family were highly conserved. Although the gRNAs have similar editing ranges, the difference in the size of the guiding region indicates that gRNAs of the same family are not the exact equivalent of each other. The shaded blocks show the range of ISSP clusters. The scatterplot shows the ISSP count at each position, while ISSPs for gRNAs from different cassettes are color-coded (I : red, II: skyblue, III: orange, IV: purple, V: green, Orphan: black). Insertions are plotted in purple blocks, and deletions are indicated below with red blocks. Below the editing sites, the coverage of gRNA families are plotted by overlapping gRNAs from the same family (initiation sequence: blue, anchor: red, guiding sequence: black).

When compiled for all gRNAs, ISSPs fall into semi-regular clusters highly conserved between the clonal *T. b. gambiense* type 1 and other sub-Saharan *T. brucei* (Figure *3-15* B, for all mRNAs, see Supplementary Figure 2). This observation suggested the concept of 'homologous' gRNAs, i.e. functionally highly similar gRNAs that were conserved across the trypanozoon group (note that the size variation between such homologous gRNAs meant that they were not exact functional equivalents of each other). Further, this conservation

allowed us to combine and quantitatively compare gRNA families from *T. b. gambiense* type 1 with those from other subspecies.

We further grouped homologous gRNAs into 'gRNA families' as follows. We defined boundaries between ISSP clusters as areas without ISSP(ISSP frequency = 0) for over four nt and identified 234 ISSP clusters. After the initial assignment, the mapping positions of the ISSPs, the anchors, and the guiding regions of the gRNAs from the same family were visually examined. Multiple peaks of initiation sequence and anchor mapping were observed in 11 ISSP clusters over 15 nt, indicating the merging of proximate gRNA families. Cutting sites were made in the valleys between the ISSP peaks to subdivide the clusters and the gRNA families.

On the other hand, we observed two instances where the gRNAs from adjacent gRNA families had highly overlapping anchors and guiding regions. Family ND8-491_498 and ND8-504_505 were combined into ND8-491_505. Family ND9-188_201 and ND9-206_207 were combined into ND9-188_207.

After subdivision and merging upon close inspection, we defined a total of 250 ISSP clusters from 13,699 gRNAs. These boundaries delimitated the ISSP clusters, and thus their corresponding gRNAs: gRNAs that belonged to the same cluster were considered one gRNA family. Each gRNA family was considered uniquely responsible for editing a defined region over the mRNA, while gRNAs from the same family were considered homologs.

Firstly, it became clear that the relative positions of ISSP clusters (and therefore gRNA families) were highly conserved between *T. b. gambiense* type 1 and other *T. brucei* subspecies (Figure *3-15* C, for all mRNAs, see Supplementary Figure 3). This conservation indicated that the prolonged genetic isolation of *T. b. gambiense* type 1 due to clonal reproduction had not altered the arrangement of editing blocks, despite a highly unique minicircle population distinct from other subspecies. In addition, *T. b. gambiense* type 1 had significantly fewer families ($\bar{x}$ = 190) compared to other subspecies (*T. b. brucei*: 200, *T. b. gambiense type 2*: 200, *T. b. rhodesiense*: 199), consistent with the less complete editing coverage determined earlier (Figure *3-16*).

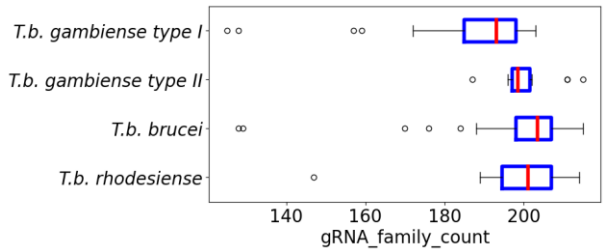


Figure 3-16. **gRNA gene family counts in *T. brucei* subspecies.**

T. b. gambiense type 1 isolates has significantly lower gRNA family counts compared to other subspecies (unpaired t test, p< 0.001)). Mean gRNA family count: T. b. gambiense type 1: 190 *T. b. brucei*: 200, *T. b. gambiense type 2*: 200, *T. b. rhodesiense*: 199. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point

3.6.2 Conservation of gRNA families

By definition, a gRNA family contains homologs directing editing of a defined but not perfectly conserved mRNA region. To visualize the precise editing ranges of homologs, we overlapped gRNAs from the same family and aligned them against edited mRNAs. This revealed the conservation of the relative positions of the initiation sequence, anchors, and the coding sequence of the families based on the delimitation of ISSP clusters (Supplementary Figure 3).

We next assessed the conservation of gRNA families amongst different isolates by determining the percentage of isolates in which they could be detected. The prevalence of gRNA gene families had two peaks, a minor one representing 30 gRNA gene families found in less than 10% of samples and a much larger peak representing 189 gRNA gene families shared by over 80% of isolates (Figure *3-17* A). We considered gRNA families corresponding to the second peak (shared by ≥80% isolates) as highly conserved. They accounted for 95.7 % (13105 / 13699) of all annotated gRNAs and constituted over 87% of the gRNA population in any given isolate, for all subspecies (Figure *3-17* B). *T. b. gambiense* type 1 had a higher percentage of highly conserved gRNA families than subspecies capable of sexual reproduction (Figure *3-17* B). The major gRNA gene percentages among subspecies capable of sexual reproduction did not differ significantly from each other (unpaired t-test, 0.69<p<0.99).

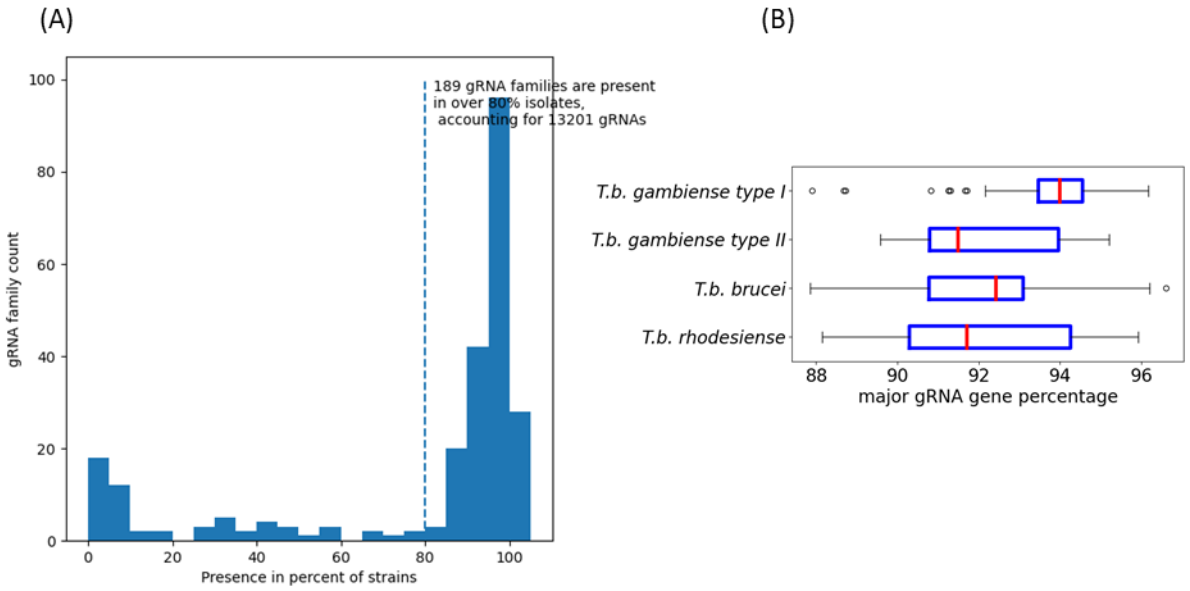


Figure 3-17. **Summary of highly conserved gRNA families.**

(A) Prevalence of gRNA gene families in the 224 *T. brucei* isolates. The distribution has two major peaks, one with 30 gRNA families found in no more than 10% of our samples and one representing the gRNA genes shared by over 85% of isolates. Bin size is 5%, i.e. 19 gRNA families are shared by 0-5% of isolates. 189 of the 250 gRNA families are present in over 80% of samples.**(B)** Percentage of highly conserved gRNA families in each isolate of the four *T. brucei* subspecies. The highly conserved gRNA families account for over 87% of the gRNA population in any given isolate. *T. b. gambiense* type 1 had higher percentage of highly conserved gRNA families compared to *T. b. brucei*, *T. b. gambiense type 2*, and *T. b. rhodesiense* (unpaired t-test, p <0.001). The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point

The less conserved families included cases where we had difficulty in gRNA detection for some isolates. For instance, although an orphan gRNA responsible for the 5' most editing site of ND7 3' editing domain has been reported in EATRO1125 [225], we detected a corresponding gRNA homolog only in 32% of our samples (always encoded outside cassettes, as in EATRO1125). Other less conserved families contained gRNAs of less than 30 nt in length, such as A6-229_230, that did not contribute to the completeness of gRNA coverage and were therefore probably false positives (Figure *3-18* B, for more examples, see Supplementary Figure 3).

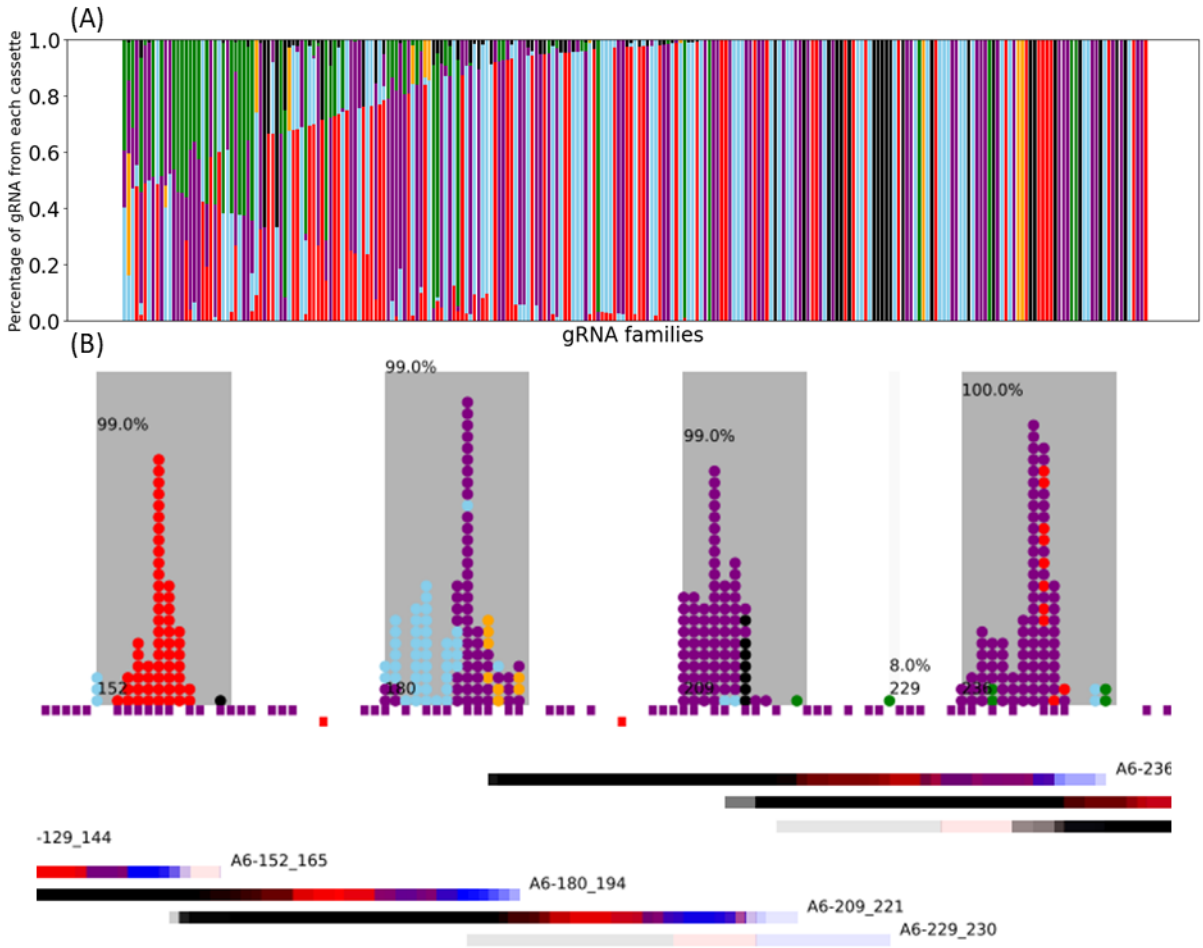


Figure 3-18. **The conservation of cassette positions within gRNA families.**

(A) Percentage of gRNAs from each cassette for each gRNA family. Cassette color coding: I : red, II: skyblue, III: orange, IV: purple, V: green, Orphan: black; **(B)** gRNA ISSPs and gRNA coverages on A6 edited mRNA (screenshot from Supplementary Figure 3). Family A6-229_230 contains a single gRNA gene present in only 8% of isolates. The gRNA is 28 nt long, shorter than the mean gRNA length of 38 nt with sd = 6 nt. Family A6-180_194 contains gRNAs from two dominant cassette positions, cassettes I and IV, while a minor subset of gRNAs are encoded in cassette III. The shaded blocks show the range of ISSP clusters. The scatterplot shows the ISSP count at each position, while ISSPs for gRNAs from different cassettes are color-coded (I : red, II: skyblue, III: orange, IV: purple, V: green, Orphan: black). Insertions are plotted in purple blocks, and deletions are indicated below with red blocks. Below the editing sites, the coverage of gRNA families are plotted by overlapping gRNAs from the same family (initiation sequence: blue, anchor: red, guiding sequence: black).

It was striking that most gRNA families were dominated by gRNAs from the same cassette (Figure *3-18* A). The conserved relative position on minicircles may indicate a common ancestry despite the highly divergent sequences, while the homologs encoded on the less

common cassettes may represent recombination events between minicircles that had shifted the gRNAs onto a novel location. However, we also observed cases where gRNAs from multiple cassettes occurred at comparable abundance in one family (Figure *3-18* B, for more examples, see Supplementary Figure 3). The ISSPs of gRNAs from the same cassettes are often clustered nearby instead of being interspersed with gRNAs from other cassettes, as exemplified by family A6-180_194. They probably represented gRNA genes with distinct origins or deep divergence from common ancestors.

3.6.3 Conservation of editing blocks

As mentioned above, gRNA families are observed at semi-regular intervals along an mRNA (Supplementary Figure 3). The effective range of editing directed by each gRNA family was considered an editing block. For each subspecies, we calculated the effective editing ranges of highly conserved gRNA families as the region covered by the mean starting and ending positions of the complementary sequences of the gRNAs present in the family of the subspecies, excluding the first 6 nt as the minimal requirement for target gRNA recognition and alignment via Watson-Crick base pairing.

The positions of editing blocks were highly conserved among the four subspecies, while they were distributed along the edited mRNA at semi-regular intervals (Figure *3-19*, see Supplementary Figure 4 for gRNA family names). The difference between the start and end positions among the four sub-Saharan *T. brucei* subspecies had mean = 3.3 and 2.8 nt and median = 2.5 and 2.2 nt respectively (Figure *3-20* A).

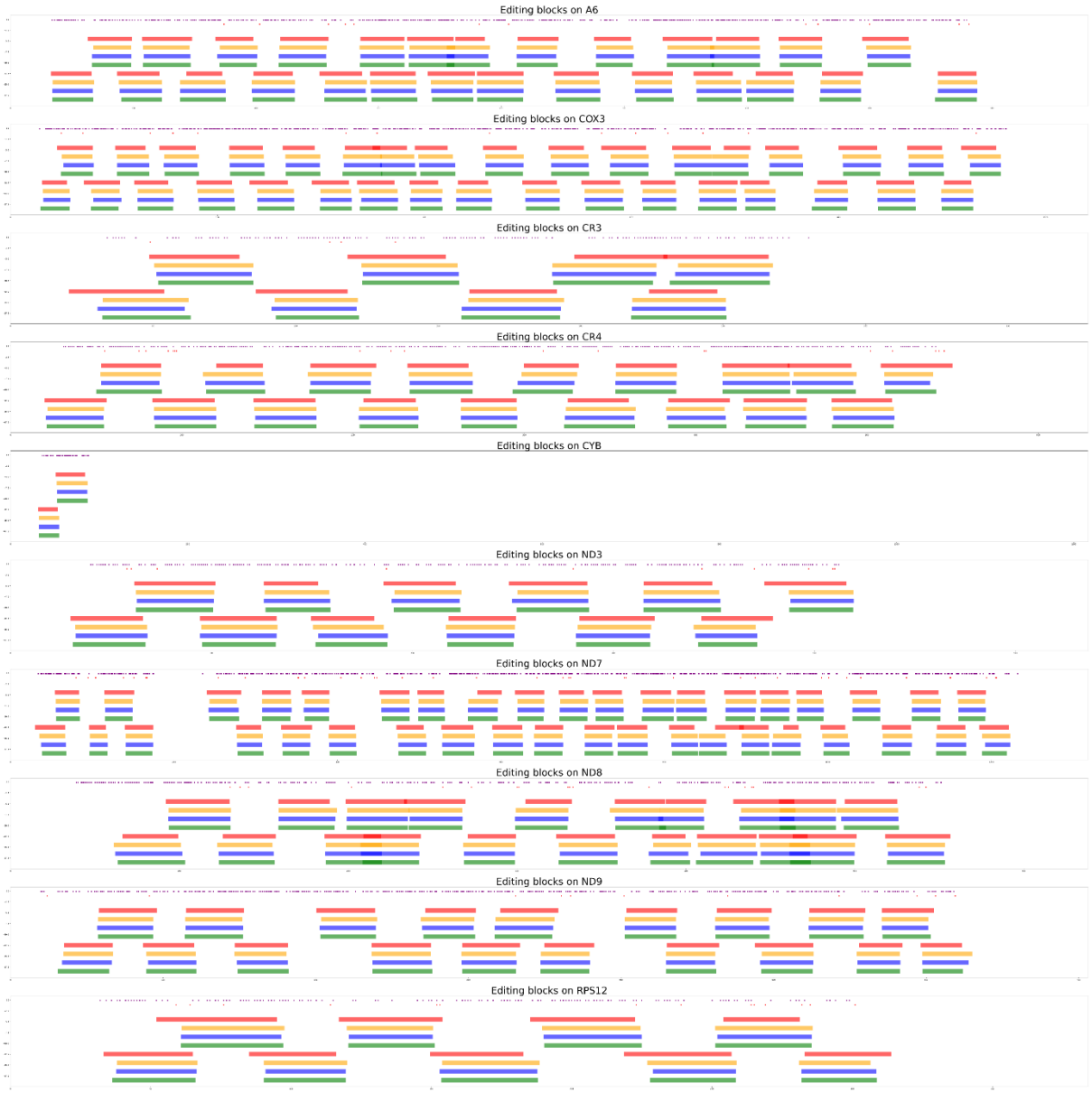


Figure 3-19. **Editing blocks of sub-Saharan *T. brucei* subspecies over all mRNAs**.

The effective editing range of each gRNA family consists of an editing block. The positions of editing blocks are highly conserved in the four *T. brucei* subspecies. In addition, the editing blocks are distributed at semi-regular intervals along the edited mRNAs. Insertions are plotted in purple blocks, and deletions are indicated below with red blocks. The bars at each editing block are drawn in the following order and color scheme: *T. b. gambiense* type 1: red, *T. b. gambiense type 2*: orange, *T. b. brucei*: blue, *T. b. rhodesiense*: green

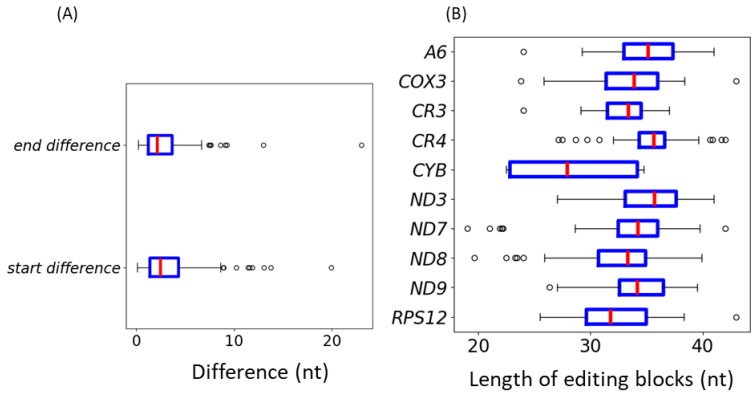


Figure 3-20. **Summary of the conserved features of the editing blocks.**

(A) The difference between the start and end positions of the effective editing ranges of the same gRNA families among sub-Saharan *T. brucei* subspecies. start difference: mean: 3.3, median: 2.5, max: 19.9, min: 0. end difference: mean: 2.8, median: 2.2, max: 23.0, min: 0. (**B)** The effective editing range representative of each gRNA family is calculated by taking the average of the start and end positions of the editing blocks from the four subspecies. Pan-edited mRNAs have conserved average editing block length between 32.45 nt in RPS12 and 35.41 nt in CR4. The minimally edited CYB has two shorter editing blocks with an average length of 28.40 nt. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point

We calculated the average editing block lengths for the edited mRNAs (Figure *3-20* B). The average lengths of the editing blocks of pan-edited mRNAs were conserved and between 32.5 nt in RPS12 and 35.4 nt in CR4 (Table *3-15*). The 3' editing block of the minimally edited CYB was around 34 nt, but the 5' editing block was only around 22 nt (Table *3-15*). Since CYB only has two editing blocks, the mean length is greatly skewed by the short 5' editing block.

Table 3-15. **Summary of descriptive statistics of the length of editing blocks on edited mRNAs**

	mean	median	max	min
A6	35.0	35.2	41.0	24.0
COX3	33.5	33.9	43.0	23.8
CR3	32.8	33.4	37.0	24.0
CR4	35.4	35.6	42.0	27.1
CYB	28.4	28.0	34.8	22.5
ND3	35.2	35.7	41.0	27.0
ND7	33.9	34.2	42.0	19.0
ND8	32.8	33.3	39.9	19.7
ND9	34.4	34.2	39.5	26.3
RPS12	32.5	31.8	43.0	25.5

To investigate the regularity of the distributions of editing blocks of the highly conserved gRNA families, we excluded CYB from the analysis as it contained only two editing blocks. The intervals between the editing blocks were defined as the distance between the start positions of the adjacent families. We observed in the four subspecies an interval distance around 30 nt over all pan-edited mRNAs (Table *3-16*). The standard deviations were > 15 in ND7 due to the distance between the two editing domains. Other mRNAs had standard deviations between 3 and 11. In summary, the editing blocks of sub-Saharan *T. brucei* were arranged around 30 nt apart on average.

Table 3-16. **The intervals between editing blocks on edited mRNAs in sub-Saharan *T. brucei* subspecies**

	***T.b. gambiense* type 1**		***T.b. gambiense* type 2**		***T.b. brucei***		***T.b. rhodesiense***	
	mean	**sd**	**mean**	**sd**	**mean**	**sd**	**mean**	**sd**
A6	27.79	9.82	27.74	9.83	27.74	9.7	27.66	9.43
COX3	26.9	9.58	27.07	7.32	27.13	7.21	27.2	6.93
CR3	29.79	12.31	28.83	9.43	28.94	8.58	28.43	9.34
CR4	28.71	5.81	28.73	5.11	28.78	4.51	28.81	4.83
ND3	31.36	3.37	32.28	6.95	32.33	5.89	32.39	5.81
ND7	30.39	15.52	30.35	15.23	30.39	15.07	30.25	15.18
ND8	22.58	9.55	22.89	10.04	22.88	9.81	22.78	9.75
ND9	31.19	9.68	31.27	10.01	31.31	9.79	31.48	9.63
RPS12	31.22	5.45	30.39	5.15	30.52	5.71	30.72	5.15

Note sd: standard deviation

3.7 Chapter conclusions

We assembled and annotated the kDNA of four sub-Saharan *T. brucei* subspecies. Compared to the groups capable of sexual reproduction, the strictly clonal *T. b. gambiense* type 1 isolates display a highly conserved and unique minicircle population. Given the characteristic minicircle profile and the clonal lifestyle, the ancestral clonal population probably had gone through a series of bottlenecks as the kDNA complexity reduced [337, 339]. The modern-day *T. b. gambiense* type 1 isolates are most likely descendants of a single cell line with reduced kDNA and human serum canonical gRNAs based on the ISSPs

resistance [155]. The minicircle profile has allowed diagnosis using primers targeting specific minicircle sequences [80].

We identified alternative editing patterns specific to *T. b. gambiense* type 1 on COX3 and ND8. Meanwhile, *T. b. gambiense* type 1 had significantly less complete gRNA coverages on all mRNAs except A6, RPS12, and ND3. While a gap shared by all isolates probably masks the difference in gRNA coverage on ND3, A6, and RPS12 are the only two edited mRNAs essential for the BSF parasite. We also detected clonal isolates with highly reduced kDNA complexity only capable of generating fully edited mRNAs of A6 and RPS12.

We annotated the minicircles pooled from the 224 *T. brucei* isolates and assigned the canonical gRNAs based on the initiation site starting positions (ISSPs) into gRNA families that reflected their functional similarities. The gRNA families were highly conserved in *T. brucei*, including the clonal *T. b. gambiense* type 1 which had a unique minicircle population. We identified 189/250 major guide RNA (gRNA) families shared by over 80% of isolates, collectively covering a significant portion of editing sites. The effective editing ranges of the major gRNA families were distributed along the edited mRNAs in semi-regular editing blocks. The length and positions of the editing blocks are conserved among isolates and subspecies, suggesting that mRNA editing occurs in a conserved cascade, probably dictated by the biophysical limitations of the editing apparatus [270, 345].

4 The maxicircle and minicircle dynamics in *T. congolense*

In this project we gave a detailed account of the kDNA composition and editing capacity of *T. congolense*, using three isolatles: IL3000, Kapeya, and UPKZN. After *de novo* kDNA assembly using Ilumina reads, we examined the features of the minicircles and compared the minicircle populations among the three isolates. To characterize the editing capacity of *T. congolense*, we annotated the *T. congolense* maxicircles via alignment to EATRO1125 maxicircles and inferred the unedited mRNAs. With Illumina reads of BSF IL3000 mRNA and PacBio reads of BSF and PCF IL3000 mRNAs, we predicted the fully edited mRNAs and identified some alternative editing patterns. The minicircles were annotated and gRNAs were identified from the minicircles, edited mRNAs, and unedited mRNAs using a published Python pipeline [225].

We wondered if *T. congolense* isolates had conserved editing blocks and if the editing block patterns were similar to what we had observed in *T. brucei*. To answer this question, we grouped gRNAs with overlapped anchors into the same family, including the three *T. congolense* isolates and EATRO1125. The editing blocks were identified using the same method as in Chapter 3. We compared the end positions of the editing blocks and showed that the editing cascade was conserved between *T. brucei* and *T. congolense.*

4.1 Annotation of the maxicircles

4.1.1 Annotation of maxicircle genes on IL3000

The Institute of Tropical Medicine (ITM), Antwerp, isolated kDNA from three *T. congolense* savannah isolates IL3000, Kapeya, and UPKZN, and sequenced the content using the Illumina MiSeq reads 150 bp pair-end method. IL3000 was originally isolated from a bovine in 1966 in Kenya [304, 305], was adapted to lab conditions decades ago, and since then has been widely used in experimental studies across many laboratories. It is therefore generally used as a reference strain for *T. congolense*. Kapeya was isolated from a bovine in Zambia in 2003 [306]. UPKZN was isolated from a buffalo in 2007 in KwaZulu-Natal province, South Africa [307].

Complete and partial maxicircle sequences have previously been reported for *T. congolense* [299, 307]. Our collaborators from ITM supplied an IL3000 maxicircle trimmed to start with the 12S rRNA gene, while the coding region extends up to the end of the ND5 gene at nt 14985. The remaining nucleotides until 27979 nt belong to the variable region. We mapped the published EATRO1125 pre-edited mRNAs onto the maxicircle for preliminary annotation [225]. The PacBio long reads were blasted against the unedited mRNAs, and the hit regions were extracted for alignment by MAFFT [327] with the annotated genes to adjust the gene boundaries. Finally, the Illumina reads were mapped to the pre-edited mRNA to validate the sequences (Table *4-1*).

Table 4-1. **Gene annotation on IL3000 maxicircle.**

Gene	Orientation	Start	End
12S rRNA	Template	1	1168
9s rRNA	Template	1210	1815
ND8	Template	1843	2178
ND9	Complement	2142	2464
MURF5/uS3m	Complement	2480	2718
ND7	Template	2717	3493
COX3	Template	3460	3938
CYB	Template	3897	5020
A6	Template	5038	5400
ND2/MURF1	Complement	5376	6707
CR3	Template	6669	6824
ND1	Complement	6816	7759
COX2	Template	7769	8398
MURF2	Template	8388	9474
COX1	Complement	9469	11117
CR4	Complement	11104	11392
ND4	Template	11468	12780
ND3	Complement	12749	13035
RPS12	Template	12986	13220
ND5	Template	13214	14985

Note: Genes on the template strand are blue. Genes on the coding strand are orange.

4.1.2 SNPs on maxicircle coding regions

We were interested in how the maxicircle coding regions differ among the three isolates. We mapped the kDNA reads of three isolates to the IL3000 maxicircle for SNP detection in the coding region. The average mapping depths for IL3000, Kapeya, and UPKZN were 4673, 4837, and 2402, respectively. The mapped IL3000 reads revealed no homogeneous SNPs and confirmed the published sequence [307]. However, we detected 17 heterogeneous SNPs that indicated a slightly heterogeneous sequence population within the network or cell population (Table *4-2*).

Interestingly, the heterogeneous SNPs were conserved among the three isolates and occurred within two confined regions, between nt 3550 and nt 3722 (within COX3 gene) and between nt 11227 and nt 11237 (in intergenic region between COX1 and CR4). SNPs between nt 3550 and nt 3722 all involved a substitution to guanine and presumably had no impact on gRNA alignment given G-U wobble base pairing, while SNPs between nt 11227 and nt 11237 all involved a substitution to cytosine.

In addition, we detected the deletion of two bases at nt 458 and 459 (within 12S rRNA gene) in Kapeya and UPKZN. There was no extensive deletion as we had observed in sub-Saharan *T. brucei* isolates (Chapter 3). Both Kapeya and UPKZN had an adenine insertion after nt 13207. The insertions occurred in the 3' UTR of RPS12 and did not affect the ORFs.

We detected 46 homogeneous SNPs between the other two isolates and IL3000 over the coding region (Table *4-3*). Kapeya had 28 SNPs, and UPKZN had 30 SNPs. At all positions

where both isolates had SNPs compared to IL3000, the base substitutions were identical in Kapeya and UPKZN. Unlike the heterogeneous SNPs, the nucleotide substitutions did not exhibit any regional preference but seemed random. Most of the SNPs were within the genes of the non-edited 12S rRNA, COX1, ND4, and ND5.

For SNPs in non-edited genes or non-edited regions of minimally edited COX2 and MURF2, we compared the protein sequences (translation table=Mold Mitochondrial) in Kapeya and UPKZN with the sequences in IL3000. Most of the SNPs are synonymous, including all the SNPs in COXI (Table *4-4*). We detected three and seven non-synonymous SNPs in Kapeya and UPKZN, respectively. The isoleucine: valine substitutions probably have little impact on the protein structure.

We were interested in the degree of variability of the *T. congolense* maxicircles. To answer this question, we used the assembly published by Kay et al. (2020), which included 85 *T. congolense* isolates. We blasted the IL3000 maxicircle coding region against the assemblies and identified 66 isolates with SID ≥ 90% and 51 isolates with SID ≥ 95% . An isolate from a bovine in Zambia in 1996 had the lowest SID of 77%. Hence, we concluded that the maxicircle coding regions were conserved among *T. congolense* isolates.

Table 4-2. **Heterogeneous SNPs on three *T. congolense* isolates by read mapping to the IL3000 reference maxicircle coding region (top to bottom: nucleotide positions, alternative bases, proportion of nucleotide in each isolate)**

	3550		3556		3561		3566		3571		3627		3630		3702		3704	
Base %	**A**	**G**	**A**	**G**	**A**	**G**	**A**	**G**	**T**	**G**	**C**	**G**	**T**	**G**	**T**	**G**	**A**	**G**
IL3000	73	21	70	25	72	23	68	26	68	25	75	21	73	20	74	23	73	20
Kapeya	73	20	70	24	72	22	69	24	66	27	75	19	70	24	75	22	72	20
UPKZN	70	23	66	28	69	26	67	28	65	28	71	24	68	28	70	27	70	23

	3711		3717		3721		3722		11227		11231		11234		11237	
Base %	**A**	**G**	**A**	**G**	**T**	**G**	**T**	**G**	**A**	**C**	**T**	**C**	**A**	**C**	**T**	**C**
IL3000	73	22	71	23	75	18	73	23	72	24	67	28	70	27	66	30
Kapeya	73	21	70	23	72	21	74	23	74	23	68	26	72	25	68	26
UPKZN	69	26	66	28	74	21	71	27	71	24	66	28	71	27	67	28

Table 4-3. **Homogeneous SNPs on *T. congolense* Kapeya and UPKZN by read mapping to the IL3000 reference maxicircle coding region.**

SNPs are color coded in the same color scheme as in IGV: A: green, T; red, C: blue, G: orange. If the SNPs are within maxicircle-encoded genes, the genes are labelled on top.

Table 4-4. **SNPs of Kapeya and UPKZN identified against IL3000 maxicircle crypto gene sequences.**

Non-synonymous SNPs in non-edited genes or non-edited regions of minimally-edited genes are highlighted: IL3000 nt: nt (IL3000 aa: alternative aa).

Gene	Kapeya	UPKZN
12S rRNA	494 A:G, 534 G:A	168 A:G, 494 A:G
9s rRNA		292 A:C
ND8		
ND9	193 A:G	
MURF5/uS3m		
ND7		
COX3		
CYB	632 C:T	113 T:A (F:L)
A6	149 A:G	149 A:G
MURF1	226 T:C, 926 A:G(I:V)	926 A:G (I:V)
CR3		
ND1		
COX2	267 T:C	563 T:C
MURF2		679 C:T
COX1	87 C:T, 180 G:A, 306 C:T, 318 A:G, 471 G:A, 633 G:A, 729 T:C, 783 T:A, 1023 T:C, 1062 T:G, 1548 C:T	80 G:A, 258 A:G, 306 C:T, 318 A:G, 684 T:C, 729 T:C, 783 T:A, 1548 C:T
CR4		
ND4	495 C:T (H:Y), 845 G:A, 903 A:G (I:V)	111 A:G (I:V), 177 T:C, 256 C:T (T:I), 417 G:A (V:I), 495 C:T (H:Y)
ND3	132 A:G	68 A:G
RPS12		
ND5	212 C:T, 1535 T:C	150 A:T (T:S), 212 C:T, 572 T:C, 797 T:C, 989 T:C, 1040 C:T

4.1.3 Edited mRNA prediction and alternatively-edited mRNAs

We aimed to predicted the edited mRNAs and identify alternative editing patterns for the subsequent minicircle annotation. Preliminary edited mRNA sequences were predicted via alignment of non-U residues with EATRO1125 edited mRNAs [225] and adjustments of U-indels to maximize the conservation of the ORFs. To confirm the predicted sequences and identify possible alternative editing patterns, crude mitochondrial fractions were purified from BSF form *T. congolense* IL3000, and RNA was extracted and sequenced using Illumina. PacBio reads for mitochondrial transcripts isolated from the BSF and the insect stage epimastigote form (EMF) were also provided from ITM.

First, the Illumina reads were mapped to the pre-edited and edited mRNA versions of the twelve maxicircle-encoded cryptogenes as well as to the eight never-edited RNAs to detect and correct SNPs while preserving the ORFs. The PacBio long reads were blasted against the polished edited mRNAs, and the hit regions were extracted for alignment by MAFFT [327] with the predicted edited mRNAs.

Our inspection of Illumina read mapping and PacBio read alignments with the edited mRNA predictions suggested alternative editing patterns for some mRNAs. Two versions of A6 and ND8 in *T. b. brucei* EATRO1125 have been reported previously [225]. The IL3000 transcriptome data used in this study detected alternative editing for A6 and ND3 but not ND8.

As in *T. brucei*, the A6 alternative editing patterns were located in the 3’ UTR and would not affect the protein sequence (Figure *4-1* A). A6_v1 had four uridine insertions between nt 725 and nt 732, while A6_v2 had five uridine insertions between nt 725 and nt 733 (the stop codon (UAG) starts at nt 699). The proportion of A6_v2 PacBio reads increased from 15.41% (98/636) in BSF to 63.83% (60/94) in EMF, probably due to stage-specific regulations via post-transcriptional editing and a mechanism probably shared between African trypanosomes. However, as only a single biological sample had been prepared for each life cycle stage, statistical significance cannot be tested and any comparisons need to be considered as preliminary observations.

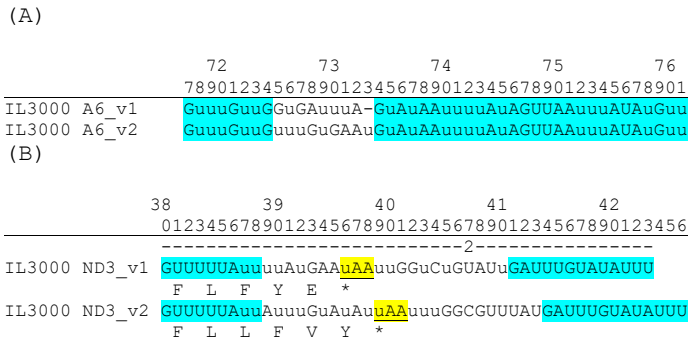


Figure 4-1. **Alternatively edited A6 and ND3 in IL3000.**

Numbers show the ones and tens digits of the nucleotide positions. Identical sequences are highlighted in blue. Alternatively inserted uridines are colored in red. **(A)** A6 mRNAs appear to be alternatively edited between nt 725 and nt 733 in IL3000. The alternative editing occurs after the stop codons of A6 mRNAs. **(B)** ND3 mRNAs appear to be alternatively edited between nt 380 and nt 413 in IL3000. The alternative editing results in slightly different protein sequences. ND3_v2 is translated to a protein sequence one amino acid longer than the one of ND3_v1. A two-uridine deletion in edited ND3_v1 is indicated in the line above the edited mRNAs.

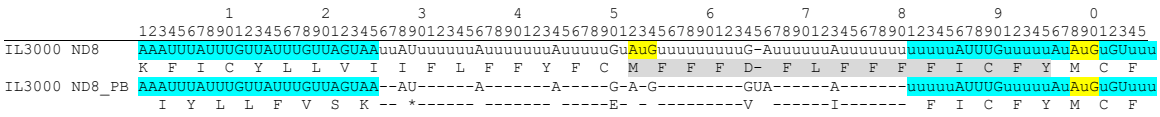
The edited ND3 sequence in *T. brucei* has been suggested to be highly heterogeneous over the 3' most editing sites, and no consensus alternative editing pattern could be determined for EATRO1125 [225]. Only one version of ND3 (ND3_v1) mRNA was observed in the *T. congolense* EMF PacBio reads. However, the editing pattern of ND3_v1 was shared by only 17.0% (18/106) unique reads in BSF PacBio data, while further inspections revealed that 66.3% (66/106) of reads had a different editing pattern (ND3_v2). We detected Illumina reads with either version of editing. The alternative editing occurred between nt 380 and nt 413 and would lead to different sequences and lengths of the C-terminal end of the protein (Figure *4-1* B). Protein folding simulation should be applied to try and predict whether the difference in the C-terminus affects protein structure and other biochemical properties.

4.1.4 Partially-edited mRNAs

Besides alternative editing, we also detected PacBio reads with partial editing that still maintained ORFs. ND8 had no Illumina read coverage for the first 41 nt, while reads extending into the region between nt 50 and nt 41 contained unedited or partially edited sequences with multiple mismatches. In addition, sequence alignment of the BSF and EMF PacBio reads with hits to the edited ND8 mRNA showed no evidence of editing on ND8 in IL3000 until nt 80 of the predicted 'fully edited' mRNA. Hence, the 5' editing sites had no read support and solely relied on U-stripped alignment with EATRO1125 edited ND8. The lack of evidence for editing might be due to the general lower coverage on the 5' end of edited mRNAs given the 3' to 5' directionality of editing. Nevertheless, we could not dismiss the possibility that IL3000, or *T. congolense* in general, no longer edited the 5' most editing sites.

The putative fully edited ND8 mRNA generated an AUG at nt 52 by uridine insertion. Another AUG occurred at nt 97 of the fully edited ND8 mRNA or nt 54 of the less edited version. If translation started from the second AUG instead, this would result in a slightly truncated but otherwise highly conserved protein sequence (Figure *4-2* A). The second start codon exists at the DNA level on some kinetoplastids that have minimally edited ND8 and has been proposed to act as an alternative for translation initiation that does not change the protein function significantly [208]. As the less edited version would not require additional or different gRNAs, we used the fully edited ND8 mRNA in the subsequent gRNA annotation.

(A)



(B)

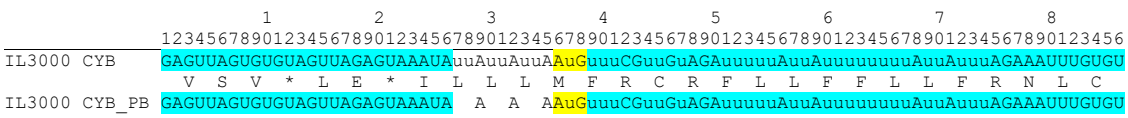


Figure 4-2. **Comparison of the initially predicted fully edited mRNAs and the partially edited consensus of PacBio reads.**

(A) No PacBio reads were founds that contained editing events between nt 26 and nt 79 of the fully edited ND8. An alternative start codon at nt 97 occurs in both versions. The postulated start codons are highlighted in yellow. Amino acids truncated by removal of the editing were shaded. **(B)** CYB is minimally edited on the 5' end. Transcripts with partial editing (without the six 5' most uridine insertions) on the 5' editing block were abundant in BSF and EMF. The postulate start codons are highlighted in yellow. Numbers show the ones and tens digits of the nucleotide positions. Identical sequences are highlighted in blue. Alternatively inserted uridines are colored in red

Although we detected PacBio reads for BSF and EMF that were consistent with the predicted fully edited CYB sequence, we also observed partially edited transcripts without the six uridine insertions in the 5' UTR (Figure *4-2* B). The shorter edited mRNAs generated the same AUG as the longer version and presumably the same protein, and it could therefore either represent an alternatively edited version or an editing intermediate. As the shorter version would not require additional or different gRNAs, we only used the longer CYB mRNA in the subsequent gRNA annotation.

4.1.5 Illumina read coverage on predicted mRNAs in IL3000

We assessed the Illumina read support for edited and unedited mRNAs via read mapping. Mapping transcriptome reads resulted in ≥ 95% coverage for all RNA sequences except the 12S rRNA, which had 79.1% coverage with a gap from nt 1 to nt 246 (Table *4-6*). The read coverages of edited mRNAs were less complete. We observed ≥ 90% coverages for COX2, CR3, CYB, MURF2, both versions of ND3, ND7, ND8, and RPS12. Nevertheless, A6_v1 and A6_v2 were not covered from the 5' end to nt 411 and nt 380 respectively. COX3, CR4, and ND9 had no coverage from the 5' ends to nt 218, nt 385, and nt 223 respectively. The declining read coverage towards the 5' ends on edited mRNAs agrees with the 3' to 5' directionality of mRNA editing and the observation that most of the transcriptomes are partially edited [249, 259, 279, 297]. Meanwhile, stage-specific gene expression due to different metabolic activities entails that some mRNAs are not actively edited in BSF *T. congolense* [42, 289].

Table 4-5. **Summary of transcriptomic Illumina read mapping for IL3000**

(Never-edited RNAs were included in both mapping)

Total reads	47299448
Mapped to pre-edited IL3000 mRNAs	257349
Mapped to edited IL3000 mRNAs	316460

Table 4-6**. Summary of Illumina read mapping to unedited IL3000 rRNAs and mRNAs (never-edited RNAs are shaded)**

unedited mRNA	Transcript length	Read count	Coverage%	Mean depth	Mean baseq
12S_rRNA	1168	594	79.1	76.2	35.0
9S_rRNA	607	20	95.2	3.2	33.1
A6	363	259	100.0	100.8	31.1
COX1	1650	2342	100.0	214.4	35.1
COX2	630	2227	100.0	525.4	35.4
COX3	479	371	99.8	114.0	31.6
CR3	156	25	100.0	21.5	35.1
CR4	289	134	100.0	68.6	32.7
CYB	1210	30981	100.0	3832.6	35.0
MURF1	1333	1509	100.0	81.0	34.2
MURF2	1209	336	96.6	40.8	34.9
MURF5	240	24	99.6	13.4	34.6
ND1	945	25365	100.0	4124.9	35.0
ND3	287	3439	100.0	1752.6	33.7
ND4	1314	4561	96.6	518.6	34.8
ND5	1773	3993	100.0	338.7	34.9
ND7	777	118391	100.0	22558.7	34.3
ND8	336	44145	100.0	19375.6	33.4
ND9	323	4407	100.0	2092.2	33.7
RPS12	235	14226	100.0	8809.1	34.0

Table 4-7**. Summary of Illumina read mapping to polished edited IL3000 mRNAs (never-edited mRNAs are shaded)**

Edited mRNA	Transcript length	Read count	Coverage%	Mean depth	Mean base q
12S_rRNA	1168	595	79.1	76.3	35.0
9S_rRNA	607	20	95.2	3.2	33.1
A6_v1	791	135	59.2	24.9	34.7
A6_v2	792	139	61.5	24.5	34.2
COX1	1650	2341	100.0	214.3	35.1
COX2	634	2230	100.0	526.1	35.5
COX3	1003	132	77.8	19.2	34.6
CR3	295	47	100.0	24.1	34.3
CR4	572	7	44.6	1.0	35.6
CYB	1160	30723	99.2	3964.7	35.0
MURF1	1333	1509	100.0	81.0	34.2
MURF2	1112	100	93.6	13.4	33.6
MURF5	240	24	99.6	13.4	34.6
ND1	945	25365	100.0	4124.9	35.0
ND3_v1	454	279	98.0	91.2	34.0
ND3_v2	457	303	97.6	98.2	34.2
ND4	1314	4561	96.6	518.6	34.8
ND5	1773	3992	100.0	338.6	34.9
ND7	1260	111289	100.0	12891.5	34.9
ND8	550	122616	94.2	33125.7	34.9
ND9	619	451	64.1	105.4	34.8
RPS12	341	9602	100.0	4190.6	34.1

4.2 PacBio read analysis

4.2.1 PacBio read coverage on predicted mRNAs

We assessed the support of PacBio reads for edited and unedited mRNAs via read mapping. The PacBio dataset included 82044 and 95524 consensus reads generated from BSF and EMF IL3000 mitochondrial RNA extractions, respectively. The BSF sample has a higher count for reads ≤1000 nt, while the EMF sample has a higher count for reads between 1000 and 4000 nt. The difference was probably generated during mRNA extraction or library preparation, which failed to capture the shorter reads in the EMF sample (Figure *4-3* A).

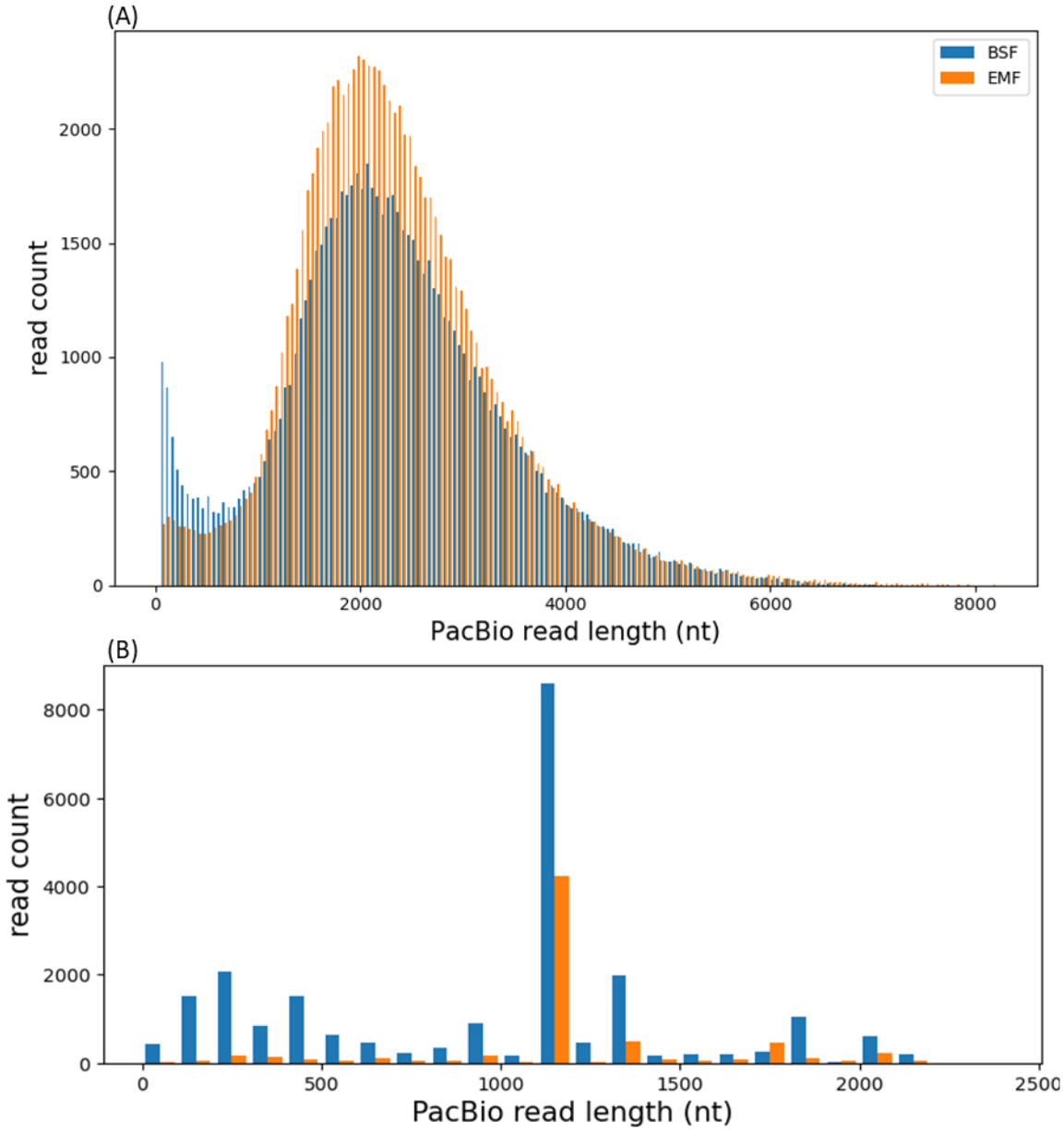


Figure 4-3. **Length distribution of PacBio reads in BSF and EMF samples.**

(A) Length distribution of unique PacBio reads. The BSF sample has a higher count for reads ≤1000 nt, while the EMF sample has a higher count for reads between 1000 and 4000 nt. (**B)** Length distribution of PacBio reads (actual number) with hits to the pre-edited or edited IL3000 mRNAs in BSF and EMF samples. The BSF sample has a much higher count for reads ≤1000 nt, while the difference becomes less pronounced for reads ≥ 1000 nt.

The BSF and EMF datasets contained 578 and 265 unique reads (22798 and 6736 actual reads before collapsing) with blastn hits to unedited and edited IL3000 mRNAs, respectively (with a cut-off of over 90% sequence identity). Although the EMF sample contained more unique reads, the BSF sample had over twice as many actual reads with hits to the IL3000 mRNAs. Comparing the distribution by read length showed that although the BSF sample had more reads at almost all length ranges, the most striking difference occurred over reads ≤ 1000 nt, reflecting the overall lack of shorter reads in the EMF sample (Figure *4-3* B). Both samples contained the most reads around 1200 nt. As the actual read counts are expected to reflect the relative abundance of transcripts in reality, we used the actual read count in the subsequent analysis.

After identifying the blast hits against the pre-edited and edited mRNAs, we annotated the PacBio reads by recording the coordinates of the hits on each read. As mentioned above, the lack of biological and technical replicates precluded making conclusive statements on stage-specific gene regulation via post-transcriptional editing. Nevertheless, the analysis of long reads helped validate the edited mRNA predictions and revealed some intriguing trends.

Mapping the PacBio hit sequences back to the corresponding unedited (or never-edited) and edited references revealed complete or nearly so transcriptome coverages over the mRNAs in at least one life stage (Table *4-8*, Table *4-9*). Due to the length of PacBio reads, reads with minor differences in alternative editing had hits on both versions of edited mRNAs. Hence, we did not distinguish the versions when reporting the PacBio coverage. The alignments were visually examined and cases of alternative or partial editing have been discussed in 4.1.3 and 4.1.4. Overall, the mapping validated the maxicircle annotations and mRNA predictions. We did not observe complete read coverage on both rRNAs and pre-edited ND3, ND7, and ND9 in both life stages. Edited complex I subunits CR3, CR4, ND3, ND7 (>99%), ND8, and ND9 did not have complete read coverage in both life stages.

Table 4-8**. Summary of PacBio read mapping to polished pre-edited IL3000 mRNAs (never-edited mRNAs are shaded)**

		BSF	EMF	BSF	EMF	BSF	EMF
	length	**Read count**		**Coverage%**		**Mean depth**	
12srRNA	1167	1029	86	98	98.1	287.3	28.5
9srRNA	606	161	12	89.9	85	92.6	8
A6	363	296	57	100.0	99.2	114.4	23.9
COX1	1649	227	556	100.0	100.0	196.2	527.9
COX2	630	25	85	100.0	100.0	20.8	81.1
COX3	479	6542	4218	100.0	100.0	616.2	362.1
CR3	156	644	58	100.0	93.6	167.6	16.3
CR4	289	649	435	100.0	100.0	159.7	70
CYB	1210	8011	4302	100.0	100.0	7233.5	3938.2
MURF1	1332	139	14	100.0	100.0	101.2	13.9
MURF2	1209	1039	178	100.0	100.0	858.7	93.8
ND1	944	84	44	100.0	100.0	76.6	43.8
ND3	287	111	0	94.8	0	36.5	0
ND4	1313	2074	445	100.0	100.0	2056.1	442.2
ND5	1772	2491	456	100.0	100.0	2273.2	438
ND7	777	441	77	99.1	87.5	161.3	19.2
ND8	336	2177	71	100.0	98.8	566.3	22.8
ND9	323	273	18	97.8	98.5	113	10.2
RPS12	235	1246	373	100.0	100.0	670.6	259.4
uS3m	239	0	2	0	100.0	0	2

Table 4-9**. Summary of PacBio read mapping to polished edited IL3000 mRNAs**

		BSF	EMF	BSF	EMF	BSF	EMF
	length	**Read count**		**Coverage%**		**Mean depth**	
A6	791	598	80	100.0	100.0	343.8	49.6
COX2	634	25	85	100.0	100.0	20.8	81.1
COX3	1003	7915	2232	100.0	100.0	1589.6	293.9
CR3	287	298	25	96.5	96.5	253.1	20.2
CR4	572	6	0	72.2	0	4.3	0
CYB	1154	7986	4296	100.0	100.0	7515.8	3819.5
MURF2	1112	1035	100	100.0	100.0	986.1	95.5
ND3	454	106	12	98.2	35	36.9	4
ND7	1260	1146	52	99.4	99.2	526.6	14.3
ND8	550	1445	44	93.8	92.4	620.2	16.8
ND9	619	471	6	93.4	68.7	182.6	2.6
RPS12	341	1897	85	100.0	100.0	542.5	38.3

4.2.2 Preliminary investigation of stage-specific RNA editing with PacBio transcriptomics

Although only a single sample was available for each life cycle stage, precluding a reliable quantitative comparison, we reasoned that such a comparison might provide preliminary

insight into the possibility of stage-specific editing events and mitochondrial gene expression.

The BSF sample had more complete read coverage than the EMF sample on all edited mRNAs. The BSF sample almost always had higher read depth than EMF, except for the minimally edited COX2 (Table *4-8*, Table *4-9*). However, despite the difference in read counts, the overall shapes of read depth over mRNAs stayed highly similar in BSF and EMF samples for both pre-edited and edited mRNAs (Supplementary Figure 5,6). For instance, we observed two read depth peaks (more pronounced in EMF than BSF) in ND7 corresponding to the two editing domains with independent initiation sites of editing, while the pan-edited A6 with a single 3' initiation site showed the typical declination in read depth from 3' to 5' end (Figure ***4-4***).

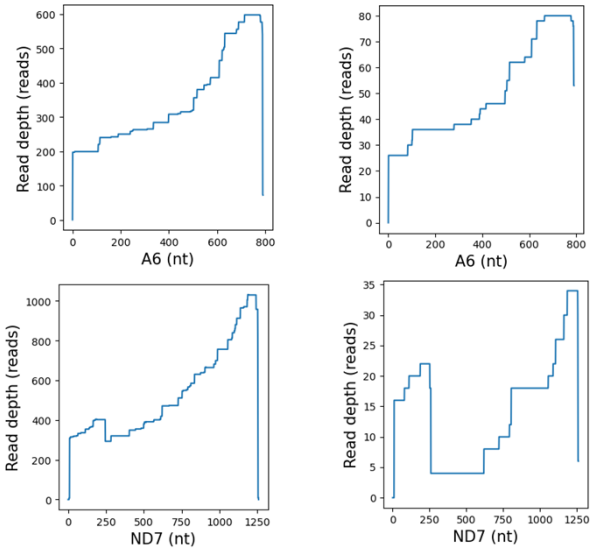


Figure 4-4. PacBio read depth on for edited A6 and ND7 IL3000 in BSF (left) and EMF (right) IL3000.

Hits from PacBio reads detected by blastn were aligned to the respective mRNAs. The read depth distributions were conserved in BSF and EMF, although the BSF sample had much higher read depth.

The metabolism of BSF *T. congolense* resembles an intermediate between BSF and insect-stage *T. brucei*, which suggests that probably the maxicircle-encoded genes normally not required in BSF *T. brucei* are actively transcribed and edited in BSF *T. congolense*. It explains the detection of fully edited reads in BSF but not the higher coverage or read depth. Alternatively, the poorer coverage in EMF may simply reflect differences in sample preparation, which may 'scale down' the read detection uniformly in EMF while preserving the similarity in read depth distributions.

Among reads with hits to edited or pre-edited mRNAs, reads with hits to CYB or COX3 mRNAs were the most abundant, although we had no explanation for their highly activated transcription. In contrast, some maxicircle-encoded transcripts had low steady state or editing levels in BSF and EMF. The blastn hit coverage over 9s rRNA was incomplete in both life stages, missing ~60 nt on the 3' end. We did not detect the never-edited uS3m (previously known as MURF5) in BSF and only captured one read with full alignment in EMF. In contrast, reads with 12S rRNA and pre-edited and edited RPS12 (uS12m) mRNAs were

readily detectable at both life stages. As the ribosomal subunits were expected to be essential at all life stages, the absence of nearly complete 9s rRNA and uS3m reads probably suggested that they were quickly degraded after translation or incorporated into the ribosomes shortly after transcription, so that complete transcripts were difficult to detect.

4.2.3 Preliminary investigation of features of the reads

Over 50% of the reads contained hits to both fully edited and unedited mRNAs; hence, they were considered partially edited (Table *4-10*). We were interested in the cases where the editing had generated junctions, i.e. transcripts with partially edited regions between fully edited 3' parts and unedited 5' parts. We filtered for reads with hits to pan-edited mRNAs and identified 14942 and 5409 reads for BSF and EMF, respectively (Table 13). Of these, 10020 and 2439 reads had junctions, i.e. they came from transcripts that contained edited and unedited sequences of the same mRNAs joined by a region that did not exhibit the canonical editing pattern.

Junctions probably indicate editing in progress or editing errors that hindered alignments of downstream gRNAs [249]. Note the counts of reads with junction in EMF (2439) did not equal the sum of counts for individual mRNAs (2443), because two transcripts spanned from COX3 to A6 and contained junctions on COX3 and A6 mRNAs. This is consistent with previous observations that the cleavage and editing of mRNAs can occur independently, i.e. editing may initiation on different mRNAs before cleavage [221].

Table 4-10. **Counts of reads with hits to edited and unedited mRNAs in BSF and EMF IL3000**

Type of hits	BSF	EMF
Only edited	3221	157
Only unedited	5953	1701
Both edited and unedited	13624	4874
Total	22798	6732

Notes: Only edited: reads with BLASTN hits for fully edited mRNAs only. Only unedited: reads with BLASTN hits for unedited mRNAs only. Both edited and unedited: reads with BLASTN hits for both fully edited and unedited versions of a transcript.

Table 4-11. **Counts of reads with junctions of editing on pan-edited mRNAs in BSF and EMF IL3000**

Edited mRNA	BSF	EMF
Hits on pan-edited	14942	5409
Only unedited	1695	2809
Only Edited	3227	161
With junction	10020	2439
A6	250	43
COX3	6344	2127
CR3	289	25
CR4	6	0
ND3	14	0
ND7	307	32
ND8	1441	44
ND9	224	6
RPS12	1120	81

Notes: Hits on pan-edited: total number of reads with BLASTN hits for pan-edited mRNAs, regardless of editing status. Unedited only: reads with BLASTN hits for unedited pan-edited mRNAs only. Edited only: reads with BLASTN hits for fully edited pan-edited mRNAs only. With junction: reads with BLASTN hits for both unedited and fully edited versions of a transcript. Each mRNA: individual counts for reads with editing junctions for each pan-edited mRNA species.

We detected 32.6% and 70.9% of reads that contained hits to two different mRNAs, regardless of the editing status, in BSF and EMF, respectively, while only 0.1% and 0.3% of reads, respectively, contained hits to three genes (Table 10). The higher percentage of reads with multiple mRNAs in the EMF sample was due to its lack of shorter reads that most likely contained transcripts from a single mRNA (Figure *4-3* A). As expected, most reads with multiple hits concerned genes encoded adjacent to each other on the maxicircle (Table *4-13*). Unexpectedly, some reads (< 0.1%) contained hits to genes not adjacent to each other.

Trypanosomatid maxicircle-encoded mRNAs have been reported to be generated as polycistronic precursors [222, 223]. More recently it was suggested that they are transcribed as independent units [224]. Although the individual mRNAs and rRNAs are independently synthesized, they extend at the 3' ends and may cross the boundary of the adjacent genes, which is concordant with the reads with hits to multiple mRNAs that we observed.

Table 4-12. **Counts of reads with matches to different genes**

Number of hits to different genes within the same PacBio read	BSF	EMF
1	15345 (67.3%)	1941 (28.8%)
2	7443 (32.6%)	4773 (70.9%)
3	10 (0.1%)	18 (0.3%)
Total	22798	6732

Table 4-13. **Counts of the top three most abundant polycistronic transcripts**

Polycistron	BSF	EMF
COX3-CYB	6027 (81.0%)	3991 (83.6%)
RPS12-ND5	1189 (16.0%)	255 (5.3%)
CR4-COX1	140 (1.9%)	407 (8.5%)

4.3 Comparison of maxicircle encoded gRNAs in trypanosomatids

So far, two maxicircle gRNAs (gCOX2, gMURF2) have been described in the salivarian trypanosome *T. brucei* [225]. Nine maxicircle-encoded gRNAs are found in *Leishmania tarentolae*, eight directing U-indels on CYB, COX2, MURF2, ND7, and RPS12, and one remaining unassigned [258, 259]. Seven maxicircle-encoded gRNA genes with conserved synteny to those in *L. tarentolae* have been described in *C. fasciculata* [346]. In addition, 19 to 21 putative gRNA genes were identified within the maxicircles of *Leishmania peruviana* and *Leishmania braziliensis* [66].

We generated a schematic representation of the maxicircle-encoded gRNAs for the five species from the literature and *T. congolense* (to be discussed in this section) to illustrate the conservation of synteny and the positions of gRNA genes relative to other maxicircle-encoded rRNAs and protein-coding genes (Figure *4-5*). The gRNAs in Figure *4-5* are labeled with Roman numerals or positions on edited mRNA according to the study in which they were first described [66, 258, 346].

To look for maxicircle encoded gRNAs, we performed multiple-sequence alignment using MAFFT for the reference maxicircles of IL3000, *T. brucei* EATRO1125, *T. brucei* Lister 427 [202], *C. fasciculata [346]*, *L. tarentolae* [258], *L. peruviana* HR78 and LCA04 [66], and *L. braziliensis* LC1412 [66] to identify regions homologous to the reported gRNAs.

The IL3000 maxicircle starts at the first nt of the 12S rRNA gene, with a coding region up to nt 14985, and variable region up to nt 27979. As the name indicates, the variable regions from different species, subspecies, and even isolates of the same species are not well conserved and align poorly against each other. Consequently, alignments in the variable regions provide little guidance for gRNA detection. We examined the maxicircle alignment for evidence of additional maxicircle-encoded IL3000 gRNAs within the coding region, from the start of the 12s rRNA gene to the end of the ND5 gene.

Ten gRNA genes reported in the literature are located within the coding region of *L. braziliensis* and *L. peruviana*. The four gRNA genes in the coding regions of *L. tarentolae* and *C. fasciculata* were homologs of four of the ten genes [66]. We examined the maxicircle alignment at the ten genes for evidence of sequence conservation. Nevertheless, 8/10 genes overlap with the non-edited regions of other maxicircle coding genes (Figure 4-5). Given the conserved genetic background, sequence conservation in IL3000 did not necessarily indicate gRNA function (Figure 4-6). In addition, since the gRNA genes overlapped other genes, sRNA mapping could be explained as fragments of other transcripts unless they were modified with untemplated U-tails, except for the cis-editing COX2.

We observed a small portion (6388/154707092) of sRNA reads mapping to *T. congolense* IL3000 maxicircle by local alignment using Bowtie 2 with no more than two mismatches. The mapping covered 57.55% of the maxicircle genome with a mean depth of 9.58 reads per nt.

We detected two sRNAs with untemplated U-tails for regions within the ND2 gene aligned with gND9_304-355 and gGR4_251-304 in *L. braziliensis* and *L. peruviana*. The region aligned with gND9_304-355 of *L. braziliensis* and *L. peruviana* spans over 6346 to 6397 nt on IL3000 maxicircle, with one read mapped from 6350 to 6399 nt without mismatch (Figure *4-7* A).

For convenience, we named the aligned region on IL3000 maxicircle gND9_IL3000. The region aligned with gGR4_251-304 of *L. braziliensis* and *L. peruviana* spans over 6492 to 6545 nt on IL3000 maxicircle, with a read mapped from 6482 to 6518 nt (Figure *4-7* B). Since the sRNA started before the putative coding area and was truncated despite the U-tail, we did not consider the mapping affirmative.

We did not detect alignments between the edited ND9 mRNAs of *L. peruviana* HR78 and IL3000, so the target region of gND9_IL3000 was identified based on ORF alignment. However, mapping gND9_IL3000 onto the target region revealed no evidence for gRNA alignment (Figure *4-8* B). It was not aligned to other sequences on edited IL3000 mRNAs. Given the detection of the transcript with an untemplated U-tail, we speculated that the gND9_IL3000 gene was once functional but has since degraded.

We examined the alignment on edited mRNA for the cis-editing COX2 gRNA. In *T. brucei*, *L. tarentolae, C. fasciculata*, *L. peruviana*, and *L. braziliensis*, four uridines are inserted into COX2 via cis-editing directed by its 3' UTR[66, 213, 225, 258, 346]. We detected the expected four U-insertions in *T. congolense* in Illumina and PacBio transcriptome data and identified the COX2 gRNA sequence from nt 8396 to nt 8406 of the IL3000 maxicircle (Figure 4-6). Alignment of the trypanosomatid maxicircles revealed sequence conservation around the putative gRNA in the 3' UTR (Figure 4-8 A).

Two gRNA genes, gMURF2-II and gCYB-II, are located mostly in the intergenic regions. gMURF2-II is encoded within the intergenic region between GR4 and ND4 and extends into the beginning of the ND4 genes (Figure 4-6). Unlike *T. brucei* and *Leishmania* which require at least two gRNAs to fully edit MURF2, *T. congolense* covered all editing sites with a single gRNA encoded on nt 11425 to nt 11468 of the maxicircle template strand (Figure 4-8 C). Although gRNAs of *T. congolense* and *T. brucei* had the same first 6 bases in the anchor regions, the guiding sequences differed because of a slight inter-specific variation of MURF2 editing pattern and the interchangeability of A-G given wobble-base-pairing. Notably, the generation of the putative start codon AUG relied on editing in both *T. brucei* and *T. congolense*. We detected 46 reads from the sequenced sRNAs that contained the complete candidate gRNA sequence with poly-U tails of different lengths.

In contrast, the gCYB-II gene overlaps the last 11 nt on the 9S rRNA gene and extends into the intergenic region upstream of ND8 in *Leishmania* (Figure 4-6) [66, 258]. In CfC1, the intergenic region is much longer probably due to an insertion, and the gCYB-II in CfC1 is fully contained in the intergenic region without overlapping the 9S rRNA gene [346]. The intergenic regions are too short in *T. brucei* and *T. congolense* in comparison to encode additional gRNA.

Besides the MURF2 and COX2 gRNAs, we identified the only ND7 3' editing domain initiation gRNA in *T. congolense* at nt 15086 - 15138 of the template strand, just downstream of the end of ND5 and at the beginning of the maxicircle variable region. That stretch of the variable region does not show homology with maxicircle sequences of other species examined (Figure 4-6). Careful inspection of the equivalent *T. brucei* and *T. congolense* maxicircle regions showed no evidence of a degraded homologous gRNA gene in *T. brucei*. Although ND7 initiation gRNAs (gND7-I) for the 5' editing domain had been reported near the end of ND5 in *L. tarentolae* and *C. fasciculata*, the ND7 mRNA does not have a 3' editing

domain in these species [258, 346, 347]. Hence, we believe the ND7 initiation gRNAs in *L. tarentolae* and *Crithidia* bear no functional homology to the *T. congolense* ND7 initiation gRNA. Although, *T. brucei*, *L. peruviana*, and *L. braziliensis* do not appear to encode ND7 gRNA near the end of ND5, these and other species encode other gRNAs in this region [66]. Perhaps that region of the maxicircle, for a yet unknown reason, lends itself to the expression of gRNAs.

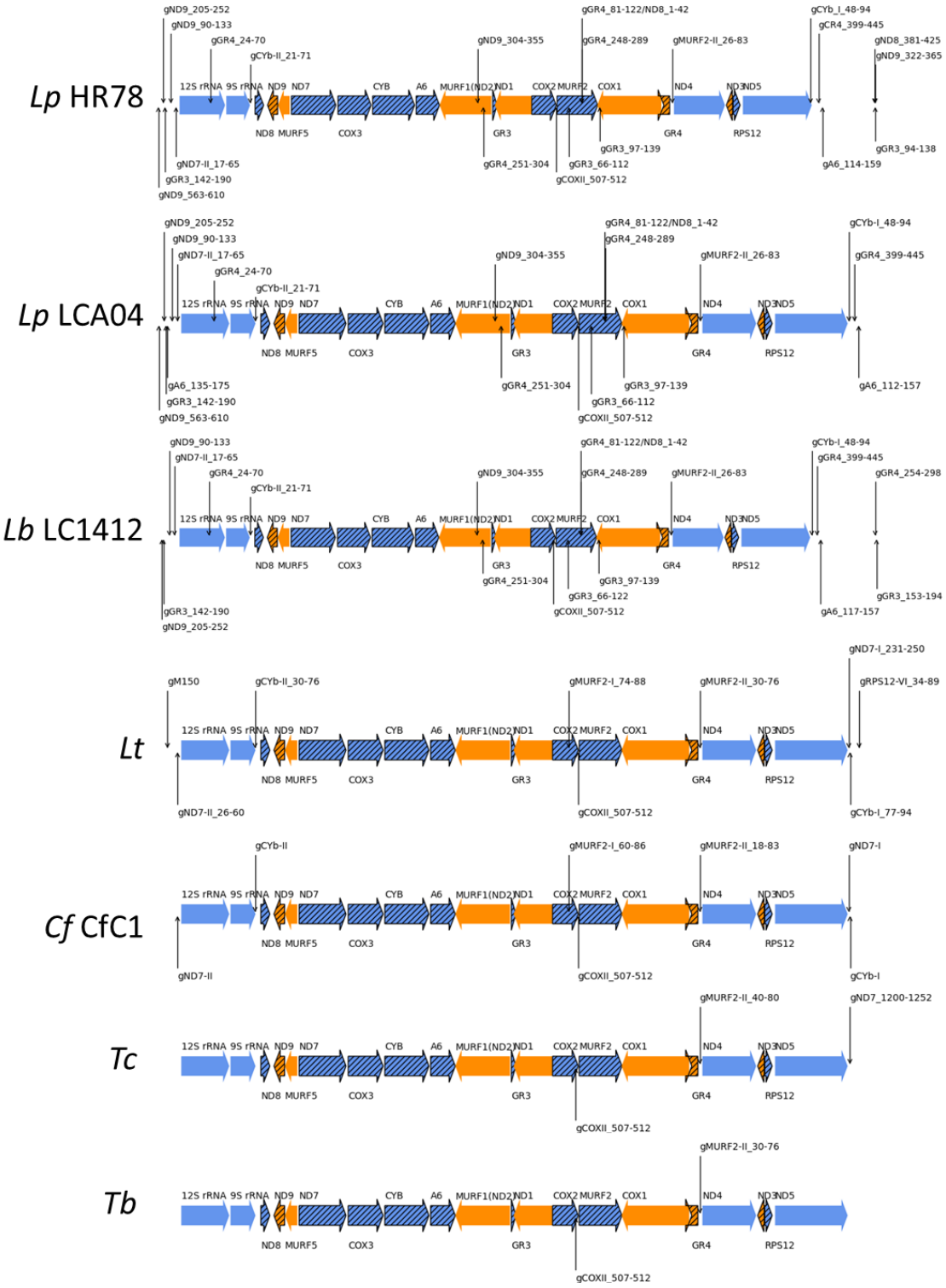
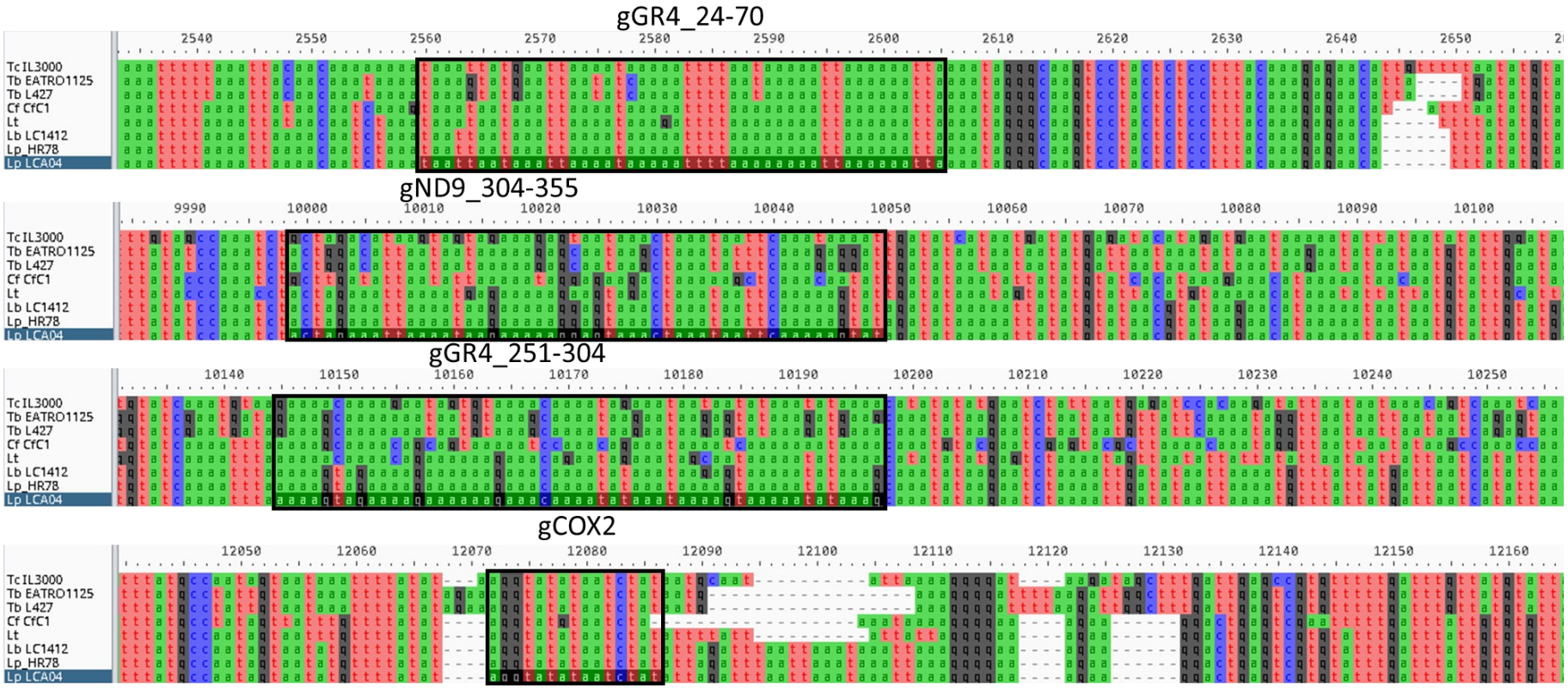
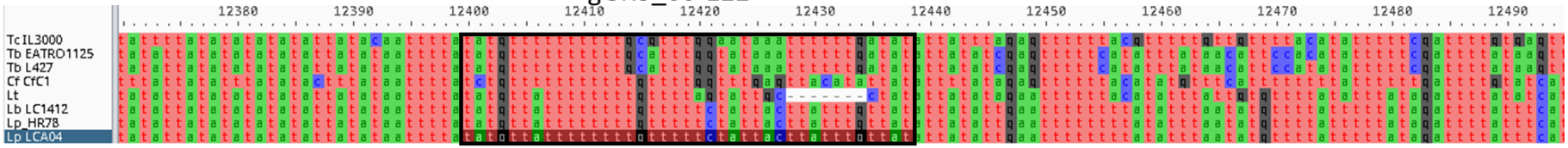


Figure 4-5. **Comparison of maxicircle-encoded gRNAs in trypanosomatids.**

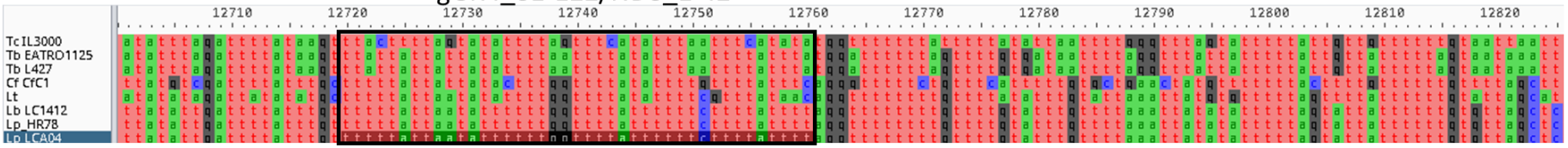
The synteny of maxicircle encoded rRNA and protein-coding genes and of some gRNA genes is conserved. The gRNAs are labeled with Roman numerals or nt position on edited mRNA based on the study in which they are first described. Abbreviation and references: *Tb*: *T. brucei* EATRO1125 [225], Cf: *C. fasciculata* CfC1 [346], Lt: *L. tarentolae* [258], *Lp*: *L. peruviana* HR78 and LCA04 [66], *Lb*: *L. braziliensis* LC1412 [66].


gGR4_24-70
gND9_304-355
gGR4_251-304
gCOX2
Tc IL3000
Tb EATRO1125
Tb L427
Cf CfC1
Lt
Lb LC1412
Lp_HR78
Lp LCA04

gGR3_66-122

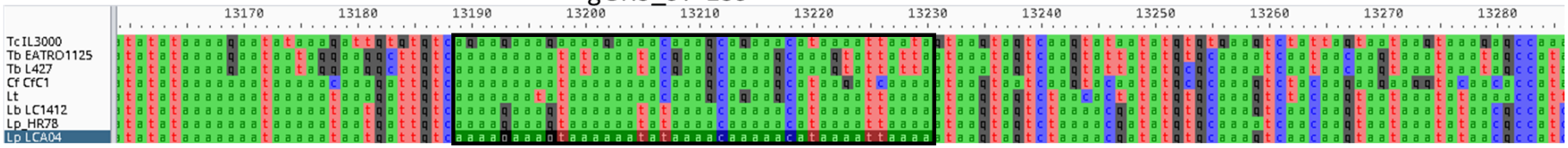


gGR4_81-122/ND8_1-42



gGR4_248-289

gGR3_97-139



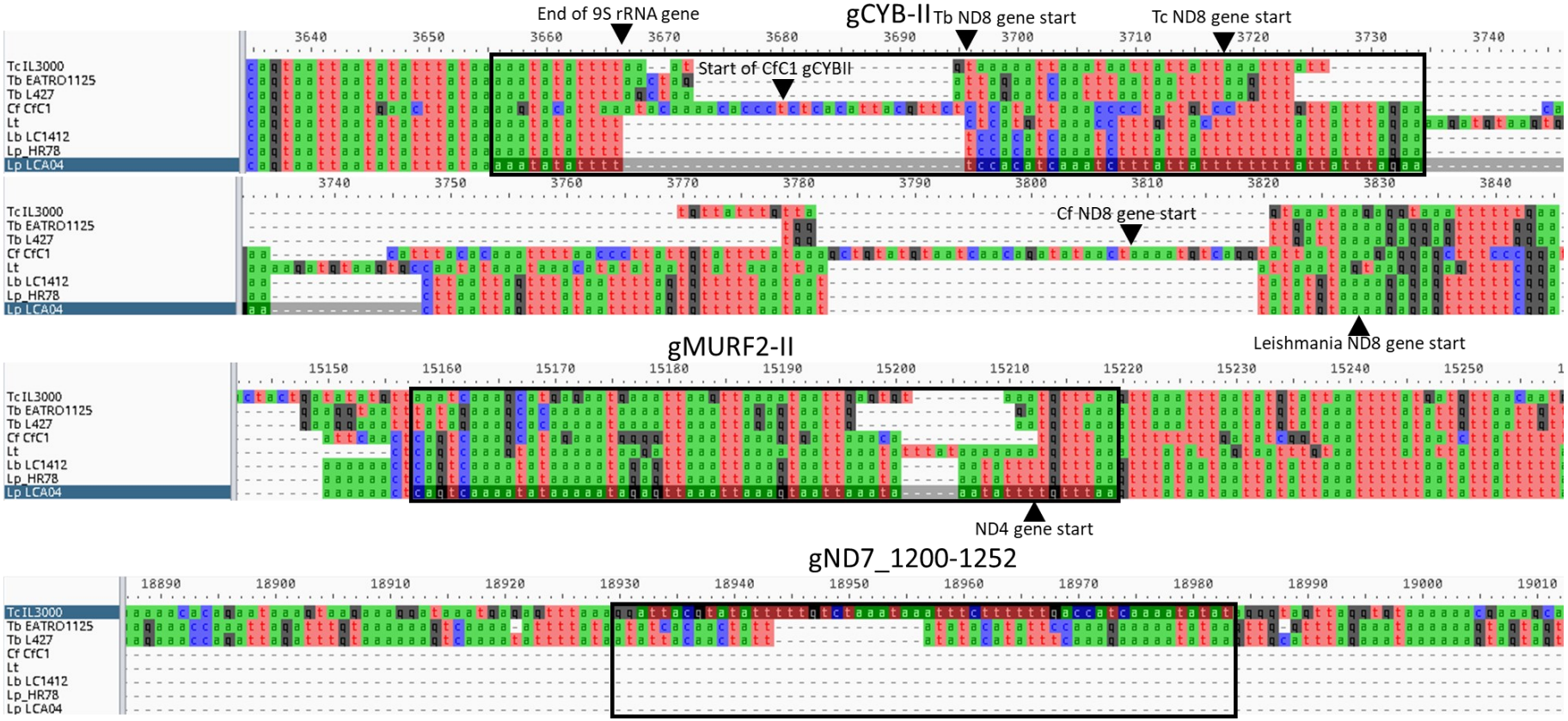


Figure 4-6**. Alignment of trypanosomatid maxicircles over the region encoding gGR3_24-70, gND9_304-355, gGR4_251-304, gCOX2, gGR3_66-122, gGR4_81-122/ND8_1-42, gGR4_248-289, gGR3_97-139, CYB-II, gMURF2-II, and gND7_1200-1252 genes in *L. braziliensis* and *L. peruviana*.**

The abbreviations are the same as in Figure 4-5.The regions that correspond to the gRNA genes are circled. The alignment, produced with Aliview [348], reveals the conservation of gene positions and sequences in the eight maxicircles examined. The gRNA genes in *L. braziliensis* LCA04 is highlighted except for gND7_1200-1250, where IL3000 is highlighted. Gene boundaries are marked with triangles.

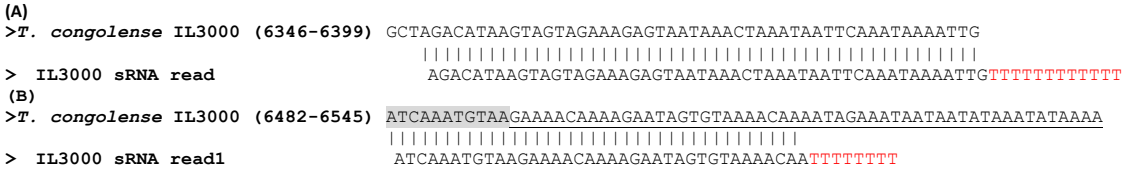


Figure 4-7. **Alignment of sRNA with untemplated U-tail (red) to the region on IL3000 maxicircle aligned to (A) gND9_304-355 and (B) gGR4_251-304 in *L. braziliensis* and *L. peruviana*.**

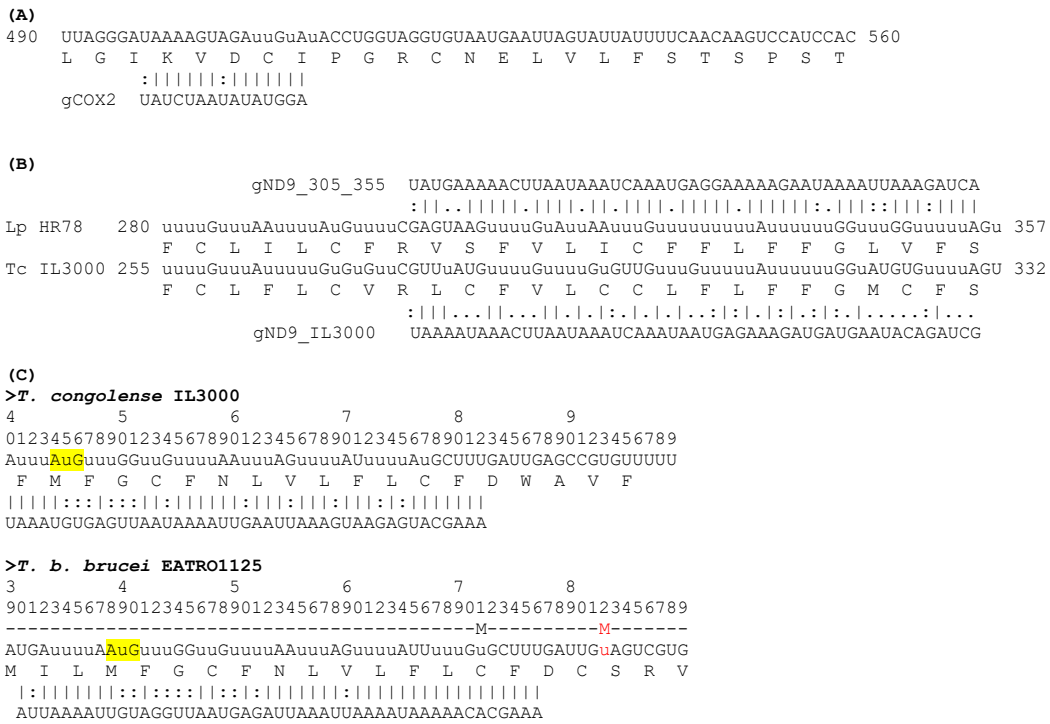


Figure 4-8. **Alignments of COX2 (A), ND9 (B), and MURF2 (C) gRNAs over edited regions in *T. congolense* IL3000.**

COX2 and MURF2 gRNAs cover all editing sites on the corresponding mRNAs in IL3000. (B) compares the alignment of gND9_IL3000 with the maxicircle-encoded gND9_304-355 in L. braziliensis and L. peruviana. The edited ND9 mRNAs from IL3000 and HR78 are not aligned due to highly divergent sequences. The potential target region of gND9_IL3000 is identified with ORF and protein sequence alignment. There is no evidence for gND9_IL3000 alignment. (C) No gRNA has been identified in T. brucei to explain the additional uridine insertion at nt 82 of the T. brucei MURF2 mRNA. The start codons are highlighted.

4.4 earMinicircle assembly and annotation of the three T. congolense isolates

4.4.1 Completeness of minicircle assembly

To characterize the kDNA complexity of *T. congolense,* we conducted minicircle assembly individually for each isolate using MEGAHIT via KOMICS [66] and recovered 180, 248, and 339 circularized minicircle contigs from IL3000, Kapeya, and UPKZN, respectively.

For each isolate, we mapped the kDNA reads back to the corresponding minicircles and maxicircle assemblies to assess the completeness of kDNA assembly by the percentage of mapped reads containing conserved sequence block CSB-3 (PMC). As expected, most of the kDNA reads (86.6% - 95.3%, depending on isolate) could be mapped (Table *4-14*). All three assemblies exhibited over 99% PMC. Full read coverage on all minicircle contigs was confirmed by mapping. We concluded that the assembly captured nearly complete kDNA networks of the three isolates. The assembly may not include minicircle classes present at extremely low frequencies in the cell population due to kDNA network heterogeneity.

Pooling minicircles from all three isolates yielded 710 distinct minicircle classes with less than 95% sequence identity shared with any other classes in the assembly, including the ~120 nt conserved region (see below). The three isolates displayed similar minicircle length distributions (Figure *4-9*). The lengths of most circularized contigs ranged in size from 890 to 1010, with a peak between 950 and 960 bp shared by three isolates. UPKZN contained the shortest minicircle class of 851 bp, and Kapeya contained the longest minicircle class of 1052 bp. All isolates had mean and mode minicircle lengths between 952 nt and 954 nt.

Table 4-14. **Completeness of minicircle assembly for three *T. congolense* isolates**

Isolate	Reads	Mapped reads	CSB-3 containing reads	Number of minicircles	Mapped CSB-3 containing reads
IL3000	6917972	6149419 (88.9%)	1489623	180	1481725 (99.5%)
Kapeya	2958684	2561293 (86.6%)	889101	248	884155 (99.4%)
UPKZN	7507852	7158050 (95.3%)	2601423	339	2592197 (99.7%)

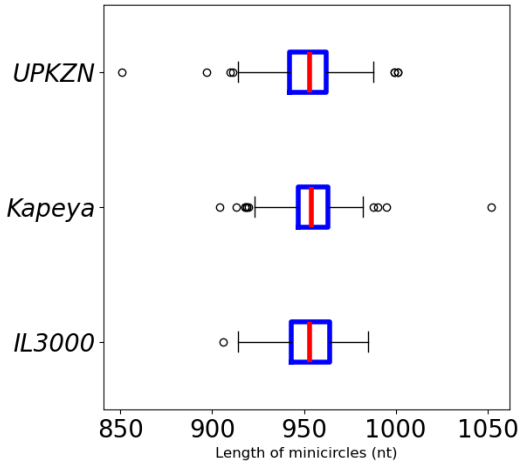


Figure 4-9. **Minicircle length distributions of *T. congolense* isolates.**

IL3000: mean: 952, sd (standard deviation): 14, median: 953.0, mode: 964, max: 985, min: 906; Kapeya: mean: 954, sd: 16, median: 954.0, mode: 949, max: 1052, min: 904; UPKZN: mean: 952, sd: 17, median: 953, mode: 950, max: 1001, min: 851. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point

4.4.2 CSB features

Similar to *T. brucei*, all *T. congolense* minicircle classes contained a single copy of the ~120 nt conserved region with CSB-1, CSB-2, and CSB-3 [225]. We derived the consensus for the conserved regions for each isolate from minicircle alignments using the Weblogo motif generator (Figure *4-10*) [324]. The conserved region included three less conserved areas at nt 18-29, nt 52-56, and nt 63-67. Using regular expression search, CSB-1 and 3 were identified in all minicircles of three isolates (Table 4-15), while CSB-2 was not detected in 24 - 31% of minicircles.

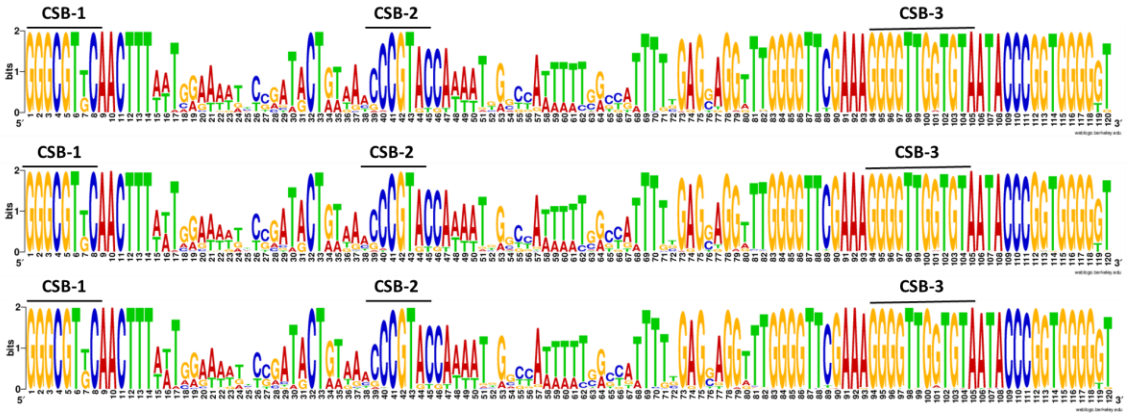


Figure 4-10. **Conserved minicircle regions of *T. congolense* isolates.**

Created with Weblogo [324]. Top to bottom: IL3000, Kapeya, UPKZN. The conserved region starting with CSB-1 spans around 120 nt. CSB-3 is located 93 nt downstream of CSB-1. CSB-1 and CSB-3 are nearly 100% conserved, while the most frequent CSB-2 variant is detected in ~50% of sequences. The consensus are highly conserved among the three isolates.

Unlike *T. brucei*, where CSB-1 appears to be 100% conserved [225, 229, 341], three *T. congolense* isolates also had major alternative CSB-1s with a G-T substitution at the 7th nucleotide (Table 4-15). While IL3000 and Kapeya had similar ratios of canonical versus alternative CSB-1 (z test, p= 0.82), UPKZN contained a significantly higher proportion of minicircles with the alternative CSB-1 than the other two isolates (z test, IL3000 p= 0.004, Kapeya p < 0.001).

Table 4-15. **Alternative CSB motifs were detected in all three isolates.**

CSB-1	**GGGCGTTCA**		**GGGCGTGCA**		**GGGCGT[AGT]CA**	
IL3000	113	(62.78%)	64	(35.56%)	3	(1.67%)
Kapeya	153	(61.69%)	88	(35.48%)	7	(2.82%)
UPKZN	180	(53.1%)	154	(45.43%)	5	(1.47%)

CSB-2	**ACCCGTAC**	**TCCCGTAC**	**ACACGTAC**	**TCCCGTGT**	**ACCCGTGC**	**TCCCGTGC**	**AACCGTAC**	**Undetected**
IL3000	95 (51.63%)	31 (16.85%)	5 (2.72%)	3 (1.63%)	1 (0.54%)	1 (0.54%)	0	44 (24.44%)
Kapeya	125 (50.4%)	30 (12.1%)	9 (3.63%)	3 (1.21%)	3 (1.21%)	0	1 (0.4%)	77 (31.05%)
UPKZN	189 (55.75%)	35 (10.32%)	6 (1.77%)	2 (0.59%)	2 (0.59%)	2 (0.59%)	0	103 (30.38%)

CSB-3	**GGGGTTGGTGT**	**GGGGTTGATGT**	**GGGGTTG[AG]TGT**
IL3000	174 (96.67%)	5 (2.78%)	1(0.56%)
Kapeya	244 (98.39%)	3 (1.2%)	1(0.4%)
UPKZN	321 (94.69%)	14 (4.13%)	4(1.18%)

Note: Nucleotides at heterogeneous sites are put in square brackets.

The three isolates all had minicircle classes that exhibited sequence heterogeneity at the substituted 7th nucleotide of CSB-1 (Table 4-15). The minor alternative CSB-1 GGGCGTACA was only observed in IL3000 and Kapeya. The detection of alternative CSB-1 and minicircle classes with heterogeneous CSB-1 suggested that the restriction on CSB-1 motif conservation was more relaxed in *T. congolense* than in *T. brucei*.

Apart from the three highly conserved bases 'CGT', CSB-2 exhibited more variability than CSB-1 and CSB-3 (Table 4-15). The three isolates shared the same most common CSB-2 motifs ACCCGTAC found in over 50% of minicircle classes. Meanwhile, over 10% of minicircles from each isolate contained the canonical CBS-2 of *T. brucei*, TCCCGTGC.

Although previously considered universally conserved, an alternative CSB-3 (GGGGTTGATGTA) has been reported for *T. b. brucei* EATRO1125 [225, 341] and in sub-Saharan *T. brucei* isolates in this project (3.3.4). The proportion of minicircle classes that homogeneously contained this alternative CSB-3 varied among *T. congolense*, ranging from 14 (4.13%) in UPKZN to 3 (1.67%) in Kapeya (Table 4-15). The three isolates also contained minicircle classes that had a mixed population of the canonical and the alternative CSB-3. Minicircle classes with heterogeneous CSB-1 and CSB-3 motifs probably represented an intermediate stage when a novel CSB had arisen in the population due to point mutation and fluctuated in abundance given the imperfect replication and random segregation of the kDNA network before being lost or fixed.

4.4.3 Minicircle copy number (MCN) and network size estimation

Mapping all reads back to the assemblies confirmed complete coverage over all minicircles, with an average depth ranging from 14 to 7935. Due to the lack of nuclear reads, we estimated MCNs per network with the ratio of minicircle read depth versus average read depth over the maxicircle coding region, assuming 30 maxicircles per network based on the estimated range of 20-50 maxicircles per network in *T. b. brucei* [195, 340]. Admittedly, the estimation could result in substantial over or underestimation of the minicircle copy number if the maxicircle copy number deviated from the presumption 30. For instance, if the actual maxicircle copy number was 15, assuming 30 maxicircles would result in doubling the estimated minicircle copy number and double the network size.

MCNs varied substantially within each isolate (Figure *4-11*). The standard deviation for UPKZN (38.28) was much higher than for the other two isolates (IL3000: 20.54, Kapeya: 17.60), although the higher variance was not statistically significant (Levene's test, p=0.4). The most abundant minicircle classes ranged from 61 copies in IL3000 and Kapeya to 128 copies in UPKZN. We also noticed minicircle classes with less than one copy per network in (1, 8, and 2 for IL3000, Kapeya, and UPKZN, respectively), suggesting that these minicircles were not present in all cells within the sampled population (if the assumption of 30 maxicircles per network is accurate). Hence, not all cells contained the entire set of minicircles from the assembly. The same observation on minicircle abundance has been made in *Leishmania* [295] and *T. brucei* [225].

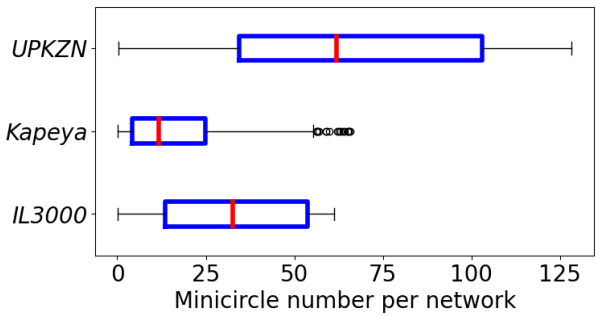


Figure 4-11. **Minicircle copy number distributions of *T. congolense* isolates.**

IL3000: mean: 32.00, median: 32.69, max: 61.23, min: 0.21. Kapeya: mean: 18.00, median: 11.63, max: 65.71, min: 0.21. UPKZN: mean: 67.00, median: 61.91, max: 128.25, min: 0.24. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point

The kDNA network sizes were estimated as the sum of MCNs plus 30 copies of maxicircles. IL3000 and Kapeya had 5859 and 4460 circles per network, respectively. Surprisingly, we estimated 22801 circles for the UPKZN network, over three times the network sizes for the other two isolates. The latter estimate did not fall within the range of earlier estimaties of 5k to 10k minicircles per network for *T. brucei* based on reassociation kinetics and restriction mapping [195, 340]. Estimated with WGS data, the MCNs of 38 *T. brucei* subspecies range between 252 and 4828 copies with a mean of 2099 in a recent study [80].

In Chapter 3, the MCNs of the sub-Saharan *T. brucei* subspecies fell between 603 and 16054 copies per network with averages between 4k and 5k. A network with nearly 23k minicircles was much larger than any of the estimations based on *T. brucei*. A most likely scenario is that UPKZN has fewer than 30 maxicircles per network and the assumption substantially scales up the MCN estimation.

4.4.4 Minicircle cassette and gRNA gene detection for *T. congolense* IL3000, Kapeya and UPKZN

Our analysis of the gRNA repertoires of *T. brucei* isolates revealed extensive conservation of editing blocks and functionally homologous gRNA genes across subspecies (see Chapter 3). We were interested in whether the *T. congolense* isolates also exhibited conservation of editing blocks, amongst themselves and perhaps even with the *T. brucei* subspecies. To address this question, we carried out the first detailed annotation of *T. congolense* kDNA with a bespoke pipeline and followed the definition for gRNAs from a previous study [225] as follows. We aligned minicircle sequences to the predicted, strain-specific edited mRNAs allowing G-U wobble base pairing to identify high-confidence canonical gRNAs encoded within cassettes. The motifs of forward and reverse repeats and the initiation sequence enabled subsequent characterization of cassettes on each minicircle and the detection of non-canonical gRNAs by nucleotide bias [225]. We defined canonical gRNA genes as gRNAs aligned to the edited mRNA without gaps for at least 25 nt with a maximum of three mismatches. As discussed previously [225], these criteria by necessity were somewhat arbitrary and may lead to a few false-positive and false-negative classifications of gRNAs with parameters close to the cutoffs. However, since the peaks of gRNA and anchor lengths lay way upstream of the cutoffs, the arbitrary cutoffs were unlikely to have significantly affected our conclusions described below.

We also modified the IL3000 edited mRNAs based on the SNPs we called in 4.1.2 to predict the edited mRNAs for Kapeya and UPKZN. The process was following the same principle we applied to the preliminary edited mRNA prediction for *T. b. gambiense* type 1 from EATRO1125 described in 3.2.2. As the SNPs did not include any insertions or deletions, we only replaced the nucleotide in IL3000 with the alternative nucleotide in the corresponding isolates. Without transcriptome data available for the other two isolates, we did not polish the isolate-specific predicted mRNAs any further.

We detected 556 gRNA genes in total in IL3000. Excluding gRNAs aligned to the non-variable regions on alternatively edited mRNAs gave 465 unique gRNA genes (Table 4-16). We detected 660 unique gRNAs in Kapeya (Table 4-17) and 860 unique gRNAs in UPKZN (Table 4-18).

Table 4-16. **IL3000 gRNA coverage on edited mRNAs.**

	total gRNAs	unique gRNAs	initiation gRNAs	missing gRNAs	all insertions	covered insertions	all deletions	covered deletions	% coverage
A6_v1	63	63	1	1	458	450	30	30	98.4
A6_v2	63	1	1	1	459	451	30	30	98.4
COX2	1	1	1	0	4	4	0	0	100
COX3	96	96	1	0	559	559	35	35	100
CR3	24	24	5	0	146	146	7	7	100
CR4	35	35	1	1	315	311	32	32	98.8
CYB	4	4	2	0	36	36	0	0	100
MURF2	1	1	1	0	25	25	1	1	100
ND3_v1	30	30	1	0	196	196	29	29	100
ND3_v2	30	1	1	0	197	197	27	27	100
ND7	106	106	1	0	558	558	75	75	100
ND8	35	35	1	2	251	227	37	37	91.7
ND9	47	47	2	0	322	322	26	26	100
RPS12	21	21	2	0	140	140	34	34	100
total	556	465	21	5	3666	3622	363	363	98.9

Table 4-17. **Kapeya gRNA coverage on mRNAs of maxicircle-encoded cryptogenes. IL3000 mRNAs predicted from transcriptome data were used.**

	total gRNAs	unique gRNAs	initiation gRNAs	missing gRNAs	all insertions	covered insertions	all deletions	covered deletions	% coverage
A6_v1	79	79	2	0	458	458	30	30	100
A6_v2	79	1	2	0	459	459	30	30	100
COX2	1	1	1	0	4	4	0	0	100
COX3	162	162	0	2	559	544	35	35	97.5
CR3	15	15	1	1	146	136	7	7	93.5
CR4	48	48	2	2	315	300	32	24	93.4
CYB	4	4	2	0	36	36	0	0	100
MURF2	1	1	1	0	25	25	1	1	100
ND3_v1	63	63	1	0	196	196	28	28	100
ND3_v2	63	1	0	1	197	192	26	26	97.8
ND7	143	143	1	1	558	555	75	75	99.5
ND8	53	53	2	2	251	225	37	37	91
ND9	61	61	1	0	322	322	26	26	100
RPS12	28	28	1	2	140	134	34	34	96.6
total	800	660	17	11	3666	3586	361	353	97.8

Table 4-18. **UPKZN gRNA coverage on mRNAs of maxicircle-encoded cryptogenes. IL3000 mRNAs predicted from transcriptome data were used.**

	total gRNAs	unique gRNAs	initiation gRNAs	missing gRNAs	all insertions	covered insertions	all deletions	covered deletions	% coverage
A6_v1	110	110	1	1	458	448	30	30	98
A6_v2	110	0	1	1	459	448	30	30	97.8
COX2	1	1	1	0	4	4	0	0	100
COX3	241	241	0	2	559	546	35	35	97.8
CR3	35	35	7	1	146	139	7	7	95.4
CR4	63	63	0	2	315	294	32	32	93.9
CYB	5	5	3	1	36	30	0	0	83.3
MURF2	1	1	1	0	25	25	1	1	100
ND3_v1	50	50	1	0	196	196	29	29	100
ND3_v2	50	1	1	0	197	197	27	27	100
ND7	178	178	1	1	558	555	75	75	99.5
ND8	63	63	2	3	251	222	37	37	89.9
ND9	71	71	1	0	322	322	26	26	100
RPS12	41	41	2	1	140	135	34	34	97.1
total	1007	860	22	13	3662	3532	363	362	96.7

We observed that gRNA genes were typically arranged as cassettes flanked by imperfectly conserved inverted repeats, as reported before for *T. brucei* and *T. congolense* [227, 228, 230, 341, 342]. Unlike *T. brucei*, where five different cassette positions are observed, including a minor cassette III, with up to four cassettes present on a minicircle [225], the overall cassette structure was less variable for *T. congolense* minicircles, with three cassette positions shared by the vast majority of minicircles (Figure *4-12*). 94.4% (170/180) IL3000 minicircles contained three cassettes, while nine had two cassettes and one had one cassette (Table *4-19*). 89.5% (222/248) Kapeya minicircles contained three cassettes, while 26 had two cassettes. 88.8% (301/339) UPKZN minicircles contained three cassettes, while 35 had two cassettes and three had one cassette. A curious trend observed in three isolates was the shift of cassette III towards the 5’ end as the minicircle lengths increased (Figure 15).

Table 4-19**. Counts of minicircles with one, two, or three cassettes in IL3000, Kapeya, and UPKZN**

Strain	Three cassettes	Two cassettes	One cassette
IL3000	170	9	1
Kapeya	222	26	0
UPKZN	301	35	3

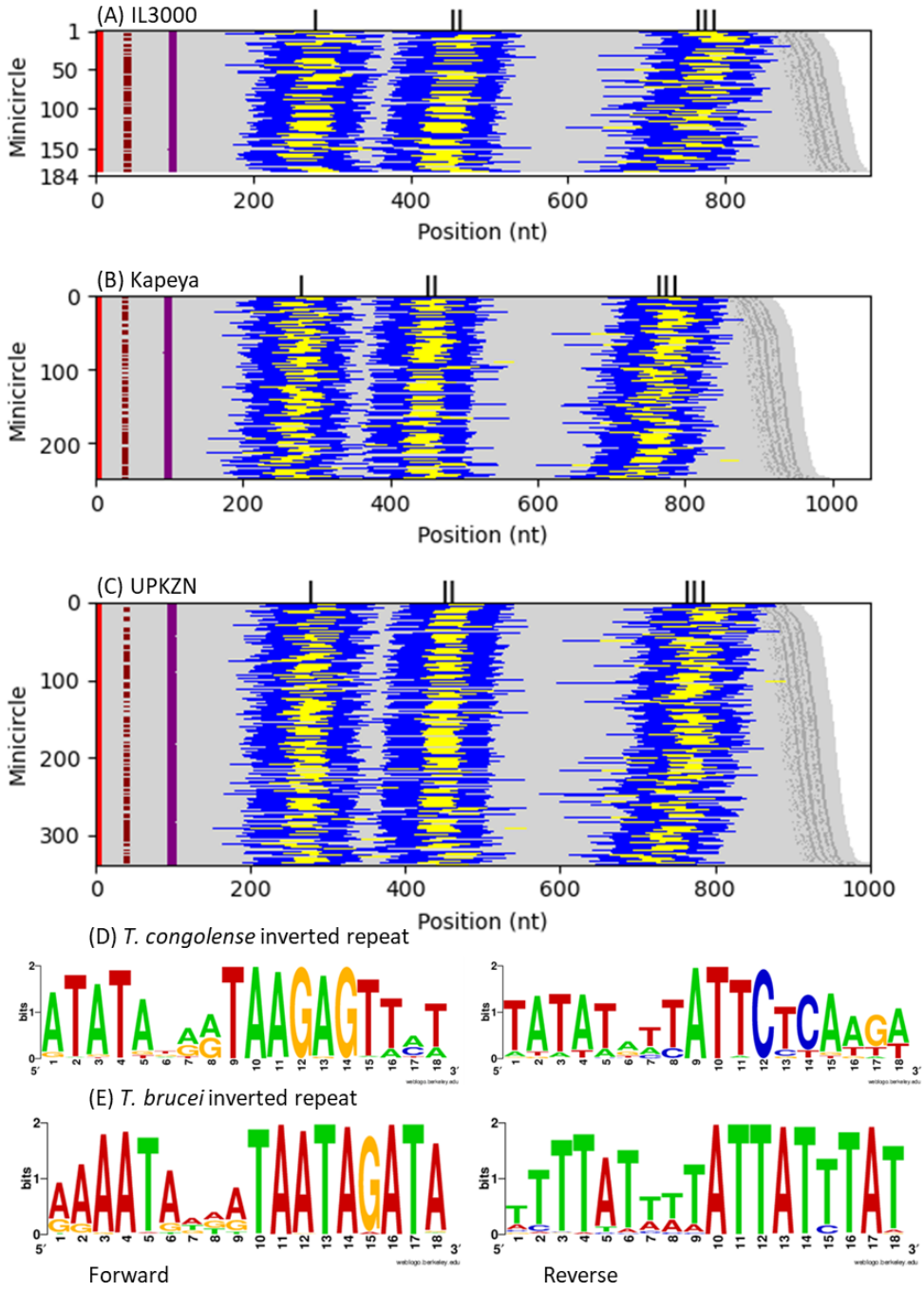


Figure 4-12. **Conserved features of *T. congolense* minicircles.**

(A)-(C) *T. congolense* minicircles show highly conserved structures with three cassettes. Red, brown, and purple represent conserved sequence blocks CSB1, CSB-2, and CSB-3, respectively. The regions between the 5' of the forward 18 bp inverted repeats to the 3' of the inverted repeats are shown as dark blue. Cassette-associated and orphan canonical gRNA genes are shown in yellow. The labels for the four gRNA cassette positions, I–IV, are located at each cassette's median center position. Dark gray bars shows A-tracts of the bend region. **(D)** The repeat sequences were extracted from 180 aligned IL3000 minicircles. The consensus motifs were distinct from **(E)** the published sequences for *T. b. brucei* strain EATRO1125 [225].

The motifs of forward and reverse repeats were derived from 529 IL3000 cassettes, and the reverse repeat was inverted for better visualization of the repeat symmetry (Figure *4-12* D). Similar to *T. brucei,* the most conserved bases in forward and reverse repeats mirrored each other almost perfectly in *T. congolense.* Despite being A-T rich and forming similar cassette

structures, the *T. congolense* forward and reverse repeats are distinct from the *T. brucei* counterparts [228] (Figure *4-12* E). The conservation of cassette structure within minicircle populations encoding distinct gRNAs and among African trypanosomatids from divergent lineages indicated the functional relevance of the structure. The inverted repeats may play roles in minicircle recombination or gRNA transcription and maturation [227, 228].

4.4.5 Completeness of gRNA coverage for *T. congolense* isolates

In IL3000, the set of predicted gRNAs covered 98.9% of all editing sites and allowed over 98% editing site coverage for all individual mRNAs except ND8 (Table *4-16*). We determined the average length of aligned gRNA sequences (the anchor and the coding region) to be 43 nt (Figure 4-13). Based on this average length, we predicted that at least four (the same gRNA could direct both versions of A6) additional gRNAs would be necessary to bridge the gaps in gRNA coverage on A6, CR4, and ND8 (Figure 4-14). Despite the pronounced gRNA redundancy observed for most editing sites (see below), and similar to what was observed for *T. brucei* [225], ten mRNAs had a single initiation gRNA, and ND9, CYB, and RPS12 had only two initiation gRNAs. The exception was CR3, with five initiation gRNAs. CR3 also has the highest initiation gRNA count in *T. brucei* [225].

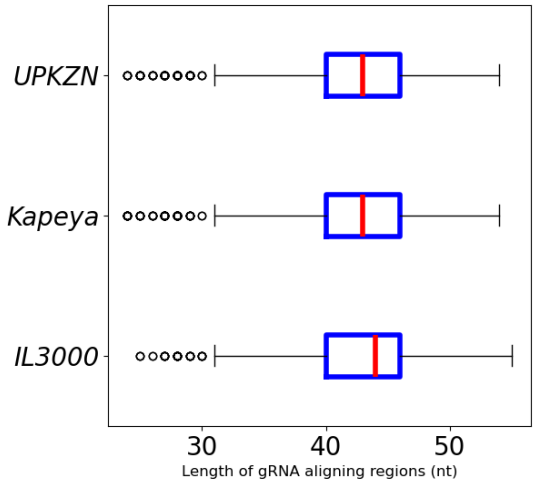


Figure 4-13. **Length distribution of the complementary regions (the anchors plus the guiding regions of gRNAs) in IL3000, Kapeya, and UPKZN.**

IL3000: mean: 43, sd (standard deivation): 6, median: 44.0, mode: 44, max: 55, min: 25; Kapeya: mean: 42, sd: 6, median: 43, mode: 45, max: 54, min: 24; UPKZN: mean: 42, sd: 6, median: 43, mode: 44, max: 54, min: 24. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point.

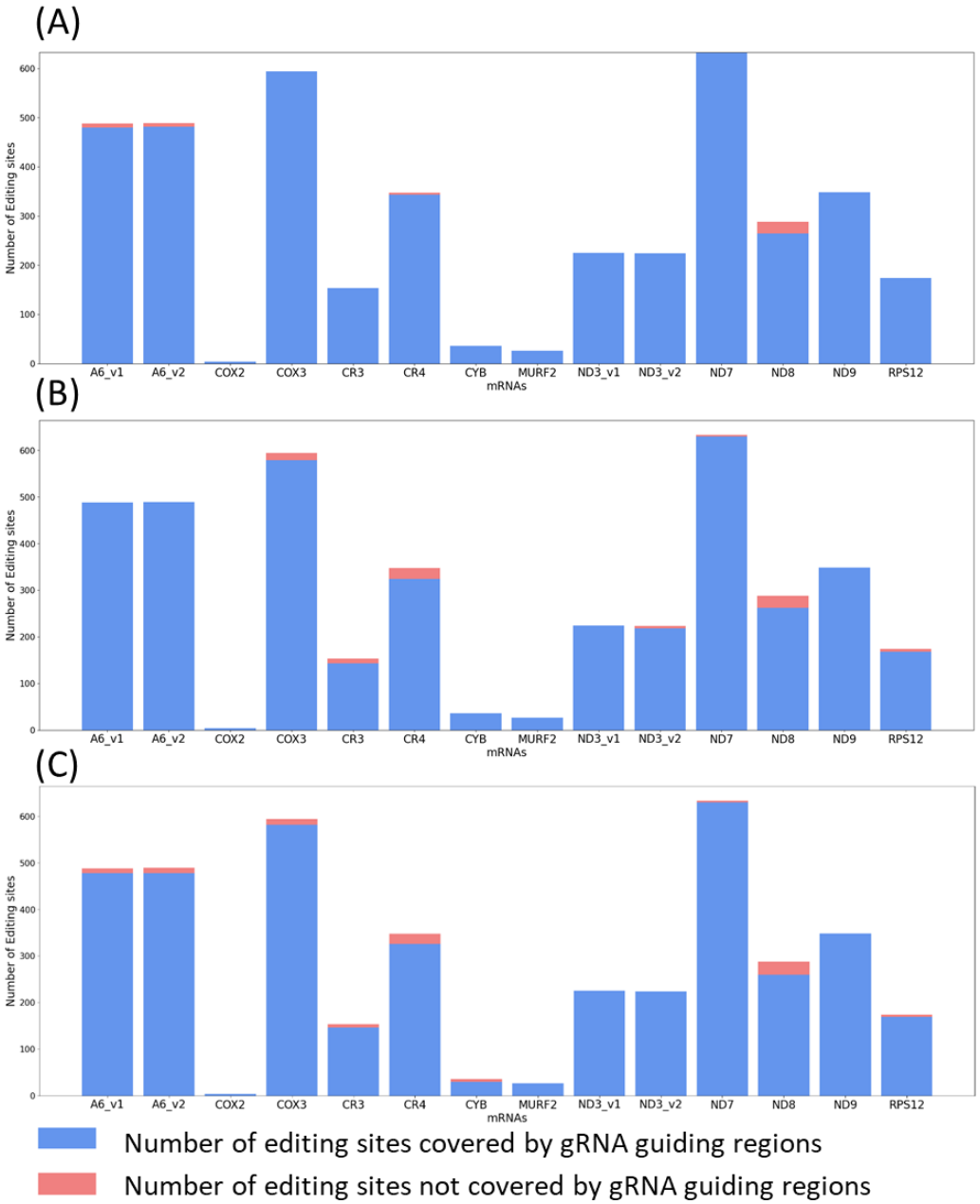


Figure 4-14. **gRNA coverage for IL3000 (A), Kapeya (B), and UPKZN (C) mRNAs.**

(A) The annotation observes 100% gRNA coverage on all mRNAs except A6, CR4, and ND8 mRNAs. A6 and CR4 have > 98% gRNA coverage. **(B)** The annotation observes 100% gRNA coverage for A6 (both versions), COX2, CYB, MURF2, ND3_v1, and ND9 mRNAs. All mRNAs have > 93% gRNA coverage. **(C)** The annotation observes 100% gRNA coverage for COX2, MURF2, both versions of ND3, and ND9 mRNAs. CYB and ND8 have the lowest gRNA coverages with 83.3% and 89.9%, respectively, while other mRNAs have gRNA coverage > 93%.

The overall gRNA coverages were slightly lower but nearly complete (>96%) in Kapeya and UPKZN (Table 4-17,Table 4-18). The lower gRNA coverages could suggest that predictions of edited sequences for Kapeya and UPKZN were slightly less robust. Given the average gRNA length 42 nt, we estimated 11 and 13 missing gRNAs in Kapeya and UPKZN, respectively. All mRNAs had one or two initiation gRNAs in Kapeya, whereas in UPKZN CR3 had seven initiation gRNAs. Note that although Kapeya and UPKZN had more minicircles and hence

more unique gRNAs, the gRNA coverage were not necessarily higher than IL3000. The additional gRNAs were mostly redundant.

In Kapeya, both versions of A6, COX2, CYB, MURF2, ND3_v1, and ND9 mRNAs had complete gRNA coverage. CR3 and CR4 had lower gRNA coverage around 93%, while other mRNAs had coverage > 96%. In UPKZN, COX2, MURF2, both versions of ND3, and ND9 mRNAs. CYB and ND8 had complete gRNA coverage. CYB, CR4, and ND8 had coverage between 83.3% and 93.9%, while other mRNAs have coverage over 95%. We concluded that we recovered a nearly complete set of gRNAs for all three isolates which allowed us to further compare the difference in their editing capacity.

4.5 Close examination of T. congolense I IL3000 gRNA alignments

In this section, we give a detailed account for the gRNA alignments on the edited mRNAs. As the mRNAs were mainly predicted with IL3000 transcriptomes and only BSF IL3000 small RNA (sRNA) was available to infer the expression of annotated gRNAs, we focused on IL3000 gRNA for the general features. After mapping sRNAs to minicircles, gRNA genes that exceeded the subjectively assigned threshold were considered expressed (0.025% of total read depth) [225]. Without biological replicates, we only used the sRNAs as evidence for gRNA expression when a transcript was detected but not vice versa. Nevertheless, the data revealed trends that could be compared with previous studies.

The concept of the gRNA family allowed us to compare the alignments of gRNAs on mRNAs between *T. congolense* isolates, and we will also mention the other two isolates when interesting comparisons could be made. All gRNA-mRNA alignment files are available on Figshare (https://doi.org/10.6084/m9.figshare.27020074)

4.5.1 Respiratory complex I

As other trypanosomatid parasites, *T. congolense* encodes eight (ND1, ND2, ND3, ND4, ND5, ND7, ND8, ND9) out of 14 core complex I subunits on kDNA. Four subunits (ND3, ND7, ND8, ND9) require post-transcriptional editing of their mRNAs, and incomplete sequence information has previously been published [204]. Meanwhile, mRNAs for two ORFs with unidentified functions (CR3, CR4) require extensive editing and have been proposed to also encode subunits of complex I [205]. They will be discussed in a separate section.

ND3 had complete gRNA coverage in IL3000 and UPKZN. Minicircle annotation using both versions of ND3 mRNA resulted in two initiation gRNAs encoded on different minicircles (Figure *4-15*). The ND3_v2 initiation gRNA was not detected in Kapeya. IL3000 BSF sRNA data supported the expression of the initiation gRNA for ND3_v2, while the sRNA coverage over ND3_v1 initiation gRNA did not exceed the subjectively assigned threshold for being expressed (0.025% of total read depth).

Figure 4-15. **The gRNA coverage of IL3000 at the 3' editing sites of ND3**.

The alternatively edited ND3_v1 and ND3_v2 require different initiation gRNAs. The numbers in the first three rows show the coordinates of the bases. Insertions not covered by gRNAs are showed as 'M' on the fourth row (no mismatch in this case), The numbers on the fifth row indicate the number of uridine deletions that occur. The inserted uridines were in lowercase in the mRNA sequence. The protein sequence represents the longest ORF. For gRNA alignment, '|' indicates Watson-Crick base pairing, ':' indicates G-U base pairing, and '.' Indicates mismatch.

The ND7 mRNA has two independent editing domains [206]. IL3000 minicircle annotation allowed complete gRNA coverage of both domains with 106 gRNAs. The three 5' most uridine insertions from nt 247 to nt 249 were not covered in Kapeya and UPKZN. Unlike other mRNAs, the only initiation gRNA for the 3' editing domain was encoded on the maxicircle from 15086 to 15138 nt on the template strand. On the other hand, at least three different IL3000 gRNAs can initiate editing of the 5' editing domain and be expressed based on sRNA mapping. While mO_082(II)_gND7(145-187) located on cassette II, mO_111(I)_gND7(144-188) and mO_120(I)_gND7(147-188) both were encoded on cassette I, yet the gRNAs encoded on other cassettes were different between mO_111 and mO_120. Hence, it suggested that the gRNAs contributing to this editing redundancy probably have arisen from distinct origins.

The amount of total edited ND8 might differ through life stages [289]. No PacBio evidence supported the editing of the 5' editing sites before nt 80, including the region covered by the orphan gRNA immediately downstream of the missing gRNA (Figure 4-15). The orphan gRNA was considered expressed based on sRNA mapping (see Figshare). It was possible that the fully edited ND8 mRNA was too scarce to be detectable for RNA-seq.

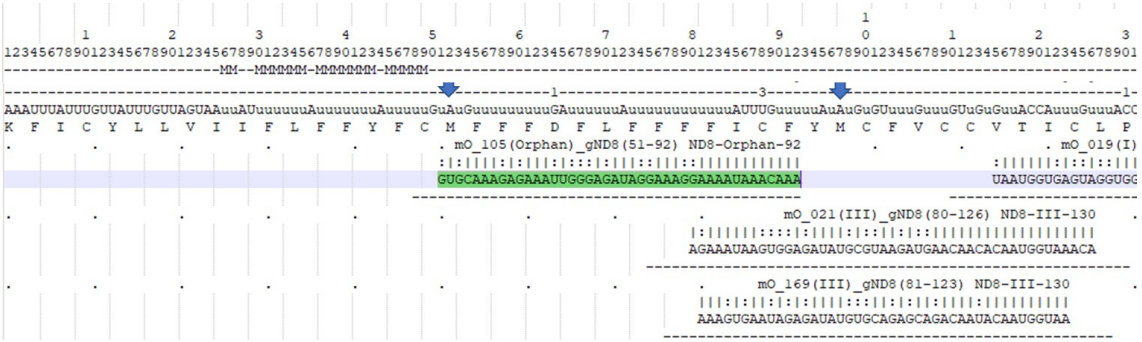


Figure 4-16. **The gRNA coverage of IL3000 over the 5' most editing sites on ND8.**

Probable start codons are marked by arrows. The uncovered editing sites are with the 5' UTR. The numbers in the first three rows show the coordinates of the bases. Insertions not covered by gRNAs are showed as 'M' on the fourth row (no mismatch in this case), The numbers on the fifth row indicate the number of uridine deletions that occur. The inserted uridines were in lowercase in the mRNA sequence. The protein sequence represents the longest ORF. For gRNA alignment, '|' indicates Watson-Crick base pairing, ':' indicates G-U base pairing, and '.' Indicates mismatch.

We observed complete coverage over ND9 in three isolates. One ND9 gRNA was encoded on the template strand in cassette I and non-redundantly covered editing sites from to 121st to 162nd nt. We also identified two slightly varied IL3000 initiation gRNAs encoded on cassette II of mO_11 and mO_47, respectively. Sharing 82.31% identity, both minicircles encoded no other canonical gRNAs, but an expressed non-canonical gRNA was detected on cassette I of both minicircles. The similarities suggested that the two minicircles descended from the same origin. However, the ND9 initiation gRNA on mO_11 was considered expressed, while the one on mO_47 was not.

4.5.2 Respiratory complex III / cytochrome bc1 complex

In *T. brucei* and *T. congolense*, the CYB mRNA is minimally edited on the 5' end in two editing blocks by orphan gRNAs located outside cassettes [280, 293, 341]. We confirmed that IL3000 covered each editing block with two orphan gRNAs (see Figshare). All IL3000 orphan gRNA genes were considered expressed besides mO_116(Orphan)_gCYB(31-57). Kapeya also had 100% gRNA coverage over CYB editing sites. In UPKZN, the 5' gRNAs could not cover the first six editing sites but were capable of generating the start codon (Figure *4-17*). We detected transcripts without the first six uridine insertions in IL3000 PacBio reads (Figure *4-2* B).

Figure 4-17. **The gRNA coverage of UPKZN over CYB.**

CYB is minimally edited on the 5' end. The shortened gRNAs can not edit six uridine insertions in the 5' UTR but still generate the start codon for mRNA translation. The numbers in the first three rows show the coordinates of the bases. Insertions not covered by gRNAs are showed as 'M' on the fourth row (no mismatch in this case), The numbers on the fifth row indicate the number of uridine deletions that occur. The inserted uridines were in lowercase in the mRNA sequence. The protein sequence represents the longest ORF. For gRNA alignment, '|' indicates Watson-Crick base pairing, ':' indicates G-U base pairing, and '.' Indicates mismatch.

4.5.3 Respiratory complex IV / cytochrome *c* oxidase

In *T. brucei* and *T. congolense*, three subunits of complex IV are kDNA encoded, mRNAs for two of which undergo post-transcriptional editing. COX2 is minimally edited via cis-editing directed by its own 3' UTR and discussed in section 4.1 We achieved complete gRNA coverage over the extensively edited COX3 with 96 gRNAs, although both Kapeya and UPKZN had uncovered editing sites. The gRNA coverage displayed substantial redundancy except for the 3' most editing sites covered by a single initiation gRNA in IL3000. The initiation gRNAs from Kapeya and UPKZN were missing.

4.5.4 Respiratory complex V / F_1F_O-ATP synthase

Trypanosoma kDNA encodes subunit 6 of complex V (A6). The edited mRNA was covered by gRNA on all except 8 uridine insertions from nt 667 to nt 680 nt in IL3000. UPKZN missed the gRNA for editing sites from nt 719 to nt 736, while Kapeya had complete A6 coverage. PacBio read mapping revealed alternative editing in the editing block downstream of the same initiation gRNA (Figure *4-1*).

We detected gRNAs for different versions of A6 encoded on minicircles with distinct cassettes and substantial sequence divergence (<65% id), indicating independent origins or prolonged divergence. IL3000 sRNA mapping provided evidence for A6_v1 gRNA expression, but we did not detect the sRNA for A6_v2 gRNA in BSF IL3000. If the gRNA for A6_v2 is expressed in EMF, the differentiated gRNA expression will agree with the low A6_v2 proportion (15.41%, 98/636) in BSF compared to 63.83% (60/94) in EMF in PacBio transcriptome data.

4.5.5 Mitoribosome

The kDNA also encodes two mitoribosomal protein subunits, uS3m and uS12m (RPS12); the mRNA for the latter requires extensive editing. Our annotation identified 21 gRNAs that completely covered all RPS12 editing sites in IL3000. However, in Kapeya and UPKZN, shorter gRNAs resulted in the lack of coverage over the five 5'-most insertions in the 5' UTR, which did not affect the downstream start codon and presumably would not interrupt mRNA translation. Kapeya also had an additional gap in gRNA coverage at nt 237. In IL3000, the two initiation gRNAs were located on cassettes II of mO_033 and mO_140, whose sequences diverged. Both gRNAs were considered expressed in BSF and located on the template strand. We also detected the RPS12 initiation gRNAs on the template strain in Kapeya and UPKZN.

4.5.6 Unidentified open reading frames

Similar to *T. brucei*, the *T. congolense* maxicircle also encodes three transcribed ORFs with unknown functions: CR3, CR4, and MURF2. The single gRNA for the minimally edited MURF2 mRNA is located on the maxicircle, while the putative complex I subunits CR3 and CR4 are extensively edited via minicircle-encoded gRNAs, despite their small size.

CR3 had complete gRNA coverage in IL3000 but not in the other two isolates. The unusual CR3 editing pattern described in *T. brucei* was also observed, as the mRNA does not depend on editing to generate its stop codon and the editing does not extend to the 3' UTR as it does in other mRNAs [225]. In *T. brucei* EATRRO1125, CR3 stands out from other pan-edited mRNAs, because eight initiation gRNAs have been identified while other mRNAs had only one or two gRNAs. Although we did not find quite as many CR3 initiation gRNAs in *T. congolense*, we detected five in IL3000 and seven in UPKZN, all encoded on cassette II of the respective minicircles, which still amounted to the highest initiation gRNA count of *T. congolense* pan-edited mRNAs (see Figshare). Surprisingly, for Kapeya we found only a single CR3 initiation gRNA on cassette II.

All three isolates had incomplete gRNA coverage over CR4 due to gaps at different editing sites. The 5' U-indels were predicted based on published *T. brucei* editing patterns to maximize conservation of ORF and protein sequences because neither Illumina nor PacBio had coverage over the first 150 nt of the predicted edited mRNA [217]. Each isolate had three CR4 gRNAs encoded in cassette I on the template strand, which had been reported in *T. brucei*.

4.5.7 Orphan gRNAs

The alignment-based method also identified in IL3000 five minicircles that contained orphan gRNAs, i.e. gRNA genes not located in the typical cassettes flanked by inverted repeats, including four encoding CYB gRNAs. While maxicircle-encoded CYB gRNAs have been found in *Leishmania braziliensis*, *L. peruviana*, *L. tarentolae*, and *C. fasciculata* (see below), all CYB gRNAs in *T. congolense* and *T. brucei* appear to minicircle-encoded as orphans [225, 343]. The orphan CYB gRNA genes on mO_022 and mO_092 were found downstream of the 120-

bp conserved region before cassette I, while CYB gRNA genes on mO_116 and mO_137 were downstream of cassette III.

A different scenario is represented by the 42-nt orphan ND8 gRNA gene on mO_105. The gRNA gene was located between cassettes I and III within the range of cassette II which was not identified for this minicircle. Close inspection of the sequence flanking the gRNA gene revealed remnants of the more conserved reverse repeat motif 'CTTA' aligned with other cassette II reverse repeats but no evidence of the forward repeat (Figure *4-18*). The gRNA gene was considered expressed based on the sRNA mapping and acted as the only gRNA that generated the first start codon at nt 51 on ND8 and non-redundantly covered nt 51 to nt 79. The same region was covered only by orphan gRNAs encoded in the same region (nt 432-490) on three minicircles each in Kapeya and UPKZN. The three minicircles in Kapeya did not encode other canonical gRNAs. However, in UPKZN, we also detected gRNAs mO_219(III)_gCOX3(546-592) and mO_235(III)_gCOX3(673-717) on cassettes III of two minicircles. We conclude that this ND8 gRNA gene is not a true orphan but represents a case where a gRNA gene probably remained functional and expressed despite the corruption of the flanking cassette. Alternatively, the degradation of cassettes may explain the scarcity of ND8 mRNAs fully edited over 5' regions.

GGGCGTGCAACTTTATTGAGATTTCTTGATACTGTAAACCCGTACCAAAATGGGCCAATTTTCACGGTTTTTGAGGAGGTTTGGGGTTCG
AAAGGGGTTGGTGTAATACCCGGTGGGGGTTTTCGGAGTTTGGAGGTGGGTGTTTATGTTCCATGGGGATTGGGGTTCTTGGGGTTCT
ATAAACCAATGGGACCAGGGAGCATACAATGATATAAAATAAGAGTTCTGGGATTAGTAGTTTATTAGTTTATTATATTTATTACATAAA
TAAAACAAACATTACCTACATAGAATCTAGCGCAGAGATCTATTAGAATAATGAACGAAGTGAATTATTATTAATATTGTGTTGTGTAAT
GTAGATTAATATCTCTTACTATATATTCTTTAATAGTCCGGTGGTGCAATAGTTGGGTGGATAGTGGTGGGAACGCATGGTTCAAGAAAC
AAATAAAAGGAAAGGATAGAGGGTTAAAGAGAAACGTGGGTGTGGTAGTAGGTGATTGGAGTTATGAGACTATGCTTAGTATGTTGA
ATAGAGAGGTGGGGTAATGAGACAGAGGTGGGGTAATGAGACAGAGGTGGGGTAATGAGACAGAGG
GTTGTGTGGAGGGCCAATGGGCGTGATGGTTATAATTGCCATGTTGGCTTAAGGCAGTGGGGTCACGGGCTAGAAAAGGATATTAGAT
AAGAGTTTATTGTTAGGTATATGCATTAGATAATATAATATAGATATAGGGGCACATCACCCCTTTATATTGATGTAATTAATTATAGATA
TTATTATAATTATAGAGCTCTTATTATATATGATTAATTAGTCAGTGCCGAGAGTCTTGTGGGGTGTAGTAATAAGTGTTCGGCTCTAGA
GGAACCCATGAAAATTTGCTGAAAATGGTCAATTTCGAAAAGGAACACTGCCGGATTTGGAA

Figure 4-18. **Annotation of mO_105.**

Highlights: red: CSBs, blue: forward repeats, yellow: reverse repeats/most conserved CTTA motif of the reverse repeat, grey: gRNA. Only the ND8 'orphan' gRNA is annotated on mO_105. The gRNA falls within the region of cassette II. The CTTA motif of the reverse repeat is detected, but no remnant of the forward repeat is found.

4.6 Organization of minicircle cassettes and gRNA genes in IL3000

We report here the organization of minicircle cassettes and gRNA genes in *T. congolense* IL3000. The analysis was performed with a custom Python pipeline. We focused on IL3000 due to the availability of sRNA transcriptome data.

4.6.1 Association between gRNA gene type, expression status, and cassette position

The 180 IL3000 minicircles contained 529 cassettes, of which 440 encode a single canonical gRNA (423) or multiple unique canonical gRNAs (17), excluding gRNAs that aligned to the common region of the two mRNAs with alternative editing (A6_v1/v2, ND3_v1/v2), and 89 encode non-canonical genes.

We investigated the expression status, position and sequence characteristics of gRNA genes using methods described in a previous study of EATRO1125, using the same definition for the transcription initiation region, initiation site, end position, and using transcriptome data of sRNAs extracted from BSF IL3000 parasites to assess transcription status [81]. After mapping sRNAs to minicircles, gRNA genes that exceeded the subjectively assigned threshold were considered expressed (0.025% of total read depth) [225]. However, the lack of biological replicates made the results of expression status tentative.

We identified 368 expressed gRNA genes and 161 non-expressed gRNA genes (Table *4-20*). Of the canonical gRNAs, 316 (72%) were expressed and 124 (28%) non-expressed. In contrast, we identified among non-canonical gRNAs 52 (58%) expressed gRNAs and 37 (42%) non-expressed gRNAs, suggesting that canonical gRNAs were significantly more likely to be expressed (X^2 =5.65, P=0.017).

Table 4-20. **Association between gRNA gene type and expression**

	expressed	**non-expressed**	**Total**
canonical	316 (72%)	124 (28%)	440
non-canonical	52 (58%)	37 (42%)	89
Total	368	161	529

Unlike *T. brucei*, which had fewer expressed canonical genes and more non-expressed non-canonical genes in cassette I, IL3000 exhibited an even distribution of gRNA gene type and expression across cassettes. The proportion of expressed canonical gRNA genes ranged between 94 (54%) in cassette II and 119 (66%) in cassette III, while other groups also had similar distributions across cassettes (Table *4-21*).

Table 4-21. **Association between cassette position and gRNA gene type and expression status**

	gRNA gene expression status/type				
cassette positions	**expressed canonical**	**non-expressed canonical**	**expressed non-canonical**	**non-expressed non-canonical**	**Total**
I	103 (59%)	43 (24%)	12 (7%)	18 (1%)	176
II	94 (54%)	43 (25%)	26 (15%)	10 (6%)	173
III	119 (66%)	38 (21%)	14 (8%)	9 (5%)	180
Total	316	124	52	37	529

We observed cassette size differences depending on gRNA type and expression (Figure *4-19*). The mean cassette sizes were slightly smaller than for *T. b. brucei* EATRO1125 for all categories [81], ranging from 137.2 nt for expressed canonical gRNA genes to 135.4 nt for non-expressed non-canonical gRNA genes (Table *4-22*). The sizes of non-expressed non-canonical cassettes were significantly more variable (Levene test for equal variance, P=0.0004) and smaller (Kruskal-Wallis test, H=10.10, P=0.018) than cassettes that contained canonical or expressed gRNA genes. The cassettes encoding expressed gRNA genes were not significantly longer than those encoding non-expressed gRNA genes (t-test, P=0.087), probably due to the limit of the sRNA data. The cassettes that encode canonical gRNAs were significantly longer (by ~1 nt) than those that encode non-canonical gRNAs (t-test, P=0.007).

Table 4-22. **Comparison of mean cassette sizes for different types of cassettes in IL3000 and EATRO1125 (nt)**

	expressed canonical	**non-expressed canonical**	**expressed non-canonical**	**non-expressed non-canonical**
IL3000	137.2	136.9	136.5	135.4
EATRO1125	142.6	140.2	138.1	136.4

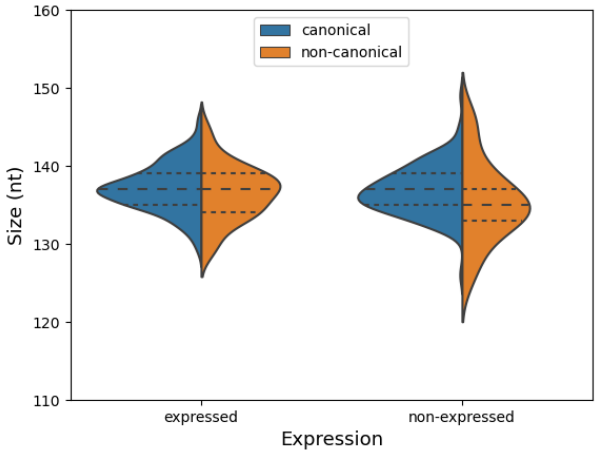


Figure 4-19. **Association between cassette size and gRNA type and expression status.**

Cassette size was measured as the distance from the 5' end of the forward 18-bp repeat to the 3' end of the reverse 18-bp repeat. Long-dashed lines: medians, short-dashed lines: first and third quartiles. Cassette sizes and standard deviations (in brackets) are as follows: expressed canonical: 137.2 nt (3.0 nt), expressed non-canonical: 136.9 nt (3.1 nt), non-expressed canonical: 136.5 nt (3.5 nt), and non-expressed non-canonical 135.4 nt (5.1 nt).

Unlike *T. brucei*, although the non-expressed and non-canonical gRNAs were shorter in *T. congolense*, their size difference did not drastically differ by type and expression. Among canonical gRNAs, the non-expressed gRNAs were less than 1 nt smaller than the expressed ones on average, while for expressed gRNAs, the canonical gRNAs were ~1.6 nt longer than the non-canonical gRNAs (Table 4-23, Table 4-24). The cutoffs for each quantile were similar between gRNAs from each type and expression status, suggesting a similar distribution. The homogeneous gRNA size in IL3000 probably contributed to the similar cassette sizes for all groups except for those encoding non-expressed non-canonical gRNAs (Table *4-22*).

Table 4-23. **Alignment length of expressed vs non-expressed canonical**

	count	mean	std	min	25%	50%	75%	max
expressed	347	42.8	5.5	25.0	41.0	44.0	47.0	55.0
non-expressed	139	42.1	4.9	28.0	40.0	43.0	45.0	54.0

Table 4-24. **Gene length of canonical vs non-canonical expressed**

	count	mean	std	min	25%	50%	75%	max
canonical	341	49.8	2.7	40.0	48.0	50.0	51.0	59.0
non-canonical	51	48.2	3.4	37.0	46.5	49.0	51.0	53.0

4.6.2 Nucleotide frequency of gRNA genes

We observed trends of the gRNA gene nucleotide frequency similar to those of *T. brucei* when aligning all expressed gRNA genes encoded in cassettes at the start of their initiation sequence (Figure *4-20*, top panels). Following a 10-nt AT-rich sequence leading to the initiation sequence start position, the subsequent 5-nt region contained most of the initiation sequences. The characteristic TA repeats in the initiation sequence (see below) gave rise to the two consecutive peaks in A/T frequencies. The anchors were aligned roughly from nt 5 to 16. The selection against GU wobble-base-pairing in anchors led to lower GT nucleotide frequency and a peak in C nucleotide frequency. The T nucleotide frequency in the anchor was higher in non-canonical gRNA genes, perhaps due to slackened selection for Watson-Crick base pairing. The region from nt 17 to 47 corresponded to the guiding sequence. We observed a steady decline in C nucleotide frequency from the peak in the anchor region. Meanwhile, G nucleotide frequency increased from around 20% to reach its peak of around 40% and then decreased again towards the 3’ end of the gRNA gene. T nucleotide frequency also rose from around 20% at the 5’ end of the guiding region to 40% at the 3’ end. In contrast, A nucleotide frequency oscillated around 40% over the entire guiding region, although non-canonical gRNA genes had greater fluctuation and overall lower A nucleotide frequency. The end of gRNA genes was marked by the abrupt decline of G nucleotide frequency after position 43, while T nucleotide frequency continued to increase. A and C nucleotide frequencies did not behave drastically differently from the guiding region.

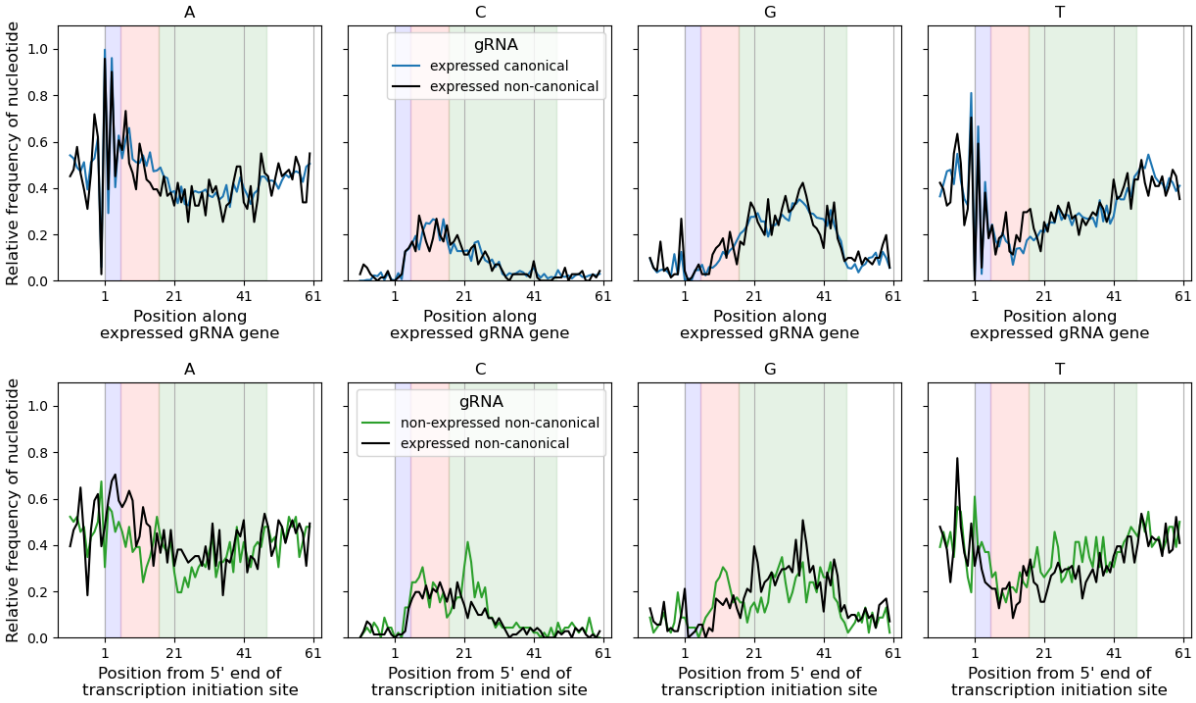


Figure 4-20. **Nucleotide frequency structure of gRNA genes.**

Initiation sequence: blue, anchor: red, guiding sequence: green. Top panels: nucleotide frequencies of expressed gRNA genes. The initiation region shows two consecutive AT peaks corresponding to the most common initiation sequence motifs. The anchor region is A-rich and G-poor. The frequency of C nucleotide decreases in the guiding region, while G and T nucleotide frequencies increase, and A nucleotide frequency remains roughly stable. Bottom panels: nucleotide frequencies of non-canonical gRNA genes. Except for the lack of defined initiation motifs, we observe similar trends in nucleotide frequency. Notably, the non-expressed non-canonical genes exhibit a C nucleotide peak from 21 to 28 nt.

The function of most non-canonical genes remains elusive, yet close inspection suggests that some non-canonical genes in *T. brucei* are homologous to known canonical genes but require gaps in alignment to be mapped onto the targeted regions on mRNA [81]. The non-canonical gRNA genes are probably responsible for uncharacterized alternative editing patterns or degraded from once functional gRNA genes due to the redundancy in gRNA gene coding on minicircles. The expressed non-canonical genes are probably at an earlier stage of degradation compared to the non-expressed non-canonical genes.

Hence, we investigated the difference in nucleotide frequency structure of expressed and non-expressed non-canonical gRNA genes to identify probable hallmarks associated with being actively transcribed. The limited sRNA data (one biological replicate at BSF) was probably not sensitive enough to capture all gRNA gene expressions, which resulted in mislabelling some expressed genes as non-expressed and obscuring the signal of gene degradation in the truly non-expressed genes. Nevertheless, we observed similar trends in IL3000 as reported in *T. brucei* that suggested further degradation of the non-expressed non-canonical gRNA genes.

Without sRNA mapping to infer the transcription initiation sites in non-expressed genes, we aligned all genes 30 nt downstream of the forward repeat, at the putative start of the initiation region proposed in *T. brucei* (Figure *4-20*, bottom panel) [81]. Because the imperfect alignment of TA repeats offset each other, we observed a smoother curve for AT

nucleotide frequency in non-expressed non-canonical genes. Similar to *T. brucei*, *T. congolense* exhibited a higher G nucleotide frequency over the anchor region in non-expressed non-canonical genes, probably due to slackened selection for Watson-Crick base-pairing in the anchors of degraded gRNA genes A decreased A nucleotide frequency over the same area compensated the higher G nucleotide frequency.

T. congolense exhibited a peak in C nucleotide frequency over 21 to 28 nt on the non-expressed genes in contrast to the overall decline of C frequency in expressed gRNA genes over the guiding region. A similar peak of C in non-expressed non-cannonical genes has been reported in *T. brucei* EATRO1125 [81]. We also observed an overall higher T and lower G nucleotide frequency over the guiding region of non-expressed non-canonical genes. The C peak, the corresponding valleys in AG frequencies, and the overall higher T frequency and lower G frequency were probably associated with a decrease in nucleotides that guided uridine insertions and the degradation of gRNA genes.

4.6.3 Characteristics of initiation sequences

The first five nucleotides of the gRNA transcript have traditionally (and somewhat arbitrarily) been defined as the transcription initiation sequence (Pollard et al, 1990; Koslowsky et al, 2014). The initiation sequence has been well characterized in *T. brucei* but remains understudied in *T. congolense*. For *T. brucei*, the consensus for gRNA gene transcription initiation sequence has evolved from a 5'-RYAYA (IUPAC codes: R = [A, G]; Y = [C, T]) [227] motif to the more specific motif of 5'-ATATA [349] present in about 60% of gRNA transcripts from expressed canonical genes in a detailed investigation of EATRO1125 [81].

In *T. congolense* IL3000, we identified 36 and 19 unique initiation sequence motifs in expressed canonical and non-canonical gRNAs, respectively. Similar to *T. brucei*, most (> 80%) expressed canonical gRNA genes had a T nucleotide and rarely (< 8%) an A nucleotide just upstream of the initiation sequence (Figure *4-20* top right and top left panels, respectively), which allows a slight extension of the motif beyond the 5' end of the transcript.

The most abundant motif in IL3000 expressed gRNA genes was 5'-ATATA, as in *T. brucei* (Table *4-25*). Nevertheless, *T. congolense* IL3000 only contained ATATA in 26% and 21% of expressed canonical and non-canonical gRNAs, respectively. We also detected two motifs with a C nucleotide in position 5 (AAAAC, ATAAC) in 12% and 17% of expressed canonical and non-canonical gRNAs respectively, while AAACA was present in 5% and 4% of expressed canonical and non-canonical gRNAs, respectively. We detected 14 motifs that followed the more lenient motif 5'-AWAHH (W = [A, T]; H = [A, C, T]) and accounted for 83% of expressed canonical gRNA genes. Compared to *T. brucei*, *T. congolense* had a greater diversity of initiation sequences, while the initiation sequence features between the canonical and non-canonical gRNA genes were more similar.

We also examined the initiation sequences of the putative non-expressed gRNA genes for the top eight most common initiation sequences to see if the lack of active transcription

may be manifested in the motif compositions. Without transcripts, we searched the 47th to 54th nt downstream of the forward repeats and detected 44 (27.33%) gRNA genes with the dominant motif ATATA and 31 (19.25%) with the second most common motif ATAAA. This is drastically different from *T. brucei*, where the presence of ATATA and other common initiation sequence motifs drops rapidly in non-expressed gRNA genes [81]. The proportions of AWAHH initiation sequence motifs in non-expressed gRNAs and expressed canonical gRNAs were around 80%, while around 60% of initiation sequences start with ATA (Table *4-25*). We attribute the similar motif composition between expressed and non-expressed gRNA genes to the lack of replica and insect-stage sRNA, as some gRNA genes may undergo stage-specific expression in the insect vectors and appear as non-expressed.

We detected sufficient sRNA transcripts mapped to all orphan gRNA genes to consider them expressed, except the one on mO_116. The expressed orphan gRNA genes on mO_022, mO_092, and mO_137 have initiation sequences AAAGA, AAATA, and AAAAT respectively, while we could not determine the initiation sequence for the non-expressed gene on mO_116. The identified orphan gRNA gene initiation sequences agree with the AWAHH consensus derived from gRNA genes associated with cassettes.

Table 4-25. **Relative frequency of the different motifs for the first three nucleotides of the initiation sequence in different types of gRNA genes**

Initiation sequence	Sequence	Expressed canonical	Expressed non-canonical	Non-expressed
Most common	ATATA	26%	21%	27%
5-nt consensus	AWAHH	83%	75%	80%
First 3 nt	ATA	63%	48%	61%
First 3 nt	AAA	28%	37%	10%

4.6.4 Characteristics of anchors

The gRNA anchor is defined as the longest, 5'-most, contiguous region complementary to the cognate unedited mRNA, following strict Watson-Crick base-pairing [350]. Interestingly, unlike *T. brucei*, most of *T. congolense* anchors along the expressed gRNA genes within cassettes overlap the initiation sequence for at least one nucleotide (Figure *4-21*A). In *T. brucei*, most anchors start around nt 5 to 7 from the 5' end [81, 349]. On the contrary, in IL3000, we observed 127 anchors that start with the first nt of the gRNA transcript, hence incorporating the initiation sequences completely into the anchor region. The next most abundant anchor starting positions are at nt 3, 4 and 5 (n = 94, 52, 92), thus still overlapping at least one nucleotide from the initiation sequences, and resulting in a median length of unmatched 5' sequence of 3 nt.

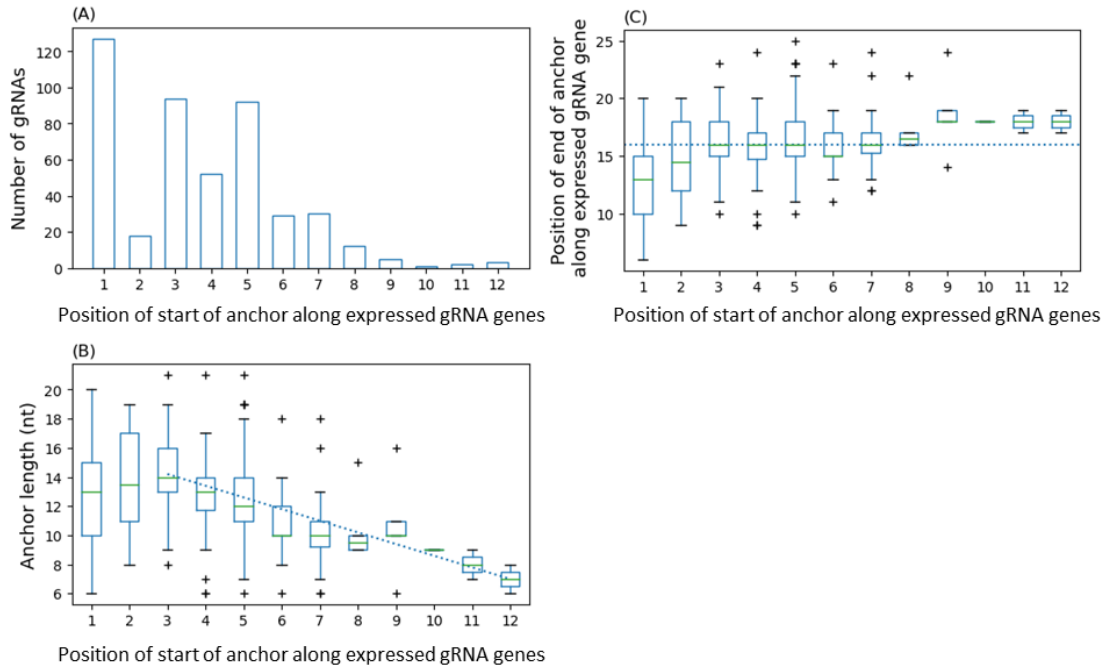


Figure 4-21. **Characteristics of anchors in expressed gRNA genes.**

(A) frequency of gRNA genes with anchors starting at each nucleotide positions along the genes. While 94, 52, and 92 genes have anchors that overlap the 5-nt initiation sequence by 3, 2, and 1 nucleotide, respectively, 127 genes have anchors completely overlapping the initiation sequences (i.e. the anchor starts at nt position one). **(B)** distribution of anchor length at each position along the gRNA gene. Green bars show the medians. Regression line: t = 27.2, P < 0.001, R^2 = -0.48. From the 3rd nucleotide, for each additional nucleotide separating the anchor from the start of the gene, mean anchor length reduces by 0.8 nt. **(C)** the distribution of the end position of anchors for anchors starting at each nucleotide along the gRNA gene. For anchors starting from the 3rd nt along the gRNA gene, over 50% have 3' positions that lie within 15 to 19 nt from the 5' end of the gene.

The mean anchor length is 12.6 nt (sd = 3.1 nt) with a range from six (the minimum cutoff of our canonical gRNA gene calling criteria) to 21 nt . For anchors starting from the third nt along the gRNA gene, for each additional nucleotide separating the anchor from the start of the gene, the mean anchor length reduces by 0.8 nt (regression line in Figure *4-21*B). The steady decrease in anchor length as the anchor starts further downstream of the gRNA gene results in a weak correlation between the 3' and 5' positions of an anchor (Figure *4-21*C, P =0.02, R^2 = 0.13). Consequently, for anchors starting from the third nt along the gRNA gene, over 50% have 3' positions within 15 to 19 nt from the 5' end of the gene, very similar to what has been observed for *T. brucei*.

We investigated the association between anchor positions in expressed canonical gRNA genes and their initiation sequences for anchors that start in the first 6 nt along the gene. Genes with anchors completely overlapping the initiation sequences have A-rich initiation sequences, with 35% AAAAN (N= [A, C, G, T]) and 78% AAANN. In 50 genes with AAAAY, 42 (84%) have anchors that overlap the initiation sequences completely. Meanwhile, the most common sequence ATATA includes 46 (54.12%), 6 (7.06%), and 15 (17.65%) genes with anchors starting at the 5th to 7th nucleotide and overlapping with the initiation sequences for < 1 nt. On the other hand, 50%, 21%, and 50% of genes with anchors starting at the 5th to 7th nucleotide contain ATATA as initiation sequences. We concluded that a substantial

overlap between the anchor and initiation sequence is associated with the A-rich initiation sequence.

4.6.5 Cassette structure

The forward and reverse repeats of each cassette (n = 529) exhibited a striking complementarity (GU wobble base-pairing was allowed) with an average of 14.7 nt (sd = 1.6 nt) of matched nucleotides. We observed an almost identical level of complementarity between randomly paired forward and reverse cassettes (n = 529, mean = 14.69, sd = 1.66).

Similar to *T. brucei*, the first two cassettes lie closer to each other than cassettes II and III (Figure *4-12*). In EATRO1125, a gap of 110.2 nt separated cassettes II and IV, where the less common cassette III could sometimes be detected [81]. Although *T. congolense* minicircles contained at most and generally three cassettes, we also observe the wider gap following cassette II. On average, cassettes I and II are 41.19 nt (sd = 14.70 nt) apart, while cassettes II and III are 167.34 nt apart (sd = 28.62 nt). Whether the structural similarity bears any biological relevance and homology between *T. congolense* and *T. brucei* remains to be explored. IL3000 also exhibited the trend of longer minicircles associated with a wider distance between the A-tracks and cassette III. In a longer minicircle, the starting positions of the A-track and cassette III were located further downstream and upstream, respectively (Figure *4-12*).

We summarized the typical structure of an expressed canonical gRNA gene within a cassette (median cassette length 134 nt) in Figure *4-22*. Admittedly, we did not confirm that the 5' ends of the sRNAs had a triphosphate and truly represented where the transcription began. The gRNA gene consisted of the unmatched 5' end sequence (blue rectangle), anchor (red rectangle), and guiding (green rectangle) sequences. The 3' end of the forward repeat and the 5' end of the transcript (blue histogram) were 30-32 nt apart (median = 30 nt). We use the 90th percentile position of mapped 3' ends of transcripts with oligo(U)-tails (orange histogram) to mark the end of the gRNA gene and define the gRNA gene as the sequence between the medians of the initiation sequence end positions and 3' ends of mapped transcripts, which had a median length of 50 nt.

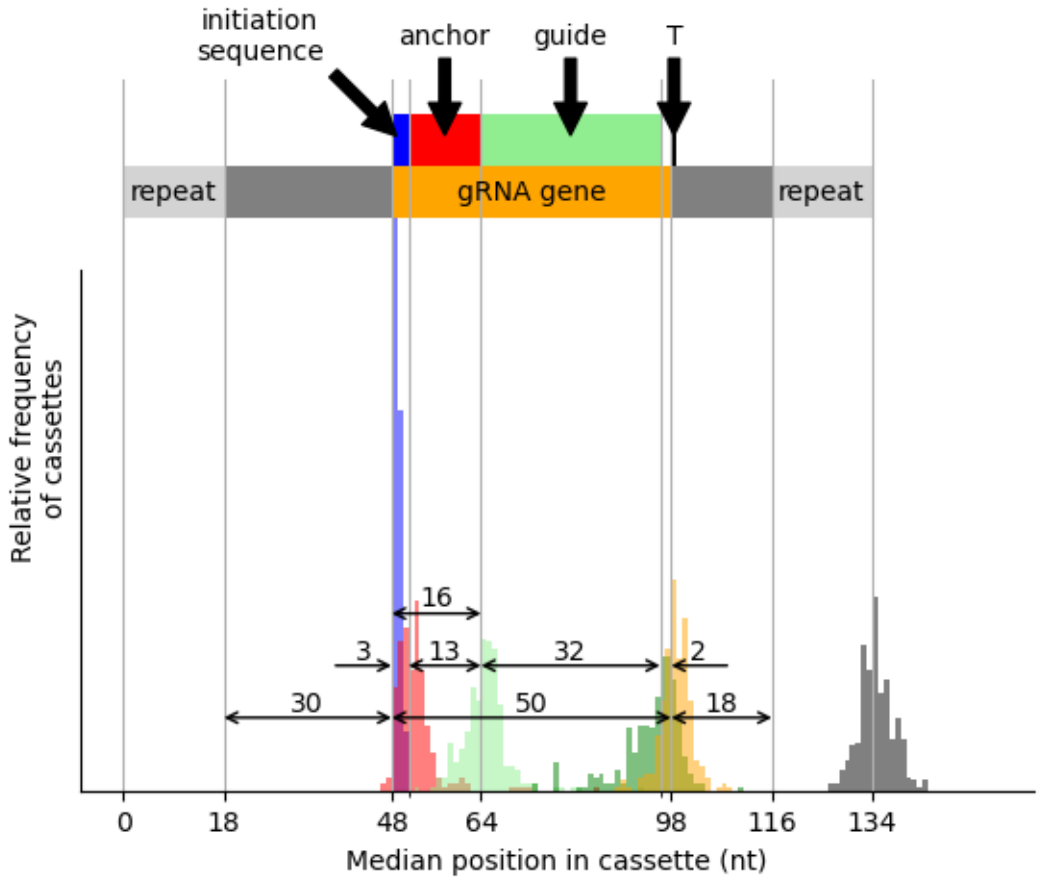


Figure 4-22. **Structure of typical cassettes encoding expressed canonical gRNA genes aligned at the 5′ end of the 18 bp forward repeat (position 0 on the x-axis).**

Histograms show the distributions of the positions and lengths of features expressed canonical gRNA genes and their transcripts, relative to the 5′ end of the forward repeat. Blue: 5′ ends of 3 nt initiation sequences (the transcription start site inferred from sRNA data); magenta: 5′ ends of anchor sequences (the stretch of Watson-Crick complementarity between the 5′ gRNA region and target mRNA, up to the first mismatch or non-Watson-Crick basepair between gRNA and unedited mRNA); light green: 5′ ends of guiding regions or the start of guiding region; dark green: 3′ ends of guiding regions (defined to start at the end of the anchor region, the guiding region is the region of complementarity between gRNA and fully edited mRNA); orange: 3′ ends of gRNA genes (the last nucleotide in the corresponding transcript before the non-templated oligo(U)-tail); and dark grey: 3′ ends of 18 bp reverse repeats. The coloured bars represent the positions of the gRNA gene components based on their median positions within cassettes. The double-headed arrows indicate median distances between cassette features.

The unmatched 5′ end sequence had a median length of 3 nt and represented part of the initiation sequences that overlapped the anchors. The anchor (magenta histogram, median length 13 nt), led to the guiding region (median length 32 nt) defined between the median of the start (light green histogram) and end (dark green histogram) positions. The anchor and the guiding region added up to the ~45 nt complementary sequence. We also observed a ~2 nt gap between the medians of the 3′ ends of the guiding region (dark green histogram) and the gene (orange histogram), similar to what was described previously for *T. b. brucei* EATRO1125 [81]. The initiation sequence (3 nt) and the 3′ end gap (2nt) explained the 5 nt length difference between the complementary sequence (45 nt) and the gRNA gene (50 nt).

4.7 Conservation of editing blocks within and between isolates and subspecies

4.7.1 Identification of gRNA families in *T. congolense*

Although we observed only a few minicircles shared among *T. congolense* isolates, we noted in the annotations that gRNAs aligned to the same areas on mRNAs often were encoded on the same cassettes on minicircles. We therefore investigated if gRNAs of *T. congolense* could also be classified into families of functional homologs, as we had observed in sub-Saharan *T. brucei*.

To define delimitations between *T. congolense* gRNA families, we grouped the gRNAs based on the positional conservation of anchor sequence alignments instead of ISSPs, as in *T. brucei*. This difference in methodology was necessitated by the smaller sample size of *T. congolense* to generate sufficient overlaps over the conserved positions for us to identify the gRNA clusters without too much arbitrary interference. This way, we did not have to define the number of gaps to be tolerated within each cluster, which would be difficult to conclude from only three samples.

In line with previous studies [225, 226], the gRNA anchor was defined as the longest undisrupted Watson-Crick base pairing starting from the 5' end of the gRNA. We aligned the anchors of gRNAs from each isolate to corresponding positions on the edited mRNAs and observed highly conserved anchor positions (Figure *4-23* shows A6 as an example, see Supplementary Figure 8 for all mRNAs). The anchor positions from the three isolates (shown as the sum anchor coverages) overlapped to form semi-regular clusters or 'islands' of non-zero values along the mRNAs. The depth of anchors at each nucleotide of the mRNA (i.e. the number of different gRNA anchors aligning with that nucleotide) was calculated for IL3000, Kapeya, and UPKZN, as was the sum of the anchor depths from the three isolates.

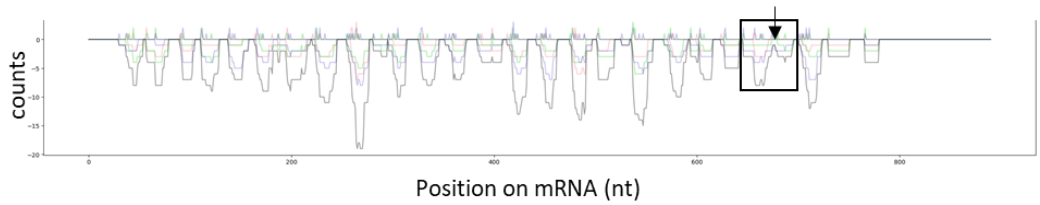


Figure 4-23. **The guiding region starting positions (above the line) and the coverage of anchors (below the line) over edited A6_v1 mRNAs**.

IL3000: red, Kapeya: green, UPKZN: purple. The peaks of anchor coverage occur with semi-conserved intervals. The sum of anchor coverage (black) from three isolates was used to assign gRNA families. The circled area exemplifies an island that required manual subdivision at the valley into separate clusters (indicated by the arrow).

We thus defined the boundaries of the islands as regions with zero anchor coverage. This method allowed us to identify 225 anchor clusters, 213 of which contained a single peak and were accepted for subsequent gRNA family assignment (Supplementary Figure 8). Eleven islands contained two peaks and one contained three peaks of anchor coverage, indicating the potential merging of at least two adjacent gRNA families. We refined the delimitations by subdividing the clusters with multiple peaks via the distinct valleys inbetween peaks (see

Figure *4-23* for an example). The delimitation broke the 12 islands into 25 clusters each with a single peak within its bound. Hence, in total we identified 238 anchor clusters.

We defined a gRNA family as a set of gRNAs with anchors within the same island and assigned 2344 gRNAs from the three *T. congolense* isolates to the 238 gRNA families. The number of families was comparable to the 250 gRNA families identified using 224 *T. brucei* isolates. IL3000, Kapeya, and UPKZN each had 220, 228, and 226 gRNA families, respectively (Table *4-26*). Although Kapeya and UPKZN had more gRNA families, their gRNA coverage still had more gaps compared to IL3000, probably due to unaccounted-for variations in editing patterns. Most of the additional gRNA families were 'out-of-phase' and probably should be dismissed as false positives (Supplementary Figure 7). The phasing and the conservation of the gRNA families will be discussed in the following sections.

Table 4-26. **Summary of gRNA families in four *T. congolense* isolates**

	IL3000	**Kapeya**	**UPKZN**
A6_v1	25	27	27
A6_v2	25	27	27
COX3	31	32	33
CR3	7	7	8
CR4	17	16	17
CYB	2	2	2
ND3_v1	12	13	12
ND3_v2	12	12	13
ND7	40	41	39
ND8	17	19	17
ND9	22	22	21
RPS12	10	10	10
Total	220	228	226

4.7.2 Template strand gRNAs

Prior to this study, gRNAs had only been found on the coding (sense) strand of minicircles in *T. congolense* [204, 223, 228, 289]. However, we found six, nine, and 13 gRNAs encoded on the template (antisense) strand in IL3000, Kapeya, and UPKZN respectively (Table *4-27*). Those gRNAs belonged to seven gRNA families, while the families on both versions of A6 and ND3 were identical. Except for the A6 (25 nt) and ND3 (27 nt) gRNAs and one CR4 gRNA (31 nt), other gRNAs were over 40 nt. Except for five gRNAs for the mitoribosomal subunit RPS12 on cassette II and two ND3 orphan gRNAs, the remaining 21 gRNAs were located on cassette I. The most antisense strand gRNAs in *T. brucei* are also found in cassette I [225].

Table 4-27. **Counts of gRNAs encoded on the template strand in *T. congolense* isolates**

gRNA families	Cassette	Length	IL3000	Kapeya	UPKZN	Total
ND9-151_169	I	43-48	1	3	1	5
RPS12-314_333	II	44-47	2	1	2	5
CR4-401_415	I	31-45	3	3	3	9
A6_v1-646_675	I	25	0	1	2	3
A6_v2-646_675	I	25	0	1	2	3
ND3_v1-115_147	Orphan	27	0	0	1	1
ND3_v1-115_147	Orphan	27	0	0	1	1
Total			6	9	13	28

4.7.3 Conservation of gRNA families

We examined the conservation of gRNA structure by aligning the gRNAs to the corresponding edited mRNA and superimposing gRNA homologs from the same family. The structural similarity could be appreciated intuitively from the clear distinction of superimposed initiation, anchor, and editing sequences, respectively (Figure *4-24* B, see Supplementary Figure 7 for all mRNAs). Taking the collective guiding regions from all gRNAs as the effective editing range of a family, we observed little overlaps, hence redundancies, between the effective editing ranges of adjacent families. The minimal overlaps strongly support that editing proceeds in a cascade of discreet blocks.

Most gRNA families were dominated by gRNAs encoded on the same cassette positions (Figure 4-24 A). 144 of the 238 gRNA families contained gRNAs from a single cassette, while 231 gRNA families contained over 50% of gRNAs encoded on the same cassette (Table 4-28). The conserved cassette positions suggested that the gRNAs and minicircles probably shared a common origin despite their apparent divergence in sequence. As in *T. brucei*, we also observed cases where a gRNA family contained gRNAs from two major cassette positions. For instance, the family ND7-1134_1153 contained two groups of gRNAs, encoded in cassettes I and II, respectively; the start positions of anchors of gRNAs from the same cassette clustered closely together (Figure 4-24 B). The clustering suggested that within a gRNA family, gRNAs encoded on the same cassettes were more closely related. We will explore this proposal in more detail in later sections.

Table 4-28. **gRNA family members tend to be encoded in the same cassette**

Maximal percentage of gRNAs from the same cassette	Count of gRNA families
50%	234
60%	221
70%	211
80%	185
90%	157
100%	148

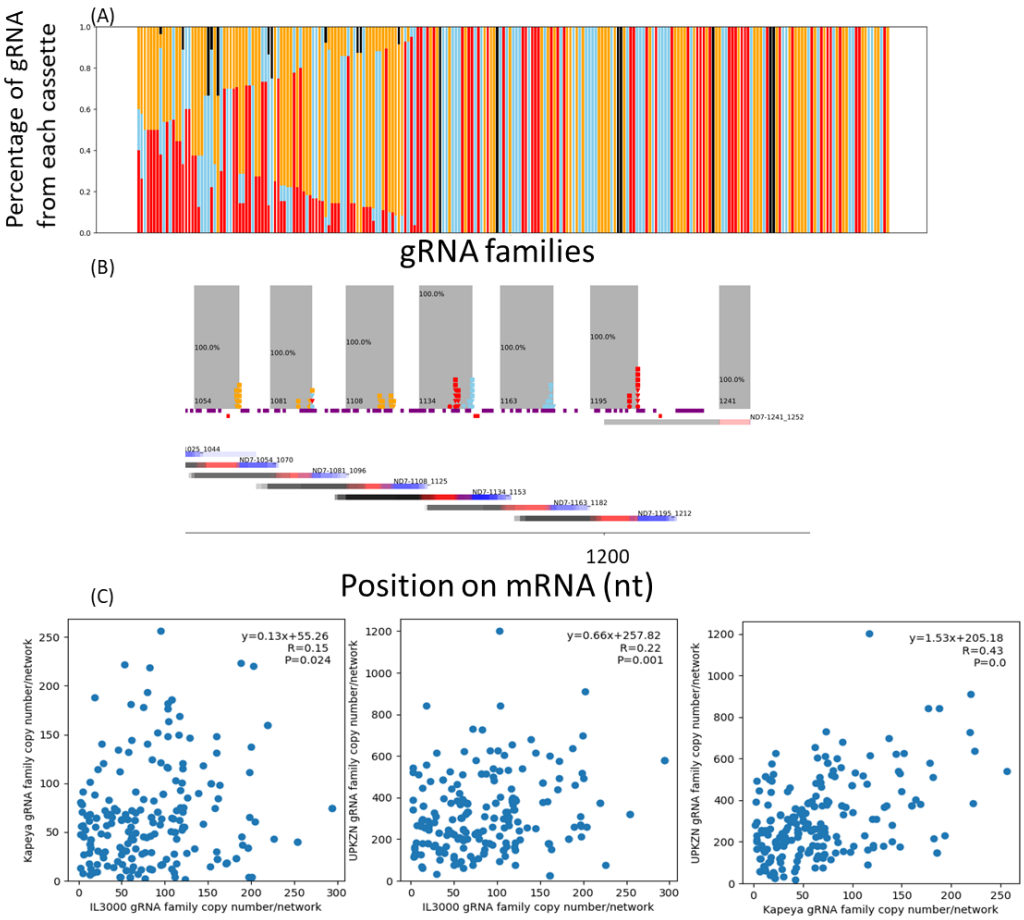


Figure 4-24.**The conservation of gRNA families.**

(A) Each bar represents a gRNA family, for which we show the percentage of gRNAs from each cassette. The gRNA families were sorted in ascending order by the percentage of the most dominant cassette. 144 gRNA families contained gRNAs encoded on the same cassette. Colour coding of cassettes: I: red, II: blue, III: orange, orphan: black. **(B)** gRNA anchor end positions and gRNA coverages of T. congolense isolates on ND7 edited mRNA (screenshot from Supplementary Figure 7). The family ND7-1134_1153 contained gRNAs from cassettes I (red) and II (sky blue). All three isolates encoded the family in both cassettes. The shaded blocks show the range of ISSP clusters. The scatterplot shows the ISSP count at each position, while ISSPs for gRNAs from different cassettes are color-coded (I : red, II: skyblue, III: orange, IV: purple, V: green, Orphan: black). Insertions are plotted in purple blocks, and deletions are indicated below with red blocks. Below the editing sites, the coverage of gRNA families are plotted by overlapping gRNAs from the same family (initiation sequence: blue, anchor: red, guiding sequence: black). **(C)** The abundance of a gRNA family with a kDNA network is estimated as the sum of the copy numbers of all the gRNAs it contains. The gRNA family abundances between any two isolates have a significant but weak linear correlation (linear least-squares regression, $0.15 \leq R \leq 0.43$, $P \leq 0.024$).

Most of the 238 gRNA families were shared by all *T. congolense* isolates. We detected 209 families in all three isolates, while 18 and 11 were detected in two isolates and one isolate, respectively. The 11 isolate-specific gRNA families all contained only one gRNA member, except family ND9-588_594, which contained two. The gRNAs from the 11 families were all shorter than 30 nt, except for COX3-972_987, a COX3 initiation gRNA only identified for IL3000. COX3-972_987 was also the only gRNA whose removal would result in additional gaps in gRNA coverage. Hence, we speculate that the other gRNA families represent false positives or sequences with little functional relevance. The conservation of 209 gRNA families suggested that the *T. congolense* isolates shared a similar arrangement of editing blocks and functionally and evolutionarily conserved gRNA gene repertoires, despite their distinct minicircle populations.

We next compared the relative abundance of copy numbers of gRNA gene families among isolates. We calculated this abundance as the sum of the copy numbers of the gRNA gene members it contained and calculated the relative abundance correlation of gRNA families shared between *T. congolense* isolates. We detected a weak linear correlation (r=0.43) between Kapeya and UPKZN, while the correlation between any two isolates was significant ($p<0.01$) (Figure *4-24* C). Hence, the relative abundance of gRNA families appeared to be conserved, suggesting that the kDNA network probably curated the relative abundance of gRNA genes based on their functionality.

4.7.4 Comparison of editing block positions between *T. congolense* and *T. brucei*

In Chapter 3 we reported the conservation of editing blocks among sub-Saharan *T. brucei*. We were interested if the editing blocks were conserved between *T. brucei* and *T. congolense*. Using the minicircle annotation of the reference *T. b. brucei* strain EATRO1125 [225], we adjusted the edited mRNAs to be aligned at the putative start codons, so the relative positions of gRNA alignments became comparable. We observed extensive overlaps of anchors between *T. brucei* and *T. congolense* isolates, which enabled us to identify cross-species gRNA families and editing blocks using the same method we employed for *T. congolense* (Figure *4-25* A, Supplementary Figure 9).

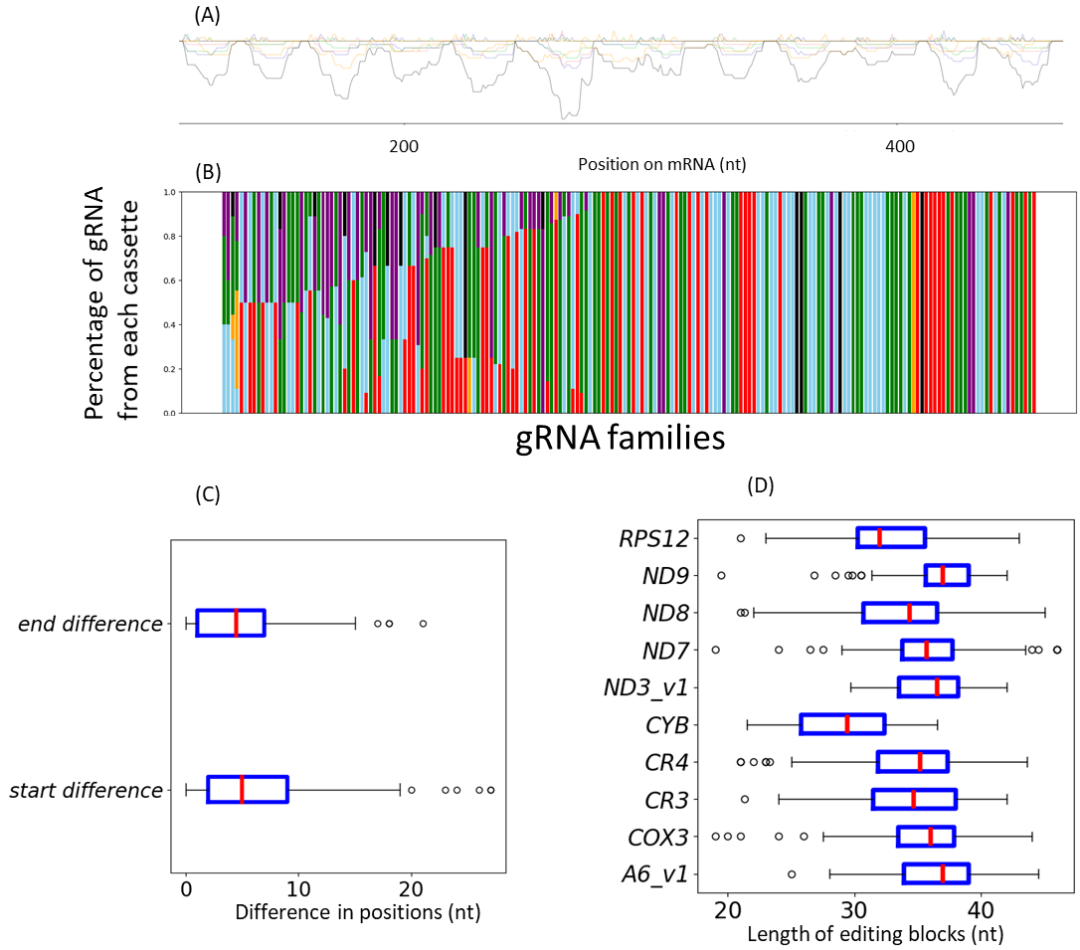


Figure 4-25. **Summary of editing block identification with *T. congolense* and *T. brucei* EATRO1125.**

(A) Comparison of anchor starting positions (top) and coverage of anchors (bottom) between *T. congolense* and *T. b. brucei* for mRNA A6. IL3000: red, Kapeya: green, UPKZN: purple, *T. b. brucei* EATRO1125: orange. The peaks of anchor coverage occur with semi-conserved intervals. The sum of anchor coverage (black) from three isolates was used to assign gRNA families. **(B)** The proportion of gRNAs encoded on each cassette in EATRO1125 gRNA families. Cassette colour coding: I: red, II: skyblue, III: orange, IV: green, V: purple, orphan: black. **(C)** Maximal difference in the start and end positions of editing blocks among the three *T. congolense* isolates and EATRO1125. Difference in start positions: mean: 6.7, median: 5.6, max: 22.7, min: 07. Difference in end positions: mean: 6.1, median: 5.5, max: 19.3, min: 0. **(D)** The effective editing range representative of each gRNA family is calculated by taking the average of the start and end positions of the editing blocks from the four subspecies. Pan-edited mRNAs have conserved average editing block length between 32.77 nt in RPS12 and 36.56 nt in ND9. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point

To avoid unnecessary ambiguity, we merged gRNAs of different versions of the same mRNAs, as we did in Chapter 3 when comparing *T. brucei* subspecies with different editing patterns. Omitting gRNA families for alternatively edited mRNAs reduced the number of gRNA families compared to 4.7.1. After careful visual examinations and manually subdividing the boundary of anchor clusters with multiple peaks at individual valleys, we assigned gRNAs into families using the same logic based on the islands of overlapping anchors, which yielded 200 gRNA families in total, of which 189 were present in EATRO1125 and 167 were shared by EATRO1125 and the three *T. congolense* isolates (Table *4-29*).

Table 4-29. **Summary of gRNA families called from three *T. congolense* isolates and *T. b. brucei* EATRO1125**

	IL3000	Kapeya	UPKZN	EATRO1125
A6	25	27	26	26
COX3	31	31	32	35
CR3	7	7	8	7
CR4	17	16	17	18
CYB	2	2	2	2
ND3	12	12	12	11
ND7	38	39	39	40
ND8	17	19	17	21
ND9	19	20	19	19
RPS12	9	10	10	10
Total	177	183	182	189

The minicircles cassette structures differed between the two species, as *T. congolense* invariably utilized three cassette positions, while *T. brucei* minicircles contained as many as five potential cassettes. Expectedly, *T. b. brucei* EATRO1125 encoded most gRNA families on cassettes different from *T. congolense* isolates. EATRO1125 encoded 76 families on cassettes IV and V, which are absent from *T. congolense* minicircles. Among the 178 gRNA families shared by *T. brucei* and *T. congolense*, only 24 had the same dominant cassettes. Within EATRO1125, we also observed the conservation of cassette positions amongst gRNAs from the same family, with 97.9% of gRNA families containing at least 50% gRNAs from the same cassette and 55.0% of families containing gRNAs from a single cassette (Figure *4-25* B, Table *4-30*).

Table 4-30. **EATRO1125 gRNA family members also tend to be encoded in the same cassette.**

Maximal percentage of gRNAs from the same cassette	Count of gRNA families
50%	185
60%	162
70%	143
80%	124
90%	107
100%	104

Note: 189 EATRO1125 gRNA families were analysed.

After grouping gRNAs into families, we calculated the range of editing blocks in each isolate using the same method as in Chapter 3. In short, we identified the editing block of a gRNA family in each isolate as the region between the mean starting and ending points of the complementary sequences of all gRNAs it contained, excluding the first six nt with respect to the gRNA. The maximal difference of the start and end positions of editing blocks among four isolates had averages of 6.7 and 6.1 nt respectively (Figure *4-25* C). The averages were about twice as high as the difference calculated from *T. congolense* isolates alone (data not shown), yet this still supported the conservation of editing block end positions across species.

The average lengths of the editing blocks were conserved among pan-edited mRNAs, ranging between 32 and 37 nt with an overall average of 35 nt (Figure *4-25* D, Table *4-31*). A similar range for editing block length was observed in sub-Saharan *T. brucei* in Chapter 3.

Similar to what we observed in sub-Saharan *T. brucei*, the minimally edited CYB (which is unique as all gRNAs are encoded in orphan cassettes) had a shorter 5' editing block around 21 nt and a longer 3' editing block around 36 nt. The minimal lengths of editing blocks in pan-edited mRNAs were also around 20 nt. Since CYB only had two editing blocks, the shorter 5' editing block substantially reduced the average block length.

Table 4-31. **Summary of descriptive statistics of the length of editing blocks on edited mRNAs**

	mean	median	max	min
A6	36.4	37.0	44.5	25.0
COX3	35.3	36.0	44.0	19.0
CR3	34.5	34.7	42.0	21.3
CR4	34.3	35.2	43.6	21.0
CYB	28.9	29.4	36.5	21.5
ND3	36.1	36.5	42.0	29.7
ND7	35.8	35.7	46.0	19.0
ND8	33.3	34.3	45.0	21.0
ND9	36.6	37.0	42.0	19.5
RPS12	32.8	32.0	43.0	21.0

We observed in four isolates an interval distance of around 30 nt (Table *4-32*). The standard deviations were high in ND7 due to the distance between the two editing domains. The short and seemingly redundant 5' most editing block and the four overlapped 3' editing blocks also increased the standard deviation in RPS12.

The editing blocks of *T. congolense* and *T. brucei* isolates covered the same regions on mRNAs, confirming that in both species the editing cascade proceeded at similar paces from the initiation gRNAs (Figure *4-26*). The distinct minicircle structures and cassette profiles of *T. brucei* and *T. congolense* gRNA families indicated prolonged divergence. Nevertheless, similar editing block patterns suggested that the progression of the editing machinery was universally conserved for African trypanosomes.

Table 4-32. **The intervals between editing blocks in EATRO1125.**

	IL3000		***Kapeya***		***UPKZN***		***EATRO1125***	
	mean	**sd**	**mean**	**sd**	**mean**	**sd**	**mean**	**sd**
A6	30.10	8.51	28.01	7.17	29.15	8.04	28.14	7.83
COX3	30.25	6.52	29.47	7.44	29.50	6.99	29.81	8.32
CR3	32.87	11.12	35.40	5.79	30.36	14.71	33.28	7.00
CR4	28.16	6.37	30.63	12.66	28.32	9.34	30.17	4.49
ND3	31.55	4.83	32.53	7.23	31.29	4.44	35.72	10.67
ND7	31.46	14.29	30.67	14.26	30.61	13.43	30.55	15.07
ND8	26.75	8.96	26.24	9.42	27.77	12.18	24.81	6.93
ND9	31.33	11.89	31.33	10.34	31.25	10.60	31.47	10.49
RPS12	31.75	13.44	27.44	13.23	27.52	12.62	26.56	15.78

Note: sd: standard deviation

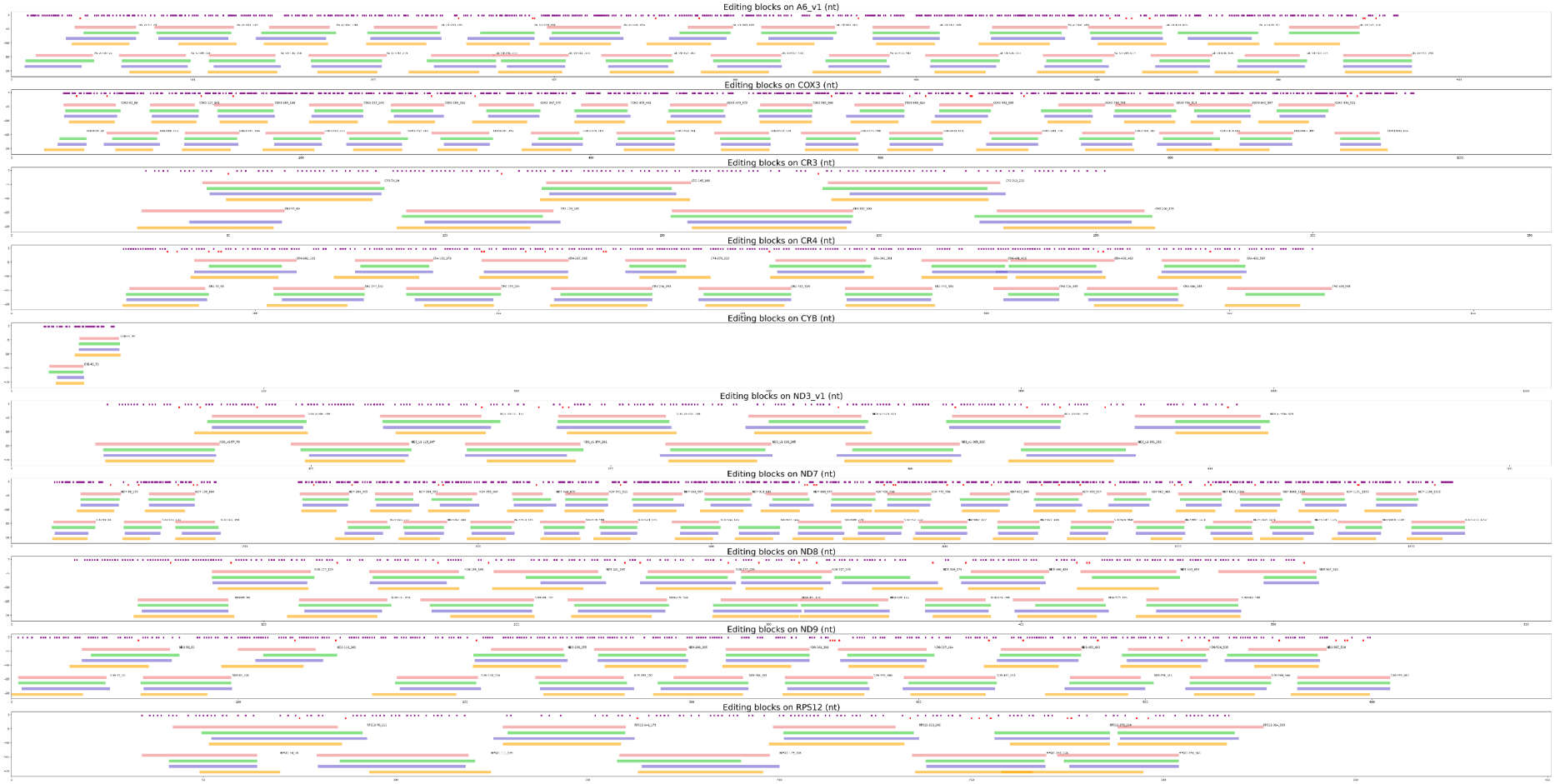


Figure 4-26. **Comparison of editing blocks of *T. congolense* and *T. b. brucei***.

Insertions are plotted in purple blocks, and deletions are indicated below with red blocks. The bars represent editing blocks and are drawn in the following order and colour scheme: IL3000: red, Kapeya: green, UPKZN: purple, EATRO1125: orange.

4.8 Chapter conclusions

We annotated the *T. congolense* maxicircles and identified SNPs among *T. congolense* isolates. In addition, having examined the gRNA genes in the maxicircle coding region of trypanosomatids, we concluded that many gRNA genes overlapped other maxicircle genes and were conserved among trypanosomatids, but they were no longer functional in *T. congolense.* We also identified evidence for gRNA gene insertion or deletion on *T. brucei* and *T. congolense*. Furthermore, using transcriptome data, we identified alternative editing patterns distinct from those in *T. brucei* on A6 and ND3 [225]. PacBio reads also suggests that although the initiations of maxicircle genes transcription are independent, the extension on 3' end allows the transcription to proceed onto the adjacent gene(s).

The annotation of minicircles gave nearly complete gRNA coverages in three isolates. We observed a similar nucleotide frequency structure of gRNA genes as in *T. brucei* [81], which suggested that the number of nucleotides that could guide uridine insertions decreased in non-expressed non-canonical gRNAs. Additionally, the initiation sequences often overlapped the anchors in *T. congolense*. We also observed a steady decrease in anchor length as the anchor starts further downstream of the gRNA gene as in *T. brucei* [81], which indicated a molecular ruler that dictates the distance between the 5' end of the gRNA genes and the 3' end of the anchors.

The conserved mapping positions of the anchors on edited mRNAs allowed us to group gRNAs into families as in sub-Saharan *T. brucei*. We observed a similar conservation of editing blocks along the mRNAs, which indicated that mRNA editing occurred in semi-regular steps with minimal overlaps to allow gRNA selection by the anchors. The editing blocks were conserved between *T. congolense* and *T. b. brucei*. We proposed that the common features in their editing complex resulted in the conservation of the editing cascade.

5. kDNA assembly and annotation of *T. b. equiperdum* and *T. b. evansi*

In this project, we tried to elucidate the impact of tsetse-independent transmission on *T. brucei* subspecies. First, we wanted to confirm the minicircle compositions in groups with homogeneous minicircle populations. To characterize the kDNA, we performed *de novo* assembly to detect and recover minicircles and maxicircles. Second, the assignment of *equiperdum* and *evansi* has been controversial historically, and type A contains isolates from both groups. We were interested in how *T. b. equiperdum* and *T. b. evansi* isolates were related within and between groups. Hence, we constructed type A phylogeny with sequences of type A minicircles from *T. b. equiperdum* and *T. b. evansi* within the group and the highly similar minicircles detected in tsetse-transmissible Sub-Saharan *T. brucei*. The phylogeny of isolates with partial or complete maxicircles is discussed in 6.4.3. Third, we detected a heterogeneous minicircle population in type OVI isolates, which raised the question of how much editing capacity they retained. We annotated the *de novo* assembled maxicircle by alignment to EATRO1125 maxicircle and extracted the unedited mRNAs. Without mRNA sequencing data, the EATRO1125 edited mRNAs were modified to reflect the SNPs unique to type OVI, using the same method as the preliminary edited mRNA predictions in 3.2.2. We then annotated the minicircles using the procedure described in Chapter 3 and 4 [225] and visualized the gRNA coverage on each mRNA.

5.1. Curation of the samples

Our collaborators at ITM Antwerp supplied WGS Illumina data including all nuclear reads of 38 *T. b. equiperdum* and *T. b. evansi* isolates. The reads were first mapped to the *T. b. brucei* EATRO1125 maxicircle coding region [225], which revealed a large number of heterogeneous SNPs in AnTat-4-1, AnTat-4-1-bis, Kenya-a, and Kenya-c. Given the expectation of homogeneous maxicircles within a kDNA network, we concluded that these samples contained kDNA from multiple isolates and were unsuitable for subsequent analysis. E28 was removed for reason elaborated in 5.3.2.

Upon their removal, the dataset consisted of 33 isolates, including 26 type A, one *T. b. evansi* type B, and six *T. b. equiperdum* (Supplementary Table 5). Type A, B, and C minicircles have been described in *T. b. evansi* and *T. b. equiperdum* with a homogeneous minicircle population, and the groups are named after the type of minicircles they contain [87, 88]. Notably, the type A group contained both *T. b. evansi* and *T. b. equiperdum* due to complications in the diagnosis history [84]. A PCR specific based on the RoTat 1.2 VSG has grouped eight historical type A *T. b. equiperdum* isolates with type A *T. b. evansi*, including SVP, ATCC-30019, ATCC-30023, STIB818, Alfort, AnTat-4-1 (removed), and isolates from Canada, Hamburg, and America [183]. The phylogenies based on random amplified polymorphic DNA (RAPD) and multiple endonuclease genotyping approach (MEGA) also group them with type A *T. b. evansi* [183]. The grouping suggests potential issues with the convention of *T. b. equiperdum* and *evansi* classification, which we will discuss in Chapter 8.

Other *T. b. equiperdum* isolates included two samples of type C isolate BoTat-1-1 [91] and four samples of closely related OVI, Te-Ap-N-D1, and Dodola (940) [90] which we collectively

termed type OVI. E28 was removed from *T. b. equiperdum* as explained later (See Supplementary Table 1 for the complete metadata for the WGS samples).

Our lab had previously sequenced the kDNA-enriched samples of 10 isolates (Table 3.2 in [73]) and these were re-assembled and re-analysed together with the above isolates. The sample included five *T. b. evansi* and five *T. b. equiperdum*. Two isolates (American, Hamburg) were labelled as type A '*T. b. equiperdum*' [351].

Table 5-1. **Summary of *T. b. equiperdum and T. b. evansi* isolates used in this study**

	WGS	kDNA	Total
***T. b. evansi* type A**	17	3	20
***T. b. evansi* type B**	1	2	3
***T. b. equiperdum* type A**	9	2	11
***T. b. equiperdum* type C**	2	1	3
***T. b. equiperdum* type OVI**	4	2	6
Total	33	10	43

5.2. Maxicircle assembly and annotation

5.2.1. Maxicircles in *T. b. equiperdum* and *T. b. evansi*

To detect maxicircles in the *T. b. equiperdum* and *T. b. evansi* isolates, reads from the 43 samples (Table *5-1*) were mapped to the coding region of the EATRO1125 maxicircle [225]. The mapping confirmed that 18 *T. b. evansi* type A, five *T. b. equiperdum* type A, and the three type B isolates did not contain maxicircles. We next performed *de novo* assembly with KOMICS [66] and detected complete or partial maxicircle in the remaining isolates. We detected reads that bridged the gaps on maxicircles with deletions, which confirmed that the deletions were not due to assembly error.

We observed two conserved deletions, a shorter 148-bp conserved deletion in the nine type C and OVI isolates and a large 8.9 kb deletion corresponding to nt 2330 to nt 11189 on the EATRO1125 reference conserved among most type A isolates (Figure *5-1*).

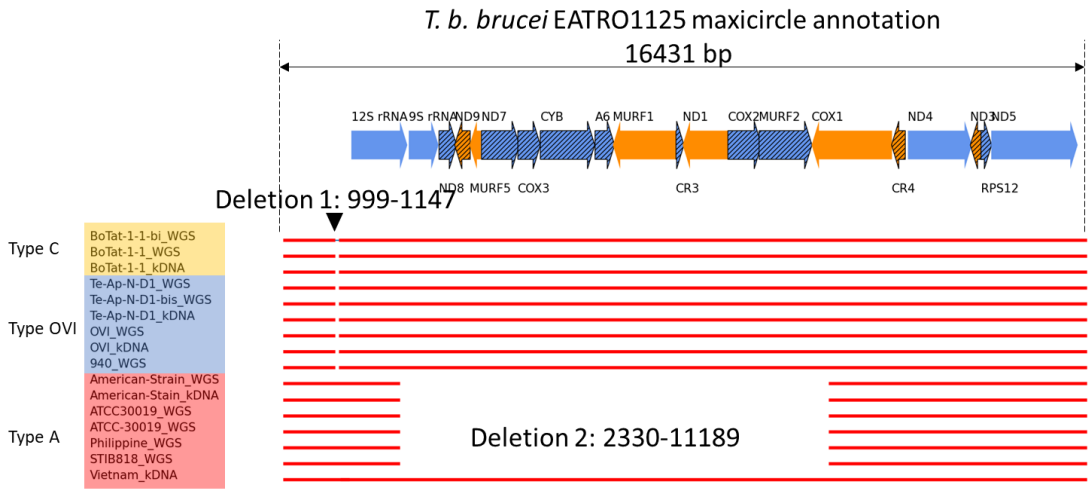


Figure 5-1**. Maxicircle alignments of *T. b. equiperdum* isolates and the maxicircle-containing type A isolates against EATRO1125 maxicircle reference.**

The alignment reveals deletions in the conserved region of maxicircle sequences from all isolates except the Vietnam isolate. The 148-bp deletion in type C and OVI *T. b. equiperdum* does not corrupt maxicircle encoded genes. The extended 8.9-kb deletion in most type A isolates results in the truncation or deletion of 15 maxicircle-encoded genes (https://www.ncbi.nlm.nih.gov/nuccore/MK584625.1 [225]).

We recovered maxicircle contigs from three *T. b. equiperdum* type C isolates and six type OVI isolates. Maxicircle alignments showed that the 148-bp deletion occurred over the sequence corresponding to nt 999 to nt 1147 on the reference EATRO1125. The deletion was located in the conserved region before the start of the 12S rRNA gene at 1364 nt and did not interrupt gene functions.

It is commonly accepted that *T. b. evansi* lacks maxicircles. However, we assembled partial or complete maxicircles from seven type A isolates[1], including five *T. b. equiperdum* and two *T. b. evansi*[2] (Figure *5-1*). The assemblies were confirmed by obtaining complete coverage when reads were mapped back to them. Seven isolates, including STIB818, exhibited an 8859 nt deletion. The deletion truncated the 12S rRNA and COX1 genes and deleted the 13 genes in between, namely 9S rRNA, ND8, ND9, MURF5, ND7, COX3, CYb, A6, ND2, CR3, ND1, COX2, and MURF2 genes. Since the loss of maxicircle genes essential for BSF and procyclic parasites is lethal for kDNA-dependent parasites, the observed gaps agree with the expectation that *T. b. evansi* isolates are kDNA independent. The deletion had previously been described for STIB818 [94], an isolate historically classified as *T. b. equiperdum* due to the presence of maxicircle but grouped with type A *T. b. evansi* [183]. However, unexpectedly, the kDNA-enriched samples of a *T. b. evansi* type A isolate from a water buffalo in Vietnam [351] contained nearly complete maxicircle. The presence of maxicircles in *T. b. evansi* and the grouping of *T. b. equiperdum* type A within *T. b. evansi* type A challenged the conventional classification.

[1] WGS: 6, kDNA-enriched: 2

[2] Isolate Philippines (WGS) and Vietnam (kDNA) 351. Gillingwater, K., P. Buscher, and R. Brun, *Establishment of a panel of reference Trypanosoma evansi and Trypanosoma equiperdum strains for drug screening.* Vet Parasitol, 2007. **148**(2): p. 114-21.

5.2.2. Copy numbers of maxicircles per network

We mapped the reads of the eleven maxicircle-containing WGS samples to the references consisting of the nuclear genome and the maxicircle from *de novo* assembly of each isolate to estimate the copy number of maxicircle per network by the ratio between their average read depth and the average read depth of the nuclear genome, assuming diploidy.

The eleven isolates had between 9 and 23 maxicircles per network, lower than previous estimates of 20-50 maxicircles per network in *T. b. brucei* (Table *5-2*) [195, 340]. In a recent analysis of sub-Saharan *T. brucei* using WGS data and the same method for calculating copy numbers per network, the maxicircle copy numbers ranged between 1 and 25 per network, with an average of 17, closer to the values that we observed in our isolates [80].

Table 5-2**. Isolates with maxicircle copy number > 1 from the assembly using the WGS dataset**

	taxon	maxicircle	Year of isolation	Country	Host	Reference
Dodola 940	*T. b. equiperdum* OVI	19.72	2008	Ethiopia	Horse	[352]
OVI	*T. b. equiperdum* OVI	14.40	1975	South Africa	Horse	[90]
Te-Ap-N-D1	*T. b. equiperdum* OVI	11.00	1991	Venezuela	Horse	[186]
Te-Ap-N-D1-bis	*T. b. equiperdum* OVI	9.04	1991	Venezuela	Horse	[186]
BoTat-1-1	*T. b. equiperdum* C	10.54	1924	Morocco	Horse	[351]
BoTat-1-1-bis	*T. b. equiperdum* C	11.00	1924	Morocco	Horse	[351]
American-Isolate	*T. b. equiperdum* A	19.75	Unknown	America?	Horse	[351]
ATCC-30019	*T. b. equiperdum* A	15.83	1903?	France	Horse	[353]
ATCC30019	*T. b. equiperdum* A	13.15	1903?	France	Horse	[353]
Philippines	*T. b. evansi* A	13.05	1996?	Philippines	Water bufflo	[351]
STIB818	*T. b. equiperdum* A	23.12	1979	China	Horse	[354]

?: records with historical uncertainty

5.2.3. Edited mRNA prediction for type OVI

Due to higher minicircle diversity observed in type OVI (to be discussed later), it became of interest to predict the editing capacity of this isolate. As a crucial first step in this process, we annotated the type OVI maxicircles by alignment with the annotated EATRO1125 maxicircle [225] and extracted the unedited mRNA sequences (Table *5-3*). OVI transcriptome data being unavailable, the edited mRNAs of type OVI were predicted by aligning non-U residues of type OVI unedited mRNAs and edited EATRO1125 mRNAs. The differences were corrected as described in the preliminary gRNA prediction of *T. b. gambiense* type 1 (Chapter 3) and *T. congolense* (Chapter 4).

The edited mRNAs predicted for type OVI had the same ORFs as the corresponding EATRO1125 mRNAs, i.e. the predicted protein products were identical. The unedited and edited mRNAs were highly conserved between type OVI and EATRO1125 (Table *5-4*). When comparing the edited sequences by blastn, we noticed that all gaps occurred in the 3' and 5' UTRs, as expected. Notably, any differences in the number of uridines in unedited COX3 were predicted to be corrected during editing, resulting in identical fully-edited COX3 in type OVI and EATRO1125.

Based on alternatively edited A6 and ND8 sequences from EATRO1125 [225] we predicted equivalent sequences for type OVI by correcting sequences based on alignments of the non-U residues. Nevertheless, no evidence indicated that type OVI edited these mRNAs alternatively. Without transcriptome data, we could not identify potential alternative editing patterns specific to type OVI. The unedited and edited type OVI mRNAs are deposited on Figshare (https://doi.org/10.6084/m9.figshare.27063367).

Table 5-3. **Comparison of type OVI and EATRO1125 un- and never edited mRNAs by Blastn**

mRNA	SID	length	mismatches	gaps
12S	98.3	1146	16	3
9S	98.7	612	6	2
A6	97.5	406	5	4
COX1	97	1649	50	0
COX2	98.1	682	12	1
COX3	99.4	464	0	3
CR3	98.2	165	2	1
CR4	96.8	284	7	2
CYB	98	1117	22	0
MURF1	98.1	1344	26	0
MURF2	98.1	1086	20	1
MURF5	97.3	225	6	0
ND1	98	959	19	0
ND3	94.5	271	11	4
ND4	97.3	1313	34	1
ND5	98.4	1771	29	0
ND7	98.3	785	4	7
ND8	98.1	362	4	3
ND9	95.9	320	10	3
RPS12	99.1	222	1	1

Table 5-4. **Comparison of type OVI and EATRO1125 edited mRNAs by Blastn**

	SID	length	mismatches	gaps
A6_v1	99.3	818	5	1
A6_v2	99.3	819	5	1
COX2	98.1	678	12	1
COX3	100	969	0	0
CR3	98.7	299	4	0
CR4	99.6	567	2	0
CYB	98.1	1151	22	0
MURF2	98.1	1107	20	1
ND3	97.2	466	11	2
ND7	99	1244	10	2
ND8_v1	99.1	576	4	1
ND8_v2	99.1	574	4	1
ND9	98.5	645	10	0
RPS12	99.4	324	2	0

5.3. Minicircle assembly and general features

5.3.1. Completeness of Assembly

We performed *de novo* minicircle assembly for the 43 isolates with KOMICS [66]. For the 34 isolates with WGS reads, for quality assessment and subsequent determination of minicircle copy numbers, we mapped the reads back to a reference consisting of the reference *T. brucei* nuclear genome and the maxicircles and minicircles pooled from all isolates. For each isolate, around 80% of WGS reads were mapped to the nuclear genome and around 3% WGS reads were mapped to the kDNA (Figure *5-2*). Our assemblies exhibited percentage of mapped CSB-3-containing reads (PMC) over 98% in 28/36 isolates and PMC over 90% in 30/36 isolates, which highly supported the near-completeness of the kDNA assemblies. The remaining six isolates had PMC ≤ 60%. We believe the six isolates were akinetoplastic, i.e. without a kinetoplast. This will be discussed in the next section.

From the 11 kDNA-enriched samples, around 65% of reads were mapped to the kDNA, and all isolates had over 98% mapped CSB-3 containing reads, which confirmed that the assemblies were (nearly) complete (Figure *5-2* B).

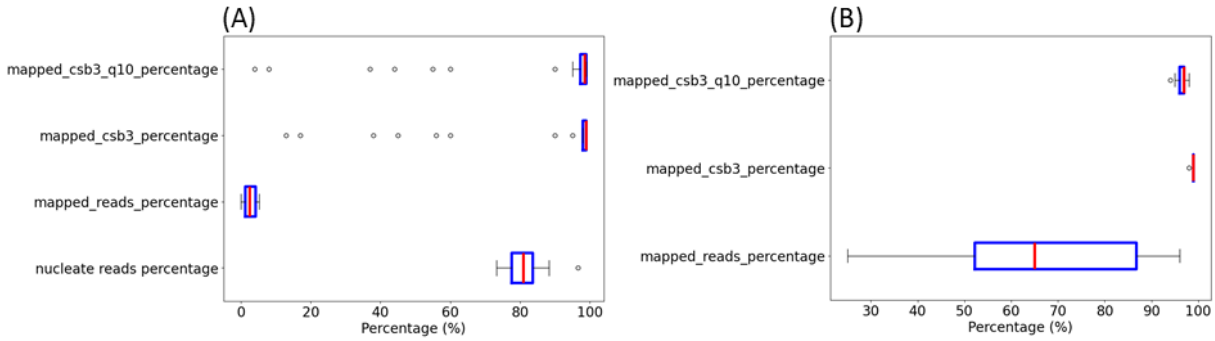


Figure 5-2. **Assessment of assembly completeness for the WGS data (A) and the kDNA-enriched data (B)**

(A) Percentage of reads mapped to the nuclear genome: mean: 80.41, median: 80.6, mode: 81.2, max: 96.5, min: 67.9; mapped reads percentage: mean: 2.88, median: 2.6, mode: 0.0, max: 10.4, min: 0.0; mapped_csb3_percentage: mean: 88.11, median: 99.0, mode: 99.0, max: 99.0, min: 13.0; mapped_csb3_q10_percentage: mean: 87.29, median: 98.0, mode: 99.0, max: 99.0, min: 4.0. **(B)** Mapped reads percentage: mean: 64.9, median: 65.0, mode: 66.0, max: 96.0, min: 25.0, mapped_csb3_percentage: mean: 98.9, median: 99.0, mode: 99.0, max: 99.0, min: 98.0; mapped_csb3_q10_percentage: mean: 96.5, median: 97.0, mode: 97.0, max: 98.0, min: 94. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point

5.3.2. Total numbers of minicircles per network

From the mapping of the 34 WGS samples, we estimated the copy number of each minicircle class per network by the ratio between their average read depth and the average read depth of the nuclear genome, assuming diploidy. Variations in copy numbers will be discussed in section 5.3.7. For each isolate, the total number of minicircles per network was then determined as the sum of the copy numbers for all minicircle classes found in that isolate. Except for type OVI, other groups of *T. b. evansi* and *T. b. equiperdum* had

homogeneous minicircle populations, if present, so the total number of minicircles per network equals the number of type A, B, or C minicircle.

The total number of minicircles per network ranged between 622 and 3808 in *T. b. equiperdum* and *T. b. evansi* isolates, lower than estimates for *T. b. brucei* based on restriction mapping (5-10k) or WGS data (6-9.5k) (Figure 5-3) [195, 225, 340, 355]. *T. b. equiperdum* and *T. b. evansi* may have smaller networks compared to other *T. brucei* subspecies. However, a recent assembly using WGS data has estimated minicircle copy numbers between 252 and 4828 per network with an average of 2100 [80]. In our analysis of sub-Saharan *T. brucei* isolates, we estimated between 603 and 16054 minicircles per network (Chapter 3). As WGS data become available for more isolates, we might realize that the network sizes are more variable and sometimes could be smaller than commonly believed.

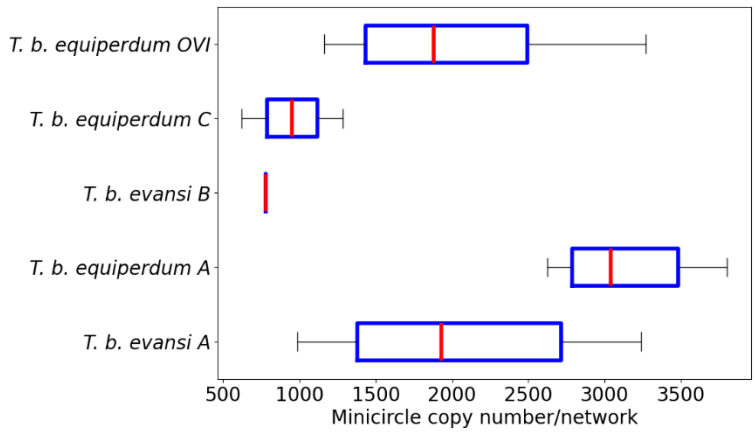


Figure 5-3. **Total minicircle numbers per network in non-akinetoplastic *T. b. equiperdum* and *T. b. evansi* isolates.**

T. b. evansi A: mean: 2051.74, median: 1929.73, mode: 987.66, max: 3239.68, min: 987.66; *T. b. equiperdum* A: mean: 3132.86, median: 3038.175, mode: 3801.75, max: 3801.75, min: 2623.83; *T. b. evansi* B: mean: 776.79, median: 776.79, mode: 776.79, max: 776.79, min: 776.79; *T. b. equiperdum* C: mean: 952.3, median: 952.30, mode: 1282.25, max: 1282.25, min: 622.34; *T. b. equiperdum* OVI: mean: 2048.42, median: 1879.84, mode: 3269.77, max: 3269.77, min: 1164.24. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point.

We estimated between 1164 and 3270 minicircles per network with a mean of 2048 in type OVI. The two type C isolates had 1282 and 622 minicircles per network, respectively. Type B had 777 minicircles per network. We noticed extremely low numbers in some type A isolates, presumably indicating akinetoplastidy. Excluding the akinetoplastic isolates, type A isolates had between 3808 and 988 minicircles per network, with an average of 2464.

No maxicircle was assembled from the putative *T. b. equiperdum* isolate E28. E28 contained only 831 reads mapped to the EATRO1125 maxicircle reference with a 97.53% coverage, while the read depth remained < 20 across the mapping, from which we estimated less than one copy of maxicircle per network. Meanwhile, E28 had a PMC of 60% and only 987 CSB-3-containing reads, compared to the mean of 110351 and median of 109707 reads in isolates with PMC ≥ 90%. The low CSB-3-containing read count indicated low minicircle abundance

within the population. We only estimated 16 minicircles per network in E28, of which 14 and 1 were type C and type A minicircles, respectively. Since the kDNA network is unlikely to exist with only 16 minicircles, we concluded that the sample of E28 contained mainly an akinetoplastic cells and a small proportion of *T. b. equiperdum* type C and unknown type A cells. We therefore did not include E28 in the subsequent analysis.

We also detected five type A isolates that had PMC ≤ 60%. The five isolates had ≤ 617 CSB-3-containing reads. The low abundance of CSB-3-containing reads led us to believe that the kDNA networks of most cells within the population did not contain minicircles, as we estimated between zero and five minicircles per network for the isolates. The five type A isolates were AnTar-7, AnTat-3-1, AnTat-3-3, ATCC-30023, and MCAM-ET-2013-MU-09. ATCC 30023 and MU-09 were known to be akinetoplastic [94, 356]. AnTat-3-1 is a variant derived from a South American *T. b. evansi* isolate from capybara [357]. Antat-3-3 is one of the first type A isolates reported [358]. Although a homogeneous minicircle population has originally been described [89, 358], some lab-cultured descendants have become akinetoplastic [35], which suggests that the cell line undergoes mitochondrial genome reduction. No information was available for AnTar-7.

5.3.3. Complexity of kDNA network in *T. b. equiperdum* and *T. b. evansi*

Type A, B, and C isolates have homogeneous minicircle populations, consisting of multiple copies of a single type A, B, or C minicircle class. However, in a group of *T. b. equiperdum* that includes three of our isolates (OVI, Te-Ap-N-D1, and Dodola 940) we detected a moderate kDNA complexity of ≥ 40 unique minicircle classes per network (Table *5-5*).

Pooling all minicircles assembled using different datasets at 95% SID resulted in 46 classes, including the type C minicircle and one minicircle closely related to type A (mO_045 in Table *5-6*). Similar to the *de novo* assembly of sub-Saharan *T. brucei*, we mapped the reads from individual isolates back to the pooled minicircles to detect minicircles not assembled probably due to lower abundance in the original assembly. Minicircles with 100% read coverage were considered present in the isolate. Using both WGS and kDNA-enriched data, we identified 44, 45, and 46 minicircle classes for OVI, Dodola 940, and Te-Ap-ND1, respectively.

Table 5-5. **Summary of unique minicircle class counts of type OVI isolates assembled from WGS and kDNA-enriched datasets**

Isolate	Year of isolation	Country	WGS	kDNA enriched	Merged (95% SID) and mapped back
Dodola 940	2008	Ethiopia	43		44
OVI	1975	South Africa	41	44	44
Te-Ap-ND1	1991	Venezuela	41,42	44	44
Total (95% SID)			45	45	46

Although type OVI isolates were isolated from different continents as many as three decades apart, the minicircle populations were highly conserved (Table *5-5*). We detected 43 minicircle classes shared by all type OVI isolates, while Dodola 940 shared one additional minicircle only with OVI. Te-Ap-ND1 and Dodola 940 each had a unique minicircle class. The sizes of the minicircles were around 1000 bp in all isolates, consistent with other *T. brucei* subspecies ([225], Chapter 3), while we detected in Dodola 940 and Te-Ap-ND1 two minicircles < 900 bp (Figure *5-4*B).

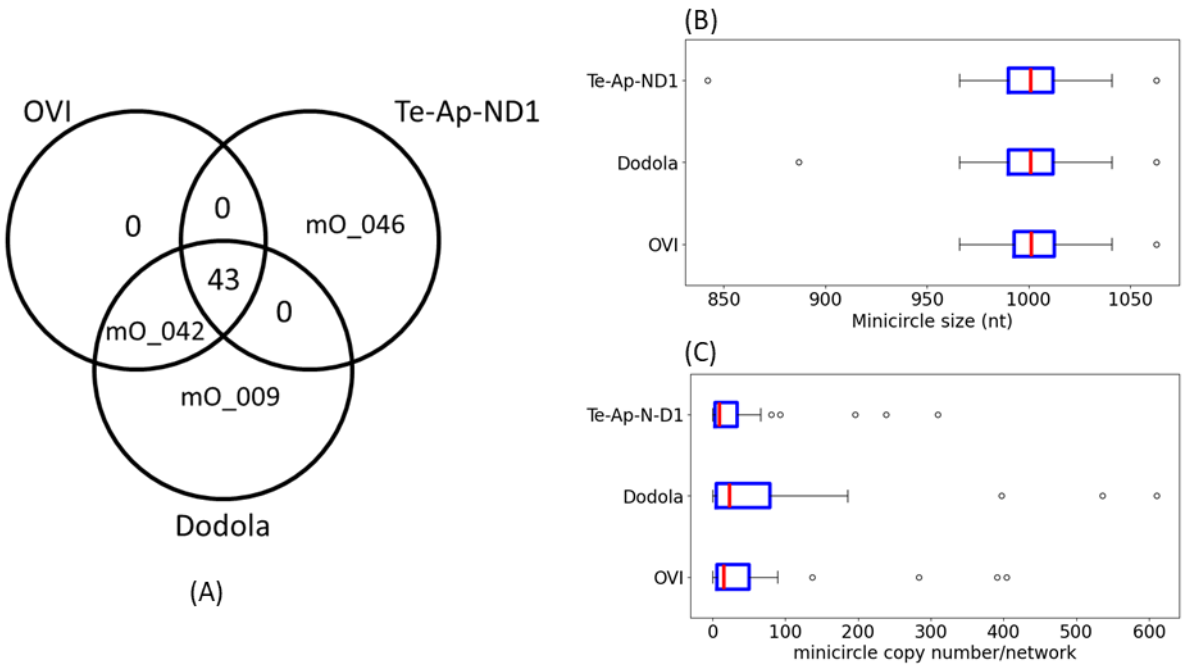


Figure 5-4. M**inicircle class compositions (A), minicircle size (B), and MCN/network (C) of *T. b. equiperdum* type OVI isolates**.

(A) The minicircle class composition was highly conserved among three type OVI isolates, which shared 43 minicircle classes. **(B)** type OVI inicircles were from 842 to 1063 bp in length, with most minicircles around 1000 bp. OVI: mean: 1003.7, median: 1001.5, mode: 1004, max: 1063, min: 966; Dodola 940: mean: 1000.13, median: 1001, mode: 1004, max: 1063, min: 887; Te-Ap-ND1: mean: 999.8, median: 1001, mode: 997, max: 1063, min: 842. **(C)** The minicircle populations are not dominated by a single class in any isolate. OVI: mean: 49.65, median: 15.38, max: 404.04, min: 0.18; Dodola 940: mean: 72.66, median: 23.56, max: 610.83, min: 0.44; Te-Ap-N-D1: mean: 33.9, median: 9.87, max: 309.98, min: 0.05; Te-Ap-N-D1: mean: 33.9, median: 9.87, max: 309.98, min: 0.05. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point

We estimated totals of 3270, 2234, 1525, and 1164 minicircles per network for Dodola (940), OVI, Te-Ap-ND1, and Te-Ap-ND1-bis, respectively. The abundance of minicircle classes within a network varied drastically (Figure *5-4*C). Dodola (940), OVI, and Te-Ap-ND1 each had three classes that accounted for ≥ 10% of the minicircle population, while 22, 19, and 21 classes, respectively, accounted for <1% of the population. The minicircles were present at between 0 and 611 copies per network, so no single class dominated the entire minicircle population, which would have suggested a strong inclination towards homogenizing minicircle populations. Minicircles present at < 1 copy per network have been reported in *T. brucei* [225]. Similarly, we reasoned that type OVI cells within a given isolate had slightly heterogeneous minicircle composition.

We compared type A, B, and C minicircles and the 46 minicircles unique to type OVI to the published 399 *T. b. brucei* EATRO1125 minicircles via blastn to identify their potential homologs in EATRO1125. The top blast hits were chosen as putative EATRO1125 homologs. The 49 minicircles had SID with their homologs between 67% and 93.6%, with a mean of 78% (Table *5-6*). Notably, four type OVI minicircles had SID ≥ 90% with the EATRO1125 homologs, and 20 minicircles, including type B and C minicircles, had SIDs ≥ 80%. For all homologs with canonical gRNAs in both EATRO1125 and type OVI, at least one if not all

cassette families were conserved. Nevertheless, we did observe different cassette families on the putative homologs.

Table 5-6 .**Blastn top hits of type OVI and type A, B, and C minicircles on published EATRO1125 minicircles (minicircles starting with mO are found in type OVI isolates)**

T. b. equiperdum T. b. evansi	**subject**	**SID**	**length**	**mismatches**	**gaps**
type_A	mO_359	68.7	1062	221	97
type_B	mO_011	83.1	1010	127	40
type_C	mO_353	81.6	1028	151	34
mO_001	mO_149	70.6	1118	218	91
mO_002	mO_257	87.2	1058	107	26
mO_003	mO_258	70.9	1107	185	114
mO_004	mO_220	79.5	1052	164	48
mO_005	mO_353	81.6	1043	137	43
mO_006	mO_215	83.5	1039	131	39
mO_007	mO_234	91.1	1028	82	10
mO_008	mO_334	81.7	1042	136	51
mO_009	mO_170	70.0	1052	226	80
mO_010	mO_318	72.6	1073	173	103
mO_011	mO_367	68.7	1074	204	112
mO_012	mO_369	73.1	1058	188	80
mO_013	mO_133	80.8	1041	140	49
mO_014	mO_360	73.1	1047	185	85
mO_015	mO_347	67.6	1054	240	94
mO_016	mO_033	92.6	1014	58	16
mO_017	mO_216	82.6	1025	138	35
mO_018	mO_238	67.3	1057	235	102
mO_019	mO_017	90.9	1025	65	22
mO_020	mO_206	78.4	1033	171	49
mO_021	mO_225	79.7	1039	155	49
mO_022	mO_199	80.5	1005	131	48
mO_023	mO_045	70.2	1055	223	84
mO_024	mO_039	83.0	1042	116	48
mO_025	mO_377	88.8	1013	90	22
mO_026	mO_209	68.4	1049	239	84
mO_027	mO_169	77.7	1021	179	45
mO_028	mO_306	91.5	1006	65	19
mO_029	mO_061	67.0	1051	244	86
mO_030	mO_130	73.0	1041	198	77
mO_031	mO_245	69.5	1043	216	90
mO_032	mO_021	67.5	1055	223	100
mO_033	mO_154	76.8	1032	179	53
mO_034	mO_108	69.9	1063	193	100
mO_035	mO_179	80.0	1018	152	47
mO_036	mO_111	86.9	1005	98	31
mO_037	mO_295	78.2	1017	172	42

mO_038	mO_245	78.8	1018	157	55
mO_039	mO_261	69.6	1052	207	102
mO_040	mO_106	87.6	999	99	24
mO_041	mO_086	70.7	1055	188	103
mO_042	mO_093	85.9	1002	109	27
mO_043	mO_345	88.7	992	95	16
mO_044	mO_104	88.3	974	95	18
mO_045	mO_359	68.6	1061	223	96
mO_046	mO_093	86.6	734	79	15

5.3.4. Variations in type A minicircles

The WGS and kDNA enriched data included 31 type A isolates (*T. b. evansi*: 21, *T. b. equiperdum*: 11), of which five were akinetoplastic. We mapped the reads from the 26 remaining samples against the published type A minicircle reference (GenBank: M57460.1 [359]) and calculated the nucleotide frequency at each position. The mapping revealed that, despite the conservation of most sequences, the isolates exhibited various SNPs against the reference sequence (Figure 5; Supplementary Figure 10).

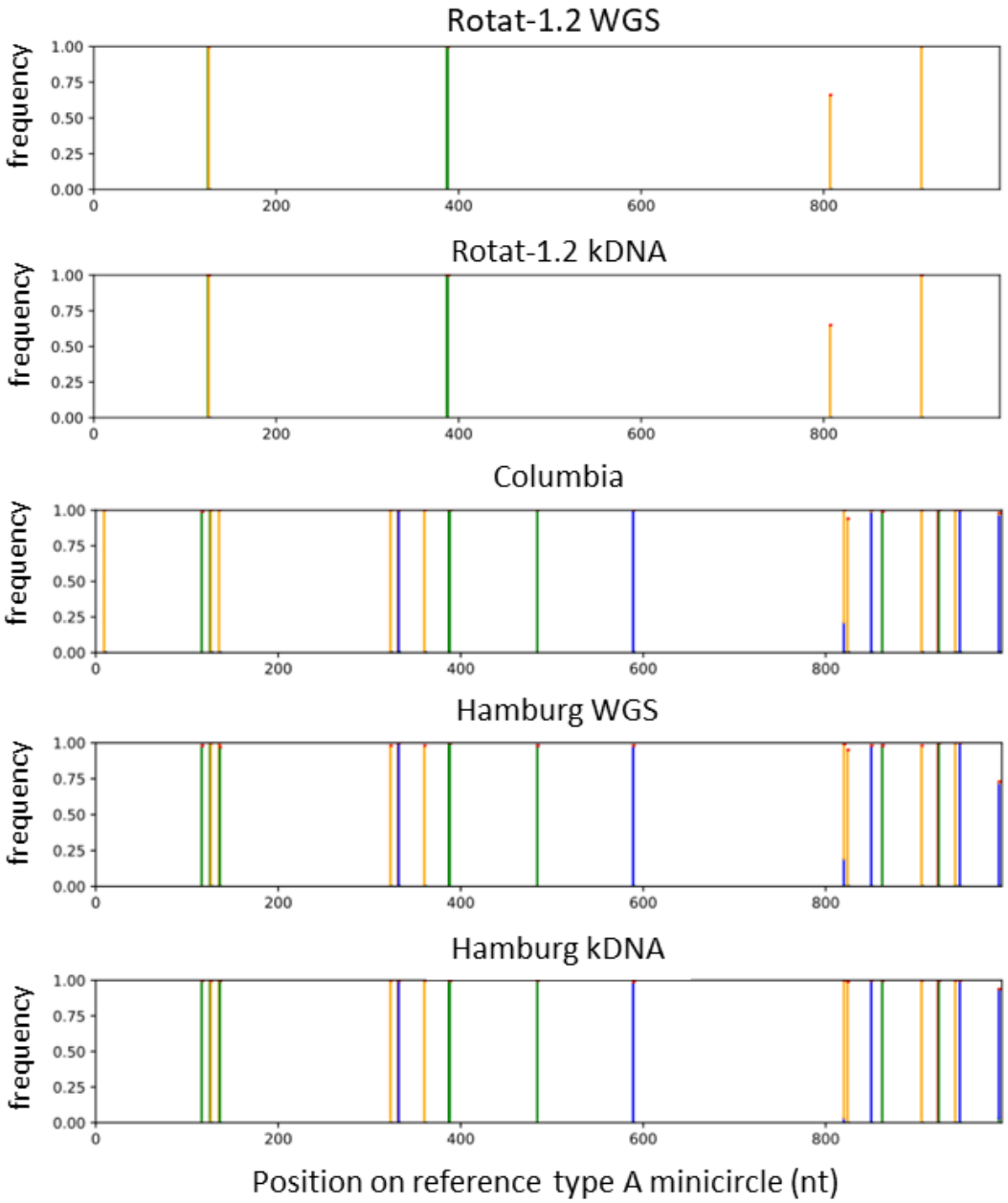


Figure 5-5. **Type A minicircle SNPs for a selection of isolates.**

SNPs for isolates RoTat-1-2, Columbia, and Hamburg (both WGS and kDNA-enriched sequencing data were available for RoTat1.2 and the Hamburg isolate) compared to GenBank M57460.1 (Ou et al., 1991) as reference are shown as examples. The mapping revealed both homogeneous and heterogeneous SNPs. The detected SNPs are shown as coloured bars: A: green; T: red, C: blue, G: orange. For heterogeneous SNPs, the frequency of the nucleotide identical to the reference is not coloured. The complete set of SNP data for type A minicircles is shown in Supplementary Figure 10.

For instance, we detected four identical, homogeneous SNPs (the first two SNPs were very close to each other) in RoTat-1-2 using WGS and kDNA-enriched sequencing (Figure *5-5*). On the other hand, isolates Columbia and Hamburg shared several identical SNPs that were different from RoTat-1-2. The mapping also revealed heterogeneous SNPs, suggesting that the minicircles within the putatively homogeneous population had slight heterogeneity or that the cells within a population had slightly different minicircles. The variations might arise from mutations in a small population of minicircles kept during replication and segregation

and randomly drifting towards higher frequency. The heterogeneous SNPs represented the mutations that have not disappeared or become fixed in the population. Nevertheless, we observed diversity of type A minicircle probably because type A samples were the most abundant in our dataset, which did not necessarily indicate that the type A group contained more variable minicircle populations compared to other *T. b. equiperdum* and *T. b. evansi* groups.

We were interested in identifying *T. brucei* isolates that contained homologs to type A minicircles. We blasted against the full collection of minicircles sub-Saharan *T. brucei* (Chapter 3) and extracted thirty homologs with >85% SID. Except for four minicircles from Zambian isolates, homologs were detected in Western and Central African isolates. The distribution of homologs agreed with the proposed Western African origin of type A *T. b. evansi* [90]. The isolates with >85% SID type A homologs included *T. b. gambiense type 2* LIGO and ABBA. Both isolates were also suggested to be closely related to type A based on analysis of genome-wide SNPs [85].

We constructed a phylogeny with iqtree using model K3Pu+F+G4 [332] to show the evolutionary relationship between the type A homologs[3]. The unrooted radial layout suggested that the minicircles formed two groups, Group 1 primarily consisting of *T. b. evansi* and *T. b. equiperdum* type A minicircles and Group 2 with only the type A homologs in tsetse-dependent *T. brucei*. We placed the root between the two groups for tree display (Figure *5-6*).

Within group 1, the homologs from the two *T. b. gambiense* type 1 (340AT, LOKO) and *T. b. rhodesiense* Etat-1-2R from Uganda were identical to the type A minicircles of three *T. b. equiperdum* isolates without maxicircle (Hamburg, Alfort, SVP). The type A homologs from five other *T. b. gambiense* type 1 and *T. b. rhodesiense* Etat-1-2R were identical to type A minicircles from Ethiopian *T. b. evansi*. The four Indonesian and Kazakhstani *T. b. evansi* isolates had identical minicircles. The seven maxicircle-containing isolates formed a basal branch in Group 1, while the Vietnam isolate with a complete maxicircle was basal to this clade. The Vietnam isolate was collected from a water buffalo and hence considered *T. b. evansi* instead of *T. b. equiperdum.* It has been proposed that *evansi*-like cell lines could emerge from *equiperdum*-like cell lines via maxicircle loss [83]. Although the hypothesis did not apply to the evolution of *T. b. equiperdum* and *T. b. evansi* as a whole given their multiple independent origins, it could explain our observations within the type A group, that the isolates without maxicircles were derived from an ancestral cell line that contained maxicircles.

[3] The minicircles of the Hamburg, American, and RoTat-1-2 isolates assembled with the WGS and kDNA-enriched data were identical. The minicircles of ATCC30019 and ATCC-30019 were identical.

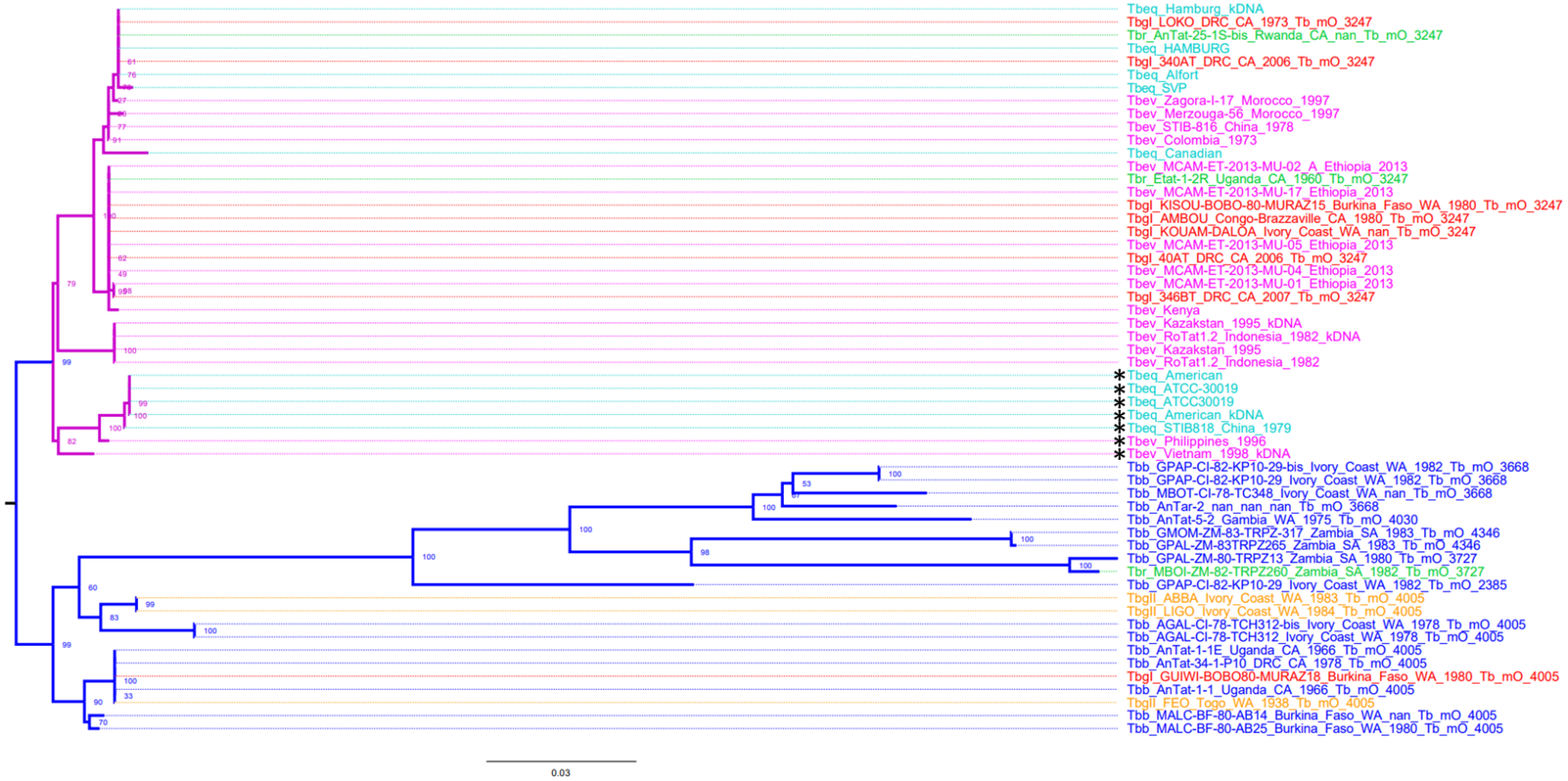


Figure 5-6. **Phylogeny based on type A minicircles in *T. b. evansi* and type A homologs in tsetse-dependent sub-Saharan *T. brucei*.**

The branch length scale represents the number of substitutions per site. Isolate prefix and colour: *T. b. gambiense* type 1 (TbgI): red; *T. b. gambiense type 2* (TbgII): orange; *T. b. brucei* (T. b. brucei): blue; *T. b. rhodesiense* (Tbr): green; *T. b. evansi* (Tbev): magenta; *T. b. equiperdum* (Tbeq): cyan . Branch colour: Group 1: magenta; Group 2: blue. Node label: bootstrap confidence. *: maxicircle containing *T. b. evansi* or *T. b. equiperdum*

5.3.5. Variations in type B minicircles

Previous studies have suggested West African origins for *T. b. evansi* type A and type B [90]. We detected in *T. b. rhodesiense* Etat-1-2R (Uganda) and *T. b. gambiense* type 1 GUIWI-BOBO80-MURAZ18 (Burkina Faso) homologs of type B with > 99% SID. The KETRI2479 minicircle we assembled had two SNPs compared to the published reference }[360]. A heterogeneous T-C SNP at 866 bp, while the A-T SNP at 726 bp was shared by all other type B isolates (Figure 5-7). Unexpectedly, we detected type B minicircle in the type A isolate MU02. Since only four type B minicircles were estimated in the kDNA network of MU02, the presence of type B minicircle was probably due to sample contamination.

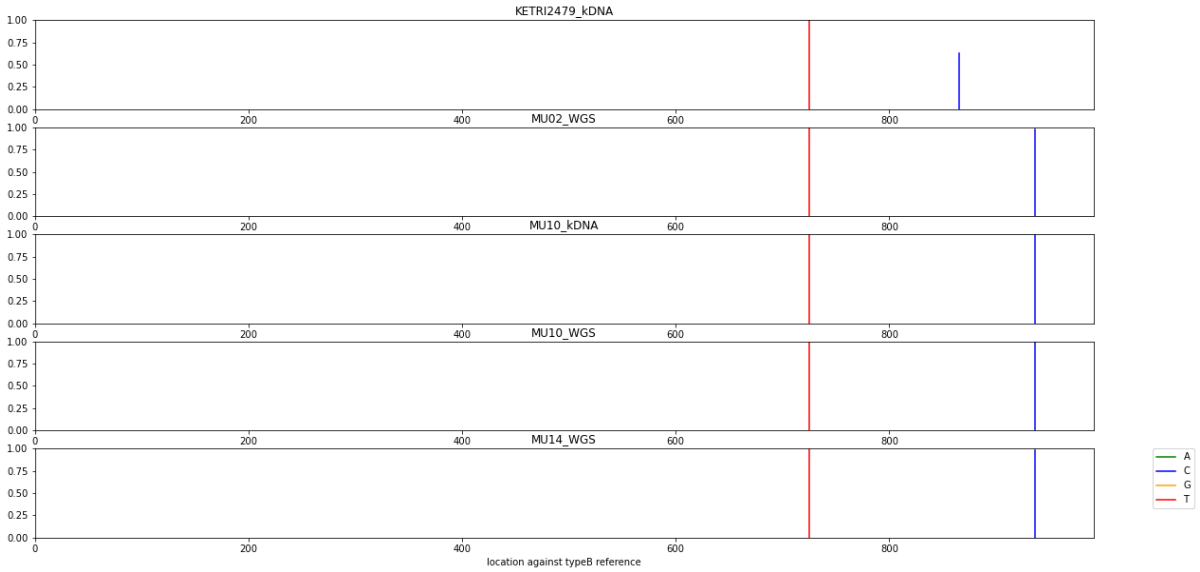


Figure 5-7. **SNP analysis of type B minicircles.**

SNP calling against type B reference (GenBank AY557604 [360]) showed little variation in minicircle sequences among type B isolates. We detected a homogeneous A-T SNP at 726 bp in all 4 isolates, and 3 isolates also share a T-C SNP at 937 bp. We detected in KETRI2479 a heterogeneous T-C SNP at 866 bp.

Compared to the plentiful records and widespread of *T. b. evansi* type A isolates, only eleven *T. b. evansi* type B isolates have been reported, all from East Africa: KETRI 2479 (Kenya 1980), MU10 (Ethiopia 2013), MU14 (Ethiopia 2013), and eight other isolates from Kenya [87, 356] [360, 361]. One Asian isolate, IVM-t1 (Mongolia 2016), was originally classified as *T. b. equiperdum* [362] but later phylogenetic analyses suggested that it is closely related to type B *evansi* [93]. Unfortunately, the isolate lost its kDNA during the isolation process and it remains unknown which type of minicircle, if any, may have been present originally [362].

Unlike type A isolates, SNP calling against the type B reference [360] shows that type B minicircle sequences of the isolates investigated here remained highly homogeneous between populations (Figure 5-7). The scarcity of *T. b. evansi* type B detection in the field suggested either insufficient sensitivity for diagnosis or less success in being epidemic. If the samples were representative of the *T. b. evansi* type B kDNA gene pool, the lack of kDNA diversity indicated that following the homogenizing of the minicircle population, *T. b. evansi* type B expansion had been undermined by a series of severe bottlenecks, which probably

explained the scarcity of its detection. Alternatively, type B may have emerged more recently. The first type B isolate (KETRI2479) is from 1980 [360], whereas type A *T. b. equiperdum* (30019 and 30023) was reported as early as 1903 [363].

5.3.6. Variations in type C minicircle and homologs

We detected two type C minicircle variants in isolate Botat-1-1. The difference arose from the 12-bp repeats (AGTGGGGAATTA) 129 bp downstream of CSB-1 and upstream of the first cassette. The published type C sequence [94](referred to as 'class B' in that study) contained six repeats, whereas the alternative sequence contained seven repeats. Given estimations based on average read depth over the repeats, the two variants coexisted at a 1:1 ratio in Botat-1-1. The type C minicircle homolog in type OVI had seven repeats at the same position, while the repeat sequence had a homogeneous A-T SNP (AGTGGGGAATTT). In contrast, the *T. b. brucei* type C minicircle homolog had six tandem repeats identical to the published sequence.

5.3.7. Correlation of minicircle copy number in type OVI

As noted above, unlike other *T. b. equiperdum* and *T. b. evansi* isolates, type OVI isolates shared a moderately complex and highly conserved population of 46 minicircle classes, of which 44 were present in all three isolates. The isolates were collected decades apart from different continents (Table 5-2).

The log values of the minicircle copy number were taken for linear correlation assessments so that the most abundant minicircles did not mask the less abundant minicircles (Figure *5-8*). Te-Ap-ND1 and Te-Ap-ND1-bis are biological replicates of the same isolate, and, as expected, the two samples had almost identical minicircle populations (linear regression, R= 1.00, P<0.0001). In the following analysis, we only considered Te-Ap-ND1 as a representative of the isolate. In addition to the correlation in log minicircle copy numbers between the biological replicates, we also observed an unexpectedly strong correlation between the minicircle copy numbers of different type OVI isolates (R >=0.88, P<0.0001, Figure *5-8*). In contrast, we detected no correlation between the copy numbers of homologous minicircle classes in *T. b. brucei* EATRO1125 [225] and any type OVI isolates (R <=0.21, P=0.26, 0.30, 0.18, Figure *5-8*). Hence, we concluded that type OVI isolates shared a highly conserved kDNA networks in terms of both the minicircle classes present and the copy numbers of each class, despite the geographical and temporal barriers.

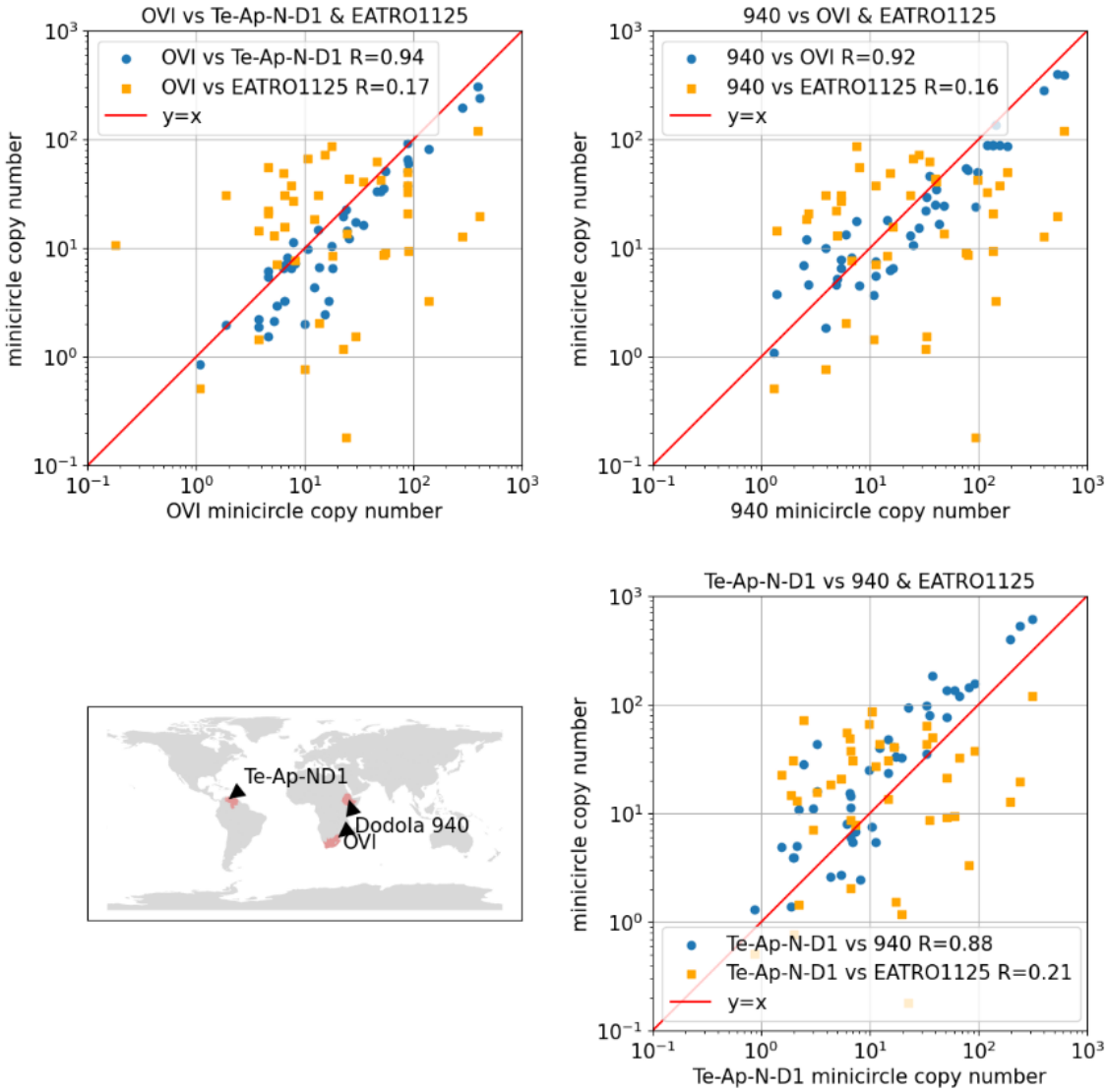


Figure 5-8. **Correlation of log minicircle copy numbers of *T. b. equiperdum* type OVI isolates from distant geographical origins decades apart.**

Relative minicircle abundances of three type OVI isolates were highly correlated (R>=0.88, P<0.0001). We did not observe a correlation in minicircle abundance when comparing type OVI minicircles to their homologs in *T. b. brucei* (R<=0.21). red line: y=x

5.3.8. Correlation of minicircle copy numbers in sub-Saharan *T. brucei*

We have observed that *T. b. gambiense* type 1 shared a highly conserved set of minicircles (see Chapter 3). However, changes in EATRO1125 BSF minicircle copy numbers had been assessed and observed after ~40 generations, suggesting that the compositions of the kDNA network change rapidly [225]. To investigate these - at first glance - contradictory observations further, we were interested in whether some tsetse-dependent isolates also exhibited a strong correlation of minicircle copy number after long divergence. From the 224 sub-Saharan *T. brucei*, we calculated the Pearson correlation coefficient (r) of the relative abundance of minicircle classes for pairs of isolates that shared over 20 classes. We calculated r < 0.85 in 10014/10047 pairs (Figure *5-9*). We concluded that they did not show evidence of correlation in minicircle copy numbers, which agreed with the observation

made with lab cultures, that the copy number of minicircles could change drastically in a cell line. Nevertheless, 33 pairs had r > 0.85, which included 14, 6, 10, and 3 pairs of *T. b. gambiense* type 1 and II, *T. b. brucei,* and *T. b. rhodesiense* respectively.

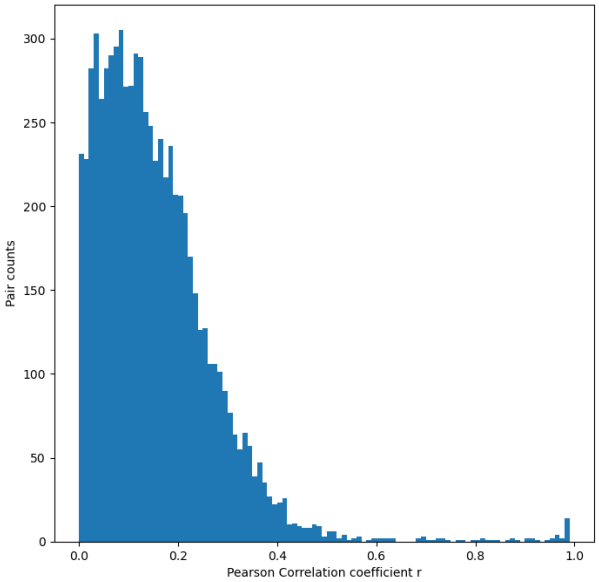


Figure 5-9. **Pearson correlation coefficient of copy number of shared minicircle classes calculated for all pairs of isolates that share ≥ 20 minicircle classes.**

10014/10047 pairs of isolates have no minicircle copy number correlation (r < 0.6). 33 pairs had r > 0.85, which indicated strong linear correlations.

Among the 33 pairs, 23 were isolated from the same year, eight contained isolates whose years of isolation were unrecorded, and two contained isolates from different years. Among the 23 pairs, nine did not have complete information on the site of isolation, 11 were isolated from the same sites, and the remaining three pairs contained isolates from different locations within geographical proximity. The two pairs of *T. b. gambiense* type 1 isolates with strong linear correlations (r=0.91, 0.99) albeit isolated one year apart were all from Mbuji-Mayi. Hence, the correlation could be explained as the isolates within each pair were most likely from the same population. We concluded that we could not detect cases in *T. brucei* where the minicircle copy numbers were conserved despite the geographical and temporal separations, which made type OVI isolates a very unique case.

.

5.4. Minicircle Annotations of type OVI isolates

The kDNA-independent dyskinetoplastic *T. b. evansi* and *T. b. equiperdum* are known to have homogeneous minicircle populations insufficient for producing any fully-edited mRNAs [87, 94, 356, 360]. The unexpected complexity of kDNA in *T. b. equiperdum* type OVI raised the question of its editing capacity, i.e. if type OVI was capable of generating any fully-edited mRNAs. We were also curious if the type OVI isolates were kDNA-dependent, which would have explained the higher kDNA complexity. To address these questions, we annotated the 46 type OVI minicircles with a bespoke pipeline [225]. The gRNA and cassette annotations and the gRNA alignments on edited mRNAs are available on Figshare (https://doi.org/10.6084/m9.figshare.27063367https://figshare.com/articles/dataset/type_OVI_gRNA_alignments/27063367).

5.4.1. Completeness of type OVI gRNA coverage

We identified 149 putative minicircle-encoded gRNA genes and two maxicircle-encoded gRNAs (gCOX2, gMURF2) with an average length of 41.5 nt (sd = 5.6) (Table 5-7, Figure *5-10*). A first pass annotation was based on the identification of matches to the predicted edited mRNA sequences (section 5.2.3). To maximize gRNA detection in *T. b. equiperdum* type OVI, we then compared EATRO1125 and type OVI minicircles that encoded similar gRNA families (see section 5.3.3). We carefully examined any non-canonical cassettes in type OVI where a canonical gRNA was detected at the same cassette position in minicircle homologs in EATRO1125. Sequence conservation with EATRO1125 gRNAs would suggest candidate canonical OVI gRNA that failed to align to the edited mRNA, for example, due to unaccounted-for editing patterns in type OVI. The aligned region of the OVI gRNA was inferred from its EATRO1125 homolog, where the candidate OVI gRNA would be aligned to examine its editing capacity. If new gRNA alignment could be obtained from U-indel modifications while preserving the ORF and protein sequence, the edited mRNAs would be modified accordingly using the gRNA as a template. The RPS12 gRNA downstream of the initiation gRNA, undetected in the first-pass annotation, was identified this way.

Table 5-7. ***T. b. equiperdum*** **type OVI gRNA coverage for mRNAs of maxicircle-encoded cryptogenes**

product	total gRNAs	unique gRNAs	total initiation gRNAs	missing gRNAs	insertions	insertions covered	deletions	deletions covered	% coverage
A6_v1	31	31	1	0	441	441	28	28	100
A6_v2	30	0	0	1	442	431	28	24	96.8
COX2	1	1	1	0	4	4	4	4	100
COX3	27	27	0	14	548	354	41	20	63.5
CR3	4	4	0	6	141	44	7	7	34.5
CR4	9	9	0	11	328	180	44	18	53.2
CYB	1	1	1	1	34	23	0	0	67.6
MURF2	1	1	0	2	25	24	5	4	93.3
ND3	5	5	0	7	208	99	13	5	47.1
ND7	11	11	0	26	549	165	88	15	28.3
ND8_v1	5	5	0	13	262	71	46	14	27.6
ND8_v2	5	1	0	13	261	81	47	14	30.8
ND9	9	9	0	12	348	171	21	6	48.0
RPS12	12	12	0	3	130	120	28	25	91.8
total	151	117	3	109	3721	2208	400	184	58.0

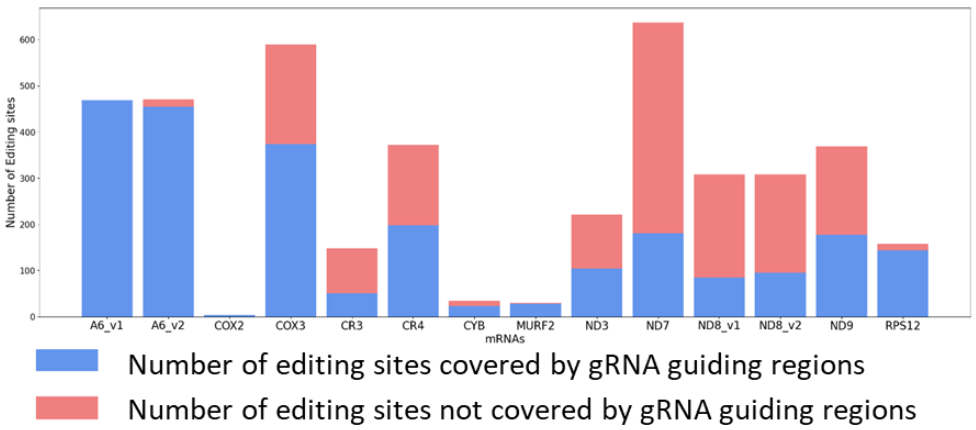


Figure 5-10. **gRNA coverage on *T. b. equiperdum* type OVI mRNAs.**

Type OVI covers 27.6% to 67.6% editing site on most cryptogenes, except for A6/RPS12. All editing sites were covered with gRNAs for A6_v1, while 91.8% editing sites were covered in RPS12.

The annotation allowed complete gRNA coverages for the edited A6_v1 mRNA and over 90% coverage for the edited RPS12 mRNA. A6 and RPS12 are the only edited mRNAs where expression is known to be essential in BSF [35, 364]. All three isolates shared most minicircles except mO_42, mO_44, and mO_46. mO_42 encoded gRNAs for ND8 and COX3 on cassettes I and IV, mO_46 encoded a COX3 gRNA on cassette III, and we did not detect any canonical gRNA genes on mO_44. Hence, the three isolates had identical sets of A6/RPS12 gRNAs and had the same gRNA coverage over their presumably edited mRNAs.

In contrast, the gRNA coverages for the other mRNAs, which are not required in BSF parasites, contained extensive gaps, ranging from 27.6% on ND8_v1 to 67.6% on minimally CYB. We only detected the initiation gRNAs for A6_v1 and CYB.

5.4.2. Assigning gRNA families in type OVI

In the kDNA analysis of sub-Saharan *T. brucei* isolates (see Chapter 3), we used the initiation site starting positions (ISSPs) to assign the identified gRNAs to gRNA families following the same principle as in chapter 3. A gRNA was considered a member of a gRNA family if its ISSP fell within the boundaries of the ISSP island corresponding to the family. We applied the same approach to the gRNA genes identified in type OVI isolates. Without adjusting the boundaries of the gRNA families, 17/149 minicircle-encoded gRNAs could not be assigned. After careful inspection of the ISSPs, we extended the boundary of each ISSP island by 2 nt on the left and 3 nt on the right to accommodate all gRNAs and identified 100 gRNA families (Table *5-8*). A6 and COX3 had the highest gRNA family counts, while all other mRNAs had less than ten families, including ND7 which had the most editing blocks in *T. brucei.* The number of gRNA families had substantially reduced compared to the 189 highly conserved families in sub-Saharan *T. brucei*, agreeing with the extensive gaps in gRNA coverage.

Table 5-8. **Comparison of minicircle-encoded gRNA and gRNA family counts on mRNAs in different isolate**

	T. b. equiperdum **type OVI**		***T. b. gambiense*** **type 1 LiTat-1-3**		***T. b. gambiense*** **type 1 Mongo**		***T. b. rhodesiense*** **Rumphi**		**Conserved editing blocks**
mRNA	**family**	**Unique gRNA**	**family**	**Unique gRNA**	**family**	**Unique gRNA**	**family**	**unique gRNA**	**Counts**
A6	26	31	28	38	27	46	28	130	27
COX3	23	27	29	43	35	76	36	227	34
CR3	3	4	3	4	8	13	8	52	8
CR4	9	9	10	14	19	29	20	78	18
CYB	1	1	1	1	2	2	2	6	2
ND3	5	5	4	4	12	18	12	38	12
ND7	9	11	23	25	40	70	43	206	39
ND8	6	6	11	14	23	40	23	94	21
ND9	9	9	11	12	21	31	22	73	19
RPS12	8	12	8	9	10	16	10	55	10
Total	99	115	128	164	197	341	204	959	190

In Chapter 3, we identified conserved gRNA families and editing blocks in sub-Saharan *T. brucei*. Here we compared the number of gRNA families in sub-Saharan *T. brucei* isolates with type OVI. The numbers of A6 and RPS12 gRNA families were close to the number of conserved editing blocks predicted with sub-Saharan *T. brucei* in all isolates examined, suggesting that the selective pressure to maintain the ability to edit both mRNAs existed in the isolates at least until recently (Table *5-8*). Similar to type OVI, *T. b. gambiense* type LiTat-1-3 had a substantially reduced kDNA network and could only produce fully-edited A6 and RPS12 mRNAs. In both type OVI and LiTat-1-3 we observed much lower gRNA family counts than the number of editing blocks on all pan-edited mRNAs except A6 and RPS12. The loss of essential gRNA families agreed with the observation of extensive gaps in their gRNA coverage. In contrast, the clonal *T. b. gambiense* type 1 Mongo had a streamlined kDNA but was presumably capable of producing all fully-edited mRNAs, while *T. b. rhodesiense* Rumphi had a highly redundant editing capacity entailed by the 408 unique minicircles. Despite the

difference in unique gRNA counts, the family counts for each mRNA in Mongo and Rumphi were slightly higher or equal to the number of conserved editing blocks on the mRNA.

5.4.3. *T. b. equiperdum/T. b. evansi* minicircle structures

Type OVI minicircles contained CSB motifs and 18-bp imperfectly conserved inverted repeats identical to the published motifs for EATRO1125 (Figure 5-11A) [225]. Most minicircles had the canonical CSB-1 motif (GGGCGTGCA), except mO_44, which had a G-T substitution on nt 7 (GGGCGTTCA). We did not detect the CBS-3 alternative reported for EATRO1125. Type A, B, and C minicircles as well as type OVI minicircles all had canonical CSB3 sequences [225]. CSB-2 was more variable, as observed in sub-Saharan *T. brucei* isolates (see Chapter 3). We detected CSB2 motifs in 42 of the 46 minicircles, with the following frequency: TCACGTGC: 31, TACCGTGC: 4, TCCCGTGC: 3, TGCCGTGC: 3, ACACGTGC: 1. We detected TCCCGTGC, the most common motif in other *T. brucei* subspecies (64.5% frequency) on only three minicircles, while TCACGTGC, the most common motif in OVI-type minicircles, was present in only 12.2% sub-Saharan *T. brucei* isolates. The change in CSB-2 motif frequencies suggested that the minicircles had undergone substantial bottleneck effects during kDNA reduction.

Table 5-9. **Summary of cassette and gRNA gene counts**

type	no canonical gRNA	one gRNA	two gRNA	total unique gRNA
Canonical cassette	0	104	4	112
Non-canonical cassette	41	0	0	0
Orphan	0	3	0	3
Maxicircle	0	2	0	2
Total unique gRNA	0	109	8	117

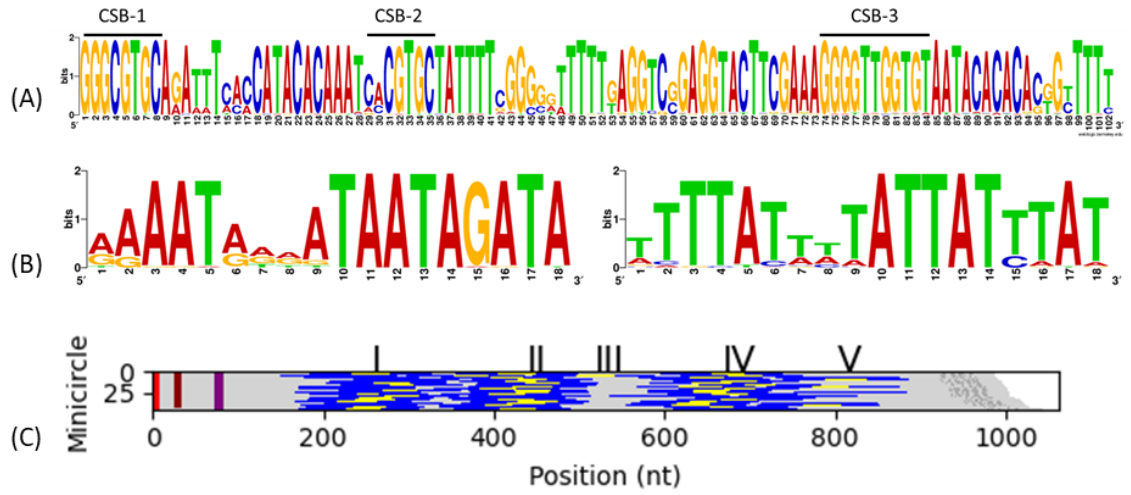


Figure 5-11. **Conserved Sequence Block (CSB) motifs of *T. b. evansi* and *T. b. equiperdum* and the structure of the 46 *T. b. equiperdum* type OVI minicircles ordered by length**

(A) the conserved region containing the origin of replication was identical to the published motif of *T. b. brucei* (EATRO1125) [225]. CSB-1 and CSB-3 were identical to EATRO1125 motifs, yet CSB-2 had a C-A substitution at the third nucleotide. (**B)** imperfectly conserved inverted repeats (left: forward, right: reverse) were identical to those described in EATRO1125 [225]. **(C)** cassette structure of type OVI minicircles. Red, brown, and purple represent conserved sequence blocks CSB1, CSB-2, and CSB-3, respectively. The regions between the 5' of the forward 18 bp inverted repeats to the 3' of the inverted repeats are shown as dark blue. Cassette-associated and orphan canonical gRNA genes are shown in yellow. The labels for the four gRNA cassette positions, I–IV, are located at each cassette's median center position. Dark gray bars shows A-tracts of the bend region.

If we did not double count the gRNAs aligned to the identical regions over the alternatively edited mRNAs, the annotation identified 117 unique gRNA genes (Table 5-7, Table *5-9*). Prior identification of gRNA cassettes based on the semi-conserved, inverted 18-bp repeats (Figure 11B) detected three, four, and two cassettes in 29, 15, and 1 minicircle(s), respectively (149 cassettes total) (Figure *5-11*C). 146 gRNA genes are located in the 108 cassettes, while the remaining 41 cassettes encoded non-canonical gRNAs. Three 'orphan gRNAs', one for CYB and two for CR4, are encoded outside of cassettes. Only four minicircles each contained a cassette with two gRNA genes. Unlike *T. brucei* or *T. congolense,* type OVI did not encode gRNAs on the anti-sense strand.

5.4.4. Respiratory complex V / F_1F_O-ATP synthase and Mitoribosome

A6 and RPS12 are necessary for BSF parasites. We therefore would expect complete gRNA coverage for A6 and RPS12 if type OVI *T. b. equiperdum* was not kDNA-independent like *T. b. evansi* type A and B or *T. b. equiperdum* type C.

Type OVI has complete gRNA coverage over A6_v2, but we did not detect an initiation gRNA corresponding to A6_v2 in EATRO1125. Most editing sites were covered by gRNAs that exhibited minimal overlaps outside the anchor region necessary for target area recognitions (Figure *5-12* A). The same trend was observed for RPS12.

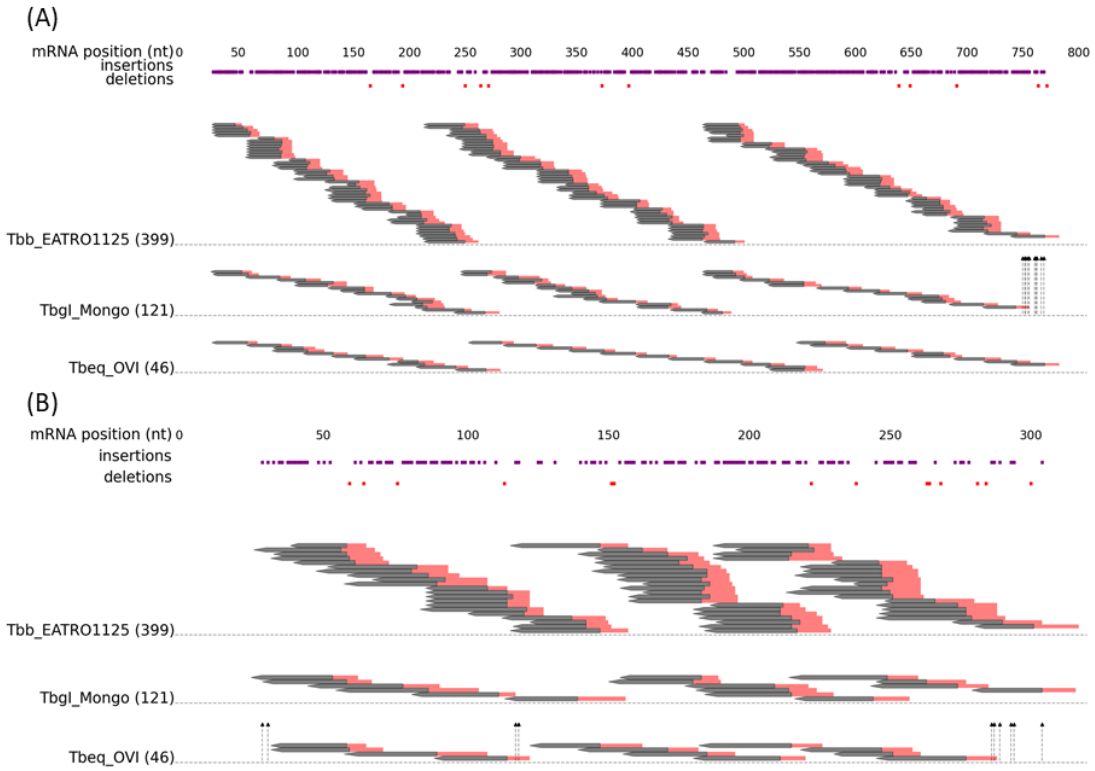


Figure 5-12.**Comparisons of A6_v1 (A) and RPS12 (B) gRNA alignments between *T. b. brucei* EATRO1125, *T. b. gambiense* type 1 Mongo, and type OVI.**

The mRNA sequence is represented on top: red: uridine deletions, purple: uridine insertions. gRNAs are aligned underneath and represented by arrows: red: anchor region, black: editing region. Dashed arrows indicate the editing sites

not covered by gRNAs. Type OVI had complete gRNA coverage over A6_v1 (panel **A**), while no initiation gRNA was identified for *T. b. gambiense* type 1 Mongo (only the A6_v2 initiation gRNA was detected in Mongo, see Chapter 3; uncovered editing sites are indicated by vertical dashed arrows). The editing pattern was more streamlined in type OVI, whose editing sites were often covered by one gRNA only, whereas in EATRO1125 multiple gRNAs encoded on different minicircle classes often aligned to the same region, hence resulting in a highly redundant editing capacity and deep gRNA coverage. **(B)** We encountered three gaps in RPS12 gRNA coverage in type OVI, including the initiation gRNA. Overall, >90% editing sites were covered in type OVI

Although we could not detect the RPS12 initiation gRNA, besides gRNAs for four additional uridine insertions, the gRNA coverage over RPS12 was >90% (Figure *5-12* B). Given the average gRNA length of ~40 nt and a minimal anchor length of 6 nt [225], we predict that at least three additional gRNAs would be necessary to bridge the gaps in RPS12 gRNA coverage (https://doi.org/10.6084/m9.figshare.27063367). The missing initiation gRNA would have been responsible for six insertions and three deletions. The two missing insertion editing sites in the 5' UTR would not affect the production of a correct ORF. Two other gaps were at nt 122 and 123, within the anchor sequence of the gRNA mO_022(II)_gRPS12(93-126).

The nearly complete gRNA coverage suggested that the RPS12 gRNAs were still maintained by selective pressure, at least until recently. If the gRNAs were truly missing, the kDNA-independence would have evolved before the geographical spread of the OVI group. However, mutations that compensate for kDNA loss have not been detected in *T. b. equiperdum* type OVI [91]. A prolonged passage in BSF can result in a significant loss of minicircle diversity and fluctuation in the copy number of the remaining minicircles within 250 generations [225]. Additionally, induced L262P mutation in F1FO–ATP synthase complex subunit γ eliminates the minicircle population in the kDNA-independent *T. b. brucei* isolate within a few generations [96]. OVI was isolated in 1975 and Dodola 940 in 2008, and the three isolates had highly conserved minicircle population. We reasoned that it was unlikely that the kDNA-independent isolates rigorously maintained the minicircle compositions for decades. Furthermore, preliminary data on drug challenges at ITM suggested that type OVI was sensitive to the kDNA-intercalating dye ethidium bromide (EtBr), which indicated kDNA dependence.

We suspect that the missing gRNAs might be encoded by low abundance minicircles and the sequencing and assembly procedures were not sensitive enough for their detection. Alternatively, not allowing any gaps in gRNA alignment might be too stringent as the exact gRNA selection mechanism *in vivo* remained unclear.

Type OVI also had a more streamlined editing capacity than EATRO1125 and even the clonal isolate *T. b. gambiense* type 1 Mongo (see Chapter 3). Most editing sites were covered by a single gRNA species, suggesting that the kDNA complexity in type OVI was approaching the minimum that allowed complete editing of A6 and, potentially, RPS12 mRNAs. This might be explained as follows. If type OVI only needed to maintain the set of gRNAs sufficient for editing two mRNAs, namely A6 and RPS12, then the reduction in linkage equilibrium between gRNA genes compared to *T. b. gambiense* type 1 would allow a more reduced kDNA network to emerge. In contrast, in *T. b. gambiense* type 1, the need to maintain a set of gRNAs sufficient for editing all mRNAs retains more redundancy, as one gRNA redundantly covered one mRNA is more likely encoded on a minicircle that also encoded non-redundant gRNAs for a different mRNA. We speculated that type OVI was not kDNA-

independent and still required A6/RPS12, besides the other maxicircle genes that are essential for BSF parasites but that do not encode RNA products that require editing, namely the other mitoribosome components uS3m, 9S rRNA and 12S rRNA.

5.4.5. A6/RPS12 gRNAs bias in type OVI

The striking difference between the gRNA coverages for A6 and RPS12 mRNAs compared to mRNAs of cryptogenes not required in BSF parasites suggested that type OVI remained kDNA-dependent at least until recently. We proceeded to explore the bias associated with A6/RPS12 gRNAs in type OVI. Among the 115 unique minicircle-encoded gRNAs, 41 (36.52%) direct editing on either A6 (26.96%) or RPS12 (9.56%). The proportion almost doubles in type OVI compared to *T. brucei*, which has 130 (17.83%) A6 and 60 (8.23%) RPS12 gRNAs, or a combined 190 gRNAs (26.06%), among the 729 (948-129-90) unique gRNAs [225]. We observed a significant enrichment of A6 (X^2 = 6.45, p=0.01) gRNAs but not RPS12 gRNAs (X^2 = 0.54, p=0.46, Figure *5-13*).

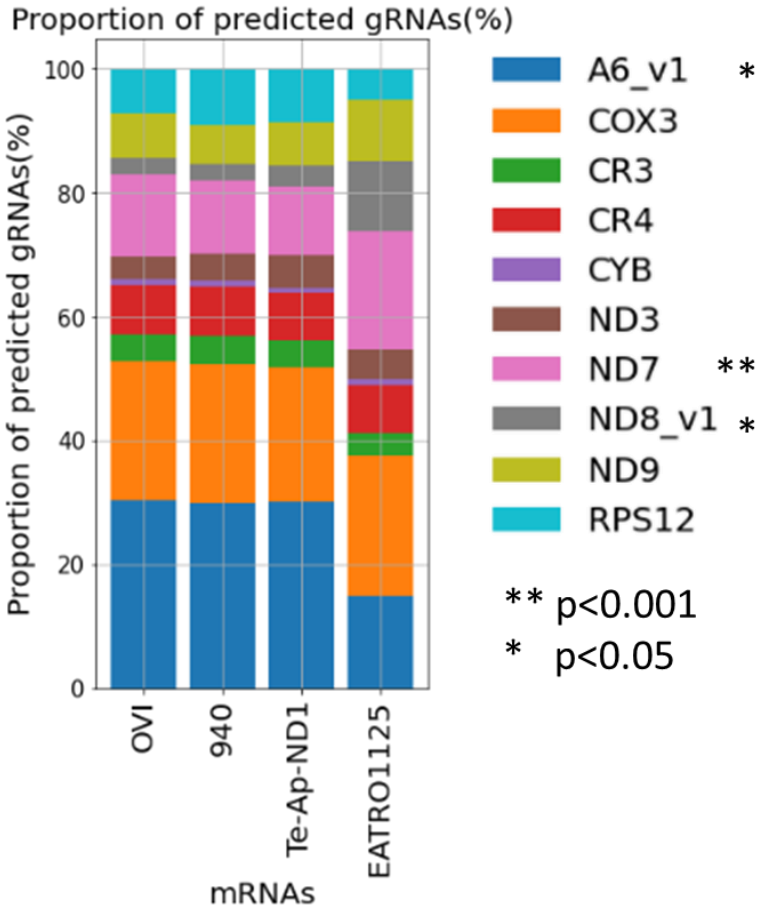


Figure 5-13. **Comparison of gRNA proportions suggest enrichment for A6/RPS12 gRNAs in OVI type.**

We compared the proportion of gRNAs for each cryptogene in *T. b. equiperdum* type OVI isolates and in EATRO1125. On average in type OVI, around 27% gRNAs edited A6, and 10% gRNAs edited RPS12. The proportion of gRNAs associated with editing of A6/RPS12 almost doubled in type OVI isolates compared to EATRO1125. Correspondingly, the proportion of ND8 and ND7 gRNAs significantly decreased.

We observed that 34 of the 46 minicircle classes encoded at least one A6/RPS12 gRNA, which accounted for 73.91% of unique type OVI minicircle classes. In *T. brucei*, only 161/399 (40.35%) unique minicircles are associated with A6/RPS12 gRNAs. Type OVI was significantly enriched in minicircles with A6/RPS12 gRNAs, compared to EATRO1125 (X^2 = 12.88, p=0.0003). Type OVI also exhibits enrichment of minicircles encoding A6 gRNAs (X^2 = 11.81, p=0.0005) but not RPS12 gRNAs alone (X^2 =0.04, p=0.83). Interestingly, 24 out of 46 minicircle classes encoded COX3 gRNAs, and 14 of these also encoded A6/RPS12 gRNAs. The linkage equilibrium may have preserved some COX3 gRNAs and the corresponding editing capacity redundancy.

We were interested in the percentage of minicircles within the network that encoded A6/RPS12 gRNA genes. In isolates Dodola 940, OVI, and Te-AP-ND1, we observed 70.1%, 70.1%, and 67.4% minicircles within the network, respectively, encoding A6/RPS12 gRNAs. We also calculated the percentage of minicircles with A6/RPS12 gRNA genes in EATRO1125 cultured in BSF for different durations from publicly available data [225]. In contrast, the percentages remained around 40% at all time points.

We were interested in whether the selective pressure to maintain A6/RPS12 gRNAs resulted in higher copy numbers of minicircles associated with those genes. However, the copy numbers of minicircles encoding at least one A6/RPS12 gRNA did not exceed those of the minicircles that contain only gRNAs for mRNAs no longer fully edited in Dodola 940, OVI, and Te-Ap-ND1 (WGS, unpaired t-test, p=0.64, 0.65, 0.50). It seemed that selection did not favour higher abundance for minicircles with gRNAs currently 'in use' but maintained all classes in the network unbiasedly. The strong correlation of the relative abundance of minicircles in type OVI isolates also indicated certain mechanisms that stabilize the minicircle populations within the kDNA network. If such a mechanism indeed exists, instead of actively regulating the abundance of individual minicircle classes, it may simply fix the relative abundance of minicircles in a viable cell line.

We also expected all cells within the population to contain all minicircles essential for A6 and RPS12 mRNA editing. The three type OVI isolates shared all the minicircles with A6/RPS12 genes. Minicircles that contained at least one A6/RPS12 gRNA gene had copy numbers greater than one per network, with a minimum of 1.39 in Dodola 940. On the contrary, some viable cells might have lost minicircles without essential gRNAs, providing evidence for the progress of kDNA reduction. Such instances were observed in all three isolates (Table *5-10*). Copy numbers per network < 1 indicated that they were absent from at least some cells of the population, and would be predicted to be lost eventually given bottlenecks and other factors for population fluctuations.

Table 5-10. **Minicircles with average copy number < 1 per network in each isolate**

Isolate	Minicircle	Copy number
Dodola 940	mO_001	0.44
OVI	mO_010	0.18
Te-Ap-ND1	mO_046	0.86

We were interested in whether the slackened selection for a subset of mRNAs may have impacted the quality of the corresponding gRNAs, manifested in gRNA length, anchor length, and number of mismatches. A6/RPS12 gRNA complementary regions had average lengths of 43.26 nt (sd=3.79) and 40.55 nt (sd=5.18 nt), respectively, and a collective mean length of 42.85 nt (sd=4.11 nt). This was over three nt longer than the remaining gRNAs (39.68 nt, sd=6.10 nt). Thus, A6/RPS12 gRNAs were significantly longer than other gRNAs in type OVI (unpaired t-test, $p<0.001$), probably because mutations on the non-essential gRNAs were not selected against (Figure *5-14*A). The mutations may interrupt the complementarity to the edited region and result in truncation of gRNAs. Hence, we also expected the non-essential gRNAs to have shorter anchors given a slackened selection against G-U wobble base pairing or mismatches in the anchor regions of non-essential gRNAs. Indeed, A6/RPS12 gRNAs had significantly longer anchors (mean=12.44 nt, sd=2.75 nt) than other gRNAs (mean=11.14 nt, sd=2.60 nt, unpaired t-test, $p=0.003$, Figure *5-14*B). We also detected significantly more mismatches in other gRNAs (unpaired t-test, $p=0.007$), in line with the slackened selection.

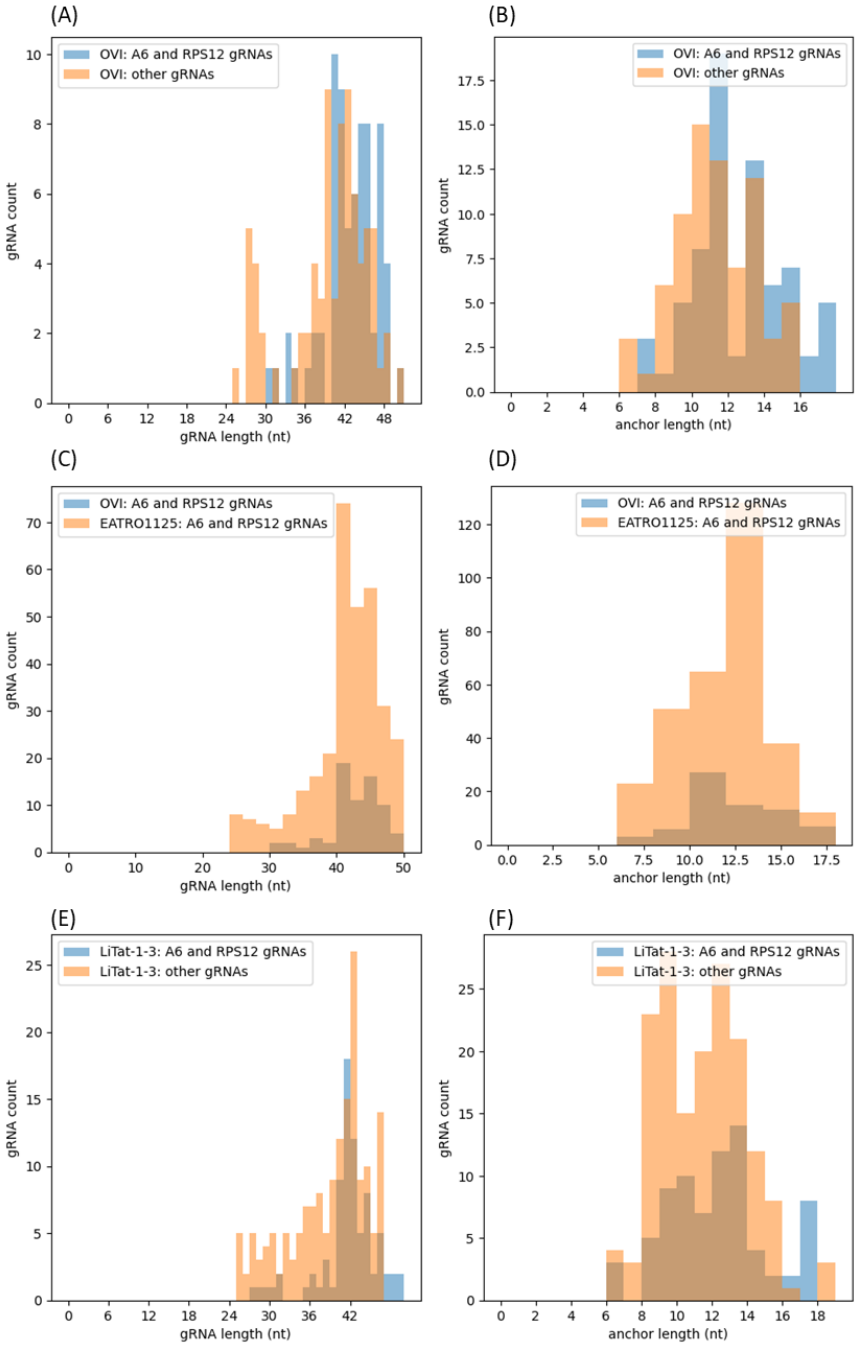


Figure 5-14. **Comparison of the quality of A6/RPS12 gRNAs with other gRNAs in type OVI (A,B), between type OVI and EATRO1125 (C, D), and in LiTat-1-3 (E,F).**

(A) length distributions of type OVI gRNAs. A6/RPS12 gRNAs (blue histogram, mean=42.61 nt, sd=4.4 nt) are significantly longer than other gRNAs (orange histogram, mean=39.38 nt, sd=6.22 nt, p=0.0003). **(B)** anchor length distributions of type OVI gRNAs. A6/RPS12 gRNAs (blue histogram, mean=12.37 nt, sd=2.80 nt) have significantly longer anchors than other gRNAs (orange histogram, mean=11.22 nt, sd=2.60 nt, p=0.003). **(C)** A6/RPS12 gRNAs in type OVI (blue histogram, mean=42.61 nt, sd=4.4 nt) are significantly longer than A6/RPS12 gRNAs in EATRO1125 (orange histogram, mean=40.92 nt, sd=5.50 nt, p=0.005). **(D)** A6/RPS12 gRNAs in type OVI (blue histogram, mean=12.37 nt, sd=2.80 nt) have significantly longer anchors compared to their EATRO1125 counterparts (orange histogram, mean=11.54 nt, sd=2.58 nt, p=0.015). **(E)**: length distributions of LiTat-1-3 gRNAs. A6/RPS12 gRNAs (blue histogram, mean=41.26 nt, sd=4.53 nt) are significantly longer than other gRNAs (orange histogram, mean=38.46 nt, sd=5.85 nt, p=0.001). **(F)** anchor length distributions of LiTat-1-3 gRNAs. A6/RPS12 gRNAs (blue histogram, mean=12.18 nt, sd=3.22 nt) have significantly longer anchors than other gRNAs (orange histogram, mean=11.07 nt, sd=2.70 nt, p=0.004). For all comparisons, the unpaired t-test was used.

We repeated the analysis for EATRO1125 gRNAs. A6/RPS12 gRNAs (mean=40.92 nt, sd=5.50 nt) were slightly longer than other gRNAs, and the difference was statistically significant (mean=40.04 nt, sd=5.32 nt, p=0.013). A probable explanation was that A6/RPS12 gRNA genes were under selective pressure constantly, while the selection on other gRNA genes may have slackened in BSF parasites. Nevertheless, the size difference was less than 1 nt, so we were unsure if it bore any biological relevance. In addition, neither the anchor length nor the number of mismatches differed between the two groups (p_{anchor} =0.265, $p_{mismatch}$=0.201). The quality of A6/RPS12 gRNAs was probably not superior to other gRNAs in an isolate with a requirement for complete editing capacity.

The higher quality of A6/RPS12 gRNAs in type OVI could result from a selection for A6/RPS12 gRNAs, from the degradation of other gRNAs, or both. To address this question, we compared the gRNAs of type OVI and EATRO1125. Type OVI had longer A6/RPS12 gRNAs than EATRO1125 (unpaired t-test, p=0.005), while the lengths of other gRNAs did not differ significantly (p=0.565). A6/RPS12 gRNAs in type OVI also had longer anchors than in EATRO1125 ($mean_{EATRO1125}$=11.54 nt, sd=2.58 nt, p=0.008) and fewer mismatches (p=0.022), while their other gRNAs do not differ significantly in anchor length (p=0.488) or mismatch count (p=0.670) (Figure *5-14*C, D). The observation thus supported a more stringent selection for A6/RPS12 gRNAs.

We speculated that editing capacity redundancy tolerated lower-quality A6/RPS12 gRNAs in EATRO1125. Nevertheless, whether there is a bias in gRNA use in EATRO1125 correlated with gRNA quality remains unknown. We suspected that the lower-quality gRNAs did not serve as functional equivalences of the high-quality ones so could be lost without undermining the editing capacity. Similar to the non-A6/RPS12 gRNAs, they were not selected during kDNA streamlining in type OVI, while the functional/active high-quality A6/RPS12 gRNAs must remain for the parasite's viability. If this hypothesis is true, the editing capacity redundancy in *T. brucei* at the kDNA level might not be as pronounced at the transcriptomic level.

5.4.6. A6/RPS12 gRNA bias in *T. b. gambiense* type 1 LiTat-1-3

We were interested in whether the A6/RPS12 bias existed in other isolates with highly reduced kDNA complexity and bias towards BSF-specific gene expression. We repeated the analysis for A6/RPS12 gRNA bias using LiTat-1-3. A6 gRNAs were enriched in LiTat-1-3 compared to EATRO1125 (X^2 = 6.31, p=0.01), but not RPS12 gRNAs (X^2 = 0.79, p=0.37). Expectedly, minicircles encoding A6 gRNAs were enriched in LiTat-1-3 (X^2 = 14.70, p=0.0001) but minicircles encoding RPS12 gRNAs were not (X^2 = 0.04, p=0.83).

However, the copy numbers of minicircles with A6/RPS12 gRNAs (mean=29) were significantly lower than the others (mean=64) in LiTat-1-3 (unpaired t-test, p=0.002), while we calculated copy number >1 for all minicircles. A possible explanation was that while all minicircles with A6 and RPS12 gRNA genes must be maintained at a sufficient level, some minicircles without those could be lost while others drift to high abundance.

Nevertheless, A6/RPS12 gRNAs also had higher quality than others in LiTat-1-3. The A6/RPS12 gRNAs were longer (p<0.001) and had longer anchors (p=0.002) than other

gRNAs, although the number of mismatches did not differ significantly (p=0.406) (Figure *5-14*E,F). The gRNA length, anchor length, and mismatch count for A6/RPS12 gRNAs did not differ between type OVI and LiTat-1-3 (p_{gRNA_length}=0.097, p_{anchor} =0.828, $p_{mismatch}$=0.597). The LiTat-1-3 non-A6/RPS12 gRNAs were longer than the EATRO1125 counterparts (unpaired t-test, p=0.0005), although comparisons of anchor lengths and mismatch counts showed no significant differences.

5.5. Chapter conclusions

We assembled and annotated the kDNA of 43 *T. b. evansi* and *T. b. equiperdum* isolates. Although *T. b. evansi* is thought to have no maxicircles [87, 183], we assembled partial maxicircles in six type A isolates with the same deletion that removed most of the maxicircle encoded genes, including a *T. b. evansi* [94]. In addition, the Vietnam *T. b. evansi* isolate had a complete maxicircle [351].

The assembly confirmed that *T. b. evansi* type A and B and *T. b. equiperdum* type C contained a homogeneous minicircle population in their kDNA networks [87, 88, 359, 360]. We detected minor variations in type B and C minicircles and homologs of type A minicircles in West African *T. brucei* isolates. The type A minicircle phylogeny placed the seven maxicircle-containing isolates as a basal group, indicating that the loss of maxicircles occurred after the homogenizing of the minicircle population.

T. b. equiperdum OVI presented a unique case of tsetse-independent *T. brucei* that maintained a moderate kDNA complexity. We assembled and annotated 46 minicircles from three type OVI isolates: Dodola 940, OVI, and Te-Ap-ND1. The isolates had conserved minicircle populations with highly correlated relative abundances. The reduced kDNA network did not encode sufficient gRNAs to provide complete coverage over most cryptogenes. However, type OVI could produce fully edited A6 mRNA and had nearly complete gRNA coverage on RPS12.

Hence, a key question now is if some RPS12 gRNAs are truly missing, which would imply kDNA independence, or whether these gRNA genes were missed in assembly or annotation and the RPS12 mRNA is in fact fully edited in OVI type *equiperdum*. Experiments to resolve this question are underway in a collaborators' laboratory. As a complementary approach, our collaborators at ITM is also examining type OVI's sensitivity to EtBr to test for kDNA independence. In the preliminary tests, type OVI isolates were found to be even more sensitive to EtBr than the known kDNA-dependent *T. brucei* isolates.

Our analysis also identified important parameters for optimal gRNA function. We observed that A6/RPS12 gRNAs in type OVI are characterised by longer complementary sequences and anchors and fewer mismatches with cognate mRNAs, compared to gRNAs for other edited OVI mRNAs and also compared to A6 and RPS12 gRNAs in *T. b. brucei* EATRO1125. This finding suggests that gRNAs with such characteristics are functionally superior, i.e. of higher quality. An implication of this is that, although EATRO1125 has a highly redundant editing capacity [225], the gRNAs aligning to the same region on the edited mRNA (i.e. members of the same gRNA family) may not function equally *in vivo*. In other words, although many seemingly functionally equivalent gRNA genes are transcribed [225], only a subset of high-quality gRNAs may be critical for mRNA editing. Hence, during the irreversible kDNA reduction, type OVI cells that had lost the minicircles with the high-quality gRNA genes were weeded out, resulting in a population with only the high-quality A6/RPS12 gRNAs.

6 kDNA complexity and Phylogeny

The definition of minicircle class is solely based on sequence identity (SID). We have discussed gRNA families in previous chapters and showed the conservation of gRNA families and editing blocks among subspecies and species. Some minicircle classes shared identical cassette families, which raised the question of whether the minicircles could be grouped based on the encoded gRNA genes instead of SID alone.

To emphasize the functional conservation of different minicircle classes and to explore potential evolutionary relationships, we introduced the concept of minicircle families and superfamilies. We defined a minicircle family as the collection of minicircle classes that contained the same gRNA families on all cassettes (Figure *6-1*). Superfamily is a more relaxed definition, that minicircle classes either with the same gRNA families or without canonical gRNAs on all cassettes formed a superfamily (Figure *6-1*).

Based on this definition, each minicircle class only belonged to one and only one minicircle family. However, the cassette families of a minicircle class might be a subset of different superfamilies so a minicircle class can belong to more than one superfamily (Figure *6-1*). Minicircle families and superfamilies revealed the functional similarities among groups with apparently highly distinct minicircle populations.

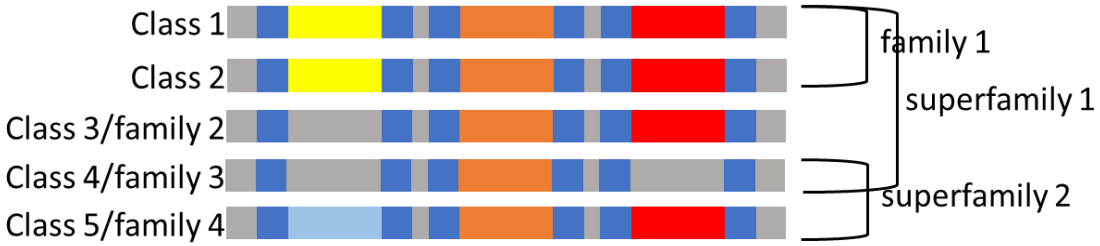


Figure 6-1. **Schematic representation of minicircle family and superfamily assignment.**

Each bar represents a unique minicircle class. Cassettes are marked with dark blue ribbons, within which coloured ribbons represent canonical gRNAs, except for grey ribbons that indicate the cassette does not contain canonical gRNA.

6.1 Minicircle family and superfamily in sub-Saharan *T. brucei*

The minicircle class compositions differed substantially among the four sub-Saharan *T. brucei* subspecies (Figure *6-2* A). Only ten classes were common to all subspecies. *T. b. gambiense* type 1 had the lowest collective minicircle diversity (193 classes), yet 132/193 classes were not found in other subspecies. In contrast, over 50% of minicircle classes in the sexual subspecies (i.e. not *T. b. gambiense* I) were present in at least one other subspecies.

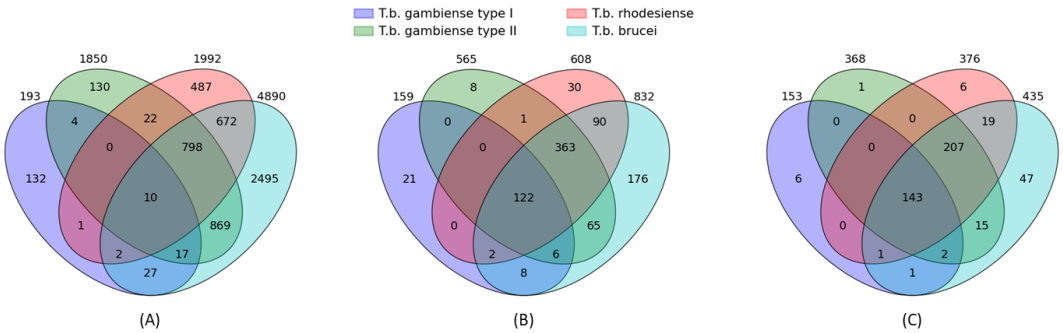


Figure 6-2**. Counts of minicircle classes (A), family (B), and superfamily (C) shared between sub-Saharan *T. brucei*.**

(A) The majority of minicircle classes were unique to *T. b. brucei* (2495). Only 10 classes were shared by all subspecies. The subspecies capable of sexual reproduction (i.e. not *T. b. gambiense* I), shared 789 classes. **(B)** The majority of minicircle families were shared by the subspecies capable of sexual reproduction (363), unique to *T. b. brucei* (176), or shared by four subspecies (122). **(C)** The majority of minicircle super-families were shared by subspecies capable of sexual reproduction (207) or shared by all four subspecies (143).

From the 224 sub-Saharan *T. b. brucei* isolates, we recovered 5666 minicircle classes and detected gRNAs on 5510. We identified 13699 minicircle-encoded canonical gRNAs and assigned them to 250 gRNA families, of which 189 highly conserved families were shared by ≥ 80% of the isolates. Cassettes in the same position that contained gRNAs from the same gRNA family were considered a cassette family.

Subsequently, we assigned the 5510 minicircle classes into 891 minicircle families. 642 (72%) minicircle families contained two or three gRNA families, while three families contained as many as five gRNA families including orphan gRNAs (Table *6-1*). We calculated the number of minicircle families in individual isolates. The subspecies capable of sexual reproduction had over 220 minicircle families per network on average, while *T. b. gambiense* type 1 isolates had only 109 families, significantly lower than other subspecies (Figure *6-3* A). In addition, the number of minicircle families varied less among *T. b. gambiense* compared to other sub-Saharan *T. brucei*, which agreed with the conserved minicircle population of the clonal subspecies (Levene test for equal variance, $p<0.0001$).

Table 6-1. **Counts of minicircle families with different numbers of cassette families**

Cassette family counts	Minicircle family counts
1	125
2	321
3	321
4	122
5	3

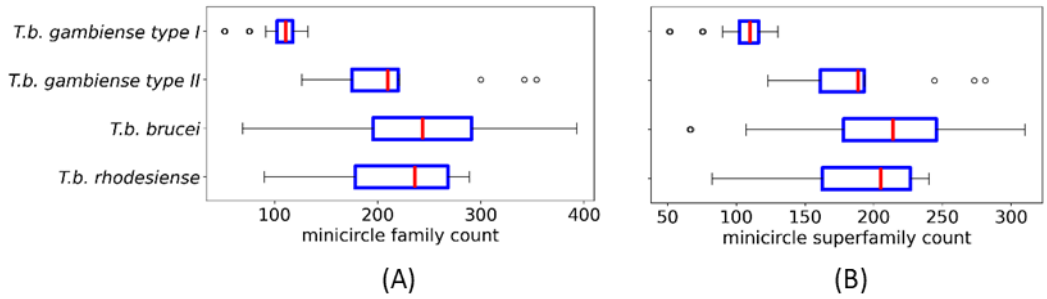


Figure 6-3 **Minicircle family (A) and superfamily (B) counts in four *T. brucei* subspecies.**

(A) T. b. gambiense type 1 has a fewer minicircle families (unpaired t-test, p<0.0001) and less intraspecific variation in kDNA complexity (Levene test for equal variance, p<0.0001) than the other three subspecies. T.b. rhodesiense: mean=224 sd=52.3 sample size=23; T.b. brucei: mean=242 sd=65.69 sample size=76; T.b. gambiense type 2: mean=220 sd=67.34 sample size=14; T.b. gambiense type 1: mean=109 sd=12.69 sample size=111; **(B)** T. b. gambiense type 1 has fewer superfamilies (unpaired t-test, p<0.0001) and less intraspecific variation in kDNA complexity (Levene test for equal variance, p<0.0001) than the other three subspecies. T.b. rhodesiense: mean=195.61 st=39.22 sample size=23; T.b. brucei: mean=209.63 st=48.04 sample size=76; T.b. gambiense type 2: mean=191.36 st=45.83 sample size=14; T.b. gambiense type 1: mean=107.57 st=12.41 sample size=111. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point

The set of minicircle families of a subspecies was obtained as the families of all the isolates combined. *T. b. gambiense* type 1, *T. b. gambiense type 2*, *T. b. rhodesiense*, and *T. b. brucei* had 159, 565, 608, and 832 minicircle families respectively (Table *6-2*). T. *b. brucei* exhibited the highest kDNA complexity and contained 832 (93.4%) of all families including 176 unique families. We identified 122 families shared by all four subspecies and 363 shared by the subspecies capable of sexual reproduction (Figure *6-2* B). Although *T. b. gambiense* type 1 had a highly conserved and unique set of minicircle classes, it shared 138/159 minicircle families with *T. b. brucei* despite the deep divergence. We explained the lack of unique minicircle families in *T. b. gambiense* type 1 by the lack of sexual recombination. Given similar mean minicircle family per isolate, the low unique minicircle family counts in *T. b. gambiense type 2* and *T. b. rhodesiense* were probably due to the smaller sample size and hybridization with *T. b. brucei*.

Table 6-2**. Counts of minicircle classes and families in sub-Saharan T. brucei**

Subspecies	Minicircle class count	Minicircle family count	Unique minicircle families
T. b. gambiense type 1	193	159	21
T. b. gambiense type 2	1850	565	8
T. b. rhodesiense	1992	608	30
T. b. brucei	4890	832	176

To account for minicircles that differed only by the presence of empty cassettes we introduce the concept of superfamily. Minicircle classes containing the same cassette families, including minicircle classes with only a subset of the cassette families, were considered a minicircle superfamily (M1, M2, M3, and M4 in Figure *6-1*). If two minicircle classes contained a different cassette family, they belonged to two minicircle superfamilies (M5 in Figure *6-1*). Some minicircles, especially minicircles with a single canonical cassette, may belong to multiple superfamilies (M4 in Figure *6-1*).

The less stringent grouping criteria derived 448 superfamilies from the 5510 minicircle classes, almost half the number of minicircle families. All superfamilies could not be treated

as a subset of other superfamilies. The average superfamily counts per network in the subspecies capable of sexual reproduction were almost twice as much as the average in the clonal *T. b. gambiense* type 1 (Figure *6-3* B). We identified 143 superfamilies shared by four subspecies and 207 shared by the subspecies capable of sexual reproduction (Figure *6-2* C). *T. b. brucei* still had the highest count of unique superfamilies.

The condensation of 5510 minicircle classes into 891 families and 448 superfamilies revealed the conservation of the functionality of minicircles within and between isolates despite the difference in sequences. Although the minicircle class compositions differed substantially among the four species, a large number of minicircle families and superfamilies were shared by sub-Saharan *T. brucei*, including *T. b. gambiense* type 1 which had a highly unique minicircle population. While 132/193 classes were unique to *T. b. gambiense* type 1, 122/159 and 143/153 families and superfamilies in *T. b. gambiense* type 1 were present in all subspecies.

.

6.2 Minicircle family and superfamily in *T. congolense*

The three *T. congolense* isolates shared 209/237 gRNA families. We were interested in the underlying similarities among the minicircle compositions of *T. congolense* isolates masked by the divergent sequences. Hence, we assigned minicircles to families and superfamilies. We would address whether the isolates had a conserved population of families and superfamilies as in sub-Saharan *T. brucei* and showed evidence of recombination.

From the three isolates, we assembled 767 minicircle classes with < 95% SID, among which 757 encoded at least one canonical gRNA. The minicircle populations differed substantially between the sampled *T. congolense* isolates (Figure *6-4* A). Only 6 classes were present in all three isolates, and only 15 to 29 classes were shared between any two isolates. The copy numbers of the shared minicircle classes also varied substantially between isolates and showed no evidence of linear correlation (Figure *6-4* D).

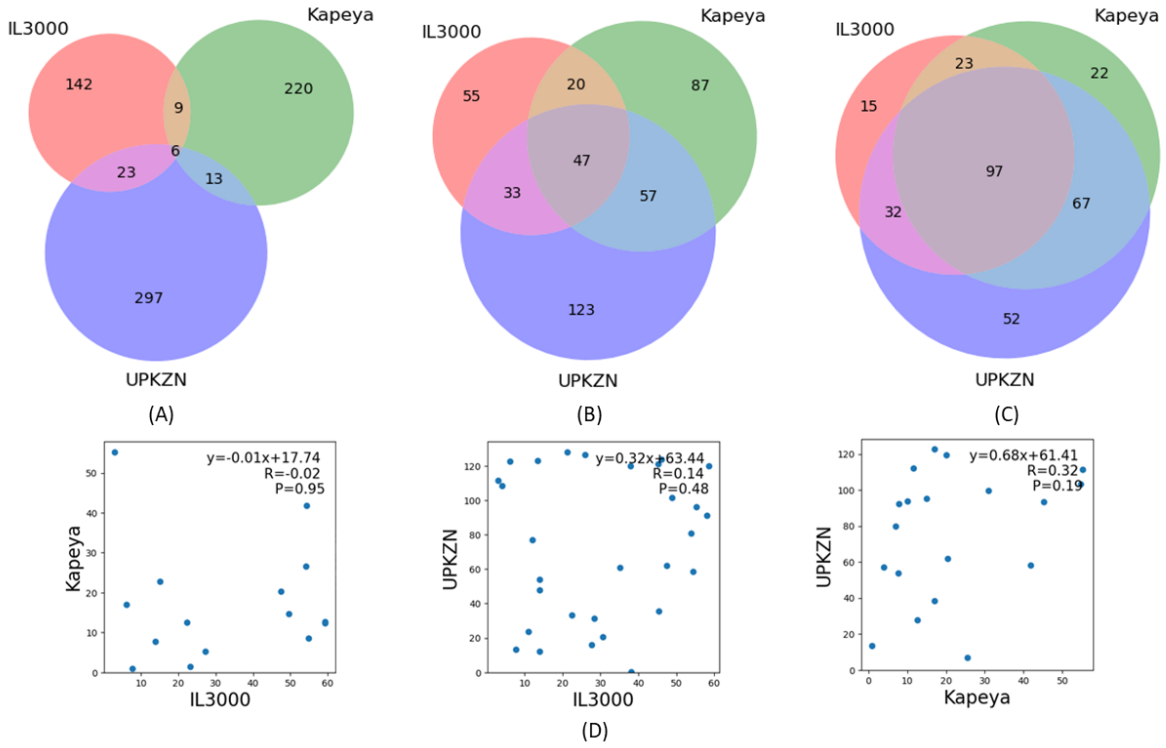


Figure 6-4. **The conservation of minicircle population between *T. congolense* isolates.**

(A) Most of the minicircle classes are unique to each isolate. **(B)** The three isolates share 47 families, and 110 families are shared by two isolates. IL3000, Kapeya, and UPKZN have 55, 87, and 123 unique families respectively, each accounting for less than 50% of the minicircle families in a given isolate. **(C) T**he three isolates share 97 superfamilies, and 122 superfamilies are shared by two isolates. **(D)** Copy number correlation of shared minicircle classes between two of the three *T. congolense* isolates. No correlation is observed between any two isolates (r: -0.02, 0.14, 0.32, p: 0.19, 0.49, 0.95)

Using the same grouping criteria, we assigned the 757 minicircle classes into 422 minicircle families that encoded unique sets of gRNA families. Most minicircle families encoded three (n=219) or two (n=154) gRNA families, with 11 encoding four gRNA families including an orphan family, and 38 encoding only one gRNA family.

IL3000, Kapeya, and UPKZN contained 155, 211, and 260 families respectively. Besides 47 families shared by all three isolates, 110 families were shared by two isolates, while each isolate contained less than 50% unique minicircle families (Figure *6-4* B). The lab-cultured IL3000 has the lowest counts of minicircle class and unique families, suggesting that some families were lost through prolonged passage in BSF due to imperfect replication and segregation.

We used the same definition of superfamilies as for *T. brucei* to derive 308 superfamilies. Minicircle classes with identical cassette families or empty cassettes belonged to the same superfamily. Most superfamilies contained one (137) or two (119) minicircle families, with 36 and 14 containing three and four families respectively. We also detected two superfamilies each including five or seven minicircle families. Most superfamilies contained one to four minicircle classes, with counts of 66, 71, 63, and 52 respectively. Most notably, the three isolates shared 97 superfamilies, while 122 superfamilies were shared by two isolates (Figure *6-4* C).

The divergence of the West and East Africa Savannah group *T. congolense* is estimated to be ~4000 years ago, while the expansion of the Savannah group *T. congolense* into East Africa occurred ~2160 years ago [92]. IL3000 has been cultured since its isolation in 1966, which may have reduced the minicircle population substantially [225]. Nevertheless, Kapeya and UPKZN were isolated more recently in 2003 and 2007, respectively. The variability of kDNA suggested that the minicircle populations were highly divergent among the East African *T. congolense* and no recent genetic exchange between the inspected lineages given the biparental inheritance of mitochondrial genome in trypanosomatids [25, 69-72, 365].

6.3 Minicircle family in *T. b. equiperdum* type OVI

We identified canonical gRNAs on 45/46 type OVI minicircles and assigned them to 45 minicircle families and 43 superfamilies. The assignments indicated that all minicircle classes encoded a different set of gRNA families, while only two could be treated as a subset of other classes. This concorded with the highly reduced kDNA that almost eliminated all redundant editing capacities. We were curious if the minicircle families were also present in tsetse-dependent *T. brucei*. To answer this question, we compared the set of minicircle families in type OVI to those of the sub-Saharan *T. brucei* isolates. We confirmed that 40/45 type OVI minicircle families were present in *T. brucei*, although five minicircle families seemed to be unique to type OVI (Table *6-3*).

Table 6-3. **Summaries of minicircles that cannot be assigned to a family in tsetse-dependent *T. brucei***

	I	II	III	IV	V
mO_042	ND8-102_104, A6-533_553	COX3-566_577			RPS12-74_86
mO_011		COX3-163_177			A6-617_626, COX3-816_830
mO_027		COX3-844_860		A6-129_144	
mO_043	ND8-368_380, ND7-500_511			COX3-118_131	
mO_047			COX3-118_131		

The type OVI minicircles had SID between 67% and 93.6% with the top blast hits in EATRO1125. Besides sequence similarities, we were interested in which *T. brucei* isolates had minicircle compositions most similar to type OVI. To our surprise, no sub-Saharan *T. brucei* contained all 40 assigned type OVI minicircle families (Figure *6-5*). T. *b. gambiense* type 1 had a highly diverged minicircle population and shared at most 13 minicircle families. The subspecies capable of sexual reproduction shared about 20 families with type OVI on average. The top three isolates with the highest counts of shared OVI families were *T. b. gambiense type 2* AnTat-25-1S (count=32), *T. b. brucei* MALC-BF-80-AB25 (count=31), and *T. b. brucei* MBOT-CI-78-TC348 (count=30). AnTat-25-1S was isolated from Rwanda in Central Africa in 1971, MALC-BF-80-AB25 was isolated from Burkina Faso in West Africa in 1981, and MBOT-CI-78-TC348 was isolated from Côte d'Ivoire in West Africa (Supplementary Table 1).

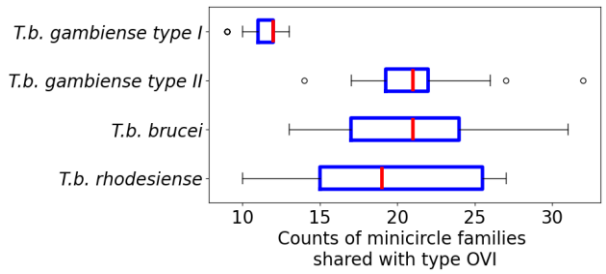


Figure 6-5. **Counts of minicircle families shared with type OVI in sub-Saharan *T. brucei*.**

T.b. rhodesiense: mean=19.87, st=5.31, max=27, sample size=23; *T.b. brucei*: mean=20.8, st=4.66, max=31, sample size=76; *T.b. gambiense* type 2: mean=21.43, st=4.48, max=32, sample size=14; *T.b. gambiense* type 1: mean=11.42, st=1.08, max=13, sample size=111. The box included the middle 50% of the data, extending from the first quartile (Q1) to the third quartile (Q3), with the red line indicating the median. Interquartile range (IQR) is the distance between Q1 and Q3. The whiskers extend from the box to the farthest data point.

6.4 Phylogeny

We were interested in the evolutionary history of Sub-Saharan *T. brucei*, in particular, in evidence of hybridizations. Hence, we investigated the *T. brucei* phylogeny using the nuclear genome, the maxicircle coding region, and the minicircle population. Phylogenies constructed from different data all supported a geographical division of sub-Saharan *T. brucei* concordant with the distribution of *T. b. gambiense* type 1 and *T. b. rhodesiense*. High-resolution phylogenies in polar layout (with bootstrap values when available) are available in Supplementary figures (DOI: 10.6084/m9.figshare.27186972).

6.4.1 Nuclear genome Phylogeny

Our collaborators' lab at ITM shared the netwick file of the sub-Saharan *T. brucei* nuclear genome phylogeny. SNPs were called against the reference nuclear genome by read mapping, with further filtering yielding 321516 homozygous variant sites. We visualised the phylogeny with figtree. The unrooted radial format showed that the isolates formed two major groups (Figure *6-6*). Group 1 mainly consisted of West and Central African isolates including all *T. b. gambiense* type 1. Group 2 primarily consisted of East African isolates including all *T. b. rhodesiense*, with a few exceptions of *T. b. brucei* and *T. b. gambiense type 2* from West or Central Africa. For display convenience, we placed the root between Group 1 and Group 2 (Figure *6-7*, for higher resolution, see Supplementary Figure 11).

The 111 *T. b. gambiense* type 1 isolates formed a monophyletic clade within the West African *T. brucei*. The deep branching indicated a substantial divergence from the subspecies capable of sexual reproduction, while little genomic variation existed among the clonal isolates.

Group 2 exhibited a mixture of *T. b. rhodesiense* and *T. b. brucei* populations. Three Central/West African *T. b. brucei* were placed in Group 2: AnTat-34-1 from DRC, Yaoundé from Cameroon, and MSUS-CI-82-TSW31-BO1 from Ivory Coast. AnTat-34-1 was clustered with *T. b. brucei* isolates AnTat-1-1 and AnTat-1-1E from Uganda, while MSUS-CI-82-TSW31-BO1 was closely related to Lister-427-AT1-KO. Five *T. b. brucei* isolates, including the kinbo-

sindo group parasites J10, and one *T. b. rhodesiense* from Zambia branched far out of Group 2, suggesting further genome diversity to be discovered in areas not frequently surveyed.

In contrast, the nuclear genomes of the West African isolates were less divergent, probably because the samples were collected from only a few locations, mostly Ivory Coast (62/81 non-gambiense-type-I isolates in Group 2), so the large sample size did not necessarily capture the genetic diversity across the continent.

While 11 *T. b. gambiense type 2* isolates were placed within the West African clade as expected, five isolates were placed within Group 2, including AnTat-25-1S, MHOM-CI-78-TH1-037, and three isolates from the patient FEO from Togo 1938 [333]. Whole genome SNP calling suggested that the FEO isolate was closely related to the isolate from Yaoundé from West Africa.

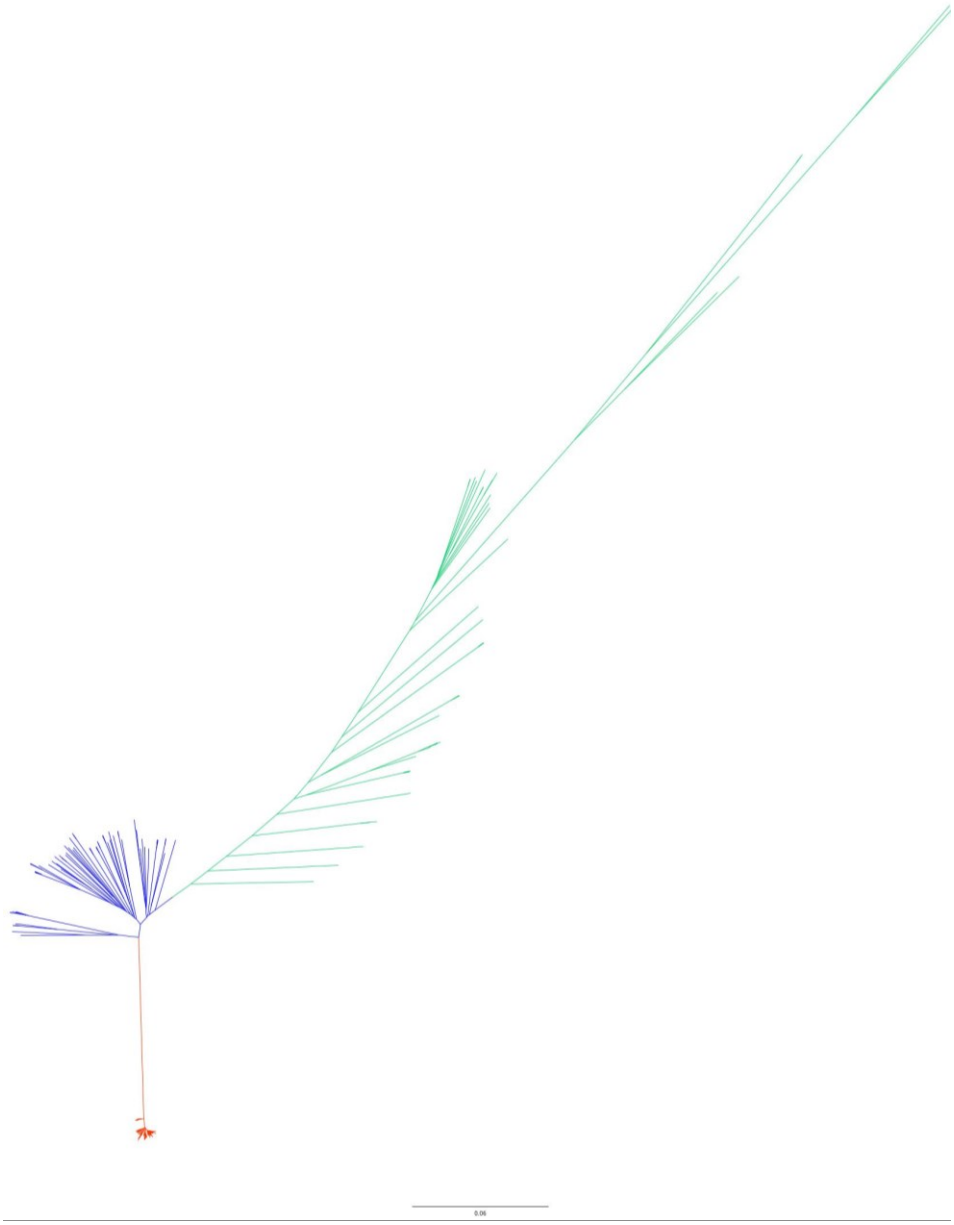


Figure 6-6. **Sub-Saharan *T. brucei* phylogeny based on whole-genome SNP calling (radial format).**

The clustering pattern reflects the division between West (Group 1) and East (Group 2) African *T. brucei*. Branch color: Group 1 non-gambiense-type-I: blue; *T. b. gambiense* type 1: red; Group 2: green. The branch length scale represents the number of substitutions per site. See Figure 6-7 for tip labels.

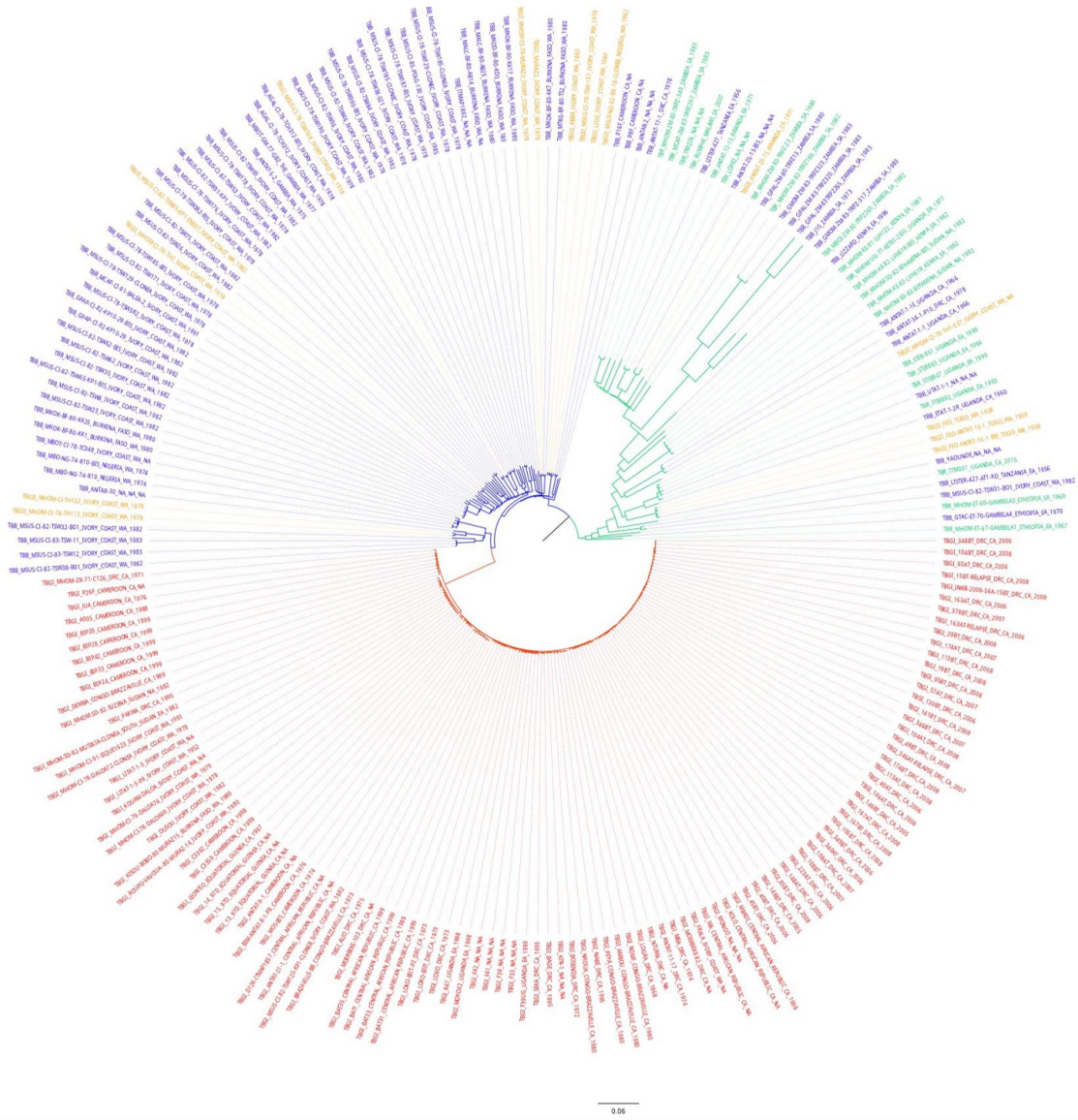


Figure 6-7. **Sub-Saharan *T. brucei* phylogeny based on whole-genome SNP calling (polar format).**

The grouping of the isolates reflected the boundary between Western and Eastern *T. brucei*. Isolate prefix and color: *T. b. gambiense* type 1 (TBGI): red; *T. b. gambiense type 2* (TBGII): orange; (T. B. BRUCEI) *T. b. brucei*: blue; (TBR) *T. b. rhodesiense*: green. Branch color: Group 1 non-gambiense-type-I: blue; *T. b. gambiense* type 1: red; Group 2: green. Geographical origin: EA: East Africa; WA: West Africa; NA: North Africa; SA: Southern Africa; CA: Central Africa. Each label is formatted as prefix_isolate_country_geographical origin_year of isolation. Missing information is indicated with 'NA'. The branch length scale represents the number of substitutions per site. The tree is rooted between Group 1 and Group 2.

6.4.2 Minicircle family Phylogeny

We constructed the minicircle family phylogeny using a discrete morphological substitution model to explore the relatedness of minicircle compositions among sub-Saharan *T. brucei*. For each isolate, we built a sequence of 891 characters each representing the existence status of a minicircle family. All minicircle families were assigned a status as 'present', 'related', or 'absent' after querying against the minicircle families present in the isolate (Figure *6-8*). We inferred the phylogeny from the discrete morphology sequences with a morphological substitution model [329-331] using iqtree [332].

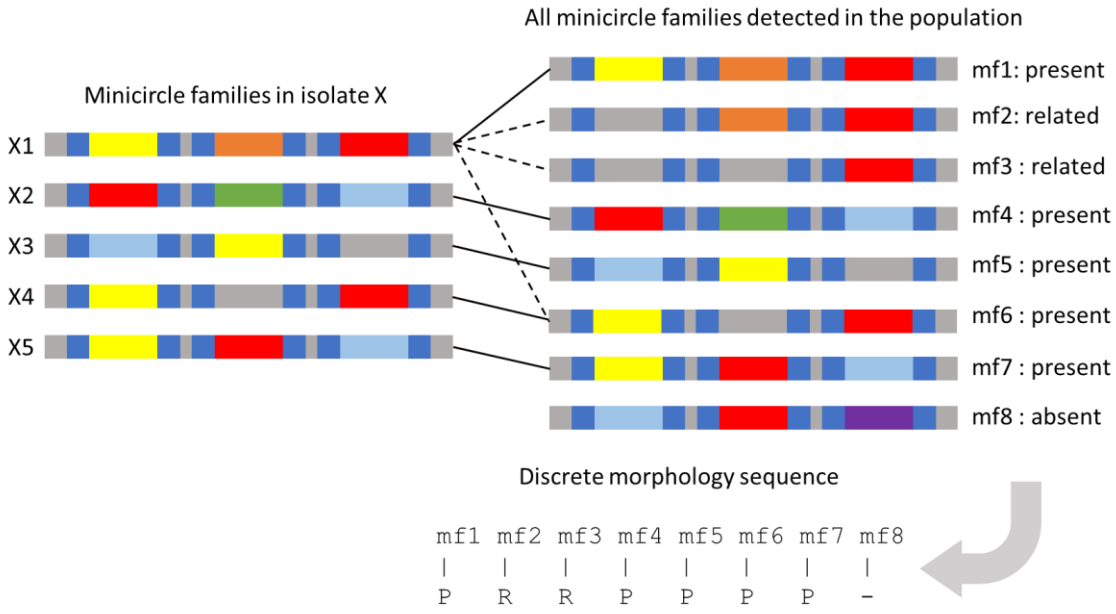


Figure 6-8: **Schematic representation of deriving the morphology sequence using minicircle families.**

The set of all minicircle families in the population is compared to the minicircle families present in a given isolate (each bar represents a minicircle family. blue: inverted repeats; colored: canonical gRNA gene; grey: non-canonical gRNA gene or empty cassettes). If the isolate contains a minicircle family with the exact cassette families (e.g. X1) as the query (mf1), the query minicircle family (mf1) is considered present. If the isolate contains a minicircle family (e.g. X1) that has identical cassette families but also encodes additional canonical gRNAs, the query (e.g. mf2, mf3) is considered related. Present overrules related (e.g. mf6). The remaining minicircle families (e.g. mf8) are absent. Hence, a morphology sequence can be translated from the existence status of minicircle families for each isolate.

The phylogeny also supported the overall division based on geographical origins (Figure *6-9*). We observed with unrooted radial format that the isolates fell into two major groups similar to the genome-wide SNP phylogeny. Group 1 mainly consisted of West and Central African *T. brucei*. Group 2 contained most of East African *T. b. rhodesiense* and *T. b. brucei*. The similar branching depth indicated that the levels of minicircle population divergence were comparable across the continent. For display, we rooted the tree between Group 1 and Group 2 (Figure *6-10*, for higher resolution, see Supplementary Figure 12).

Clonal reproduction isolated the minicircles in *T. b. gambiense* type 1, resulting in substantial divergence of the minicircle family profiles. As expected from the highly conserved and unique minicircle population, *T. b. gambiense* type 1 formed a monophyletic clade within Group 1, including the three isolates found in Uganda between 1998 and 1999. The observation confirmed that, despite colocalization with East African *T. brucei* and the human infective *T. b. rhodesiense* in Uganda, no genetic exchange occurred between *T. b. gambiense* type 1 and other subspecies.

The *T. b rhodesiense* and *T. b. brucei* isolates from the same areas also shared similar minicircle family profiles, which indicated frequent genetic exchange. Despite the opportunity for hybridization, all isolates from Uganda were located in Group 2,

suggesting a minicircle population more similar to other East African *T. brucei*. The two *T. b. rhodesiense* isolates found among West African *T. brucei* were MHOM-ET-67-GAMBELA1 and MHOM-ET-69-GAMBELA3 from Ethiopia.

We observed the same 11 *T. b. gambiense type 2* within Group 1 and the five within Group 2 as in genome-wide SNP phylogeny. The three isolates from the patient FEO had highly conserved minicircle family profiles most closely related to the Yaoundé isolate.

Most West and Central African *T. b. brucei* were found in Group 1 except MSUS-CI-82-TSW31-BO1 and AnTat-34-1-P10. The former shared a similar minicircle family profile with Lister-427-AT1-KO from Tanzania. The latter also shared a similar minicircle family profile with *T. b. brucei* isolates AnTat-1-1 and AnTat-1-1E from Uganda. Overall, the grouping of *T. b. brucei* by minicircle family profiles agreed with the East-West divisions of the parasite's distribution.

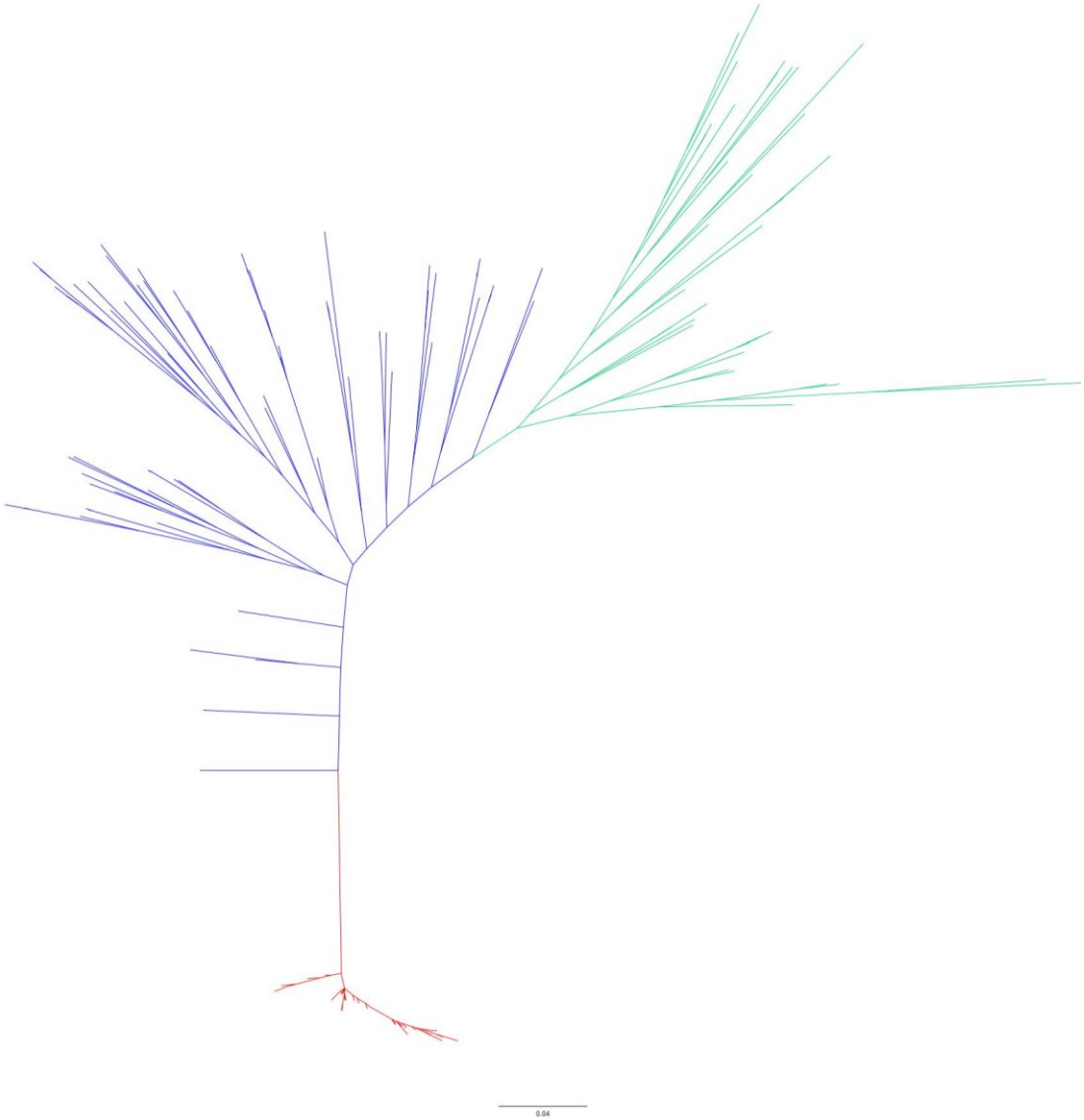


Figure 6-9. **Sub-Saharan *T. brucei* phylogeny based on minicircle family composition (radial format).**

The clustering pattern reflects the division between West (Group 1) and East (Group 2) African *T. brucei*. Branch color: Group 1 non-gambiense-type-I: blue; *T. b. gambiense* type 1: red; Group 2: green. The branch length scale represents the number of substitutions per site. See Figure 6-10 for tip labels.

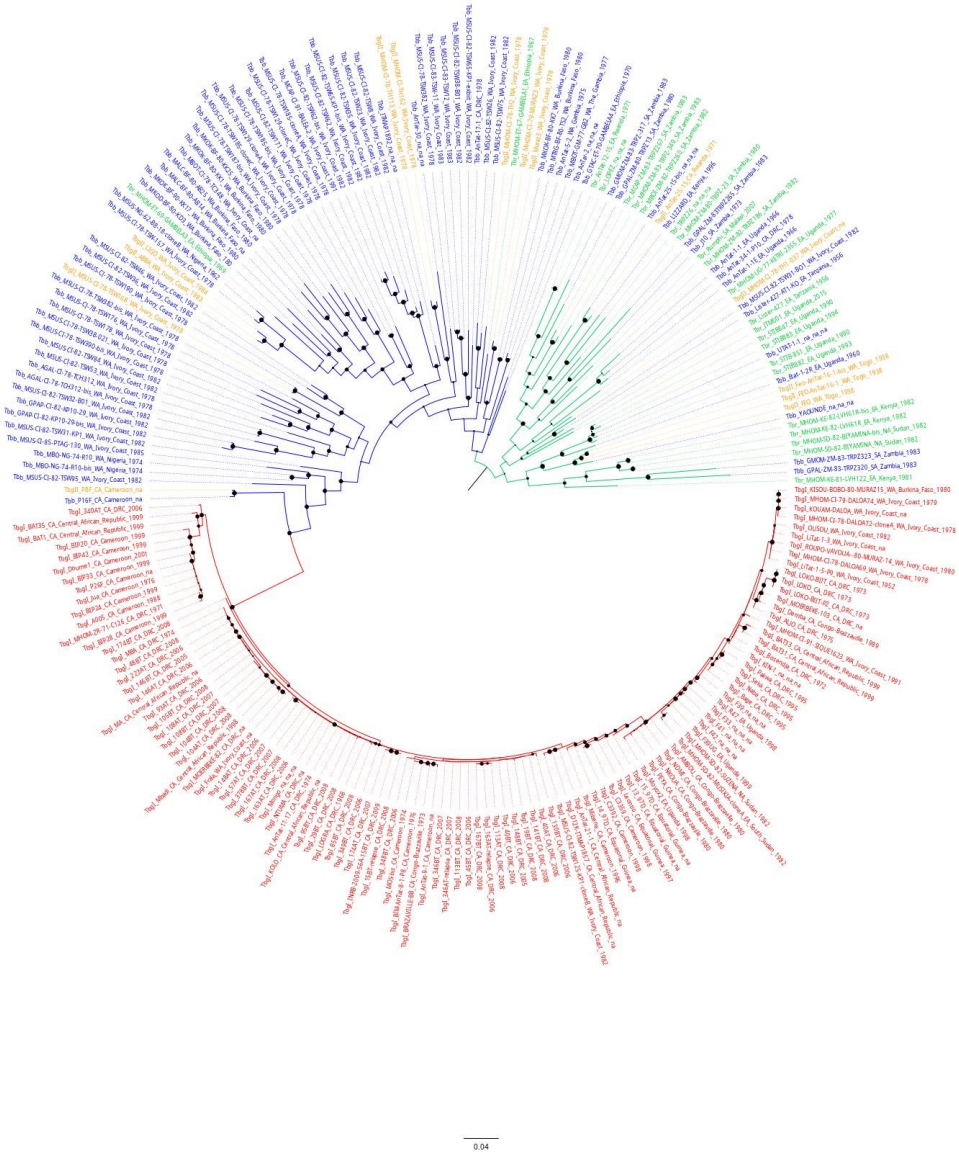


Figure 6-10. **Sub-Saharan *T. brucei* phylogeny based on minicircle family composition (polar format).**

The isolates form two major groups. Group 1 contained mostly West and Central African *T. brucei*, including the monophyletic *T. b. gambiense* type 1. Group 2 primarily consisted of East African *T. b. brucei* and *T. b. rhodesiense. T. b. gambiense type 2* is found in both groups. Bootstrap confidence is indicated by the sizes of the nodes. Isolate prefix and color: *T. b. gambiense* type 1 (TBGI): red; *T. b. gambiense type 2* (TBGII): orange; (T. B. BRUCEI) *T. b. brucei*: blue; (TBR) *T. b. rhodesiense*: green. Branch color: Group 1 non-gambiense-type-I: blue; *T. b. gambiense* type 1: red; Group 2: green. Geographical origin: EA: East Africa; WA: West Africa; NA: North Africa; SA: Southern Africa; CA: Central Africa. Each label is formatted as prefix_isolate_country_geographical origin_year of isolation. Missing information is indicated with 'NA'. The branch length scale represents the number of substitutions per site. The tree is rooted between Group 1 and Group 2.

6.4.3 Maxicircle Phylogeny

We inferred a phylogeny based on SNPs called against the coding region of EATRO1125 minicircle for the 224 sub-Saharan *T. brucei* isolates, three *T. b. equiperdum* type C, three *T. b. equiperdum* type OVI, and six *T. b. evansi* type A isolates that contained maxicircles (Figure *6-11*). The maxicircle phylogeny exhibited the same division between West African (Group 1) and East African (Group 2) isolates. Similar to genome-wide SNPs, we observed the most variation and divergence in the East African isolates. The clade that contains *T. b. equiperdum* and J10 from the kiboko-sindo group had maxicircles highly divergent from other sub-Saharan *T. brucei*. For display convenience, we placed the root between Group 1 and Group 2 (Figure *6-12*, for higher resolution, see Supplementary Figure 13).

Group 2 contained no isolates from West or Central Africa except MSUS-CI-82-TSW31-BO1 and AnTat-17-1 from DRC. *T. b. gambiense* type 1 formed a monophyletic group within West African *T. brucei* and shared highly conserved maxicircles. Group 1 also contained all *T. b. gambiense type 2* except AnTat-25-1-S and 10/12 isolates from Uganda, including four *T. b. rhodesiense* isolates, three *T. b. brucei*, and three *T. b. gambiense* type 1.

All East African *T. b. brucei* and *T. b. rhodesiense* that did not belong to Group 2 formed a monophyletic clade in Group 1 (Table 6-4). The clade also included the three *T. b. gambiense type 2* FEO isolates and the Yaoundé isolate. All isolates within this clade had identical maxicircle coding regions except the *T. b. rhodesiense* isolate MHOM-ET-67-GAMBELA1 from Ethiopia.

Table 6-4. **Summary of the isolates assigned to Group 1**

isolate	Taxon	Country of isolation	Year of isolation
Etat-1-2R	*T. b. brucei*	Uganda	1960
Yaoundé	*T. b. brucei*	Cameroon	Na
UTAT-1-1	*T. b. brucei*	Na	Na
MHOM-SD-82-BIYAMINA	*T. b. rhodesiense*	Sudan	1982
MHOM-SD-82-BIYAMINA-bis	*T. b. rhodesiense*	Sudan	1982
MHOM-KE-82-LVH61R	*T. b. rhodesiense*	Kenya	Na
MHOM-KE-82-LVH61R-bis	*T. b. rhodesiense*	Kenya	1982
MHOM-KE-81-LVH122	*T. b. rhodesiense*	Kenya	Na
STIB882	*T. b. rhodesiense*	Uganda	1993
STIB883	*T. b. rhodesiense*	Uganda	1994
STIB851	*T. b. rhodesiense*	Uganda	1990
STIB847	*T. b. rhodesiense*	Uganda	1990
MHOM-ET-67-GAMBELA1	*T. b. rhodesiense*	Ethiopia	1967
FEO	*T. b. gambiense type 2*	Togo	1938
FEO-AnTat-16-1	*T. b. gambiense type 2*	Togo	1938
Feo-AnTat-16-1-bis	*T. b. gambiense type 2*	Togo	1938

We detected complete or partial maxicircles in *T. b. equiperdum* and some *T. b. evansi* type A isolates. The *T. b. equiperdum* type C and type OVI isolates were closely related to the five Zambian isolates from Group 2. *T. b. equiperdum* type C was grouped closely with J10 from kiboko-sindo group, while the type OVI isolates occupied an adjacent branch as a sister clade.

We did not expect the maxicircle-containing type A isolates to fall within Group 2 given the West African origin of *T. b. evansi* type A. The type A isolates formed a monomorphic clade with the Vietnam isolate residing on the outermost branch. The isolate with the most closely related maxicircle to type A was *T. b. brucei* AnTat-17-1 from DRC (SID=98.33% with the Vietnam isolate). AnTat-17-1 was placed among Group 1 by genome-wide SNPs and minicircle family profile, while the clade also diverged far from other Group 2 isolates.

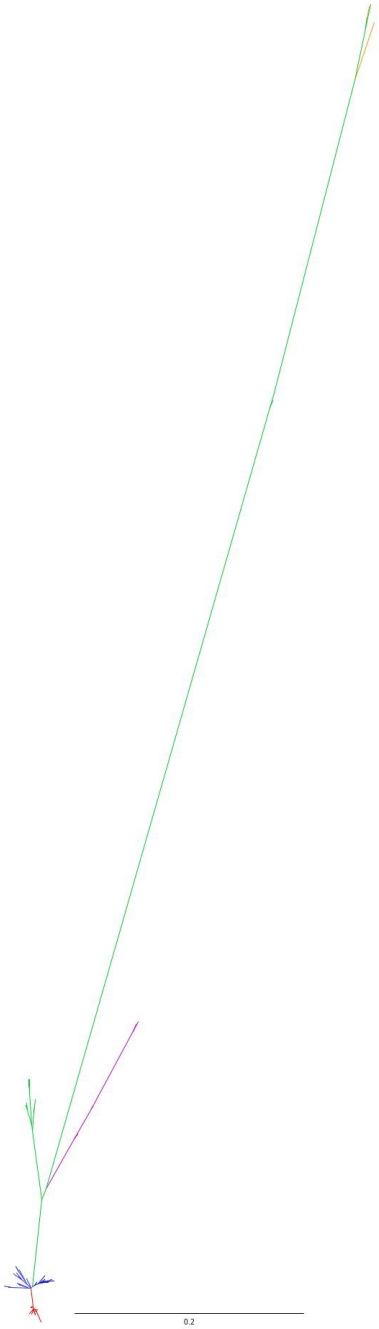


Figure 6-11. **Sub-Saharan and tsetse-independent *T. brucei* phylogeny based on SNPs called against maxicircle coding region of EATRO1125 (radial format).**

The clustering pattern reflects the division between West (Group 1) and East (Group 2) African *T. brucei*. Within Group 2, the branches that contain *T. b. evansi* type A are in magenta, and the branches that contain *T. b. equiperdum* are in orange. Branch color: Group 1 non-gambiense-type-I: blue; *T. b. gambiense* type 1: red; Group 2: green. The branch length scale represents the number of substitutions per site. See Figure 6-12 for tip labels.

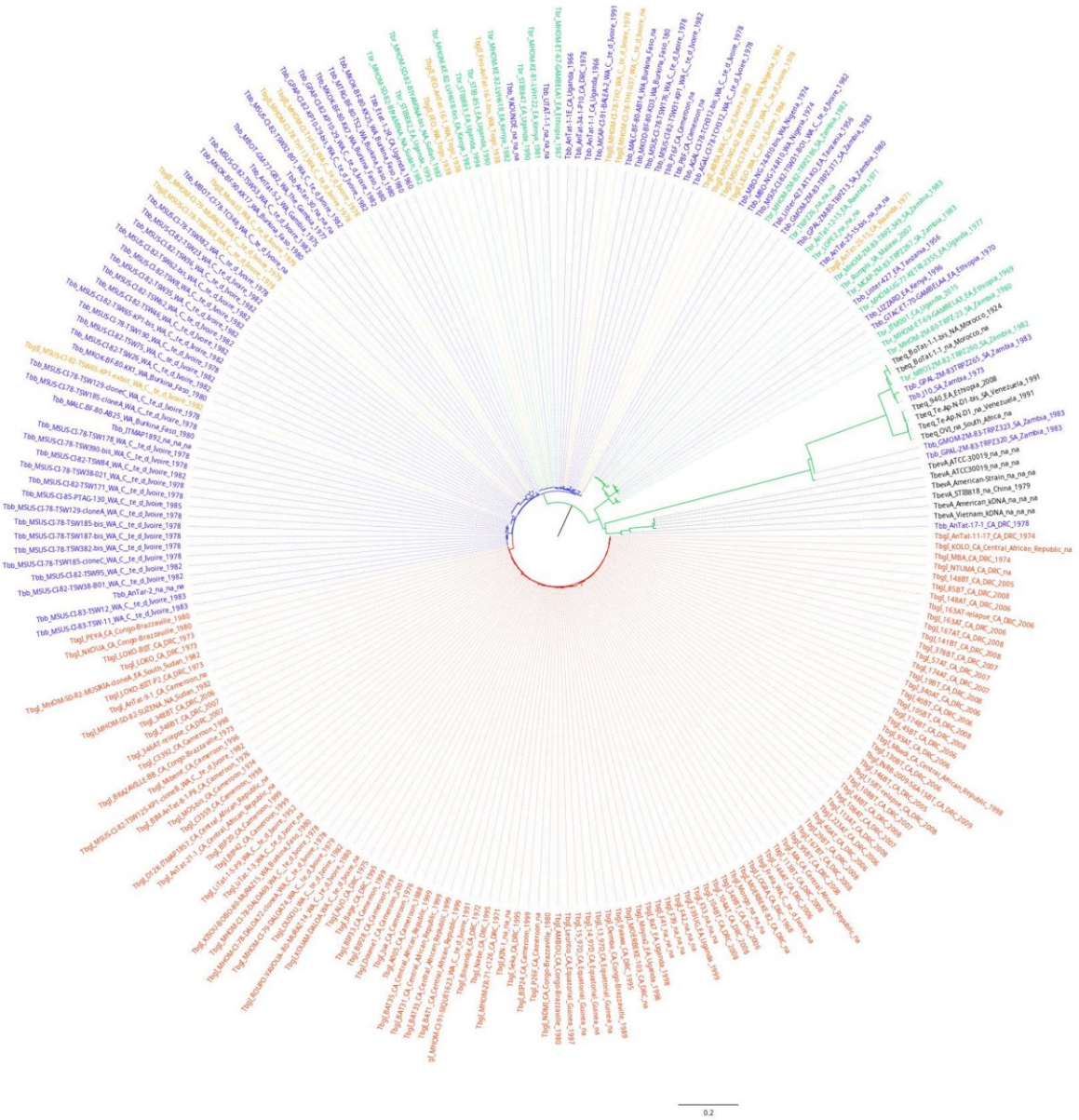


Figure 6-12. **Sub-Saharan and tsetse-independent *T. brucei* phylogeny based on SNPs called against maxicircle coding region of EATRO1125.**

The phylogeny exhibits two major groups consisting of East and West African *T. brucei*. *T. b. gambiense* type 1 forms a monophyletic clade within the West African *T. brucei*. Both *T. b. equiperdum* and *T. b. evansi* showed deep divergence from most non-kiboko-sindo *T. brucei. T. b. equiperdum* and maxicircle-containing *T. b. evansi* type A isolates were placed within the East African group. Isolate prefix and color: *T. b. gambiense* type 1 (TBGI): red; *T. b. gambiense type 2* (TBGII): orange; (T. B. BRUCEI) *T. b. brucei*: blue; (TBR) *T. b. rhodesiense*: green. Branch color: Group 1 non-gambiense-type-I: blue; *T. b. gambiense* type 1: red; Group 2: green. Geographical origin: EA: East Africa; WA: West Africa; NA: North Africa; SA: Southern Africa; CA: Central Africa. Each label is formatted as prefix_isolate_country_geographical origin_year of isolation. Missing information is indicated with 'NA'. The branch length scale represents the number of substitutions per site. The tree is rooted between Group 1 and Group 2.

6.4.4 Comparison between phylogenies using different markers

Phylogenies generated with the nuclear and the mitochondrial genomes agreed on the general patterns. *T. b. gambiense* type 1 isolates formed a monophyletic clade within the

West African group. The minicircle phylogeny placed *T. b. gambiense* type 1 isolates more dispersedly than in the nuclear and maxicircle phylogeny, suggesting that the minicircle composition was more variable than the other two parameters. This observation demonstrated the sensitivity of minicircle composition in resolving the evolutionary relationships among closely related cell lines with highly conserved nuclear genome and maxicircles.

We observed no segregation between *T. b. brucei* and *T. b. rhodesiense*, which suggested frequent genetic exchange and sexual recombination. The three phylogenies demonstrated the genetic division between the East and West African *T. brucei*, which agreed with the boundary between *T. b. gambiense* type 1 and *T. b. rhodesiense* distributions.

However, contradicting the commonly believed West African origin, four *T. b. gambiense* type 2 isolates reported in the literature: three FEO cell lines and MHOM-CI-78-TH1-037, were found among East African *T. brucei* based on nuclear and minicircle phylogeny. The maxicircle phylogeny, in contrast, grouped them among the West African isolates as expected. The apparent discrepancy probably reflected the unreliable history of type 2 classification, sample mixing, or computational error. On the other hand, we could not rule out the possibility that the assay was not sensitive enough to detect the SRA markers in AnTat-25-1S, which would explain its presence in East Africa among *T. b. rhodesiense*.

Nevertheless, we observed discrepancies between the origins of the isolates and the placements within the phylogenies explainable by the movement of infected humans or livestock and hybridization between cell lines with different genetic backgrounds (Table *6-5*). The discrepancies between the minicircle family profiles and the maxicircle sequences probably originated from the biparental inheritance of minicircles [69, 70] and the biparental inheritance of maxicircles followed by the loss of maxicircles from one parent [70, 71]. Admittedly, sample mixing-up and mislabelling may explain some of the discrepancies.

Table 6-5. **Summary of isolates with discrepancies between their origins and the placements within the phylogenies**

isolate	taxon	Country	Genome-wide SNPs	Minicircle family profile	Maxicircle coding region SNPs
AnTat-1-1	*T. b. brucei*	Uganda	East	East	West
AnTat-1-1E	*T. b. brucei*	Uganda	East	East	West
AnTat-17-1	*T. b. brucei*	DRC	West	West	East
AnTat-34-1	*T. b. brucei*	DRC	East	East	West
Etat-1-2R	*T. b. brucei*	Uganda	East	East	West*
MSUS-CI-82-TSW31-BO1	*T. b. brucei*	Ivory Coast	East	East	East
UTAT-1-1	*T. b. brucei*	na	East	East	West*
Yaoundé	*T. b. brucei*	Cameroon	East	East	West*
AnTat-25-1S	*T. b. gambiense type 2*	Rwanda	East	East	East
FEO	*T. b. gambiense type 2*	Togo	East	East	West*
FEO-AnTat-16-1	*T. b. gambiense type 2*	Togo	East	East	West*
Feo-AnTat-16-1-bis	*T. b. gambiense type 2*	Togo	East	East	West*
MHOM-CI-78-TH1-037	*T. b. gambiense type 2*	Ivory Coast	East	East	West
MHOM-ET-67-GAMBELA1	*T. b. rhodesiense*	Ethiopia	East	West	West *
MHOM-ET-69-GAMBELA3	*T. b. rhodesiense*	Ethiopia	East	West	East
MHOM-KE-82-LVH61R-bis	*T. b. rhodesiense*	Kenya	East	East	West*
MHOM-KE-81-LVH122	*T. b. rhodesiense*	Kenya	East	East	West*
MHOM-KE-82-LVH61R	*T. b. rhodesiense*	Kenya	East	East	West*
MHOM-SD-82-BIYAMINA	*T. b. rhodesiense*	Sudan	East	East	West*
MHOM-SD-82-BIYAMINA-bis	*T. b. rhodesiense*	Sudan	East	East	West*
STIB847	*T. b. rhodesiense*	Uganda	East	East	West*
STIB851	*T. b. rhodesiense*	Uganda	East	East	West*
STIB882	*T. b. rhodesiense*	Uganda	East	East	West*
STIB883	*T. b. rhodesiense*	Uganda	East	East	West*

*Note: *: isolates listed in Table* 6-4*, East African: Green, West African: Blue.*

Some isolates received unexpected group assignments based on their locations of isolation. Although these may represent interesting cases of relocation and/or hybridization, we must be aware that mislabelling, contamination, and other errors, could also explain the apparent inconsistency.

Sample mixing or relocation could explain the discrepancies in AnTat-25-1S from Rwanda and *T. b. brucei* MSUS-CI-82-TSW31-BO1 from Ivory Coast. All three phylogenies grouped the two isolates with the East African *T. brucei*. The phylogenies suggested that MSUS-CI-82-TSW31-BO1 was closely related to Lister-427-AT1-KO from Tanzania.

Genetic introgression of East and West African *T. brucei* occurs in Uganda [75]. Three out of the 12 isolates from Uganda were *T. b. gambiense* type 1 and did not show evidence of genetic exchanges [77, 78]. Phylogenies by nuclear and mitochondrial DNA consistently placed *T. b. rhodesiense* MHOM-UG-77-KETRI-2355 and ITM001 within the East African group. However, while the nuclear genomes and minicircle compositions of the remaining seven isolates suggested they were closely related to the East African *T. brucei*, they had maxicircles more similar to the West African group. The discrepancy provided evidence of hybridization, which might introduce maxicircles from West African *T. brucei*. Surprisingly, the four *T. b. rhodesiense* isolates and *T. b. brucei* Etat-1-2R had identical maxicircle coding regions, suggesting they probably shared closely related West African parents.

The FEO patient remained in the first stage for 25 years, and two FEO strains were isolated from rodent inoculation, one sensitive to normal human serum (NHS) and one resistant to NHS [333]. The NHS-sensitive cell line is not expected to be able to survive in the patient for long. Probable explanations include sample contamination, mislabelling, and coinfection during mouse inoculation. The similar nuclear and mitochondrial genomes suggested that one of the strains present in the FEO isolates from Togo was probably closely related to the Yaoundé isolate.

Six isolates were collected from West or Central Africa, grouped within the East African clade by genome-wide SNPs and minicircle family profiles, yet grouped within the West African group by maxicircle phylogeny. A probable speculation to address the discrepancy between the nuclear genome, minicircle, and maxicircle phylogenies could be the coinfection and hybridization of isolates with different genetic backgrounds in Uganda, followed by subsequent migrations of humans and livestock.

Despite the placement in the East African group by the nuclear genome phylogeny, seven East African *T. b. rhodesiense* isolates had minicircle family profiles or maxicircles clustered among the West African *T. b. brucei*. The minicircle population and maxicircle coding region of MHOM-ET-67-GAMBELA1 were similar to the West African group, suggesting a more recent hybridization so that the West African signatures in minicircles were not yet diluted by a series of subsequent sexual recombination with East African cell lines. MHOM-ET-69-GAMBELA3 had a minicircle population indicative of a West African origin but a maxicircle more similar to the East African *T. brucei*. The maxicircle coding regions of the remaining five isolates were identical and belonged to the same clade as five other isolates from Uganda (Table 6-4). Hybridization with the same West African strain in Uganda probably contributed to the discrepancy in the mitochondrial genomes of the East African isolates, while human movements spread the maxicircle across East Africa.

6.5 Chapter Conclusions

We assembled 5666 minicircle classes from the 224 sub-Saharan *T. brucei* isolates and assigned them to 891 families and 448 superfamilies. While we observed a few shared minicircle classes, 122 families and 143 superfamilies were present in the four subspecies. An additional 363 families and 207 superfamilies were shared by the subspecies capable of sexual reproduction. We also examined minicircle families and superfamilies in *T. congolense* and type OVI *T. b. equiperdum* and observed similarity among the isolates.

We compared the three phylogenies based on nuclear and mitochondrial genomes. The phylogenies agreed on the general trends, that the gene pools of the Western and Eastern trypanosomes were distinct, that *T. b. gambiense* type I formed a monophyletic group within the West African *T. brucei*, and that *T. b. rhodesiense* and *T. b. brucei* formed a single breeding population. We observed indications of hybridization in Uganda, along with other discrepancies probably due to movements of humans and animals.

The maxicircle-containing type A formed a monophyletic clade with AnTat-17-1 from DRC, which agreed with the hypothesis of a West African ancestor. *T. b. equiperdum* type C and OVI resided on the most divergent branch (based on whole-genome SNP and maxicircle phylogeny) in the East African group composed of the isolates from Zambia. The type C isolates were closely related to the Kiboko-Sindo isolate J10 as reported [92].

7 Discussion

7.1 Key findings

7.1.1 The editing cascade and editing blocks are conserved among African trypanosome isolates, subspecies, and species

Guide RNA alignments on the edited mRNAs reveal semi-regular gRNA clusters, each cluster composed of functionally similar gRNA homologs. We grouped these functionally similar gRNAs into gRNA families and refined the definition of an RNA editing block as the effective editing range of the corresponding gRNA family.

We found that the gRNA families and editing blocks, i.e. the average editing range of the gRNAs from the same editing block, are conserved among sub-Saharan *T. brucei* subspecies. *T. b. gambiense* type 1 and *T. b. equiperdum* type OVI share the same editing blocks as the groups capable of sexual reproduction. Furthermore, the editing blocks are also conserved between *T. brucei* and *T. congolense*, which have distinct minicircle structures and encode gRNA families on different cassettes. This suggests that the delimitation of editing blocks predates the divergence of the African trypanosomes, and that the phasing and length of blocks are quite stable. This finding raised the question of the nature of the underlying stabilising condition.

The editing blocks are ~34 nt in length in sub-Saharan *T. brucei* and ~35 in *T. congolense*. We excluded the 3' most 6 nt from the complementary sequences as the minimal requirement for anchor recognition when inferring the editing blocks from the ~40 nt complementary regions. Since RESC loads a single gRNA each time [269, 270, 345], it probably prefers gRNAs with complementary sequences ~40, which agrees with the median length of the complementary region of gRNAs associated with the stabilizing RESC-A [269]. The average anchor length is 12.6 nt in *T. congolense* and 11.4 nt in *T. brucei*, concordant with the observed length of the gRNA-mRNA duplex [269]. Finally, the length of the remaining guiding region predicted from our annotation is ~30 nt, which also agrees with the observed values in the gRNA-stabilizing RESC-A [269].

Most adjacent editing blocks overlap minimally, spaced at an interval of 30 nt, which equals the length of edited mRNA that fits in RESC (~40 nt) minus the length of anchor (~10 nt), to maximize the coverage of the guiding regions. Based on these considerations, we present possible mechanisms for the editing cascade to account for the strikingly economical overlapping patterns and the conservation of editing block positions. A possibility is that the new RESC-B assembles on the unedited mRNA protruding from the current RESC-B, and the unedited mRNA shielded by the current RESC-B will become exposed and subsequently edited. The length of mRNA shielded by RESC-B is not reported in [269]. If the shielded unedited mRNA is significantly longer than 30 nt, there is probably an additional step to partially dissemble RESC-B into a smaller intermediate (a likely candidate is RESC-C) before the assembly of the new RESC-B.

Another possibility is that after the completion of editing with one gRNA and the detachment of RECC, the current RESC-B presents part of the 5' region of the newly edited

mRNA to facilitate the assembly of the subsequent RESC-B. A likely mechanism is that the current RESC-B remodels into RESC-C, which contains subunits with gRNA-mRNA duplex and is preferably associated with edited mRNAs [269]. In RESC-B, the RESC-C subunits, RESC5, RESC8, RESC10, and RESC14, are also adjacent to the 5' end of the newly edited mRNA [269] and thus can probably interact with the mRNA sequence recognizable by the anchor of the new gRNA. Hence, we hypothesize that RESC-C facilitates the gRNA-mRNA recognition and recruits the RESC-A that carries a gRNA with the correct anchor. As the RESC-A remodels into the new RESC-B for the next round of editing, the RESC-C from the previous RESC-B dissembles. Note that this possibility that RESC-C facilitates RESC-A recruitment is not mutually exclusive with the possibility that RESC-C or RESC-B acts as a ruler to control the length of the newly edited mRNA.

In summary, we reason that the editing cascade begins with the initiation gRNAs and proceeds in steps of ~40 nt, generating ~30 nt of newly edited mRNA each time while keeping ~10 nt from the previous round of editing for gRNA selection by the anchors. This mechanism explains the conservation of anchor and editing block positions between *T. congolense* and *T. brucei* as long as they have editing complexes with similar biophysical limits. The semi-regular steps also entail that the minor 'out-of-phase' gRNAs probably do not participate in mRNA editing, as their anchors are not complementary to the anchor aligning sequences presented during the mainstream editing cascade.

An editing mechanism that keeps the gRNAs in semi-regular clusters brings a selective advantage. The daughter cells receive a mixed minicircle population from the parents during sexual reproduction. Without a universal phasing of editing cascades and gRNA alignment positions, combining a subset from each parent may not provide complete mRNA coverage when the parental cell lines have different gRNA arrangements. Hence, gamete fusion with a cell line with incompatible gRNA arrangements is unproductive at best and fatal at worst. This problem is largely avoided by keeping the editing cascade and gRNA alignment in the same phase. The compatibility of gRNA facilitates sexual reproduction. We propose that the phasing of gRNA is fundamental to the benefit of sexual reproduction, including stabilizing the minicircle population and combating kDNA decay during clonal proliferation, which we will discuss in the next section.

7.1.2 The clonally propagating *T. b. gambiense* type 1 is susceptible to kDNA decay

Vector-borne *T. brucei* cell lines newly established in the mammalian host will possess complete gRNA coverage on all mRNAs of genes essential for the parasite in the insect vector [37, 38]. Nevertheless, the distribution of minicircles into daughter networks is imperfect or even random, leading to the fluctuation in minicircle copy number and making some minicircles and corresponding gRNA gene families present at extremely low abundance in a cell line [241, 244].

A risk associated with having essential minicircles at low abundance is the probability of minicircle loss during clonal reproduction in the mammalian bloodstream. In the mammalian host, the parasites no longer rely on the ETC, which makes most of the maxicircle encoded

genes non-essential, except for the ribosomal and F_1F_o-ATP synthase subunit genes [35, 364]. RPS12 and A6 are the only essential maxicircle-encoded genes whose mRNAs require post-transcriptional editing. Consequently, minicircles without A6 or RPS12 gRNA genes are not vital to the BSF parasites. If some of these minicircles are underrepresented and then lost in some of the daughter cells, these cells will still survive in BSF but can no longer produce all fully edited mRNAs essential at insect stage. This results in a heterogeneous cell population with different capacities for editing and for survival and development in the vector. In other words, the kDNA of a recently transmitted cell line will start to decay in the mammalian bloodstream.

The decay raises a question on how long a *T. b. gambiense* type 1 cell line remains tsetse-transmissible in a patient. Given *T. b. gambiense* type 1 causes chronic HAT, the parasites propagate within the patients and remain in stage 1 for years [366]. Our observation suggests that the parasites will undergo kDNA decay throughout the infection and the proportion of tsetse-transmissible cells will fluctuation and decline in a long run. The rate of kDNA decay is probably inferable by sequencing cell lines cultured in BSF over generations. The data has been generated by our lab, but due to the scope of this study, they will be analysed in future projects.

Sexual reproduction in *trypanosomatids* likely lowers the risk of kDNA decay by replenishing the underrepresented minicircles due to the biparental inheritance of minicircles [53, 73]. On the other hand, we expect strictly clonal *T. brucei*, such as *T. b. gambiense* type 1 [77, 78], to be more susceptible to kDNA decay. In other words, we expect a higher proportion of cells with compromised editing capacity. If the cells capable of producing all fully-edited mRNAs are rare in a population, bulk sequencing may not be sufficiently sensitive to capture the essential underrepresented minicircles. Alternatively, a gRNA family might be missing from all cells in the cell line. One way or the other, our assembly and annotation will not detect the associated gRNA genes. This will manifest as gaps gRNA coverage.

All the *T. brucei* isolates in our study with known hosts were collected from mammals, including human patients inflicted with the human-infective *T. brucei* subspecies. Hence, prior to isolation the cells will have proliferated in the bloodstream (and other tissues) for a period of time. We observed significant decreases in the completeness of editing site coverage in *T. b. gambiense* type 1 over pan-edited mRNAs except A6, RPS12, and ND3. The observation suggests that only a small proportion of cells in a given *T. b. gambiense* type 1 population retain all essential gRNA genes for the insect stage, while all cells are capable of editing A6 and RPS12. Estimating the proportion of cells capable of tsetse transmission would be enlightening. However, we could not conclude without single-cell sequencing data. From bulk sequencing, we could only infer that most, if not all, of the cells were no longer tsetse-transmissible, as the minicircles encoding the gRNAs essential for the insect stage were too rare to be detectable.

An impact of kDNA decay that has progressed to the complete loss of a gRNA family with insect-specific function from an individual cell is the loss of that cell's capacity to survive in insect vectors. The increase in the proportion of cells with degraded kDNA within a mammalian host results in a decline in the transmissibility of the population. *T.b. rhodesiense*

infections in tsetse salivary glands have been detected in field samples, with field isolates readily infecting tsetse in laboratory assays [335]. In contrast, although *T. b. gambiense* type 1 is transmitted by tsetse of the *palpalis* group, surveys of parasite load in adult flies reveal a very low vector infection rate and undetectable salivary gland infection necessary for transmission [367, 368]. In addition, laboratory tsetse infection assays have shown highly variable infection rates in *T. b. gambiense* type 1 isolates recently sampled from the field [369]. The variation likely reflects fluctuations in the proportion of cells containing all essential minicircles for editing insect-stage-specific mRNAs, which can be attributed to differences in the length of passage time in mammalian hosts and the initiation minicircle population. This suggests that *T. b. gambiense* type 1 has difficulty in completing insect-stage development [370]. While cases of vector transmission are rare and instances of vertical transmission have been documented, it has been proposed that *T. b. gambiense* type 1 is primarily transmitted from mother to offspring during gestation instead of exclusively vector-borne [367, 370].

Our study shows that without sexual reproduction to circulate minicircles and gRNA genes through the population, the clonal *T. b. gambiense* type 1 population experiences more substantial kDNA decay during repeated proliferation in the mammalian bloodstream. Our observation is consistent with the hypothesis that *T. b. gambiense* type 1 has variable and generally low success of reinfecting the insect vector. We could expect a higher likelihood of *T. b. gambiense* cell lines losing the ability to produce fully edited mRNAs essential for insect-stage parasites and becoming trapped in the mammalian bloodstream. If some cell lines are adapted to alternative transmission modes, they will have a chance to evade extinction at the death or cure of the current host, potentially giving rise to new tsetse-independent groups and expanding beyond the tsetse-belt.

We also observed low kDNA complexity in some *T. b. brucei* isolates, suggesting potential strictly clonal isolates that are not human infective. Field *T. b. brucei* isolates with low kDNA complexity may shed light on the mechanism that drives the abandonment of sexual reproduction in trypanosomatids and even the emergence of tsetse-independent groups, including *T. b. evansi* and *T. b. equiperdum* [83, 84, 91, 94, 96].

7.1.3 *T.* brucei and *T. congolense* probably have lost most maxicircle-encoded gRNAs

There are two major clades of trypanosomatids, the trypanosomes and the clade that includes *Leishmania*, *Crithidia*, *Leptomonas*, *Phytomonas*, *Herpetomonas*, *Blastocrithidia* and others [295, 371]. While *T. brucei* and *T. congolense* have two and three maxicircle encoded gRNAs, respectively, as many as 20 gRNA genes have been detected in maxicircles of *Leishmania* species [66].

We are interested in the origin of the maxicircle-encoded gRNAs that are absent from *T. congolense* and *T. brucei*, as this line of investigation may provide insight into gRNA gene evolution in general. Several scenarios seem plausible: (i) an ancient gRNA gene predates the divergence of the two major clades and subsequently lost in the trypanosome lineages due to changes in the mRNA editing patterns or gRNA gene degradations and deletions; (ii)

subsequent acquisition after the divergence of the two clades, for example via the proposed mechanism of cDNA insertion (retroposition) [295].

We have identified three maxicircle encoded gRNA genes in *T. congolense* IL3000: gCOX2_507-512, gMURF2-II_40-80, and gND7_1200-1252, the latter so far being unique to *T. congolense*. The COX2 and MURF2 gRNA genes are found in all species examined and therefore probably predate the divergence of the two trypanosomatid clades. Eight genes reported in *L. peruviana* and *L. braziliensis* overlap other maxicircle genes [66]. Due to the conserved genetic background and general synteny over the gene-coding region in trypanosomatid maxicircles, we could precisely align the *T. congolense* and *Leishmania* sequences. None of the aligned regions in IL3000 show any recognisable complementarity to edited mRNAs in IL3000. This argues against a previous function as gRNA genes. Nevertheless, we detected a single IL3000 read with an untemplated U-tail over the region aligned with *L. braziliensis* gND9_304-355, which suggests that the aligned region is transcribed and may have once been functional, but the functionality is lost in lineages including *T. brucei* and *T. congolense*.

Trypanosomes minicircles encode the functional equivalents of the *Leishmania* maxicircle gRNAs [250]. While the evolution of maxicircle gRNA genes that overlap other maxicircle genes is restricted, minicircle encoded gRNAs allow the editing patterns to change by introducing more U-insertions [257, 372]. Such changes may compromise the complementarity of the maxicircle-encoded gRNAs and the edited mRNAs. Subsequently, the maxicircle-encoded gRNAs can no longer align with the edited mRNAs in the trypanosomes despite being transcribed. Eventually, the transcription and post-transcriptional modification stop on the non-functional gRNAs. This hypothesis also entails that complementary mRNA fragments may serve as one source of gRNAs [295].

Among gRNA genes in the maxicircle coding region, the gCYB-II gene is found in intergenic regions in examined *Leishmania* species and *C. fasciculata.* The corresponding intergenic region is too short for gene encoding in *T. brucei* and *T. congolense*. Hence, we propose that the gCYB-II gene represents a gRNA gene insertion event in the non-trypanosome clade or a gene deletion in the trypanosome clade after the divergence of their common ancestor. Notably, all CYB gRNAs in *T. brucei* and *T. congolense* are encoded on minicircles as orphan gRNA genes unbounded by cassettes. More gene insertion and deletions may occur in the maxicircle variable region, but the lack of sequence homology hinders further investigation without availability of complementary transcriptomics data.

7.1.4 *T. b.* equiperdum type OVI is probably kDNA dependent

Type OVI isolates (OVI, Dodola, TeApND1) form a monophyletic clade based on genome-wide SNPs [85, 90, 93] and maxicircles [92]. Unlike type A, B, and C isolates with a single minicircle class specific to each group, *T. b. equiperdum* type OVI isolates have as many as 45 distinct minicircle classes per network. These minicircle classes encode gRNAs sufficient for directing complete editing of the A6 and nearly complete editing of RPS12 mRNAs. We did not detect all RPS12 gRNA genes, probably because of errors in the edited mRNA

prediction due to the lack of transcriptome data or the arbitrary criteria for gRNA alignment was too strict in the annotation pipeline. In addition, mutations that compensate for kDNA loss have not been detected in *T. b. equiperdum* type OVI [91], suggesting that type OVI may not be kDNA-independent as commonly believed.

Our collaborators at ITM conducted *in vitro* ethidium bromide (EtBr) challenges on type OVI type C, type A, and kDNA-dependent *T. b. brucei* EATRO1125 [300]. Introduced in the 1950s, EtBr is still used as an anti-trypanosomal drug for African cattle [373]. EtBr is known to cause the loss of kDNA and most likely kills kDNA-dependent parasites by inhibiting minicircle replication initiation [373]. At higher concentrations, EtBr also inhibits nuclear DNA replication and kills the a/dyskinetoplastic parasites [373]. In the experiment at ITM, the cells were cultured at 0.01 nM, 0.1nM, and 1nM EtBr for 4-25 days, and their mobility and viability were monitored. Type OVI stopped proliferating at 0.1 nM EtBr. In contrast, the kDNA-independent controls proliferated at all EtBr concentrations. Additionally, all type OVI isolates were more sensitive to EtBr than the kDNA-dependent EATRO1125.

Based on these results, we propose that *T. b. equiperdum* type OVI is indeed kDNA-dependent. Whether the loss of essential gRNAs for cryptogenes necessary in insect stage parasites has occurred as a cause or result of abandoning tsetse transmission is unclear. Nevertheless, it has been reported that the strictly clonal yet tsetse-transmissible *T. b. gambiense* type 1 has a low tsetse-infection rate in the field [368]. We could hypothesize a scenario for the emergence of *T. b. equiperdum* OVI, where clonal tsetse-transmissible cell lines became trapped in the mammalian hosts due to the loss of essential minicircle classes through imperfect replication and segregation. Some cells within the population may acquire mutations that allow them to transmit directly between mammalian hosts and manage to pass down the cell line [85]. Subsequently, they lose the now redundant minicircles that only encoded gRNAs required in the insect stage.

The higher sensitivity to EtBr suggests that the type OVI isolates are more susceptible to disturbance of kDNA replications. A possibility is that they have reached the limit of kDNA reduction, that minor disruptions to the kDNA network can result in the loss of essential gRNA genes and undermine the production of fully edited A6 and RPS12. We have shown that the minicircles encoding A6/RPS12 gRNA genes were not more abundant than other minicircles and that some A6/RPS12 gRNA-encoding minicircles were estimated to be present at between one and two copies per network. A few unsuccessful replications could eliminate one of these minicircle classes, and their loss would be lethal to the daughter cells as fully edited A6 and RPS12 mRNAs can no longer be produced. In contrast, the redundant editing capacity of AnTat 1.1E can buffer some minicircle loss before it starts affecting the editing capacity.

7.1.5 The definition of *T. b. evansi* and *T. b. equiperdum* is controversial

T. b. evansi and *T. b. equiperdum* are thought to have evolved from *T. b. brucei* by switching from tsetse transmission to mechanical and sexual transmission [83, 90, 374, 375]. Both groups are morphologically indistinguishable from *T. b. brucei* [183]. The classification of the

two groups has historically been made from a veterinary perspective with little awareness of the evolutionary history of the parasites. With the emergence of molecular evidence, the definition of *T. b. evansi* and *T. b. equiperdum* remains controversial. Although some argue that the current nomenclature should be retained for practicality [179], many diagnostic traits for the two groups are unstable and ambiguous [376].

First, *T. b. evansi* is considered a blood parasite and *T. b. equiperdum* a tissue parasite found in equines in nature [83]. Nevertheless, when *T. b. evansi* exists at very high parasitaemia in camels, horses, and dogs, it invades the central nervous system [179, 377]. Cases of transmission during coitus also suggest invasion into genital mucosae [83]. On the other hand, while the establishment of *T. b. equiperdum* in the blood of laboratory animals (mice) is extremely challenging, once established, the murine-adapted clones can be maintained by serial passages and cause acute infection similar to *T. b. evansi* [83, 378]. The observation suggests that *evansi*-like parasites could be selected from *T. b. equiperdum* cells [83]. In addition, as an extracellular blood parasite, the presumably ancestral *T. brucei* also invades nervous [105], skin [379], and adipose tissue [105]. In summary, tissue tropism is probably not a reliable and distinguishable trait of *T. b. evansi* and *T. b. equiperdum*.

Second, *T. b. evansi* is transmitted mechanically via biting insects and vampire bats, while *T. b. equiperdum*, present in the seminal fluid and mucous membranes of the genitalia of the infected donor animal, is transmitted between horses during copulation. However, sexual transmission is not unique to *T. b. equiperdum*, as *T. b. evansi* can also be directly transmitted during coitus [83]. On the other hand, while not all factors contributing to mechanical transmission are characterized, it has been proposed that the efficacy of mechanical transmission is directly proportional to parasitaemia [380]. The observation that the murine-adapted laboratory clones of *T. b. equiperdum* can be maintained through serial passages also suggests a flexibility in the transmission mechanism of *T. b. equiperdum* [83].

Third, first described in 1880 as a parasite of camel, *T. b. evansi* has substantially expanded its host range and infects a large variety of wild and domestic animals, including equines [179]. As the name suggested, *T. b. equiperdum* is considered an exclusive parasite of horses. However, some propose that the dourine might also occur in donkeys and mules without obvious clinical signs [83]. Furthermore, sheep and goats can be infected with the murine-adapted *T. b. equiperdum* strains and exhibit typical dourine symptoms [83].

Finally, *T. b. evansi* and *T. b. equiperdum* sometimes exhibit similar molecular markers [83]. PCR based on the RoTat 1.2 VSG, random amplified polymorphic DNA (RAPD), and multiple endonuclease genotyping approach (MEGA) group eight historical type A *T. b. equiperdum* isolates with type A *T. b. evansi* [183]. In addition, characterization of 16 enzymes has suggested that the *T. b. evansi* and *T. b. equiperdum* (both type A) from China form a homogeneous group [83, 354, 381].

Based on nuclear genome and maxicircle analyses, it was recognised that at least four independent events have led to the emergences of different groups of *T. b. equiperdum* and *T. b. evansi*, namely the type A, B, C, and OVI groups [90-93]. We have confirmed in this project that the four groups also have unique minicircle compositions. The observation

further supports the notion that the conventional definition of *T. b. evansi* and *T. b. equiperdum* is inevitably paraphyletic and polyphyletic. From an evolutionary perspective, it is probably more accurate to treat each group as a unique lineage derived from *T. b. brucei* that has abandoned tsetse transmission and adapted to alternative transmission modes. Although kDNA-independence [179] and the single point mutations in ATP synthase that compensate for mitochondrial genome loss in trypanosome [95] have been suggested to play a role in the transition, the observation that type OVI is probably kDNA-dependent supports the involvement of other factors. Given the high level of parasitaemia in the blood critical for mechanical transmission [380] and the need for the hosts to survive long enough for transmission, the parasite may have uncontrolled proliferation in some hosts and more controlled proliferation in others [91]. Hence, reaching equilibrium between the survival of the parasite and the host is probably key to the success of mechanical transmission [91]. Alternatively, using skin as an additional reservoir, the parasite probably lowers the threshold parasitaemia for mechanical transmission and allows longer proliferation time within one host [382].

Brun and colleagues have proposed a model where *T. b. evansi* emerged from *T. b. equiperdum* [83]. Although this hypothesis is incompatible with the multiple origins of the two groups, it could explain the evolution route of type A parasites. Some type A isolates lacked maxicircles as expected for *T. b. evansi*, while seven type A isolates contained partial or complete maxicircles. In a phylogeny based on type A minicircles, the maxicircle-containing isolates form a basal monophyletic group, while the Vietnam isolate from a buffalo with a complete maxicircle is basal to other isolates with a conserved truncation on maxicircles. We propose that the ancestral type A has been through a stage when the minicircle population has become homogeneous yet the maxicircles remain present, similar to *T. b. equiperdum* type C. Subsequently, a deletion occurred in the no-longer-functional maxicircles, eventually leading to the loss of maxicircles in the typical '*T. b. evansi* type A'.

To reconcile both models, we hypothesize that each tsetse-independent *T. b. brucei* lineage emerges when a *T. b. brucei* cell line acquires other means of transmission. The cell line is initially kDNA-dependent and experiences a reduction in kDNA complexity as the minicircles without A6 or RPS12 gRNA genes become redundant, which we have observed in type OVI. Subsequently, the cell line may acquire kDNA-independence. It undergoes further reduction in kDNA complexity that culminates in a homogeneous minicircle population as in type C and type A *T. b. equiperdum*. Potentially initiating with partial maxicircle sequence deletions, the ultimately complete loss of maxicircles results in a dyskinetoplastic state similar to type B and most type A isolates. Finally, the loss of minicircles leads to the akinetoplastic cells. We propose that this process is repeated each time a tsetse-independent *T. b. brucei* cell line occurs, resulting in the unique minicircle marker observed in each group.

7.1.6 Some *T. brucei* isolates can only produce a single version of A6

We observed alternative editing patterns unique to closely related *T. brucei* subspecies and *T. congolense*. The significance of these alternative products remains uncertain despite

various observation and hypothesis [225, 281, 282]. More group-specific alternative editing patterns are probably yet to be revealed by RNA transcriptomics.

Nevertheless, A6 alternative editing in different *T. brucei* subspecies sheds light on this issue, suggesting that not all alternative editing events play a critical role. Two versions of fully edited A6 have so far been reported that are conserved in at least some strains or isolates of *T. brucei* [225], and no other patterns were found in this study. Both versions are present in the *T. b. brucei* reference strain EATRO1125 [225]. The two versions differ in the editing sites within the 3' UTR and require two different initiation gRNAs. However, the gRNAs are encoded on minicircles of the same family, with identical cassette families on other positions, which suggests that the alternative gRNAs have a common origin and at some point in time were homologous. Hence, the alternative editing probably results from a functional mutant gRNA. Whether the alternative 3' UTR editing patterns have functional consequences remains to be elucidated.

Here we found that not all isolates are capable of producing both versions of A6. *T. b. equiperdum* type OVI only produces A6_v1, but not v2. In contrast, *T. b. gambiense* type 1 produces A6_v2 but not A6_v1. The sub-Saharan *T. brucei* subspecies capable of sexual reproduction include isolates with initiation gRNAs for both versions of A6, but the majority, namely 68/78 *T. b. brucei* and 12/23 *T. b. rhodesiense* isolates, are only capable of producing A6_v2. Alternative editing of A6 is therefore not essential for these isolates. Hence, we propose that alternative editing of A6 is not essential and probably associated with a tolerated gRNA gene mutation. However, it does not preclude the possibility that alternative editing may regulate gene expression or serve other beneficial purposes.

However, we also detected alternative editing at the 3' end of A6 in *T. congolense*. Although the BSF and EMF samples have different proportions of PacBio reads with each version of A6, we cannot draw conclusions about the stage-specific expression because of the lack of replicates. Similar to *T. brucei*, the alternative editing occurs in the 3' UTR and does not lead to different proteins. However, unlike *T. brucei*, the alternative editing sites occur upstream of the initiation gRNA, while the minicircles encoding the alternative gRNAs have distinct cassette families and substantial sequence divergence. Hence, we cannot attribute the emergence of alternative A6 in *T. congolense* simply to gRNA gene mutations. A few different scenarios may explain this situation: (1) recombination of minicircles followed by prolonged divergence has erased the homology, or (2) random mutations in unrelated minicircles have altered the function of a gRNA gene and allowed it to edit A6.

Nevertheless, the presence of alternative editing on the 3' end of A6 in *T. brucei* and *T. congolense* is intriguing. If the alternative editing of A6 bears no functional relevance, the coincidence indicates that certain features of the 3' end of A6 encourage or tolerate alternative editing. Alternatively, the alternative editing may not be essential but beneficial, which favors its maintenance once it emerges.

7.1.7 kDNA replication could probably be more precise than expected

OVI was isolated in 1975 in South Africa [90], Te-Ap-ND1 in 1990 in Venezuela [186], and Dodola 940 in 2008 in Ethiopia [352]. It was striking that the MCNs were highly correlated in isolates collected decades apart from different continents. This observation raises the question of how type OVI faithfully replicates the kDNA into the daughter cells, considering other studies have suggested a highly dynamic network [241, 244].

It has been proposed that the minicircle segregation into daughter cells is random in *T. brucei* [241, 244]. Passaging BSF *T. brucei* over 250 generations results in a significant loss of minicircle diversity and fluctuation in the copy number of the remaining minicircles within [225], demonstrating that kDNA replication and segregation are clearly not entirely faithful processes. However, type OVI seems to have minimized such changes to the kDNA network for at least three decades.

One possibility is that the replication of type OVI kDNA is not *per se* more accurate than in other *T. brucei.* Instead, the intolerance to changes in minicircle copy numbers may reflect a strong selection against cells with altered networks, leaving only cells with conserved kDNA networks in the population. Such purging selection might entail that a large proportion of daughter cells will not be viable. Interestingly, we also observed the conservation of MCNs among the 12 minicircles not encoding essential A6/RPS12 gRNAs. This observation might suggest that these minicircles have a structural role that remains to be identified. A possible explanation is that the type OVI kDNA has reached its limit of reduction and any further disruption, including the loss of seemingly redundant minicircles, would incur unwanted chain effect that undermines the entire network. This agrees with the observed higher sensitivity to EtBr.

Alternatively, type OVI isolates replicate their kDNA network with higher accuracy. A popular model implies an unknown sorting and transport complex that imperfectly assigns the daughter minicircles to the antipodal sites [22, 240]. Recently, the loose-diploid model tried to account for the segregation of minicircles without assuming such a sorting apparatus [236]. However, the intrinsic sloppiness of the loose-diploid model makes it hard to explain the conservation of MCN in type OVI. Otherwise, if the former scenario is true, a considerable number of daughter cells will not be viable after division.

We propose that the assignment of minicircles into daughter cells is imperfect but not random. The redundancy of gRNA genes in EATRO1125 masks the sorting mechanism behind the segregation or provides wiggle-rooms for multiple ‘solutions’ to be acceptable. In contrast, the highly streamlined type OVI kDNA removes the noises from the redundancy and forces the precision of the minicircle curation to emerge.

7.2 Limitations and Prospects

The detection of minicircles at less than one copy per network indicates kDNA heterogeneity within cell lines. The population-level sequencing used in all projects could not capture the variation of the minicircle composition among cells within a population. The hypothesis proposed in the previous sections could be addressed with single-cell resolution. Firstly, it could be tested if the *T. b. gambiense* type 1 isolates are more susceptible to kDNA decay by determining the percentage of cells incapable of producing fully-edited mRNAs of insect-stage-specific genes within each isolate. The test is also key in addressing the low field tsetse infection rate and the variable experimental tsetse infection rate [369]. To test this possibility, future experiments could compare isolates capable of sexual reproduction, for which we expect a higher percentage of cells with healthy kDNA networks. Secondly, it could be tested if the cells within type OVI cell lines have highly correlated MCNs per network and compare it with a cell line with redundant editing capacity such as EATRO1125. This would shed light on how the parasites regulate the replication and segregation of the kDNA network. With current advances in single-cell technologies, one can characterize the population structure at higher resolution. Combined with the data available from population-based methods to compensate for the insufficient read depth often associated with single-cell sequencing, this technology would unfold more mysteries on the dynamics of trypanosome kDNA.

Besides the sequencing technique, the minicircle assembly pipeline probably limited the detection of minicircles with atypical CSB-3 sequences. We allowed a single mismatch in our CSB-3 motif search while extracting candidate minicircle contigs. This method may miss minicircles with less conserved CSB-3 (i.e. with two mismatches) and lead to the wrong conclusion about the conservation of CSB-3. An alternative is to align and examine all circularizable contigs from the primary assembly. Nevertheless, we believe that our search criteria did not miss a significant amount of minicircles if any, given the conservation of CSB-3 reported in literature. Conserved CSB-3 motifs are detected in the divergent trypanosome and *Leishmania* clade [19, 197, 229]. Its critical involvement in UMSBP binding for minicircle replication [19, 230] suggests that minicircles with substantially divergent CSB-3 motifs would be unexpected.

The prediction of fully edited mRNAs experienced a similar problem. The alignment method for preliminary edited mRNA prediction prevents us from identifying alternative editing patterns with more differences from the canonical sequences. However, an exhaustive analysis of probable alternative editing patterns was not the goal of this project. A more detailed examination of the ORFs of the *T. congolense* PacBio reads combined with greater read depths may identify candidates with highly divergent editing patterns. Nevertheless, the current annotation suggests that most transcripts contain a fully edited region, probably with minor sequence variations from the canonical editing, and an unedited region, sometimes joined by a short junction with noncanonical editing. Meanwhile, similar data should be generated for other (sub)species.

For *T. b. equiperdum*, the lack of transcriptome data restricts the edited mRNA prediction to inference based on U-striped alignments and limits its accuracy. The conclusion of this study

could be greatly improved by the availability of type OVI transcriptome data (ideally generated with both PacBio and Illumina technology), which would also provide insights into the editing activities and the expression of A6 and RPS12 genes. The mRNA predictions might be corrected to allow detection of the missing RPS12 gRNAs. Although the preliminary test has indicated kDNA dependence in type OVI, more rigorous experiments should be conducted to confirm that, probably using proteomics to detect the respective protein products. Transcriptome data of sRNA could probably help to identify the currently missing RPS12 gRNAs while answering questions on minicircle expression.

In addition, for the 224 *T. brucei* isolates, we did not correct the unedited and edited mRNAs for individual isolates but used EATRO1125 mRNAs for all isolates capable of sexual reproduction. Ideally, we would annotate the maxicircle of each isolate and extract the unedited mRNAs for edited mRNA corrections by U-stripped alignment as described in preliminary edited mRNA prediction of *T. congolense*, *T. b. gambiense* type 1 isolate Mongo, and type OVI. The analysis could also benefit from transcriptome data, especially for isolates with highly divergent maxicircle sequences, such as the Zambian isolates including J10 that were closely related to *T. b. equiperdum*. This was not done due to time and financial restrictions but would be an important avenue to pursue in future studies as it would add confidence to gRNA predictions.

Unfortunately, we could not draw robust conclusions about the expression of gRNAs due to the lack of sRNA data (*T. b. gambiense* type 1) or the lack of biological replicates (*T. congolense*). Small-RNA sequencing of mitochondrial RNAs would be vital in determining the expression status of the predicted gRNA genes and corroborating the annotation pipeline. Interesting topics to be explored include the expression and modification of non-canonical gRNAs, the differential expression of gRNA genes within one family, stage-specific gRNA expression, and the expression of gRNA genes on the same minicircle. Epigenetic modification on minicircle for gRNA regulation is probably another fruitful path to pursue.

The kDNA annotation pipeline used arbitrary cut-offs for minimum anchor lengths, mismatches, and minimum gRNA length. A previous study using the same pipeline has reported gRNAs not detected due to mismatches exceeding the threshold, G: U wobble base pairing in the anchor regions, and gaps required for complementarity [225]. Unfortunately, the exact gRNA selection and alignment to mRNA is still unclear. The stringency probably generates false negatives and obscures the flexibility in the gRNA function. Nevertheless, the higher quality A6/RPS12 gRNA genes in type OVI suggest that gRNAs with longer anchors and fewer mismatches are favoured or even crucial despite the presence of lower-quality homologs.

To elucidate the gRNA-mRNA interaction, future studies may consider techniques including crosslinking and sequencing of hybrids (CLASH) [383], RIC-seq (Cai, Z. et al. RIC-seq for global in situ profiling of RNA–RNA spatial interactions. Nature 582, 432–437 (2020).), miRNA Trapping by RNA in Vitro Affinity Purification (miTRAP) [384], or RNA Hybrid and Individual-Nucleotide Resolution Ultraviolet Crosslinking and Immunoprecipitation (hiCLIP) [385]. These methods exploit the *in vivo* protein-RNA crosslinking and generate chimeric reads of the ligated interacting RNAs, in our case, the gRNA-mRNA chimera. This would enable the

precise mapping of gRNA on mRNA and answer the following questions: what gRNAs are linked to mRNAs and what is the criteria for gRNA selection by RESC, are gRNAs aligned to mRNAs in a step-wise cascade as predicted by the editing blocks, and does the gRNA affinity differ across the mRNAs.

8 Reference s

1. Kostygov, A.Y., et al., *Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses.* Open Biol, 2021. **11**(3): p. 200407.
2. Richardson, J.B., et al., *Genomic analyses of African Trypanozoon strains to assess evolutionary relationships and identify markers for strain identification.* PLoS Negl Trop Dis, 2017. **11**(9): p. e0005949.
3. Kimaro, E.G., et al., *Occurrence of trypanosome infections in cattle in relation to season, livestock movement and management practices of Maasai pastoralists in Northern Tanzania.* Vet Parasitol Reg Stud Reports, 2018. **12**: p. 91-98.
4. Echodu, R., et al., *Genetic diversity and population structure of Trypanosoma brucei in Uganda: implications for the epidemiology of sleeping sickness and Nagana.* PLoS Negl Trop Dis, 2015. **9**(2): p. e0003353.
5. Simpson, A.G., J.R. Stevens, and J. Lukes, *The evolution and diversity of kinetoplastid flagellates.* Trends Parasitol, 2006. **22**(4): p. 168-74.
6. Cooper, C., et al., *The marsupial trypanosome Trypanosoma copemani is not an obligate intracellular parasite, although it adversely affects cell health.* Parasit Vectors, 2018. **11**(1): p. 521.
7. Godfrey, S.S., et al., *Trypanosome co-infections increase in a declining marsupial population.* Int J Parasitol Parasites Wildl, 2018. **7**(2): p. 221-227.
8. Bailes, E.J., et al., *Host density drives viral, but not trypanosome, transmission in a key pollinator.* Proc Biol Sci, 2020. **287**(1918): p. 20191969.
9. Smit, N.J., et al., *Morphological and molecular characterization of an African freshwater fish trypanosome, including its development in a leech vector.* Int J Parasitol, 2020. **50**(10-11): p. 921-929.
10. Lehmann, D.L., *A new species of trypanosome from the salamander Ambystoma gracile, with notes on a collection of amphibian blood smears.* J Parasitol, 1954. **40**(6): p. 656-9.
11. Minter-Goedbloed, E., et al., *First record of a reptile trypanosome isolated from Glossina pallidipes in Kenya.* Z Parasitenkd, 1983. **69**(1): p. 17-26.
12. Votypka, J., et al., *Trypanosoma culicavium sp. nov., an avian trypanosome transmitted by Culex mosquitoes.* Int J Syst Evol Microbiol, 2012. **62**(Pt 3): p. 745-754.
13. Fernandes, A.P., K. Nelson, and S.M. Beverley, *Evolution of nuclear ribosomal RNAs in kinetoplastid protozoa: perspectives on the age and origins of parasitism.* Proc Natl Acad Sci U S A, 1993. **90**(24): p. 11608-12.
14. Deschamps, P., et al., *Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids.* Mol Biol Evol, 2011. **28**(1): p. 53-8.
15. Maslov, D.A., et al., *Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed.* Trends Parasitol, 2013. **29**(1): p. 43-52.
16. Teixeira, M.M., et al., *Phylogenetic validation of the genera Angomonas and Strigomonas of trypanosomatids harboring bacterial endosymbionts with the description of new species of trypanosomatids and of proteobacterial symbionts.* Protist, 2011. **162**(3): p. 503-24.
17. Hamilton, P.B., et al., *Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA.* Int J Parasitol, 2004. **34**(12): p. 1393-404.
18. Santos, D.O., et al., *Infection of mouse dermal fibroblasts by the monoxenous trypanosomatid protozoa Crithidia deanei and Herpetomonas roitmani.* J Eukaryot Microbiol, 2004. **51**(5): p. 570-4.
19. Abu-Elneel, K., I. Kapeller, and J. Shlomai, *Universal minicircle sequence-binding protein, a sequence-specific DNA-binding protein that recognizes the two replication origins of the kinetoplast DNA minicircle.* J Biol Chem, 1999. **274**(19): p. 13419-26.

20. Drew, M.E. and P.T. Englund, *Intramitochondrial location and dynamics of Crithidia fasciculata kinetoplast minicircle replication intermediates.* J Cell Biol, 2001. **153**(4): p. 735-44.
21. Ferguson, M., et al., *In situ hybridization to the Crithidia fasciculata kinetoplast reveals two antipodal sites involved in kinetoplast DNA replication.* Cell, 1992. **70**(4): p. 621-9.
22. Guilbride, D.L. and P.T. Englund, *The replication mechanism of kinetoplast DNA networks in several trypanosomatid species.* J Cell Sci, 1998. **111 (Pt 6)**: p. 675-9.
23. Melendy, T., C. Sheline, and D.S. Ray, *Localization of a type II DNA topoisomerase to two sites at the periphery of the kinetoplast DNA of Crithidia fasciculata.* Cell, 1988. **55**(6): p. 1083-8.
24. Perez-Morga, D.L. and P.T. Englund, *The attachment of minicircles to kinetoplast DNA networks during replication.* Cell, 1993. **74**(4): p. 703-11.
25. Simpson, L., A.M. Simpson, and R.D. Wesley, *Replication of the kinetoplast DNA of Leishmania tarentolae and Crithidia fasciculata.* Biochim Biophys Acta, 1974. **349**(2): p. 161-72.
26. Soh, B.W. and P.S. Doyle, *Deformation Response of Catenated DNA Networks in a Planar Elongational Field.* Acs Macro Letters, 2020. **9**(7): p. 944-949.
27. Klotz, A.R., B.W. Soh, and P.S. Doyle, *Equilibrium structure and deformation response of 2D kinetoplast sheets.* Proc Natl Acad Sci U S A, 2020. **117**(1): p. 121-127.
28. He, P.Y., et al., *Single-Molecule Structure and Topology of Kinetoplast DNA Networks.* Physical Review X, 2023. **13**(2).
29. Englund, P.T., *Free minicircles of kinetoplast DNA in Crithidia fasciculata.* J Biol Chem, 1979. **254**(11): p. 4895-900.
30. Sugisaki, H. and D.S. Ray, *DNA sequence of Crithidia fasciculata kinetoplast minicircles.* Mol Biochem Parasitol, 1987. **23**(3): p. 253-63.
31. Yasuhira, S. and L. Simpson, *Minicircle-encoded guide RNAs from Crithidia fasciculata.* RNA, 1995. **1**(6): p. 634-43.
32. Ramakrishnan, S., et al., *Single-Molecule Morphology of Topologically Digested Olympic Networks.* PRX Life, 2024. **2**(1): p. 013009.
33. Smith, T.K., et al., *Metabolic reprogramming during the Trypanosoma brucei life cycle.* F1000Res, 2017. **6**.
34. Matthews, K.R., *Trypanosome Signaling-Quorum Sensing.* Annu Rev Microbiol, 2021. **75**: p. 495-514.
35. Schnaufer, A., et al., *The F1-ATP synthase complex in bloodstream stage trypanosomes has an unusual and essential function.* EMBO J, 2005. **24**(23): p. 4029-40.
36. Ooi, C.P. and P. Bastin, *More than meets the eye: understanding Trypanosoma brucei morphology in the tsetse.* Front Cell Infect Microbiol, 2013. **3**: p. 71.
37. van Hellemond, J.J., F.R. Opperdoes, and A.G. Tielens, *The extraordinary mitochondrion and unusual citric acid cycle in Trypanosoma brucei.* Biochem Soc Trans, 2005. **33**(Pt 5): p. 967-71.
38. Bringaud, F., L. Riviere, and V. Coustou, *Energy metabolism of trypanosomatids: adaptation to available carbon sources.* Mol Biochem Parasitol, 2006. **149**(1): p. 1-9.
39. Naguleswaran, A., et al., *Developmental changes and metabolic reprogramming during establishment of infection and progression of Trypanosoma brucei brucei through its insect host.* PLoS Negl Trop Dis, 2021. **15**(9): p. e0009504.
40. Dolezelova, E., et al., *Cell-based and multi-omics profiling reveals dynamic metabolic repurposing of mitochondria to drive developmental progression of Trypanosoma brucei.* PLoS Biol, 2020. **18**(6): p. e3000741.
41. Gibson, W. and L. Peacock, *Fluorescent proteins reveal what trypanosomes get up to inside the tsetse fly.* Parasit Vectors, 2019. **12**(1): p. 6.

42. Steketee, P.C., et al., *Divergent metabolism between Trypanosoma congolense and Trypanosoma brucei results in differential sensitivity to metabolic inhibition.* PLoS Pathog, 2021. **17**(7): p. e1009734.
43. Silvester, E., A. Ivens, and K.R. Matthews, *A gene expression comparison of Trypanosoma brucei and Trypanosoma congolense in the bloodstream of the mammalian host reveals species-specific adaptations to density-dependent development.* PLoS Negl Trop Dis, 2018. **12**(10): p. e0006863.
44. Peacock, L., et al., *The life cycle of Trypanosoma (Nannomonas) congolense in the tsetse fly.* Parasit Vectors, 2012 Jun 27. **5**: p. 109.
45. Peacock, L., et al., *Experimental genetic crosses in tsetse flies of the livestock pathogen Trypanosoma congolense savannah.* Parasites & Vectors, 2024. **17**(1).
46. Boundenga, L., et al., *Molecular Identification of Trypanosome Diversity in Domestic Animals Reveals the Presence of Trypanosoma brucei gambiense in Historical Foci of Human African Trypanosomiasis in Gabon.* Pathogens, 2022. **11**(9).
47. Bienen, E.J., P. Webster, and W.R. Fish, *Trypanosoma (Nannomonas) congolense: changes in respiratory metabolism during the life cycle.* Exp Parasitol, 1991. **73**(4): p. 403-12.
48. Tait, A., *Evidence for diploidy and mating in trypanosomes.* Nature, 1980. **287**(5782): p. 536-8.
49. Jenni, L., *Sexual stages in trypanosomes and implications.* Ann Parasitol Hum Comp, 1990. **65 Suppl 1**: p. 19-21.
50. Jenni, L., et al., *Hybrid Formation between African Trypanosomes during Cyclical Transmission.* Nature, 1986. **322**(6075): p. 173-175.
51. Gibson, W., et al., *Chapter 24 - Genetic Exchange in Trypanosomatids and its Relevance to Epidemiology*, in *Genetics and Evolution of Infectious Diseases (Third Edition)*, M. Tibayrenc, Editor. 2024, Elsevier. p. 607-634.
52. Peacock, L., M. Bailey, and W. Gibson, *Dynamics of gamete production and mating in the parasitic protist Trypanosoma brucei.* Parasit Vectors, 2016. **9**(1): p. 404.
53. Gibson, W., et al., *The use of yellow fluorescent hybrids to indicate mating in Trypanosoma brucei.* Parasit Vectors, 2008. **1**(1): p. 4.
54. Peacock, L., et al., *Dynamics of infection and competition between two strains of Trypanosoma brucei brucei in the tsetse fly observed using fluorescent markers.* Kinetoplastid Biol Dis, 2007. **6**: p. 4.
55. Peacock, L., et al., *Intraclonal mating occurs during tsetse transmission of Trypanosoma brucei.* Parasit Vectors, 2009. **2**(1): p. 43.
56. Gibson, W.C., *Analysis of a genetic cross between Trypanosoma brucei rhodesiense and T. b. brucei.* Parasitology, 1989. **99 Pt 3**: p. 391-402.
57. Peacock, L., et al., *Identification of the meiotic life cycle stage of Trypanosoma brucei in the tsetse fly.* Proc Natl Acad Sci U S A, 2011. **108**(9): p. 3671-6.
58. Peacock, L., et al., *Meiosis and haploid gametes in the pathogen Trypanosoma brucei.* Curr Biol, 2014. **24**(2): p. 181-186.
59. Peacock, L., et al., *Sequential production of gametes during meiosis in trypanosomes.* Commun Biol, 2021. **4**(1): p. 555.
60. Inbar, E., et al., *Whole genome sequencing of experimental hybrids supports meiosis-like sexual recombination in Leishmania.* Plos Genetics, 2019. **15**(5).
61. Sadlova, J., et al., *Visualisation of Leishmania donovani fluorescent hybrids during early stage development in the sand fly vector.* PLoS One, 2011. **6**(5): p. e19851.
62. Akopyants, N.S., et al., *Demonstration of genetic exchange during cyclical development of Leishmania in the sand fly vector.* Science, 2009. **324**(5924): p. 265-8.
63. Belli, A.A., M.A. Miles, and J.M. Kelly, *A Putative Leishmania-Panamensis Leishmania-Braziliensis Hybrid Is a Causative Agent of Human Cutaneous Leishmaniasis in Nicaragua.* Parasitology, 1994. **109**: p. 435-442.

64. Delgado, O., et al., *Cutaneous leishmaniasis in Venezuela caused by infection with a new hybrid between Leishmania (Viannia) braziliensis and L. (V.) guyanensis.* Memorias Do Instituto Oswaldo Cruz, 1997. **92**(5): p. 581-582.
65. Ravel, C., et al., *First report of genetic hybrids between two very divergent Leishmania species: Leishmania infantum and Leishmania major.* International Journal for Parasitology, 2006. **36**(13): p. 1383-1388.
66. Van den Broeck, F., et al., *Ecological divergence and hybridization of Neotropical Leishmania parasites . doi: 10.1073/pnas.1920136117. .* Proc Natl Acad Sci U S A., 2020 Oct 6. **117**(40): p. 25159-25168.
67. Chargui, N., et al., *Population structure of Tunisian Leishmania infantum and evidence for the existence of hybrids and gene flow between genetically different populations.* International Journal for Parasitology, 2009. **39**(7): p. 801-811.
68. Rougeron, V., et al., *Extreme inbreeding in Leishmania braziliensis.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(25): p. 10224-10229.
69. Kay, C., et al., *Signatures of hybridization in Trypanosoma brucei.* PLoS Pathog, 2022. **18**(2): p. e1010300.
70. Gibson, W., M. Crow, and J. Kearns, *Kinetoplast DNA minicircles are inherited from both parents in genetic crosses of Trypanosoma brucei.* Parasitology Research, 1997. **83**(5): p. 483-488.
71. Turner, C.M., et al., *Trypanosoma brucei: inheritance of kinetoplast DNA maxicircles in a genetic cross and their segregation during vegetative growth.* Exp Parasitol, 1995. **80**(2): p. 234-41.
72. Ferreira, T.R., et al., *Self-Hybridization in Leishmania major.* mBio, 2022. **13**(6): p. e0285822.
73. Wadsworth, E., *Bi-parental inheritance of kinetoplast DNA following sexual reproduction maintains mitochondrial genome complexity in Trypanosoma and Leishmania parasites*, in *School of Biological Sciences*. 2020, The University of Edinburgh. p. 228.
74. Gibson, W., et al., *Genetic recombination between human and animal parasites creates novel strains of human pathogen.* PLoS Negl Trop Dis, 2015. **9**(3): p. e0003665.
75. Goodhead, I., et al., *Whole-genome sequencing of Trypanosoma brucei reveals introgression between subspecies that is associated with virulence.* mBio, 2013. **4**(4).
76. Gibson, W., *Will the real Trypanosoma brucei rhodesiense please step forward?* Trends Parasitol, 2002. **18**(11): p. 486-90.
77. Simo, G., et al., *Population genetic structure of Central African Trypanosoma brucei gambiense isolates using microsatellite DNA markers.* Infect Genet Evol, 2010. **10**(1): p. 68-76.
78. Koffi, M., et al., *Population genetics of Trypanosoma brucei gambiense, the agent of sleeping sickness in Western Africa.* Proc Natl Acad Sci U S A, 2009. **106**(1): p. 209-14.
79. Weir, W., et al., *Population genomics reveals the origin and asexual evolution of human infective trypanosomes.* Elife, 2016. **5**: p. e11473.
80. Geerts, M., et al., *Deep kinetoplast genome analyses result in a novel molecular assay for detecting Trypanosoma brucei gambiense-specific minicircles.* NAR Genom Bioinform, 2022. **4**(4): p. lqac081.
81. Cooper, S., et al., *Organization of minicircle cassettes and guide RNA genes in Trypanosoma brucei.* RNA, 2022. **28**(7): p. 972-992.
82. Hoare, C.A., *The trypanosomes of mammals. A zoological monograph*. 1972: Blackwell Scientific Publications, 5 Alfred Street, Oxford. xvii + 749 pp.
83. Brun, R., H. Hecker, and Z.R. Lun, *Trypanosoma evansi and T. equiperdum: distribution, biology, treatment and phylogenetic relationship (a review).* Vet Parasitol, 1998. **79**(2): p. 95-107.

84. Desquesnes, M., et al., *A review on the diagnosis of animal trypanosomoses.* Parasit Vectors, 2022. **15**(1): p. 64.
85. Oldrieve, G.R., et al., *Mechanisms of life cycle simplification in African trypanosomes.* bioRxiv, 2024: p. 2024.07.12.603250.
86. Artama, W.T., M.W. Agey, and J.E. Donelson, *DNA comparisons of Trypanosoma evansi (Indonesia) and Trypanosoma brucei spp.* Parasitology, 1992. **104 Pt 1**: p. 67-74.
87. Borst, P., F. Fase-Fowler, and W.C. Gibson, *Kinetoplast DNA of Trypanosoma evansi.* Mol Biochem Parasitol, 1987. **23**(1): p. 31-8.
88. Barrois, M., G. Riou, and F. Galibert, *Complete nucleotide sequence of minicircle kinetoplast DNA from Trypanosoma equiperdum.* Proc Natl Acad Sci U S A, 1981. **78**(6): p. 3323-7.
89. Domingo, G.J., et al., *Dyskinetoplastic Trypanosoma brucei contains functional editing complexes.* Eukaryot Cell, 2003. **2**(3): p. 569-77.
90. Cuypers, B., et al., *Genome-Wide SNP Analysis Reveals Distinct Origins of Trypanosoma evansi and Trypanosoma equiperdum.* Genome Biol Evol, 2017. **9**(8): p. 1990-1997.
91. Carnes, J., et al., *Genome and phylogenetic analyses of Trypanosoma evansi reveal extensive similarity to T. brucei and multiple independent origins for dyskinetoplasty.* PLoS Negl Trop Dis, 2015. **9**(1): p. e3404.
92. Kay, C., T.A. Williams, and W. Gibson, *Mitochondrial DNAs provide insight into trypanosome phylogeny and molecular evolution.* BMC Evol Biol, 2020. **20**(1): p. 161.
93. Oldrieve, G., et al., *Monomorphic Trypanozoon: towards reconciling phylogeny and pathologies.* Microb Genom, 2021. **7**(8).
94. Lai, D.H., et al., *Adaptations of Trypanosoma brucei to gradual loss of kinetoplast DNA: Trypanosoma equiperdum and Trypanosoma evansi are petite mutants of T. brucei.* Proc Natl Acad Sci U S A, 2008. **105**(6): p. 1999-2004.
95. Dean, S., et al., *Single point mutations in ATP synthase compensate for mitochondrial genome loss in trypanosomes.* Proc Natl Acad Sci U S A, 2013. **110**(36): p. 14741-6.
96. Schnaufer, A., G.J. Domingo, and K. Stuart, *Natural and induced dyskinetoplastic trypanosomatids: how to live without mitochondrial DNA.* Int J Parasitol, 2002. **32**(9): p. 1071-84.
97. Feasey, N., et al., *Neglected tropical diseases.* Br Med Bull, 2010. **93**: p. 179-200.
98. Houweling, T.A., et al., *Socioeconomic Inequalities in Neglected Tropical Diseases: A Systematic Review.* PLoS Negl Trop Dis, 2016. **10**(5): p. e0004546.
99. Llovet, I., G. Dinardi, and F.G. De Maio, *Mitigating social and health inequities: community participation and Chagas disease in rural Argentina.* Glob Public Health, 2011. **6**(4): p. 371-84.
100. Buscher, P., et al., *Human African trypanosomiasis.* Lancet, 2017. **390**(10110): p. 2397-2409.
101. Franco, J.R., et al., *The elimination of human African trypanosomiasis: Achievements in relation to WHO road map targets for 2020.* PLoS Negl Trop Dis, 2022. **16**(1): p. e0010047.
102. Pays, E., et al., *The molecular arms race between African trypanosomes and humans.* Nat Rev Microbiol, 2014. **12**(8): p. 575-84.
103. Vanhollebeke, B., et al., *Distinct roles of haptoglobin-related protein and apolipoprotein L-I in trypanolysis by human serum.* Proc Natl Acad Sci U S A, 2007. **104**(10): p. 4118-23.
104. Pays, E., *Apolipoprotein-L1 (APOL1): From Sleeping Sickness to Kidney Disease.* Cells, 2024. **13**(20).
105. Trindade, S., et al., *Trypanosoma brucei Parasites Occupy and Functionally Adapt to the Adipose Tissue in Mice.* Cell Host Microbe, 2016. **19**(6): p. 837-48.
106. Mehlitz, D. and D.H. Molyneux, *The elimination of Trypanosoma brucei gambiense? Challenges of reservoir hosts and transmission cycles: Expect the unexpected.* Parasite Epidemiol Control, 2019. **6**: p. e00113.
107. Alfituri, O.A., et al., *To the Skin and Beyond: The Immune Response to African Trypanosomes as They Enter and Exit the Vertebrate Host.* Frontiers in Immunology, 2020. **11**.

108. Kristensson, K., et al., *African trypanosome infections of the nervous system: parasite entry and effects on sleep and synaptic functions.* Prog Neurobiol, 2010. **91**(2): p. 152-71.
109. Steverding, D., *The history of African trypanosomiasis.* Parasit Vectors, 2008. **1**(1): p. 3.
110. Sodeman, W.A., Jr., *A note on the early history of African trypanosomiasis.* Am J Trop Med Hyg, 1974. **23**(4): p. 712-3.
111. Capewell, P., et al., *Human and animal Trypanosomes in Cote d'Ivoire form a single breeding population.* PLoS One, 2013. **8**(7): p. e67852.
112. Arenas, M., et al., *Trypanosoma cruzi genotypes of insect vectors and patients with Chagas of Chile studied by means of cytochrome b gene sequencing, minicircle hybridization, and nuclear gene polymorphisms.* Vector Borne Zoonotic Dis, 2012. **12**(3): p. 196-205.
113. Olivera, M.J., et al., *Addressing Chagas disease from a One Health perspective: risk factors, lessons learned and prevention of oral transmission outbreaks in Colombia.* Sci One Health, 2024. **3**: p. 100066.
114. Gomes, C., et al., *American trypanosomiasis and Chagas disease: Sexual transmission.* Int J Infect Dis, 2019. **81**: p. 81-84.
115. Cevallos, A.M. and R. Hernandez, *Chagas' disease: pregnancy and congenital transmission.* Biomed Res Int, 2014. **2014**: p. 401864.
116. Santos, E. and L. Menezes Falcao, *Chagas cardiomyopathy and heart failure: From epidemiology to treatment.* Rev Port Cardiol (Engl Ed), 2020. **39**(5): p. 279-289.
117. Swett, M.C., et al., *Chagas Disease: Epidemiology, Diagnosis, and Treatment.* Curr Cardiol Rep, 2024.
118. Akhoundi, M., et al., *A Historical Overview of the Classification, Evolution, and Dispersion of Leishmania Parasites and Sandflies.* PLoS Negl Trop Dis, 2016. **10**(3): p. e0004349.
119. McGwire, B.S. and A.R. Satoskar, *Leishmaniasis: clinical syndromes and treatment.* QJM, 2014. **107**(1): p. 7-14.
120. Alvar, J., et al., *Leishmaniasis worldwide and global estimates of its incidence.* PLoS One, 2012. **7**(5): p. e35671.
121. Okwor, I. and J.E. Uzonna, *The immunology of Leishmania/HIV co-infection.* Immunol Res, 2013. **56**(1): p. 163-71.
122. Jain, P., V. Goyal, and R. Agrawal, *An atypical Trypanosoma lewisi infection in a 22-day-old neonate from India: An emergent zoonosis.* Indian J Pathol Microbiol, 2023. **66**(1): p. 199-201.
123. Truc, P., et al., *Atypical human infections by animal trypanosomes.* PLoS Negl Trop Dis, 2013. **7**(9): p. e2256.
124. Truc, P., et al., *[Atypical human trypanosomoses].* Med Sante Trop, 2014. **24**(3): p. 249-52.
125. Kumar, R., et al., *Atypical human trypanosomosis: Potentially emerging disease with lack of understanding.* Zoonoses Public Health, 2022. **69**(4): p. 259-276.
126. *Guidelines for the treatment of human African trypanosomiasis*, in *Guidelines for the treatment of human African trypanosomiasis*. 2024: Geneva.
127. Fairlamb, A.H., *Chemotherapy of human African trypanosomiasis: current and future prospects.* Trends Parasitol, 2003. **19**(11): p. 488-94.
128. Steverding, D., *The development of drugs for treatment of sleeping sickness: a historical review.* Parasit Vectors, 2010. **3**(1): p. 15.
129. Torreele, E., et al., *Fexinidazole--a new oral nitroimidazole drug candidate entering clinical development for the treatment of sleeping sickness.* PLoS Negl Trop Dis, 2010. **4**(12): p. e923.
130. Agency, E.M. *Product information 14/12/2023: fexinidazole Winthrop - opinion on medicine for use outside EU. Amsterdam*. 2024 [cited 2024 2 May 2024]; Available from: https://www.ema.europa.eu/en/opinion-medicine-use-outside-EU/human/fexinidazole-winthrop.
131. Hidalgo, J., et al., *Efficacy and Toxicity of Fexinidazole and Nifurtimox Plus Eflornithine in the Treatment of African Trypanosomiasis: A Systematic Review.* Cureus, 2021. **13**(8): p. e16881.
132. Deeks, E.D., *Fexinidazole: First Global Approval.* Drugs, 2019. **79**(2): p. 215-220.

133. Bernhard, S., et al., *Fexinidazole for Human African Trypanosomiasis, the Fruit of a Successful Public-Private Partnership.* Diseases, 2022. **10**(4).
134. Wyllie, S., et al., *Nitroheterocyclic drug resistance mechanisms in Trypanosoma brucei.* J Antimicrob Chemother, 2016. **71**(3): p. 625-34.
135. Organization, P.A.H. *Guidelines for the diagnosis and treatment of Chagas disease*. 2019.
136. Pinazo, M.J., et al., *Interventions to bring comprehensive care to people with Chagas disease: Experiences in Bolivia, Argentina and Colombia.* Acta Trop, 2020. **203**: p. 105290.
137. Aronson, N., et al., *Diagnosis and Treatment of Leishmaniasis: Clinical Practice Guidelines by the Infectious Diseases Society of America (IDSA) and the American Society of Tropical Medicine and Hygiene (ASTMH).* Am J Trop Med Hyg, 2017. **96**(1): p. 24-45.
138. Moore, E.M. and D.N. Lockwood, *Treatment of visceral leishmaniasis.* J Glob Infect Dis, 2010. **2**(2): p. 151-8.
139. Bekhit, A.A., et al., *Leishmania treatment and prevention: Natural and synthesized drugs.* Eur J Med Chem, 2018. **160**: p. 229-244.
140. Altamura, F., et al., *The current drug discovery landscape for trypanosomiasis and leishmaniasis: Challenges and strategies to identify drug targets.* Drug Dev Res, 2022. **83**(2): p. 225-252.
141. Pinheiro, A.C. and M.V.N. de Souza, *Current leishmaniasis drug discovery.* RSC Med Chem, 2022. **13**(9): p. 1029-1043.
142. Horn, D., *Antigenic variation in African trypanosomes.* Mol Biochem Parasitol, 2014. **195**(2): p. 123-9.
143. Bangs, J.D., *Evolution of Antigenic Variation in African Trypanosomes: Variant Surface Glycoprotein Expression, Structure, and Function.* Bioessays, 2018. **40**(12).
144. Pereira, S.S., A.P. Jackson, and L.M. Figueiredo, *Evolution of the variant surface glycoprotein family in African trypanosomes.* Trends in Parasitology, 2022. **38**(1): p. 23-36.
145. Freymann, D., et al., *2.9 a Resolution Structure of the N-Terminal Domain of a Variant Surface Glycoprotein from Trypanosoma-Brucei.* Journal of Molecular Biology, 1990. **216**(1): p. 141-160.
146. Bartossek, T., et al., *Structural basis for the shielding function of the dynamic trypanosome variant surface glycoprotein coat.* Nature Microbiology, 2017. **2**(11): p. 1523-1532.
147. Umaer, K., et al., *Dynamic, variable oligomerization and the trafficking of variant surface glycoproteins of.* Traffic, 2021. **22**(8): p. 274-283.
148. Pinger, J., et al., *African trypanosomes evade immune clearance by glycosylation of the VSG surface coat.* Nature Microbiology, 2018. **3**(8): p. 932-938.
149. Dakovic, S., et al., *A structural classification of the variant surface glycoproteins of the African trypanosome.* PLoS Negl Trop Dis, 2023. **17**(9): p. e0011621.
150. Degreef, C. and R. Hamers, *The Serum Resistance-Associated (Sra) Gene of Trypanosoma-Brucei-Rhodesiense Encodes a Variant Surface Glycoprotein-Like Protein.* Molecular and Biochemical Parasitology, 1994. **68**(2): p. 277-284.
151. Xong, H.V., et al., *A expression site-associated gene confers resistance to human serum in.* Cell, 1998. **95**(6): p. 839-846.
152. Stephens, N.A. and S.L. Hajduk, *Endosomal Localization of the Serum Resistance-Associated Protein in African Trypanosomes Confers Human Infectivity.* Eukaryotic Cell, 2011. **10**(8): p. 1023-1033.
153. Vanhamme, L., et al., *Apolipoprotein L-I is the trypanosome lytic factor of human serum.* Nature, 2003. **422**(6927): p. 83-87.
154. Oli, M.W., et al., *Serum resistance-associated protein blocks lysosomal targeting of trypanosome lytic factor in Trypanosoma brucei.* Eukaryotic Cell, 2006. **5**(1): p. 132-139.
155. Berberof, M., D. Perez-Morga, and E. Pays, *A receptor-like flagellar pocket glycoprotein specific to Trypanosoma brucei gambiense.* Mol Biochem Parasitol, 2001. **113**(1): p. 127-38.

156. Capewell, P., et al., *The TgsGP gene is essential for resistance to human serum in Trypanosoma brucei gambiense.* PLoS Pathog, 2013. **9**(10): p. e1003686.
157. Cisarovsky, G., P. Schmid-Hempel, and B.M. Sadd, *Robustness of the outcome of adult bumblebee infection with a trypanosome parasite after varied parasite exposures during larval development.* J Evol Biol, 2012. **25**(6): p. 1053-9.
158. Gonzalez, E., R. Molina, and M. Jimenez, *Rabbit trypanosome detection in Phlebotomus perniciosus sand flies from the leishmaniasis outbreak in Madrid, Spain.* Acta Trop, 2018. **187**: p. 201-206.
159. Mossaad, E., et al., *The incrimination of three trypanosome species in clinically affected German shepherd dogs in Sudan.* Parasitol Res, 2017. **116**(11): p. 2921-2925.
160. Fakae, B.B. and S.N. Chiejina, *The prevalence of concurrent trypanosome and gastrointestinal nematode infections in west African dwarf sheep and goats in Nsukka area of eastern Nigeria.* Vet Parasitol, 1993. **49**(2-4): p. 313-8.
161. Lobsiger, L., et al., *An autochthonous case of cutaneous bovine leishmaniasis in Switzerland.* Vet Parasitol, 2010. **169**(3-4): p. 408-14.
162. Muller, N., et al., *Occurrence of Leishmania sp. in cutaneous lesions of horses in Central Europe.* Vet Parasitol, 2009. **166**(3-4): p. 346-51.
163. Hodo, C.L., et al., *Trypanosome species, including Trypanosoma cruzi, in sylvatic and peridomestic bats of Texas, USA.* Acta Trop, 2016. **164**: p. 259-266.
164. Hanotte, O., et al., *Mapping of quantitative trait loci controlling trypanotolerance in a cross of tolerant West African N'Dama and susceptible East African Boran cattle.* Proc Natl Acad Sci U S A, 2003. **100**(13): p. 7443-8.
165. Chadenga, V., *Epidemiology and control of trypanosomosis.* Onderstepoort J Vet Res, 1994. **61**(4): p. 385-90.
166. Odeniran, P.O., et al., *Bovine and small ruminant African animal trypanosomiasis in Nigeria - A review.* Vet Parasitol Reg Stud Reports, 2018. **13**: p. 5-13.
167. Kristjanson, P.M., et al., *Measuring the costs of African animal trypanosomosis, the potential benefits of control and returns to research.* Agricultural Systems, 1999. **59**(1): p. 79-98.
168. Abro, Z., et al., *The potential economic benefits of controlling trypanosomiasis using waterbuck repellent blend in sub-Saharan Africa.* PLoS One, 2021. **16**(7): p. e0254558.
169. Staal, S., Poole, J., Baltenweck, I., Mwacharo, J., Notenbaert, A., Randolph, T., Thorpe, W., Nzuma, J. and Herrero, M. *Targeting strategic investment in livestock development as a vehicle for rural livelihoods*. 2009; Available from: https://hdl.handle.net/10568/35206.
170. Shaw, A.P., *Assessing the economics of animal trypanosomosis in Africa--history and current perspectives.* Onderstepoort J Vet Res, 2009. **76**(1): p. 27-32.
171. Meyer, A., et al., *Past and Ongoing Tsetse and Animal Trypanosomiasis Control Operations in Five African Countries: A Systematic Review.* PLoS Negl Trop Dis, 2016. **10**(12): p. e0005247.
172. Richards, S., et al., *Pharma to farmer: field challenges of optimizing trypanocide use in African animal trypanosomiasis.* Trends Parasitol, 2021. **37**(9): p. 831-843.
173. Bouchet, A., et al., *[Parasitism in zebus in the west of the Central African Republic. I. Parasitism in suckling calves].* Rev Elev Med Vet Pays Trop, 1969. **22**(3): p. 373-83.
174. Desrotour, J., et al., *[Trypanotolerant cattle: their breeding in the Central African Republic].* Rev Elev Med Vet Pays Trop, 1967. **20**(4): p. 589-94.
175. Graber, M., et al., *[Parasitism in zebus in the west of the Central African Republic. 2. Parasitism in steers and adult zebus].* Rev Elev Med Vet Pays Trop, 1969. **22**(4): p. 509-19.
176. Yvore, P., R. Lacotte, and P. Finelle, *[Study on the biology and ecology of Glossina fusca congolensis Newst and Evans in the Central African Republic. I. Influence of the climate and vegetation on the distribution and density of the Glossinae].* Rev Elev Med Vet Pays Trop, 1965. **18**(2): p. 151-64.
177. Desquesnes, M. and C. Gutiérrez, *Animal Trypanosomosis: An important constraint fro livestock in tropical and sub-tropical regions*. 2011. p. 127-144.

178. Giordani, F., et al., *The animal trypanosomiases and their chemotherapy: a review.* Parasitology, 2016. **143**(14): p. 1862-1889.
179. Desquesnes, M., et al., *Trypanosoma evansi and surra: a review and perspectives on transmission, epidemiology and control, impact, and zoonotic aspects.* Biomed Res Int, 2013. **2013**: p. 321237.
180. Sumba, A.L., S. Mihok, and F.A. Oyieke, *Mechanical transmission of Trypanosoma evansi and T. congolense by Stomoxys niger and S. taeniatus in a laboratory mouse model.* Med Vet Entomol, 1998. **12**(4): p. 417-22.
181. Aregawi, W.G., et al., *Systematic review and meta-analysis on the global distribution, host range, and prevalence of Trypanosoma evansi.* Parasit Vectors, 2019. **12**(1): p. 67.
182. Desquesnes, M., et al., *Trypanosoma evansi and surra: a review and perspectives on origin, history, distribution, taxonomy, morphology, hosts, and pathogenic effects.* Biomed Res Int, 2013. **2013**: p. 194176.
183. Claes, F., et al., *Trypanosoma equiperdum: master of disguise or historical mistake?* Trends Parasitol, 2005. **21**(7): p. 316-21.
184. Gizaw, Y., M. Megersa, and T. Fayera, *Dourine: a neglected disease of equids.* Trop Anim Health Prod, 2017. **49**(5): p. 887-897.
185. Yasine, A., et al., *Reduction of Trypanosoma equiperdum from equine semen by single layer centrifugation.* Exp Parasitol, 2019. **200**: p. 79-83.
186. Sanchez, E., et al., *Molecular characterization and classification of Trypanosoma spp. Venezuelan isolates based on microsatellite markers and kinetoplast maxicircle genes.* Parasit Vectors, 2015. **8**: p. 536.
187. Perrone, T.M., et al., *Molecular profiles of Venezuelan isolates of Trypanosoma sp. by random amplified polymorphic DNA method.* Vet Parasitol, 2009. **161**(3-4): p. 194-200.
188. Hoare, C.A. and F.G. Wallace, *Developmental Stages of Trypanosomatid Flagellates - a New Terminology.* Nature, 1966. **212**(5068): p. 1385-+.
189. Sinclair, A.N. and C.L. de Graffenried, *More than Microtubules: The Structure and Function of the Subpellicular Array in Trypanosomatids.* Trends in Parasitology, 2019. **35**(10): p. 760-777.
190. Field, M.C. and M. Carrington, *The trypanosome flagellar pocket.* Nature Reviews Microbiology, 2009. **7**(11): p. 775-786.
191. Amodeo, S., et al., *Characterization of the novel mitochondrial genome segregation factor TAP110 in Trypanosoma brucei.* J Cell Sci, 2021. **134**(5).
192. Schimanski, B., et al., *p166 links membrane and intramitochondrial modules of the trypanosomal tripartite attachment complex.* PLoS Pathog, 2022. **18**(6): p. e1010207.
193. Sharma, R., et al., *Asymmetric cell division as a route to reduction in cell length and change in cell morphology in trypanosomes.* Protist, 2008. **159**(1): p. 137-151.
194. Langousis, G. and K.L. Hill, *Motility and more: the flagellum of.* Nature Reviews Microbiology, 2014. **12**(7): p. 505-518.
195. Stuart, K., *Kinetoplast DNA OF Trypanosoma brucei: physical map of the maxicircle.* Plasmid, 1979. **2**(4): p. 520-8.
196. Jensen, R.E. and P.T. Englund, *Network news: the replication of kinetoplast DNA.* Annu Rev Microbiol, 2012. **66**: p. 473-91.
197. Yaffe, N., et al., *Direct monitoring of the stepwise condensation of kinetoplast DNA networks.* Sci Rep, 2021. **11**(1): p. 1501.
198. Shapiro, T.A. and P.T. Englund, *The structure and replication of kinetoplast DNA.* Annu Rev Microbiol, 1995. **49**: p. 117-43.
199. Ferguson, M.L., et al., *Kinetoplast DNA replication: mechanistic differences between Trypanosoma brucei and Crithidia fasciculata.* J Cell Biol, 1994. **126**(3): p. 631-9.
200. Lukes, J., et al., *Kinetoplast DNA network: evolution of an improbable structure.* Eukaryot Cell, 2002. **1**(4): p. 495-502.

201. Urrea, D.A., O. Triana-Chavez, and J.F. Alzate, *Mitochondrial genomics of human pathogenic parasite Leishmania (Viannia) panamensis.* PeerJ, 2019. **7**: p. e7235.
202. Eperon, I.C., et al., *The major transcripts of the kinetoplast DNA of Trypanosoma brucei are very small ribosomal RNAs.* Nucleic Acids Res, 1983. **11**(1): p. 105-25.
203. de la Cruz, V.F., N. Neckelmann, and L. Simpson, *Sequences of six genes and several open reading frames in the kinetoplast maxicircle DNA of Leishmania tarentolae.* J Biol Chem, 1984. **259**(24): p. 15136-47.
204. Read, L.K., et al., *Sequences of three Trypanosoma congolense maxicircle genes allow prediction of regions encoding transcripts that undergo extensive RNA editing.* Mol Biochem Parasitol, 1993. **60**(2): p. 337-41.
205. Duarte, M. and A.M. Tomas, *The mitochondrial complex I of trypanosomatids--an overview of current knowledge.* J Bioenerg Biomembr, 2014. **46**(4): p. 299-311.
206. Koslowsky, D.J., et al., *The MURF3 gene of T. brucei contains multiple domains of extensive editing and is homologous to a subunit of NADH dehydrogenase.* Cell, 1990. **62**(5): p. 901-11.
207. Souza, A.E., P.J. Myler, and K. Stuart, *Maxicircle CR1 transcripts of Trypanosoma brucei are edited and developmentally regulated and encode a putative iron-sulfur protein homologous to an NADH dehydrogenase subunit.* Mol Cell Biol, 1992. **12**(5): p. 2100-7.
208. Gerasimov, E.S., et al., *From cryptogene to gene? ND8 editing domain reduction in insect trypanosomatids.* Eur J Protistol, 2012. **48**(3): p. 185-93.
209. Kannan, S. and G. Burger, *Unassigned MURF1 of kinetoplastids codes for NADH dehydrogenase subunit 2.* BMC Genomics, 2008. **9**: p. 455.
210. Benne, R., et al., *The nucleotide sequence of a segment of Trypanosoma brucei mitochondrial maxi-circle DNA that contains the gene for apocytochrome b and some unusual unassigned reading frames.* Nucleic Acids Res, 1983. **11**(20): p. 6925-41.
211. Benne, R., et al., *Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA.* Cell, 1986. **46**(6): p. 819-26.
212. Feagin, J.E., J.M. Abraham, and K. Stuart, *Extensive editing of the cytochrome c oxidase III transcript in Trypanosoma brucei.* Cell, 1988. **53**(3): p. 413-22.
213. Golden, D.E. and S.L. Hajduk, *The 3'-untranslated region of cytochrome oxidase II mRNA functions in RNA editing of African trypanosomes exclusively as a cis guide RNA.* RNA, 2005. **11**(1): p. 29-37.
214. Hensgens, L.A., et al., *The sequence of the gene for cytochrome c oxidase subunit I, a frameshift containing gene for cytochrome c oxidase subunit II and seven unassigned reading frames in Trypanosoma brucei mitochondrial maxi-circle DNA.* Nucleic Acids Res, 1984. **12**(19): p. 7327-44.
215. Bhat, G.J., et al., *An extensively edited mitochondrial transcript in kinetoplastids encodes a protein homologous to ATPase subunit 6.* Cell, 1990. **61**(5): p. 885-94.
216. Maslov, D.A., et al., *An intergenic G-rich region in Leishmania tarentolae kinetoplast maxicircle DNA is a pan-edited cryptogene encoding ribosomal protein S12.* Mol Cell Biol, 1992. **12**(1): p. 56-67.
217. Corell, R.A., P. Myler, and K. Stuart, *Trypanosoma brucei mitochondrial CR4 gene encodes an extensively edited mRNA with completely edited sequence only in bloodstream forms.* Mol Biochem Parasitol, 1994. **64**(1): p. 65-74.
218. Myler, P.J., et al., *Structural organization of the maxicircle variable region of Trypanosoma brucei: identification of potential replication origins and topoisomerase II binding sites.* Nucleic Acids Res, 1993. **21**(3): p. 687-94.
219. de Vries, B.F., et al., *The variable region of the Trypanosoma brucei kinetoplast maxicircle: sequence and transcript analysis of a repetitive and a non-repetitive fragment.* Mol Biochem Parasitol, 1988. **27**(1): p. 71-82.

220. Gibson, W., P. Borst, and F. Fase-Fowler, *Further analysis of intraspecific variation in Trypanosoma brucei using restriction site polymorphisms in the maxi-circle of kinetoplast DNA.* Mol Biochem Parasitol, 1985. **15**(1): p. 21-36.
221. Koslowsky, D.J. and G. Yahampath, *Mitochondrial mRNA 3' cleavage/polyadenylation and RNA editing in Trypanosoma brucei are independent events.* Mol Biochem Parasitol, 1997. **90**(1): p. 81-94.
222. Parsons, M., et al., *Advancing Trypanosoma brucei genome annotation through ribosome profiling and spliced leader mapping.* Mol Biochem Parasitol, 2015. **202**(2): p. 1-10.
223. Read, L.K., J. Lukes, and H. Hashimi, *Trypanosome RNA editing: the complexity of getting U in and taking U out.* Wiley Interdiscip Rev RNA, 2016. **7**(1): p. 33-51.
224. Sement, F.M., et al., *Transcription initiation defines kinetoplast RNA boundaries.* Proc Natl Acad Sci U S A, 2018. **115**(44): p. E10323-E10332.
225. Cooper, S., et al., *Assembly and annotation of the mitochondrial minicircle genome of a differentiation-competent strain of Trypanosoma brucei.* Nucleic Acids Res, 2019. **47**(21): p. 11304-11325.
226. Suematsu, T., et al., *Antisense Transcripts Delimit Exonucleolytic Activity of the Mitochondrial 3' Processome to Generate Guide RNAs.* Mol Cell, 2016. **61**(3): p. 364-378.
227. Pollard, V.W., et al., *Organization of minicircle genes for guide RNAs in Trypanosoma brucei.* Cell, 1990. **63**(4): p. 783-90.
228. Read, L.K. and K. Stuart, *Conservation of gRNA gene cassette structure in African trypanosomes despite divergence in the defining flanking repeats.* Mol Biochem Parasitol, 1993. **60**(2): p. 333-5.
229. Ray, D.S., *Conserved sequence blocks in kinetoplast minicircles from diverse species of trypanosomes*, in *Mol Cell Biol*. 1989. p. 1365-7.
230. Nasir, A., G.A. Cook, and J.E. Donelson, *Sequences of two kinetoplast minicircle DNAs of Trypanosoma (Nannomonas) congolense.* Mol Biochem Parasitol, 1987. **24**(3): p. 295-300.
231. Li, S.J., et al., *Novel organization of mitochondrial minicircles and guide RNAs in the zoonotic pathogen Trypanosoma lewisi.* Nucleic Acids Res, 2020. **48**(17): p. 9747-9761.
232. Blom, D., et al., *Mitochondrial minicircles in the free-living bodonid Bodo saltans contain two gRNA gene cassettes and are not found in large networks.* RNA, 2000. **6**(1): p. 121-35.
233. Krasnow, M.A. and N.R. Cozzarelli, *Catenation of DNA rings by topoisomerases. Mechanism of control by spermidine.* J Biol Chem, 1982. **257**(5): p. 2687-93.
234. Borst, P., *Why kinetoplast DNA networks?* Trends Genet, 1991. **7**(5): p. 139-41.
235. Simpson, L. and P. Braly, *Synchronization of Leishmania tarentolae by hydroxyurea.* J Protozool, 1970. **17**(4): p. 511-7.
236. Amodeo, S., I. Bregy, and T. Ochsenreiter, *Mitochondrial genome maintenance-the kinetoplast story.* FEMS Microbiol Rev, 2023. **47**(6).
237. Kapeller, I., et al., *Interactions of a replication initiator with histone H1-like proteins remodel the condensed mitochondrial genome.* J Biol Chem, 2011. **286**(47): p. 40566-74.
238. Povelones, M.L., *Beyond replication: division and segregation of mitochondrial DNA in kinetoplastids.* Mol Biochem Parasitol, 2014. **196**(1): p. 53-60.
239. Jackson, A.P., *Genome evolution in trypanosomatid parasites.* Parasitology, 2015. **142 Suppl 1**(Suppl 1): p. S40-56.
240. Hoeijmakers, J.H. and P.J. Weijers, *The segregation of kinetoplast DNA networks in Trypanosoma brucei.* Plasmid, 1980. **4**(1): p. 97-116.
241. Savill, N.J. and P.G. Higgs, *A theoretical study of random segregation of minicircles in trypanosomatids.* Proc Biol Sci, 1999. **266**(1419): p. 611-20.
242. Simpson, A.M. and L. Simpson, *Pulse-labeling of kinetoplast DNA: localization of 2 sites of synthesis within the networks and kinetics of labeling of closed minicircles.* J Protozool, 1976. **23**(4): p. 583-7.

243. Thiemann, O.H., D.A. Maslov, and L. Simpson, *Disruption of RNA editing in Leishmania tarentolae by the loss of minicircle-encoded guide RNA genes.* EMBO J, 1994. **13**(23): p. 5689-700.
244. Savill, N.J. and P.G. Higgs, *Redundant and non-functional guide RNA genes in Trypanosoma brucei are a consequence of multiple genes per minicircle.* Gene, 2000. **256**(1-2): p. 245-52.
245. Knoop, V., *When you can't trust the DNA: RNA editing changes transcript sequences.* Cell Mol Life Sci, 2011. **68**(4): p. 567-86.
246. Bazak, L., et al., *A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes.* Genome Res, 2014. **24**(3): p. 365-76.
247. Liscovitch-Brauer, N., et al., *Trade-off between Transcriptome Plasticity and Genome Evolution in Cephalopods.* Cell, 2017. **169**(2): p. 191-202 e11.
248. Bar-Yaacov, D., Y. Pilpel, and O. Dahan, *RNA editing in bacteria: occurrence, regulation and significance.* RNA Biol, 2018. **15**(7): p. 863-867.
249. Simpson, R.M., et al., *High-throughput sequencing of partially edited trypanosome mRNAs reveals barriers to editing progression and evidence for alternative editing.* RNA, 2016. **22**(5): p. 677-95.
250. Simpson, L. and D.A. Maslov, *Evolution of the U-insertion/deletion RNA editing in mitochondria of kinetoplastid protozoa.* Ann N Y Acad Sci, 1999. **870**: p. 190-205.
251. Blom, D., et al., *RNA editing in the free-living bodonid Bodo saltans.* Nucleic Acids Res, 1998. **26**(5): p. 1205-13.
252. Maslov, D.A. and L. Simpson, *RNA editing and mitochondrial genomic organization in the cryptobiid kinetoplastid protozoan Trypanoplasma borreli.* Mol Cell Biol, 1994. **14**(12): p. 8174-82.
253. Lukes, J., et al., *Novel pattern of editing regions in mitochondrial transcripts of the cryptobiid Trypanoplasma borreli.* EMBO J, 1994. **13**(21): p. 5086-98.
254. Gerasimov, E.S., et al., *Mitochondrial RNA editing in Trypanoplasma borreli: New tools, new revelations.* Comput Struct Biotechnol J, 2022. **20**: p. 6388-6402.
255. Leung, S.S. and D.J. Koslowsky, *RNA editing in Trypanosoma brucei: characterization of gRNA U-tail interactions with partially edited mRNA substrates.* Nucleic Acids Res, 2001. **29**(3): p. 703-9.
256. McManus, M.T., et al., *Trypanosoma brucei guide RNA poly(U) tail formation is stabilized by cognate mRNA.* Mol Cell Biol, 2000. **20**(3): p. 883-91.
257. Stoltzfus, A., *Constructive neutral evolution: exploring evolutionary theory's curious disconnect.* Biol Direct, 2012. **7**: p. 35.
258. Blum, B., N. Bakalara, and L. Simpson, *A model for RNA editing in kinetoplastid mitochondria: "guide" RNA molecules transcribed from maxicircle DNA provide the edited information.* Cell, 1990. **60**(2): p. 189-98.
259. Maslov, D.A. and L. Simpson, *The polarity of editing within a multiple gRNA-mediated domain is due to formation of anchors for upstream gRNAs by downstream editing.* Cell, 1992. **70**(3): p. 459-67.
260. Gao, G., et al., *Guide RNAs of the recently isolated LEM125 strain of Leishmania tarentolae: an unexpected complexity.* RNA, 2001. **7**(9): p. 1335-47.
261. Verner, Z., et al., *Complex I (NADH:ubiquinone oxidoreductase) is active in but non-essential for procyclic Trypanosoma brucei.* Mol Biochem Parasitol, 2011. **175**(2): p. 196-200.
262. Feagin, J.E., D.P. Jasmer, and K. Stuart, *Apocytochrome b and other mitochondrial DNA sequences are differentially expressed during the life cycle of Trypanosoma brucei.* Nucleic Acids Res, 1985. **13**(12): p. 4577-96.
263. McDermott, S.M., J. Carnes, and K. Stuart, *Identification by Random Mutagenesis of Functional Domains in KREPB5 That Differentially Affect RNA Editing between Life Cycle Stages of Trypanosoma brucei.* Mol Cell Biol, 2015. **35**(23): p. 3945-61.

264. Nolan, D.P. and H.P. Voorheis, *The mitochondrion in bloodstream forms of Trypanosoma brucei is energized by the electrogenic pumping of protons catalysed by the F1F0-ATPase.* Eur J Biochem, 1992. **209**(1): p. 207-16.
265. Gahura, O., C. Hierro-Yap, and A. Zikova, *Redesigned and reversed: architectural and functional oddities of the trypanosomal ATP synthase.* Parasitology, 2021. **148**(10): p. 1151-1160.
266. Aphasizhev, R. and I. Aphasizheva, *Uridine insertion/deletion editing in trypanosomes: a playground for RNA-guided information transfer.* Wiley Interdiscip Rev RNA, 2011. **2**(5): p. 669-85.
267. Weng, J., et al., *Guide RNA-binding complex from mitochondria of trypanosomatids.* Mol Cell, 2008. **32**(2): p. 198-209.
268. Aphasizhev, R. and I. Aphasizheva, *Mitochondrial RNA editing in trypanosomes: small RNAs in control.* Biochimie, 2014. **100**: p. 125-31.
269. Liu, S., et al., *Structural basis of gRNA stabilization and mRNA recognition in trypanosomal RNA editing.* Science, 2023. **381**(6653): p. eadg4725.
270. Dolce, L.G., et al., *Structural basis for guide RNA selection by the RESC1-RESC2 complex.* Nucleic Acids Res, 2023. **51**(9): p. 4602-4612.
271. McDermott, S.M., J. Carnes, and K. Stuart, *Editosome RNase III domain interactions are essential for editing and differ between life cycle stages in Trypanosoma brucei.* RNA, 2019. **25**(9): p. 1150-1163.
272. McDermott, S.M., et al., *Differential Editosome Protein Function between Life Cycle Stages of Trypanosoma brucei.* J Biol Chem, 2015. **290**(41): p. 24914-31.
273. Carnes, J., S.M. McDermott, and K. Stuart, *RNase III Domain of KREPB9 and KREPB10 Association with Editosomes in Trypanosoma brucei.* mSphere, 2018. **3**(1).
274. Carnes, J., et al., *Domain function and predicted structure of three heterodimeric endonuclease subunits of RNA editing catalytic complexes in Trypanosoma brucei.* Nucleic Acids Res, 2022. **50**(17): p. 10123-10139.
275. Cruz-Reyes, J., et al., *Dynamic RNA holo-editosomes with subcomplex variants: Insights into the control of trypanosome editing.* Wiley Interdiscip Rev RNA, 2018. **9**(6): p. e1502.
276. Aphasizhev, R., et al., *Isolation of a U-insertion/deletion editing complex from Leishmania tarentolae mitochondria.* EMBO J, 2003. **22**(4): p. 913-24.
277. McDermott, S.M., et al., *The Architecture of Trypanosoma brucei editosomes.* Proc Natl Acad Sci U S A, 2016. **113**(42): p. E6476-E6485.
278. Blum, B., et al., *Chimeric gRNA-mRNA molecules with oligo(U) tails covalently linked at sites of RNA editing suggest that U addition occurs by transesterification.* Cell, 1991. **65**(4): p. 543-50.
279. Zimmer, S.L., R.M. Simpson, and L.K. Read, *High throughput sequencing revolution reveals conserved fundamentals of U-indel editing.* Wiley Interdiscip Rev RNA, 2018 Jun 11. **e1487**.
280. Tylec, B.L., et al., *Intrinsic and regulated properties of minimally edited trypanosome mRNAs.* Nucleic Acids Res, 2019. **47**(7): p. 3640-3657.
281. Ochsenreiter, T., M. Cipriano, and S.L. Hajduk, *Alternative mRNA editing in trypanosomes is extensive and may contribute to mitochondrial protein diversity.* PLoS One, 2008. **3**(2): p. e1566.
282. Kirby, L.E. and D. Koslowsky, *Cell-line specific RNA editing patterns in Trypanosoma brucei suggest a unique mechanism to generate protein variation in a system intolerant to genetic mutations.* Nucleic Acids Res, 2020. **48**(3): p. 1479-1493.
283. Meehan, J., et al., *KREH2 helicase represses ND7 mRNA editing in procyclic-stage Trypanosoma brucei by opposite modulation of canonical and 'moonlighting' gRNA utilization creating a proposed mRNA structure.* Nucleic Acids Res, 2024.

284. Meehan, J., et al., *Trypanosome RNA helicase KREH2 differentially controls non-canonical editing and putative repressive structure via a novel proposed 'bifunctional' gRNA in mRNA A6.* Nucleic Acids Res, 2023. **51**(13): p. 6944-6965.
285. Druzhyna, N.M., G.L. Wilson, and S.P. LeDoux, *Mitochondrial DNA repair in aging and disease.* Mech Ageing Dev, 2008. **129**(7-8): p. 383-90.
286. Cucchi, D., A. Gibson, and S.A. Martin, *The emerging relationship between metabolism and DNA repair.* Cell Cycle, 2021. **20**(10): p. 943-959.
287. Rong, Z., et al., *The Mitochondrial Response to DNA Damage.* Front Cell Dev Biol, 2021. **9**: p. 669379.
288. Lukes, J., H. Hashimi, and A. Zikova, *Unexplained complexity of the mitochondrial genome and transcriptome in kinetoplastid flagellates.* Curr Genet, 2005. **48**(5): p. 277-99.
289. Read, L.K., et al., *Developmental regulation of RNA editing and polyadenylation in four life cycle stages of Trypanosoma congolense.* Mol Biochem Parasitol, 1994. **68**(2): p. 297-306.
290. Michaeli, S., *Non-coding RNA and the complex regulation of the trypanosome life cycle.* Curr Opin Microbiol, 2014. **20**: p. 146-52.
291. Inbar, E., et al., *The Transcriptome of Leishmania major Developmental Stages in Their Natural Sand Fly Vector.* mBio, 2017. **8**(2).
292. Filosa, J.N., et al., *Dramatic changes in gene expression in different forms of Crithidia fasciculata reveal potential mechanisms for insect-specific adhesion in kinetoplastid parasites.* PLoS Negl Trop Dis, 2019. **13**(7): p. e0007570.
293. Clement, S.L., M.K. Mingler, and D.J. Koslowsky, *An intragenic guide RNA location suggests a complex mechanism for mitochondrial gene expression in Trypanosoma brucei.* Eukaryot Cell, 2004. **3**(4): p. 862-9.
294. Landweber, L.F. and W. Gilbert, *RNA editing as a source of genetic variation.* Nature, 1993. **363**(6425): p. 179-82.
295. Simpson, L., et al., *Evolution of RNA editing in trypanosome mitochondria.* Proc Natl Acad Sci U S A, 2000 Jun 20. **97**(13): p. 6986-93.
296. Stoltzfus, A., *On the possibility of constructive neutral evolution.* J Mol Evol, 1999. **49**(2): p. 169-81.
297. Zimmer, S.L., *Revisiting Trypanosome Mitochondrial Genome Mysteries: Broader and Deeper.* Trends Parasitol, 2019 Feb. **35**(2): p. 102-104.
298. Tihon, E., et al., *Discovery and genomic analyses of hybridization between divergent lineages of Trypanosoma congolense, causative agent of Animal African Trypanosomiasis.* Mol Ecol, 2017. **26**(23): p. 6524-6538.
299. Van den Broeck, F., et al., *Mitonuclear genomics challenges the theory of clonality in Trypanosoma congolense: Reply to Tibayrenc and Ayala.* Mol Ecol, 2018. **27**(17): p. 3425-3431.
300. Ochieng, G.O., *Evaluation of kinetoplast DNA loss in dyskinetoplastic trypanosomes: in vitro adaptation and ethidium bromide challenge.*, in *Educational Commission Biomedical Sciences* 2024, University of Antwerp: Educational Commission Biomedical Sciences p. 49.
301. Aerts, D., et al., *A kit for in vitro isolation of trypanosomes in the field: first trial with sleeping sickness patients in the Congo Republic.* Trans R Soc Trop Med Hyg, 1992. **86**(4): p. 394-5.
302. Birkenmeyer, L., H. Sugisaki, and D.S. Ray, *The majority of minicircle DNA in Crithidia fasciculata strain CF-C1 is of a single class with nearly homogeneous DNA sequence.* Nucleic Acids Res, 1985. **13**(19): p. 7107-18.
303. Li, D., et al., *MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.* Bioinformatics, 2015. **31**(10): p. 1674-6.
304. Gibson, W., *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, 2012. **5**: p. 71.

305. Wellde, B., et al., *Trypanosoma congolense. I. Clinical observations of experimentally infected cattle.* Exp Parasitol, 1974. **36**(1): p. 6-19.
306. Masumu, J., D. Geysen, and P. Van den Bossche, *Endemic type of animal trypanosomiasis is not associated with lower genotype variability of Trypanosoma congolense isolates circulating in livestock.* Res Vet Sci, 2009. **87**(2): p. 265-9.
307. Van den Bossche, P., et al., *Virulence in Trypanosoma congolense Savannah subgroup. A comparison between strains and transmission cycles.* Parasite Immunol, 2011. **33**(8): p. 456-60.
308. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
309. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
310. *FastQC*. 2015.
311. Chen, S., et al., *fastp: an ultra-fast all-in-one FASTQ preprocessor.* Bioinformatics, 2018. **34**(17): p. i884-i890.
312. Berriman, M., et al., *The genome of the African trypanosome Trypanosoma brucei.* Science, 2005. **309**(5733): p. 416-22.
313. Abbas, A.H., et al., *The Structure of a Conserved Telomeric Region Associated with Variant Antigen Loci in the Blood Parasite Trypanosoma congolense.* Genome Biol Evol, 2018. **10**(9): p. 2458-2473.
314. Camacho, E., et al., *Gene Annotation and Transcriptome Delineation on a De Novo Genome Assembly for the Reference Leishmania major Friedlin Strain.* Genes (Basel), 2021. **12**(9).
315. Gonzalez-de la Fuente, S., et al., *Resequencing of the Leishmania infantum (strain JPCM5) genome and de novo assembly into 36 contigs.* Sci Rep, 2017. **7**(1): p. 18050.
316. Rogers, M.B., et al., *Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania.* Genome Research, 2011. **21**(12): p. 2129-2142.
317. Peacock, C.S., et al., *Comparative genomic analysis of three Leishmania species that cause diverse human disease.* Nature Genetics, 2007. **39**(7): p. 839-847.
318. Nasereddin, A., et al., *Characterization of Leishmania (Leishmania) tropica axenic amastigotes.* Acta Trop, 2010. **113**(1): p. 72-9.
319. Downing, T., et al., *Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance.* Genome Res, 2011. **21**(12): p. 2143-56.
320. Chen, Y., et al., *High speed BLASTN: an accelerated MegaBLAST search tool.* Nucleic Acids Res, 2015. **43**(16): p. 7762-8.
321. Rognes, T., et al., *VSEARCH: a versatile open source tool for metagenomics.* PeerJ, 2016. **4**: p. e2584.
322. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.
323. Danecek, P., et al., *Twelve years of SAMtools and BCFtools.* Gigascience, 2021. **10**(2).
324. Crooks, G.E., et al., *WebLogo: a sequence logo generator.* Genome Res, 2004. **14**(6): p. 1188-90.
325. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.
326. Gerasimov, E.S., et al., *Trypanosomatid mitochondrial RNA editing: dramatically complex transcript repertoires revealed with a dedicated mapping tool.* Nucleic Acids Res, 2018. **46**(2): p. 765-781.
327. Katoh, K., J. Rozewicki, and K.D. Yamada, *MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization.* Brief Bioinform, 2019. **20**(4): p. 1160-1166.
328. Kalyaanamoorthy, S., et al., *ModelFinder: fast model selection for accurate phylogenetic estimates.* Nat Methods, 2017. **14**(6): p. 587-589.

329. Hohna, S., et al., *RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language.* Syst Biol, 2016. **65**(4): p. 726-36.
330. Lewis, P.O., *A likelihood approach to estimating phylogeny from discrete morphological character data.* Syst Biol, 2001. **50**(6): p. 913-25.
331. Wright, A.M. and D.M. Hillis, *Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data.* PLoS One, 2014. **9**(10): p. e109210.
332. Nguyen, L.T., et al., *IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.* Mol Biol Evol, 2015. **32**(1): p. 268-74.
333. Jamonneau, V., et al., *Trypanosoma brucei gambiense Group 2: The Unusual Suspect.* Trends Parasitol, 2019. **35**(12): p. 983-995.
334. Agbo, E.C., et al., *Population genetic structure and cladistic analysis of Trypanosoma brucei isolates.* Infect Genet Evol, 2003. **3**(3): p. 165-74.
335. Hide, G., et al., *Epidemiological relationships of Trypanosoma brucei stocks from south east Uganda: evidence for different population structures in human infective and non-human infective isolates.* Parasitology, 1994. **109 (Pt 1)**: p. 95-111.
336. Moore, A. and M. Richer, *Re-emergence of epidemic sleeping sickness in southern Sudan.* Trop Med Int Health, 2001. **6**(5): p. 342-7.
337. Felu, C., et al., *Diagnostic potential of a conserved genomic rearrangement in the Trypanosoma brucei gambiense-specific TGSGP locus.* Am J Trop Med Hyg, 2007. **76**(5): p. 922-9.
338. Cuypers, B., et al., *Apolipoprotein L1 Variant Associated with Increased Susceptibility to Trypanosome Infection.* mBio, 2016. **7**(2): p. e02198-15.
339. Paindavoine, P., et al., *Different allele frequencies in Trypanosoma brucei brucei and Trypanosoma brucei gambiense populations.* Mol Biochem Parasitol, 1989. **32**(1): p. 61-71.
340. Simpson, L., *The mitochondrial genome of kinetoplastid protozoa: genomic organization, transcription, replication, and evolution.* Annu Rev Microbiol, 1987. **41**: p. 363-82.
341. Hong, M. and L. Simpson, *Genomic organization of kinetoplast DNA minicircles.* Protist, 2003. **154**(2): p. 265-279.
342. Jasmer, D.P. and K. Stuart, *Sequence organization in African trypanosome minicircles is defined by 18 base pair inverted repeats.* Mol Biochem Parasitol, 1986. **18**(3): p. 321-31.
343. Riley, G.R., R.A. Corell, and K. Stuart, *Multiple guide RNAs for identical editing of Trypanosoma brucei apocytochrome b mRNA have an unusual minicircle location and are developmentally regulated.* J Biol Chem, 1994. **269**(8): p. 6101-8.
344. Büscher, P., et al., *Improved latex agglutination test for detection of antibodies in serum and cerebrospinal fluid of infected patients.* Acta Tropica, 1999. **73**(1): p. 11-20.
345. Aphasizheva, I., et al., *RNA binding and core complexes constitute the U-insertion/deletion editosome.* Mol Cell Biol, 2014. **34**(23): p. 4329-42.
346. van der Spek, H., et al., *Conserved genes encode guide RNAs in mitochondria of Crithidia fasciculata.* EMBO J, 1991. **10**(5): p. 1217-24.
347. Simpson, L., et al., *Comparison of the Mitochondrial Genomes and Steady State Transcriptomes of Two Strains of the Trypanosomatid Parasite, Leishmania tarentolae.* PLoS Negl Trop Dis, 2015. **9**(7): p. e0003841.
348. Larsson, A., *AliView: a fast and lightweight alignment viewer and editor for large datasets.* Bioinformatics, 2014. **30**(22): p. 3276-8.
349. Koslowsky, D., et al., *The insect-phase gRNA transcriptome in Trypanosoma brucei.* Nucleic Acids Res, 2014. **42**(3): p. 1873-86.
350. Sturm, N.R. and L. Simpson, *Kinetoplast DNA minicircles encode guide RNAs for editing of cytochrome oxidase subunit III mRNA.* Cell, 1990. **61**(5): p. 879-84.

351. Gillingwater, K., P. Buscher, and R. Brun, *Establishment of a panel of reference Trypanosoma evansi and Trypanosoma equiperdum strains for drug screening.* Vet Parasitol, 2007. **148**(2): p. 114-21.
352. Hagos, A., et al., *Efficacy of Cymelarsan and Diminasan against Trypanosoma equiperdum infections in mice and horses.* Vet Parasitol, 2010. **171**(3-4): p. 200-6.
353. Hajduk, S.L., *Demonstration of kinetoplast DNA in dyskinetoplastic strains of Trypanosoma equiperdum.* Science, 1976. **191**(4229): p. 858-9.
354. Lun, Z.R., R. Brun, and W. Gibson, *Kinetoplast DNA and molecular karyotypes of Trypanosoma evansi and Trypanosoma equiperdum from China.* Mol Biochem Parasitol, 1992. **50**(2): p. 189-96.
355. Steinert, M. and S. Van Assel, *Sequence heterogeneity in kinetoplast DNA: reassociation kinetics.* Plasmid, 1980. **3**(1): p. 7-17.
356. Birhanu, H., et al., *New Trypanosoma evansi Type B Isolates from Ethiopian Dromedary Camels.* PLoS Negl Trop Dis, 2016. **10**(4): p. e0004556.
357. Baltz, T., et al., *Cultivation in a semi-defined medium of animal infective forms of Trypanosoma brucei, T. equiperdum, T. evansi, T. rhodesiense and T. gambiense.* EMBO J, 1985. **4**(5): p. 1273-7.
358. Songa, E.B., et al., *Evidence for kinetoplast and nuclear DNA homogeneity in Trypanosoma evansi isolates.* Mol Biochem Parasitol, 1990. **43**(2): p. 167-79.
359. Ou, Y.C., C. Giroud, and T. Baltz, *Kinetoplast DNA analysis of four Trypanosoma evansi strains.* Mol Biochem Parasitol, 1991. **46**(1): p. 97-102.
360. Njiru, Z.K., et al., *Characterization of Trypanosoma evansi type B.* Infect Genet Evol, 2006. **6**(4): p. 292-300.
361. Ngaira, J.M., et al., *The detection of non-RoTat 1.2 Trypanosoma evansi.* Exp Parasitol, 2005. **110**(1): p. 30-8.
362. Suganuma, K., et al., *Isolation, cultivation and molecular characterization of a new Trypanosoma equiperdum strain in Mongolia.* Parasit Vectors, 2016. **9**(1): p. 481.
363. Tobie, E.J., *Loss of the Kinetoplast in a Strain of Trypanosoma equiperdum.* Transactions of the American Microscopical Society, 1951. **70**(3): p. 251-254.
364. Schnaufer, A., et al., *An RNA ligase essential for RNA editing and survival of the bloodstream form of Trypanosoma brucei.* Science, 2001. **291**(5511): p. 2159-62.
365. Wadsworth, E., *Bi-parental inheritance of kinetoplast DNA following sexual reproduction maintains mitochondrial genome complexity in Trypanosoma and Leishmania parasites* 2020.
366. Checchi, F., et al., *The natural progression of Gambiense sleeping sickness: what is the evidence?* PLoS Negl Trop Dis, 2008. **2**(12): p. e303.
367. Krafsur, E.S. and I. Maudlin, *Tsetse fly evolution, genetics and the trypanosomiases - A review.* Infect Genet Evol, 2018. **64**: p. 185-206.
368. Bemba, I., et al., *Tsetse Flies Infected with Trypanosomes in Three Active Human African Trypanosomiasis Foci of the Republic of Congo.* Pathogens, 2022. **11**(11).
369. Ravel, S., et al., *Cyclical transmission of Trypanosoma brucei gambiense in Glossina palpalis gambiensis displays great differences among field isolates.* Acta Trop, 2006. **100**(1-2): p. 151-5.
370. Welburn, S.C., D.H. Molyneux, and I. Maudlin, *Beyond Tsetse--Implications for Research and Control of Human African Trypanosomiasis Epidemics.* Trends Parasitol, 2016. **32**(3): p. 230-241.
371. Kostygov, A.Y., et al., *Phylogenetic framework to explore trait evolution in Trypanosomatidae.* Trends Parasitol, 2024. **40**(2): p. 96-99.
372. Lukes, J., et al., *How a neutral evolutionary ratchet can build cellular complexity.* IUBMB Life, 2011. **63**(7): p. 528-37.

373. Roy Chowdhury, A., et al., *The killing of African trypanosomes by ethidium bromide.* PLoS Pathog, 2010. **6**(12): p. e1001226.
374. Lun, Z.R., et al., *Trypanosoma brucei: two steps to spread out from Africa.* Trends Parasitol, 2010. **26**(9): p. 424-7.
375. Schnaufer, A., *Evolution of dyskinetoplastic trypanosomes: how, and how often?* Trends Parasitol, 2010. **26**(12): p. 557-8.
376. Buscher, P., et al., *Equine trypanosomosis: enigmas and diagnostic challenges.* Parasit Vectors, 2019. **12**(1): p. 234.
377. Holland, W.G., et al., *The effect of Trypanosoma evansi infection on pig performance and vaccination against classical swine fever.* Vet Parasitol, 2003. **111**(2-3): p. 115-23.
378. Haig, D.A. and A.S. Lund, *Transmission of the South African strain of dourine to laboratory animals.* Onderstepoort J Vet Sci Anim Ind, 1948. **23**(1-2): p. 59-61.
379. Capewell, P., et al., *The skin is a significant but overlooked anatomical reservoir for vector-borne African trypanosomes.* Elife, 2016. **5**.
380. Desquesnes, M., et al., *Development of a mathematical model for mechanical transmission of trypanosomes and other pathogens of cattle transmitted by tabanids.* Int J Parasitol, 2009. **39**(3): p. 333-46.
381. Lun, Z.R., et al., *The isoenzyme characteristics of Trypanosoma evansi and Trypanosoma equiperdum isolated from domestic stocks in China.* Ann Trop Med Parasitol, 1992. **86**(4): p. 333-40.
382. Reuter, C., et al., *Vector-borne Trypanosoma brucei parasites develop in artificial human skin and persist as skin tissue forms.* Nat Commun, 2023. **14**(1): p. 7660.
383. Helwak, A. and D. Tollervey, *Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH).* Nat Protoc, 2014. **9**(3): p. 711-28.
384. Braun, J., et al., *Rapid identification of regulatory microRNAs by miTRAP (miRNA trapping by RNA in vitro affinity purification).* Nucleic Acids Res, 2014. **42**(8): p. e66.
385. Sugimoto, Y., et al., *hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1.* Nature, 2015. **519**(7544): p. 491-4.