Deep kinetoplast genome analyses result in a novel molecular assay for detecting *Trypanosoma brucei gambiense*-specific minicircles

Manon Geerts[1], Zihao Chen[2], Nicolas Bebronne[1], Nicholas J. Savill[2], Achim Schnaufer[2], Philippe Büscher[1], Nick Van Reet[1], Frederik Van den Broeck[1,3]

[1] Department of Biomedical Sciences, Institute of Tropical Medicine, 2000 Antwerp, Belgium

[2] Institute of Immunology and Infection Research, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom

[3] Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, Katholieke Universiteit Leuven, 3000 Leuven, Belgium

Co-senior authors: Frederik Van den Broeck and Nick Van Reet

Corresponding author: Frederik Van den Broeck

KEYWORDS

next-generation sequencing; comparative genomics; mitochondrial minicircles; African trypanosomiasis; diagnostics; animal reservoir, *Trypanosoma brucei gambiense*

ABSTRACT

The World Health Organization targeted *Trypanosoma brucei gambiense* (*Tbg*) human African trypanosomiasis for elimination of transmission by 2030. Sensitive molecular markers that specifically detect *Tbg* type 1 (*Tbg1)* parasites will be important tools to assist in reaching this goal. Here, we aim at improving molecular diagnosis of *Tbg*1 infections by targeting the abundant mitochondrial minicircles within the kinetoplast of *Trypanosoma brucei* parasites. Using Next-Generation Sequencing of total cellular DNA extracts, we assembled and annotated the kinetoplast genome and investigated minicircle sequence diversity in 38 animal- and human-infective trypanosome strains. Computational analyses recognized a total of 241 Minicircle Sequence Classes as *Tbg*1-specific, of which three were shared by the 18 studied *Tbg*1 strains. We then developed a novel multiplex quantitative PCR assay (*g*-qPCR3) targeting one *Tbg*1-specific minicircle and three *Tbg*1-specific or Trypanozoon-specific markers. Molecular analyses revealed that the minicircle-based assay is applicable on animals and is as specific as the *TgsGP*-based assay, the current golden standard for molecular detection of *Tbg*1. The median copy number of the targeted minicircle was equal to eight, suggesting that our minicircle-based assay may be used for the sensitive detection of *Tbg*1 parasites. Finally, annotation of the targeted minicircle sequence indicated that it encodes genes essential for the survival of the parasite, and will thus likely be preserved in natural *Tbg*1 populations. These results demonstrated that our minicircle-based assay is a promising new tool for reliable and sensitive detection of *Tbg*1 infections in humans and animals.

INTRODUCTION

Human African trypanosomiasis (HAT), also known as sleeping sickness, is a vector-borne disease caused by two *Trypanosoma brucei* (*Tb*) subspecies and transmitted by tsetse flies. *Trypanosoma brucei rhodesiense* (*Tbr*) causes acute infections in East Africa, whereas *Trypanosoma brucei gambiense* (*Tbg*) causes chronic infections in West and Central Africa (1). *Trypanosoma brucei rhodesiense* and *Tbg* type I (*Tbg*1) are defined by the presence of truncated variant surface glycoprotein (VSG) genes, respectively the serum-resistance-associated (SRA) gene (2) and the *Tbg*-specific glycoprotein (TgsGP) gene (3) that play a role in infectivity to humans (4). Unlike *Tbr*, *Tbg*1 parasites are genetically homogeneous and form a monophyletic group (5, 6). They are responsible for the vast majority of the HAT-cases (85% of the 663 newly reported HAT-cases in 2020) (7). Infections by a third group of human-infective trypanosomes - *Tbg* type II (*Tbg*2) that lack the SRA and TgsGP genes - are extremely rare (8). Like the non-human-infective *T. b. brucei (Tbb)*, *Tbg*2 is genetically and phenotypically highly diverse (8).

The WHO targeted *Tbg*-HAT (gHAT) for elimination as a public health problem by 2020 and for global elimination of transmission (EOT) to humans (i.e. zero reported cases) by 2030 (9). Elimination of gHAT as a public health problem has been reached in several countries and HAT foci, and recently Togo and Côte d'Ivoire have been validated as such by WHO (10, 11). However, EOT remains challenging due to imperfect diagnostics and the risk of re-emergence from asymptomatic human infections and/or a possible animal reservoir (12–14). Current serological tests based on the *Tbg*-specific LiTat 1.3 and LiTat 1.5 VSG antigens (15–20) still require validation in different animal species. The golden standard for molecular detection of *Tbg*1 involves the single-copy *TgsGP* gene (21–23), but the analytical sensitivity of these tests is limited because they target a hemizygous single-copy gene (24). Among the molecular tests, several studies have proposed various genotyping techniques such as isoenzymes, ribosomal genes, VSGs, SNPs and microsatellites (25–32), but these have some disadvantages such as the requirement of multiple PCR reactions or high

amounts of input material, without necessarily increasing the sensitivity of *Tbg1* detection. To this end, alternative genetic markers are much needed to reliably and with improved sensitivity demonstrate *Tbg*1 infection in humans and in animals, including the tsetse fly vector.

Previous reports indicated that *Tbg1*-specific minicircle sequences exist in the kinetoplast DNA (33, 34). The kinetoplast DNA (kDNA), unique to the single mitochondrion of unicellular flagellates of the order Kinetoplastida, is a giant network of dozens of homogenous maxicircles (20-30 kb) interlaced with hundreds to thousands of heterogeneous minicircles (0.5-2.5 kb) (35). Maxicircles are homologous to the mitochondrial genome of other eukaryotes and encode components of the respiratory chain complexes and the mitoribosome. Minicircles generally consist of a ~100 bp conserved sequence region that contains hyper conserved sequences named Conserved Sequence Blocks (CSBs), and a variable region including genes encoding guide RNAs (gRNAs) that are responsible for directing post-transcriptional modification of the maxicircle-encoded messenger RNAs (35). A complete assembly and annotation of the kinetoplast genome of a lab-adapted *Tbb* strain identified 391 minicircle classes, encoding ~1,000 gRNA genes (36). Analysis of kDNA minicircles is already used for *Leishmania* detection and differentiation (37, 38) and allows subtypes of *Trypanosoma evansi* to be distinguished (39). It has also been proposed for sensitive and specific detection of *Tbg*1 infection in humans (34) and animals (33). However, the exact nature of these sequences is unknown and their use for *Tbg1* diagnosis is cumbersome, requiring both nested PCR and DNA hybridization.

Recent advances in the development of bioinformatic tools now facilitate the assembly and annotation of the structurally complex kinetoplast genome (36, 40), allowing us to investigate minicircle sequence diversity in tens to hundreds of samples simultaneously (41). Therefore, we used next-generation sequencing of total cellular DNA extracts as a strategy to investigate minicircle sequence diversity in 38 animal- and human-infective trypanosome strains from diverse geographical origins. Following a series of computational analyses, we

identify minicircles that were present in all *Tbg*1 strains and absent in all *Tbb*, *Tbg2* and *Tbr* strains. Using a newly developed quantitative PCR (qPCR), we demonstrate that our minicircle-based assay reliably identifies *Tbg1* within the Trypanozoon subgenus. Furthermore, we show that the copy number of the targeted minicircle was on average 4-fold higher compared to the hemizygous single-copy *TgsGP* gene, suggesting that our minicircle-based assay may be used for sensitive detection of *Tbg*1 infections in humans and animals.

METHODS

Ethics statement

Expansion of bloodstream form trypanosome populations in mice received approval from the Animal Ethics Committee of the Institute of Tropical Medicine (DPU2017-1).

DNA extraction and sequencing

To cover a wide geographical range, we included a total of 18 *Tbg*1 strains isolated from humans between 1952 and 2008 from Cameroon (n = 2), the Democratic Republic of the Congo (n = 10), the Republic of Congo (n = 2), Côte d'Ivoire (n = 3) and South Sudan (n = 1) (Supplementary Table 1). For comparative purposes, we also included nine *Tbb*, five *Tbg*2 and six *Tbr* strains (Supplementary Table 1).

All 38 strains were propagated as bloodstream form populations in OF1 mice (Charles River, Belgium) and purified from the infected mouse blood via DEAE ion exchange chromatography (42). Purified trypanosomes were sedimented by centrifugation (17,000 g, 10 min at 4°C). DNA of 50 µL pure trypanosome sediment was extracted using the standard phenol:chloroform method (43), aliquoted at 1 ng/µl and stored at -20°C. The concentration of extracted DNA was determined using a Qubit 4 Fluorometer (Invitrogen by Thermo Fisher Scientific). Paired-end 150 bp sequences were generated using the DNA nanoball sequencing technology (DNBSEQ^{TM}) at the Beijing Genomics Institute (BGI) in Hongkong, China.

Genomic analyses

Paired-end reads were aligned against the *Tbb* TREU927 v4.6 reference genome (available on https://tritrypdb.org) using SMALT v0.7.6 (https://www.sanger.ac.uk/tool/smalt-0/). A hash index of the reference genome was built using *k*-mer words of length 13 that were sampled every other position in the genome. Mapping was done using an exhaustive search for

alignments with a minimum identity threshold (-y) of 80% and a maximum insert size for paired reads of 1,500 bp.

Variant calling and filtering was performed using the Genome Analysis Toolkit (GATK) v4.1.4.1 (44). First, reads were assigned to a single read-group with *AddOrReplaceReadGroups* and duplicated reads were marked with *MarkDuplicates.* Variants were then called for each strain separately with *HaplotypeCaller* using default parameters. The resulting gVCF files for all strains were combined with *CombineGVCFs* to allow joint genotyping with *GenotypeGVCFs*. Single Nucleotide Polymorphisms (SNPs) were extracted from the resulting VCF file with *SelectVariants* and filtered with *VariantFiltration* using the following parameters: QUAL < 500, DP < 5, QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, --cluster_window_size 10 and --cluster_size 3. Finally, we used BCFtools v1.10.2 (45) to extract bi-allelic SNP sites that were called in all *Tb* strains. Using the resulting set of genome-wide SNPs, we reconstructed a phylogenetic network with SplitsTree v4 (46) to infer the ancestral relationship among the 38 *Tb* strains.

The species identity of the *Tbg*1 and *Tbr* strains was confirmed *in silico* by investigating the presence of the TgsGP and SRA genes, respectively. To this end, MEGAHIT v1.2.9 (47) was used for *de novo* assembly of genomes of all 38 *Tb* strains using default parameters. The presence of TgsGP and SRA genes in the assembled contigs was then confirmed through a local BLAST search (48) using publicly available nucleotide sequences of TgsGP (21) (NCBI accession number: FN555993) and SRA (2) (NCBI accession number: Z37159) with the following parameters: minimum 90% identity, minimum e-value of 0.0001 and minimum alignment length of 500bp.

Assembly of the kinetoplast genome

Reads that did not align to the *Tbb* TREU927 nuclear reference genome were extracted using SAMtools v1.9 (45) and converted to FASTQ format using GATK *SamToFastq*. These

unmapped reads were aligned against the 23 kb maxicircle sequence of *Tbb* Lister 427 (GenBank accession id M94286) using SMALT and the same parameters as described above except that the hash index was built with *k*-mer words of length six and the reads were mapped with a minimum identity threshold (-y) of 90% and a maximum insert size (-i) of 500. SNP calling was done using GATK as described above, and we extracted only those SNPs that passed the quality criteria (see above) and that were present within the maxicircle coding region (1.3-16.3 kb). Similar to the analyses above for the nuclear genome, a phylogenetic network analysis was done using maxicircle coding SNPs with SplitsTree.

Reads that did not align to the maxicircle sequence of *Tbb* Lister 427 were extracted from the alignment file as described above, and used for the assembly of minicircle contigs. Before assembly, sequence reads were trimmed for high quality with fastp v0.20.0 (49) using the following parameters: allow for a maximum of 10% of bases per read that have a phred-scaled base quality below 30, trim bases at either end of the read when their phred-scaled quality is below 30, move a sliding window of 10 bp from front to tail and cut the read once the average phred-scaled base quality drops below 30, and only retain reads with a minimum and maximum length of 100 bp and 155 bp after trimming, respectively. Using KOMICS v1.8 *assemble (40)*, trimmed reads were used for *de novo* assembly of contigs using a *k*-mer list of 99, 109 and 119, and putative minicircle contigs were extracted based on the highly conserved sequence block 3 (CSB3) dodecamer (GGGGTTG[G/A]TGTA) (36, 50) Five minicircles had a slight variation of CSB3 (one with GGGGGTGGTGTA found in *Tbg*2 strain FEO and four with GGGGTTAGTGTA found in *Tbg*1 strains 15BT-relapse, OUSOU, NDIMI and ROUPO-VAVOUA--80-MURAZ-14). These minicircles were also retained. KOMICS *circularize* was then used to identify circular minicircle contigs by searching for overlapping fragments at either end of each contig; when a minicircle contig was classified as circular, the overlapping fragment at the start of the contig was removed. Using KOMICS *polish*, all circular minicircle contigs were oriented by putting the conserved sequence block 1 (CSB1) (GGGCGT[T/G]C) (50) at the start of each contig. One minicircle

had a slight variation of CSB1 (GGGCGTGT found in *Tbg*2 strain MSUS-CI-78-TSW-157), which was specified to KOMICS to allow proper reorientation for this minicircle. Finally, KOMICS *polish* was also used to remove duplicate sequences, which was achieved by extracting the representative sequences (cluster centroids) of all clusters identified at 97% identity with VSEARCH v2.14.2 (51).

The quality of the minicircle assembly was assessed by re-aligning the unmapped reads to the assembled minicircles using SMALT. Before mapping, we first extended the circularized minicircle sequences by copying the last 150 bp at the start of each sequence to minimize the number of clipped reads at either end of the assembled minicircles, using a custom python script implemented in KOMICS. Following mapping with SMALT with exhaustive search (-x) and a percent identity of 97% (-y), we have calculated the following metrics using a bash script implemented in KOMICS: number of reads, number of mapped reads, number of properly paired reads, number of reads with mapping quality >=20, number of CSB3-containing reads, number of mapped CSB3-containing reads and number of perfectly aligned CSB3-containing reads (i.e. alignments without any insertions or deletions). The proportion of perfect alignments of CSB3-containing reads serves as a proxy for the total number of minicircles that were initially present within the DNA sample. All metrics were processed and visualized using the R function *msc.quality* as implemented in the R package rKOMICS (41). In addition, the quality of the assembly was further verified by calling SNPs with BCFtools mpileup/call, retaining only SNPs with QUAL >= 60 and DP >= 30, and assuming that high-quality assemblies should yield relatively low number of homozygous SNPs.

Finally, using the rKOMICS function *msc.depth*, minicircle copy numbers (MCN) were estimated as the median read depth per minicircle contig divided by the median genome-wide read depth times two (assuming diploidy in all *Tb* strains).

Identification of minicircle sequences unique to *Tbg*1

The diversity and similarity of minicircle sequences within all *Tb* strains were examined with the R package rKOMICS (41). Following visual inspection of length distributions using *msc.length*, we used *preprocess* to retain minicircle sequences that had the expected length (800-1200 bp) and that were successfully circularized. Retained sequences of all samples were concatenated into a single FASTA file and clustered into Minicircle Sequence Classes (MSCs) based on a minimum percent identity (MPI) of 70, 80, 90 and 95-100 with VSEARCH. In order to choose an appropriate MPI for downstream analyses, we processed VSEARCH clustering results with *msc.uc* and inspected - at each MPI - the number of MSCs, the number of perfect alignments and the number of 2-nt and 3-nt gaps. In addition, VSEARCH clustering results were stored into a matrix using the rKOMICS function *msc.matrix*, which records the presence (1) or absence (0) of all MSCs (rows) for each strain (columns). This matrix was subsequently used to document the number of MSCs per strain with *msc.richness*, to calculate the proportion of MSCs shared between the different *Tb* subspecies with *msc.similarity* and to investigate the ancestry among all *Tb* strains with *msc.pca*. Finally, we used the rKOMICS function *msc.subset* to find MSCs that were present in the *Tbg*1 strains and absent in the strains belonging to the other *Tb* subspecies.

Development of minicircle-based quantitative PCR assays

For each of the common *Tbg*1-specific MSC, we extracted the assembled minicircle sequences for all 18 *Tbg*1 strains from the alignment using the rKOMICS function *msc.seq*, generated consensus sequences with Jalview v2.11.1.4 (52) and designed primers and probes (Table 1) with the RealTimeDesign qPCR assay design software (LGC, Biosearch Technologies). Probes targeting *Tbg*1-specific MSCs were modified with a FAM dye label at the 5' end and paired with BHQ-1 *plus* at the 3' end.

A 20 µl reaction mixture contained 1X PerfeCTa qPCR Toughmix (Quantabio), 100 nM of each primer (LGC, Biosearch Technologies), 300 µM of each probe (LGC, Biosearch Technologies) and 5 µl of template DNA. The thermal cycling profile consisted of an initial

230 denaturation step at 95°C for 10 min followed by 40 cycles at 95°C for 15 s and 60°C for 1
231 min. qPCR was conducted on a Q-qPCR Instrument (Quantabio), and detection of the
232 quantification cycle (Cq) was calculated using the Q-qPCR Instrument Software v1.0.2 with
233 the automatic threshold enabled.

234 Each qPCR targeting *Tbg*1-specific MSCs was multiplexed with the *Trypanozoon*-specific
235 q18S-assay targeting the multi-copy 18S rRNA gene (53), the *Trypanozoon*-specific qGPI-
236 PLC-assay targeting the single-copy glycosylphosphatidylinositol-specific phospholipase C
237 (GPI-PLC) gene (53) and a *Tbg*1-specific qTgsGP-assay, designed to target the single-copy
238 TgsGP gene and avoid amplification of TgsGP-like genes (21). The multi-copy 18S rRNA
239 gene was used as an internal standard for the sensitive detection of *Trypanozoon* DNA. The
240 single-copy GPI-PLC gene was used as an internal standard to determine if sufficient
241 *Trypanozoon* DNA was present to detect a single-copy sequence (54), and for the
242 calculation of relative copy numbers (RCN) of each *Tbg*1-specific MSC (see below). The
243 single-copy TgsGP gene was used as a golden standard for the specific detection of *Tbg*1
244 DNA. The q18S-assay contains a CAL Fluor Orange 56 dye labeled probe paired with BHQ-
245 1 *plus* (53). The qGPI-PLC-assay contains a CAL Fluor Red 610 dye labeled probe paired
246 with BHQ-2 *plus (53)*. The qTgsGP-assay contains a Quasar 670 dye labeled probe paired
247 with BHQ-2 *plus*.

248 The qPCR efficiency and analytical sensitivity were calculated for each qPCR targeting
249 *Tbg*1-specific MSCs in simplex and in quadruplex format. This was done using
250 phenol:chloroform extracted DNA (see above) of two *Tbg*1 strains. From these DNA
251 extracts, ten-fold serial dilutions in DEPC-Treated Water, ranging from 100 pg/µl to 1 fg/µl,
252 were prepared. Each qPCR was run in quadruplicate for each DNA dilution. A reaction was
253 considered positive if at least three out of four replicates were positive.

254 The specificity of the quadruplex qPCR assays was assessed with the phenol:chloroform
255 extracted genomic DNA of 34 *Tbb*, 49 *Tbg*1, 7 *Tbg*2, 15 *Tbr*, 2 *T. equiperdum* and 5 *T.*

evansi strains (Supplementary Table 2). The 49 *Tbg*1 samples originated from Burkina Faso (2), Cameroon (2), Côte d'Ivoire (6), Congo Brazzaville (3), Democratic Republic of the Congo (35) and South Sudan (1). Note that 26 strains from the Democratic Republic of the Congo were sampled within the context of a treatment outcome study in Mbuji-Mayi (55), with 14 strains sampled from seven patients before (sample name include 'BT') and after ('AT') treatment. Quadruplex qPCR assays were run in duplicate for each DNA extract. The specificity of the assays was further assessed on DNA prepared from man, cattle, dog, goat, horse, sheep and tsetse (*Glossina fuscipes* from Kwamouth, Democratic Republic of the Congo, 2018), all known hosts of *Tbg*1 (12).

Relative copy numbers (RCN) of each target were calculated using the ΔCq-method with qGPI-PLC as reference. This was done by subtracting the Cq-values obtained for each *Tbg*1-specific MSC from the Cq-value obtained for qGPI-PLC. The resulting ΔCq-value were averaged between replicates and transformed (2^ΔCq) to yield RCNs for each target.

Annotation of minicircle sequences targeted by qPCR assays

Strain *Tbg*1 340AT (MHOM/CD/INRB/2006/21B) isolated in 2006 in Mbuji-Mayi (DRC) (55) (Supplementary Table 1) was selected for representative minicircle annotation because of its comparatively high minicircle complexity (see results). DNA extraction, sequencing and alignment of sequence reads against the *Tbb* TREU927 v4.6 reference genome was done as described above for the other 38 *Tb* isolates initially included in our study. Reads that did not align to the nuclear reference genome were used for the assembly of mitochondrial maxicircles and minicircles with KOMICS, as described above.

Due to the lack of transcriptomic data for *Tbg*1 strain 340AT, edited mRNA sequences were predicted following a similar approach as in (40). Edited mRNA sequences for *Tbb* strains Lister 427, EATRO 164 and EATRO 1125 (36) were obtained from GenBank and manually corrected for changes in non-T residues based on alignment of the *Tbg*1 340AT maxicircle with the annotated *Tbb* EATRO 1125 maxicircle (36).

Guide RNA prediction and minicircle annotation were performed with python3.7 package *kDNA annotation (36)*. The alignments of the gRNAs encoded on the minicircles targeted by the diagnostic assay to their cognate mRNAs were carefully inspected to identify any 'non-redundant' gRNAs, i.e. gRNAs that direct editing events not covered by any other gRNAs. Furthermore, gRNAs 3' of such non-redundant gRNAs were checked for potential premature truncations by the gRNA calling algorithm by carefully examining the editing capacity of their 3' end extensions. Any gRNA genes that were confirmed to be non-redundant were considered essential, as were the minicircles that encoded them.

RESULTS

Genome analyses confirms the taxon identity of *Tbg*1 strains

The genomes of 38 *Tb* strains were sequenced at a median 159x depth (mean = 155, min = 126, max = 178) (Supplementary Table 3). On average 86.3% of the reads (min = 81.25%, max = 92.05%) aligned to the *Tbb* TREU927 nuclear reference genome (Supplementary Table 3). Initial variant discovery with GATK identified a total of 1,558,963 SNPs across the 38 *Tb* strains. Strict quality filtering and the exclusion of multiallelic sites reduced the data set to 316,287 genome-wide bi-allelic SNPs, of which 310,701 SNPs (98.23%) were located within the 11 megabase chromosomes. In addition, joint genotyping identified a total of 150 SNPs within the maxicircle coding region.

Phylogenetic analyses based on genome-wide SNPs and SNPs from the maxicircle coding region confirmed that all 18 *Tbg*1 parasites clustered together in a monophyletic group, a prerequisite for downstream analyses that aim at identifying *Tbg*1-specific minicircles (Supplementary Figure 1). Using a local BLAST search of assembled contigs, we also confirmed the presence of the *Tbg*1-specific TgsGP gene for the 18 *Tbg*1 strains, and the *Tbr*-specific SRA gene for the six *Tbr* strains. These two genes were absent for the remaining nine *Tbb* strains and five *Tbg*2 strains (Supplementary Table 4).

Assembly and circularization of kinetoplast minicircles

Mitochondrial minicircles were *de novo* assembled, circularized and reoriented for each of the 38 *Tb* strains using the Python package KOMICS. A total of 9,076 minicircle contigs were assembled across all 38 *Tb* strains, of which 7,156 (78,85%) were successfully circularized (Supplementary Table 5). The length of the majority of circularized minicircles (7,111 contigs, 99.37%) showed a unimodal distribution around ~1,000 bp (Supplementary Figure 2), which is comparable to the minicircle length found in *Tbb (36)*.

To validate the quality of the assembly process, sequence reads were aligned to the assembled minicircle contigs and several mapping and genotyping statistics were calculated. First, high-quality assemblies should result in relatively low numbers of homozygous SNPs when reads are aligned against the assembled contigs. Here, we identified a total of 302 homozygous SNPs within 127 minicircle contigs (Supplementary Table 5), which is only 1.4% of all assembled minicircle contigs. Only 48 homozygous SNPs were identified within 17 circularized contigs (0.2% of all circularized minicircles) (Supplementary Table 5). These results show that homozygous SNPs were found for only a fraction of the assembled contigs. Second, on average 96.99% of the sequence reads mapped in proper pairs and 94.46% aligned with a mapping quality larger than 20 (Supplementary Table 5), indicating that the large majority of mapped reads aligned with a high quality and with the expected orientation to the minicircle assemblies. Third, we calculated the number of aligned reads containing the CSB3 12-mer as a proxy for the total number of minicircles initially present within the DNA sample. This revealed that on average 93.24% of the CSB3-containing reads aligned against the assembled minicircles, and 88.77% aligned perfectly (Supplementary Table 5), suggesting that we were able to retrieve the vast majority of the minicircles.

Estimation of maxicircle and minicircle copy numbers and network size

To calculate the average number of maxicircles and minicircles per kinetoplast network, we used the coverage (i.e. median read depth) of the diploid nuclear genome. The average

334 genome-wide coverage was 155 per diploid cell (two copies), and the average coverage per
335 haploid sequence was equal to 78 (Supplementary Table 3). The coverage of the kinetoplast
336 maxicircles (coding region only) and minicircles was estimated based on median read
337 depths per 0.1kb. Average coverage of the maxicircle was 1,262 (median = 1,325, min = 99,
338 max = 2,320) (Supplementary Table 6). This equaled an average copy number of 17
339 maxicircles (median = 18, min = 1, max = 28) per network (Supplementary Table 6), which is
340 slightly lower compared with previous estimates of 20-50 copies per network (36, 56, 57).
341 The average copy number of minicircles (MCN) ranged from 2.1 to 37.9 copies per network
342 in *Tbg*1 strains, from 0.8 to 32.7 copies per network in *Tbb* strains, from 4 to 11.9 copies per
343 network in *Tbg*2 strains and from 3.1 to 11.1 copies per network in *Tbr* strains
344 (Supplementary Table 7). Thus, copy numbers for minicircles within each network varied
345 substantially. The size of each kDNA network was estimated by adding up the estimated
346 copy numbers for all minicircles. These calculations resulted in an estimated average of
347 ~2,100 minicircles per network, ranging between 252 minicircles in the Nabe strain and
348 4,828 minicircles in the MSUS-CI-78-TSW-157 strain (Supplementary Table 7). These
349 numbers are slightly lower compared to earlier estimates of 5,000-10,000 minicircles per
350 network (36, 56–58).

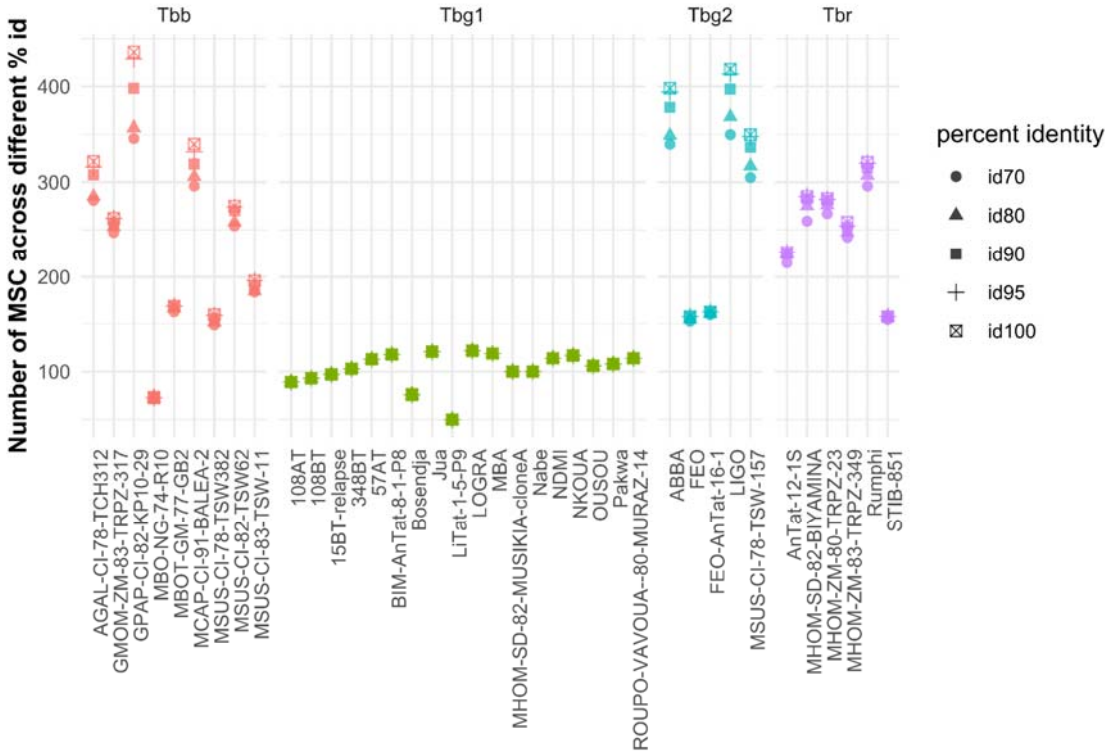
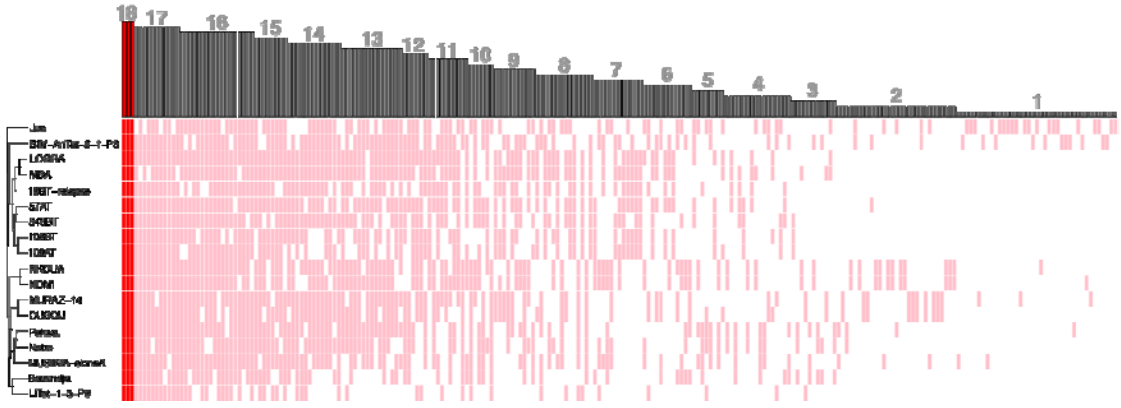


Figure 1. Number of Minicircle Sequence Classes (MSCs) for 38 *Tb* strains. Following *de novo* assembly and circularization, mitochondrial minicircle sequences were clustered into groups of sequences sharing a minimum percent identity (MSCs). Here, we summarized the number of MSCs identified within each *Tb* strain for a range of percent identities (70, 80, 90, 95 and 100). Strains were grouped as *T. b. brucei* (*Tbb*), *T. b. gambiense* type 1 (*Tbg*1), *T. b. gambiense* type 2 (*Tbg*2) and *T. b. rhodesiense* (*Tbr*).

Sequence diversity and similarity of kinetoplast minicircles

Minicircle sequence diversity was examined using a clustering approach, whereby minicircles were grouped into MSCs based on a minimum percent identity. This was done on the 7,111 circularized minicircle contigs of the expected length, as these would produce the most robust alignments. At 100% identity, a total of 5,883 unique MSCs were identified across the 38 *Tb* strains, leaving 1,228 MSCs that are shared between two or more isolates.

365 The number of MSCs decreased sharply with decreasing percent identities to a total of 719
366 MSCs at 70% identity (Supplementary Figure 3A). Regardless of the percent identity used,
367 *Tbg*1 parasites contained an average of 103 MSCs per strain (median = 107), which is on
368 average 2.48-fold lower compared to other taxa of the *Trypanozoon* subgenus (Figure 1).
369 Most of the *Tbg*1 strains displayed a fairly similar number of MSCs, ranging between 89 and
370 122 MSCs (Figure 1), with the exception of LiTat 1.5 (50 MSCs) and Bosendja (76 MSCs).
371 The composition of minicircle sequences was further investigated by quantifying the
372 proportion of MSCs unique to *Tbg*1. At 98%-100% identity, the 18 *Tbg*1 parasites did not
373 share any MSC with the non-*Tbg*1 subspecies (*Tbr*, *Tbg*2 and *Tbb*) (Supplementary Figure
374 3B). Below 98%, there was a steady increase in the proportion of shared MSCs between
375 *Tbg*1 and non-*Tbg*1 subspecies (Supplementary Figure 3B). At the 98% identity threshold,
376 the *Tbg*1 group contained 241 MSCs, none of which were found in the other *Tb* subspecies
377 and three of which were found in all 18 *Tbg*1 strains (Figure 2). These three *Tbg*1-specific
378 MSCs were retained as candidate markers for our new molecular test.
379



380

381 **Figure 2. Minicircle sequence diversity in 18 *Tbg*1 strains.** Clustering at the 98% identity threshold
382 revealed a total of 241 MSCs that were present in at least one of 18 *Tbg*1 strains and absent in the
383 *Tbb, Tbr* and *Tbg2* strains. Heatmap summarizes the presence/absence of these 241 MSCs
384 (columns) in the 18 *Tbg1* strains (rows). Gray lines indicate the presence of a given MSC in a given
385 *Tbg*1 strain. Dark red lines indicate the three MSCs that were found in all 18 *Tbg*1 strains. Barplot on
386 top of the plot shows the number of strains harboring a given MSC.

387

388 **Novel multiplex qPCRs including *Tbg*1-specific kinetoplast minicircles as target**

389 Using the RealTimeDesign qPCR assay design software, we successfully designed three
390 simplex qPCR assays (here-after referred to as qMini1, qMini2 and qMini3) targeting two
391 *Tbg*1-specific MSCs (here-after referred to as mO_078 and mO_104) (Table 1). Due to
392 limitations such as ambiguous base count, low GC percentage and low melting
393 temperatures, we were unable to design primers and probes for the third *Tbg*1-specific MSC.
394 The *Tbg*1 specificity of the designed primers and probes was confirmed with a BLAST
395 search (48) on TriTrypDB (https://tritrypdb.org), with a Primer BLAST search on NCBI
396 (https://www.ncbi.nlm.nih.gov) and with the command line search tool grep in sequencing
397 reads generated by this study. The three simplex qPCR assays qMini1, qMini2 and qMini3
398 were each multiplexed with q18S, qGPI-PLC and qTgsGP (Table 1) to produce three
399 quadruplex reactions (here-after referred to as *g*-qPCR1, *g*-qPCR2 and *g*-qPCR3 when
400 qMini1, qMini2 and qMini3 were included, respectively).

401 The efficiency and analytical sensitivity of the simplex and quadruplex qPCR assays were
402 investigated using DNA from two *Tbg1* strains (Supplementary Table 8), one with relatively
403 low MCNs (Nabe; average MCN = 1) and one with relatively high MCNs (LOGRA; average
404 MCN = 8). The lower detection limit of 0.05 pg DNA was reached in both strains for qMini1,
405 qMini2 and qMini3 in their respective simplex reactions (Supplementary Figure 4), for q18S
406 in all three *g*-qPCR assays (Supplementary Figure 4), and for qMini1 and qMini3 in their
407 respective *g*-qPCR assays, with the exception of qMini3 that achieved the detection limit of
408 0.5 pg DNA in the Nabe strain (Supplementary Figure 4). This detection limit of 0.5 pg DNA
409 was also reached by qTgsGP and/or qGPI-PLC in all three *g*-qPCR assays with the LOGRA
410 strain and in the *g*-qPCR2 and *g*-qPCR3 assays with the Nabe strain (Supplementary Figure
411 4). The analytical sensitivity of qMini2 was greatly reduced in the *g*-qPCR2 assay to 100 pg
412 DNA with the Nabe strain and 0.5 pg DNA with the LOGRA strain (Supplementary Figure 4).

The qPCR efficiency of *g*-qPCR1 and g-qPCR3 was estimated between 93% and 106% (Supplementary Figure 4), which is considered acceptable (https://www.thermofisher.com/content/dam/LifeTech/global/Forms/PDF/real-time-pcr-handbook.pdf).

Table 1. Summary of the various qPCR assays used in this study.

assay		sequence	target	position on target	expected amplicon length	reference
qMini1	F	5' TGAGGTCTGAGGTACTTCGAAAG 3'	mO_104	52-152	151	this paper
	R	5' TGGATTACTTGGTGTTTTCTATTGATAA 3'				
	P	5' FAM-TTTTCCTGGAGAAAACTGTAT-BHQ-1 plus 3'				
qMini2	F	5' TCTTATGACTGATTTTACGAGAATA 3'	mO_078	244-431	188	this paper
	R	5' GACATAACAGAGGAAAGTGCTC 3'				
	P	5' FAM-TTGTGGTAAGAGTGATTTAGTAAT-BHQ-1 plus 3'				
qMini3	F	5' AAACCAACAGAAAGAGATTGCTTA 3'	mO_078	626-810	185	this paper
	R	5' ATGGTGATAGAAGTTAGAGATGTGTAG 3'				
	P	5' FAM-TAGATGTAGTATAAGAATTTAAAAT-BHQ-1 plus 3'				
qTgsGP	F	5' GAAGCAGTGGGACCTTAGC 3'	*TgsGP*	753-840	87	this paper
	R	5' TTTGTGCTCTTGCTTGCTATTAC3'				
	P	5' Quasar 670 -CTCTCCGAACACAGCAGCGACATC-BHQ-21 plus 3'				
q18S	F	5' CGTAGTTGAACTGTGGGCCACGT 3'	*18S*	679--829	150	(53)
	R	5' ATGCATGACATGCGTGAAAGTGAG3'				
	P	5' CAL Fluor Orange 560 TCGGACGTGTTTTGACCCACGC-BHQ-1 plus 3'				
qGPI-PLC	F	5' CCCACAACCGTCTCTTTAACC 3'	*GPI-PLC*	520-626	106	(53)

	R	5' GGAGTCGTGCATAAGGGTATTC3'				
	P	5' CAL Fluor Red 610-ACACCACTTTGTAACCTCTGGCAGT-BHQ-1 plus 3'				

F = forward primer, R = reverse primer, P = probe sequence.

The qMini3 assay displays a similar specificity as the qTgsGP assay

To assess the taxon-specificity of qMini1 and qMini3 within the *Trypanozoon* subgenus, a total of 118 DNA extracts were tested with the *g*-qPCR1 and *g*-qPCR3 assays (Supplementary Table 9). Here, qMini2 was excluded because of its low analytical sensitivity in the *g*-qPCR2 assay (see above) and because it targets the same minicircle as qMini3 (Table 1). Six of the 118 DNA extracts were excluded as they didn't react in duplicate with qGPI-PLC, which was used as an internal standard to determine if sufficient *Trypanozoon* DNA was present to detect a single-copy sequence (Supplementary Table 10). The remaining 112 DNA extracts reacted with the *Trypanozoon*-specific assays q18S and qGPI-PLC (Supplementary Table 9). All 49 *Tbg*1 DNA extracts reacted with qTgsGP, qMini1 and qMini3, with the exception of MSUS/CI/82/TSW125_KP1_cloneB (*Tbg*1 isolated from a pig in Côte d'Ivoire), ALJO (*Tbg*1 isolated from a human patient in DRC) and GUIWI-BOBO80-MURAZ18 (*Tbg*1 isolated from a patient in Burkina Faso) that remained negative for qMini1. The qMini1 assay also showed one cross reaction with DNA extracted from the *Tbr* strain Etat 1.2 R, with a Cq-value 5x lower than that of the qGPI-PLC. The qMini3 assay showed no cross reactions or false negative results. In addition, the *g*-qPCR3 assay did not amplify DNA from human and non-human vertebrates that are known to be susceptible for *Tbg*1 infection, *in casu* horse, cattle, goat, sheep, pig and dog, and DNA from *Glossina fuscipe*s. Also, DNA from other livestock affecting trypanosomes like *T. congolense*, *T. theileri* and *T. vivax* was not amplified (Supplementary Table 9).

The qMini3 assay targets a minicircle with a relatively high, but variable copy number

Relative Copy Numbers (RCN) of the minicircle sequence targeted by qMini3 (RCN_{mO_078}), the target sequence of qTgsGP (RCN_{qTgsGP}) and the target sequence of q18S (RCN_{q18S}) were calculated for each of the 49 *Tbg*1 DNA extracts using the ΔCq-method (Supplementary Table 11). The median RCN_{mO_078} (7.98) was 4.29x higher than the median RCN_{qTgsGP} (1.86) and 2.06x higher than the median RCN_{q18S} (3.88) (Figure 3A). However, RCN_{mO_078} also displayed a larger variation (SD = 9.01, min = 0.77, max = 40.58) compared to RCN_{qTgsGP} (SD = 0.62, min = 0.94, max = 3.47) and RCN_{q18S} (SD = 0.70, min = 1.51, max = 5.47). The RCN_{mO_078} was lower compared to RCN_{qTgsGP} for 4/49 *Tbg*1 strains and to RCN_{q18S} for 12/49 *Tbg*1 strains. There was no association between RCN_{mO_078} with the year of isolation (Pearson correlation test; cor= -0.1925748, t = -1.331, df = 46, p-value = 0.1897) or the geographical origin of the strain (Kruskal-Wallis test; country: chi-squared = 4.0896, df = 5, *p*-value = 0.5366; region: chi-squared = 3.0453, df = 2, *p*-value = 0.2181). High variation in RCN_{mO_078} was found within one group of *Tbg*1 strains isolated from humans between 2005 and 2009 in Mbuji-Mayi (Democratic Republic of the Congo), with a minimum RCN_{mO_078} of 0.8 in the 186BT strain and a maximum RCN_{mO_078} of 36.7 in the 93AT strain. Here, RCN_{mO_078} was significantly higher in strains sampled after treatment (mean RCN_{mO_078} = 14.9) compared to the RCN_{mO_078} in strains sampled before treatment (mean RCN_{mO_078} = 6.9), although this difference was not significant (Welch two sample t-test on all strains, t=2.0305, df = 11.944, *p*=0.06519 and paired t-test on paired strains, t=1.5476, df=5, *p*=0.183). The RCN_{mO_078} as calculated using the ΔCq-method was strongly associated with MCN_{mO_078} as calculated using standardized read depths, with a coefficient of determination of 0.91, a slope of 0.32 and a y-intercept of 0.10 (Figure 3B).

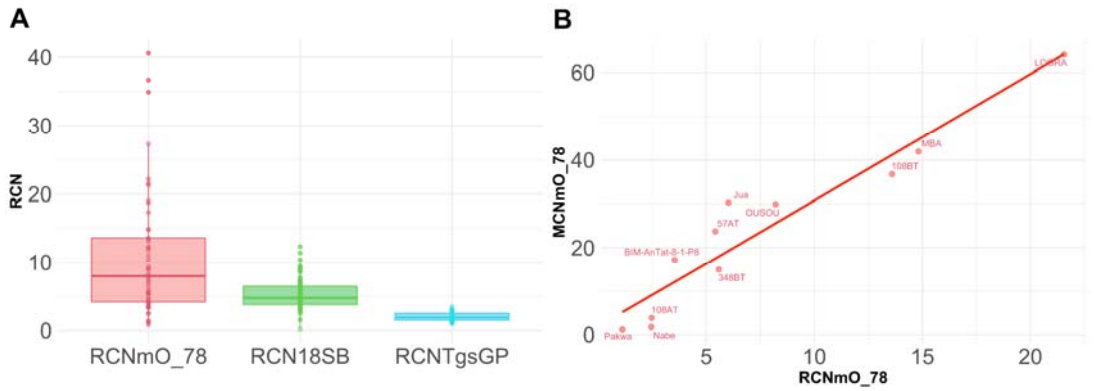


Figure 3. **Relative Copy Numbers of the qMini3 target sequence**. **(A)** Relative copy numbers (RCN) were estimated for the qMini3, q18S and qTgsGP target sequences using the ΔCq-method with the qGPI-PLC as reference. Boxplots summarize the RCN estimates as calculated for 49 *Tbg*1 strains. **(B)** Scatter plot showing the relationship between RCN (x-axis) and MCN (y-axis). To test whether RCN as calculated using the ΔCq-method is comparable to the Minicircle Copy Numbers (MCN) as calculated using standardized read depths, RCN and MCN were calculated for the qMini3 target sequence for 11 out of 18 *Tbg*1 strains. For these 11 strains, there was sufficient DNA to allow both whole genome sequencing (for MCN calculation) and a qPCR run (for RCN calculation) on the same DNA extract. The remaining seven strains were excluded here as there wasn't sufficient DNA left for a qPCR run following whole genome sequencing.

The qMini3 assay targets a minicircle containing non-redundant guide RNA genes

Annotation of minicircles was done for strain 340AT. A local BLAST search of assembled contigs revealed the presence of the *Tbg*1-specific *TgsGP* gene, confirming that 340AT is a *Tbg*1 strain (Supplementary Table 3). The kinetoplast maxicircle and minicircles were assembled with KOMICS using sequence reads that did not align to the nuclear reference genome. This resulted in a maxicicle contig of 21,287bp long (including the entire coding region) and a total of 143 minicircle contigs (including 129 circularized contigs). Hence, 340AT has the highest number of minicircles when compared to the 18 *Tbg*1 strains (max.

132 minicircles) initially sequenced in this study (Supplementary Table 7), which was the main motivation for using the 340AT data for representative minicircle annotation.

Annotation of the 143 minicircles revealed that the minicircle targeted by qMini3 encodes four gRNA genes (Figure 4). These gRNAs are involved in editing of the maxicircle genes cytochrome c oxidase subunit 3 (gCOX3(616-656) and gCOX3(341-369)), ATPase subunit 6 (gA6(415-452)) and NADH dehydrogenase subunit 7 (gND7(847-888)) (Figure 4). As expected, the minicircles contain the semi-conserved region characterized by conserved sequence blocks CSB1, CSB2 and CSB3 (36, 50), and the gRNA genes are framed by imperfect 18bp inverted repeats (36, 59–61).

Next, we investigated if any of the gRNAs encoded by minicircle mO_078 are non-redundant, i.e. whether they direct the editing of sites not covered by any of the other gRNAs encoded in this strain's kDNA. Minicircles encoding only redundant gRNAs might be more prone to loss due to lack of selective pressure, which could make a diagnostic assay based on such minicircles less reliable. Our analyses showed that gRNAs gA6(415-452) and gND7(847-888) direct editing events not covered by any other gRNA (Figure 5), and are thus non-redundant. The two COX3 gRNAs are redundant with gRNAs encoded by other minicircles (results not shown).

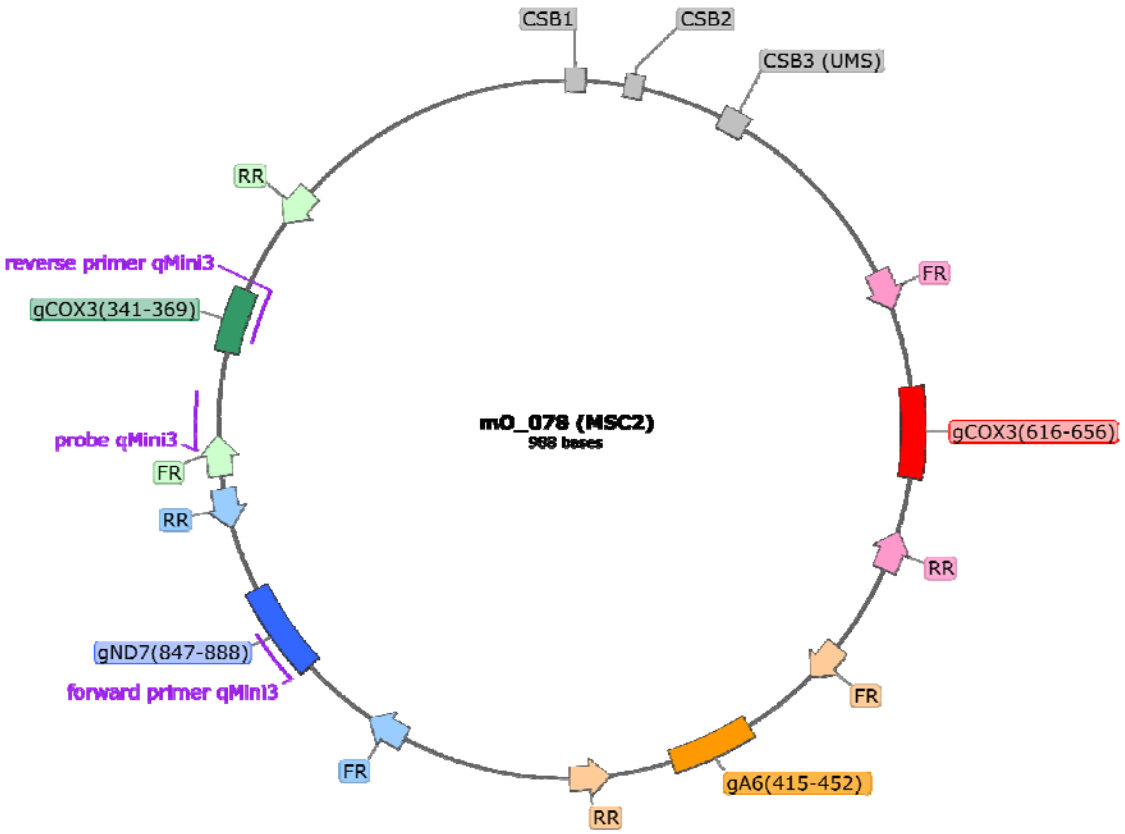
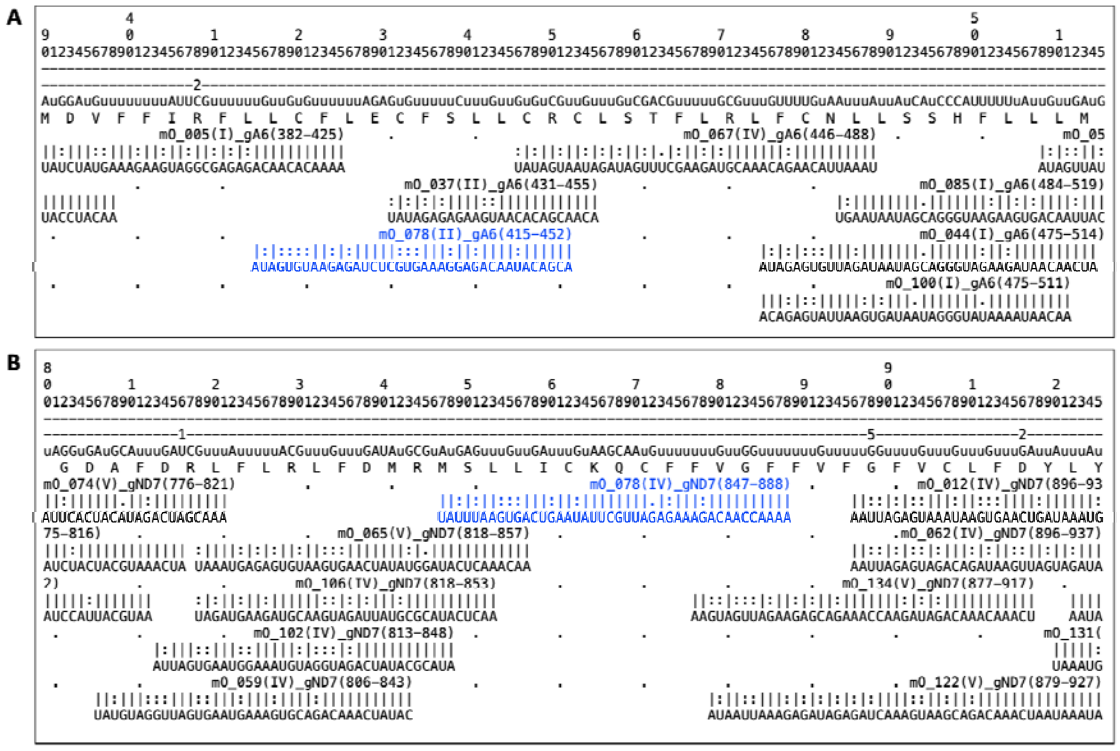


Figure 4. Annotation of minicircle mO_78 targeted by qMini3. Binding sites for the diagnostic PCR primer pair and the corresponding probe for qMini3 are indicated with purple lines. Conserved sequence blocks CSB1, CSB2 and CSB3 (or universal minicircle sequence, UMS) (gray boxes), the four encoded gRNA genes (red, orange, blue and green boxes), and the 18-bp inverted repeats (block arrows) that flank the gRNA genes are also indicated. This figure was generated with SnapGene (http://www.snapgene.com).



511

512 **Figure 5. Alignment of the non-redundant gRNAs from the qMini3 targeted minicircle mO_078**
513 **to their target mRNAs.** The alignment shows the non-redundant gRNAs in blue (**A**, gA6(415-452); **B**,
514 gND7(847-888)) as well as the mRNA sequences immediately upstream and downstream, along with
515 the neighbouring gRNAs encoded by other minicircles. Lines 1-3: mRNA position (hundreds, tens,
516 ones); line 4: number of Us that have been deleted from the pre-edited mRNA at this position; line 5:
517 edited mRNA sequence 5' to 3' (lowercase "u"s represent insertions). Editing events directed by the
518 non-redundant gRNA are shown in red. Note that the 'anchor' sequence at the 5' end of each gRNA
519 cannot direct editing events; line 6: protein sequence. For each gRNA: Line 1: name
520 (mO_name(cassette position)_mRNA(start-end of alignment on mRNA)), underscore characters
521 denote the anchor; line 2: base-pairing: "|": Watson-Crick basepair, ":": GU basepair, ".": mismatch
522 basepair; line 3: gRNA sequence 3' to 5'.

523

524

525

DISCUSSION

This study presents a computational investigation of mitochondrial minicircle sequence diversity in trypanosome isolates, resulting in the development of a qPCR assay as a promising new tool for sensitive diagnosis of *Tbg*1 infections in humans and animals.

Computational and phylogenetic analyses using genome sequencing data revealed that the 18 sequenced *Tbg*1 strains are monophyletic and contain the *TgsGP* gene, an essential precondition for identifying *Tbg*1-specific minicircles. The latter was achieved by grouping *de novo* assembled and circularized minicircles into MSCs according to sequence similarity. This uncovered a variable number of MSCs within the *Trypanozoon* subgenus, with a considerably lower number of MSCs in *Tbg*1 compared to the other subspecies. The comparatively lower number of MSC within *Tbg*1 may be the result of its asexual evolution (5) that results in the inevitable loss of redundant minicircles due to random genetic drift (62). Hence, the monophyletic origin and the asexual evolution of *Tbg*1 may explain the less complex minicircle populations in isolates for this subspecies, with conserved presence of some minicircle classes, which facilitated the identification of *Tbg*1-specific minicircles. Identification of taxon-specific minicircles may prove more challenging for trypanosomatid parasites experiencing occasional recombination (63–65), leading to heterogeneous minicircle populations as a result of biparental inheritance of mitochondrial minicircles (40, 66–68).

A total of 241 MSCs were recognized as *Tbg*1-specific, of which three were shared by the 18 studied *Tbg*1 strains. For two of the three *Tbg*1-specific minicircles (mO_078 and mO_104), three molecular assays could be successfully developed and tested. While two qPCR assays were discontinued because of false-negative results or low analytical sensitivities, one qPCR assay targeting minicircle mO_078 was fully specific (i.e. no false positives or false negatives) when tested on DNA of 112 different *Trypanosoma sp*. strains. These results show that the minicircle-based assay is as specific as the assay targeting the *TgsGP*

552 gene, the current golden standard for molecular detection of *Tbg*1 parasites (69–75),
553 confirming the taxon-specificity of some kDNA minicircles in *T.b. gambiense* (33, 34) and
554 their exploitability in molecular tests as has been described for *Leishmania* (37, 38) and *T.*
555 *evansi* type A and B (39, 76, 77). Annotation of the mO_078 minicircle demonstrated that it
556 encodes two non-redundant gRNAs that are essential for completing the editing of ATP
557 synthase subunit A6, a gene required for survival in both the bloodstream stage and the
558 insect stage of *Tb* (78, 79), and NADH dehydrogenase subunit 7, respectively. Hence, our
559 results indicate that minicircle mO_078 is essential for the survival of the parasite and will
560 most likely be preserved in all natural *Tbg*1 strains, ensuring the reliability of a diagnostic
561 assay targeting this minicircle.

562 Molecular analyses revealed that the *Tbg*1-specific minicircle mO_078 is a multicopy marker
563 for the large majority of strains tested in this study, with a median copy number equal to
564 eight and a maximum copy number of 41. Minicircle mO_078 copy number exceeded that of
565 the multicopy *18S* gene in ¾ of the *Tbg*1 strains and was as low as the *TgsGP* copy number
566 in only four of the 49 tested *Tbg*1 strains. This finding confirmed the multicopy nature of
567 mitochondrial minicircles in *Trypanosoma* and *Leishmania* parasites (36, 80). However, our
568 results also indicated that there is a relatively large variation in minicircle copy numbers
569 across strains, suggesting that the detection limit of the minicircle-based assay may depend
570 on the strain being investigated. The detection limit may also depend on the DNA extraction
571 method, as DNA-extraction based on spin columns causes a random loss of small molecules
572 like minicircles (36). Here, we avoided such biases by using the phenol-chloroform extraction
573 method that captures all nucleic acids, although this method may be less amenable for high-
574 throughput processing of human and animal specimens.

575 The high specificity and generally high copy number of minicircle mO_078 makes this a
576 promising new marker for sensitive diagnosis of *Tbg*1 infections. Specifically, within the
577 context of reaching EOT of *g*HAT by 2030 (9), our minicircle-based assay may prove
578 valuable for studying the role of an animal reservoir in the epidemiology of *g*HAT (12).

579 Therefore, we proposed a new multiplex qPCR assay (*g*-qPCR3) that targets the *Tbg*1-
580 specific minicircle mO_078 (serving as sensitive detection of *Tbg*1) in combination with the
581 *Trypanozoon*-specific *18S* (serving as sensitive detection of *Tb s.l.*) and the *Tbg*1-specific
582 *TgsGP* gene (which confirms the trypanosome is human-infective and thus of
583 epidemiological significance). We have shown that the *g*-qPCR3 assay is applicable on
584 animals, as it does not amplify DNA from other livestock affecting trypanosomes like *T.*
585 *congolense*, *T. theileri* and *T. vivax*, from the tsetse fly *Glossina fuscipe*s and from six
586 livestock species that are known to be susceptible to *Tbg*1 infection. Based on these data,
587 we also trust that the specificity will not be compromised when testing blood from wild
588 *Bovidae* and non-human primates (12), although we had no access to specimens from wild
589 fauna to formally test this. The main limitation of our study is that the *g*-qPCR3 assay was
590 tested on only one *Tbg*1 strain isolated from a host other than human, *in casu* a pig from
591 Côte d'Ivoire (81, 82), mainly because of the scarcity of such samples. However, given that
592 minicircle mO_078 is non-redundant and universal (computational analyses confirmed that
593 mO_078 was present in 193 sequenced *Tbg*1 strains sampled from humans in 11 different
594 countries (data not shown)), we are confident that the *g*-qPCR3 assay will be successful at
595 amplifying *Tbg*1 DNA from livestock specimens.

596 In conclusion, this study exemplifies the power of genome assembly and annotation for
597 identifying species-specific multicopy genetic markers. We developed a minicircle-based
598 assay that is as specific as the current golden standard for molecular detection of *Tbg*1
599 infections, and argued that the *g*-qPCR3 has the diagnostic potential for assessing the
600 importance of an animal reservoir in the epidemiology of *g*HAT.

601

FUNDING

PB received financial support from the Bill & Melinda Gates Foundation (grant number OPP1174221) and the Flemish Government EWI SOFI-2018 “Cryptic human and animal reservoirs compromise the sustained elimination of gambiense-human African trypanosomiasis in the Democratic Republic of the Congo”. FVdB was supported by the Department of Economy, Science and Innovation in Flanders and by the Research Foundation Flanders (Grants 1226120N and 1528117N). AS is supported by the UK Medical Research Council Fellowship MR/L019701/1. ZC is supported by an EASTBIO PhD studentship from the UK Biotechnology and Biological Sciences Research Council.

ACKNOWLEDGEMENTS

We thank Prof. Van Den Abbeele for critical reading of the manuscript, Isabel Saldanha and Steve Torr from the Liverpool School of Tropical Medicine for providing tsetse fly DNA and Dr. Vet. Xanthe Helsen and Dr. Vet. Wauter Van Deun for providing us with cattle, dog, goat, horse, pig and sheep blood.

DATA AND SCRIPT AVAILABILITY

Sequence reads generated within the context of this study have been deposited in the European Nucleotide Archive under accession number PRJEB49966 (https://www.ebi.ac.uk/ena/browser/view/PRJEB49966). The sequence of the *Tbg*1-specific minicircle mO_078 was deposited to NCBI under accession number OM238297. Scripts used for processing WGS data in this study are available at https://github.com/GeertsManon/g-qPCR. All graphical analyses were performed using R v4.0.3 in RStudio v1.4.1103.

626

627 **REFERENCES**

628 1. Gibson,W.C. (1986) Will the real Trypanosoma b. gambiense please stand up. *Parasitol.*
629 *Today*, **2**, 255–257.

630 2. De Greef,C. and Hamers,R. (1994) The serum resistance-associated (SRA) gene of
631 Trypanosoma brucei rhodesiense encodes a variant surface glycoprotein-like protein.
632 *Mol. Biochem. Parasitol.*, **68**, 277–284.

633 3. Berberof,M., Pérez-Morga,D. and Pays,E. (2001) A receptor-like flagellar pocket
634 glycoprotein specific to Trypanosoma brucei gambiense. *Mol. Biochem. Parasitol.*, **113**,
635 127–138.

636 4. Pays,E., Vanhollebeke,B., Uzureau,P., Lecordier,L. and Pérez-Morga,D. (2014) The
637 molecular arms race between African trypanosomes and humans. *Nat. Rev. Microbiol.*,
638 **12**, 575–584.

639 5. Weir,W., Capewell,P., Foth,B., Clucas,C., Pountain,A., Steketee,P., Veitch,N., Koffi,M.,
640 De Meeûs,T., Kaboré,J., *et al.* (2016) Population genomics reveals the origin and
641 asexual evolution of human infective trypanosomes. *Elife*, **5**, e11473.

642 6. Mathieu-Daudé,F., Stevens,J., Welsh,J., Tibayrenc,M. and McClelland,M. (1995) Genetic
643 diversity and population structure of Trypanosoma brucei: clonality versus sexuality.
644 *Mol. Biochem. Parasitol.*, **72**, 89–101.

645 7. Number of reported cases of human African trypanosomiasis (T.b. gambiense).

646 8. Jamonneau,V., Truc,P., Grébaut,P., Herder,S., Ravel,S., Solano,P. and De Meeus,T.
647 (2019) Trypanosoma brucei gambiense Group 2: The Unusual Suspect. *Trends*
648 *Parasitol.*, **35**, 983–995.

649 9. Franco,J.R., Cecchi,G., Priotto,G., Paone,M., Diarra,A., Grout,L., Simarro,P.P., Zhao,W.
650 and Argaw,D. (2020) Monitoring the elimination of human African trypanosomiasis at

651 continental and country level: Update to 2018. *PLoS Negl. Trop. Dis.*, **14**, e0008261.

652 10. Togo is first African country to end sleeping sickness as a public health problem.

653 11. WHO validates Côte d'Ivoire for eliminating sleeping sickness as a public health
654 problem.

655 12. Büscher,P., Bart,J.-M., Boelaert,M., Bucheton,B., Cecchi,G., Chitnis,N., Courtin,D.,
656 Figueiredo,L.M., Franco,J.-R., Grébaut,P., *et al.* (2018) Do Cryptic Reservoirs Threaten
657 Gambiense-Sleeping Sickness Elimination? *Trends Parasitol.*, **34**, 197–207.

658 13. Rock,K.S., Ndeffo-Mbah,M.L., Castaño,S., Palmer,C., Pandey,A., Atkins,K.E.,
659 Ndung'u,J.M., Hollingsworth,T.D., Galvani,A., Bever,C., *et al.* (2018) Assessing
660 Strategies Against Gambiense Sleeping Sickness Through Mathematical Modeling.
661 *Clin. Infect. Dis.*, **66**, S286–S292.

662 14. Mehlitz,D. and Molyneux,D.H. (2019) The elimination of Trypanosoma brucei
663 gambiense? Challenges of reservoir hosts and transmission cycles: Expect the
664 unexpected. *Parasite Epidemiol Control*, **6**, e00113.

665 15. Bisser,S., Lumbala,C., Nguertoum,E., Kande,V., Flevaud,L., Vatunga,G., Boelaert,M.,
666 Büscher,P., Josenando,T., Bessell,P.R., *et al.* (2016) Sensitivity and specificity of a
667 prototype rapid diagnostic test for the detection of Trypanosoma brucei gambiense
668 infection: A multi-centric prospective study. *PLoS Negl. Trop. Dis.*, **10**, e0004608.

669 16. Geerts,M., Van Reet,N., Leyten,S., Berghmans,R., S Rock,K., Ht Coetzer,T., E-A
670 Eyssen,L. and Büscher,P. (2020) Trypanosoma brucei gambiense-iELISA: a promising
671 new test for the post-elimination monitoring of human African trypanosomiasis. *Clin.*
672 *Infect. Dis.*, 10.1093/cid/ciaa1264.

673 17. Büscher,P., Gilleman,Q. and Lejon,V. (2013) Rapid diagnostic test for sleeping sickness.
674 *N. Engl. J. Med.*, **368**, 1069–1070.

675 18. Lejon,V., Büscher,P., Magnus,E., Moons,A., Wouters,I. and Van Meirvenne,N. (1998) A
676 semi-quantitative ELISA for detection of Trypanosoma brucei gambiense specific
677 antibodies in serum and cerebrospinal fluid of sleeping sickness patients. *Acta Trop.*,
678 **69**, 151–164.

679 19. Magnus,E., Vervoort,T. and Van Meirvenne,N. (1978) A card-agglutination test with
680 stained trypanosomes (CATT) for the serological diagnosis of T. b. gambiense
681 trypanosomiasis. In *Annales de la Société belge de Médecine Tropicale*. Societe Belge
682 de Medecine Tropicale, Vol. 58, pp. 169–176.

683 20. Van Meirvenne,N., Magnus,E. and Büscher,P. (1995) Evaluation of variant specific
684 trypanolysis tests for serodiagnosis of human infections with Trypanosoma brucei
685 gambiense. *Acta Trop.*, **60**, 189–199.

686 21. Gibson,W., Nemetschke,L. and Ndung'u,J. (2010) Conserved sequence of the TgsGP
687 gene in Group 1 Trypanosoma brucei gambiense. *Infection, Genetics and Evolution*, **10**,
688 453–458.

689 22. Radwanska,M., Magnus,E., Claes,F., Pérez-Morga,D., Magez,S., Pays,E. and
690 Büscher,P. (2002) Novel primer sequences for polymerase chain reaction-based
691 detection of Trypanosoma brucei gambiense. *The American Journal of Tropical*
692 *Medicine and Hygiene*, **67**, 289–295.

693 23. Compaoré,C.F.A., Ilboudo,H., Kaboré,J., Kaboré,J.W., Camara,O., Bamba,M.,
694 Sakande,H., Koné,M., Camara,M., Kaba,D., *et al.* (2020) Analytical sensitivity of
695 loopamp and quantitative real-time PCR on dried blood spots and their potential role in
696 monitoring human African trypanosomiasis elimination. *Exp. Parasitol.*, **219**, 108014.

697 24. Felu,C., Pasture,J., Pays,E. and Pérez-Morga,D. (2007) Diagnostic potential of a
698 conserved genomic rearrangement in the Trypanosoma brucei gambiense-specific
699 TGSGP locus. *Am. J. Trop. Med. Hyg.*, **76**, 922–929.

25. Agbo,E.E.C., Majiwa,P.A.O., Claassen,H.J.H. and Te Pas,M.F.W. (2002) Molecular variation of Trypanosoma brucei subspecies as revealed by AFLP fingerprinting. *Parasitology*, **124**, 349–358.

26. Jamonneau,V., Garcia,A., Ravel,S., Cuny,G., Oury,B., Solano,P., N'Guessan,P., N'Dri,L., Sanon,R., Frezil,J.L., *et al.* (2002) Genetic characterization of Trypanosoma brucei gambiense and clinical evolution of human African trypanosomiasis in Cote d'Ivoire. *Tropical Medicine and International Health*, **7**, 610–621.

27. Gibson,W. (2001) Molecular characterization of field isolates of human pathogenic trypanosomes. *Trop. Med. Int. Health*, **6**, 401–406.

28. Auty,H., Anderson,N.E., Picozzi,K., Lembo,T., Mubanga,J., Hoare,R., Fyumagwa,R.D., Mable,B., Hamill,L., Cleaveland,S., *et al.* (2012) Trypanosome Diversity in Wildlife Species from the Serengeti and Luangwa Valley Ecosystems. *PLoS Neglected Tropical Diseases*, **6**, e1828.

29. Biteau,N., Bringaud,F., Gibson,W., Truc,P. and Baltz,T. (2000) Characterization of Trypanozoon isolates using a repeated coding sequence and microsatellite markers. *Molecular and Biochemical Parasitology*, **105**, 187–202.

30. Agbo,E.C., Majiwa,P.A.O., Claassen,E.J.H.M. and Roos,M.H. (2001) Measure of Molecular Diversity within the Trypanosoma brucei Subspecies Trypanosoma brucei brucei and Trypanosoma brucei gambiense as Revealed by Genotypic Characterization. *Exp. Parasitol.*, **99**, 123–131.

31. Bromidge,T., Gibson,W., Hudson,K. and Dukes,P. (1993) Identification of Trypanosoma brucei gambiense by PCR amplification of variant surface glycoprotein genes. *Acta Tropica*, **53**, 107–119.

32. Richardson,J.B., Lee,K.-Y., Mireji,P., Enyaru,J., Sistrom,M., Aksoy,S., Zhao,H. and Caccone,A. (2017) Genomic analyses of African Trypanozoon strains to assess

725 evolutionary relationships and identify markers for strain identification. *PLoS Negl. Trop.*
726 *Dis.*, **11**, e0005949.

727 33. Schares,G. and Mehlitz,D. (1996) Sleeping sickness in Zaire: a nested polymerase chain
728 reaction improves the identification of Trypanosoma (Trypanozoon) brucei gambiense
729 by specific kinetoplast DNA probes. *Tropical Medicine & International Health*, **1**, 59–70.

730 34. Mathieu-Daudé,F., Bicart-See,A., Tibayrenc,M., Brenière,S.-F. and Bosseno,M.-F.
731 (1994) Identification of Trypanosoma brucei gambiense Group I by a Specific
732 Kinetoplast DNA Probe. *The American Journal of Tropical Medicine and Hygiene*, **50**,
733 13–19.

734 35. Lukes,J., Guilbride,D.L., Votýpka,J., Zíková,A., Benne,R., Englund,P.T., Voty,J. and
735 Zíkova,A. (2002) Kinetoplast DNA Network : Evolution of an Improbable Structure.
736 *Eukaryot. Cell*, **1**, 495–502.

737 36. Cooper,S., Wadsworth,E.S., Ochsenreiter,T., Ivens,A., Savill,N.J. and Schnaufer,A.
738 (2019) Assembly and annotation of the mitochondrial minicircle genome of a
739 differentiation-competent strain of Trypanosoma brucei. *Nucleic Acids Res.*, **47**, 11304–
740 11325.

741 37. Singh,N., Curran,M.D., Rastogil,A.K., Middleton,D. and Sundar,S. (1999) Diagnostic
742 PCR withLeishmania donovanispecificity using sequences from the variable region of
743 kinetoplast minicircle DNA. *Tropical Medicine & International Health*, **4**, 448–453.

744 38. Ceccarelli,M., Galluzzi,L., Diotallevi,A., Andreoni,F., Fowler,H., Petersen,C., Vitale,F.
745 and Magnani,M. (2017) The use of kDNA minicircle subclass relative abundance to
746 differentiate between Leishmania (L.) infantum and Leishmania (L.) amazonensis.
747 *Parasit. Vectors*, **10**, 239.

748 39. Borst,P., Fase-Fowler,F. and Gibson,W.C. (1987) Kinetoplast DNA of Trypanosoma
749 evansi. *Molecular and Biochemical Parasitology*, **23**, 31–38.

750 40. Van den Broeck,F., Savill,N.J., Imamura,H., Sanders,M., Maes,I., Cooper,S., Mateus,D.,
751 Jara,M., Adaui,V., Arevalo,J., *et al.* (2020) Ecological divergence and hybridization of
752 Neotropical Leishmania parasites. *Proc. Natl. Acad. Sci. U. S. A.*, **117**.

753 41. Geerts,M., Schnaufer,A. and Van den Broeck,F. (2021) rKOMICS: an R package for
754 processing mitochondrial minicircle assemblies in population-scale genome projects.
755 *BMC Bioinformatics*, **under review**.

756 42. Lanham,S.M. and Godfrey,D.G. (1970) Isolation of salivarian trypanosomes from man
757 and other mammals using DEAE-cellulose. *Exp. Parasitol.*, **28**, 521–534.

758 43. Sambrook,J. and Russell,D.W. (2006) Purification of nucleic acids by extraction with
759 phenol:chloroform. *CSH Protoc.*, **2006**.

760 44. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A.,
761 Garimella,K., Altshuler,D., Gabriel,S., Daly,M., *et al.* (2010) The genome analysis toolkit:
762 A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome*
763 *Res.*, **20**, 1297–1303.

764 45. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A.,
765 Keane,T., McCarthy,S.A., Davies,R.M., *et al.* (2021) Twelve years of SAMtools and
766 BCFtools. *Gigascience*, **10**.

767 46. Huson,D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data.
768 *Bioinformatics*, **14**, 68–73.

769 47. Li,D., Liu,C.M., Luo,R., Sadakane,K. and Lam,T.W. (2015) MEGAHIT: An ultra-fast
770 single-node solution for large and complex metagenomics assembly via succinct de
771 Bruijn graph. *Bioinformatics*, 10.1093/bioinformatics/btv033.

772 48. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local
773 alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

774 49. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ
775 preprocessor. *Bioinformatics*, **34**, i884–i890.

776 50. Ray,D.S. (1989) Conserved sequence blocks in kinetoplast minicircles from diverse
777 species of trypanosomes. *Mol. Cell. Biol.*, **9**, 1365–1367.

778 51. Rognes,T., Flouri,T., Nichols,B., Quince,C. and Mahé,F. (2016) VSEARCH: A versatile
779 open source tool for metagenomics. *PeerJ*, **4**, e2584.

780 52. Waterhouse,A.M., Procter,J.B., Martin,D.M.A., Clamp,M. and Barton,G.J. (2009) Jalview
781 Version 2—a multiple sequence alignment editor and analysis workbench.
782 *Bioinformatics*, **25**, 1189–1191.

783 53. Van Reet,N., Patient Pyana,P., Dehou,S., Bebronne,N., Deborggraeve,S. and
784 Büscher,P. (2021) Single nucleotide polymorphisms and copy-number variations in the
785 Trypanosoma brucei repeat (TBR) sequence can be used to enhance amplification and
786 genotyping of Trypanozoon strains. *PLoS One*, **16**, e0258711.

787 54. Picozzi,K., Carrington,M. and Welburn,S.C. (2008) A multiplex PCR that discriminates
788 between Trypanosoma brucei brucei and zoonotic T. b. rhodesiense. *Exp. Parasitol.*,
789 **118**, 41–46.

790 55. Pyana Pati,P., Van Reet,N., Mumba Ngoyi,D., Ngay Lukusa,I., Karhemere Bin
791 Shamamba,S. and Büscher,P. (2014) Melarsoprol sensitivity profile of Trypanosoma
792 brucei gambiense isolates from cured and relapsed sleeping sickness patients from the
793 Democratic Republic of the Congo. *PLoS Negl. Trop. Dis.*, **8**, e3212.

794 56. Simpson,L. (1987) The mitochondrial genome of kinetoplastid protozoa: genomic
795 organization, transcription, replication, and evolution. *Annu. Rev. Microbiol.*, **41**, 363–
796 382.

797 57. Stuart,K. (1979) Kinetoplast DNA OF Trypanosoma brucei: physical map of the

maxicircle. *Plasmid*, **2**, 520–528.

58. Steinert,M. and Van Assel,S. (1980) Sequence heterogeneity in kinetoplast DNA: reassociation kinetics. *Plasmid*, **3**, 7–17.

59. Pollard,V.W., Rohrer,S.P., Michelotti,E.F., Hancock,K. and Hajduk,S.L. (1990) Organization of minicircle genes for guide RNAs in Trypanosoma brucei. *Cell*, **63**, 783–790.

60. Hong,M. and Simpson,L. (2003) Genomic organization of Trypanosoma brucei kinetoplast DNA minicircles. *Protist*, **154**, 265–279.

61. Jasmer,D.P. and Stuart,K. (1986) Sequence organization in African trypanosome minicircles is defined by 18 base pair inverted repeats. *Mol. Biochem. Parasitol.*, **18**, 321–331.

62. Savill,N.J. and Higgs,P.G. (1999) A theoretical study of random segregation of minicircles in trypanosomatids. *Proceedings of the Royal Society B: Biological Sciences*, **266**, 611–620.

63. Goodhead,I., Capewell,P., Wendi Bailey,J., Beament,T., Chance,M., Kay,S., Forrester,S., MacLeod,A., Taylor,M., Noyes,H., *et al.* (2013) Whole-genome sequencing of Trypanosoma brucei reveals introgression between subspecies that is associated with virulence. *MBio*, **4**.

64. Tihon,E., Imamura,H., Dujardin,J.-C., Van Den Abbeele,J. and Van den Broeck,F. (2017) Discovery and genomic analyses of hybridization between divergent lineages of Trypanosoma congolense, causative agent of Animal African Trypanosomiasis. *Mol. Ecol.*, **26**.

65. Van den Broeck,F., Tavernier,L.J.M., Vermeiren,L., Dujardin,J.C. and Van Den Abbeele,J. (2018) Mitonuclear genomics challenges the theory of clonality in

822 Trypanosoma congolense: Reply to Tibayrenc and Ayala. *Mol. Ecol.*,
823 10.1111/mec.14809.

824 66. Gibson,W. and Garside,L. (1990) Kinetoplast DNA minicircles are inherited from both
825 parents in genetic hybrids of Trypanosoma brucei. *Mol. Biochem. Parasitol.*, **42**, 45–53.

826 67. Gibson,W., Crow,M. and Kearns,J. (1997) Kinetoplast DNA minicircles are inherited from
827 both parents in genetic crosses of Trypanosoma brucei. *Parasitol. Res.*, **83**.

828 68. Rusman,F., Tomasini,N., Yapur,N.-F., Puebla,A.F., Ragone,P.G. and Diosque,P. (2019)
829 Elucidating diversity in the class composition of the minicircle hypervariable region of
830 Trypanosoma cruzi: New perspectives on typing and kDNA inheritance. *PLoS Negl.*
831 *Trop. Dis.*, **13**, e0007536.

832 69. Balyeidhusa,A.S.P., Kironde,F.A.S. and Enyaru,J.C.K. (2012) Apparent lack of a
833 domestic animal reservoir in Gambiense sleeping sickness in northwest Uganda. *Vet.*
834 *Parasitol.*, **187**, 157–167.

835 70. Cordon-Obras,C., Berzosa,P., Ndong-Mabale,N., Bobuakasi,L., Buatiche,J.N., Ndongo-
836 Asumu,P., Benito,A. and Cano,J. (2009) Trypanosoma brucei gambiense in domestic
837 livestock of Kogo and Mbini foci (Equatorial Guinea). *Trop. Med. Int. Health*, **14**, 535–
838 541.

839 71. Cordon-Obras,C., García-Estébanez,C., Ndong-Mabale,N., Abaga,S., Ndongo-
840 Asumu,P., Benito,A. and Cano,J. (2010) Screening of Trypanosoma brucei gambiense
841 in domestic livestock and tsetse flies from an insular endemic focus (Luba, Equatorial
842 Guinea). *PLoS Negl. Trop. Dis.*, **4**, e704.

843 72. Cordon-Obras,C., Rodriguez,Y.F., Fernandez-Martinez,A., Cano,J., Ndong-Mabale,N.,
844 Ncogo-Ada,P., Ndongo-Asumu,P., Aparicio,P., Navarro,M., Benito,A., *et al.* (2015)
845 Molecular evidence of a Trypanosoma brucei gambiense sylvatic cycle in the human
846 african trypanosomiasis foci of Equatorial Guinea. *Front. Microbiol.*, **6**, 765.

847 73. Simo,G., Fongho,P., Farikou,O., Ndjeuto-Tchouli,P.I.N., Tchouomene-Labou,J.,
848 Njiokou,F. and Asonganyi,T. (2015) Trypanosome infection rates in tsetse flies in the
849 'silent' sleeping sickness focus of Bafia in the Centre Region in Cameroon. *Parasit.*
850 *Vectors*, **8**, 528.

851 74. Umeakuana,P.U., Gibson,W., Ezeokonkwo,R.C. and Anene,B.M. (2019) Identification of
852 Trypanosoma brucei gambiense in naturally infected dogs in Nigeria. *Parasit. Vectors*,
853 **12**, 420.

854 75. Vourchakbé,J., Tiofack,A.A.Z., Mbida,M. and Simo,G. (2020) Trypanosome infections in
855 naturally infected horses and donkeys of three active sleeping sickness foci in the south
856 of Chad. *Parasit. Vectors*, **13**, 323.

857 76. Birhanu,H., Gebrehiwot,T., Goddeeris,B.M., Büscher,P. and Van Reet,N. (2016) New
858 Trypanosoma evansi Type B Isolates from Ethiopian Dromedary Camels. *PLOS*
859 *Neglected Tropical Diseases*, **10**, e0004556.

860 77. Birhanu,H., Fikru,R., Said,M., Kidane,W., Gebrehiwot,T., Hagos,A., Alemu,T., Dawit,T.,
861 Berkvens,D., Goddeeris,B.M., *et al.* (2015) Epidemiology of Trypanosoma evansi and
862 Trypanosoma vivax in domestic animals from selected districts of Tigray and Afar
863 regions, Northern Ethiopia. *Parasites & Vectors*, **8**.

864 78. Schnaufer,A., Panigrahi,A.K., Panicucci,B., Igo R.P.,J., Salavati,R. and Stuart,K. (2001)
865 An RNA ligase essential for RNA editing and survival of the bloodstream form of
866 Trypanosoma brucei. *Science*, **291**, 2159–2162.

867 79. Zíková,A., Schnaufer,A., Dalley,R.A., Panigrahi,A.K. and Stuart,K.D. (2009) The
868 F(0)F(1)-ATP synthase complex contains novel subunits and is essential for procyclic
869 Trypanosoma brucei. *PLoS Pathog.*, **5**, e1000436.

870 80. Simpson,L., Douglass,S.M., Lake,J.A., Pellegrini,M. and Li,F. (2015) Comparison of the
871 mitochondrial genomes and steady state transcriptomes of two strains of the

872 trypanosomatid parasite, leishmania tarentolae. *PLoS Negl. Trop. Dis.*, **9**, e0003841.

873 81. Mehlitz,D. (1986) Le réservoir animal de la maladie du sommeil à Trypanosoma brucei
874 gambiense.

875 82. Paindavoine (1988) Variabilité génétique chez Trypanosoma brucei.