Congopain genes diverged to become specific to Savannah, Forest and Kilifi subgroups of *Trypanosoma congolense,* and are valuable for diagnosis, genotyping and phylogenetic inferences

Adriana C. Rodrigues[a#]; Paola A. Ortiz[a#]; André G. Costa-Martins[a]; Luis Neves[b,c]; Herakles A. Garcia[a,d]; João M. P. Alves[a]; Erney P. Camargo[a] ;Silvia C. Alfieri[a]; Wendy Gibson[e] & Marta M. G. Teixeira[a] *

# Both authors contributed equally to this work

[a] Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, SP, Brazil; [b] Biotechnology Centre, Eduardo Mondlane University, Maputo, Mozambique; [c] Faculty of Veterinary Science, University of Pretoria, South Africa; [d] Department of Veterinary Pathology, Faculty of Veterinary Sciences, Central University of Venezuela, Maracay, Venezuela; [e] School of Biological Sciences University of Bristol, Bristol BS8 1UG, UK.

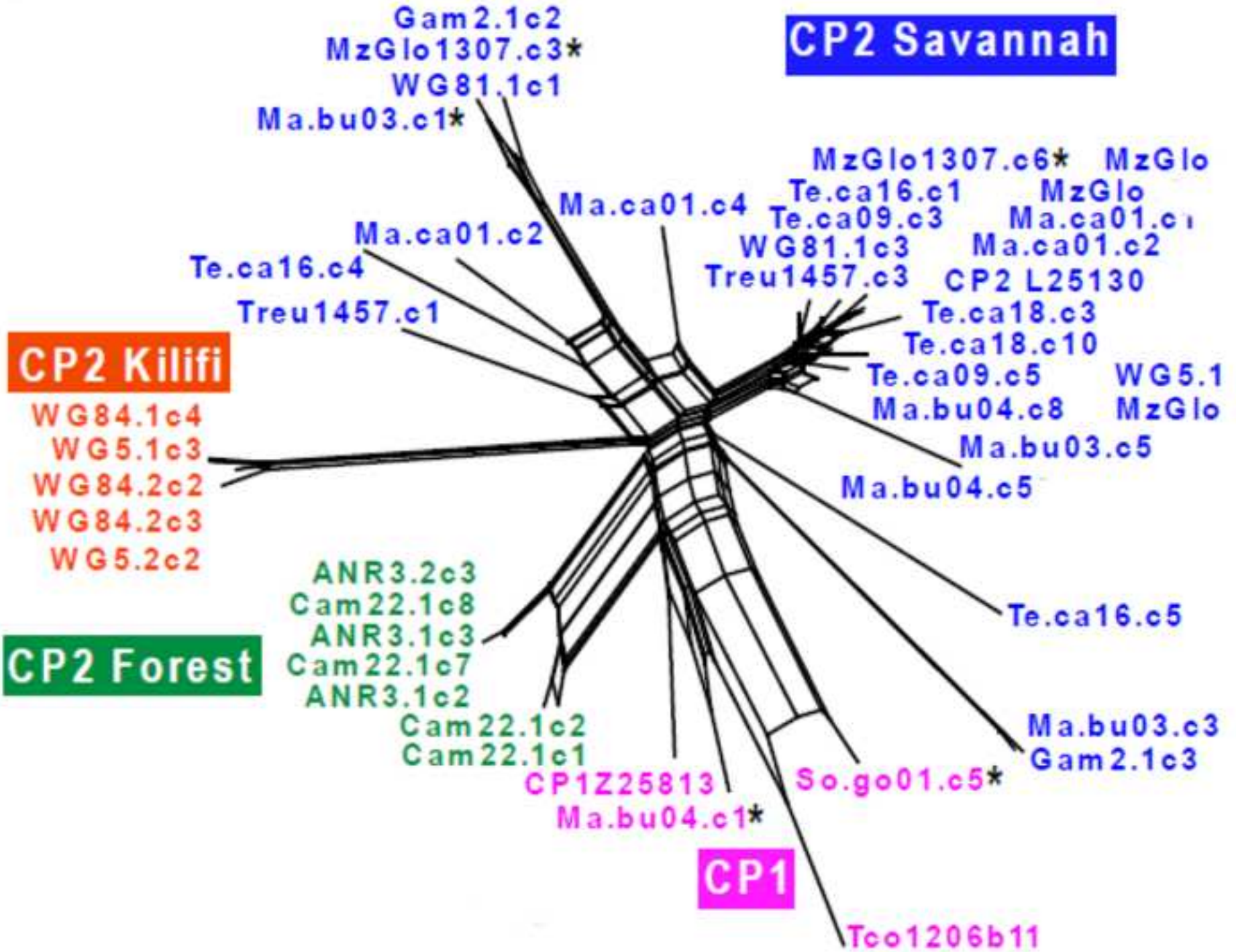
* corresponding author: Tel. 551130917429, FAX 551130917417; *Email address*: mmgteix@icb.usp.br

Abstract

Trypanosoma congolense is the most important agent of nagana, a wasting livestock trypanosomosis in sub-Saharan Africa. This species is a complex of three subgroups (Savannah, Forest and Kilifi) that differ in virulence, pathogenicity, drug resistance, vectors, and geographical distribution. Congopain, the major Cathepsin L-like cysteine protease (CP2) of *T. congolense,* has been extensively investigated as a pathogenic factor and target for drugs and vaccines, but knowledge about this enzyme is mostly restricted to the reference strain IL3000, which belongs to the Savannah subgroup. In this work we compared sequences of congopain genes from IL3000 genome database and isolates of the three subgroups of *T. congolense*. Results demonstrated that the congopain genes diverged into three subclades consistent with the three subgroups within *T. congolense*. Laboratory and field isolates of Savannah exhibited a highly polymorphic repertoire both inter- and intra-isolates: sequences sharing the archetypical catalytic triad clustered into SAV1-SAV3 groups, whereas polymorphic sequences that, in general, exhibited unusual catalytic triad (variants) assigned to SAV4 or not assigned to any group. Congopain homologous genes from Forest and Kilifi isolates showed, respectively, moderate and limited diversity. In the phylogenetic tree based on congopain and homologues, Savannah was closer to Forest than to Kilifi. All *T. congolense* subgroup nested into a single clade, which together with the sister clade formed by homologues from *T. simiae* and *T. godfreyi* formed a clade supporting the subgenus *Nannomonas.* A single PCR targeting congopain sequences was developed for the diagnosis of *T. congolense* isolates of the three subgroups. Our findings demonstrated that congopain genes are valuable targets for the diagnosis, genotyping, and phylogenetic and taxonomic inferences among *T. congolense* isolates and other members of the subgenus *Nannomonas.*

Highlights

- Genetic repertoire and genealogy of Congopain (CP2) encoding genes
- Diagnosis and genotyping of *T. congolense* Savannah, Forest, and Kilifi based on congopain sequences
- The repertoire of congopain encoding genes is subgroup-specific, and highly heterogeneous in Savannah, moderate in Forest and more homogeneous in the isolates of the Kilifi subgroup.


Gam2.1c2
MzGlo1307.c3*
WG81.1c1
Ma.bu03.c1*
CP2 Savannah
MzGlo1307.c6*
MzGlo
Te.ca16.c1
MzGlo
Ma.ca01.c4
Te.ca09.c3
Ma.ca01.c
Ma.ca01.c2
WG81.1c3
Ma.ca01.c2
Te.ca16.c4
Treu1457.c3
CP2 L25130
Treu1457.c1
Te.ca18.c3
Te.ca18.c10
CP2 Kilifi
Te.ca09.c5
WG5.1
Ma.bu04.c8
MzGlo
WG84.1c4
WG5.1c3
Ma.bu03.c5
WG84.2c2
Ma.bu04.c5
WG84.2c3
WG5.2c2
ANR3.2c3
Te.ca16.c5
Cam22.1c8
ANR3.1c3
CP2 Forest
Cam22.1c7
ANR3.1c2
Ma.bu03.c3
Cam22.1c2
Gam2.1c3
Cam22.1c1
CP1Z25813
So.go01.c5*
Ma.bu04.c1*
CP1
Tco1206b11

Keywords: *T. congolense*; Animal trypanosomosis, Congopain; cysteine protease; cathepsin L; Savannah; Forest; Kilifi; genotyping; diagnosis

1. Introduction

Livestock trypanosomosis (Nagana) is a chronic wasting disease that poses a major constraint to livestock productivity in sub-Saharan Africa. The causative agents are tsetse-borne trypanosomes, of which *Trypanosoma congolense* is the most prevalent and widespread. *T. congolense* comprises three morphologically indistinguishable but genetically recognisable subgroups — Savannah, Forest, and Kilifi (Gibson, 2002, 2007) — which vary in virulence, pathogenicity, and geographical distribution. The Savannah and Forest subgroups were originally evidenced by isoenzymes (Young and Godfrey, 1983; Gashumba et al., 1988), whereas RFLP and karyotyping disclosed a further subgroup: Kenya Coast or Kilifi (Knowles et al., 1988; Majiwa et al., 1985, 1986). Methods based on repetitive DNA sequences were developed to identify these three subgroups (Gibson et al. 1988; Masiga et al., 1992), all further corroborated by other molecular markers (Gibson et al., 2001; Desquesnes et al., 2001; Gibson, 2007, 2011; Adams et al., 2010).

The Savannah subgroup of *T. congolense* is the most widespread. Field investigations associated the Savannah subgroup with a range of tsetse (*Glossina*) species (morsitans, palpalis and fusca groups) and a broad range of ungulates and carnivore hosts across the whole of sub-Saharan Africa. In contrast, *T. congolense* Forest appears to be largely restricted to the palpalis group of tsetse flies and, consequently, to riverine-forest biomes. It has been recorded in pigs, goats, cattle, and dogs in West and Central Africa, and also at low prevalence in parts of East Africa. *T. congolense* Kilifi was first isolated from Kenya but has since been widely reported throughout south-eastern Africa; it is associated with tsetse of the morsitans group and has been reported in cattle, sheep, and goats (Majiwa et al., 1993, 1985; Reinfenberg et al., 1997; Knowles et al., 1988; Masiga et al., 1996; Njiru et al., 2004; Malele et al., 2011; Simo et al., 2012; 2013). Infections with a mixture of subgroups are frequent in ungulates and tsetse flies: co-infections with Savannah and Forest subgroups are common in West and Central Africa (Seck et al., 2010; Simo et al., 2012,2013), whereas Savannah and Kilifi subgroups mixed infections occur in East and South Africa (Mekata et al., 2008; Mamabolo et al., 2009). The three subgroups coexist in Zambia, Kenya, and Tanzania (Njiru et al., 2004; Mekata et al., 2008; Malele et al., 2011). Infection of susceptible zebuine cattle revealed Kilifi as non-pathogenic, Forest of low pathogenicity, and Savannah as the most virulent subgroup (Bengaly et al., 2002a,b). Isolates of Savannah differed markedly in virulence and drug resistance, even in the same location (Seck et al., 2010; Van den Bossche et al., 2011; Vitoulay et al., 2011; Moti et al., 2012).

Phylogenetic analyses based on SSU rRNA and gGAPDH genes showed that the three subgroups of *T. congolense* clustered together, forming a clade within a monophyletic assemblage corresponding to the subgenus *Nannomonas* that also includes *T. simiae* and *T. godfreyi* (Hamilton et al., 2004). However, these genes were

unable to resolve the relationships among the subgroups of *T. congolense*. Previous studies have suggested a closer relationship between Savannah and Forest than between these subgroups and Kilifi. *T. congolense* Savannah and Forest share lengths of the major satellite DNA repeat, kDNA minicircles, and mini-exon gene repeats and also polymorphisms in the beta tubulin and rRNA genes, whereas these markers are significantly different in Kilifi (Garside and Gibson, 1995). Sequence analysis of the gene coding a major surface glycoprotein, glutamate- and alanine-rich protein (GARP), demonstrated a similar relationship among the subgroups of *T. congolense* (Asbeck et al., 2000).

The Cathepsin L (CATL)-like cysteine proteases (CPs) have been extensively studied in trypanosomes due to their important roles in pathogenicity, virulence, cell differentiation, and immune evasion. These CPs belong to the papain family (clan CA, family C1) that typically consists of a signal peptide, pro-peptide, catalytic domain (cd), and a C-terminal extension of variable size unique to kinetoplastid CATL (Sajid and Mckerrow 2002; Atkinson et al., 2009; Alvarez et al., 2012). The two main CATL-like CPs (CP1 and CP2) characterized in *T. congolense* can be distinguished by polymorphisms in the cds, which result in functional differences. CP2, usually referred to as congopain, is the major CP of *T. congolense* (Fish et al., 1995; Jaye et al., 1993; Authié et al., 1992, 1994, 2001; Boulangé et al., 2001). Congopain is an important antigen in the development of vaccines and target for chemotherapy (Authié et al., 1992; 1994; 2001; Boulangé et al., 2001; Huson et al., 2009; Kateregga et al., 2013; Lalmanach et al., 2002). In addition, an unusual CP identified in *T. congolense* differs from CP1 and CP2 by a serine replacing cysteine in the catalytic triad (Downey and Donelson, 1999; Pillay et al., 2010); in addition, other variants have been reported among the CP enzymes of this species (Kakundi 2008).

Our studies of other trypanosome species demonstrated that CATL-like genes are useful markers for diagnosis, genotyping, and phylogenetic reconstruction at species and genotype levels in *T. vivax*, *T. theileri*, *T. rangeli*, and *T. cruzi* and allied species (Cortez et al., 2009; Garcia et al., 2011a,b; Lima et al., 2012; Ortiz et al., 2009; Rodrigues et al., 2010). While congopain have been very well characterised in the laboratory strain IL3000 of *T. congolense* Savannah, there is an absence of data on CPs from Forest and Kilifi subgroups. In this study, we characterised the catalytic domains of genes encoding congopain from *T. congolense* isolates of Savannah, Forest and Kilifi subgroups, including samples from west, central, and east Africa. Our main goals were: a) to compare the congopain genetic repertoires of the three subgroups; b) to infer the genealogy and to evaluate the suitability of congopain sequences for diagnosis, genotyping, and population structure analysis; c) to infer the phylogenetic relationships among isolates from the three subgroups of *T. congolense, T. simiae* and *T. godfreyi* (subgenus *Nannomonas*) based on congopain and homologous genes.

2. Material and methods

2.1. *Trypanosomes and PCR amplification of the catalytic domains of congopain genes*

Table 1. Geographic and host origin of *T. congolense* isolates, CATL-like sequences of *T. congolense*, and homologous sequences from other trypanosome species.

Trypanosoma	Subgroup	Host Origin	Geographic origin	GenBank and GeneDB accession number	CATL-like sequence
T. congolense					
Laboratory isolates					
Cam22	Forest	Goat	Cameroon	KF414001-KF414024	Forest
ANR3	Forest	Tsetse	The Gambia	KF414025-KF414036	Forest
WG5	Kilifi	Goat	Kenya	KF413898-KF413922	Kilifi
WG84	Kilifi	Sheep	Kenya	KF413923-KF413933	Kilifi
WG81	Savannah	-	Kenya	KF413934-KF413940	SAV1, SAV2, SAV3
Gam2	Savannah	Cow	The Gambia	KF413941-KF413948	SAV2, SAV3
IL1180	Savannah	Lion	Tanzânia	KF413977-KF413983	SAV1, SAV2
TREU1457	Savannah	Cow	Nigeria	KF413949-KF413951	SAV1, SAV2
TRUM 183 [a]	Savannah	-	-	AF139913	SAV1
CP2 (archetype)	Savannah	-	-	L25130	SAV1
IL3000 (CP1 archetype)	Savannah	Cow	Kenya	Z25813	CP1
IL3000 [b]	Savannah	Cow	Kenya	TcIL3000: 0.28270, 0.53250, 0.12120, 0.11840, 0.44210, 0.17860, 0.25670	CP1
				TcIL3000.0.26770	SAV1
				TcIL3000.0.31720	SAV2
				TcIL3000.0.60020	SAV3
				TcIL3000: 0.26780, 0.49190, 0.47610, 0.18880	SAV4
				TcIL3000: 0.28390, 0.55820, 0.55830, 0.55780, 0.37240, 0.31730	SAVna [c]
Field isolates					
Ma.ca01	Savannah	Cattle	Mozambique Maputo	KF413965-KF413976	SAV1, SAV3
Ma.ca06	Savannah	Cattle	Mozambique Maputo	KF413984-KF413986	SAV1
Ma.bu03	Savannah	Buffalo	Mozambique Maputo	KF414051-KF414053	Savannah [d]
Ma.bu04	Savannah	Buffalo	Mozambique Maputo	KF414048-KF414048	Savannah [d]
Ma.bu05	Savannah	Buffalo	Mozambique Maputo	KF413987-KF413994	SAV1, SAV2, SAV3
Te.ca09	Savannah	Cow	Mozambique Tete	KF413957-KF413960	SAV1
Te.ca016	Savannah	Cow	Mozambique Tete	KF413952-KF413956	SAV1
Te.ca018	Savannah	Cow	Mozambique Tete	KF413961-KF413964	SAV1, SAV3
So.go01	Savannah	Goat	Mozambique Sofala	KF413995-KF414000, KF414054	SAV1, SAV4
MzGlo92	Savannah	Tsetse	Mozambique Sofala	KF414041-KF414042	SAV1, SAV3
MzGlo93	Sanannah	Tsetse	Mozambique Sofala	KF414043-KF414047	Savannah [d]
Other species					
T. vivax	Y486	Cow	Nigeria	Tviv534d01.q1k7, Tviv290f05.q1k11 [d]	
T. vivax	TviMzNy	Nyala	Mozambique	EU753814	
T. b. brucei	427	Sheep	Uganda	EU753820	
T. b. brucei	Star	-	-	X16465	
T. b. gambiense	TB26	Pig	Congo	EU753821	
T. b. rhodesiense	AntTat1.12	Human	-	EU753822	
T. equiperdum	Botat1.	Horse	-	EU753819	
T. evansi	Ted2	Dog	Brazil	EU753818	
T. simiae	Ken14	Tsetse	The Gambia	KF414037-KF414038	
T. godfreyi	Ken7	Tsetse	The Gambia	KF414039-KF414040	
T. rangeli	(AM80)	Human	Brazil	FJ997560	
T. theileri	Tthc12	Cow	Brazil	GU299366	
T. carassi	-	Fish	-	EF538803	

a - CP sequence from T. congolense TRUM 183 retrivied from Genebank; b - *T. congolense* sequence data obtained from the Sanger Institute website at http://www.sanger.ac.uk/Projects/T_congolense/; c - SAVna, na = sequence not assigned to SAV1-4 groups; d, diagnosed as *T. congolense* and genotyped into the Savannah subgroup by the method of TcoCATL-PCR followed by sequencing of the amplified fragments; The acession numbers of sequences determined in this work and deposited at the Genbank are KF413898-414047.

Table 2. Biogeographical characteristics and congopain repertoires of *T. congolense* isolates from Savannah, Forest and Kilifi subgropus.

***T. congolense* Subgroup**	**Distribution**	**Host range**	**Tsetse transmission**	**Pathogenicity [a]**	**Drug susceptibility [b]**	**CP2 repertoire**
Savannah	Tropical Africa	numerous species of ungulates and other mammals	morsitans, palpalis and fusca groups	High	Susceptible Resistant	SAV1-SAV4, SAVna [c]
Forest	West and Central Africa	pigs, goats, cattle, dogs	palpalis group	Low	Susceptible	Forest CP2
Kilifi	Southeast Africa	Cattle, sheep, goats	morsitans group	Non-pathogenic	Not-determined	Kilifi CP2

a - Differential pathogenicity among *T. congolense* subgroups, and variable virulence within Savannah, revealed by studies in mice and cattle (Bengaly et al., 2002a,b; Masumu et al., 2006). b - Resistance to diminazene has been reported for Savannah in field infected cattle and mice (Van den Bossche et al., 2011; Moti et al., 2012). c - SAVna, na = sequence not assigned to SAV1-4 groups.

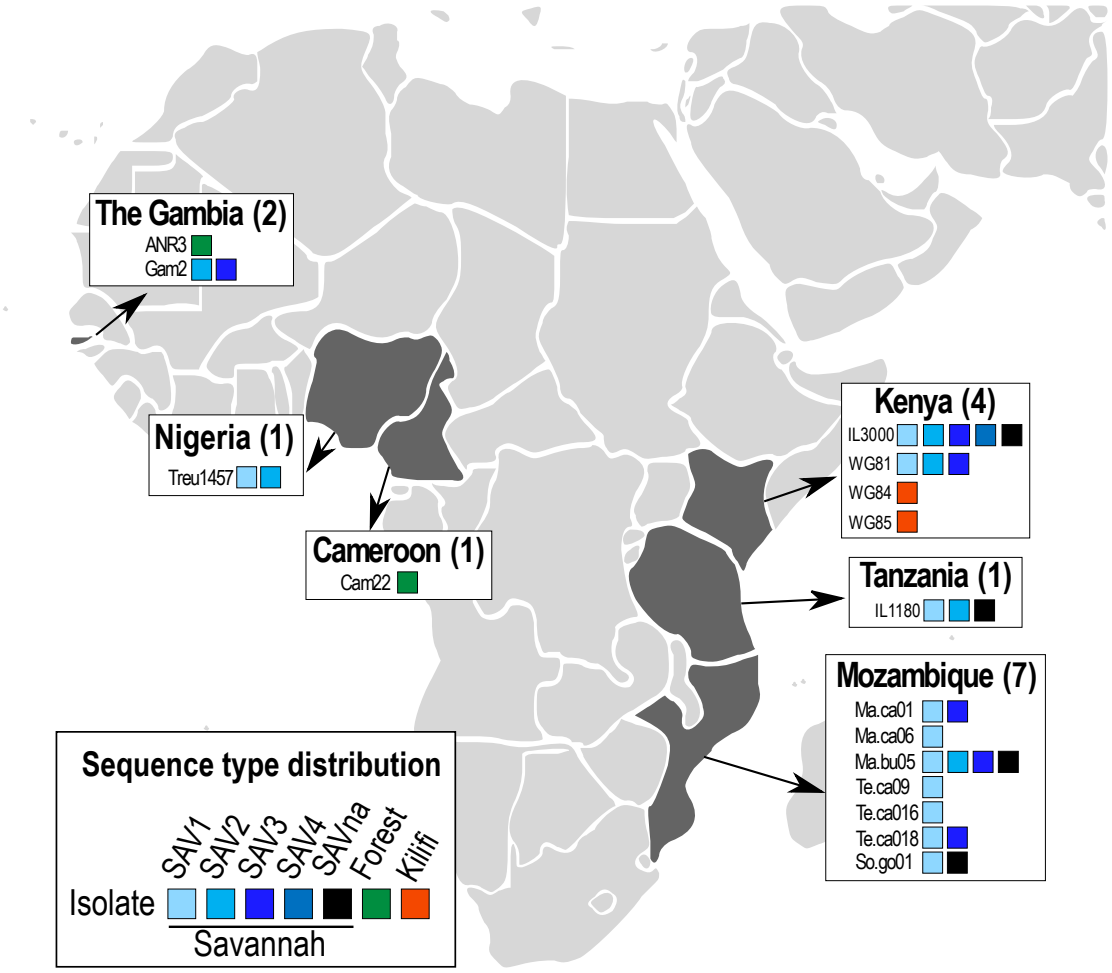


Fig. 1. Geographical origin of the T. congolense Savannah, Forest, and Kilifi subgroups characterised in this study and the respective genetic repertoire of CP2 sequences. Seven groups of CATL-like genes (indicated by different colours) were defined within T. congolense by the genealogy of the catalytic domain sequences and specific amino acid signatures: SAV1-SAV4 plus SAVna (= sequences not assigned to any group) for Savannah, and only one group of sequences each for the Forest and Kilifi subgroups. The number of isolates from each country is indicated within parentheses.

T. congolense clone IL3000 represents the Savannah subgroup and was selected for the genome project. This clone, obtained in 1966, has been maintained for decades by successive passages in mice (Gibson, 2013). Additional *T. congolense* isolates characterised in this study were from different hosts and geographic origins (Table 1). DNA templates were prepared from *T. congolense* Savannah (WG81, Gam2, IL1180, TREU1457), Forest (ANR3, Cam22), and Kilifi (WG5, WG84) laboratory stocks. Field-collected blood samples were obtained from infected cattle, water buffalo, and goats from Mozambique collected from endemic settlings in the provinces of Maputo (isolates Ma.ca01, Ma.ca06, Ma.bu03, Ma.bu05, Ma.bu04), Sofala (So.go01) and Tete (isolates Te.ca09, Te.ca016, Te.ca018) in the Southern, Central, and Northern regions, respectively (Table 2, Fig. 1). CATL homologous sequences from *T. simiae* (Ken14) and *T. godfreyi* (Ken7) were determined in this study, and included in the phylogenetic analysis of *Trypanosoma*. The DNA from *T. congolense* isolates, *T. simiae* (Ken14) and *T. godfreyi* (Ken7) were used for the PCR-amplification of partial sequences (477 bp) corresponding to the cds of CP (Fig. 2) using primers and conditions previously described for the amplification of CATL-like encoding genes from several trypanosome species (Lima et al., 1994; Cortez et al., 2009; Garcia et al., 2011a,b; Lima et al., 2012; Rodrigues et al., 2010; Ortiz et al., 2009). From 6 to 10 cloned cdCP sequences were determined for each isolate included in this study.

2.2. Genealogies of CATL-like nucleotide and deduced amino acid sequences

The genealogies of sequences encoding CATL-like enzymes were inferred by analyses of either nucleotide or deduced amino acid sequence data sets. The present study benefited from congopain sequences retrieved from the on-going genome project of *T. congolense* IL3000 (http://www.genedb.org). In addition to sequences determined in this study and those retrieved from the genome project, we included in the alignments: a) the prototype sequences available from Genbank of CP1 (Z25813), CP2 (L25130), and the variant CP2-like from the strain TRUM 183 (AF139913); b) CATL-like genes determined in this study from *T. simiae* and *T. godfreyi;* c) CATL-like genes from *T. b. brucei*, *T. b. rhodesiense*, *T. b. gambiense*, *T. evansi*, *T. equiperdum*, *T. vivax*, *T. theileri*, *T. rangeli*, and *T. carassii,* all from previous studies (Table 2).

The alignments created for this study included: a) amino acid sequences comprising the pre-, pro-, and catalytic domains (cd) (Fig. 2) of 12 genes encoding CP2, and 7 encoding CP1, all from the IL3000 genome; b) 72 cd amino acid sequences (477 bp) from Savannah (5 isolates), including sequences from IL3000 (13), livestock blood samples and one from tsetse fly from Mozambique (CP1 sequences were excluded from this alignment); c) amino acid sequences of cdCP2 from 7 isolates of Savannah, Forest and Kilifi subgroups, and homologous genes from *T. simiae*, *T. godfreyi,* and other species of *Trypanosoma;* d) sequences of 211 bp fragments from cdCP genes obtained by TcoCATL-PCR (see below). Genealogies were inferred with the neighbour-joining (NJ) algorithm in Mega 5 software, treating gaps as deletions, and maximum-likelihood (ML) analysis was carried out

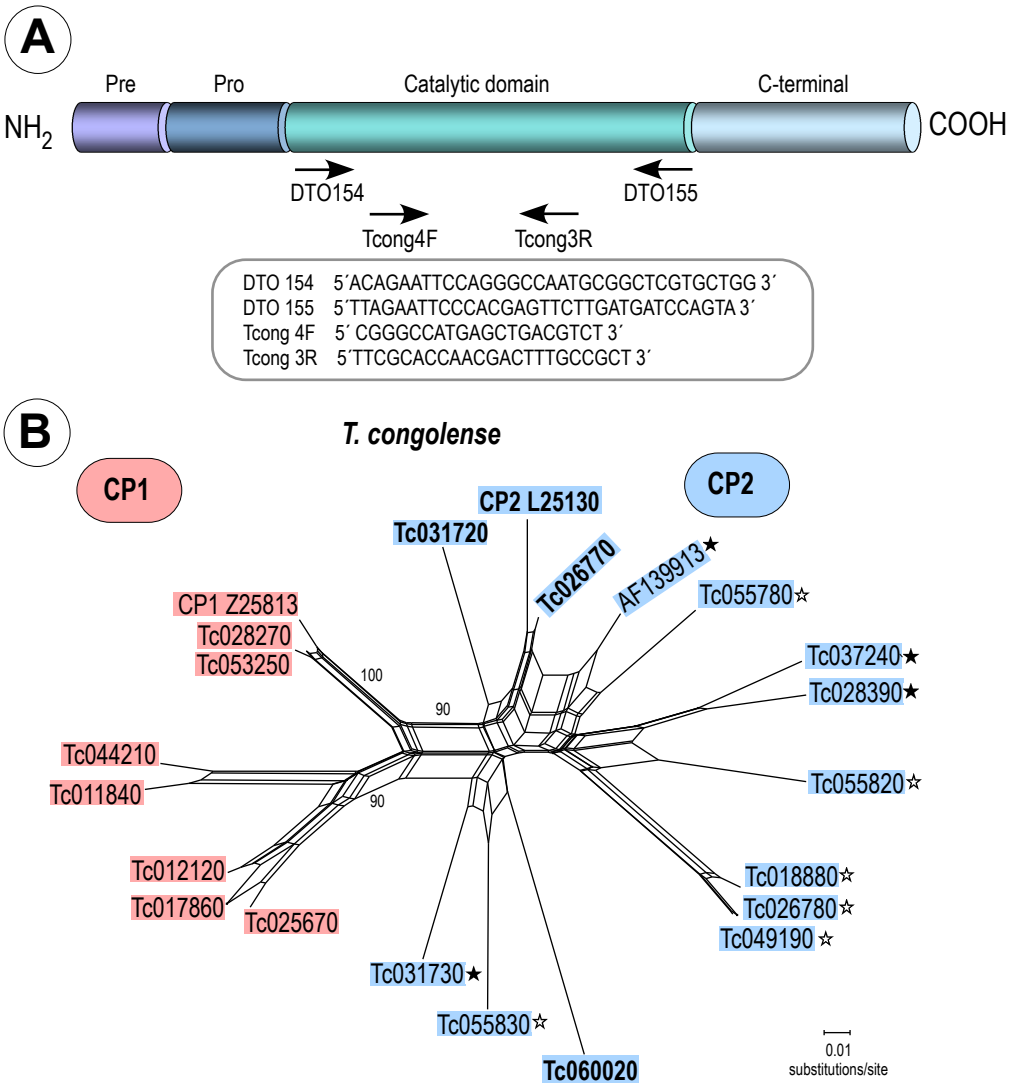


Fig. 2. (A) Schematic representation of the congopain gene indicating the protein domains and the primers employed for PCR amplification of the catalytic domains (primers DTO 154 and DTO 155) and the T. congolense specific fragment of 211 bp (primers Tcong4F and Tcong 3R).
(B) Network genealogy of CP predicted amino acid sequences (pre-, pro-, and catalytic domain) from the genome database of T. congolense IL3000 and from GenBank, inferred using the Neighbour-Net method with the K2P parameter and 1000 bootstrap replicates. The archetypical CP1 and CP2 (in bold) and variant sequences: the CP2-like and new CP2-like sequences are indicated by filled and unfilled stars, respectively.

using RAxML v.7.0 as described in our previous studies (Garcia et al., 2011b; Lima et al., 2012; Rodrigues et al., 2010).Phylogenetic trees were constructed using the Neighbor-Net method with Kimura 2 parameters implemented in Splits Tree4 V4.10 (Huson and Bryant, 2006). Internode support was estimated by performing 100 bootstrap replicates using the same parameters optimised for tree inferences. To provide a visual representation of the distance matrix, we used the multidimensional scaling (MDS) plot with two dimensions (2D). The MDS statistical analysis and graphing were performed using the Bios2mds package of the R language and environment for statistical computing (Pelé et al., 2012). To find conserved motifs, "Multiple EM for Motif Elicitation" (MEME) version 3.5.4 (Bailey et al., 2006) was used. The parameters used for the analysis were number of repetitions – any, maximum number of motifs – 50, and optimum width of motif ≥3 and ≤5.

2.3. Codon usage and recombination analyses.

The ratio of non-synonymous to synonymous (dN/dS) amino acid changes was calculated according to Yang and Nielsen (2000) using PAML v.4.2 software to infer relative selection pressures (Yang, 2007). A positive value for this test indicates an overabundance of nonsynonymous substitutions, and in this case, the probability of rejecting the null hypothesis of neutral evolution (P-value) is calculated. The existence of putative recombination events in the genes encoding congopains was investigated using the RDP3 programme (Martin et al., 2010). All eight methods available were employed, and recombination events were considered valid if detected by at least four methods, with a minimum significance P-value of 0.05.

2.4. Standardisation of PCR targeting CATL-like sequences for the diagnosis of T. congolense

An alignment including nucleotide sequences of cdCATL genes from *T. congolense* Savannah, Forest, Kilifi and homologues from closely related *T. simiae* and *T. godfreyi* and other trypanosome species was used to design the *T. congolense*-specific primers Tco4F and Tco3R (Fig. 2A). A PCR assay, designated TcoCATL-PCR, was developed for the amplification of a 211bp DNA sequence specific for *T. congolense* using the following PCR conditions: 35 cycles of 94°C (1 min), 63°C (1 min) and 72°C (1 min), with a final extension of 10 min at 72°C.

To assess the species-specificity, TcoCATL-PCR was tested using DNA from *T. congolense* isolates of all subgroups (Table1) and *T. vivax* from Brazil (TviBrMi, TviBrCa), West (Y486), and East (TviMzNy) Africa; *T. b. brucei* (427, 8195, AnTat1.1), *T. b. gambiense*group 2(TB26), *T. b. rhodesiense* (AnTat 1.12), *T. equiperdum* (BoTat1.1), *T. evansi* from Brazil (Ted1, Tec2,Teh1) and Africa (TeET); and *T. theileri* from cattle (TthATCC, Tthc3, Tthc17) and water buffalo (Tthb4, Tthb6). *T. simiae* (Ban2, Ken14) and *T. godfreyi* (Ken7) were also tested. DNA from field collected blood samples preserved in filter paper or in ethanol was obtained as previously described (Rodrigues et al., 2010; Garcia et al., 2011a), and tested using the method of TcoCATL-PCR. PCR-amplified DNA fragments were separated in 2% agarose gels, and stained with ethidium bromide.

3. Results

3.1. Diversity of CATL-like genes in T. congolense IL3000 genome database

To investigate the repertoire of all potential CATL-like genes encoded by *T. congolense*, we performed a BLAST search for proteins with high sequence similarity to the archetypes of CP2 (GenBank accession number L25130) and CP1 (GenBank Z25813) in the *T. congolense* IL3000 genome database available from http://www.genedb.org. All sequences sharing high similarity with CP2 (14) or CP1 (8 sequences) and containing the catalytic domain were downloaded, aligned and employed for phylogenetic analysis (Table 1, Fig. 1). Our analysis of the catalytic triads disclosed: a) three sequences exhibiting the archetypical catalytic triad (CHN) of CP2; b) four sequences showing SYN or SHN triad that, unexpectedly, encoded congopain-like enzymes active against classical CP substrates but differing slightly from one another and also from CP2 in substrate preferences (Pillay et al., 2010); c) 7 sequences showing SSN and PHN triads, designated herein as "new congopain-like" because although sequences with a serine or tyrosine replacing histidine have been reported (Kakundi 2008; Pillay et al. 2010), the enzymatic activities of these variants have not investigated to date.

The amino acid sequences from the cds of CATL-like genes retrieved from *T. congolense* IL3000 genome database showed relevant polymorphisms (~18%) between the archetypical and variant CP sequences. The network genealogy of the 14 cd sequences found in the IL3000 genome confirmed and improved the divergent repertoire previously demonstrated by analysing CP sequences from a cosmid library of this strain, or PCR-amplified using degenerate primers (Kakundi, 2008; Pillay et al., 2010). Unfortunately, the sequences from CPs determined by Kakundi (2008), even those encoding congopain-like enzymes characterized by Pillay et al. (2010), are not available from public databases.

3.2. Repertoire and genealogy of the catalytic domains of CATL-like *sequences in T. congolense isolates of Savannah, Forest, and Kilifi subgroups*

We determined 100 sequences (477 bp) of the CATL-like catalytic domains from isolates of Savannah (WG81, Gam2, IL1180 and TREU1457), Forest (Cam22 and ANR3), and Kilifi (WG5 and WG84) (Fig. 1A). We also determined 44 sequences from blood samples of naturally infected cattle (5), water buffalo (1) and goats (1), plus one sample from a tsetse fly, all from Mozambique. Congopain analyses included several cloned sequences from each isolate, ranging from 10 sequences, for most Savannah laboratory isolates, to ≈30 for some isolates from Forest and Kilifi subgroups. Excluding the identical sequences, we generated alignments containing 70 different nucleotide sequences or 57 different predicted amino acid sequences (36 from Savannah, 9 from Kilifi, and 12 from Forest isolates), and the IL3000 genome sequences. The genealogy branching pattern not only segregated the sequences according to the *T. congolense* subgroups but also supported 4 subclades of sequences within the Savannah subgroup.

We selected sequences representative of the genetic diversity to illustrate the repertoire of CATL-like genes in *T. congolense* (Fig. 3A,B). Sequences from Forest and Kilifi subgroups always exhibited the typical catalytic triad and, respectively, QQLD or QQLN residues, preferentially, in the S2' subsite. In contrast, sequences from the Savannah isolates diverged highly in both the catalytic triad and S2´ subsite (Fig. 4 and Fig S1 available as Supplementary online documentation).

To more specifically evaluate the genetic repertoire of congopain and homologues in *T. congolense* Savannah, we compared the 36 new sequences (from 4 laboratory and 7 field isolates) determined in this study with 13 sequences from the IL3000 genome. The analysis revealed a high level of polymorphism among and within the isolates. In the network genealogy, CATL-like sequences were subdivided into 4 subclades (SAV1 to SAV4 subgroups). Most sequences were homologous to CP2 and assigned to SAV1 (21 sequences), SAV2 (10) or SAV3 (7). Sequences of SAV1 predominated and, despite preferentially showing typical triad (CHN), this group included the first variant (SHN) reported in *T. congolense* (Downey and Donelson, 1999) in the strain TRUM 183. The group SAV4 consisted of sequences from the IL3000 genome with the SSN variant triad previously described by Pillay et al., (2010) (Figs. 3A, B). Each group, SAV1-4, exhibited specific amino acid signatures characterized by two or three motifs, including the S2 and S2´ subsite regions (Fig. 4). Nine sequences (6 from IL3000 and 3 from field isolates) did not share these signatures and did not cluster into the groups SAV1-4; these sequences were provisionally denominated as "SAV not assigned" (SAVna). From field samples, a single sample of Savannah (So.go01) showed a SAVna sequence exhibiting a unique variant catalytic triad (CSN) (Figs. 3, 4; Fig. 1 supplementary material).

To quantify the overall divergence of the whole repertoire of CP amino acid sequences from the three subgroups of *T. congolense,* we calculated the mean divergences and the number of polymorphic sites (PS) in amino acid (95 PS) or nucleotide (222 PS) sequences. The amino acid sequences among the Savannah isolates were highly divergent (≈14% internal divergence and 87 PS) compared to Forest (≈4.0% internal divergence and 32 PS) or Kilifi (≈1.3% internal divergence and 11 PS) isolates. However, the mean divergences of amino acid sequences between the three subgroups were comparable (12-16%).

Only one copy of CP1 was obtained from three isolates of the Savannah subgroup (Ma.ca06, Ma.bu05, and Te.ca16), and no CP1 homologues were found among 82 sequences determined in this study for Kilifi and Forest isolates (Fig. 3B).

3.3. Molecular evolution of CP repertoires

To evaluate the role of positive selection in the molecular evolution of the CP repertoires of *T. congolense* Savannah, Forest and Kilifi, we analysed the ratio of non-synonymous to synonymous substitutions (dN-dS), considering each different nucleotide triplet; the analysis included 72 CP2 sequences and a total of 159 amino

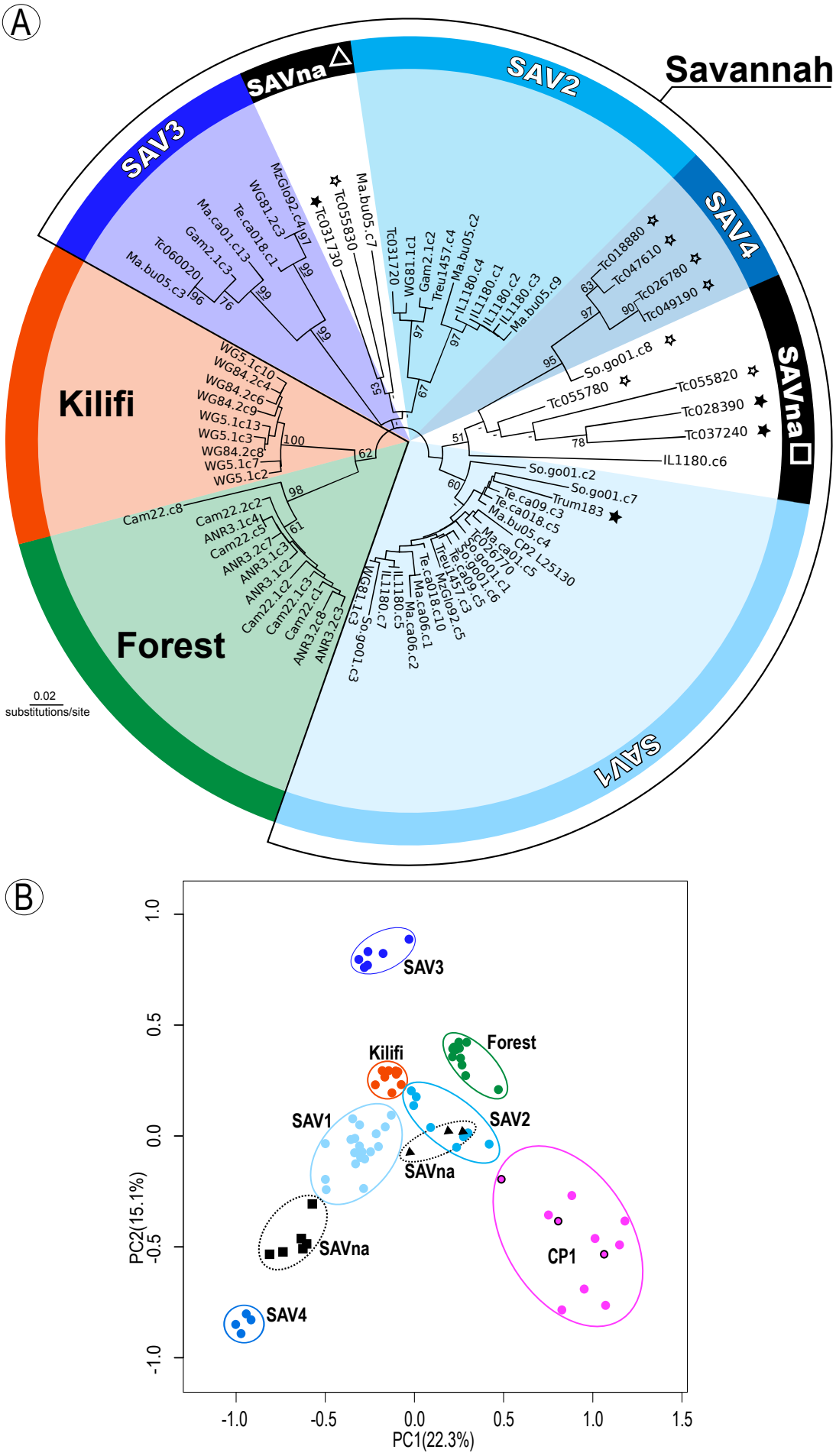


Fig. 3. (A) Genealogy of T. congolense catalytic domain of CATL-like inferred with 72 amino acid sequences from isolates representative of the Savannah, Forest, and Kilifi subgroups, including 36 sequences from livestock blood samples and tsetse flies determined in this study plus 13 sequences from the IL3000 genome. The four groups of sequences within the Savannah subgroup are indicated by different colours (SAV1-SAV4), and SAVna (Savannah21 not assigned) sequences are indicated by triangles and squares; sequences exhibiting variant catalytic triads corresponding to CP2-like (SAV4) and new CP2-like (SAVna) sequences are indicated by filled and unfilled stars, respectively. **(B)** The 2D sequence space of CP1 (pink circles with black stroke correspond to sequences determined in this study) and CP2 sequences were defined by the first two components (PC1 and PC2) of multidimensional scaling (MDS) plot constructed using the pairwise alignment of the 72 sequences, and the K-means method to define the groups of sequences.

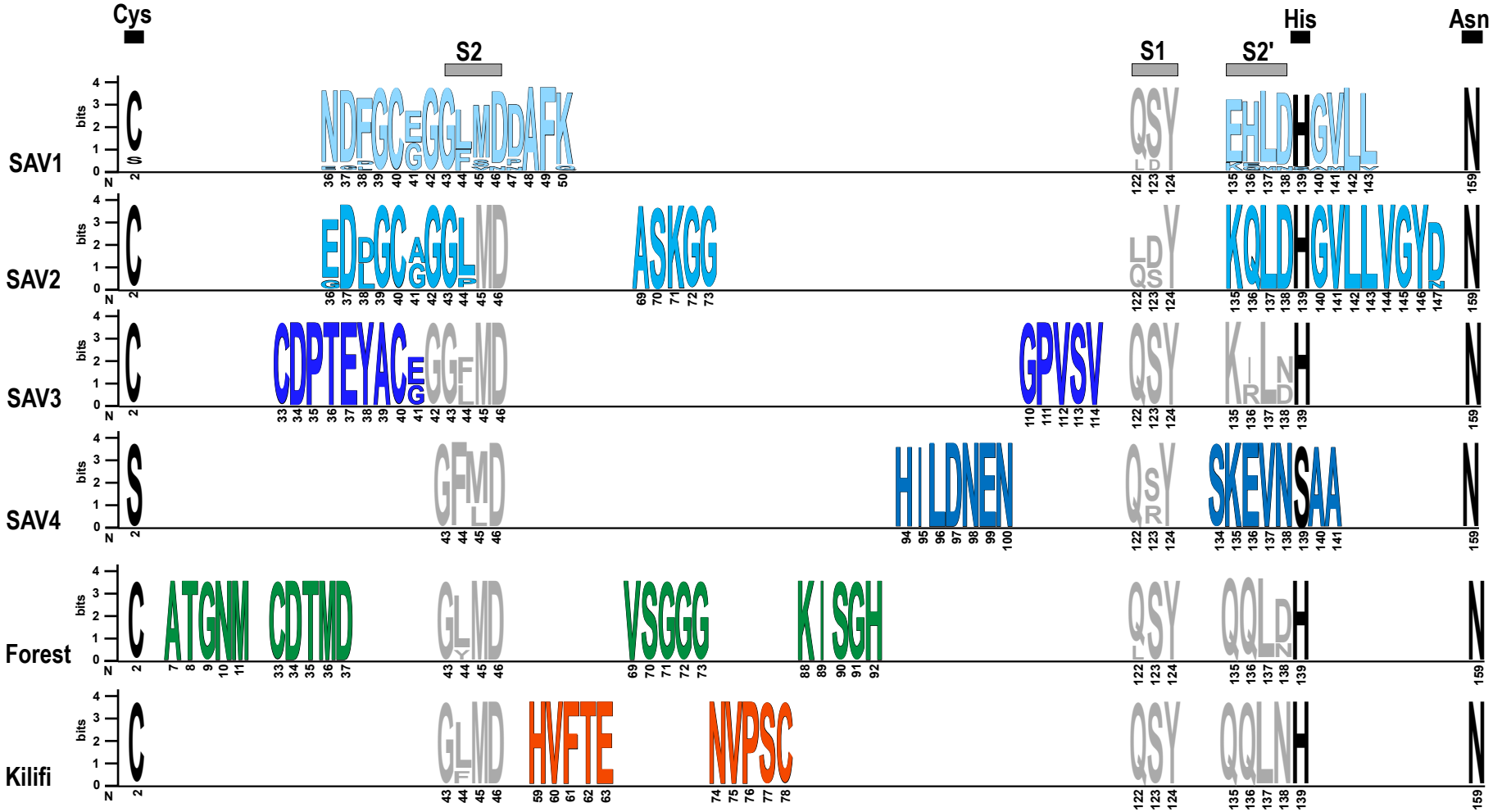


Fig. 4. Representation of signature residues within the catalytic domains of CATL-like of T. congolense defining he groups of sequences found within Savannah (SAV1-SAV4), Forest, and Kilifi subgroups. The amino acid motifs unique to each group of sequences were identified by analysing 155 amino acid sequences and used to design the logos. The CP2 catalytic triad, cysteine (Cys), histidine (His) and asparagine (Asn), and the subsites S2, S1, and S2' are indicated in black and grey, respectively. The numbers indicate the position in the CP catalytic domain amino acid sequence.

acids corresponding to 477 nucleotide positions in the final dataset. Analysis of the overall inferred substitutions was performed for all subgroups. The number of non-synonymous substitutions was always higher than synonymous substitutions, regardless of the subgroup analysed. A large proportion of the substitutions were within the subgroup Savannah, and no significant difference was observed when variant CP sequences were excluded from the analysis. A specific analysis of codons revealed a high degree of conservation for the S2 and S1 subsites, with more polymorphisms for the S2' subsite when sequences from each subgroup were compared separately. Despite non-synonymous substitutions in all subsites, non-significant P-values were estimated within Kilifi (0 to 0.45), Forest (0.45 to 0.70), and Savannah (0.08 to 0.70) subgroups. However, a comparison between the subgroups gave significant P-values for S1 and S2' subsites varying from 0.007 to 0.042. In the dN-dS test, P-values < 0.05 suggest rejection of the null hypothesis of neutral evolution. Therefore, these results suggest that CP genes are not subject to highly constrained evolution and diverged to constitute subgroup-specific subclades within *T. congolense*. In contrast to the highly polymorphic repertoires among and within the isolates of the Savannah subgroup, our findings suggest that congopain genes are moderately divergent in Forest, and more homogeneous in the Kilifi isolates.

The occurrence of recombination events in CP genes was evaluated by eight detection methods using the RDP3 program; recombination events were considered valid if detected by at least four methods, with a minimum significance P-value of 0.05. Six recombination methods indicated that one sequence (Ma.bu05.c6 from a buffalo isolate from Mozambique), which nested into CP1, was a product of recombination between CP1 and CP2-like sequence (P-value ranging from 2.5×10^{-9} to 3.0×10^{-2}). In addition, at least one sequence from each of the SAV1 and SAV2 groups also appears to be derived from recombination events (indicated by 4 methods) (supplementary Fig.2). Although preliminary, these findings suggest that recombination may be an important process in generating the diverse repertoire of CP sequences within the Savannah subgroup, providing additional insights to the mating capability of this subgroup (Morrison et al., 2009).

3.4. Relationships of T. congolense and other trypanosome species in phylogenetic trees based on CATL-like homologous genes

We inferred phylogenies based on catalytic domains of congopain genes from the three subgroups of *T. congolense* and homologues from *T. simiae, T. godfreyi,* and several other *Trypanosoma* species. Only sequences sharing the catalytic triad and active sites with the archetype congopain (CP2) were included in the analysis. The results revealed an heterogeneous clade comprising all and exclusively the *T. congolense* sequences, with a sister clade formed by sequences from *T. simiae* and *T. godfreyi,* which are more closely related than those from the isolates of distinct subgroups of *T. congolense.* Together, these three species formed a monophyletic assemblage that strongly validated the subgenus *Nannomonas.* The phylogenetic positioning of

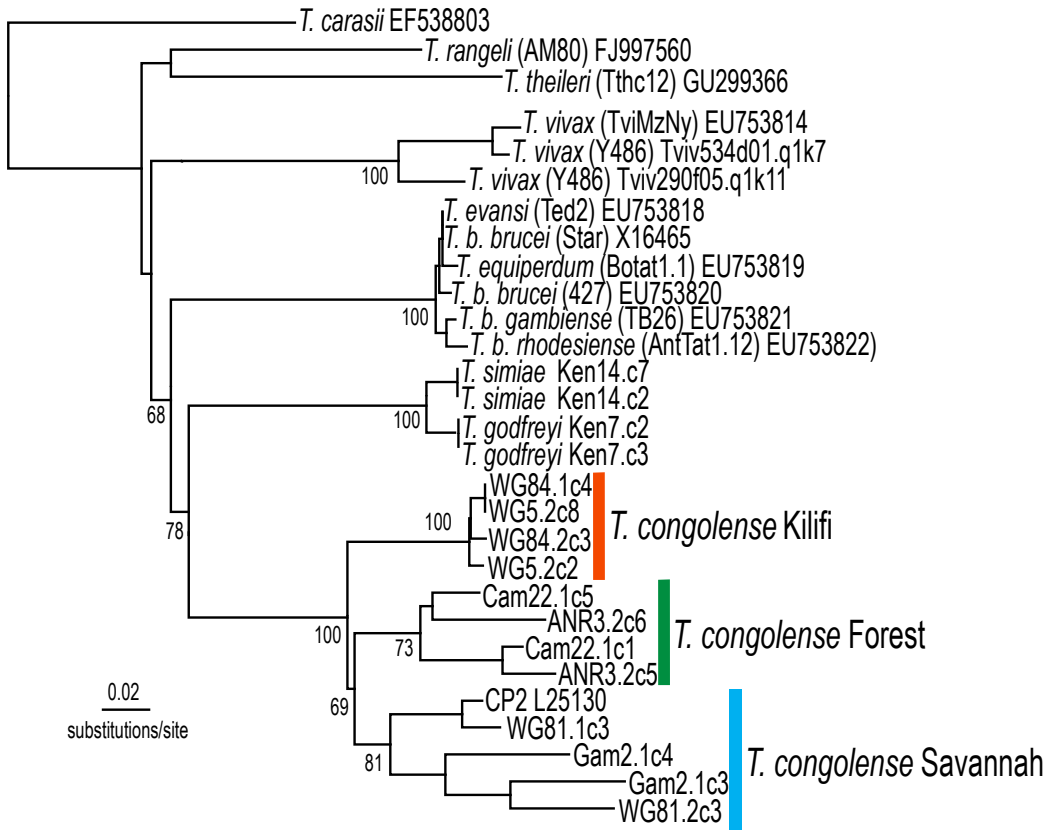


Fig. 5. Trypanosoma phylogenetic tree (neighbour joining tree) using congopain (CP2) amino acid sequences of T. congolense Savannah, Forest and Kilifi, and CATL-like homologous sequences from other trypanosome species. The numbers at the nodes are bootstrap support values from 500 replicates.

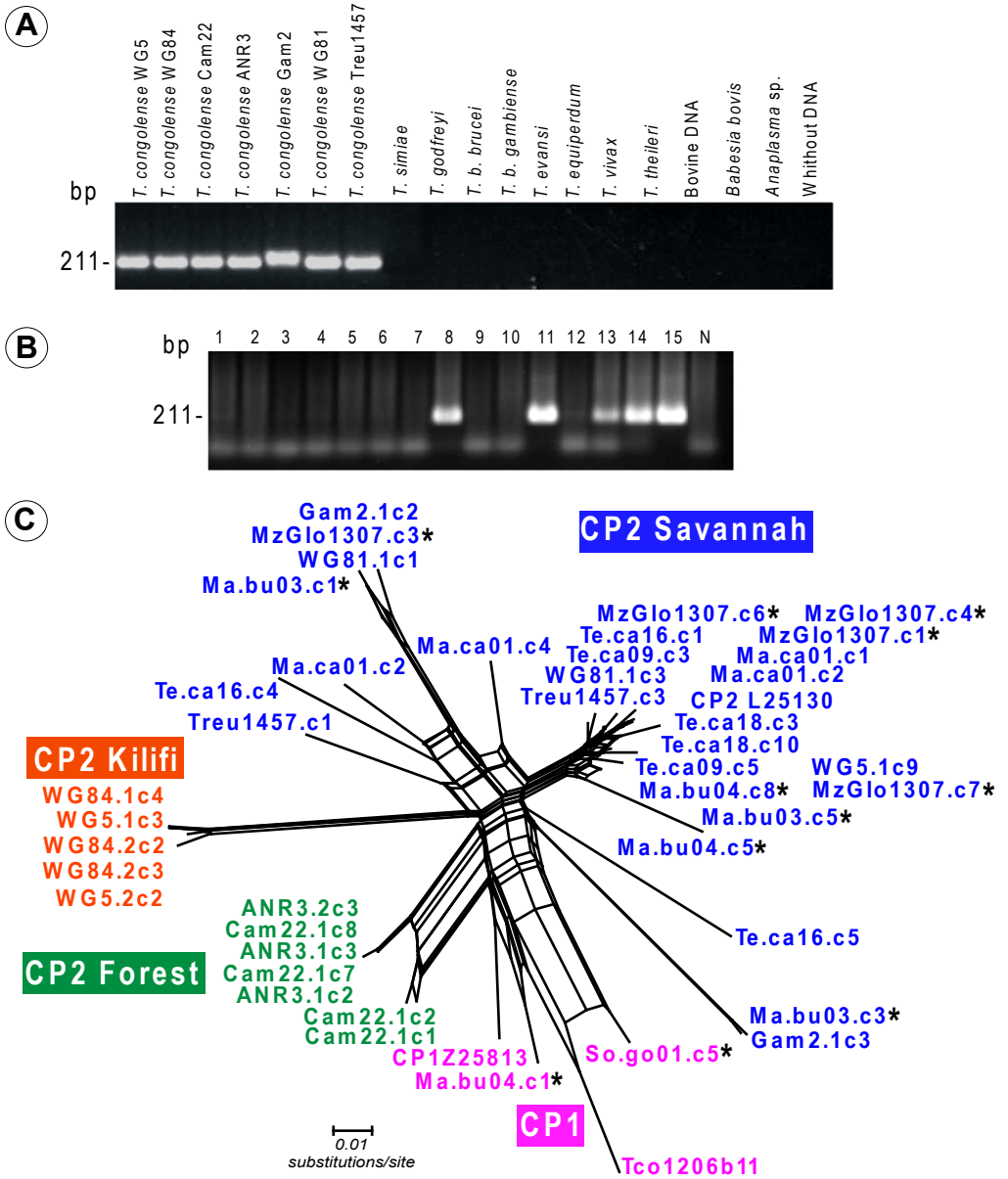


Fig. 6. Agarose gels stained with ethidium bromide (EtBr) showing DNA fragments amplified by TcoCATL-PCR: **(A)** specificity analysis using the DNA of isolates from the subgroups Savannah, Forest, and Kilifi, and other trypanosome species; **(B)** illustrative results from the evaluation of the suitability of the TcoCATL-PCR using crude DNA preparations from field-collected blood samples from sheep (1-5), buffalo (6-9), and cattle (11-14); positive (15) and negative (N) controls; **(C)** network genealogy of sequences amplified by TcoCATL-PCR (*) aligned with the corresponding region from T. congolense of all subgroups; analysis performed with the Neighbour-Net method with the K2P parameter and 1000 bootstrap replicates. **(D)** Selected region of aligned sequences illustrating the polymorphic sites useful for the genotyping of T. congolense subgroups.

this subgenus closer to the clade *T. brucei* than to *T. vivax* is in agreement with previous phylogenetic trees of *Trypanosoma* inferred with SSU rRNA and gGAPDH genes (Hamilton et al., 2004).

The congopain homologous genes from isolates of the same *T. congolense* subgroup always clustered together, forming three well-supported subclades corresponding to Savannah, Forest, and Kilifi. In the best resolved phylogenetic tree, obtained using archetypical congopain genes, Savannah was more related to Forest than to Kilifi (Fig. 5). However, in the network genealogy inferred with all sequences including those from variant CPs (Fig. 3A), Forest appears to be more related to Kilifi than to Savannah subgroup. Most likely, inconsistencies in the phylogenetic relationships among the *T. congolense* subgroups observed when applying different inference methods (data not shown) are due to their comparable genetic distances between the nucleotide/amino acid sequences: Savannah and Forest (14%/12%), Savannah and Kilifi (14%/12%), and Forest and Kilifi (12%/12%).

The comparison of cd sequences showed large divergences separating congopain genes and all other CATL-like homologous genes, even those from the closest phylogenetically related *T. simiae* and *T. godfreyi* (26%). Sequences from *T. congolense* also largely diverged from the highly homogeneous CATL-like genes from *T. brucei* ssp. and *T. evansi* of the subgenus *Trypanozoon* (30% sequence divergence), and from the moderately variable CATL-like sequences from *T. vivax* isolates (32% sequence divergence).

3.5. T. congolense diagnosis, genotyping, and preliminary epidemiological study targeting CATL-like *sequences*

To standardise the PCR assay targeting cdCATL-like sequences (TcoCATL-PCR), the primers Tco3F and Tco4R were designed to be complementary to sequences conserved in all congopain sequences from *T. congolense* of the three subgroups, and non-complementary to sequences from any other pathogenic African trypanosomes and *T. theileri* (non-pathogenic species of ruminants). This method amplified a *T. congolense*-specific DNA fragment of ≈211 bp for isolates of all subgroups (Fig. 6A). No amplified products were detected using DNA from *T. simiae*, *T. godfreyi, T. vivax* (South American and West and East African genotypes), *T. brucei* ssp., *T. evansi*, *T. equiperdum*, and *T. theileri.* Negative results were obtained using DNA templates from the hemoprotozoans *Babesia bovis*, *B. bigemina*, and *Anaplasma* sp. (Fig. 6A).

The suitability of TcoCATL-PCR for epidemiological studies was evaluated using blood samples preserved in ethanol or spotted on filter paper from cattle (97), goats (28) and water buffalo (6), and the gut contents of tsetse flies (16), all from Mozambique (Fig.6B). The TcoCATL-PCR was able to detect *T. congolense* in cattle blood sample that tested negative by microhaematocrit, and in tsetse samples that exhibited predominantly *T. simiae* and *T. godfreyi* mixed with very low amounts of *T. congolense* previously identified using the highly sensitive method of FFLB (Hamilton et al., 2008; Garcia et al., in preparation). Confirmation of PCR-amplified DNA bands was performed by sequencing randomly selected DNA fragments, and the results revealed exclusively isolates of the subgroup Savannah in livestock (Fig. 6C, D).

The sequences from the PCR-amplified DNA fragments (211 bp) were aligned with the corresponding sequences of congopain catalytic domains determined from the subgroups Savannah, Forest and Kilifi to evaluate their suitability for *T. congolense* genotyping. Small (211 bp) (Fig. 6C) and large (477 bp) (Fig. 3) congopain sequences resulted in similar groups corroborating the value of the TcoCATL-PCR–amplified sequences for *T. congolense* genotyping. Short sequences were sufficient for the identification of all subgroups by assessing the polymorphic sites (Fig. 6D), and the genealogy pattern (Fig. 6C).

The first epidemiological survey using TcoCATL-PCR followed by sequence analysis of the selected amplified DNA revealed that ≈27.5% of the livestock from Mozambique was infected with *T. congolense* Savannah (Fig. 6C). Our in progress study based on FFLB barcoding (Hamilton et al., 2008) has revealed high prevalence of both Savannah and Kilifi in tsetse flies from Mozambique (Garcia et al., in preparation).

Discussion

The present study examined the repertoire of CATL-like genes of isolates from the three subgroups of *T. congolense* (Savannah, Kilifi and Forest). This is the most comprehensive comparative study using protein coding genes from isolates of these three subgroups, and it was carried out with the aim of investigating the genetic repertoire of congopain-encoding genes (catalytic domains), and to assess the suitability of these sequences for diagnosis, genotyping, and phylogenetic inferences.

The analysis of CP sequences demonstrated significant variability among *T. congolense* Savannah, Forest and Kilifi subgroups, with extensive polymorphism within Savannah, moderate polymorphism within Forest, and relative homogeneity within Kilifi. From subgroup Savannah, we evaluated 9 laboratory isolates plus 7 field samples from cow, buffalo, goat and sheep, and tsetse; the isolates were collected in sites differing in ecological traits and separated by large geographical distances. The two Forest isolates were obtained from tsetse and goat from The Gambia and Cameroon, respectively, while the two Kilifi isolates were from the same farm at Matuga, Kenya. Therefore, the high diversity of Savannah may reflect sampling from wider geographic and host ranges, compared with limited sampling from the other groups. Nevertheless, in contrast to isolates from Forest and Kilifi, all Savannah isolates showed very polymorphic sequences, even when derived from animals living in sympatry.

In general, the CP sequences were conserved in regions involved in both substrate specificity and enzymatic activity regardless of subgroup affiliation. However, several sequences from isolates of the subgroup Savannah, mostly from the IL3000 genome, exhibited a polymorphic S2´subsite and unusual catalytic triads, corroborating previous reports of variant triads and the expression of congopain-like enzymes (Kakundi, 2008; Pillay et al., 2010).

The *T. congolense* IL3000 strain was selected for the genome project, and has been the subject of many studies regarding drugs and vaccines. Therefore, it is important to determine whether the remarkably diverse

genetic repertoires of both congopain and congopain-like enzymes found in this strain are common to other Savannah strains, and to strains of Forest and Kilifi subgroups. With this aim, using degenerate primers designed specifically for PCR amplification of variant CPs, Kakundi (2008) obtained sequences varying in the catalytic triad, in three other Savannah strains and in one Forest strain; a single sequence was reported from each strain. The primers we have employed in this study, regardless of the polymorphisms at the primer DTO154 annealing region, allowed for the amplification of highly polymorphic sequences from Savannah isolates (Sav1-4 and SAVna sequences), from isolates of the three subgroups of *T. congolense,* and also of CATL-like from all trypanosome species examined to date (Cortez et al., 2009; Garcia et al., 2011; Lima et al., 2012; Ortiz et al., 2009; Rodrigues et al., 2010). However, we cannot rule out the possibility that the primers employed for PCR amplification of CATL-like genes could have hampered the amplification of variant genes. Therefore, further studies are still necessary for a better appraisal of the congopain repertoire within *T. congolense*.

Phylogenetic analysis demonstrates that CP genes have diverged in specific subgroups, with a highly heterogeneous genetic repertoire among and within the isolates of the subgroup Savannah. The positive selection shaping the subgroup-specific and intra-Savannah genetic diversity suggests that CP2-encoding genes are not subject to highly constrained evolution among subgroups or within the subgroup Savannah. This process may have prevented extensive homogenisation, allowing for the emergence of subgroup-specific and highly divergent CP2 and CP2-like genes within Savannah. Differences in the ability to recombine may account for the higher diversity within Savannah. Microsatellite analyses suggested high variability, most likely resulting from mating in Savannah (Morrison et al., 2009), and low genetic variability and predominant clonal reproduction in Forest (Simo et al., 2013). Compared with data from *T. congolense*, there is a limited polymorphism in CATL-like genes from *T. brucei* ssp. and *T. evansi* and moderate diversity in *T. vivax* (Cortez et al., 2009). Interestingly, analysis of cathepsin B genes from *T. congolense* IL3000 revealed 13 gene copies with unusual polymorphisms in contrast to the single-copy gene from other trypanosome species (Mendoza-Palomares et al., 2008).

The results from this study provided new insights into the diversity of *T. congolense* CATL-like enzymes. Differences regarding the development in tsetse fly of different species and in experimental (livestock and mice) and field-infected animals have been observed among *T. congolense* Savannah, Forest and Kilifi isolates, and also within Savannah isolates (Bengaly et al., 2002a,b; Masumu et al., 2006; Seck et al., 2010; Van den Bossche et al., 2011; Vitoulay et al., 2011; Moti et al., 2012). Evidence from this study showing that highly virulent (Savannah), moderate (Forest) and non-virulent (Kilifi) isolates differ in their CP2 gene repertoires deserve to be better investigated regarding the association of virulence with distinct enzymes.

Findings from this study demonstrated for the first time that congopain genes are valuable markers for genotyping and phylogenetic inferences in *T. congolense* Savannah, Forest and Kilifi. Inferred phylogenetic trees based on CATL-like genes were similar to those based on SSU rRNA and gGAPDH genes (Hamilton et al., 2004)

clustering *T. congolense* together with *T. simiae* and *T. godfreyi* in the clade corresponding to subgenus *Nannomonas*. The assemblages comprising all and exclusively sequences from *T. congolense* were formed by three well-supported subclades corresponding to the three known subgroups: Savannah, Forest, and Kilifi. The best resolved phylogenetic analysis showed that Savannah and Forest isolates were more closely related and distant from isolates of the subgroup Kilifi, consistent with the results based on GARP (glutamate- and alanine-rich protein) gene sequences (Asbeck et al., 2000). Sequences of ribosomal RNA genes also supported a closer relationship between Savannah and Forest (Auty et al., 2012). We provide additional genetic evidence based on congopain genes corroborating that the three subgroups of *T. congolense* diverged enough to be separated into phylogenetically supported species. The genetic distances separating the subgroups of *T. congolense* are larger than the divergences between *T. simiae* and *T. godfreyi.* However, the characterization of more samples of Forest and Kilifi subgroups are required for a better appraisal of diversity, recombination, and taxonomic status of subgroups within *T. congolense* throughout sub-Saharan Africa (Gibson, 2007).

We developed a *T. congolense*-specific PCR assay targeting CATL-like sequences using crude DNA templates from field-collected blood samples preserved in ethanol at room temperature and, hence, amenable to epidemiological studies. The diagnostic PCR assay generated the same-sized fragments for the three subgroups of *T. congolense,* which can be genotyped by sequencing the small PCR-amplified fragment. The first method developed for this purpose, based on repetitive DNA, required three independent PCR reactions, one for each subgroup (Masiga et al., 1992), whereas a single PCR based on ITS1 rDNA can distinguish Savannah (700 bp), Forest (710 bp) and Kilifi (620 bp) (Desquesnes et al., 2001). However, epidemiological studies have required further sequencing of amplified ITS rDNA to distinguish between Savannah and Forest, which are often found as mixed infections (Malele et al., 2011; Auty et al., 2012).

Conclusions

We showed for the first time *T. congolense* Savannah, Forest and Kilifi specific repertoires of genes encoding congopain enzymes. The knowledge of genetic repertoires of CP enzymes and, specifically, of congopain, are valuable for studies about the roles of these enzymes in pathogenicity and virulence, and in the design of targets for the development of polyvalent vaccines and enzyme inhibitors useful as drugs against infections caused by isolates of the three subgroups of *T. congolense*. The method of PCR developed in this study can be helpful to improve the diagnosis and genotyping of *T. congolense* subgroups using crude DNA preparations from field-samples from livestock, wild reservoirs and tsetse flies.

Acknowledgements: We are indebted to the several students, colleagues, and local people who helped with the fieldwork in Mozambique. We are grateful to The Wellcome Trust for making available sequences from the genome of *T. congolense*. Sequence data were obtained from the Sanger Institute at the website of Genedb, http://www.genedb.org. This work was supported by the Brazilian agency CNPq within the PROAFRICA and UNIVERSAL programs. Adriana C. Rodrigues is a postdoctoral fellow of PNPD-CAPES, and Paola A. Ortiz and André G. Martins are recipients of scholarships from PROTAX-CNPq.

References

Adams, E.R., Hamilton, P.B., Gibson, W.C., 2010. African trypanosomes: celebrating diversity. Trends Parasitol. 26, 324–328.

Alvarez, V.E., Niemirowicz, G.T., Cazzulo, J.J., 2012. The peptidases of *Trypanosoma cruzi*: digestive enzymes, virulence factors, and mediators of autophagy and programmed cell death. Biochim. Biophys. Acta 1824, 195–206.

Asbeck, K., Ruepp, S., Roditi, I., Gibson, W., 2000. GARP is highly conserved among *Trypanosoma congolense* Savannah, Forest and Kilifi subgroups. Mol. Biochem. Parasitol. 106, 303–306.

Atkinson, H.J., Babbitt, P.C., Sajid, M., 2009. The global cysteine peptidase landscape in parasites. Trends Parasitol. 25, 573–581.

Authié, E., 1994. Trypanosomiasis and trypanotolerance in cattle: a role for congopain? Parasitol. Today 10, 360–364.

Authié, E., Boulangé, A., Muteti, D., Lalmanach, G., Gauthier, F., Musoke, A.J., 2001. Immunisation of cattle with cysteine proteinases of *Trypanosoma congolense*: targetting the disease rather than the parasite. Int. J. Parasitol. 31, 1429–1433.

Authié, E., Muteti, D.K., Mbawa, Z.R., Lonsdale-Eccles, J.D., Webster, P., Wells, C.W., 1992. Identification of a 33-kilodalton immunodominant antigen of *Trypanosoma congolense* as a cysteine protease. Mol. Biochem. Parasitol. 56, 103–116.

Auty, H., Anderson, N.E., Picozzi, K., Lembo, T., Mubanga, J., Hoare, R., Fyumagwa, R.D., Mable, B., Hamill, L., Cleaveland, S., Welburn, S.C., 2012. Trypanosome diversity in wildlife species from the serengeti and Luangwa Valley ecosystems. PLoS Negl Trop Dis 6, 1828.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S., 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37, 202–208.

Bengaly, Z., Sidibe, I., Boly, H., Sawadogo, L., Desquesnes, M., 2002a. Comparative pathogenicity of three genetically distinct *Trypanosoma congolense*-types in inbred Balb/c mice. Vet. Parasitol. 105, 111–118.

Bengaly, Z., Sidibe, I., Ganaba, R., Desquesnes, M., Boly, H., Sawadogo, L., 2002b. Comparative pathogenicity of three genetically distinct types of *Trypanosoma congolense* in cattle: clinical observations and haematological

changes. Vet. Parasitol. 108, 1–19.

Boulangé, A., Serveau, C., Brillard, M., Minet, C., Gauthier, F., Diallo, A., Lalmanach, G., Authié, E., 2001. Functional expression of the catalytic domains of two cysteine proteinases from *Trypanosoma congolense*. Int. J. Parasitol. 31, 1435–1440.

Cortez, A.P., Rodrigues, A.C., Garcia, H.A., Neves, L., Batista, J.S., Bengaly, Z., Paiva, F., Teixeira, M.M.G., 2009. Cathepsin L-like genes of *Trypanosoma vivax* from Africa and South America--characterization, relationships and diagnostic implications. Mol. Cell. Probes 23, 44–51.

Desquesnes, M., McLaughlin, G., Zoungrana, A., Dávila, A.M., 2001. Detection and identification of *Trypanosoma* of African livestock through a single PCR based on internal transcribed spacer 1 of rDNA. Int. J. Parasitol. 31, 610–614.

Downey, N., Donelson, J.E., 1999. Expression of foreign proteins in *Trypanosoma congolense.* Mol. Biochem. Parasitol. 104, 39–53.

Fish, W.R., Nkhungulu, Z.M., Muriuki, C.W., Ndegwa, D.M., Lonsdale-Eccles, J.D. and Steyaert, J. 1995. Primary structure and partial characterization of a life-cycle-regulated cysteine protease from *Trypanosoma (Nannomonas) congolense*. Gene. 161, 125–128.

Garcia, H.A., Kamyingkird, K., Rodrigues, A.C., Jittapalapong, S., Teixeira, M.M.G., Desquesnes, M. 2011a. High genetic diversity in field isolates of *Trypanosoma theileri* assessed by analysis of cathepsin L-like sequences disclosed multiple and new genotypes infecting cattle in Thailand. Vet. Parasitol. 180, 363–367.

Garcia, H.A., Rodrigues, A.C., Martinkovic, F., Minervino, A.H., Campaner, M., Nunes, V.L., Paiva, F., Hamilton, P.B., Teixeira, M.M.G. 2011b. Multilocus phylogeographical analysis of *Trypanosoma* (*Megatrypanum*) genotypes from sympatric cattle and water buffalo populations supports evolutionary host constraint and close phylogenetic relationships with genotypes found in other ruminants. Int. J. Parasitol. 41, 1385–96.

Garside, L.H., Gibson, W.C., 1995. Molecular characterization of trypanosome species and subgroups within subgenus *Nannomonas*. Parasitology 111, 301–312.

Gashumba, J.K., Baker, R.D., Godfrey, D.G., 1988. *Trypanosoma congolense*: the distribution of enzymic variants in east and west Africa. Parasitology 96, 475–486.

Gibson, W., 2002. Epidemiology and diagnosis of African trypanosomiasis using DNA probes. Trans. R. Soc. Trop. Med. Hyg. 96, 141–143.

Gibson, W., 2007. Resolution of the species problem in African trypanosomes. Int. J. Parasitol. 37, 829–838.

Gibson, W., 2012. *The origins of the trypanosome genome strains Trypanosoma brucei brucei TREU 927, T. b. gambiense* DAL 972, *T. vivax* Y486 and *T. congolense* IL3000. Parasit Vectors 5, 1.

Gibson, W.C., Dukes, P., Gashumba, J.K., 1988. Species-specific DNA probes for the identification of African trypanosomes in tsetse flies. Parasitology 97, 63-73.

Gibson, W.C., Stevens, J.R., Mwendia, C.M., Ngotho, J.N., Ndung'u, J.M., 2001. Unravelling the phylogenetic relationships of African trypanosomes of suids. Parasitology 122, 625–631.

Hamilton, P.B., Adams, E.R., Malele, I.I., Gibson, W.C., 2008. A novel, high-throughput technique for species identification reveals a new species of tsetse-transmitted trypanosome related to the *Trypanosoma brucei* subgenus *Trypanozoon*. Infect. Genet. Evol. 8, 26–33.

Hamilton, P.B., Stevens, J.R., Gaunt, M.W., Gidley, J., Gibson, W.C., 2004. Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA. Int. J. Parasitol. 34, 1393–404.

Huson, L.E., Authié, E., Boulangé, A.F., Goldring, J.P., Coetzer, T.H., 2009. Modulation of the immunogenicity of the *Trypanosoma congolense* cysteine protease, congopain, through complexation with alpha(2)-macroglobulin. Vet. Res. 40, 52.

Huson, D. H., Bryant, D. 2006. Application of Phylogenetic Networks in Evolutionary Studies, Mol. Biol. Evol. 23, 254-267.

Jaye, A.B., Nantulya, V.M., Majiwa, P.A., Urakawa, T., Masake, R.A., Wells, C.W., ole-MoiYoi_O.K., 1993. EMBL/GenBank/DDBJ databases.

Kateregga, J., Lubega, G.W., Lindblad, E.B., Authié, E., Coetzer, T.H., Boulangé, A.F., 2012. Effect of adjuvants on the humoral immune response to congopain in mice and cattle. Vet. Res. 8, 63.

Kakundi, E. M., 2008. Molecular Analysis of the Congopain Gene Family, School of Biochemistry, Genetics, Microbiology and Plant Pathology, University of Kwa Zulu-Natal, Pietermaritzburg (MSc Dissertation).

Knowles, G., Betschart, B., Kukla, B.A., Scott, J.R., Majiwa, P.A., 1988. Genetically discrete populations of *Trypanosoma congolense* from livestock on the Kenyan coast. Parasitology 96, 461–474.

Lalmanach, G., Boulangé, A., Serveau, C., Lecaille, F., Scharfstein, J., Gauthier, F., Authié, E., 2002. Congopain from *Trypanosoma congolense*: drug target and vaccine candidate. Biol. Chem. 383, 739–749.

Lima, A. P., Tessier, D.C., Thomas, D. Y., Scharfstein, J., Storer, A. C., Vernet, T. 1994 Identification of new cysteine protease gene isoforms in *Trypanosoma cruzi*. Mol. Biochem. Parasitol. 67, 333–8.

Lima, L., Ortiz, P.A., da Silva, F.M., Alves, J.M.P., Serrano, M.G., Cortez, A.P., Alfieri, S.C., Buck, G.A., Teixeira, M.M.G., 2012. Repertoire, genealogy and genomic organization of cruzipain and homologous genes in *Trypanosoma cruzi*, *T. cruzi*-like and other trypanosome species. PLoS ONE 7, e38385.

Majiwa, P.A., Hamers, R., Van Meirvenne, N., Matthyssens, G., 1986. Evidence for genetic diversity in *Trypanosoma (Nannomonas) congolense*. Parasitology 93, 291–304.

Majiwa, P.A., Maina, M., Waitumbi, J.N., Mihok, S., Zweygarth, E., 1993. *Trypanosoma (Nannomonas) congolense*: molecular characterization of a new genotype from Tsavo, Kenya. Parasitology 106, 151–162.

Majiwa, P.A., Masake, R.A., Nantulya, V.M., Hamers, R., Matthyssens, G., 1985. *Trypanosoma (Nannomonas) congolense*: identification of two karyotypic groups. EMBO J. 4, 3307–3313.

Malele, I.I., Magwisha, H.B., Nyingilili, H.S., Mamiro, K.A., Rukambile, E.J., Daffa, J.W., Lyaruu, E.A., Kapange, L.A., Kasilagila, G.K., Lwitiko, N.K., Msami, H.M., Kimbita, E.N., 2011. Multiple *Trypanosoma* infections are common amongst *Glossina* species in the new farming areas of Rufiji district, Tanzania. Parasit Vectors 4, 217.

Mamabolo, M.V., Ntantiso, L., Latif, A., Majiwa, P.A.O., 2009. Natural infection of cattle and tsetse flies in South Africa with two genotypic groups of *Trypanosoma congolense*. Parasitology 136, 425–431.

Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., Lefeuvre, P., 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26, 2462–2463.

Masiga, D.K., McNamara, J.J., Laveissière, C., Truc, P., Gibson, W.C., 1996. A high prevalence of mixed trypanosome infections in tsetse flies in Sinfra, Côte d'Ivoire, detected by DNA amplification. Parasitology 112, 75-80.

Masiga, D.K., Smyth, A.J., Hayes, P., Bromidge, T.J., Gibson, W.C., 1992. Sensitive detection of trypanosomes in tsetse flies by DNA amplification. Int. J. Parasitol. 22, 909–918.

Masumu, J., Marcotty, T., Geysen, D., Geerts, S., Vercruysse, J., Dorny, P., den Bossche, P.V., 2006. Comparison of the virulence of *Trypanosoma congolense* strains isolated from cattle in a trypanosomiasis endemic area of eastern Zambia. Int. J. Parasitol. 36, 497–501.

Mekata, H., Konnai, S., Simuunza, M., Chembensofu, M., Kano, R., Witola, W.H., Tembo, M.E., Chitambo, H., Inoue, N., Onuma, M., Ohashi, K., 2008. Prevalence and source of trypanosome infections in field-captured vector flies (*Glossina pallidipes*) in southeastern Zambia. J. Vet. Med. Sci. 70, 923–928.

Mendoza-Palomares, C., Biteau, N., Giroud, C., Coustou, V., Coetzer, T., Authié, E., Boulangé, A., Baltz, T., 2008. Molecular and biochemical characterization of a cathepsin B-like protease family unique to *Trypanosoma congolense*. Eukaryotic Cell 7, 684–697.

Morrison, L.J., Tweedie, A., Black, A., Pinchbeck, G.L., Christley, R.M., Schoenefeld, A., Hertz-Fowler, C., MacLeod, A., Turner, C.M.R., Tait, A., 2009. Discovery of mating in the major African livestock pathogen *Trypanosoma congolense*. PLoS ONE 4, e5564.

Moti, Y., Fikru, R., Van Den Abbeele, J., Büscher, P., Van den Bossche, P., Duchateau, L., Delespaux, V., 2012. Ghibe river basin in Ethiopia: present situation of trypanocidal drug resistance in *Trypanosoma congolense* using tests in mice and PCR-RFLP. Vet. Parasitol. 189, 197–203.

Njiru, Z.K., Makumi, J.N., Okoth, S., Ndungu, J.M., Gibson, W.C., 2004. Identification of trypanosomes in *Glossina pallidipes* and *G. longipennis* in Kenya. Infect. Genet. Evol. 4, 29–35.

Ortiz, P.A., Maia da Silva, F., Cortez, A.P., Lima, L., Campaner, M., Pral, E.M.F., Alfieri, S.C., Teixeira, M.M.G., 2009. Genes of cathepsin L-like proteases in *Trypanosoma rangeli* isolates: markers for diagnosis, genotyping and phylogenetic relationships. Acta Trop. 112, 249–259.

Pelé, J., Bécu, J.-M., Abdi, H., Chabbert, M., 2012. Bios2mds: an R package for comparing orthologous protein families by metric multidimensional scaling. BMC Bioinformatics 13, 133.

Pillay, D., Boulangé, A.F., Coetzer, T.H.T., 2010. Expression, purification and characterisation of two variant cysteine peptidases from *Trypanosoma congolense* with active site substitutions. Protein Expr. Purif. 74, 264–271.

Reifenberg, J.M., Cuisance, D., Frezil, J.L., Cuny, G., Duvallet, G., 1997. Comparison of the susceptibility of different *Glossina* species to simple and mixed infections with *Trypanosoma (Nannomonas) congolense* savannah and riverine forest types. Med. Vet. Entomol. 11, 246–252.

Rodrigues, A.C., Garcia, H.A., Ortiz, P.A., Cortez, A.P., Martinkovic, F., Paiva, F., Batista, J.S., Minervino, A.H., Campaner, M., Pral, E.M., Alfieri, S.C., Teixeira, M.M.G, 2010. Cysteine proteases of *Trypanosoma (Megatrypanum) theileri*: cathepsin L-like gene sequences as targets for phylogenetic analysis, genotyping diagnosis. Parasitol. Int. 59, 318–25.

Sajid, M., McKerrow, J.H., 2002. Cysteine proteases of parasitic organisms. Mol. Biochem. Parasitol. 2002. 120, 1–21.

Seck, M.T., Bouyer, J., Sall, B., Bengaly, Z., Vreysen, M.J.B., 2010. The prevalence of African animal trypanosomoses and tsetse presence in Western Senegal. Parasite 17, 257–265.

Simo, G., Silatsa, B., Flobert, N., Lutumba, P., Mansinsa, P., Madinga, J., Manzambi, E., De Deken, R., Asonganyi, T., 2012. Identification of different trypanosome species in the mid-guts of tsetse flies of the Malanga (Kimpese) sleeping sickness focus of the Democratic Republic of Congo. Parasit. Vectors 5, 201.

Simo, G., Sobgwi, P.F., Njitchouang, G.R., Njiokou, F., Kuiate, J.R., Cuny, G., Asonganyi, T., 2013. Identification and genetic characterization of *Trypanosoma congolense* in domestic animals of Fontem in the South-West region of Cameroon. Infect. Genet. Evol. 18, 66–73.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.

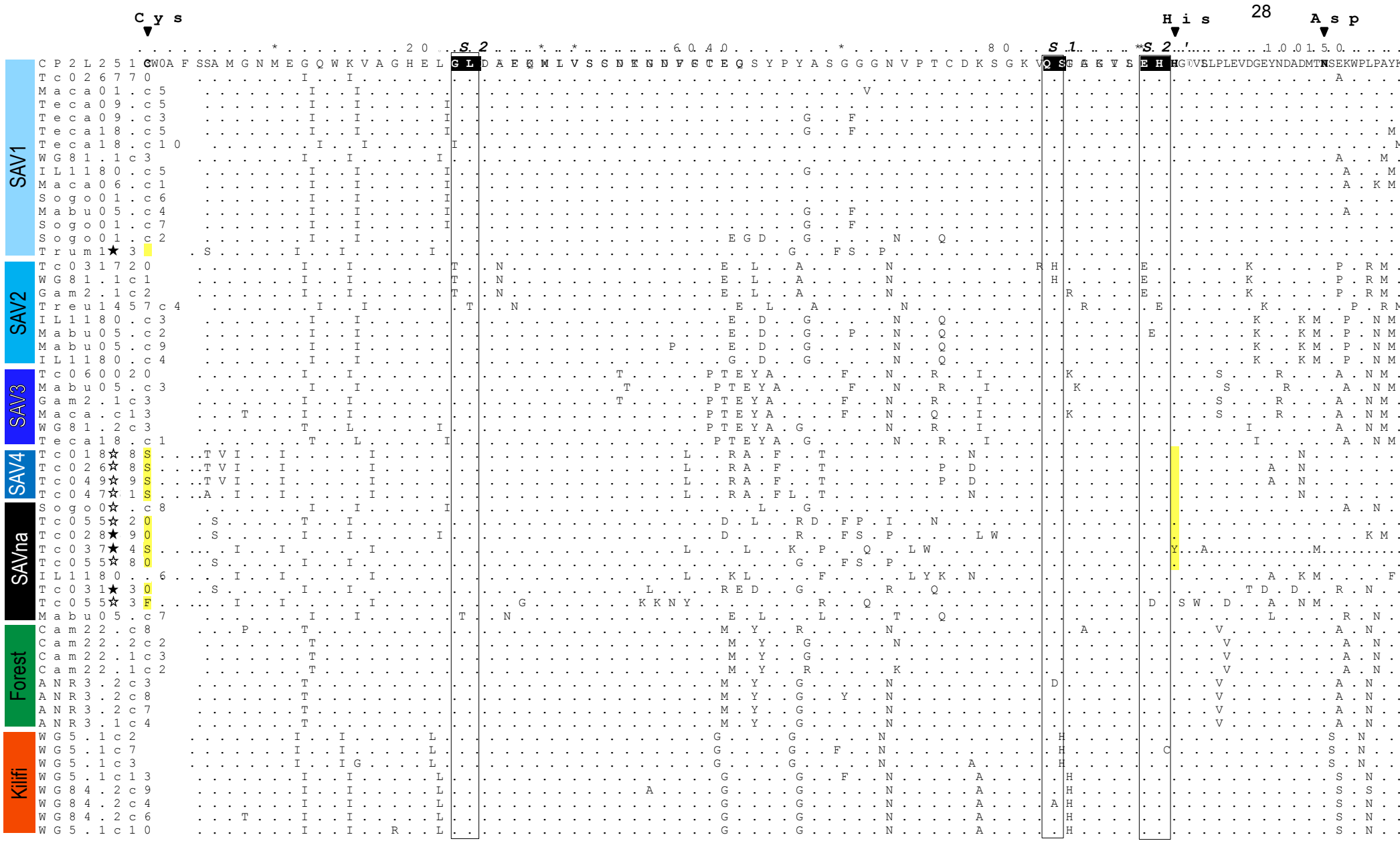
Van den Bossche, P., Chitanga, S., Masumu, J., Marcotty, T., Delespaux, V., 2011. Virulence in *Trypanosoma congolense* Savannah subgroup. A comparison between strains and transmission cycles. Parasite Immunol. 33, 456–460.

Vitouley, H.S., Mungube, E.O., Allegye-Cudjoe, E., Diall, O., Bocoum, Z., Diarra, B., Randolph, T.F., Bauer, B., Clausen, P.-H., Geysen, D., Sidibe, I., Bengaly, Z., Van den Bossche, P., Delespaux, V., 2011. Improved PCR-RFLP for the detection of diminazene resistance in *Trypanosoma congolense* under field conditions using filter papers for sample storage. PLoS Negl. Trop. Dis. 5, e1223.

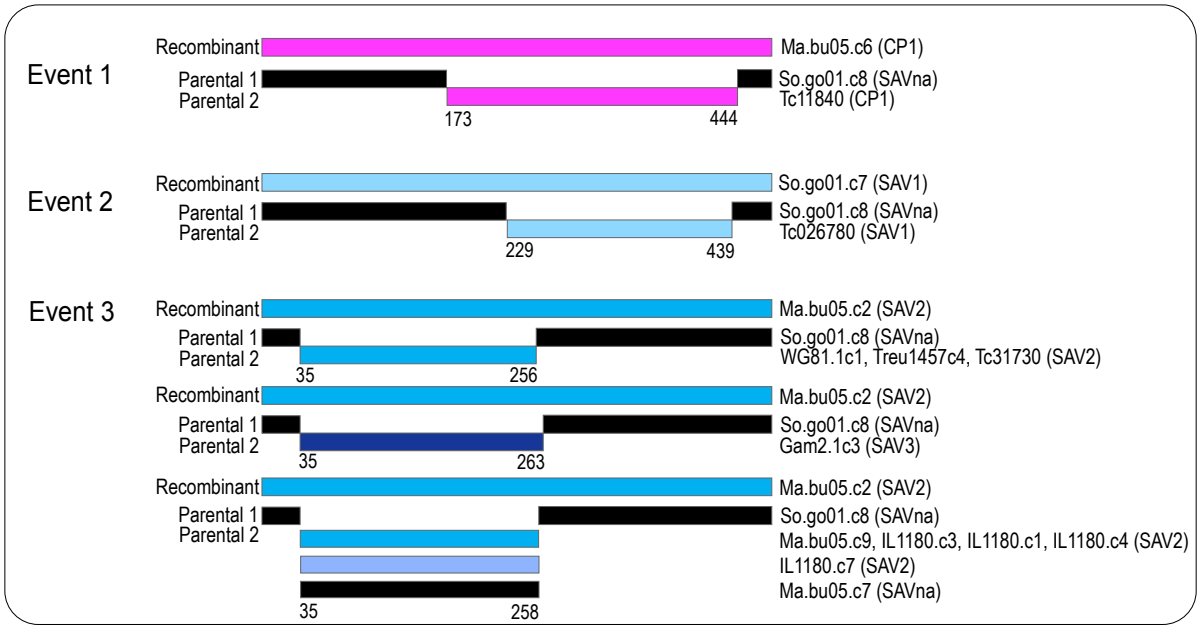
Yang, Z. 2007. PAML4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 8, 1586–1591.

Yang,Z., Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17, 32–43.

Young, C.J., Godfrey, D.G., 1983. Enzyme polymorphism and the distribution of *Trypanosoma congolense* isolates. Ann. Trop. Med. Parasitol. 77, 467–481.



Supplementary data 1. Alignment of deduced amino acid sequences selected to illustrate the polymorphism in the catalytic domains of T. congolense CP genes in the Savannah (SAV1, SAV2, SAV3, SAV4, and SAVna), Forest, and Kilifi subgroups. Genes retrieved from the genome draft of T. congolense IL3000 were aligned with the archetype CP2 (Genbank L25130) sequences (catalytic triad CHN) and with sequences determined in this study. Dots represent identical amino acids. The residues Cys, His and Asn of the active triad are indicated by arrowheads; the subsites S1, S2, and S2′ are depicted in rectangular blocks; amino acids shaded in yellow correspond to residue substitutions (Cys, His or Phe); the variant CP2-like sequences (filled stars) showed SYN and SHN, and the new CP2-like sequences (unfilled stars) SSN/PHN/CSN triads.



Supplementary data 2. Representation of the main putative recombination events in CP genes of T. congolense by analysing 155 nucleotide sequences using the RDP3 programme and testing 9 methods. Event 1 suggests a putative recombination between CP1, CP2 and CP2-like sequences and is supported by six methods, and events 2 and 3 suggest recombination between CP2 sequences and are supported by four methods.