BMC Evolutionary Biology

RESEARCH ARTICLE

Open Access

Mitochondrial DNAs provide insight into trypanosome phylogeny and molecular evolution

C. Kay[*], T. A. Williams and W. Gibson

Abstract

Background: Trypanosomes are single-celled eukaryotic parasites characterised by the unique biology of their mitochondrial DNA. African livestock trypanosomes impose a major burden on agriculture across sub-Saharan Africa, but are poorly understood compared to those that cause sleeping sickness and Chagas disease in humans. Here we explore the potential of the maxicircle, a component of trypanosome mitochondrial DNA to study the evolutionary history of trypanosomes.

Results: We used long-read sequencing to completely assemble maxicircle mitochondrial DNA from four previously uncharacterized African trypanosomes, and leveraged these assemblies to scaffold and assemble a further 103 trypanosome maxicircle gene coding regions from published short-read data. While synteny was largely conserved, there were repeated, independent losses of Complex I genes. Comparison of pre-edited and non-edited genes revealed the impact of RNA editing on nucleotide composition, with non-edited genes approaching the limits of GC loss. African tsetse-transmitted trypanosomes showed high levels of RNA editing compared to other trypanosomes. The gene coding regions of maxicircle mitochondrial DNAs were used to construct time-resolved phylogenetic trees, revealing deep divergence events among isolates of the pathogens *Trypanosoma brucei* and *T. congolense*.

Conclusions: Our data represents a new resource for experimental and evolutionary analyses of trypanosome phylogeny, molecular evolution and function. Molecular clock analyses yielded a timescale for trypanosome evolution congruent with major biogeographical events in Africa and revealed the recent emergence of *Trypanosoma brucei gambiense* and *T. equiperdum*, major human and animal pathogens.

Keywords: Trypanosome, Kinetoplast, Maxicircle, Mitochondrial DNA, Phylogeny, RNA editing

Background

Trypanosomes are a group of single-celled eukaryotic flagellates, including important pathogens of humans and their livestock (*Trypanosoma* and *Leishmania*), plants (*Phytomonas*) and insects (*Crithidia*). A distinctive feature of trypanosomes is the compartmentalization of the mitochondrial DNA into an organelle located at the proximal end of the flagellum, the kinetoplast, which contains a network of interlocked circular DNAs of two types: maxicircles which are equivalent to the mitochondrial genome of other eukaryotes, and minicircles that encode guide RNAs (gRNAs) used to edit the maxicircle transcripts [1, 2]. Thus both mini- and maxicircles are essential for expression of mitochondrial genes. In trypanosomes, mitochondrial transcripts are edited by the insertion or deletion of uridine residues at positions demarcated by gRNAs to yield mRNAs that can be correctly translated [3–5]. Why this energetically costly and potentially error prone mRNA processing step evolved, and how, are unanswered questions in trypanosome biology, but

[*]Correspondence: chris.kay@bristol.ac.uk
School of Biological Sciences, University of Bristol, Bristol, UK

RNA editing is found throughout the Kinetoplastea [1, 6].

Mitochondrial DNA is widely used in evolutionary, phylogenetic and population genetics analyses and has proved particularly useful as a molecular clock to date speciation events, but the extensive RNA editing of the trypanosome maxicircle might potentially undermine it's use. Within the Kinetoplastea, trypanosomes are monophyletic according to phylogenetic trees constructed from nuclear-encoded 18S ribosomal RNA (rRNA) and glycosomal GAPDH genes [7, 8], but it has proved difficult to date the emergence of particular lineages, as trypanosomes have no fossil record and are not sufficiently host specific to allow dating by co-speciation with their hosts. Nevertheless, the divergence date of two major groups of pathogenic trypanosomes in Africa (*T. brucei* clade) and South America (*T. cruzi* clade) has been linked to the breakup of Gondwana during the Cretaceous, ~100 Mya [7]. The *T. brucei* clade comprises the Salivaria, trypanosomes transmitted via the mouthparts of bloodsucking tsetse flies (*Glossina*) in sub-Saharan Africa, while the *T. cruzi* clade contains the agent of Chagas disease, *T. cruzi*, and related New World trypanosomes [9]. The 100 Mya date has been used to calibrate subsequent trees, e.g. Lewis et al. [10] estimated that *T. cruzi* lineages radiated 3.35 Mya and dated the emergence of two hybrid lineages of *T. cruzi* to < 60,000 years ago. However, the discovery of trypanosomes from wild animals in Africa that belong to the *T. cruzi* clade suggested the possibility of intercontinental transfer more recently via bats or rodents [11] so dating the emergence of trypanosome lineages remains uncertain. A means to infer origins independent of sparse historical information would give valuable insights into the emergence of different pathogens, as well as provide information on how quickly trypanosomes can switch hosts and vectors, with implications for the emergence of new diseases.

Here we have examined the potential of the trypanosome maxicircle for phylogenetic inference and dating. We used long-read sequencing to assemble complete maxicircles from four previously uncharacterized African trypanosomes, including the repetitive non-coding variable region. These assemblies were leveraged to scaffold and assemble a further 103 maxicircle gene coding regions, exploiting the wealth of published short read data. We show that time-resolved phylogenetic trees based on the maxicircle genecoding region can be used to explore events in the recent history of *Trypanosoma brucei* and *T. congolense*, and infer ages which fit well with historical evidence. Our analyses of the pre-edited and non-edited maxicircle genes indicate very high levels of RNA editing in salivarian trypanosomes, limiting further evolution in this direction without incurring functional costs.

Results

Whole maxicircle sequences

We sequenced (PacBio Sequel) mitochondrial DNA from four African trypanosomes (*T. congolense* savannah and kilifi, *T. simiae*, and *T. godfreyi*), and assembled complete maxicircles (including the variable region, Additional file 1: Table S2) using the long-read assemblers Canu and Flye [12, 13]. These novel data include the first complete maxicircles for *T. simiae*, *T. godfreyi* and the divergent *T. congolense* kilifi subgroup. We assembled two additional complete maxicircles for the genome strains *T. congolense* IL3000 and *T. vivax* Y486 from published data. We then assembled a further 101 maxicircle gene coding regions from public genome sequence data, using reference sequences or our new assemblies to recover maxicircle reads. In total, we obtained 51 complete maxicircle coding regions for *Trypanosoma brucei*, 34 for *T. congolense*, 3 for *T. equiperdum*, 2 for *T. godfreyi*, 1 for *T. grayi*, 2 for *T. simiae*, and 14 for *T. vivax* (Additional file 2: Table S1). No significant heteroplasmy was detected during sequence assembly.

Complete maxicircles ranged between 19.8 kbp (*T. vivax* Y486) and 27.6 kbp (*T. congolense* IL3000), with most of the size variation occurring in the variable region (4.6 kbp in *T. vivax* Y486 to 12.6 kbp in *T. congolense* IL3000; Additional file 1: Table S2). The overall GC content was 20.9–23.7%, but the GC% of the variable region was much lower (14.1% *T. godfreyi* KEN7 to 17.2% in *T. vivax* Y486). No significant correlation was found between gene coding and variable region GC% ($n = 6$, $\rho = -0.20$, $P = 0.70$), suggesting that changes to the composition of the variable/gene coding regions are independent. Dot plots of the variable regions typically showed two domains: one densely repetitive with short repeats and the other with longer period self-similarity (Additional file 3: Figure S2). Whilst the organisation of the variable region was similar between isolates of the same species (*T. vivax*, *T. congolense*), variation was seen in the fine structure and repeat copy number.

Complete maxicircles were identified by the assembly of circular sequences. For PacBio data, individual reads spanning the entire maxicircle (Additional file 4: Figure S1) were used to validate the assembled sequence.

Reference sequences for the complete variable region of *T. vivax* MT1 as well as the truncated (red line) variable region for *T. b. brucei* Lister 427, are shown to scale against other assembled salivarian variable regions.

Independent deletions of Complex I genes

Alignment of the gene coding regions showed overall conservation of synteny (Fig. 1); however, major gene deletions were evident in *T. godfreyi* (ND1) and *T. theileri* (ND4), as well as previously described deletions in New World *T. vivax* [14], *T. cruzi* [15], and *T. equiperdum* [16]. The deletion of ND1 in *T. godfreyi* was surprising, as this species undergoes full cyclical development in tsetse flies unlike New World *T. vivax* and *T. equiperdum*, which have both adapted to non-tsetse transmission and evidently do not require a fully functional mitochondrion. The deletion of ND4 in *T. theileri* has also eroded neighbouring genes, CR4 and ND3. Like *T. godfreyi*, *T. theileri* is predicted to require a functional mitochondrion as it completes development to mammal-infective forms in the gut of tabanid flies [17]. One possibility is that these deletions represent an early stage of mtDNA reduction in which mitochondrial function is reduced but not abolished.

Discounting sequences with segmental gene deletions, the size of the region containing both pre-edited and non edited genes (whole coding region, WCR) showed variation across *Trypanosoma* (Table 1) and trypanosome WCRs were approximately 1 kbp smaller than the reference sequences for related trypanosomatids *Crithidia* and *Leishmania*. Among the salivarian trypanosomes (lower portion of Table 1), gene coding regions without deletions (n = 10) are significantly smaller (Kruskal–Wallis one-way ANOVA on ranks, H = 8.0, $P = 0.05$) than those of non-salivarian trypanosomes (n = 4). These size differences can be traced to changes in the length of the pre-edited genes, for if they are summed (Table 1 'Σedited') and subtracted from the whole (Table 1 **'**WCR-Σedited'), the remaining region is relatively invariant in length (~ 12.4 kbp).

Gene coding regions frequently contained long homopolymers of either A or T, which appear to relate to reading direction of the gene (Fig. 1), indicating a strand specific bias. In contrast, the untranslated rRNA genes have low GC content but no directional bias in poly-A/T. Comparatively little expansion and contraction was observed in the intergenic regions, although in *T. vivax* a putative microsatellite was identified between the 9S and ND8 genes (Additional file 5: Figure S3).

High levels of RNA editing in salivarian trypanosomes

The overall GC% of the maxicircle gene coding region was low (~ 25%) in trypanosomes and other trypanosomatids, but these mean values conceal the fact that pre-edited genes are far more GC-rich than non-edited genes (Table 2). Comparison of these genes indicates that they are significantly more GC-rich in salivarian compared to non-salivarian trypanosomes (**ND8**, **ND9**, **COIII**, **A6** (n = 12,6); **CR3**, **ND7** (n = 11,6); Kruskal–Wallis one-way ANOVA on ranks all H > 14, P = < 0.001). These genes also show variation in T:C ratios, with particularly high T:C ratios in ND9, CR3 and CR4 (Table 2).

Some pre-edited genes showed a large variation of sequence length (Table 2), which was inversely proportional to GC% (ρ = $\mathbf{ND8}_{(n=20)}$ − 0.91, $\mathbf{ND9}_{(n=20)}$ − 0.93, $\mathbf{ND7}_{(n=19)}$ − 0.98, $\mathbf{COIII}_{(n=20)}$ − 0.96, $\mathbf{A6}_{(n=20)}$ − 0.95, $\mathbf{CR3}_{(n=19)}$ − 0.80, all $P = < 0.001$, Fig. 2, Additional file 6: Figure S4). Analysis of base composition reveals a strong proportional correlation of sequence length to T% (Additional file 6: Figure S4) whilst A% has a weaker inverse correlation (ρ = A%/T%, $\mathbf{ND8}_{(n=20)}$ − 0.72/0.98, $\mathbf{ND9}_{(n=20)}$ − 0.65/0.96, $\mathbf{ND7}_{(n=19)}$ − 0.72/0.99, $\mathbf{COIII}_{(n=20)}$ − 0.84/0.97, $\mathbf{A6}_{(n=20)}$ − 0.79/0.96, $\mathbf{CR3}_{(n=19)}$ − 0.75/0.97, all $P = < 0.005$). Providing that the translated product remains similar in size, shorter genes indicate a greater extent of editing in salivarian compared to non-salivarian

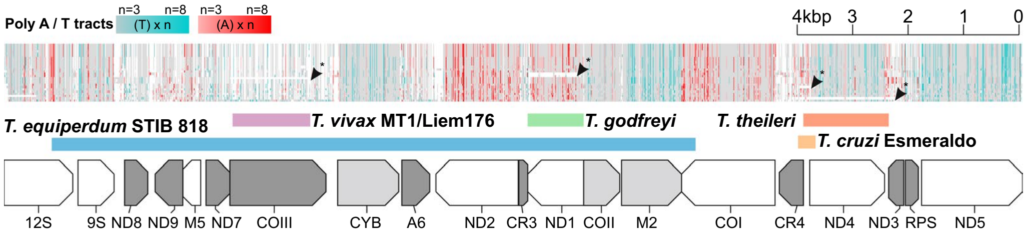


Fig. 1 Global overview of maxicircle gene coding region. Top: alignment of the maxicircle mitochondrial DNA gene coding regions from 27 isolates (top to bottom, *Trypanosoma brucei* H866, 1829 ALJO, Lister 427, TREU 927, TSW 55, J10, LF1; *T. congolense* IL3000, WG81, GAM2, IL3900, IL3578, ERA D1; *T. simiae* ERA C2; *T. godfreyi* KEN7, ERA F1; *T. vivax* Liem 176, Y486, Tv2323, *T. cruzi*, CL Brener, Esmeraldo; *T. lewisi*, *T. conorhini*, T. *copemani*, *T. grayi*, *T. theileri*, *Leishmania tarentolae*). Tracts of poly-T or poly-A are shown coloured turquoise or red respectively. An approximate scale is shown. Segmental gene deletions in the alignment are indicated by arrows and are also shown below as coloured bars; the deletion from *T. equiperdum* STIB 818 is also shown for comparison. Bottom: gene order in the maxicircle gene coding region. Non-edited genes in are shown in white, minimally edited genes in light grey and extensively edited (pan-edited) genes in grey. Editing categories are on the basis of *T. vivax* [14]

Table 1 Characteristics of pre-edited and non-edited mitochondrial maxicircle genes in trypanosomatids

Species		Edited genes (bp)											Non-edited genes					
													Reducible GC%			%AU codons		
	WCR (bp)	ND8	ND9	ND7	COIII	A6	CR3	CR4	ND3	RPS12	∑edited	WCR-∑edited	COI	ND4	ND5*	COI	ND4	ND5*
C. fasciculata Cf-CI	16118	316	293	339	1565	592	101	208	174	131	3618	12500	18		17	35		43
L. tarentolae	16169	276	313	339	1595	556	103	208	188	153	3628	12541	13	9	8	37	59	49
T. theileri Edinburgh TM35	13727	312	370	311	956	331	106			165	2445	11282	19			34		
T. conorhini 025E	15415	303	402	307	1001	337	122	235	219	163	2967	12448		15			51	
T. copemani G1	15215	293	365	290	958	318	114	241	174	157	2796	12419	22	18	22	32	46	41
T. lewisi	15274	295	360	289	989	304	121	235	183	167	2822	12452	16	15	13	36	49	48
T. cruzi CL Brener	15255	289	349	293	959	337	115	238	202	167	2834	12421	19	14	15	35	47	44
T. grayi ANR4	14919	296	369	323	983	340	119	256	203	170	2940	11979	22	18	19	32	48	40
T. vivax Y486	15182	282	295	278	886	321	120	227	224	170	2683	12499	14	9	7	38	58	54
T. godfreyi KEN7	14436	250	288	280	855	311		270	220	160	2634	11802	13	13	11	37	51	49
T. simiae ERA C2	14975	248	275	284	842	298	104	253	218	160	2578	12397	17	20		37	53	
T. simiae tsavo KETRI 3436	14587	247	271		850	297	100	328	213	236	2442	12145	12	9	9	38	55	53
T. brucei brucei Lister 427	14874	276	259	282	868	307	102	222	196	144	2554	12320	18	12	12	36	51	49
T. brucei gambiense 1829 ALJO	14882	276	259	281	868	307	102	224	197	144	2556	12326	19	12	12	35	52	48
T. brucei rhodesiense H866	14880	276	259	281	868	307	102	224	196	144	2555	12325	18	12	12	35	51	48
T. equiperdum BoTat	14883	278	259	285	868	309	103	225	197	144	2565	12318	18	13	12	36	52	48
T. congolense IL3578 (s)	14939	263	286	282	844	295	98	227	232	154	2583	12356	18	11	10	34	54	50
T. congolense IL3900 (f)	15166	262	293	283	1034	299	100	222	226	157	2776	12390	15	12	11	34	54	51
T. congolense ERA D1 (k)	15016	253	299	282	852	374	97	229	224	150	2663	12353	19	14	12	34	51	50
T. congolense IL3000 (s)	14941	263	286	283	846	294	98	227	231	154	2584	12357	17	11	11	34	53	50

Left: Variation in coding region size relates to the sequence contribution of pre-edited genes. Trypanosome gene coding regions (WCR) are shorter than for related trypanosomatids Leishmania and Crithidia. These size differences reflect the contraction of pre-edited regions (Σedited); the remaining gene coding region (WCR-Σedited) is relatively invariant in length. All numbers present the base length of ungapped sequences. Right: Non-edited genes show trends of GC loss which suggest vulnerability to loss of function. Universally non-edited genes COI, ND4 and the first 500 codons of ND5 (ND5*) were analysed for codon usage (% AU codons) and possible composition changes (reducible GC%), which is the percentage of alternate synonymous codons with reduced GC content. Coding regions which are [incomplete] or have [gene deletions] that impact calculations are highlighted. Numbers have been shaded by value order on a column by column basis. Trypanosomes below the solid line, with non-salivarian above and salivarian below the dashed line.

trypanosomes; presumably the increase in GC% and decline specifically in T% is offset by U insertion during RNA editing.

Non-edited genes are approaching limits to GC loss

Unlike pan-edited genes, non-edited and lightly edited genes are characterised by low GC% (Table 2). However the T:C ratios for some genes vary significantly between trypanosomes (Table 2), especially between salivarian and non-salivarian trypanosomes (**12S**$_{(n=12,5)}$, **9S**$_{(n=12,6)}$, **CYB**$_{(n=11,6)}$, **M2**$_{(n=12,6)}$, **COI**$_{(n=12,6)}$ **ND4**$_{(n=12,5)}$, **ND5**$_{(n=12,6)}$, one-way ANOVA all $F>70$, $P<0.001$; **COII**$_{(n=12,6)}$, $F=3.6$, $P=0.06$), whilst other genes do not show significant change (**M5**$_{(n=11,6)}$ $F=0.56$, $P=0.46$; **ND2**$_{(n=12,6)}$ $F=2.92$, $P=0.10$; **ND1**$_{(n=11,6)}$ $F=5.48$, $P=0.03$).

The low GC% of non-edited genes suggests that further reduction might lead to non-synonymous substitutions. Indeed, the total number of AU codons (no G or C) is already high, exceeding 50% in ND4 and ND5, and only a small proportion of remaining codons could be converted to AU codons without incurring translational changes (Table 1). Six amino acids (F, I, K, L, N, Y) solely use AU codons, and for amino acids encoded by GC or AU codons, the AU codon was strongly preferred. Thus further GC loss would either result in non-synonymous mutations or introduce a compositional bias in the gene product, suggesting that non-edited genes have reached the limit of GC loss, particularly in salivarians.

Time-resolved phylogeny of African trypanosomes

To test whether the discrete mechanisms driving sequence change in the maxicircle gene coding region would affect phylogenetic analysis between species of trypanosomes, alignments were prepared of (a) individual genes, (b) sets of pre-edited and non-edited genes, and (c) the whole gene coding region (WCR) with and without pre-edited genes (Additional file 7: Figure S5). Trees inferred from individual genes (e.g. COI) were congruent, providing no evidence of recombination between maxicircle loci, and strongly supported the monophyly of salivarians, although they showed weak resolution in the

Table 2 Individual maxicircle genes show different trends for GC composition and T:C ratios

Species	GC%																					T:C																				
	WCR	12S	9S	ND8	ND9	M5	ND7	COIII	CYB	A6	ND2	CR3	ND1	COII	M2	COI	CR4	ND4	ND3	RPS12	ND5	WCR	12S	9S	ND8	ND9	M5	ND7	COIII	CYB	A6	ND2	CR3	ND1	COII	M2	COI	CR4	ND4	ND3	RPS12	ND5
C. fasciculata Cf-Cl	26.0	16	19	40	41	15	34	29	25	22	21	38	31	28	21	31	42	23	47	46	25	3	6	5	2	2	7	3	6	6	10	6	5	4	4	8	4	2	5	2	1	6
L. tarentolae	22.9	16	18	50	42	13	32	25	24	21	16	38	26	26	15	29	44	18	45	39	20	4	7	6	2	5	7	3	8	6	11	7	5	4	5	14	4	5	9	2	3	9
T. theileri Edinburgh TM35	29.4	21	19	47	37	22	42	46	28	42	24	41	31	27	22	30				45	27	2	5	5	2	12	4	3	3	5	4	6	8	5	5	5	5				3	3
T. conorhini 025E	26.8	17	17	48	33	19	42	43	27	41	19	33	29	26	18	31	43	21	36	43	24	3	6	5	2	11	6	2	3	5	4	8	12	5	5	8	5	20	5	8	2	5
T. copemani G1	28.7	17	17	50	39	20	44	47	28	45	22	34	29	27	22	32	44	23	47	47	27	3	6	6	2	9	5	2	3	5	3	6	11	4	5	5	4	29	5	4	2	4
T. lewisi	26.4	18	18	49	39	14	44	44	24	45	19	36	27	28	18	29	43	21	42	43	22	3	5	5	2	9	11	2	3	7	3	7	11	4	5	7	5	27	5	5	3	6
T. cruzi CL Brener	26.1	18	17	49	34	17	43	44	26	37	19	31	26	28	18	30	34	22	36	42	23	3	6	5	2	8	5	2	3	6	4	5	11	4	5	8	4	27	5	6	3	7
T. grayi ANR4	28.0		16	49	38	18	40	46	25	41	21	33	30	25	20	31	44	22	44	41	25	3		6	2	8	4	3	3	5	5	7	12	4	5	6	5	35	4	6	3	5
T. vivax Y486	24.1	14	16	49	47	15	48	52	22	47	15	35	23	22	15	28	44	17	36	49	18	4	10	8	2	5	6	2	3	9	4	9	12	5	6	13	4	25	9	10	2	11
T. godfreyi KEN7	25.2	17	15	55	47	13	48	52	23	46	19			22	16	28	36	20	38	43	22	4	7	9	1	4	5	2	2	8	3	5			8	14	4	19	8	6	3	9
T. simiae ERA C2	25.3	15	15	55	47	13	46	52	23	49	18	41	26	22	16	28	41	18	40	45	20	4	9	8	1	4	5	2	2	8	3	5	8	4	8	17	4	11	10	6	3	9
T. simiae tsavo KETRI 3436	24.6	15	15	56	48			51		49	17	39	25	22	15	27	35	18	41	48	20	4	9	9	1	4			2		2	6	7	4	7	18	4	11	10	6	2	9
T. brucei brucei Lister 427	26.1	17	17	53	51	15	46	50	23	45	18	38	26	24	15	29	46	20	46	49	21	4	7	7	2	3	4	2	2	7	3	5	8	4	6	15	4	16	6	5	2	9
T. brucei gambiense 1829 ALJO	26.1	17	17	53	52	15	46	50	23	46	18	38	26	24	15	30	47	20	46	49	21	4	8	7	2	3	4	2	2	7	3	5	8	4	6	14	4	16	6	5	2	10
T. brucei rhodesiense H866	26.2	17	17	53	52	15	46	50	23	45	18	38	26	24	15	30	47	20	46	49	21	4	8	7	2	3	4	2	2	7	3	5	8	4	6	14	4	16	6	5	2	10
T. equiperdum BoTat	26.2	17	17	52	51	16	46	50	24	45	18	38	26	25	15	30	46	20	46	49	21	4	8	8	2	3	4	2	2	7	3	5	8	4	6	14	4	17	7	5	2	9
T. congolense IL3578 (s)	25.8	16	16	52	47	15	46	51	23	49	18	41	27	24	16	29	44	19	41	44	20	4	8	9	1	5	3	2	2	8	3	5	7	3	6	14	4	18	8	7	2	10
T. congolense IL3900 (f)	25.6	16	16	53	45	10	46	48	23	48	19	37	25	23	16	29	44	19	39	43	21	4	8	8	1	5	7	2	2	8	3	5	7	4	6	15	4	17	8	7	3	10
T. congolense ERA D1 (k)	26.6	17	15	55	42	14	47	52	25	45	20	40	28	24	17	30	41	20	42	46	21	4	8	9	1	6	5	2	2	7	3	4	6	3	7	14	3	15	7	5	2	11
T. congolense IL3000 (s)	25.8	17	16	52	47	15	46	52	23	49	18	41	26	24	16	29	44	19	40	44	20	4	8	9	1	5	3	2	2	8	3	5	7	3	6	13	4	18	8	7	2	10

From an alignment of gene coding regions, aligned sequence regions were extracted and analysed in the reading direction for GC% (left) and the ratio of T:C (right). Extensively edited genes ([dark grey]) have greater GC% than lightly ([light grey]) or non-edited (white) genes. Likewise, the T:C ratio is very high in some edited genes e.g. ND9, CR3, CR4. Shading as Table 1.

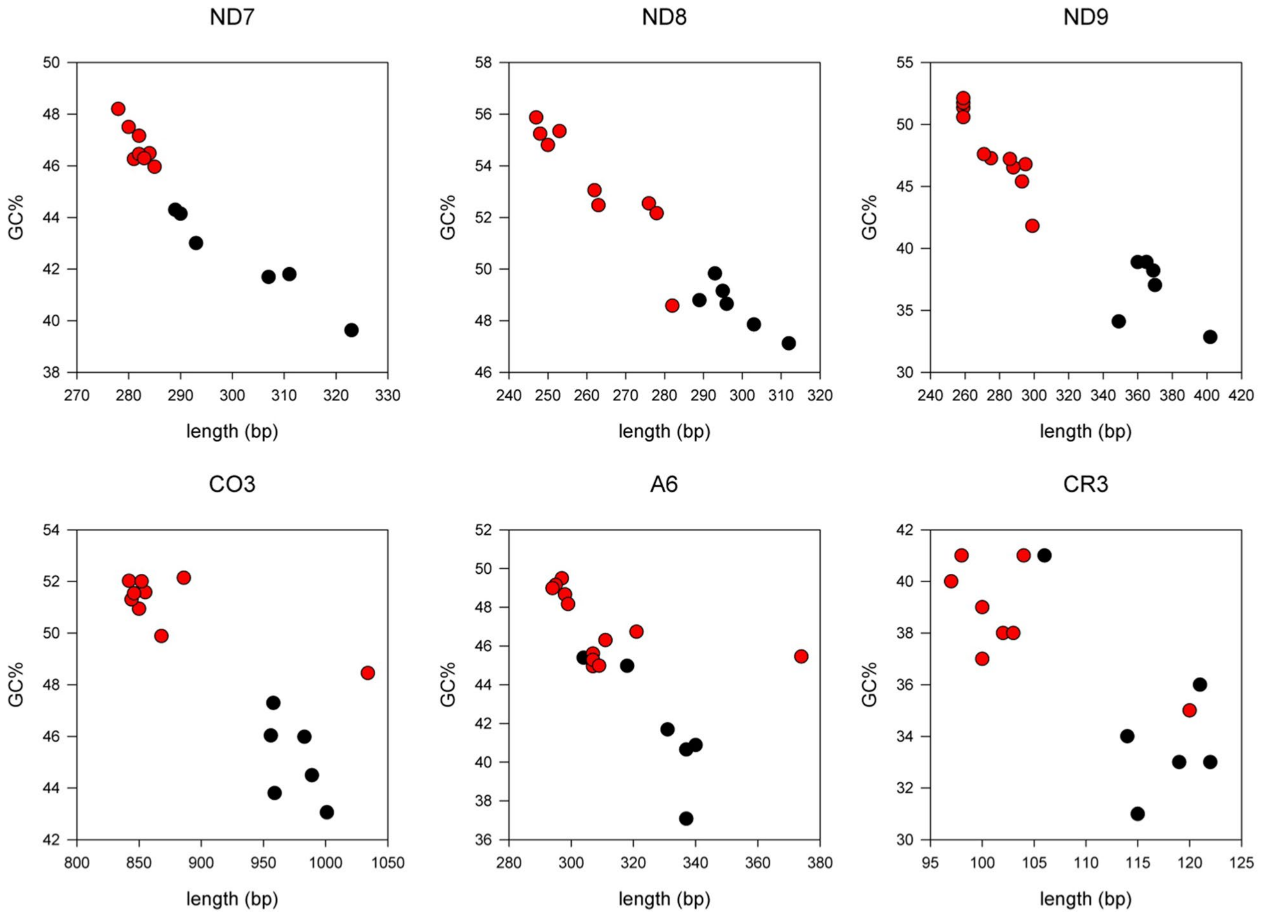


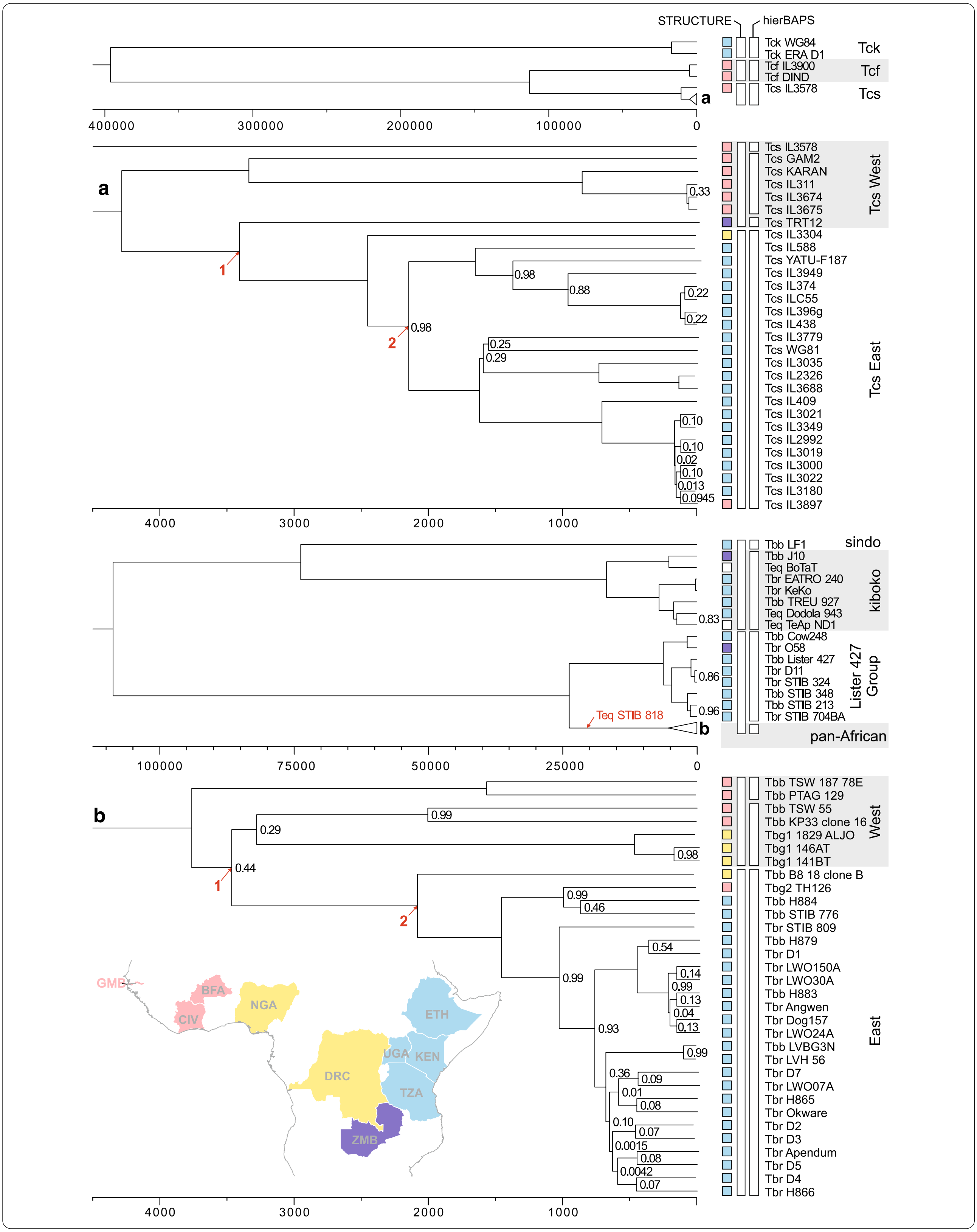
Fig. 2 Correlation between sequence length and GC% for six pre-edited maxicircle genes. There is an inverse correlation between the length and GC% of pre-edited genes in trypanosome maxicircles, where salivarian trypanosomes (red filled circle) generally have shorter, more GC-rich pre-edited genes than non-salivarian trypanosomes (black filled circle).This relationship also holds true for other trypanosomatid species (Additional file 6: Figure S4)

topology of deeper branches. Sets of non-edited genes had consistent topology for deeper branches, but were so conserved that the resolving power within species was limited. Topologies inferred from pre-edited regions alone, which as a whole are faster-evolving, resolved intraspecific groups confidently but presented conflicting topologies for deeper branches. Using the entire gene coding redion, including pre-edited genes, gave better resolving power (in terms of bootstrap support) than individual genes or sets of non-edited genes. Therefore, the use of the entire gene coding region including pre-edited genes and intergenic sequence appears to be a useful phylogenetic marker for trypanosome evolution [18] and were used in subsequent analyses.

Aligned WCRs of *T. congolense* (including savannah, forest and kilifi subgroups) and *T. brucei* (including *T. equiperdum* strains with complete coding regions) were used to construct time-resolved phylogenies (Fig. 3). To date species trees, the molecular clock was calibrated using tip isolation dates (Additional file 8: Data S1). Best marginal likelihoods were obtained with birth–death models using strict molecular clocks. Clock rates for *T.*

(See figure on next page.)

Fig. 3 Time resolved phylogenies of *T. congolense* and *T. brucei*. The savannah (**a**) and pan-African (**b**) clades are expanded below their respective trees. The coloured boxes correspond to countries of origin on the map of Africa (inset). STRUCTURE and hierBAPS groups are indicated by the white boxes. Timelines are in years before present and node values are posterior probabilities < 1. Arrowed nodes 1 and 2 are discussed in the text. The putative position of *T. equiperdum* STIB818 inferred from an independent ML tree (Additional file 10: Figure S6) is indicated in **b**


STRUCTURE
hierBAPS
Tck WG84
Tck ERA D1
Tcf IL3900
Tcf DIND
Tcs IL3578
Tck
Tcf
Tcs
a
400000
300000
200000
100000
0
a
Tcs IL3578
Tcs GAM2
Tcs KARAN
Tcs IL311
Tcs IL3674
Tcs IL3675
Tcs TRT12
Tcs West
0.33
1
2
0.98
0.88
0.98
0.22
0.22
0.25
0.29
0.10
0.10
0.02
0.10
0.013
0.0945
Tcs IL3304
Tcs IL588
Tcs YATU-F187
Tcs IL3949
Tcs IL374
Tcs ILC55
Tcs IL396g
Tcs IL438
Tcs IL3779
Tcs WG81
Tcs IL3035
Tcs IL2326
Tcs IL3688
Tcs IL409
Tcs IL3021
Tcs IL3349
Tcs IL2992
Tcs IL3019
Tcs IL3000
Tcs IL3022
Tcs IL3180
Tcs IL3897
Tcs East
4000
3000
2000
1000
Tbb LF1
Tbb J10
Teq BoTaT
Tbr EATRO 240
Tbr KeKo
Tbb TREU 927
Teq Dodola 943
Teq TeAp ND1
Tbb Cow248
Tbr O58
Tbb Lister 427
Tbr D11
Tbr STIB 324
Tbb STIB 348
Tbb STIB 213
Tbr STIB 704BA
sindo
kiboko
Lister 427 Group
pan-African
0.83
0.86
0.96
Teq STIB 818
b
100000
75000
50000
25000
0
b
0.99
0.29
0.98
0.44
1
2
0.99
0.46
0.54
0.14
0.99
0.13
0.04
0.13
0.99
0.93
0.99
0.36
0.09
0.01
0.08
0.10
0.07
0.0015
0.08
0.0042
0.07
Tbb TSW 187 78E
Tbb PTAG 129
Tbb TSW 55
Tbb KP33 clone 16
Tbg1 1829 ALJO
Tbg1 146AT
Tbg1 141BT
West
Tbb B8 18 clone B
Tbg2 TH126
Tbb H884
Tbb STIB 776
Tbr STIB 809
Tbb H879
Tbr D1
Tbr LWO150A
Tbr LWO30A
Tbb H883
Tbr Angwen
Tbr Dog157
Tbr LWO24A
Tbb LVBG3N
Tbr LVH 56
Tbr D7
Tbr LWO07A
Tbr H865
Tbr Okware
Tbr D2
Tbr D3
Tbr Apendum
Tbr D5
Tbr D4
Tbr H866
East
GMB
BFA
CIV
NGA
ETH
UGA
KEN
DRC
TZA
ZMB
4000
3000
2000
1000

brucei (median 1.81×10^{-7} substitutions/site/year, s/s/y, 95% HPD interval 1.00×10^{-6}–5.56×10^{-7} s/s/y) and *T. congolense* (median 7.45×10^{-7} s/s/y, 95% HPD interval 1.43×10^{-8}–2.89×10^{-6} s/s/y) were found to be similar in magnitude but significantly different from each other (Kruskal–Wallis one-way ANOVA on ranks, H = 680, $P < 0.001$). Clock rates calculated alone for the *T. brucei* pan-African clade and *T. congolense* savannah have similar (median *T. brucei* 2.35×10^{-7} s/s/y, *T. congolense* 1.15×10^{-6} s/s/y) but significantly faster (Kruskal–Wallis one-way ANOVA on ranks, H > 80, $P < 0.001$) rates compared to the species as a whole (Additional file 9: Data S2). The rates reported here are similar to those reported for other mitochondrial clocks [19], and faster than the estimated rate of trypanosome nuclear evolution based on 18S data (~ 1×10^{-10} s/s/y) [20]. Rates calculated from regions with non-edited protein coding genes (ND2 < > COI) have lower substitution rates (median, *T. brucei* 1.62×10^{-7} s/s/y, *T. congolense* 4.50×10^{-7} s/s/y) compared to the maxicircle gene coding region as a whole. However rates for this region are faster than rates predicted for the *T. cruzi* COII–ND1 region (~ 2×10^{-8} s/s/y) [21].

The inferred tree for *T. congolense* shows three clades, deeply separated in time, corresponding to the three known subgroups (Fig. 3). The kilifi subgroup (Tck) diverged approximately 400 kya (95% HPD interval 29–1200 kya), and the forest (Tcf) and savannah (Tcs) subgroups approximately 115 kya (95% HPD interval 20–680 kya). Most isolates fell in the Tcs clade (Fig. 3), which was further subdivided into two clades with a divergence date of ~ 4 kya; these clades broadly comprise East (Kenya, Uganda, Ethiopia and Zambia) and West (The Gambia and Burkina Faso) African isolates. Results from hierBAPS and STRUCTURE analyses confirmed these results, and in addition hierBAPS resolved TRT12 from Zambia and IL3578 from Burkina Faso as separate individuals (Fig. 3); for these analyses, runs with and without admixture models had the same best predicted K, except in the resolution of *T. congolense* subgroups.

The *T. brucei* tree revealed two deeply separated clades, which diverged ~ 108 kya (95% HPD interval 8–325 kya); these two clades also emerged from STRUCTURE analysis (Fig. 3). Separate from the main clade of *T. brucei* subspecies isolates is a clade containing isolates previously identified as belonging to kiboko (J10, 927) and sindo (LF1) groups on the basis of maxicircle restriction fragment length polymorphisms [22] and COI haplotypes [23]; interestingly this clade also contains Old and New World *T. equiperdum* isolates (BoTat, Dodola 943, TeAPND1). The main *T. brucei* sspp. clade is further subdivided, separating a group of East African isolates containing Lister 427 from a pan-African group ~ 23 kya. A group of largely East African isolates emerge from the pan-African clade ~ 3.5 kya, and this group is also present in STRUCTURE and hierBAPS analyses (Fig. 3).

The inferred trees for *T. brucei* give insights into the evolution of *T. equiperdum* and *T. b. gambiense*. For *T. equiperdum* only three isolates with whole coding regions were included in time-resolved analysis, as sequences of other isolates are either incomplete or have large deletions [STIB841 and STIB842 are truncated in ND5; STIB818 has lost most of the maxicircle (Fig. 1)]. The positions of these other isolates were inferred from maximum likelihood trees on shared sequence (partial 12S, partial COI to partial ND5, ~ 4.7 kbp of sequence), which indicate that STIB841 and STIB842 group with the other *T. equiperdum* isolates with full length sequences, while STIB818 is placed separately (Additional file 10: Figure S6). The divergence date for this main group of *T. equiperdum* from *T. b. brucei* t is around 5000 ya (*T. b. brucei* J10 and *T. equiperdum* BoTat 5190 ya (95% HPD interval 360–15,800 ya); *T. b. brucei* 927 and *T. equiperdum* Dodola 943/TeAPND1 4310 ya (95% HPD interval 232–9810 ya)). However, the position of STIB818 suggested by maximum likelihood trees could support a much older origin for this lineage.

Origins of both Type 1 and 2 T. *b. gambiense* (*Tbg1* and *Tbg2*) can be estimated from the inferred trees: *Tbg1* 3,240 ya (95% HPD interval 222–8380 ya); *Tbg2* ~ 1000 ya (95% HPD interval 80–3020 ya). This result is consistent with a published emergence date of 750–9500 years ago for *Tbg1*, based on estimated mutation rates and the observed number of mutations accumulated per genome in this asexual lineage [24].

Despite the difference in clock rates for each species, the time resolved phylogenies indicate that both *T. congolense* and *T. brucei* underwent a major divergence events simultaneously (Fig. 3: 1, West to Central Africa, *T. congolense* 3410 ya, 95% HPD interval 294–13,400 ya; *T. brucei* 3430 ya, 95% HPD interval 236–9610 ya; 2, expansion into East Africa *T. congolense* 2160 ya, 95% HPD interval 160–6800 ya; *T. bruce*i 2070 ya, 95% HPD interval 141–6210 ya), posing intriguing questions about the causes. Possible reasons include the major climatic changes that have affected the African continent in the past few thousand years, including the gradual desiccation of the Sahara desert (~ 3.0 kya) and the closure of the Dahomey gap (~ 4.5 kya) [25], overlaid by movements of wild animals, humans and their livestock in response to ecological changes.

Discussion

The trypanosome maxicircle presents a complex evolutionary system, with several discrete mechanisms bringing about sequence change. Synteny in the gene coding region is largely conserved, except for segmental gene deletions in some lineages (*T. equiperdum*, New World *T. vivax*, *T. godfreyi*, *T. theileri*), leading to presumed loss of function. Whether complete maxicircle loss, as seen in *T. evansi*, is the inevitable fate for maxicircles with small deletions remains unclear, but the fact that several maxicircles with deletions have been found suggests that maxicircle loss may not happen as a single event.

Assembled maxicircles have low GC content in both gene coding and variable regions. For non-edited genes the remaining permissive mutational space for further GC loss is small, particularly in salivarians, as few mutations would be synonymous and protein composition might already be compromised. The true extent of GC loss in pre-edited genes is cryptic as additional coding information comes from the minicircle gRNAs, however the declines in pre-edited gene length indicate that genes are more extensively edited in salivarian compared to non-salivarian trypanosomes. Given the recent emergence of the salivarian clade this would conflict with the idea that RNA editing is a primitive kinetoplastid feature that is always "on the way out" [1, 6].

The observed base composition biases in the maxicircle could be driven by the loss of recombination, as GC loss is a feature commonly observed in non-recombining populations [26]. Alternatively base composition biases could reflect the metabolic cost and availability of these nucleobases [27]. In non-edited genes a strand-specific bias for poly-T as well as selection for AU codons suggests that selection acting at the level of the transcript, such as for translational efficiency or against transcript cost, influences the evolution of these sequences. The rRNA genes have low GC content but as they are not translated are not expected to share the same codon selection pressures. The variable region has the lowest GC content, but wide variations in the size even within the same species, suggesting that it is not being streamlined for a reduced cost.

The Salivaria appear as a distinct group in the analyses presented, sharing properties of increased T:C ratio in their non-edited genes and shorter edited genes compared to non-salivarian trypanosomes. This disjunction suggests that the salivarians have undergone a period of evolutionary change, perhaps associated with their adaptation to transmission via the salivary route in tsetse. Unfortunately there are no intermediate taxa to sample. Although *T. grayi* is also transmitted by tsetse, this is by the posterior rather than salivarian route, and *T. grayi* is not a close relative of salivarian trypanosomes in phylogenetic analyses [8, 9, 28].

Despite the different evolutionary processes at work in the maxicircle gene coding region, our analyses demonstrate that it is a useful tool for phylogenetic analysis and a good molecular clock within a species. From population genetics analyses and the consistent phylogeny of isolates using different portions of the maxicircle, recombination appears rare or is restricted to very closely related sequences in the salivarian trypanosomes *T. congolense* and *T. brucei*. This contrasts with *T. cruzi,* where evidence for recombination and heteroplasmy has been presented [29, 30]. Our analysis of *T. brucei* and *T. congolense* suggests that the maxicircle can be used to probe the recent history and distribution of a species using isolation dates without other assumptions. The dates inferred for the *T. brucei* group fit well with estimations for the date of emergence of the human pathogen *T. b. gambiense* Type 1, previously calculated as 750–9500 ya, based on estimated mutation rates and the observed number of mutations accumulated per genome in this asexual lineage [24]. These dates fit with the development of settled agriculture and burgeoning centres of population in West Africa in the past 10,000 years that favoured the evolution of parasites adapted to human to human transmission. As shown by previous studies [31], *T. equiperdum* is polyphyletic. A new finding here was the emergence of one clade of *T. equiperdum* from the divergent group of *T. b. brucei* associated with wild animal-tsetse transmission cycles in East Africa, referred to as kiboko/sindo group [22]. This puts a new perspective on the evolution of *T. equiperdum* from *T. b. brucei,* with an estimated emergence date of *T. equiperdum* of ~ 5000 ya. The kiboko/sindo clade itself is estimated to have diverged from the main *T. brucei* clade > 108,000 ya.

Besides the kiboko/sindo clade, a small group of East African *T. b. brucei* and *T. b. rhodesiense* isolates was clearly separate from the majority of *T. brucei* isolates from sub-Saharan Africa. The human pathogen *T. b. rhodesiense* is characterised by a unique gene, the *SRA* (serum resistance associated) gene, which confers the trait of human infectivity [32]. Two major sequence variants of this gene have been identified that distinguish *T. b. rhodesiense* isolates from northern and southern East Africa. Here, *T. b. rhodesiense* LVH 56 (northern *SRA* variant) and *T. b. rhodesiense* 058 (southern *SRA* variant) were found in separate clades in the tree (Fig. 3), with an estimated divergence time of ~ 23,000 ya, placing the emergence of the *SRA* gene, and consequently *T. b. rhodesiense,* earlier than this date.

The dated phylogeny also has ramifications for the evolution of the important livestock pathogen, *T. congolense.* Of the three subgroups of *T. congolense,* kilifi is the earliest diverging, estimated to have split from the forest and savannah subgroups ~ 400,000 ya. *T. congolense* savannah and forest subgroups diverged more recently around 115,000 ya. The more extensive sampling of the savannah subgroup provides evidence of a split between East and West African isolates about 4000 ya. The position of TRT12 from Zambia on a long branch at the edge of the East African clade suggests that further subdivisions may emerge with more sampling of the savannah subgroup throughout its geographical range, as already suggested in other studies [33].

From the difference in calculated clock rates for *T. brucei* and *T. congolense,* it is clear that clock rates vary between trypanosome species, which fits with the observation that the rate of nuclear evolution in salivarians is seven to tenfold higher than non-salivarians [20]. It is also reasonable to assume that there is rate variation across the coding region. At present, the geological timescale of salivarian divergence is poorly constrained, with most published studies based on a single calibration of divergence between New World and Old World trypanosomes at 100 Mya, coincident with the splitting of Africa and South America. However, it is difficult to exclude the possibility that trypanosome exchange between continents might have occurred much more recently [11], which would have a major impact on inferred rates. The alternative of using isolation dates may provide a useful complementary approach for investigating more recent divergences within trypanosome clades. We show here that isolation dates can be used to explore events in the recent history of a species, and infer ages which fit well with historical evidence (*T. equiperdum*, *T. b. gambiense*). Rate calculations for *T. brucei* and *T. congolense* from different sets of isolation dates are in strong agreement for geographically shared events in recent history, and could be tested further by future analysis of *T. vivax*. Future analyses of deeper trypanosome evolution must address assumptions on how rates are calculated, how rate varies between species, and our confidence in using geological events for speciation barriers. This would put us in a better position to understand the evolution of the salivarian trypanosomes and the genus as a whole, and infer accurate dates for the origins of the group.

Conclusions

The maxicircle data we present represents a new resource for experimental and evolutionary analyses of trypanosome phylogeny, molecular evolution and function. Despite the different evolutionary processes at work in the maxicircle coding region, our analyses demonstrate that it is a useful tool for phylogenetic analysis and a good molecular clock within a species. Molecular clock analyses yielded a timescale for trypanosome evolution congruent with major biogeographical events in Africa and revealed the recent emergence of *Trypanosoma brucei gambiense* and *T. equiperdum*, major human and animal pathogens.

Methods

Genomic DNA extraction

High molecular weight DNA for genome sequencing was purified from axenically-grown procyclic trypanosomes using a Blood and cell culture kit (Qiagen) and a modification of the manufacturer's yeast cell protocol. Briefly, approximately 5×10^8 trypanosomes were pelleted by centrifugation, washed once with PBS and resuspended in 5 ml lysis buffer containing proteinase and RNAase as per the manufacturer's protocol. Following 1 h incubation at 50 °C, lysates were centrifuged at 5000 rpm for 5 min at room temperature in a microfuge to pellet debris before the supernatant was applied to a Genomic-tip 100/G column (Qiagen). Subsequent processing followed the manufacturer's protocol; after isopropanol precipitation, DNA was resuspended in 200 µl 10 mM Tris, 0.1 mM EDTA, pH 8 and stored at 4 °C.

Sequence data

Long read data was obtained on a PacBio Sequel II System, using 1 or 2 cells of a 4 reaction SMRT Cell 1M v2 plate per sample, and prepared using the SMRTbell® Template Prep Kit 1.0. Short read sequence data was obtained on an Illumina NovaSeq producing approximately 20 Gbp of 150 bp paired end reads per sample. Reference sequences were obtained from NCBI Refseq database [34]. Data in the Sequence Read Archive [35] was recovered using the SRA Toolkit [36]. All the data sources used for assembly are listed in Additional file 2: Table S1.

Assembly

Illumina assembly

For species with reference maxicircles, Illumina sequence data was searched using Magic-BLAST v1.4.0 [37] for aligning reads. Pooled reads were then assembled using SPAdes v3.13.1 [38] and maxicircle contigs identified by BLAST v2.2.31+ [39]. Where assembly yielded multiple maxicircle contigs, those of > 1000 bp were oriented and scaffolded using MeDuSa v1.6 [40]. For species without close references (e.g. *T. grayi*) NOVOplasty v3.3 [41] was used to extend the COI seed region of a related species to yield partial maxicircle sequences.

PacBio assembly

Long PacBio reads spanning the maxicircle were identified by BLAST; an example maxicircle spanning read is shown in Additional file 4: Figure S1. These reads were then used to fish additional sequences from the read pool. Reads were then corrected using Canu v1.8 [13] and split to less than 12 kbp before being assembled with Flye v2.5 [12]. Illumina read data, where available, were used to polish Flye assembled maxicircles.

Sanger assembly

Maxicircle reads were identified by BLAST against a reference maxicircle and assembled using CAP3 [42].

Assembly assessment

Reads were aligned to the assembled maxicircle sequences using BWA MEM v0.7.17 [43] and visualised in Tablet v1.19.09.03 [44]. Dot plots were produced in Flexidot v1.06 [45].

T. theileri

The *T. theileri* maxicircle sequence was identified from the assembled contig pool by BLAST.

Additional information on assembly if given in Additional file 11: Methods S1.

Gene annotation

For partially assembled maxicircles BLAST was used to recover individual non-edited genes. Complete coding regions were prepared using BLAST to crop assembled maxicircles between 12S rRNA and ND5 genes. Sequences were aligned using MAFFT v7.427 [46] (coding sequence; G-INS-i, PAM 200, k = 2, individual genes); short sequences were discarded. An approximation of gene boundaries for edited genes was made by aligning an annotated coding region of *T. vivax* Y486 to the coding region alignment and cropping sub-alignments on the basis of these annotations. For non-edited protein coding genes, gene boundaries could be determined by predicted open reading frame using translation Table 4 (https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi).

Phylogenetics

Maximum likelihood trees were inferred using IQ-Tree v1.6.12 [47], using ModelFinder [48] to find the best-fitting nucleotide substitution models. Parameters from these runs were used to inform a time-resolved phylogeny using BEAST2 v2.6.1 [49]. A birth–death model using isolation dates as tip dates and a strict molecular clock was used for both *T. congolense* and *T. brucei*, on the basis of marginal likelihood estimation using the BEAST2 path sampling and ModelTest packages. Each run was sampled every 100 iterations over a chain length of 10,000,000 with the first 10% discarded as burn-in; analyses were examined in Tracer v1.7.1 [50]. Treeannotator v1.10.4 was used to extract a consensus tree from the sampled population, and trees were visualised in FigTree v1.4.2.

Population genetics

Clustering of isolates into groups was performed by first extracting variable positions from aligned coding regions using SNP-sites [51]. Appropriately formatted files were then prepared using PGDSpider v2.1.1.5 [52] for later use in hierBAPS [53] and STRUCTURE [54]. Job runs in STRUCTURE used 10 iterations of admixture and no admixture models between K 1–8, with 5000 generations of burn-in and 5000 sampled, and assumed the maxicircle as a haploid allele. STRUCTURE HARVESTER v0.6.94 [54, 55] was then used to determine K. Runs in hierBAPS used 4 levels and 20 initial clusters and were run until convergence.

Statistics

Comparison of sequence properties between species used a representative from each species and species subgroup. For comparing clock rates, 900 evenly sampled clock rates after a 10% burn-in on a MCMC chain length of 10,000,000 were used for the basis of analysis. Where gene or sequence properties are compared, tests for normality (Shapiro–Wilk) and equal variance were first applied to determine an appropriate test of variance (normal, one-way ANOVA; non-normal, Kruskal–Wallis one-way ANOVA on ranks). All correlations used the Pearson correlation coefficient (ρ =).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12862-020-01701-9.

Additional file 1: Table S2. Properties of the six complete salivarian maxicircles.

Additional file 2: Table S1. Strain information and source data.

Additional file 3: Figure S2. Assembled variable regions of salivarian maxicircle mitochondrial DNAs.

Additional file 4: Figure S1. An example of a PacBio read spanning the entire sequence of the trypanosome mitochondrial DNA (maxicircle). A single read from the *T. congolense* GAM2 readpool is shown dot-plotted against itself. The highly repetitive short period portion of the variable region is visualised as a densely self-similar region between 20-26 kbp, whilst the longer period portion of the variable region begins at 15 kbp. The remainder of the sequence shown belongs to the gene coding region. The complete length of the maxicircle is seen from 0-26.3 kbp, and thereafter begins to repeat. The assembled sequence is shown in Additional file 3: Figure S2.

Additional file 5: Figure S3. Trypanosoma vivax, an alignment of the intergenic region between 9S and ND8 containing a putative microsatellite. Bases are shown as coloured bands with the top line tick showing 20 bp increments. The sequence [ATATA] is tandemly repeated between 18 and 51 times in the selected isolates.

Additional file 6: Figure S4. Correlation between sequence length, GC% and T% for six pre-edited maxicircle genes. Some pre-edited maxicircle genes exhibit transcript length variation with strong correlation between length and T% as well as an inverse correlation for GC%. The weak negative correlation for A% indicates that this is a strand specific phenomenon consistent with RNA editing, where uridines are inserted back into the transcript. Key: *Crithidia, Leishmania* (open circle), salivarian (red filled circle) and non-salivarian trypanosomes (black filled circle).

Additional file 7: Figure S5. A comparison of maximum likelihood trees inferred from different regions of the trypanosome maxicircle mitochondrial DNA. In general using more sequence contributes to higher bootstrap support for the inferred maximum likelihood topology. If individual genes are used, confidence for the deepest branches is reduced, and topological variances are observed. Collections of non-edited genes have a consistent topology but fail to resolve well within species. Use of the entire gene coding region (WCR), with or without pre-edited genes, provides better supported trees. If pre-edited genes alone are used, structure within species is well supported, but multispecies relationships are poorly resolved.

Additional file 8: Data S1. Distribution of isolation dates used for inferring time-resolved phylogeny. A spread of isolation dates for strains of *T. brucei*, *T. congolense*, *T. equiperdum* and *T. vivax* are shown. Complete gene coding regions used for time resolved phylogeny are indicated in red. Multiple complete coding regions were obtained for *T. vivax* but clocks were not calculated based on the limited range of isolation dates.

Additional file 9: Data S2. Distribution of clock rates sampled from BEAST2 for trypanosome species and subgroups. Nine hundred evenly sampled clock rates from timed phylogeny runs are shown for *T. brucei* (Tb) and the pan-African subgroup, as well for *T. congolense* (Tc) and the savannah subgroup (Tcs). Box and whisker plots show the 10th, 25th, 75th and 90th percentiles with the midline representing the median.

Additional file 10: Figure S6. Inferred polyphyly of *T. equiperdum*.T. equiperdum isolates in red font. A maximum likelihood tree inferred from the shared common sequence from the reference sequences of STIB818, STIB841 and STIB842, which have incomplete coding region sequences, and BoTat, Dodola 943 and TeAp ND1, which all have complete maxicircle coding regions. Node values represent bootstrap support.

Additional file 11. Additional methods.

Additional file 12. Assembled maxicircle FASTA formatted sequence data.

Abbreviations

Ribosomal genes

12S: 12S rRNA; 9S: 9S rRNA; RPS12: Ribosomal protein 12.

Respiratory complex genes

ND(1–9): NADH dehydrogenase subunits (1–9); CYB: Apo-cytochrome *b*; CO(I–III): Cytochrome oxidase subunits (I–III); A6: ATPase subunit 6.

Genes of unknown function

M(2. 5): Maxicircle Unidentified Reading Frame (MURF, 2, 5); CR(3, 4): C-rich Reading frame (3, 4.

Authors' contributions

CK conceived the study, CK, TAW and WG contributed to the study design, CK analysed the data. WG provided materials. CK wrote the paper with contributions from all authors. All authors read and approved the final manuscript.

Funding

This study was funded by the UK Biotechnology and Biological Sciences Research Council project Grant BB/R016437/1. T.A.W. is supported by a Royal Society University Research Fellowship.

Availability of data and materials

Public data used to assemble sequences in this study, and the assembled sequences are freely available, and listed by run experiment, Bioproject, and Genbank accession in Additional file 2: Table S1. Assembled sequences used for our analyses are collected together in Additional file 12.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

No competing interests to declare.

Received: 6 May 2020 Accepted: 12 October 2020
Published online: 09 December 2020

References

1. Lukes J, Hashimi H, Zíková A. Unexplained complexity of the mitochondrial genome and transcriptome in kinetoplastid flagellates. Curr Genet. 2005;48(5):277–99.
2. Jensen RE, Englund PT. Network news: the replication of kinetoplast DNA. Annu Rev Microbiol. 2012;66:473–91.
3. Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, Tromp MC. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell. 1986;46(6):819–26.
4. Sturm NR, Simpson L. Partially edited mRNAs for cytochrome b and subunit III of cytochrome oxidase from leishmania tarentolae mitochondria: RNA editing intermediates. Cell. 1990;61:871–8. https://doi.org/10.1016/0092-8674(90)90197-m.
5. Stuart K, Panigrahi AK. RNA editing: complexity and complications. Mol Microbiol. 2002;45(3):591–6.
6. Simpson L, Maslov DA. Ancient origin of RNA editing in kinetoplastid protozoa. Curr Opin Genet Dev. 1994;4(6):887–94.
7. Stevens JR, Noyes HA, Dover GA, Gibson WC. The ancient and divergent origins of the human pathogenic trypanosomes *Trypanosoma brucei* and *T. cruzi*. Parasitology. 1999;18:107–16. https://doi.org/10.1017/s0031182098003473.
8. Hamilton PB, Stevens JR, Gaunt MW, Gidley J, Gibson WC. Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA. Int J Parasitol. 2004;34(12):1393–404.
9. Stevens JR, Gibson WC. The evolution of pathogenic trypanosomes. Cad Saude Publica. 1999;15(4):673–84.
10. Lewis MD, Llewellyn MS, Yeo M, Acosta N, Gaunt MW, Miles MA. Recent, independent and anthropogenic origins of *Trypanosoma cruzi* hybrids. PLoS Negl Trop Dis. 2011;5(10):e1363.
11. Hamilton PB, Adams ER, Njiokou F, Gibson WC, Cuny G, Herder S. Phylogenetic analysis reveals the presence of the *Trypanosoma cruzi* clade in African terrestrial mammals. Infect Genet Evol. 2009;9(1):81–6.
12. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–6.
13. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36. https://doi.org/10.1101/gr.215087.116.
14. Greif G, Rodriguez M, Reyna-Bello A, Robello C, Alvarez-Valin F. Kinetoplast adaptations in American strains from *Trypanosoma vivax*. Mutat Res. 2015;773:69–82.
15. Westenberger SJ, Cerqueira GC, El-Sayed NM, Zingales B, Campbell DA, Sturm NR. *Trypanosoma cruzi* mitochondrial maxicircles display species- and strain-specific variation and a conserved element in the non-coding region. BMC Genomics. 2006;7:60.
16. Lai D-H, Hashimi H, Lun Z-R, Ayala FJ, Lukes J. Adaptations of *Trypanosoma brucei* to gradual loss of kinetoplast DNA: *Trypanosoma equiperdum* and *Trypanosoma evansi* are petite mutants of *T. brucei*. Proc Natl Acad Sci USA. 2008;105(6):1999–2004.
17. Hoare CA. The trypanosomes of mammals: a zoological monograph. Hoboken: Wiley-Blackwell; 1972. p. 749.
18. Kaufer A, Barratt J, Stark D, Ellis J. The complete coding region of the maxicircle as a superior phylogenetic marker for exploring evolutionary

relationships between members of the Leishmaniinae. Infect Genet Evol. 2019;70:90–100.

19. Molak M, Ho SYW. Prolonged decay of molecular rate estimates for metazoan mitochondrial DNA. PeerJ. 2015;3:e821.
20. Stevens J, Rambaut A. Evolutionary rate differences in trypanosomes. Infect Genet Evol. 2001;1(2):143–50.
21. Machado CA, Ayala FJ. Nucleotide sequences provide evidence of genetic exchange among distantly related lineages of *Trypanosoma cruzi*. Proc Natl Acad Sci USA. 2001;98(13):7396–401.
22. Gibson W, Borst P, Fase-Fowler F. Further analysis of intraspecific variation in *Trypanosoma brucei* using restriction site polymorphisms in the maxicircle of kinetoplast DNA. Mol Biochem Parasitol. 1985;15(1):21–36.
23. Balmer O, Beadell JS, Gibson W, Caccone A. Phylogeography and taxonomy of *Trypanosoma brucei*. PLoS Negl Trop Dis. 2011;5(2):e961.
24. Weir W, Capewell P, Foth B, Clucas C, Pountain A, Steketee P, et al. Population genomics reveals the origin and asexual evolution of human infective trypanosomes. Elife. 2016;5:e11473.
25. Demenou BB, Doucet J-L, Hardy OJ. History of the fragmentation of the African rain forest in the Dahomey Gap: insight from the demographic history of Terminalia superba. Heredity. 2018;120(6):547–61.
26. Lynch M. The origins of genome architecture. Sunderland: Sinauer Associates Incorporated; 2007. p. 494.
27. Seward EA, Kelly S. Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. Genome Biol. 2016;17(1):226.
28. Kelly S, Ivens A, Manna PT, Gibson W, Field MC. A draft genome for the African crocodilian trypanosome *Trypanosoma grayi*. Sci Data. 2014. https://doi.org/10.1038/sdata.2014.24.
29. Barnabé C, Brenière SF. Scarce events of mitochondrial introgression in *Trypanosoma cruzi*: new case with a Bolivian strain. Infect Genet Evol. 2012;12(8):1879–83.
30. Messenger LA, Llewellyn MS, Bhattacharyya T, Franzén O, Lewis MD, Ramírez JD, et al. Multiple mitochondrial introgression events and heteroplasmy in trypanosoma cruzi revealed by maxicircle MLST and next generation sequencing. PLoS Negl Trop Dis. 2012;6(4):e1584.
31. Cuypers B, Van den Broeck F, Van Reet N, Meehan CJ, Cauchard J, Wilkes JM, et al. Genome-wide SNP analysis reveals distinct origins of *Trypanosoma evansi* and *Trypanosoma equiperdum*. Genome Biol Evol. 2017;9(8):1990–7.
32. De Greef C, Imberechts H, Matthyssens G, Van Meirvenne N, Hamers R. A gene expressed only in serum-resistant variants of *Trypanosoma brucei* rhodesiense. Mol Biochem Parasitol. 1989;36(2):169–76.
33. Tihon E, Imamura H, Dujardin J-C, Van Den Abbeele J, Van den Broeck F. Discovery and genomic analyses of hybridization between divergent lineages of *Trypanosoma congolense*, causative agent of Animal African Trypanosomiasis. Mol Ecol. 2017;26(23):6524–38.
34. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.
35. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011;39(1):D19-21.
36. SRA Toolkit Development Team. https://ncbi.github.io/sra-tools/
37. Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. BMC Bioinform. 2019;20(1):405.
38. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
40. Bosi E, Donati B, Galardini M, Brunetti S, Sagot M-F, Lió P, et al. MeDuSa: a multi-draft based scaffolder. Bioinformatics. 2015;31(15):2443–51.
41. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty:de novoassembly of organelle genomes from whole genome data. Nucleic Acids Res. 2016. https://doi.org/10.1093/nar/gkw955.
42. Huang X. CAP3: a DNA sequence assembly program. Genome Res. 1999;9:868–77. https://doi.org/10.1101/gr.9.9.868.
43. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
44. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, et al. Using Tablet for visual exploration of second-generation sequencing data. Brief Bioinform. 2013;14(2):193–202.
45. Seibt KM, Schmidt T, Heitkam T. FlexiDot: highly customizable, ambiguity-aware dotplots for visual sequence analyses. Bioinformatics. 2018;34(20):3575–7.
46. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
47. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.
48. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14(6):587–9.
49. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2019;15(4):e1006650.
50. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol. 2018;67:901–4. https://doi.org/10.1093/sysbio/syy032.
51. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom. 2016. https://doi.org/10.1101/038190.
52. Lischer HEL, Excoffier L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics. 2012;28(2):298–9.
53. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol. 2013;30(5):1224–8.
54. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59.
55. Earl DA, vonHoldt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour. 2012;4:359–61. https://doi.org/10.1007/s12686-011-9548-7.

Publisher's Note