

1 **Genomic evidence for population specific selection in Nilo-Saharan and Niger-**  
2 **Congo linguistic groups in Africa**

3

4 Julius Mulindwa<sup>1</sup>, Harry Noyes<sup>2</sup>, Hamidou Ilboudo<sup>3</sup>, Oscar Nyangiri<sup>1</sup>, Mathurin  
5 Koffi<sup>4</sup>, Dieudonne Mumba<sup>5</sup>, Gustave Simo<sup>6</sup>, John Enyaru<sup>1</sup>, John Chisi<sup>7</sup>, Martin  
6 Simuunza<sup>8</sup>, Pius Alibu<sup>1</sup>, Veerle Lejon<sup>9</sup>, Vincent Jamonneau<sup>9</sup>, Annette Macleod<sup>10</sup>,  
7 Bruno Bucheton<sup>9,11</sup>, Christiane Hertz-Fowler<sup>2</sup>, Issa Sidibe<sup>3</sup>, Enock Matovu<sup>1</sup> for the  
8 TrypanoGEN Research Group, as members of The H3Africa Consortium.

9

10

11 <sup>1</sup>Makerere University, Kampala, Uganda, Bobo-Dioulasso, Burkina Faso; <sup>2</sup>Centre for Genomic  
12 Research, University of Liverpool, UK; <sup>3</sup>Centre International de Recherche-Développement sur  
13 l'Élevage en zone Subhumide (CIRDES); <sup>4</sup>Université Jean Lorougnon Guédé (UJLoG), Daloa, Côte  
14 d'Ivoire; <sup>5</sup>Institut National de Recherche Biomedicale, Kinshasa, Democratic Republic of Congo;  
15 <sup>6</sup>Faculty of Science, University of Dschang, Cameroon; <sup>7</sup>University of Malawi, College of Medicine,  
16 Department of Basic Medical Sciences, Blantyre, Malawi; <sup>8</sup>Department of Disease Control, School of  
17 Veterinary Medicine, University of Zambia, Lusaka, Zambia; <sup>9</sup>Institut de Recherche pour le  
18 Développement (IRD), IRD-CIRAD 177, Montpellier, France; <sup>10</sup>Wellcome Center for Molecular  
19 Parasitology, University Place, Glasgow, UK; <sup>11</sup>Programme National de Lutte contre la  
20 Trypanosomose Humaine Africaine, Conakry, Guinea

21

22 **1.0 Abstract**

23 **Background:** There are over 2000 genetically diverse ethno-linguistic groups in  
24 Africa that could help decipher human evolutionary history and the genetic basis of  
25 phenotypic variation. We have analysed 298 genomes from Niger-Congo populations  
26 from six sub-Saharan African countries (Uganda, Democratic Republic of Congo,  
27 Cameroon, Zambia, Ivory Coast, Guinea) and a Nilo-Saharan population from

28 Uganda. These samples were collected as part of the TrypanoGEN consortium project  
29 <http://www.trypanogen.net>.

30 **Results:** The population genetic structure of the 298 individuals revealed four clusters  
31 which correlated with ethno-linguistic group and geographical latitude, that is, West  
32 African Niger-Congo A, Central African Niger Congo, East African Niger-Congo B  
33 and the Nilo-Saharan.

34 We observed a spatial distribution of positive natural selection signatures in genes  
35 associated with AIDS, Tuberculosis, Malaria and Human African Trypanosomiasis  
36 among the TrypanoGEN samples. Having observed a marked difference between the  
37 Nilo-Saharan Lugbara and Niger-Congo populations, we identified four genes  
38 [APOBEC3G, TOP2B, CAPN9, LANCL2, ( $\pi$   $-\log p > 3.0$ ,  $R_{sb} -\log p > 3.0$ ,  $F_{st}$   
39  $> 0.1$  bonferroni  $p > 1.8 \times 10^4$ )], which are highly differentiated between the two  
40 ethnic groups and under positive selection in the Lugbara population.

41 **Conclusion:** The signatures that differentiate ethnically distinct populations provide  
42 information on the specific ecological adaptations with respect to disease history and  
43 susceptibility/ resistance; as demonstrated in this study where *APOBEC3G* is believed  
44 to be involved in the susceptibility of the Nilo-Saharan Lugbara population to  
45 Hepatitis B virus infection.

46

## 47 **2.0 Background**

48 The African continent's ethno-lingual groups have been classified into four major  
49 families, Afro-Asiatic, Nilo-Saharan, Niger-Congo, and Khoisan (Blench 2006). The  
50 Afro-Asiatic which includes the Semitic, Cushitic, and ancient Egyptian languages, is  
51 spoken predominantly by northern and eastern African pastoralists and agro-  
52 pastoralists; the Nilo-Saharan, which includes the Central Sudanic and Eastern

53 Sudanic (Nilotic) languages, is spoken predominantly by eastern and central Saharan  
54 pastoralists; the Niger-Congo languages are subdivided into the Niger-Congo A in  
55 West Africa and the Niger-Congo B or Bantu in Central, Southern and Eastern Africa  
56 (Greenberg 1963; Lewis et al. 2009). Fourteen ancestral population clusters have been  
57 identified amongst these groups that correlate with shared cultural and linguistic  
58 affiliations (Tishkoff et al. 2009). These 14 ancestral populations break down further  
59 into over 2000 ethnically diverse linguistic groups (Bryc et al. 2010; Tishkoff and  
60 Williams 2002).

61 The diversity of ethno-linguistic groups can be used to study human evolutionary  
62 history and the genetic basis of phenotypic variation (Tishkoff et al. 2009),  
63 complementing studies of African genotype variations  
64 (Tishkoff et al. 2009; Gurdasani et al. 2015; Busby et al. 2016; Patin et al. 2017)  
65 which have contributed to the understanding of human origins and disease  
66 susceptibility markers.

67 However, samples from sufficient individuals for population analysis have been  
68 sequenced from relatively few African populations. The 1000 genome project  
69 generated data from five Niger-Congo populations, The African Variome project  
70 added Afro-Asiatic populations and there have been small scale studies of the  
71 Khoisan hunter-gatherers (Kim et al. 2014; Mallick et al. 2016; Tishkoff et al. 2009;  
72 Gurdasani et al. 2015). However no sequences of Nilotic populations have been  
73 published to date although one previous study used 200,000 SNP loci to examine  
74 genetic diversity of the Nilo-Saharan speaking population of southern Sudan  
75 Darfurian and Nuba people (Dobon et al. 2015). In the present study we present the  
76 first genome sequences of a Nilo-Saharan population and genome sequences from six  
77 new Niger-Congo populations.

78

## 79 **3.0 Results**

### 80 **3.1 Samples and sequencing**

81 The samples used for this study were collected by the TrypanoGEN consortium and  
82 consisted of 298 individuals from 19 linguistic groups residents of Guinea, Ivory  
83 Coast, Cameroon, Democratic Republic of Congo, Uganda and Zambia (Table1). The  
84 DNA from the study participant's blood samples was extracted and genomes were  
85 sequenced on the Illumina 2500 at 10X coverage, except for the Zambia and Cameroon  
86 samples that were sequenced at 30X coverage.

87 Following mapping and SNP calling, we identified approximately 34.1 million single  
88 nucleotide polymorphisms (SNPs) and 5.3 million insertion/deletion polymorphisms  
89 (Table 2). We identified 2.02 million variants that did not have rsIDs and we hence  
90 consider them 'novel'. The SNPs had a transition-transversion ratio of 2.0  
91 (Supplementary figure S1), implying good quality SNP calls (DePristo et al. 2011).  
92 Prior to population analysis, variants (SNPs and Indels) were filtered by removing  
93 loci with >10% missing data, MAF < 0.05 or Hardy Weinberg Equilibrium (HWE) P-  
94 value < 0.01. 13 individuals with > 10% SNP loci missing were removed from the  
95 data (Table 2). To our 298 samples, 504 additional samples from five African  
96 populations from the 1000 genomes project (Esan and Yoruba from Nigeria, Mende  
97 from Sierra Leone, Mandinka from Gambia and Luhya from Kenya), were included in  
98 some of our analyses.

99

### 100 **3.2 Population stratification by Multiple Dimensional Scaling**

101 Multiple Dimensional Scaling (MDS) implemented in Plink 1.9 was used to help  
102 visualise genetic distances between samples (Figure 1). All TrypanoGEN samples

103 clustered by country except those in Uganda, where the Nilo-Saharan Lugbara  
104 samples formed a distinct cluster from the Basoga samples. When the samples from  
105 the six TrypanoGEN and the four African 1000 genomes project countries were  
106 merged, five groups representing five major geographic groups were observed (Figure  
107 1B): the Uganda Nilo-Saharan; East African Bantu speakers from Uganda and  
108 Kenya; Central African Bantu speakers from Cameroon, DRC and Zambia; Nigerian  
109 Niger-Congo A speakers (Esan and Yoruba);. West African Niger-Congo A speakers  
110 from the Ivory Coast, Gambia, Sierra Leone and Guinea. The African and European  
111 samples were very distinct (Figure 1C). Since all samples except Ugandan Bantu and  
112 Nilo-Saharan clustered by country by MDS, samples were grouped by country for  
113 subsequent analyses, except for the Uganda samples which were grouped by both  
114 country and linguistic group.

115

### 116 **3.3 Population Admixture and differentiation**

117 The amount of shared genetic ancestry within the samples was estimated using  
118 Admixture (Alexander et al. 2009). Admixture was run on 2-8 population clusters (K)  
119 in triplicate; with K=4, K=5 and K=6 having the lowest cross validation errors and  
120 hence the most probable numbers of ancestral components represented in the data  
121 (Supplementary Figure S2). At K=6 The Niger-Congo populations exhibited 17-60%  
122 admixture with minor ancestries, whilst the Ugandan Nilo-Saharan population had 7%  
123 admixture with Niger Congo ancestries (Figure 2A).

124 At K4 one European and three ancestral African populations were observed which  
125 corresponded to Nilo-Saharan, Niger-Congo-B (East African) and Niger-Congo-A  
126 (West African). At K5 a homogeneous group of seven samples emerged within the  
127 Zambia population with no admixture with other populations in our data set and were

128 also outliers on the MDS plot (Figure 1B). These seven were recorded as  
129 Soli/Chikunda speakers, which are Bantu languages but they had no admixture at  
130 (K=5 and K=6) with the other speakers of this language group from Zambia or any  
131 other group included in this study, suggesting that they were from a quite distinct  
132 population. At K6, a major group appeared that contributed ancestry to both East  
133 African Niger-Congo B and West African Niger-Congo A but does not correspond to  
134 any existing linguistic group.

135 The genetic variation within the populations that are part of the TrypanoGEN project  
136 was estimated using the pairwise  $F_{ST}$  (Wright 1949) (Figure 2B, supplementary fig  
137 S3).  $F_{ST}$  was relatively high between the Nilo-Saharan Lugbara samples and the  
138 African Bantu populations (Figure 2B) except the East African Basoga (population  
139 mean  $F_{ST}$  = 0.012) and Luhya (population mean  $F_{ST}$  = 0.011), presumably due to the  
140 30% admixture of Nilo-Saharan origin within these populations. The pattern of the  
141 observed genetic variation was consistent with the relative geographic distance from  
142 the Nilo-Saharan population (Figure 2C). In addition, a phylogenetic tree based on the  
143 genetic distances between populations ( $F_{ST}$ ) showed clustering of populations by  
144 geographic region on the African continent (Figure 2D).

145

#### 146 **3.4 Population size over time and timing of population isolation.**

147 The multiple sequentially Markovian coalescent (MSMC) was used to estimate  
148 population sizes over time and times at which populations became isolated (Figure 3).  
149 Effective population sizes ( $N_e$ ) were relatively stable at around 13,000 in all  
150 populations tested from 100 thousand years ago (kya) until about 50kya when they  
151 started to decline reaching a nadir of about 8,000 about 13kya coinciding with the dry  
152 period at the end of the last ice age (Figure 3A, Supplementary table S6). All

153 population sizes increased rapidly thereafter but the Niger-Congo populations  
154 increased to an  $N_e$  of around 200,000, whilst the Nilotic population only increased to  
155 60,000. The Ugandan Bantu population was intermediate in  $N_e$  presumably due to  
156 admixture with the Nilotics. This post glacial population increase was briefly reversed  
157 in the Central and West African populations which suffered declines of 6-23%  
158 between 1500 and 750 years ago before recovering to even higher levels at the present  
159 time. This decline in  $N_e$  was not observed in the Ugandan Bantu population, although  
160 the growth rate declined, and in the Nilotic population a decline was observed at a  
161 later time point after 750 years ago.

162 Population separation data is less clear and may be more sensitive to admixture  
163 (Figure 3B). The Guinea and Ivory Coast populations were the least admixed and  
164 appeared panmictic until about 10kya, and had become isolated by about 3kya. The  
165 Ugandan Bantu and Ugandan Nilotic appeared to begin separating from other  
166 populations about 23 and 47kya, respectively and became isolated about 3kya but  
167 these estimations may be confounded by admixture.

168

### 169 **3.5 Genome-wide screen for extended haplotypes under selection**

#### 170 *Signatures within population*

171 In order to identify alleles under selection pressure, we used the within population  
172 Extended Haplotype Homozygosity (EHH) test (Sabeti et al. 2002). Similar patterns  
173 of loci with extreme positive and negative  $iHS$  scores were observed across all groups  
174 (Supplementary Figure S4A). The  $iHS$  values for all groups had an approximate  
175 normal distribution (Supplementary Figure S5) implying that the sizes of  $iHS$  signals  
176 from different SNPs in all the populations were comparable (Voight et al. 2006). The  
177 mean number of loci with extreme positive and negative  $iHS$  score ( $-\log p > 3$ ) from

178 all groups was 8,984, Guinea had the largest number of loci with extreme iHS score  
179 (11,401) and Zambia had the least (5,570) (Table 3, Supplementary Table S1). These  
180 extreme loci were classified by the Ensembl annotation of the nearest gene.  
181 Approximately 34% of these annotations were for protein coding genes; a mean of  
182 3,058 SNPs in protein coding genes per population were associated with extreme iHS  
183 scores. Some protein coding genes with extreme iHS SNP loci were shared between  
184 different Countries whereas some occurred only in a single Country population  
185 (Supplementary Table S1, sheet 'ALLpop.protein\_coding'). We observed strong iHS  
186 signatures in genes that have been previously identified in other African populations  
187 as being under strong selection (Voight et al. 2006; Gurdasani et al. 2015; Sabeti et al.  
188 2007). These included *SYTI*, a synaptosomal protein implicated in Alzheimer's  
189 disease (Yoo et al. 2001) was found in all Country populations, *LARGE* a glycosylase  
190 involved in Lassa fever virus binding (Andersen et al. 2012) (Zambia, Cameroon,  
191 Ivory Coast), *CDK5RAP2*, a microcephaly gene controlling brain size (Bond et al.  
192 2005) (Ugandan Bantu), *NCOA1* a transcriptional co-activator associated with  
193 Lymphoma (Guinea, Ivory Coast, DRC), *SIGLEC12* involved in immune responses  
194 (Crocker et al. 2007) (Zambia, Cameroon). Using the DAVID annotation (Huang et  
195 al. 2008) we observed that all of the Country populations had strong signals that have  
196 been implicated in communicable diseases such as HIV/AIDS, Malaria and  
197 Tuberculosis that have the highest burden on the African continent (Bhutta et al.  
198 2014) (Table 4), suggesting an adaptive role of these genes to infection.  
199 Having collected samples from Human African Trypanosomiasis endemic regions, we  
200 identified signatures that have been implicated in Trypanosome infection. These  
201 signatures were observed in genes overlapping the KEGG calcium signalling pathway  
202 (<http://www.kegg.jp/>)(Kanehisa et al. 2016); *F2RL1* (Guinea, Ivory Coast), *GNAI4*



203 (Zambia), **GNAQ** (Cameroon), **GNAL** (Guinea, Cameroon), **GNAS** (Zambia),  
204 identified mainly from mice studies (Grab 2009). The calcium signalling pathway  
205 regulates permeability of the blood brain barrier to trypanosome parasites during CNS  
206 disease (Nikolskaia et al. 2006). In addition, we observed signatures in genes  
207 overlapping the Mitogen-activated protein kinase MAPK pathway **MAPK1**  
208 (Cameroon), **MAPK10** (Ugandan Nilo-Saharan, DRC, Ugandan Bantu), **MAPK9**  
209 (Zambia); which is targeted by trypanosomatids in order to modulate the host's  
210 immune response (Soares-Silva et al. 2016). These host signalling pathways have  
211 been shown to play a role in host immunity against trypanosome infection in mice and  
212 cattle (Noyes et al. 2011).

213

#### 214 *Signatures unique to Nilo-Saharans*

215 In order to determine which signatures are unique to the Nilo-Saharan Lugbara, we  
216 first ascertained which extreme iHS loci ( $-\log p > 3$ ) were common to the Nilo-  
217 Saharan and one or more Niger-Congo groups. We observed that approximately 15%  
218 of the protein coding gene associated extreme iHS SNPs of the Ugandan Bantu, DRC,  
219 Ivory Coast and Guinea groups were common with the Nilo-Saharan group, whereas  
220 Cameroon and Zambian groups had 2.7% in common (Table 3, supplementary figure  
221 S4B). 149 extreme SNPs associated with protein coding genes were unique to the  
222 Uganda Nilo-Saharan (Supplementary table S2). Using the PANTHER Gene ontology  
223 database (Thomas et al. 2003), we observed that these unique genes were mainly  
224 enriched for cellular and metabolic process proteins (approximately 50.8%)  
225 (Supplementary figure S6). Amongst these were SNPs associated with genes that  
226 have also been shown by other studies to be under positive selection including,  
227 **APOBEC3G**, which is involved in innate anti-viral immunity (Sawyer et al. 2004;

228 Zhang and Webb 2004), has protective alleles against HIV-1 in Biaka and Mbuti  
229 pygmies of Central African Republic and DRC respectively (Zhao et al. 2012); *IFIH1*  
230 (also called *MDA5*) is a cytoplasmic RNA receptor that mediates antiviral responses  
231 by activating type I interferon signalling (Rice et al. 2014) but is also implicated in  
232 protection against type 1 diabetes ((Nejentsev et al. 2009; Fumagalli et al. 2010);  
233 *OR2L13* olfactory receptor involved in activation of signal transduction pathway for  
234 odorant recognition and discrimination (Sharon et al. 1999), and is associated with  
235 Diabetic nephropathy in African Americans (Bailey et al. 2014).

236

### 237 *Nilo-Saharan versus Niger-Congo cross population signatures*

238 There were 299 SNP with high  $F_{ST}$  (above 99<sup>th</sup> percentile) and XPEHH ( $R_{sb} -\log p$   
239  $> 3$ ) in the regions of protein coding genes that were also highly differentiated  
240 between the Nilo-Saharan and Niger-Congo populations (Supplementary table S3).  
241 We then compared SNP loci with derived alleles that are unique to the Nilo-Saharans  
242 and occur in highly differentiated genes (extreme  $R_{sb}$ , high  $F_{ST}$ ) between the Nilo-  
243 Saharan and Niger-Congo groups. From this we identified 12 genes (Table 5,  
244 Supplementary figure S8B) including the *APOBEC3G* gene that are highly  
245 differentiated between the Nilo-Saharan and Niger-Congo groups (mean  $F_{ST}$  0.11,  
246  $R_{sb} -\log p$  4.1). *APOBEC3G* also contains the SNP rs112077004, which was  
247 observed to be under positive selection in the Nilo-Saharans (Figure 4, Supplementary  
248 figure S9).

249

## 250 **4. Discussion**

251 We have analysed the genomes of 298 individuals from seven major groups of  
252 samples from six Sub-Saharan Africa Countries, investigating their admixture profile,

253 demographic histories and signatures of selection that differentiate the major  
254 linguistic groups. The MDS analysis identified five major clusters: Nilo-Saharan, two  
255 Niger-Congo A groups from Nigeria and West Africa and two groups of Niger-Congo  
256 B (Bantu) speakers from Central and East Africa, which were consistent with  
257 previous studies (Tishkoff et al. 2009; Gomez et al. 2014; Gurdasani et al. 2015). The  
258 samples represented three of the five major linguistic groups in Africa. Afro-Asiatic  
259 speakers are found across North and North-East Africa in regions adjacent to Nilo-  
260 Sharan and Bantu speakers. Afro-Asiatic reference populations were not included in  
261 this study and we are therefore not able to detect any admixture from this source.  
262 However a SNP genotype based analysis of Nilotic populations indicated that Nilotic  
263 populations only contain a trace of Afro-Asiatic ancestry and therefore our  
264 observations on East African populations may not be significantly limited by the  
265 absence of Afro-Asiatic data (Dobon et al. 2015).

266 **Admixture:** Niger-Congo speaking hunter-gathers are believed to have originated  
267 from the Kordofanian speakers of the Nuba mountains of Sudan and then traversed  
268 the Sahel to Mali (Figure 5). They then colonised the coast from Senegal to Nigeria  
269 and Cameroon, over several thousand years forming multiple linguistic groups. The  
270 Bantu (Niger-Congo-B) speaking people emerged as another linguistic group amongst  
271 the greater than 60 Niger-Congo-A groups in the Nigeria/Cameroon region about  
272 3,000 years ago. Bantu speaking peoples then spread South and East along savannah  
273 corridors through the Congo basin and emerged in the Great Lakes region and spread  
274 North to the Lake Victoria region and South down the East Side of Africa  
275 (Grollemund et al. 2015; Patin et al. 2017). This rapid expansion is believed to have  
276 been enabled by the development of agriculture and later enhanced by the acquisition  
277 of iron tools (Tishkoff et al. 2009).

278 The admixture analysis at  $K=4$  is consistent with this linguistic history and recent  
279 genetic analyses (Patin et al. 2017; Gurdasani et al. 2015) with three African  
280 Ancestral allele clusters (AAC) which can be interpreted as representing Niger-Congo  
281 A languages in West Africa, Niger-Congo B (Bantu) in Central and East Africa and  
282 Nilo-Saharan in Northern Uganda. The Niger-Congo-A speakers in extreme West  
283 Africa appear to have approximately 10% Nilo-Saharan ancestry, consistent with an  
284 ancestral relationship with Nilo-Saharans and this declines towards the East. The  
285 Bantu speakers are a mix of Niger-Congo-A and a distinct putative Bantu ancestral  
286 cluster that it at highest frequency in Nigeria and Cameroon, the Niger-Congo-A  
287 component is displaced by a Nilotic component with easterly latitude whilst the  
288 “Bantu” component remains constant. At  $K=5$  a small AAC of 7 Bantu speakers from  
289 Zambia emerges, who evidently have a genetic heritage that does not match their self-  
290 declared linguistic affiliation, and may be of Khoisan descent. At  $K=6$  a fourth major  
291 African AAC appears (green in Figure 2) with strongest representation in the Nigerian  
292 Yoruba and Esan then tapering off east and west into Central and West Africa. This  
293 does not correspond to any linguistic group and displaces the Niger-Congo-A ancestry  
294 to the east of Nigeria and Niger-Congo-B (Bantu) in Nigeria and to the West. This  
295 ancestral cluster could represent a secondary movement out of Nigeria of migrants  
296 who adopted their hosts language. One possible driver for such a migration, if it  
297 occurred, was the development of iron smelting which may have originated in Nigeria  
298 about 2,500 years ago (Vansina 2006). Irrespective of the true number of ancestral  
299 allele clusters there is evidence of back migration of people with Bantu ancestral  
300 alleles into West Africa as has been observed before (Gomez et al. 2014). This  
301 migration to the west was not accompanied by language expansion as it was to the  
302 east.

303 **Population History:** The estimates of current  $N_e$  obtained from our data with MCMS  
304 (Fig 3A) of around 200,000 in West and Central Africa and 57,000-125,000 in East  
305 Africa (Supplementary Table S6) was consistent with previous observations on other  
306 African samples using the same method (Schiffels and Durbin 2014) but ten times  
307 higher than the estimates of around 20,000 obtained from SNP chip genotype data  
308 (Shriner et al. 2014). The faster growth in the Niger-Congo A and B than the Nilotic  
309 populations appears to predate the Bantu expansion. The Niger-Congo A population  
310 was believed to be expanding through West Africa as the climate became wetter after  
311 10kya, consistent with the separation times between the Guinea and Ivory Coast  
312 populations observed on the Cross-Coalescence Plot (Figure 3B). The Nilotics  
313 population developed a pastoralist economy probably after 6kya but their expansion  
314 into the tsetse belt may have been delayed by trypanosomiasis and other diseases until  
315 the cattle developed tolerance (Gifford-Gonzalez 2000) (Smetko et al. 2015) (Chritz  
316 et al. 2015) and the effective population size did not grow so fast as that of the Niger-  
317 Congo-A populations. The brief population decline dated at ~1340CE by MSMC  
318 coincides with the timing of the Black Death (1343-1353), however time resolution is  
319 low and the decrease was only observed at a single time point. There is evidence of  
320 abandonment of multiple large settlements throughout West Africa around the time of  
321 the Black Death and there is speculation that this was caused by the disease (Chouin  
322 2015). The decrease at this time appears to have impacted the West and Central  
323 African Niger-Congo but not the East African populations. Both Bantu and Nilotic  
324 populations in East Africa were cattle keepers and pastoralists to varying degrees  
325 (Chritz et al. 2015) and the concomitant lower population density and mobile lifestyle  
326 may have made them less vulnerable than the more settled and urbanised West  
327 Africans to plague infection. The more recent decline in the Nilotic Lugbara effective

328 population size is unexplained, but the catastrophic Rinderpest outbreak in the 1880's  
329 and 1890's that killed up 90% of indigenous cattle, which lead to the depopulation of  
330 the East African savannahs and may have ended the dominance of the Nilotic  
331 speaking Maasai over the Bantu Kikuyu could have been a contributory factor (Mack  
332 1970).

333 The Cross-Coalescence plots for comparison between populations other than the  
334 Guinea and Ivory Coast Niger-Congo-A show long periods of separation. This is not  
335 consistent with the Ugandan Bantu populations having separated from Niger-Congo-  
336 A populations even more recently than the separation between Guinea and Ivory  
337 Coast populations, and is presumably due to the extensive admixture with the Nilotics  
338 observed in this population. The Central African cross-coalescence data also indicated  
339 older separation times than linguistic evidence suggests (not shown) and although  
340 there was less evidence of admixture in this population these data should be treated  
341 with caution.

342 **Selective Sweeps:** We identified selective sweeps in genes that have been associated  
343 with HIV/AIDS, Tuberculosis and Malaria. Given the high prevalence of these  
344 infections on the continent (Bhutta et al. 2014), there is increased frequency of these  
345 beneficial heritable traits hence positive natural selection. However not all these genes  
346 occurred in all the populations demonstrating spatially varying selection probably due  
347 to differing environmental pressures (Gillespie 1994; Thorne et al. 1998).

348 We identified signatures in genes that are involved in pathways implicated in  
349 trypanosome infection: calcium signalling, (Grab et al. 2009; Nikolskaia et al. 2006),  
350 the MAPK pathway (Noyes et al. 2011), *HPR*, *APOL1*, *IL6* and *HLA-G*, (Hardwick et  
351 al. 2013; Genovese et al. 2010; Cooper et al. 2017; Courtin et al. 2013; 2007)  
352 (Supplementary figure S10, Supplementary table S5). We only found evidence for

353 selection for the calcium signalling and MAPK pathway genes. This suggests that  
354 HAT may have had a selective force in these populations.,

355 In order to determine signatures of selection unique to the Nilo-Saharan population,  
356 we used a combination of linkage disequilibrium-based method (iHS and Rsb) and  
357 population differentiation based method ( $F_{ST}$ ) (2013a). Using this approach we  
358 identified 12 loci associated with coding genes, which are unique to the Nilo-Saharan  
359 Lugbara population and highly differentiated from the Niger-Congo population.  
360 Among these was the variant associated with *APOBEC3G* that demonstrated  
361 significant positive selection in the Lugbara Nilo-Saharan population. This protein is  
362 involved in viral innate immunity (2003a), by inducing a high rate of dC to dU  
363 mutations in the nascent reverse transcripts leading to the degradation of the viral  
364 genome (2004c; 2004a). The Lugbara have relatively low prevalence of HIV (4%) in  
365 comparison to the Basoga (6.4%) and Baganda (10.7%) Bantu groups of Uganda but  
366 relatively high prevalence of Hepatitis B suggesting that either *APOBEC3G* does not  
367 control both these viruses or it has different effects on each (2011b)(2003b; 2003c)  
368 (2009a; 2013b). (2017).

369 We also identified the missense variant rs10930046 (T/C) located in the *IFIH1* CDS,  
370 which was unique to the Nilo-Saharan Lugbara and highly selected (iHS  $-\log$  p-value  
371 3.264). This gene is associated with up regulation of type I interferon signalling  
372 occurring in a spectrum of human diseases (2014a) and is believed to be involved in  
373 the suppression of Hepatitis B viral replication (2013d). Being a nonsynonymous  
374 variant, rs10930046 could alter the functioning of IFIH1 and thus increase  
375 susceptibility to HBV in the Lugbara population, something that could be tested by a  
376 candidate gene study for DNA virus infections. Northern Uganda is considered to  
377 have one of the highest prevalence of Hepatitis B virus in the world (2015a) which

378 has perhaps resulted in a unique adaption of the Lugbara Nilo-Saharan population to  
379 infection.

380

## 381 **5. Conclusion**

382 We have incorporated a Nilo-Saharan population into a analysis of genomic  
383 sequences of Niger-Congo populations for the first time and show extensive  
384 admixture between Nilo-Saharan ancestry and Niger-Congo B (Bantu) populations.  
385 We show evidence for signatures of selection the Nilo-Saharan population in genes  
386 associated with communicable diseases that have different prevalences from  
387 surrounding Bantu (Niger-Congo B) populations.

388

## 389 **6. Materials and Methods**

### 390 **Ethical approval and sample collection**

391 The samples used for this study are part of the TrypanoGEN biobank (Ilboudo et al.  
392 2017), which describes ethics approval, recruitment, sample processing and the meta  
393 data collected. The ethical approval for the study was provided by the national ethics  
394 councils of the TrypanoGEN consortium countries involved in the sample collection  
395 which are: Uganda (HS 1344), Zambia (011-09-13), Democratic Republic of Congo  
396 (No 1/2013), Cameroon (2013/364/L/CNERSH/SP), Cote d'Ivoire (2014/No  
397 38/MSLS/CNER-dkn), and Guinea (1-22/04/2013). All the participants in the study  
398 were guided through the consent forms, and written consent was obtained to collect  
399 biological specimens. Peripheral blood was collected from the participants at the field  
400 sites, transported to reference laboratories from where DNA extraction was carried  
401 out using the Whole blood MidiKit (Qiagen). The DNA was quantified using the  
402 Qubit (Qiagen) and approximately 1µg was shipped from each country to the



403 University of Liverpool, UK except for Cameroon and Zambia from where DNA was  
404 shipped to Baylor College, USA.

405

## 406 **6.1 Sequencing and SNP calling**

407 The whole genome sequencing libraries were prepared using the Illumina Truseq  
408 PCR-free kit and sequencing done using the Illumina Hiseq2500. The samples from  
409 Guinea, Cote D'Ivoire, Uganda and DRC were sequenced to 10x coverage at the  
410 Center for Genomic Research at the University of Liverpool. The samples from  
411 Zambia and Cameroon were sequenced to 30X at the Baylor College of Medicine  
412 Sequencing Facility.

413 The sequenced reads were mapped onto the 1000 genomes project  
414 human\_g1k\_v37\_decoy reference genome using BWA. The SNP calling on all the  
415 samples was done using the genome analysis tool kit GATK v3.4. The SNPs were  
416 then filtered by; a) removing loci with > 10% missing SNP, b) removing individuals  
417 with > 10% missing SNP loci and c) removing loci with Hardy Weinberg P value <  
418 0.01. In addition, loci with MAF < 0.05 were also removed for the PCA and  
419 Admixture analysis. The variant annotation was done using snpEff  
420 ([www.snpeff.sourceforge.net](http://www.snpeff.sourceforge.net)).

421

## 422 **6.2 PCA analysis**

423 The principal component analyses (PCA) were performed using Plink 1.9 and R v  
424 3.2.1. Data were filtered using the following criteria: a) removing loci with > 10%  
425 missing SNP, b) removing individuals with > 10% missing SNP loci and c) removing  
426 loci with Hardy Weinberg P value < 0.01, removing loci with minor allele frequencies  
427 (MAF) < 0.05. SNP loci less than 2000bp apart were removed in order to reduce the

428 linkage disequilibrium (LD) between adjacent SNP. PCA was carried out for (i) all  
429 TrypanoGEN data , (ii) all TrypanoGEN data plus African 1000 genome data, (iii) all  
430 TrypanoGEN data including 50 European and all African 1000 genome data  
431 excluding African Caribbean in Barbados (ACB) and African Southwest USA (ASW)  
432 populations.

433

### 434 **6.3 Population Admixture**

435 The population ancestry of each individual was obtained using Admixture 1.23  
436 (Alexander et al., 2009) on the filtered PLINK .bed files on the same TrypanoGEN,  
437 one thousand genome African and European population data sets analysed by PCA.  
438 Admixture was run on K1 to K8 for which three replicates were done for each run.  
439 The Admixture plots were drawn using the R tool ‘strplot’ (Ramasamy et al. 2014).

440

### 441 **6.4 Genetic diversity: Fst**

442 The genetic diversity due to difference in allele frequency among populations was  
443 analysed by the inter-population Wright’s Fst (Wright, 1951) in PLINKv1.9. The Fst  
444 estimates were made between TrypanoGEN (UGN, UGB, DRC, CIV, GUI) and one  
445 thousand genome African (LWK, YRI, ESN, MSL, GWD) populations. The  $F_{ST}$   
446 dendrogram was generated using Fitch in Phylip3.685 (1993). The geographic  
447 distance matrix between populations was calculated based on their global position  
448 system (GPS) coordinates (2011a).

449

### 450 **6.5 Population History.**

451 Population sizes and divergence times were calculated using MSMC (Schiffels and  
452 Durbin 2014). Since PCA and Admixture analysis had indicated little difference

453 between linguistic groups in each country with the exception of the Ugandan Bantu  
454 and Nilotic populations, samples from each country with highest coverage were  
455 analysed together except for Uganda where Bantu and Nilotic samples were analysed  
456 as separate populations. For population size estimates output from 3 independent runs  
457 each using 8 different haplotypes were combined. Using 8 haplotypes rather than 4  
458 gives higher resolution at more recent time points. For estimates of relative cross  
459 coalescence rate, three replicate runs were done, each using 2 different samples (4  
460 haplotypes) from each pairwise comparison between populations. Results presented  
461 are the means of the replicates.

462

## 463 **6.6 Signatures of selection**

464 The estimation of haplotypes was carried out by Phasing of the genotyped SNPs using  
465 SHAPEIT v2.2 software (Delaneau et al., 2013). The extended haplotype  
466 homozygosity (EHH) was then analysed using the R software package *rehh* (Gautier  
467 et al., 2012). Two main EHH derived statistics were calculated from the phased  
468 haplotype data, that is, intra-population integrated haplotype Score (iHS) (Voight et  
469 al., 2006) and inter-population  $R_{sb}$  (Tang et al, 2007). Bedtools v2.26.0 was used to  
470 identify the intersection of the  $F_{st}$  and  $R_{sb}$  loci.

471

## 472 **Acknowledgements:**

473 The authors would like to acknowledge the study participants who donated their  
474 specimen, the personnel involved in the community engagement and coordinating  
475 sample collection and processing, the National sleeping sickness control programmes of  
476 the participating Countries. Z Lombard (University of Witwatersrand) and D  
477 Adeyemo (NHGRI) for facilitating sequencing of samples from Zambia and

478 Cameroon at Baylor College of Medicine. NIH grant XXXX for sequencing at  
479 Baylor.

480

481 **Disclosure declaration:**

482 The authors declare no competing interests.

483

484

485 **References:**

486 Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry  
487 in unrelated individuals. *Genome Res* **19**: 1655–1664.

488 Busby G, Band G, Si Le Q, Jallow M, Bougama E, Mangano V, Amenga-Etego L,  
489 Emil A, Apinjoh T, Ndila C, et al. 2016. *Admixture into and within sub-Saharan*  
490 *Africa*.

491 Chouin GLF. 2015. Fossés, enceintes et peste noire en Afrique de l’Ouest forestière  
492 (500-1500 AD). *Afrique : Archeologie et Arts* 43–66.

493 Chritz KL, Marshall FB, Zagal ME, Kirera F, Cerling TE. 2015. Environments and  
494 trypanosomiasis risks for early herders in the later Holocene of the Lake Victoria  
495 basin, Kenya. *Proc Natl Acad Sci USA* 201423953–6.

496 Cooper A, Ilboudo H, Alibu VP, Ravel S, Enyaru J, Weir W, Noyes H, Capewell P,  
497 Camara M, Milet J, et al. 2017. APOL1 renal risk variants have contrasting  
498 resistance and susceptibility associations with African trypanosomiasis. *Elife* **6**:  
499 56.

500 Courtin D, Milet J, Jamonneau V, Yeminanga CS, Kumeso VKB, Bilengue CMM,  
501 Betard C, Garcia A. 2007. Association between human African trypanosomiasis  
502 and the IL6 gene in a Congolese population. *Infection, Genetics and Evolution* **7**:  
503 60–68.

504 Courtin D, Milet J, Sabbagh A, Massaro JD, Castelli EC, Jamonneau V, Bucheton B,  
505 Sese C, Favier B, Rouas-Freiss N, et al. 2013. HLA-G 3’ UTR-2 haplotype is  
506 associated with Human African trypanosomiasis susceptibility. *Infection,*  
507 *Genetics and Evolution* **17**: 1–7.

508 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis  
509 AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation  
510 discovery and genotyping using next-generation DNA sequencing data. *Nat*  
511 *Genet* **43**: 491–498.

512 Dobon B, Hassan HY, Laayouni H, Luisi P, Ricaño-Ponce I, Zhernakova A,

- 513 Wijmenga C, Tahir H, Comas D, Netea MG, et al. 2015. The genetics of East  
514 African populations: a Nilo-Saharan component in the African genetic landscape.  
515 *Sci Rep* **5**: 9996.
- 516 Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, Bowden  
517 DW, Langefeld CD, Oleksyk TK, Knob AU, et al. 2010. Association of  
518 Trypanolytic ApoL1 Variants with Kidney Disease in African-Americans.  
519 *Science (New York, NY)*.
- 520 Gifford-Gonzalez D. 2000. Animal Disease Challenges to the Emergence of  
521 Pastoralism in Sub-Saharan Africa. *African Archaeological Review* **17**: 95–139.
- 522 Gomez F, Hirbo J, Tishkoff SA. 2014. Genetic Variation and Adaptation in Africa:  
523 Implications for Human Evolution and Disease. *Cold Spring Harbor Perspectives*  
524 *in Biology* **6**: a008524–a008524.
- 525 Grollemund R, Branford S, Bostoen K, Meade A, Venditti C, Pagel M. 2015. Bantu  
526 expansion shows that habitat alters the route and pace of human dispersals. *Proc*  
527 *Natl Acad Sci U S A* **112**: 13296–13301.
- 528 Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas  
529 K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, et al. 2015. The African  
530 Genome Variation Project shapes medical genetics in Africa. *Nature* **517**: 327–  
531 332.
- 532 Hardwick RJ, Ménard A, Sironi M, Milet J, Garcia A, Sese C, Yang F, Fu B, Courtin  
533 D, Hollox EJ. 2013. Haptoglobin (HP) and Haptoglobin-related protein (HPR)  
534 copy number variation, natural selection, and trypanosomiasis. *Human Genetics*  
535 **133**: 69–83.
- 536 Huang DW, Sherman BT, Lempicki RA. 2008. Systematic and integrative analysis of  
537 large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**: 44–  
538 57.
- 539 Ilboudo H, Noyes H, Mulindwa J, Kimuda MP, Koffi M, Kabore JW, Ahouty B,  
540 Ngoyi DM, Fataki O, Simo G, et al. 2017. Introducing the TrypanoGEN biobank:  
541 A valuable resource for the elimination of human African trypanosomiasis. *PLoS*  
542 *Negl Trop Dis* **11**: e0005438.
- 543 Kim HL, Ratan A, Perry GH, Montenegro A, Miller W, Schuster SC. 2014. Khoisan  
544 hunter-gatherers have been the largest population throughout most of modern-  
545 human demographic history. *Nat Commun* **5**: 5692–8.
- 546 Mack R. 1970. The great African cattle plague epidemic of the 1890's. *Tropical*  
547 *animal health and production*.
- 548 Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri  
549 N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project:  
550 300 genomes from 142 diverse populations. *Nature* **538**: 201–206.
- 551 Noyes H, Brass A, Obara I, Anderson S, Archibald AL, Bradley DG, Fisher P,  
552 Freeman A, Gibson J, Gicheru M, et al. 2011. Genetic and expression analysis of

- 553 cattle identifies candidate genes in pathways responding to *Trypanosoma*  
 554 *congolense* infection. *Proc Natl Acad Sci USA* **108**: 9304–9309.
- 555 Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, Laval G, Perry GH,  
 556 Barreiro LB, Froment A, et al. 2017. Dispersals and genetic adaptation of Bantu-  
 557 speaking populations in Africa and North America. *Science (New York, NY)* **356**:  
 558 543–546.
- 559 Ramasamy RK, Ramasamy S, Bindroo BB, Naik VG. 2014. STRUCTURE PLOT: a  
 560 program for drawing elegant STRUCTURE bar plots in user friendly interface.  
 561 *Springerplus* **3**: 431.
- 562 Sabeti PC, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH,  
 563 McCarroll SA, Gaudet R, Schaffner SF, et al. 2007. Genome-wide detection and  
 564 characterization of positive selection in human populations. *Nature* **449**: 913–918.
- 565 Schiffels S, Durbin R. 2014. Inferring human population size and separation history  
 566 from multiple genome sequences. *Nat Genet* **46**: 919–925.
- 567 Shriner D, Tekola-Ayele F, Adeyemo A, Rotimi CN. 2014. Genome-wide genotype  
 568 and sequence-based reconstruction of the 140,000 year history of modern human  
 569 ancestry. *Sci Rep* **4**.
- 570 Smetko A, Soudre A, Silbermayr K, Müller S, Brem G, Hanotte O, Boettcher PJ,  
 571 Stella A, Mészáros G, Wurzinger M, et al. 2015. Trypanosomosis: potential  
 572 driver of selection in African cattle. *Front Gene* **6**: 1–8.
- 573 Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB,  
 574 Awomoyi AA, Bodo JM, Doumbo O, et al. 2009. The Genetic Structure and  
 575 History of Africans and African Americans. *Science (New York, NY)* **324**: 1035–  
 576 1044.
- 577 Vansina J. 2006. Linguistic Evidence for the Introduction of Ironworking into Bantu-  
 578 Speaking Africa. *History in Africa* **33**: 321–361.
- 579 Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A Map of Recent Positive  
 580 Selection in the Human Genome ed. L. Hurst. *PLoS Biol* **4**: e72.

581

## 582 Tables

583 **Table 1.** Table showing the Ethnic groups and number of individual from each  
 584 Country that were used for Whole genome sequencing

585

Country	District(s)	Ethno-linguistic group(s)	No. of samples	Branch	Family
Uganda (UGN)	Maracha	Lugbara	50	Central Sudanic	Nilo-Saharan
Uganda (UGB)	Iganga	Basoga	33	Bantu	Niger-Congo B
Zambia (ZAM)	Chama, Rufunsa	Soli/Chikunda (28), Tumbuka (14), Bemba(8)	41	Bantu	

Democratic Republic of Congo (DRC)	Bandundu	Ngongo, Songo, Yansi, Mbala	50	Bantu	Niger Congo A
Cameroon (CAM)	Campo, Fontem, Bipindi	Bamilike(6), Mundani(8)	26	Bantoid	
		Ngoumba(12)		Bantu	
Ivory Coast (CIV)	Bonon, Sinfra	Baoule (11),	50	Kwa	
		More (12), Senoufo (4),		Gur	
		Gouro (21), Malinke (1), Koyaka (1)		Mande	
Guinea (GUI)	Forecariah, Boffa, Dubreka	Soussou	48	Mande	

586  
587

588 **Table 2.** The number of SNPs and Indels obtained from the mapping and variant  
589 calling pipeline. The SNPs were filtered for HWE, MAF and missing genotypes

590  
591  
592  
593

\* *Identified 2,023,049 SNPs without rsIDs {Total SNPs (30,591,165) – SNPs with rsIDs (28,568,116)}*

594

595 **Table 3.** Extreme iHS loci that overlap with the UGN population

596

Population	Number	SNPs before filtering	SNPs after filtering	Indels before filtering	Indels after filtering
CIV	50	18,780,913	16,066,827	3,069,408	1,583,594
DRC	50	19,188,537	16,449,696	3,146,802	1,626,826
GUI	48	18,831,834	16,075,002	3,063,080	1,579,352
UGB	33	17,671,306	14,987,699	2,889,915	1,426,646
UGN	50	18,986,243	15,598,629	3,130,979	1,536,490
CAM	26	17,183,994	14,579,603	3,283,543	1,539,459
ZAM	41	18,232,386	15,548,110	3,448,501	1,651,467
<b>Total</b>	<b>298</b>	<b>34,116,333</b>	<b>30,591,165</b>	<b>5,336,622</b>	<b>3,166,196</b>

CIV=Côte d'Ivoire, DRC= democratic republic of Congo, GUI= Guinea, UGB= Uganda Bantu, UGN= Uganda Nilotic, CAM = Cameroun, ZAM= Zambia,

Pop	Extreme iHS SNPs (-log p > 3.0)	Extreme iHS SNPs associated with protein coding genes	Extreme iHS SNPs overlapping with UGN
<b>UGN</b>	8454	2613	2613
<b>UGB</b>	9617	3326	512
<b>DRC</b>	10037	3790	535
<b>ZAM</b>	5570	1990	86
<b>CIV</b>	10129	3541	534
<b>CAM</b>	7686	2597	82
<b>GUI</b>	11401	3741	382

597

598 **Table 4.** DAVID (Huang et al., 2009) analysis of Genes that are highly selected within TrypanoGEN population and associated with HIV,  
599 Tuberculosis, and Malaria. The Fisher's exact test *P*-values indicate significant gene enrichment in the associated disease (1991; 2010; 2001;  
600 2015b; 2014b; 2004b; 2009b; 2016c; 2015c; 2005a; 2013c)  
601

Gene	Chr	Populations affected	Associated Disease	P value	Reference
<b>HLA-DRB1</b>	6p21.32	ZAM,CAM,CIV,DRC,UGB	HIV/TB/Malaria	1.63E-09	Ranasinghe et al., 2013, Hill et al., 1991
<b>NLRP1</b>	17p13.2	ZAM,CIV,DRC,GUI,UGB	HIV	2.42E-07	Pontillo et al., 2010
<b>VPRBP</b>	3p10.6	UGB,CIV,DRC,GUI	HIV	2.42E-07	Zhang et al, 2001
<b>TRIM5</b>	11p15.4	UGN,CAM,GUI	HIV	7.30E-07	Deng et al., 2015
<b>ANKRD30A</b>	10p11.21	DRC,CIV	HIV	2.42E-07	Meyerson et al., 2014
<b>HLA-A</b>	6p22.1	ZAM,CAM	HIV/TB	4.70E-06	Louie et al., 2004
<b>HLA-DQA1</b>	6p21.32	UGB,DRC	HIV/TB	4.70E-06	Louie et al., 2004
<b>HLA-DQB1</b>	6p21.32	UGB,DRC	HIV/TB	4.70E-06	Louie et al., 2004
<b>KIR3DL1</b>	19q13.42	UGN,CIV	Malaria	1.63E-09	Taniguchi et al., 2009, Norman et al., 2013
<b>CD36</b>	7q21.11	UGN,CAM,CIV	Malaria	1.55E-06	Hsieh et al., 2016
<b>DDC</b>	7p12.2	UGB,DRC,GUI	Malaria	1.63E-09	Manjurano et al., 2015
<b>HBE1</b>	11p15.4	UGB,CAM,DRC	Malaria	5.48E-07	Patrinos et al., 2005
<b>ADCY9</b>	16p13.3	UGN,CIV	Malaria	5.48E-07	Maiga et al., 2013

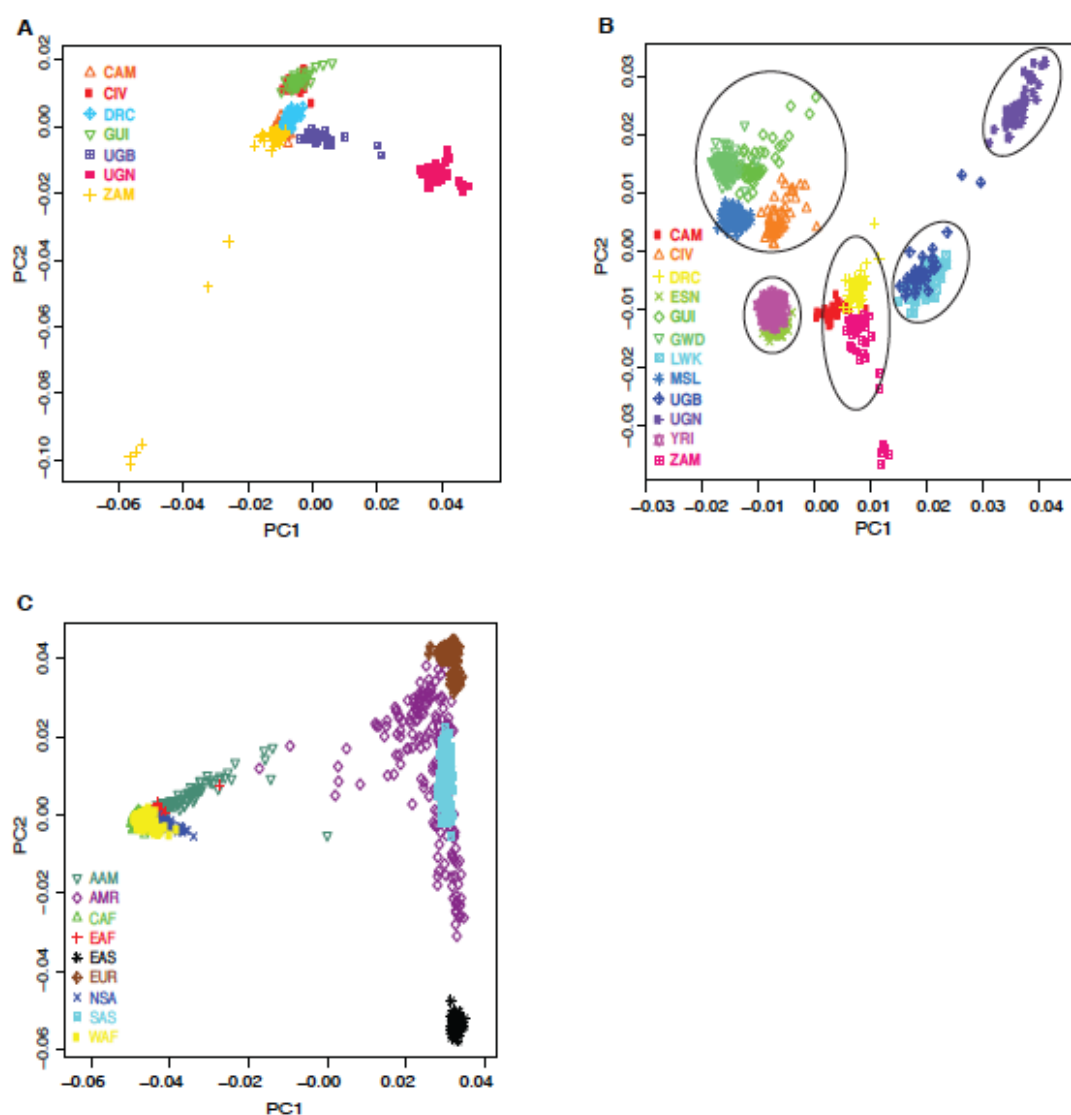


602  
603  
604

**Table 5.** Genes that are highly differentiated between the Nilo-Saharan and Trypanogen Niger congo populations that contain SNPs unique to UGN population

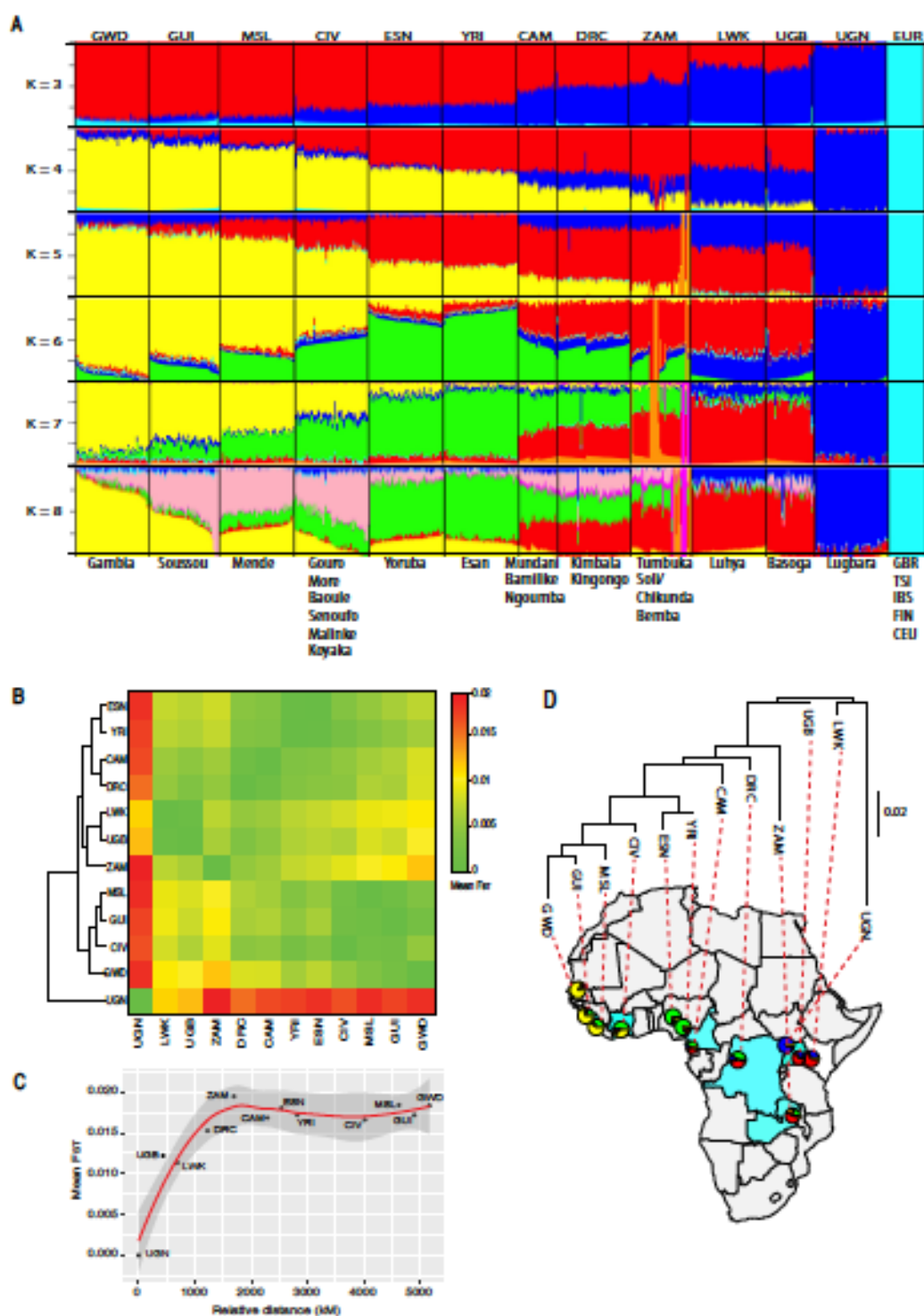
<b>Chr</b>	<b>Gene</b>	<b>Position of Unique SNP</b>	<b>UGN unique SNP</b>	<b>iHS pvalue</b>	<b>Differentiated loci</b>	<b>Mean Fst</b>	<b>Bonferroni pvalue</b>	<b>Rsb pvalue</b>
<b>22</b>	APOBEC3G	22:39453783	rs112077004	4.002	rs5757467	0.110	8.17E-23	4.116
<b>3</b>	TOP2B	3:25670166	rs11712723	3.000	rs6786520	0.115	4.80E-11	4.151
<b>1</b>	CAPN9	1:230886378	rs113802713	3.362	rs16852681	0.105	1.89E-08	5.632
<b>7</b>	LANCL2	7:55476708	rs62457872	3.019	rs3807360	0.109	1.89E-08	3.269
<b>3</b>	NEK4	3:52762698	rs11130321	3.101	rs6445535	0.111	1.51E-04	6.871
<b>20</b>	GDAP1L1	20:42907542	rs1884607	3.187	rs4810417	0.114	0.003070493	3.931
<b>2</b>	NBAS	2:15527280	rs6723183	3.763	rs4668447	0.123	0.053555484	3.020
<b>3</b>	PBRM1	3:52698560	rs12488527	3.130	rs2878632	0.113	0.053555484	6.742
<b>17</b>	ZPBP2	17:38031164	rs11658278	3.115	rs9903250	0.113	0.053555484	3.384
<b>12</b>	MGAT4C	12:86435551	rs11513957	3.134	rs1502802	0.113	0.060467781	3.520
<b>11</b>	FAT3	11:92291634	rs675654	3.063	rs2852859	0.112	0.061504234	3.582
<b>9</b>	MEGF9	9:123462573	rs75959206	3.046	rs1530370	0.113	0.061504234	4.958

605 **Figures**



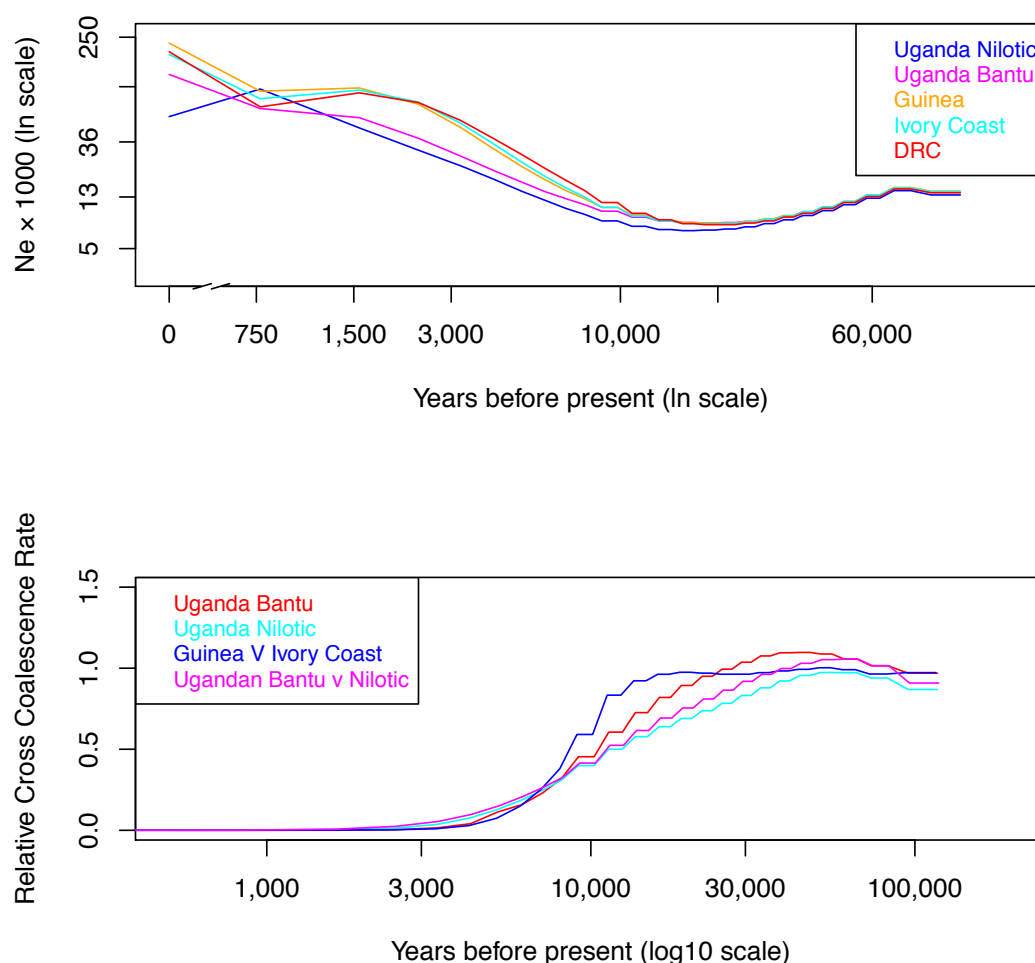
606

607 **Figure 1.** Principal component analysis (PCA) of the sequenced TrypanoGEN  
608 samples, Guinea (GUI), Ivory Coast (CIV), Cameroon (CAM), Democratic Republic  
609 of Congo (DRC), Uganda (Nilotics, UGN, Bantu, UGB) and Zambia (ZAM), (A); B,  
610 TrypanoGEN and selected 1000 genomes African samples Nigeria (ESN, YRI),  
611 Sierra Leone (MSL), Gambia (GWD), Kenya (LWK); C, 1000 genomes samples  
612 from Africa and the rest of the world. AAM, African Americans; AMR, indigenous  
613 Americans; CAF, Central Africa; EAF, East Africa; EAS, East Asia; EUR, Europe;  
614 NSA, Nilo-Saharan; SAS, South Asia; WAF, West Africa;  
615



616

617 **Figure 2.** Genetic admixture and diversity between TrypanoGEN and selected 1000  
 618 genome populations. **A.** Admixture plot of the K populations of the TrypanoGEN,  
 619 1000 genome African and European populations. **B.** Heatmap of mean Fst between  
 620 TrypanoGEN and 1000 genome African populations. **C.** Polynomial regression plot of the mean Fst against the relative geographical distance of the African Niger-Congo  
 621 populations from the Uganda Nilotic population. **D.** Phylogeographic plot of the mean  
 622 Fst distances on the Trypanogen populations and selected 1000 kgenome African  
 623 populations; the pie charts represent the population sample size and admixture.  
 624



625

626 **Figure 3.** Population sizes and cross-coalescence rates compiled by MSMC. **A**  
627 Effective population sizes for each population since 75kya. The Ugandan Bantu and  
628 Nilotic populations have grown continuously but at a slower rate than the West and  
629 Central African populations. These latter populations experienced declines of 6-  
630 23% between 1500 and 800 years ago. **B** Cross-coalescence rates for pairs of  
631 populations. At 1.0 populations are panmictic and at 0.0 there is no gene flow. The  
632 Guinea and Ivory Coast populations were panmictic until about 10 kya and then  
633 became separated by 3kya. Other populations appear to have separated more  
634 gradually but these may be confounded by admixture.

635

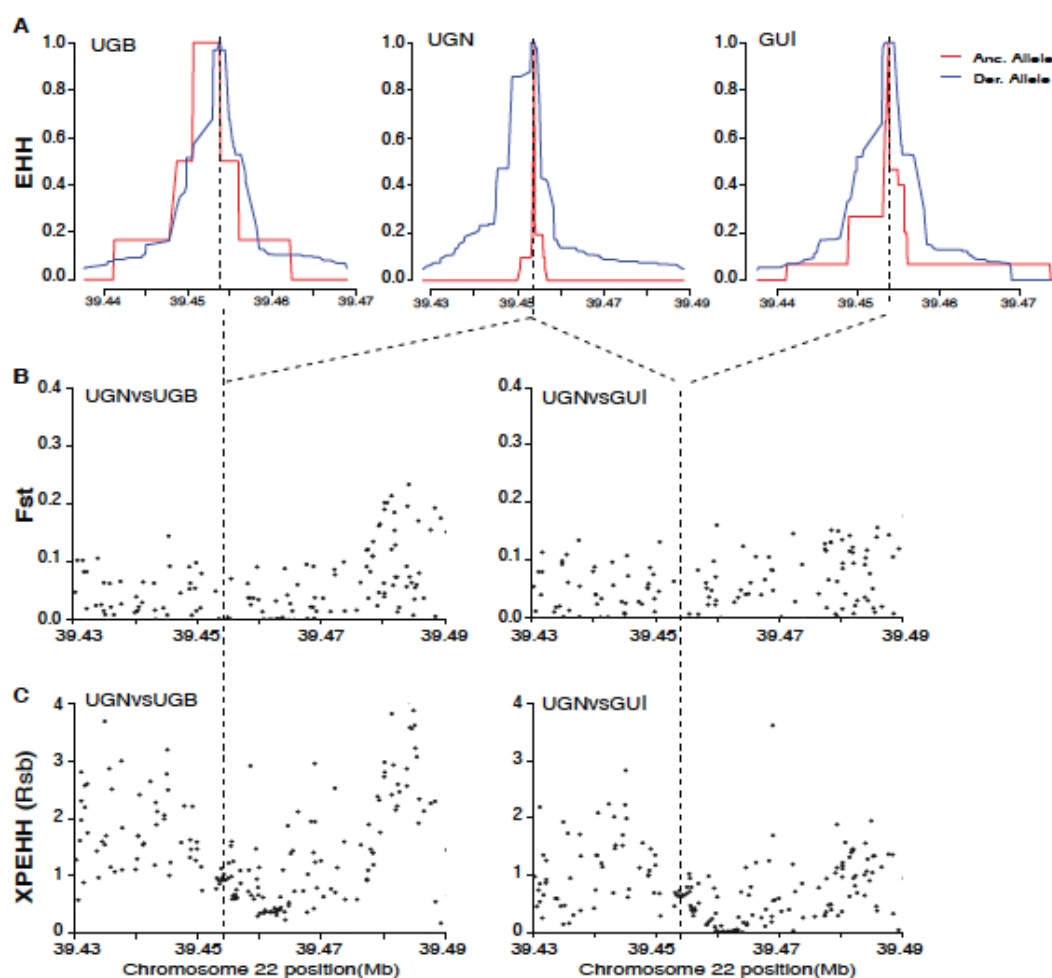
636

637

638

639

640



641

642

643 **Figure 4.** Illustration of signatures unique to the Uganda Nilotic population. Signal of  
644 positive selection within the *APOBEC3G* gene on Chromosome 22 at the  
645 rs112077004 loci of the Uganda Nilo-saharan Lugbara population, in comparison  
646 with the Niger-Congo B populations of Uganda (UGB) and Niger-Congo A  
647 population of Guinea (GUI). **A.** The calculated site specific extended haplotype  
648 homozygosity (EHH) within a population. **B.** Between population Fst analysis. **C.**  
649 Across population (XPEHH) analysis.

650

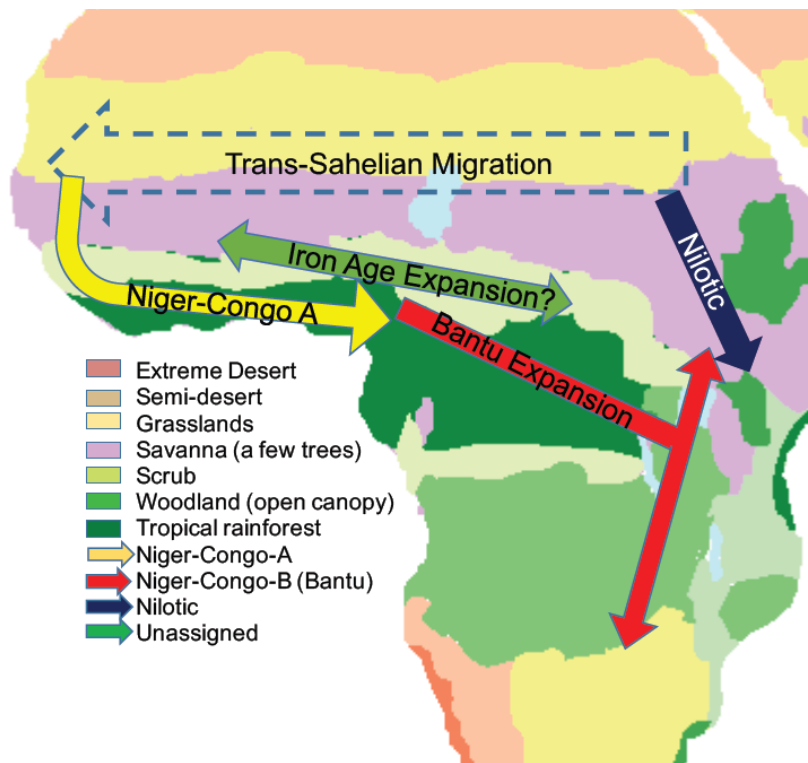
651

652

653

654

655



656

657 **Figure 5.** Migrations of Niger-Congo speakers. Map colours show vegetation  
658 coverage approximately 10kya (Adams, 1998). Colours for linguistic groups as for fig  
659 4. Blue Nilo-Saharan; Yellow, Niger-Congo A; Red, Niger-Congo-B (Bantu); Green  
660 putative expansion of an ancestral group out of modern Nigeria. Blue dotted arrow,  
661 suspected route of proto-Niger-Congo-A speakers from Nuba mountains of Sudan to  
662 Senegal (1966) when it was much wetter than at present.

663

664

665