Evidence of population specific selection inferred from 289 genome sequences of Nilo-Saharan and Niger-Congo linguistic groups in Africa

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

Abstract

Background

There are over 2000 genetically diverse ethno-linguistic groups in Africa that could help decipher human evolutionary history and the genetic basis of phenotypic variation. We have sequenced 300 genomes from Niger-Congo populations from six sub-Saharan African countries (Uganda, Democratic Republic of Congo, Cameroon, Zambia, Ivory Coast, Guinea) and a Nilo-Saharan population from Uganda. Of these, we analysed 289 samples for population structure, genetic admixture, population history and signatures of selection. These samples were collected as part of the TrypanoGEN consortium project [1].

Results

The population genetic structure of the 289 individuals revealed four clusters, which correlated with ethno-linguistic group and geographical latitude. These were the West African Niger-Congo A, Central African Niger-Congo B, East African Niger-Congo B and the Nilo-Saharan. We observed a spatial distribution of positive natural selection signatures in genes previously associated with AIDS, Tuberculosis, Malaria and Human African Trypanosomiasis among the TrypanoGEN samples. Having observed a marked difference between the Nilo-Saharan Lugbara and Niger-Congo

populations, we identified four genes (*APOBEC3G*, *TOP2B*, *CAPN9*, *LANCL2*), which are highly differentiated between the two ethnic groups and under positive selection in the Lugbara population (_iHS –log p > 3.0, Rsb –log p > 3.0, Fst > 0.1 bonferroni p > 1.8x10e4).

Conclusion

The signatures that differentiate ethnically distinct populations could provide information on the specific ecological adaptations with respect to disease history and susceptibility/ resistance. For instance in this study we identified *APOBEC3G* which is believed to be involved in the susceptibility of the Nilo-Saharan Lugbara population to Hepatitis B virus infection.

Keywords

Whole genome sequencing, Sub-Saharan African genomes, population genetic variation, signatures of selection

Background

The African continent's ethno-linguistic groups have been classified into four major families, Afro-Asiatic, Nilo-Saharan, Niger-Congo, and Khoisan [2]. The Afro-Asiatic which includes the Semitic, Cushitic, and ancient Egyptian languages, is spoken predominantly by northern and eastern African pastoralists and agro-pastoralists; the Nilo-Saharan, which includes the Central Sudanic and Eastern Sudanic (Nilotic) languages, is spoken predominantly by eastern and central Saharan pastoralists; the Niger-Congo languages are subdivided into the Niger-Congo A in West Africa and the Niger-Congo B or Bantu in Central, Eastern and Southern Africa [3, 4]. Fourteen ancestral population clusters, which correlate with shared cultural and

linguistic affiliations, have been identified amongst these groups [5]. These 14 ancestral populations further subdivide into over 2000 ethnically diverse linguistic groups [6, 7].

The genetic diversity of ethno-linguistic groups can be used to study human evolutionary history and the genetic basis of phenotypic variation [5] and complement studies of African genotype variations [5, 8-10] that have contributed to the understanding of human origins and disease susceptibility markers. However, samples from sufficient individuals for population analysis have been sequenced from relatively few African populations. The 1000 genome project generated data from five Niger-Congo populations, The African Variome project added Afro-Asiatic populations and there have been small scale studies of the Khoisan hunter-gatherers [5, 8, 11, 12]. To date, no sequences of Nilotic populations have been published, although one previous study used 200,000 SNP loci to examine genetic diversity of the Nilo-Saharan speaking population of southern Sudan Darfurian and Nuba peoples [13]. In the present study we present the first genome sequences of a Nilo-Saharan population and genome sequences from six new Niger-Congo populations.

Results

Samples and sequencing

The samples used for this study were collected by the TrypanoGEN consortium and consisted of 300 individuals from 17 linguistic groups who were residents of Guinea, Ivory Coast, Cameroon, Democratic Republic of Congo, Uganda and Zambia (Table 1, Additional file 1: Table_S0). DNA was extracted from blood and genomes were sequenced on the Illumina 2500 platform at 10X coverage, except for the Zambia and Cameroon samples that were sequenced at 30X coverage.

77 Following mapping and SNP calling, we identified approximately 34.1 million single
78 nucleotide polymorphisms (SNPs) and 5.3 million insertion/deletion (Indel)
79 polymorphisms (Table 2). We identified 2.02 million variants that did not have rsIDs
80 and were considered 'novel'. The SNPs had a transition-transversion ratio of 2.0
81 (Additional file 2: Figure S1), implying good quality SNP calls [14, 15]. Prior to
82 population analysis, variants (SNPs and Indels) were filtered by removing loci with
83 >10% missing data, Minor Allele Frequency (MAF) < 0.05 or Hardy Weinberg
84 Equilibrium (HWE) P-value < 0.01. Eleven individuals with > 10% loci missing were
85 removed from the original 300 individuals dataset (Table 2), leaving 289 samples for
86 downstream analysis. An additional 504 samples from five African populations in the
87 1000 genomes project (Esan and Yoruba from Nigeria, Mende from Sierra Leone,
88 Mandinka from Gambia and Luhya from Kenya), were included in some of our
89 analyses.
90
91 **Population stratification by Multiple Dimensional Scaling**
92 Multiple Dimensional Scaling (MDS) implemented in Plink 1.9 software was used to
93 help visualise genetic distances between samples (Figure 1). All TrypanoGEN
94 samples clustered by country except those in Uganda, where the Nilo-Saharan
95 Lugbara samples formed a distinct cluster from the Niger-Congo B Basoga samples.
96 When the samples from the six TrypanoGEN and the four African 1000 genomes
97 project countries were merged, five groups representing five major geographic groups
98 were observed (Figure 1B): the Uganda Nilo-Saharan; East African Bantu speakers
99 from Uganda and Kenya; Central African Bantu speakers from Cameroon, DRC and
100 Zambia; Nigerian Niger-Congo A speakers (Esan and Yoruba); West African Niger-
101 Congo A speakers from the Ivory Coast, Gambia, Sierra Leone and Guinea. The

African and European samples were very distinct (Figure 1C). Since all samples except Ugandan Bantu and Nilo-Saharan clustered by country by MDS, samples were grouped by country for subsequent analyses except for the Uganda samples, which were grouped by both country and linguistic group.

Population Admixture and differentiation

The amount of shared genetic ancestry within the samples was estimated using the Admixture software [16]. Admixture was run on 2-8 population clusters (K) in triplicate; with K=4, K=5 and K=6 having the lowest cross validation errors and hence the most probable numbers of ancestral components represented in the data (Additional file 2: Figure S2). At K=6 the Niger-Congo populations exhibited 17-60% admixture with minor ancestries, whilst the Ugandan Nilo-Saharan population had a platry 7% admixture with Niger-Congo ancestries (Figure 2A). At K4 one European and three ancestral African populations were observed, which corresponded to Nilo-Saharan, Niger-Congo-B (East African) and Niger-Congo-A (West African). At K5 a homogeneous group of seven samples emerged within the Zambia population with no admixture with other populations in our data set and were also outliers on the MDS plot (Figure 1B). These seven were recorded as Soli/Chikunda speakers, which are Bantu languages but they had no admixture at (K=5 and K=6) with the other speakers of this language group from Zambia or any other group included in this study, suggesting that they had a quite distinct ancestry. At K6, a major group appeared that contributed ancestral components to both East African Niger-Congo B and West African Niger-Congo A but did not correspond to any existing linguistic group.

Genetic variation within the populations that are part of the TrypanoGEN project was estimated using the pairwise F_{ST} [17] (Figure 2B, Additional file 2: Figure S3). F_{ST}

was relatively high between the Nilo-Saharan Lugbara samples and the African Bantu populations (Figure 2B) except the East African Basoga (population mean F_{ST} = 0.012) and Luhya (population mean F_{ST} = 0.011), presumably due to the 30% admixture of Nilo-Saharan origin within these populations. The pattern of the observed genetic variation was consistent with the relative geographic distance from the Nilo-Saharan population (Figure 2C). In addition, a classification based on the genetic distances between populations (F_{ST}) showed clustering of populations by geographic region on the African continent (Figure 2D).

Population size over time and timing of population isolation.

The Multiple Sequentially Markovian Coalescent (MSMC) method was used to estimate population sizes over time and times at which populations became isolated (Figure 3). Effective population sizes (N_e) were relatively stable at around 13,000 in all populations tested from 100 thousand years ago (kya) until about 50kya when they started to decline reaching a nadir of about 8,000 about 13kya coinciding with the dry period at the end of the last ice age (Figure 3A, Additional file 3: Table S1). All population sizes increased rapidly thereafter but the Niger-Congo populations increased to an N_e of around 200,000, whilst the Nilotic population only increased to 60,000. The Ugandan Bantu population was intermediate in N_e presumably due to admixture with the Nilotics. This post glacial population increase was briefly reversed in the Central and West African populations which suffered declines of 6-23% between 1500 and 750 years ago before recovering to even higher levels at the present time. This absolute decline in N_e was not observed in the Ugandan Bantu population, although the growth rate declined. In the Nilotic population, a decline was observed at a later time point after 750 years ago.

152 Population separation data is less clear and may be more sensitive to admixture
153 (Figure 3B). The Guinea and Ivory Coast populations were the least admixed and
154 appeared panmictic until about 10kya, and had become isolated by about 3kya. The
155 Ugandan Bantu and Ugandan Nilotic appeared to begin separating from other
156 populations about 23 and 47kya, respectively and became isolated about 3kya but
157 these estimates may be confounded by admixture.
158
159 **Genome-wide screen for extended haplotypes under selection**
160 ***Signatures within population***
161 In order to identify alleles under selection pressure, we used the within population
162 Extended Haplotype Homozygosity (EHH) test [2, 18]. Similar patterns of loci with
163 extreme positive and negative iHS scores were observed across all groups (Additional
164 file 4: Figure S4A). The iHS values for all groups had an approximate normal
165 distribution (Additional file 4: Figure S4C) implying that the sizes of iHS signals
166 from different SNPs in all the populations were comparable [3, 4, 19]. The mean
167 number of loci with extreme positive and negative iHS score (–log p > 3) from all
168 groups was 8,984, Guinea had the largest number of loci with extreme iHS score
169 (11,401) and Zambia had the least (5,570) (Table 3, Additional file 5: Table S2).
170 These extreme loci were classified by the Ensembl annotation of the nearest gene.
171 Approximately 34% of these annotations were for protein coding genes; a mean of
172 3,058 SNPs in protein coding genes per population were associated with extreme iHS
173 scores. Some protein coding genes with extreme iHS SNP loci were shared between
174 different countries whereas some occurred only in a single country population
175 (Additional file 5: Table S2, sheet 'ALLpop.protein_coding'). We observed strong
176 iHS signatures in genes that have been previously identified in other African

177 populations as being under strong selection [5, 8, 19, 20]. These included ***SYT1,*** a
178 synaptosomal protein implicated in Alzheimer's disease [6, 7, 21] that was found in
179 all country populations; ***LARGE***, a glycosylase involved in Lassa fever virus binding
180 [5, 22] (Zambia, Cameroon, Ivory Coast); ***CDK5RAP2***, a microcephaly gene
181 controlling brain size [5, 8-10, 23] (Ugandan Bantu); ***NCOA1***, a transcriptional co-
182 activator associated with Lymphoma (Guinea, Ivory Coast, DRC); ***SIGLEC12***;
183 involved in immune responses [5, 8, 11, 12, 24] (Zambia, Cameroon). Using the
184 DAVID annotation database [13, 25] we observed that all of the country populations
185 had strong signals that have been implicated in communicable diseases such as
186 HIV/AIDS, Malaria and Tuberculosis which have high prevalence in Africa [14, 15,
187 26] (Table 4), suggesting an adaptive role of these genes to infection.
188 Having collected samples from Human African Trypanosomiasis (HAT) endemic
189 regions, we identified signatures that have been implicated in trypanosome infection.
190 These signatures were observed in genes overlapping the KEGG calcium signalling
191 pathway [16, 27, 28] that were previously identified mainly from mouse studies [29];
192 ***F2RL1*** (Guinea, Ivory Coast), ***GNA14*** (Zambia), ***GNAQ*** (Cameroon), ***GNAL***
193 (Guinea, Cameroon), ***GNAS*** (Zambia). The calcium-signalling pathway regulates
194 permeability of the blood brain barrier to trypanosome parasites during central
195 nervous system (CNS) disease [30]. In addition, we observed signatures in genes
196 overlapping the Mitogen-activated protein kinase MAPK pathway ***MAPK1***
197 (Cameroon), ***MAPK10*** (Ugandan Nilo-Saharan, DRC, Ugandan Bantu), ***MAPK9***
198 (Zambia); which is targeted by trypanosomatids in order to modulate the host's
199 immune response [17, 31]. These host signalling pathways have been shown to play a
200 role in host immunity against trypanosome infection in mice and cattle [32-34].
201

202 ***Signatures unique to Nilo-Saharans***

203 In order to determine which signatures are unique to the Nilo-Saharan Lugbara, we
204 first ascertained which extreme iHS loci (-log p > 3) were common to the Nilo-
205 Saharan and one or more Niger-Congo groups. We observed that approximately 15%
206 of the protein coding gene associated extreme iHS SNPs of the Ugandan Bantu, DRC,
207 Ivory Coast and Guinea groups were common with the Nilo-Saharan group, whereas
208 Cameroon and Zambian groups had 2.7% in common (Table 3, Additional file 4:
209 Figure S4B). We identified 149 extreme SNPs associated with protein coding genes
210 and were unique to the Uganda Nilo-Saharan (Additional file 6: Table S3). Using the
211 PANTHER Gene ontology database [35], we observed that these unique genes were
212 mainly enriched for cellular and metabolic process proteins (approximately 50.8%)
213 (Additional file 7: Figure S5). Amongst these were SNPs associated with genes that
214 have also been shown by other studies to be under positive selection, including
215 ***APOBEC3G,*** which is involved in innate anti-viral immunity [36, 37], with
216 demonstrated protective alleles against HIV-1 in Biaka and Mbuti pygmies of Central
217 African Republic and DRC respectively [38]; ***IFIH1*** (also called ***MDA5***), a
218 cytoplasmic RNA receptor that mediates antiviral responses by activating type I
219 interferon signalling [39] but is also implicated in protection against type 1 diabetes
220 ([40, 41]; ***OR2L13,*** an olfactory receptor involved in activation of signal transduction
221 pathway for odorant recognition and discrimination [42], and is associated with
222 Diabetic nephropathy in African Americans [43].

223

224 ***Nilo-Saharan versus Niger-Congo cross population signatures***

225 There were 299 SNPs with high F_{ST} (above 99th percentile) and XPEHH (Rsb –log p
226 > 3) in the regions of protein coding genes that were also highly differentiated

227 between the Nilo-Saharan and Niger-Congo populations (Additional file 8: Table S4).
228 We then compared SNP loci with derived alleles that are unique to the Nilo-Saharan
229 Lugbara and occur in highly differentiated genes (extreme Rsb, high Fst) between the
230 Nilo-Saharan and Niger-Congo groups. From this we identified 12 genes (Table 5,
231 Additional file 9: Figure S6D) including the ***APOBEC3G*** gene, that are highly
232 differentiated between the Nilo-Saharan Lugbara and Niger-Congo groups (mean F_{ST}
233 0.11, Rsb –log p = 4.1). *APOBEC3G* also contains the SNP rs112077004, which was
234 observed to be under positive selection in the Nilo-Saharan Lugbara (Figure 4,
235 Additional file 10: Figure S7, Additional file 11: Table S5).
236
237 **Discussion**
238 We have analysed the genomes of 289 individuals from seven populations from six
239 Sub-Saharan Africa Countries, investigating their admixture profile, demographic
240 histories and signatures of selection that differentiate the major linguistic groups. The
241 MDS analysis identified five major clusters: Nilo-Saharan, two Niger-Congo A
242 groups from Nigeria and West Africa and two groups of Niger-Congo B (Bantu
243 speakers) from Central and East Africa, which were consistent with previous studies
244 [5, 8, 44]. The samples represented three of the five major linguistic groups in Africa,
245 omitting the Afro-Asiatic and Khoisan speakers. Afro-Asiatic speakers are found
246 across North and North-East Africa in regions adjacent to Nilo-Sharan and Bantu
247 speakers. Afro-Asiatic reference populations were not included in this study and we
248 therefore could not detect any admixture from this source. A SNP genotype based
249 analysis of Nilotic populations indicated that Nilotic populations only contain a trace
250 of Afro-Asiatic ancestry and therefore our observations on East African populations
251 may not be significantly limited by the absence of Afro-Asiatic data [13].

Linguistic analyses suggest that Niger-Congo speaking hunter-gathers originated from the Kordofanian speakers of the Nuba mountains of Sudan and then traversed the Sahel to Mali (Figure 5). They colonised the coast from Senegal to Nigeria and Cameroon over several thousand years, forming multiple linguistic groups. The Bantu (Niger-Congo-B) speaking people emerged as another linguistic group amongst the greater than 60 Niger-Congo-A groups in the Nigeria/Cameroon region about 3,000 years ago. Bantu speaking peoples then spread South-East along savannah corridors through the Congo basin and emerged in the Great Lakes region and spread North to the Lake Victoria region and South down the East Side of Africa[10, 45, 46]. This rapid expansion is believed to have been enabled by the development of agriculture and later enhanced by the acquisition of iron tools [5].

Our admixture analysis at K=4 was consistent with this linguistic history and recent genetic analyses [8, 10] with three African ancestral allele clusters (AAC) which can be interpreted as representing Niger-Congo A languages in West Africa, Niger-Congo B (Bantu) in Central and East Africa and Nilo-Saharan in Northern Uganda. The Niger-Congo-A speakers in extreme West Africa appear to have approximately 10% Nilo-Saharan ancestry and this declines towards the East. On the other hand, the Bantu speakers are a mix of Niger-Congo-A and a distinct putative Bantu ancestral cluster that it at highest frequency in Nigeria and Cameroon, despite the Nigerian Yoruba and Esan not being Bantu languages. The Niger-Congo-A component is displaced by a Nilotic component with easterly latitude whilst the “Bantu” component remains constant. At K=5 a small AAC of 7 Bantu speakers from Zambia emerges, who evidently have a genetic heritage that does not match their self-declared linguistic affiliation, and are of unknown descent. At K=6 a fourth major African AAC appears (green in Figure 2) with strongest representation in the Nigerian Yoruba

277 and Esan then tapering off east and west into Central and West Africa. This does not
278 correspond to any linguistic group and displaces the Niger-Congo-A ancestry
279 observed at K=4 and 5, to the east of Nigeria and Niger-Congo-B (Bantu) in Nigeria
280 and to the West. This ancestral cluster could represent a secondary movement out of
281 Nigeria of migrants who adopted their host's language. One possible driver for such a
282 migration, if it occurred, was the development of iron smelting which may have
283 originated in Nigeria about 2,500 years ago [47]. Irrespective of the true number of
284 ancestral allele clusters there is evidence of back migration of people with Bantu
285 ancestral alleles into West Africa as has been observed before [44]. Interestingly this
286 migration to the west was not accompanied by language expansion as it was to the
287 east.
288 Regarding population history, the estimates of current N_e obtained from our data with
289 MCMS (Figure 3A) of around 200,000 in West and Central Africa and 57,000-
290 125,000 in East Africa (Supplementary Table S6) was consistent with previous
291 observations on other African samples using the same method [48] [12], but our larger
292 dataset provides higher resolution at recent time points than a previous analysis [12].
293 The faster growth in the Niger-Congo A and B than the Nilotic populations appears to
294 predate the Bantu expansion. The Niger-Congo A population was believed to be
295 expanding through West Africa as the climate became wetter after 10kya, consistent
296 with the separation times between the Guinea and Ivory Coast populations observed
297 on the Cross-Coalescence Plot (Figure 3B). The Nilo-Saharan population developed a
298 pastoralist economy probably after 6kya but their expansion into the tsetse belt may
299 have been delayed by trypanosomiasis and other diseases until the cattle developed
300 tolerance [49-51]. The effective population size did not grow as fast as that of the
301 Niger-Congo-A populations. The brief population decline dated at ~1340CE by

302 MSMC coincides with the timing of the Black Death (1343-1353), however time
303 resolution is low and the decrease was only observed at a single time point. There is
304 evidence of abandonment of multiple large settlements throughout West Africa
305 around the time of the Black Death and there is speculation that this was caused by
306 the plague [52]. The decrease at this time appears to have impacted the West and
307 Central African Niger-Congo but not the East African populations. Both Bantu and
308 Nilotic populations in East Africa were cattle keepers and pastoralists to varying
309 degrees [51] and the concomitant lower population density and mobile lifestyle may
310 have made them less vulnerable to plague than the more settled and urbanised West
311 Africans. The more recent decline in the Nilotic Lugbara effective population size is
312 unexplained, but the catastrophic Rinderpest outbreak in the 1880's and 1890's that
313 killed up 90% of indigenous cattle, which lead to the depopulation of the East African
314 savannahs and may have ended the dominance of the Nilotic speaking Maasai over
315 the Bantu Kikuyu could have been a contributory factor [53].
316 The Cross-Coalescence plots for comparison between populations other than the
317 Guinea and Ivory Coast Niger-Congo-A show long periods of separation (not shown).
318 This is not consistent with previous observations [12] or with the Ugandan Bantu
319 populations having separated from Niger-Congo-A populations even more recently
320 than the separation between Guinea and Ivory Coast populations, and is presumably
321 due to the extensive admixture with the Nilotics observed in this population. The
322 Central African cross-coalescence data also indicated older separation times than
323 linguistic evidence suggests (not shown) and although there was less evidence of
324 admixture in this population, these data should be treated with caution.
325 We identified evidence of selection in genes that have previously been associated
326 with HIV/AIDS, Tuberculosis and Malaria. Given the high prevalence of lethal

327 infections on the continent [26], the finding of positive natural selection at disease
328 susceptibility loci is not surprising. However not all these genes occurred in all the
329 populations, demonstrating spatially varying selection probably due to differing
330 environmental pressures [54, 55]. We looked for signatures of selection in genes and
331 pathways that are implicated in trypanosome infection, including the calcium
332 signalling pathway [29, 30], the MAPK pathway [31-34], and *HPR*, *APOL1, IL6,*
333 *HLAG,* genes [56-61] (Additional file 12: Figure S8, Additional file 13: Table S6).
334 However we only found evidence for selection for the calcium signalling and MAPK
335 pathway genes. This suggests that Trypanosomiasis may have had a selective force in
336 these populations.
337 In order to determine signatures of selection unique to the Nilo-Saharan Lugbara
338 population, we used a combination of linkage disequilibrium-based method (iHS and
339 Rsb) and population differentiation based method (F_{ST}) [62]. Using this approach we
340 identified 12 loci associated with coding genes, which are unique to the Nilo-Saharan
341 Lugbara and highly differentiated from the Niger-Congo population. Among these
342 was the variant associated with *APOBEC3G* that demonstrated significant positive
343 selection in the Nilo-Saharan Lugbara population. This protein is involved in viral
344 innate immunity [63], by inducing a high rate of dC to dU mutations in the nascent
345 reverse transcripts leading to the degradation of the viral genome [36, 37]. The
346 Lugbara have relatively low prevalence of HIV (4%) in comparison to the Basoga
347 (6.4%) and Baganda (10.7%) Bantu groups of Uganda but relatively high prevalence
348 of Hepattis B suggesting that either *APOBEC3G* has different effects on each of these
349 viruses [64-69].
350 We also identified the missense variant rs10930046 (T/C) located in the *IFIH1* CDS,
351 which was unique to the Nilo-Saharan Lugbara and highly selected (iHS –log p

352 3.264). This gene is associated with up regulation of type I interferon signalling
353 occurring in a spectrum of human diseases [39] and is believed to be involved in the
354 suppression of Hepatitis B viral replication [70]. Being a non-synonymous variant,
355 rs10930046 could alter the functioning of IFIH1 and thus increase susceptibility to
356 HBV in the Lugbara population, something that could be tested by a candidate gene
357 study for DNA virus infections. Northern Uganda is considered to have one of the
358 highest prevalence of Hepatitis B virus in the world [71] which has perhaps resulted
359 in a unique adaption of the Nilo-Saharan Lugbara population to infection.
360

361 **Conclusion**

362 We have incorporated a Nilo-Saharan population into the analysis of genomic
363 sequences of Niger-Congo populations for the first time and show extensive
364 admixture between Nilo-Saharan ancestry and Niger-Congo B (Bantu) populations.
365 We show evidence for signatures of selection within the Nilo-Saharan population in
366 genes associated with infectious diseases that have different prevalence from
367 surrounding Bantu (Niger-Congo B) populations.
368

369 **Methods**

370 **Sample collection**

371 The samples used for this study are part of the TrypanoGEN biobank [72], which
372 describes ethics approval, recruitment, sample processing and the meta data collected.
373 The ethical approval for the study was provided by the national ethics councils of the
374 TrypanoGEN consortium countries involved in the sample collection.
375 The sampled populations were assigned a 3 letter code with the following criteria:
376 Country name first letter, was used as the first letter of the code; For countries where

377 more than one linguistic group was sampled, we used abbreviations of the country
378 name, ZAM-Zambia, CAM-Cameroon, CIV-Ivory Coast, DRC-Democratic republic
379 of Congo; For countries where a homogeneous linguistic group was sampled, we used
380 the Country, language group and language, UNL-Uganda Nilotic Lugbara, UBB-
381 Uganda, Niger-congo B (Bantu), Basoga, GAS-Guinea, Niger-Congo A, Soussou.
382 Peripheral blood was collected from the participants at the field sites, frozen, and
383 transported to reference laboratories from where DNA extraction was carried out
384 using the Whole blood MidiKit (Qiagen). The DNA was quantified using the Qubit
385 (Qiagen) and approximately 1μg was shipped from each country for sequencing at the
386 University of Liverpool, UK except for Cameroon and Zambia whose DNA was
387 shipped to Baylor College, USA.
388
389 **Sequencing and SNP calling**
390 The whole genome sequencing libraries were prepared using the Illumina Truseq
391 PCR-free kit and sequenced on the Illumina Hiseq2500. The samples from Guinea,
392 Cote D'Ivoire, Uganda and DRC were sequenced to 10x coverage at the Center for
393 Genomic Research at the University of Liverpool. The samples from Zambia and
394 Cameroon were sequenced to 30X at the Baylor College of Medicine Human Genome
395 Sequencing Center.
396 The sequenced reads were mapped onto the 1000 genomes project
397 human_g1k_v37_decoy reference genome using BWA. The SNP calling on all the
398 samples was done using the genome analysis tool kit GATK v3.4. The SNPs were
399 then filtered by; a) removing loci with > 10% missing SNP, b) removing individuals
400 with > 10% missing SNP loci and c) removing loci with Hardy Weinberg P value <
401 0.01. In addition, loci with MAF < 0.05 were also removed for the population

402 stratification and Admixture analysis. Variant annotation was done using snpEff
403 (www.snpeff.sourceforge.net).
404
405 **MDS analysis**
406 Population stratification was done using Multi dimensional scaling tool in Plink 1.9
407 and R v 3.2.1 tools. For this analysis, in addition to the filtering mentioned above,
408 SNP loci less than 2000bp apart were removed in order to reduce the linkage
409 disequilibrium (LD) between adjacent SNP. MDS analysis was carried out for (i) all
410 TrypanoGEN data , (ii) all TrypanoGEN data plus African 1000 genome data, (iii) all
411 TrypanoGEN data including 50 European and all African 1000 genome data
412 excluding African Caribbean in Barbados (ACB) and African Southwest USA (ASW)
413 populations.
414
415 **Population Admixture**
416 The population ancestry of each individual was obtained using Admixture 1.23 [16]
417 on the filtered PLINK .bed files on the same TrypanoGEN, one thousand genome
418 African and European population data sets analysed by MDS. Admixture was run on
419 K1 to K8 with three replicates for each run. The Admixture plots were drawn using
420 the R tool 'strplot' [73].
421
422 **Genetic diversity: F_{ST}**
423 The genetic diversity due to difference in allele frequency among populations was
424 analysed by the inter-population Wright's F_{ST} [17] in PLINKv1.9. The F_{ST} estimates
425 were made between TrypanoGEN (UNL, UBB, DRC, CIV, GAS) and one thousand
426 genome African (LWK, YRI, ESN, MSL, GWD) populations. The F_{ST} dendrogram

427 was generated using Fitch in Phylip3.685 [74]. The geographic distance matrix
428 between populations was calculated based on their global position system (GPS)
429 coordinates [75].
430
431 **Population History.**
432 Population sizes and divergence times were calculated using MSMC [48] on data
433 phased with Beagle [76]. Since PCA and Admixture analysis had indicated little
434 difference between linguistic groups in each country with the exception of the
435 Ugandan Bantu and Nilotic populations, samples from each country with highest
436 coverage were analysed together except for Uganda where Bantu and Nilotic samples
437 were analysed as separate populations. For population size estimates, output from 3
438 independent runs each using 8 different haplotypes were combined. Using 8
439 haplotypes rather than 4 gives higher resolution at more recent time points. For
440 estimates of relative cross coalescence rate, three replicate runs were done, each using
441 2 different samples (4 haplotypes) from each pairwise comparison between
442 populations. Results presented are the means of the replicates.
443
444 **Signatures of selection**
445 The estimation of haplotypes was carried out by first Phasing of the genotyped SNPs
446 using SHAPEIT v2.2 software [77]. The extended haplotype homozygosity (EHH)
447 was then analysed using the R software package *rehh* [78]. Two main EHH derived
448 statistics were calculated from the phased haplotype data, that is, intra-population
449 integrated haplotype Score (iHS) [19] and inter-population Rsb [79]. Bedtools v2.26.0
450 was used to identify the intersection of the iHS, Fst and Rsb loci.
451

Declarations

Ethical approval

The ethical approval for the study was provided by the national ethics councils of the TrypanoGEN consortium countries involved in the sample collection which are: Uganda (HS 1344), Zambia (011-09-13), Democratic Republic of Congo (No 1/2013), Cameroon (2013/364/L/CNERSH/SP), Cote d'Ivoˆıre (2014/No 38/MSLS/CNER-dkn), and Guinea (1-22/04/2013). All the participants in the study were guided through the consent forms, and written consent was obtained to collect biological specimens.

Consent of publication

Informed consent was provided by the study participants, for sharing and publishing their anonymised data.

Availability of data and material

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request. The sequenced data will be submitted to the EGA by H3ABionet under the study accession number EGAS00001002602.

Competing interests

The authors declare that they have no competing interests.

Funding

The study was under the TrypanoGEN project, funded by the Wellcome Trust (099310/Z/12/Z). The funders had no roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Author's contributions

EM, AM, IS, BB, CHZ conceived the study; HI, MK, DM, GS, JE, JC, MS, PA, collected samples; JM and HN analysed the data, interpreted output and wrote the manuscript; All authors read and approved the final manuscript.

Acknowledgements

The authors would like to acknowledge the study participants who donated their specimens, the personnel involved in the community engagement and coordinating sample collection and processing, the National sleeping sickness control programmes of the participating Countries. Z Lombard (University of Witwatersrand) and D Adeyemo (NHGRI) for facilitating sequencing of samples from Zambia and Cameroon at Baylor College of Medicine. The H3ABionet for training and support on data analysis. Also Fiona Marshall and Rebecca Grollemund for helpful discussions of African History.

Authors' Affiliations

1. Makerere University, Kampala, Uganda
2. Centre for Genomic Research, University of Liverpool, UK
3. Centre International de Recherche-Développement sur l'Elevage en zone Subhumide (CIRDES), Bobo-Dioulasso, Burkina Faso
4. Université Jean Lorougnon Guédé (UJLoG), Daloa, Côte d'Ivoire

5. Institut National de Recherche Biomedicale, Kinshasa, Democratic Republic of Congo
6. Faculty of Science, University of Dschang, Cameroon
7. University of Malawi, College of Medicine, Department of Basic Medical Sciences, Blantyre, Malawi
8. Department of Disease Control, School of Veterinary Medicine, University of Zambia, Lusaka, Zambia
9. Institut de Recherche pour le Développement (IRD), IRD-CIRAD 177, Montpellier, France
10. Wellcome Center for Molecular Parasitology, University Place, Glasgow, UK
11. Programme National de Lutte contre la Trypanosomose Humaine Africaine, Conakry, Guinea

References:

1. *TRYPANOGEN: an H3 Africa Funded Project* [http://www.trypanogen.net]

2. Blench R: *Archaeology, Language, and the African Past*. Rowman Altamira; 2006.

3. Greenberg J: *Greenberg: the Languages of Africa (International... - Google Scholar*. Bloomington; 1963.

4. Lewis MP, Simons GF, Fennig CD: *Ethnologue: Languages of the World, 16th Edition | SIL International*. 2009.

5. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM: **The genetic structure and history of Africans and African Americans.** *Science* 2009, **324**:1035–1044.

6. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo J-M, Wambebe C, Tishkoff SA, Bustamante CD: **Genome-wide patterns of population structure and admixture in West Africans and African Americans.** *Proc Natl Acad Sci USA* 2010, **107**:786–791.

530 7. Tishkoff SA, Williams SM: **Genetic analysis of African populations: human**
531 **evolution and complex disease.** *Nat Rev Genet* 2002, **3**:611–621.

532 8. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I,
533 Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, Ritchie GRS, Xue
534 Y, Asimit J, Nsubuga RN, Young EH, Pomilla C, Kivinen K, Rockett K, Kamali A,
535 Doumatey AP, Asiki G, Seeley J, Sisay-Joof F, Jallow M, Tollman S, Mekonnen E,
536 Ekong R, Oljira T, Bradman N, Bojang K, et al.: **The African Genome Variation**
537 **Project shapes medical genetics in Africa**. *Nature* 2014, **517**:327–332.

538 9. Busby G, Band G, Si Le Q, Jallow M, Bougama E, Mangano V, Amenga-Etego L,
539 Emil A, Apinjoh T, Ndila C, Manjurano A, Nyirongo V, Doumbo O, Rockett K,
540 Kwiatkowski D, Spencer C, The Malaria Genomic Epidemiology Network:
541 *Admixture Into and Within Sub-Saharan Africa*. 2016.

542 10. Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, Laval G, Perry
543 GH, Barreiro LB, Froment A, Heyer E, Massougbodji A, Fortes-Lima C, Migot-
544 Nabias F, Bellis G, Dugoujon J-M, Pereira JB, Fernandes V, Pereira L, Van der Veen
545 L, Mouguiama-Daouda P, Bustamante CD, Hombert J-M, Quintana-Murci L:
546 **Dispersals and genetic adaptation of Bantu-speaking populations in Africa and**
547 **North America.** *Science* 2017, **356**:543–546.

548 11. Ratan A, Perry GH, Montenegro A, Miller W, Kim HL, Schuster SC: **Khoisan**
549 **hunter-gatherers have been the largest population throughout most of modern-**
550 **human demographic history**. *Nat Commun* 2014, **5**:1–8.

551 12. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M,
552 Chennagiri N, Nordenfelt S, Tandon A, Skoglund P, Lazaridis I, Sankararaman S, Fu
553 Q, Rohland N, Renaud G, Erlich Y, Willems T, Gallo C, Spence JP, Song YS, Poletti
554 G, Balloux F, van Driem G, de Knijff P, Romero IG, Jha AR, Behar DM, Bravi CM,
555 Capelli C, et al.: **The Simons Genome Diversity Project: 300 genomes from 142**
556 **diverse populations.** *Nature* 2016, **538**:201–206.

557 13. Dobon B, Hassan HY, Laayouni H, Luisi P, Ricaño-Ponce I, Zhernakova A,
558 Wijmenga C, Tahir H, Comas D, Netea MG, Bertranpetit J: **The genetics of East**
559 **African populations: a Nilo-Saharan component in the African genetic**
560 **landscape.** *Sci Rep* 2015, **5**:9996.

561 14. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C,
562 Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ,
563 Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: **A**
564 **framework for variation discovery and genotyping using next-generation DNA**
565 **sequencing data.** *Nat Genet* 2011, **43**:491–498.

566 15. Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y: **Genome measures used for**
567 **quality control are dependent on gene function and ancestry.** *Bioinformatics*
568 2015, **31**:318–323.

569 16. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry**
570 **in unrelated individuals**. *Genome Res* 2009.

571 17. Wright S: **The Genetical Structure Of Populations**. *Annals of Human Genetics*
572 1949, **15**:323–354.

573 18. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel
574 SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler
575 D, Cooper R, Kwiatkowski D, Ward R, Lander ES: **Detecting recent positive**
576 **selection in the human genome from haplotype structure.** *Nature* 2002, **419**:832–
577 837.

578 19. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A Map of Recent Positive**
579 **Selection in the Human Genome**. *PLoS Biol* 2006, **4**:e72.

580 20. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne
581 EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Frazer KA, Ballinger DG,
582 Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P,
583 Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H,
584 et al.: **Genome-wide detection and characterization of positive selection in human**
585 **populations**. *Nature* 2007, **449**:913–918.

586 21. Yoo BC, Cairns N, Fountoulakis M, Lubec G: **Synaptosomal proteins, beta-**
587 **soluble N-ethylmaleimide-sensitive factor attachment protein (beta-SNAP),**
588 **gamma-SNAP and synaptotagmin I in brain of patients with Down syndrome**
589 **and Alzheimer's disease.** *Dement Geriatr Cogn Disord* 2001, **12**:219–225.

590 22. Andersen KG, Shylakhter I, Tabrizi S, Grossman SR, Happi CT, Sabeti PC:
591 **Genome-wide scans provide evidence for positive selection of genes implicated in**
592 **Lassa fever**. *Philos Trans R Soc Lond, B, Biol Sci* 2012, **367**:868–877.

593 23. Bond J, Roberts E, Springell K, Lizarraga SB, Lizarraga S, Scott S, Higgins J,
594 Hampshire DJ, Morrison EE, Leal GF, Silva EO, Costa SMR, Baralle D, Raponi M,
595 Karbani G, Rashid Y, Jafri H, Bennett C, Corry P, Walsh CA, Woods CG: **A**
596 **centrosomal mechanism involving CDK5RAP2 and CENPJ controls brain size.**
597 *Nat Genet* 2005, **37**:353–355.

598 24. Crocker PR, Paulson JC, Varki A: **Siglecs and their roles in the immune system**
599 **- ProQuest**. ... *reviews Immunology* 2007.

600 25. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of**
601 **large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4**:44–57.

602 26. Bhutta ZA, Sommerfeld J, Lassi ZS, Salam RA, Das JK: **Global burden,**
603 **distribution, and interventions for infectious diseases of poverty.** *Infect Dis*
604 *Poverty* 2014, **3**:21.

605 27. *KEGG: Kyoto Encyclopedia of Genes and Genomes* [http://www.kegg.jp/kegg/]

606 28. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M: **KEGG as a**
607 **reference resource for gene and protein annotation.** *Nucleic Acids Res* 2016,
608 **44**:D457–62.

609 29. Grab DJ, Garcia-Garcia JC, Nikolskaia OV, Kim YV, Brown A, Pardo CA, Zhang
610 Y, Becker KG, Wilson BA, de A Lima APC, Scharfstein J, Dumler JS: **Protease**

611 **activated receptor signaling is required for African trypanosome traversal of**
612 **human brain microvascular endothelial cells.** *PLoS Negl Trop Dis* 2009, **3**:e479.

613 30. Nikolskaia OV, de A Lima APC, Kim YV, Lonsdale-Eccles JD, Fukuma T,
614 Scharfstein J, Grab DJ: **Blood-brain barrier traversal by African trypanosomes**
615 **requires calcium signaling induced by parasite cysteine protease.** *J Clin Invest*
616 2006, **116**:2739–2747.

617 31. Soares-Silva M, Diniz FF, Gomes GN, Bahia D: **The Mitogen-Activated Protein**
618 **Kinase (MAPK) Pathway: Role in Immune Evasion by Trypanosomatids.** *Front*
619 *Microbiol* 2016, **7**:183.

620 32. Mansfield JM, Paulnock DM: **Regulation of innate and acquired immunity in**
621 **African trypanosomiasis.** *Parasite Immunol* 2005, **27**:361–371.

622 33. Noyes H, Brass A, Obara I, Anderson S, Archibald AL, Bradley DG, Fisher P,
623 Freeman A, Gibson J, Gicheru M, Hall L, Hanotte O, Hulme H, McKeever D, Murray
624 C, Oh SJ, Tate C, Smith K, Tapio M, Wambugu J, Williams DJ, Agaba M, Kemp SJ:
625 **Genetic and expression analysis of cattle identifies candidate genes in pathways**
626 **responding to Trypanosoma congolense infection.** *Proc Natl Acad Sci USA* 2011,
627 **108**:9304–9309.

628 34. Kuriakose SM, Singh R, Uzonna JE: **Host Intracellular Signaling Events and**
629 **Pro-inflammatory Cytokine Production in African Trypanosomiasis.** *Front*
630 *Immunol* 2016, **7**:181.

631 35. Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, Ladunga I,
632 Ulitsky-Lazareva B, Muruganujan A, Rabkin S, Vandergriff JA, Doremieux O:
633 **PANTHER: a browsable database of gene products organized by biological**
634 **function, using curated protein family and subfamily classification.** *Nucleic Acids*
635 *Res* 2003, **31**:334–341.

636 36. Sawyer SL, Emerman M, Malik HS: **Ancient adaptive evolution of the primate**
637 **antiviral DNA-editing enzyme APOBEC3G.** *PLoS Biol* 2004, **2**:E275.

638 37. Zhang J, Webb DM: **Rapid evolution of primate antiviral enzyme**
639 **APOBEC3G.** *Hum Mol Genet* 2004, **13**:1785–1791.

640 38. Zhao K, Ishida Y, Oleksyk TK, Winkler CA, Roca AL: **Evidence for selection at**
641 **HIV host susceptibility genes in a West Central African human population.** *BMC*
642 *Evol Biol* 2012, **12**:237.

643 39. Rice GI, del Toro Duany Y, Jenkinson EM, Forte GMA, Anderson BH, Ariaudo
644 G, Bader-Meunier B, Baildam EM, Battini R, Beresford MW, Casarano M,
645 Chouchane M, Cimaz R, Collins AE, Cordeiro NJV, Dale RC, Davidson JE, De
646 Waele L, Desguerre I, Faivre L, Fazzi E, Isidor B, Lagae L, Latchman AR, Lebon P,
647 Li C, Livingston JH, Lourenço CM, Mancardi MM, Masurel-Paulet A, et al.: **Gain-**
648 **of-function mutations in IFIH1 cause a spectrum of human disease phenotypes**
649 **associated with upregulated type I interferon signaling.** *Nat Genet* 2014, **46**:503–
650 509.

651 40. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA: **Rare variants of IFIH1,**

652 **a gene implicated in antiviral responses, protect against type 1 diabetes.** *Science*
653 2009, **324**:387–389.

654 41. Fumagalli M, Cagliani R, Riva S, Pozzoli U, Biasin M, Piacentini L, Comi GP,
655 Bresolin N, Clerici M, Sironi M: **Population genetics of IFIH1: ancient population**
656 **structure, local selection, and implications for susceptibility to type 1 diabetes.**
657 *Molecular Biology and Evolution* 2010, **27**:2555–2566.

658 42. Sharon D, Glusman G, Pilpel Y, Khen M, Gruetzner F, Haaf T, Lancet D:
659 **Primate evolution of an olfactory receptor cluster: diversification by gene**
660 **conversion and recent emergence of pseudogenes.** *Genomics* 1999, **61**:24–36.

661 43. Bailey JNC, Palmer ND, Ng MCY, Bonomo JA, Hicks PJ, Hester JM, Langefeld
662 CD, Freedman BI, Bowden DW: **Analysis of coding variants identified from exome**
663 **sequencing resources for association with diabetic and non-diabetic nephropathy**
664 **in African Americans.** *Hum Genet* 2014, **133**:769–779.

665 44. Gomez F, Hirbo J, Tishkoff SA: **Genetic Variation and Adaptation in Africa:**
666 **Implications for Human Evolution and Disease**. *Cold Spring Harbor Perspectives*
667 *in Biology* 2014, **6**:a008524–a008524.

668 45. Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M,
669 Salie T, Rohland N, Mallick S, Peltzer A, Heinze A, Olalde I, Ferry M, Harney E,
670 Michel M, Stewardson K, Cerezo-Román JI, Chiumia C, Crowther A, Gomani-
671 Chindebvu E, Gidna AO, Grillo KM, Helenius IT, Hellenthal G, Helm R, Horton M,
672 López S, Mabulla AZP, Parkington J, Shipton C, et al.: **Reconstructing Prehistoric**
673 **African Population Structure**. *Cell* 2017, **171**:59–71.e21.

674 46. Grollemund R, Branford S, Bostoen K: **Bantu expansion shows that habitat**
675 **alters the route and pace of human dispersals**. 2015.

676 47. Vansina J: **Linguistic evidence for the introduction of ironworking into**
677 **Bantu-Speaking Africa**. *History in Africa* 2006.

678 48. Schiffels S, Durbin R: **Inferring human population size and separation history**
679 **from multiple genome sequences.** *Nat Genet* 2014, **46**:919–925.

680 49. Gifford-Gonzalez D: **Animal Disease Challenges to the Emergence of**
681 **Pastoralism in Sub-Saharan Africa**. *African Archaeological Review* 2000, **17**:95–
682 139.

683 50. Smetko A, Soudré A, Silbermayr K, Müller S, Brem G, Hanotte O, Boettcher PJ,
684 Stella A, Mészáros G, Wurzinger M, Curik I, Müller M, Burgstaller J, Sölkner J:
685 **Trypanosomosis: potential driver of selection in African cattle.** *Front Gene* 2015,
686 **6**:137.

687 51. Chritz KL, Marshall FB, Zagal ME, Kirera F, Cerling TE: **Environments and**
688 **trypanosomiasis risks for early herders in the later Holocene of the Lake**
689 **Victoria basin, Kenya**. *Proc Natl Acad Sci USA* 2015:201423953.

690 52. Chouin GLF: **Fossés, enceintes et peste noire en Afrique de l'Ouest forestière**
691 **(500-1500 AD). Réflexions sous canopée**. *http://aaarevuesorg* 2013:43–66.

692 53. Mack R: **The great African cattle plague epidemic of the 1890's**. *Tropical*
693 *Animal Health and Production* 1970.

694 54. Gillespie JH: *The Causes of Molecular Evolution*. Newyork: Oxford Univ. Press;
695 New York; 1994.

696 55. Thorne JL, Kishino H, Painter IS: **Estimating the rate of evolution of the rate of**
697 **molecular evolution**. *Molecular Biology and Evolution* 1998, **15**:1647–1657.

698 56. Hardwick RJ, Ménard A, Sironi M, Milet J, Garcia A, Sese C, Yang F, Fu B,
699 Courtin D, Hollox EJ: **Haptoglobin (HP) and Haptoglobin-related protein (HPR)**
700 **copy number variation, natural selection, and trypanosomiasis.** *Hum Genet* 2013.

701 57. Pays E, Vanhollebeke B, Uzureau P, Lecordier L, Perez-Morga D: **The molecular**
702 **arms race between African trypanosomes and humans.** *Nat Rev Microbiol* 2014,
703 **12**:575–584.

704 58. Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI,
705 Bowden DW, Langefeld CD, Oleksyk TK, Uscinski Knob AL, Bernhardy AJ, Hicks
706 PJ, Nelson GW, Vanhollebeke B, Winkler CA, Kopp JB, Pays E, Pollak MR:
707 **Association of trypanolytic ApoL1 variants with kidney disease in African**
708 **Americans.** *Science* 2010, **329**:841–845.

709 59. Cooper A, Ilboudo H, Alibu VP, Ravel S, Enyaru J, Weir W, Noyes H, Capewell
710 P, Camara M, Milet J, Jamonneau V, Camara O, Matovu E, Bucheton B, Macleod A:
711 **APOL1 renal risk variants have contrasting resistance and susceptibility**
712 **associations with African trypanosomiasis.** *Elife* 2017, **6**.

713 60. Courtin D, Milet J, Sabbagh A, Massaro JD, Castelli EC, Jamonneau V, Bucheton
714 B, Sese C, Favier B, Rouas-Freiss N, Moreau P, Donadi EA, Garcia A: **HLA-G 3'**
715 **UTR-2 haplotype is associated with Human African trypanosomiasis**
716 **susceptibility.** *Infect Genet Evol* 2013.

717 61. Courtin D, Milet J, Jamonneau V, Yeminanga CS, Kumeso VKB, Bilengue
718 CMM, Betard C, Garcia A: **Association between human African trypanosomiasis**
719 **and the IL6 gene in a Congolese population.** *Infect Genet Evol* 2007, **7**:60–68.

720 62. Vitti JJ, Grossman SR, Sabeti PC: **Detecting Natural Selection in Genomic**
721 **Data**. *Annu Rev Genet* 2013, **47**:97–120.

722 63. Mangeat B, Turelli P, Caron G, Friedli M, Perrin L, Trono D: **Broad**
723 **antiretroviral defence by human APOBEC3G through lethal editing of nascent**
724 **reverse transcripts.** *Nature* 2003, **424**:99–103.

725 64. Ministry of Health Uganda, ICF, CDC, USAID, WHO: *Uganda AIDS Indicator*
726 *Survey (AIS)*. AIDS Indicator Survey (AIS); 2011.

727 65. Stopak K, de Noronha C, Yonemoto W, Greene WC: **HIV-1 Vif blocks the**
728 **antiviral activity of APOBEC3G by impairing both its translation and**
729 **intracellular stability.** *Mol Cell* 2003, **12**:591–601.

730 66. Sheehy AM, Gaddis NC, Malim MH: **The antiretroviral enzyme APOBEC3G**

731 **is degraded by the proteasome in response to HIV-1 Vif.** *Nat Med* 2003, **9**:1404–
732 1407.

733 67. Bwogi J, Braka F, Makumbi I, Mishra V, Bakamutumaho B, Nanyunja M, Opio
734 A, Downing R, Biryahwaho B, Lewis RF: **Hepatitis B infection is highly endemic**
735 **in Uganda: findings from a national serosurvey.** *Afr Health Sci* 2009, **9**:98–108.

736 68. Ochola E, Ocama P, Orach CG, Nankinga ZK, Kalyango JN, McFarland W,
737 Karamagi C: **High burden of hepatitis B infection in Northern Uganda: results of**
738 **a population-based survey.** *BMC Public Health* 2013, **13**:727.

739 69. Chen R, Zhao X, Wang Y, Xie Y, Liu J: **Hepatitis B virus X protein is capable**
740 **of down-regulating protein level of host antiviral protein APOBEC3G.** *Sci Rep*
741 2017, **7**:40783.

742 70. Lu H-L, Liao F: **Melanoma differentiation-associated gene 5 senses hepatitis B**
743 **virus and activates innate immune signaling to suppress virus replication.** *J*
744 *Immunol* 2013, **191**:3264–3276.

745 71. Teshale EH, Kamili S, Drobeniuc J, Denniston M, Bakamutamaho B, Downing R:
746 **Hepatitis B virus infection in northern Uganda: Impact of pentavalent hepatitis**
747 **B vaccination.** *Vaccine* 2015, **33**:6161–6163.

748 72. Ilboudo H, Noyes H, Mulindwa J, Kimuda MP, Koffi M, Kaboré JW, Ahouty B,
749 Ngoyi DM, Fataki O, Simo G, Ofon E, Enyaru J, Chisi J, Kamoto K, Simuunza M,
750 Alibu VP, Lejon V, Jamonneau V, Macleod A, Camara M, Bucheton B, Hertz-Fowler
751 C, Sidibe I, Matovu E, TrypanoGEN Research Group as members of The H3Africa
752 Consortium: **Introducing the TrypanoGEN biobank: A valuable resource for the**
753 **elimination of human African trypanosomiasis.** *PLoS Negl Trop Dis* 2017,
754 **11**:e0005438.

755 73. Ramasamy RK, Ramasamy S, Bindroo BB, Naik VG: **STRUCTURE PLOT: a**
756 **program for drawing elegant STRUCTURE bar plots in user friendly interface.**
757 *Springerplus* 2014, **3**:431.

758 74. Felsenstein J: **{PHYLIP}: phylogenetic inference package, version 3.5c.** *(1993)*
759 1993.

760 75. Ersts PJ: *Geographic Distance Matrix Generator (Version 1.2.3). American*
761 *Museum of Natural History*. Center for Biodiversity and Conservation; 2011.

762 76. Williams AL, Patterson N, Glessner J: **Phasing of Many Thousands of**
763 **Genotyped Samples - ScienceDirect**. *The American Journal of …* 2012.

764 77. Delaneau O, Zagury J-F, Marchini J: **Improved whole-chromosome phasing for**
765 **disease and population genetic studies.** *Nat Meth* 2013, **10**:5–6.

766 78. Gautier M, Vitalis R: **rehh: an R package to detect footprints of selection in**
767 **genome-wide SNP data from haplotype structure**. *Bioinformatics* 2012, **28**:1176–
768 1177.

769 79. Tang K, Thornton KR, Stoneking M: **A new approach for using genome scans**

770 **to detect recent positive selection in the human genome.** *PLoS Biol* 2007, **5**:e171.

771 80. *Ethnologue: Languages of the World* [https://www.ethnologue.com]

772 81. Ranasinghe S, Cutler S, Davis I, Lu R, Soghoian DZ, Qi Y, Sidney J, Kranias G,
773 Flanders MD, Lindqvist M, Kuhl B, Alter G, Deeks SG, Walker BD, Gao X, Sette A,
774 Carrington M, Streeck H: **Association of HLA-DRB1-restricted CD4+ T cell**
775 **responses with HIV immune control.** *Nat Med* 2013, **19**:930–933.

776 82. Hill AV, Allsopp CE, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA,
777 Bennett S, Brewster D, McMichael AJ, Greenwood BM: **Common west African**
778 **HLA antigens are associated with protection from severe malaria.** *Nature* 1991,
779 **352**:595–600.

780 83. Pontillo A, Brandão LA, Guimarães RL, Segat L, Athanasakis E, Crovella S: **A**
781 **3'UTR SNP in NLRP3 gene is associated with susceptibility to HIV-1 infection.** *J*
782 *Acquir Immune Defic Syndr* 2010, **54**:236–240.

783 84. Zhang S, Feng Y, Narayan O, Zhao LJ: **Cytoplasmic retention of HIV-1**
784 **regulatory protein Vpr by protein-protein interaction with a novel human**
785 **cytoplasmic protein VprBP.** *Gene* 2001, **263**:131–140.

786 85. Deng J, Chen Y, Ding D, Lu P, Yang X, Song Z, Zhu H: **TRIM5α H43Y**
787 **Polymorphism and Susceptibility to HIV-1 Infection: A Meta-Analysis.** *AIDS Res*
788 *Hum Retroviruses* 2015, **31**:1213–1218.

789 86. Meyerson NR, Rowley PA, Swan CH, Le DT, Wilkerson GK, Sawyer SL:
790 **Positive selection of primate genes that promote HIV-1 replication.** *Virology*
791 2014, **454-455**:291–298.

792 87. Louie LG, Hartogensis WE, Jackman RP, Schultz KA, Zijenah LS, Yiu CH-Y,
793 Nguyen VD, Sohsman MY, Katzenstein DK, Mason PR: **Mycobacterium**
794 **tuberculosis/HIV-1 coinfection and disease: role of human leukocyte antigen**
795 **variation.** *J Infect Dis* 2004, **189**:1084–1090.

796 88. Taniguchi M, Kawabata M: **KIR3DL1/S1 genotypes and KIR2DS4 allelic**
797 **variants in the AB KIR genotypes are associated with Plasmodium-positive**
798 **individuals in malaria infection.** *Immunogenetics* 2009, **61**:717–730.

799 89. Norman PJ, Hollenbach JA, Nemat-Gorgani N, Guethlein LA, Hilton HG, Pando
800 MJ, Koram KA, Riley EM, Abi-Rached L, Parham P: **Co-evolution of human**
801 **leukocyte antigen (HLA) class I ligands with killer-cell immunoglobulin-like**
802 **receptors (KIR) in a genetically diverse population of sub-Saharan Africans.**
803 *PLoS Genet* 2013, **9**:e1003938.

804 90. Hsieh F-L, Turner L, Bolla JR, Robinson CV, Lavstsen T, Higgins MK: **The**
805 **structural basis for CD36 binding by the malaria parasite.** *Nat Commun* 2016,
806 **7**:12837.

807 91. Manjurano A, Sepúlveda N, Nadjm B, Mtove G, Wangai H, Maxwell C, Olomi R,
808 Reyburn H, Drakeley CJ, Riley EM, Clark TG, in Collaboration With MalariaGEN:
809 **USP38, FREM3, SDC1, DDC, and LOC727982 Gene Polymorphisms and**

Differential Susceptibility to Severe Malaria in Tanzania. *J Infect Dis* 2015, **212**:1129–1139.

92. Patrinos GP, Kollia P, Papadakis MN: **Molecular diagnosis of inherited disorders: lessons from hemoglobinopathies.** *Hum Mutat* 2005, **26**:399–412.

93. Maiga B, Dolo A, Touré O, Dara V, Tapily A, Campino S, Sepúlveda N, Risley P, Silva N, Silva N, Corran P, Rockett KA, Kwiatkowski D, MalariaGEN Consortium, Clark TG, Troye-Blomberg M, Doumbo OK: **Human candidate polymorphisms in sympatric ethnic groups differing in malaria susceptibility in Mali.** *PLoS ONE* 2013, **8**:e75675.

94. *Africa During the Last 150,000 Years* [http://www.esd.ornl.gov/projects/gen/nercAFRICA.html]

95. Greenberg JH: *Greenberg: the Languages of Africa - Google Scholar*. 1966.

843 **Tables**

844

845 **Table 1.** Table showing the Ethnic groups and number of individual from each
846 Country that were used for Whole genome sequencing. The Ethnicity codes were
847 obtained from the Ethnologue languages of the World catalogue [80].

848

Country	Code	Ethnicity	Ethnologue code	No. of Samples
Uganda	UNL	Lugbara	lgg	50
	UBB	Basoga	xog	33
Zambia	ZAM	Soli/Chikunda	sby; kdn	28
		Tumbuka	tum	14
		Bemba	bem	8
Congo	DRC	Kimbala	mdp	20
		Kingongo	noq	30
Cameroon	CAM	Bamilike	fmp	6
		Mundani	mnf	8
		Ngoumba	nmg	12
Ivory Coast	CIV	Baoule	bci	11
		Gouro	goa	21
		More	Moa	12
		Senoufo	sef	4
		Malinke	loi	1
		Koyaka	kga	1
Guinea	GAS	Soussou	sus	50

849

850 **Table 2**. The number of SNPs and Indels obtained from the mapping and variant
851 calling pipeline. The SNPs were filtered for HWE, MAF and missing genotypes
852

Pop	Samples (WGS)	Samples after filtering	SNPs before filtering	SNPs after filtering	Indels before filtering	Indels after filtering
CIV	50	50	18,780,913	16,066,827	3,069,408	1,583,594
DRC	50	49	19,188,537	16,449,696	3,146,802	1,626,826
GAS	50	47	18,831,834	16,075,002	3,063,080	1,579,352
UBB	33	32	17,671,306	14,987,699	2,889,915	1,426,646
UNL	50	44	18,986,243	15,598,629	3,130,979	1,536,490
CAM*	26	26	17,183,994	14,579,603	3,283,543	1,539,459
ZAM*	41	41	18,232,386	15,548,110	3,448,501	1,651,467
Total	**300**	**289**	**34,116,333**	**30,591,165**	**5,336,622**	**3,166,196**

* Samples Sequenced at 30X coverage, rest of the samples were at 10X coverage. Identified 2,023,049 SNPS without rsIDs {Total SNPs (30,591,165) – SNPs with rsIDs (28,568,116)}

853

854

855 **Table 3.** Extreme iHS loci that overlap with the UNL population
856

Pop	Extreme iHS SNPs (–log p > 3.0)	Extreme iHS SNPS associated with protein coding genes	Extreme iHS SNPs overlapping with UNL
UNL	8454	2613	2613
UBB	9617	3326	512
DRC	10037	3790	535
ZAM	5570	1990	86
CIV	10129	3541	534
CAM	7686	2597	82
GAS	11401	3741	382

857

858

859
860
861
862

863 **Table 4**. DAVID [25] analysis of Genes that are highly selected within TrypanoGEN population and associated with HIV, Tuberculosis, and
864 Malaria. The Fisher's exact test *P*-values indicate significant gene enrichment in the associated disease
865

Gene	Chr	Populations affected	Associated Disease	p value	Reference
HLA-DRB1	6p21.32	ZAM, CAM, CIV, DRC, UBB	HIV/TB/Malaria	1.63E-09	[81, 82]
NLRP1	17p13.2	ZAM, CIV, DRC, GAS, UBB	HIV	2.42E-07	[83]
VPRBP	3p10.6	UBB, CIV, DRC, GAS	HIV	2.42E-07	[84]
TRIM5	11p15.4	UNL, CAM, GAS	HIV	7.30E-07	[85]
ANKRD30A	10p11.21	DRC, CIV	HIV	2.42E-07	[86]
HLA-A	6p22.1	ZAM, CAM	HIV/TB	4.70E-06	[87]
HLA-DQA1	6p21.32	UBB, DRC	HIV/TB	4.70E-06	[87]
HLA-DQB1	6p21.32	UBB, DRC	HIV/TB	4.70E-06	[87]
KIR3DL1	19q13.42	UNL, CIV	Malaria	1.63E-09	[88, 89]
CD36	7q21.11	UNL, CAM, CIV	Malaria	1.55E-06	[90]
DDC	7p12.2	UBB, DRC, GAS	Malaria	1.63E-09	[91]
HBE1	11p15.4	UBB, CAM, DRC	Malaria	5.48E-07	[92]
ADCY9	16p13.3	UNL, CIV	Malaria	5.48E-07	[93]

866
867
868
869
870
871

872 **Table 5**. Genes that are highly differentiated between the Nilo-Saharan and Trypanogen Niger-Congo populations that contain SNPs
873 unique to UNL population
874
875

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

876 **Figures**

877 **Figure 1**. Multi dimensional scaling (MDS) analysis of the sequenced TrypanoGEN
878 samples, Guinea (GAS), Ivory Coast (CIV), Cameroon (CAM), Democratic Republic
879 of Congo (DRC), Uganda (Nilotics, UNL, Bantu, UGB) and Zambia (ZAM), (**A**); **B**,
880 TrypanoGEN and selected 1000 genomes African samples Nigeria (ESN, YRI),
881 Sierra Leone (MSL), Gambia (GWD), Kenya (LWK); The black circles indicate the
882 clustering of the samples into the West African Niger-Congo-A, Nigerian Niger-
883 Congo-A, Central African Niger-Congo-B, East African Niger-Congo-B and the Nilo-
884 Saharan Lugbara (NSL); **C,** 1000 genomes samples from Africa and the rest of the
885 world. AAM, African Americans; AMR, indigenous Americans; CAF, Central Africa;
886 EAF, East Africa; EAS, East Asia; EUR, Europe; NSA, Nilo-Saharan; SAS, South
887 Asia; WAF, West Africa;

888

889 **Figure 2**. Genetic admixture and diversity between TrypanoGEN and selected 1000
890 genome populations. **A**. Admixture plot of the K populations of the TrypanoGEN,
891 1000 genome African and European populations. **B**. Heatmap of mean Fst between
892 TrypanoGEN and 1000 genome African populations. **C**. Polynomial regression plot
893 of the mean Fst against the relative geographical distance of the African Niger-Congo
894 populations from the Uganda Nilotic population. **D**. Phylogeographic plot of the mean
895 Fst distances on the Trypanogen populations and selected 1000 genome African
896 populations; the pie charts represent the population sample size and admixture.

897

898 **Figure 3.** Population sizes and cross-coalescence rates compiled by MSMC. **A**
899 Effective population sizes for each population since 75kya. The Ugandan Bantu and
900 Nilotic populations have grown continuously but at a slower rate than the West and

901 Central African populations since 13kya. These latter populations experienced
902 declines of 6-23% between 1500 and 800 years ago. **B** Cross-coalescence rates for
903 pairs of populations. At 1.0 populations are panmictic and at 0.0 there is no gene flow.
904 The Guinea and Ivory Coast populations were panmictic until about 10 kya and then
905 became separated by 3kya. Other populations appear to have separated more
906 gradually but these may be confounded by admixture.

907

908 **Figure 4.** Illustration of signatures unique to the Uganda Nilotic population. Signal of
909 positive selection within the *APOBEC3G* gene on Chromosome 22 at the
910 rs112077004 loci of the Uganda Nilo-Saharan Lugbara population, in comparison
911 with the Niger-Congo B populations of Uganda (UGB) and Niger-Congo A
912 population of Guinea (GAS). **A**. The calculated site specific extended haplotype
913 homozygosity (EHH) within a population. **B**. Between population Fst analysis. **C.**
914 Across population (XPEHH) analysis.

915

916 **Figure 5.** Migrations of Niger-Congo speakers. Map colours show vegetation
917 coverage approximately 10kya [94]. Colours for linguistic groups as for Figure 2.
918 Blue Nilo-Saharan; Yellow, Niger-Congo A; Red, Niger-Congo-B (Bantu); Green
919 putative expansion of an ancestral group out of modern Nigeria. Blue dotted arrow,
920 suspected route of proto-Niger-Congo-A speakers from Nuba mountains of Sudan to
921 Senegal across the Sahel [95] when it was much wetter than at present.

922

923 **Additional files**

924

925 **Additional file 1:** Table S0.xlsx. List of participant samples sequenced indicating
926 their sequencing ID, country of origin, language and sex of the individual.
927
928 **Additional file 2**: Figure S1-S3.pdf: Figure S1. Variant annotation of the population
929 whole genome sequence (WGS) data using snpEff. **a.** count of number of
930 substitutions for each base combination, **b.** count of number of Insertion-Deletions
931 with corresponding InDel length of affected loci **c.** Count of number of sites with
932 Transition-Tranversion ratios with the corresponding allele frequency, **d.** Percentage
933 count of the variant annotation classifications. Figure S2. Plot of the coefficient of
934 variation (CV) error verses the K population (number of clusters). The K population
935 with the lowest CV error is considered as the most appropriate K for the admixture.
936 Figure S3. Bar plot of the mean F_{ST} between TrypanoGEN populations.
937
938 **Additional file 3**: Table S1.xlsx. Table showing MSMC output of effective
939 populations sizes at different time points
940
941 **Additional file 4**: Figure S4.pdf. Extended haplotype homozygosity analysis of the
942 TrypanoGEN populations for signatures of natural selection. **A.** Manhattan plot of the
943 Intra-population integrated haplotype score (iHS) showing SNPs with extreme
944 positive (> 2.5) and extreme negative (< -2.5). **B.** Bar plot showing the number of
945 SNPs associated with protein coding with iHS > 3.0 in each population (black) and
946 those within each population that overlap with the UNL population (grey).
947 Distribution of the iHS scores compared by **C.** Histograms and **D.** QQ plots.
948
949 **Additional file 5**: Table S2.xlsx. List of SNPs with Ensembl annotation of the nearest

950 protein coding gene and iHS > -log p 3.0. It contains sheets for the TrypanoGEN
951 populations and combined sheet for all population genes and DAVID derived Genetic
952 associated diseases (GAD) and KEGG pathway affiliation
953
954 **Additional file 6**: Table S3.xlsx. List of signature SNPs associated with protein
955 coding genes, with _iHS > -log p 3.0 and unique to the Nilo-Saharan Lugbara
956 population.

957 **Additional file 7**: Figure S5.pdf. PANTHER Gene ontology database classification of
958 the Nilo-Saharan Lugbara unique genes based on the Biological process and protein
959 class they belong to. X-axis indicates the process and Y-axis the number of genes
960 corresponding to the process.

961 **Additional file 8**: Table S4.xlsx. List of SNPs that are highly differentiated between
962 the Niger-Congo and Nilo-Saharan Lugbara populations.

963 **Additional file 9**: Figure S6.pdf. Manhattan plot of the Rsb (**A**) and F_{ST} (**B**) against
964 the whole genome (chromosome 1 - 22) between the Nilo-Saharan Lugbara (UNL)
965 and TrypanoGEN Niger-Congo populations. **C.** Scatter plot of inter-population F_{ST}
966 against Rsb scores for the SNPs having a high F_{ST} (> 0.99 percentile) and high Rsb (-
967 log p-value > 3.0) between UGN and the Niger-Congo TrypanoGEN populations. **D**.
968 Venn diagram showing the number of intersecting protein coding SNP loci between
969 the high Rbs-Fst (UNL verses All populations (CAM, CIV, GAS, UBB, ZAM) and
970 high iHS (-log p-value > 3) protein coding loci unique to the UNL population only.

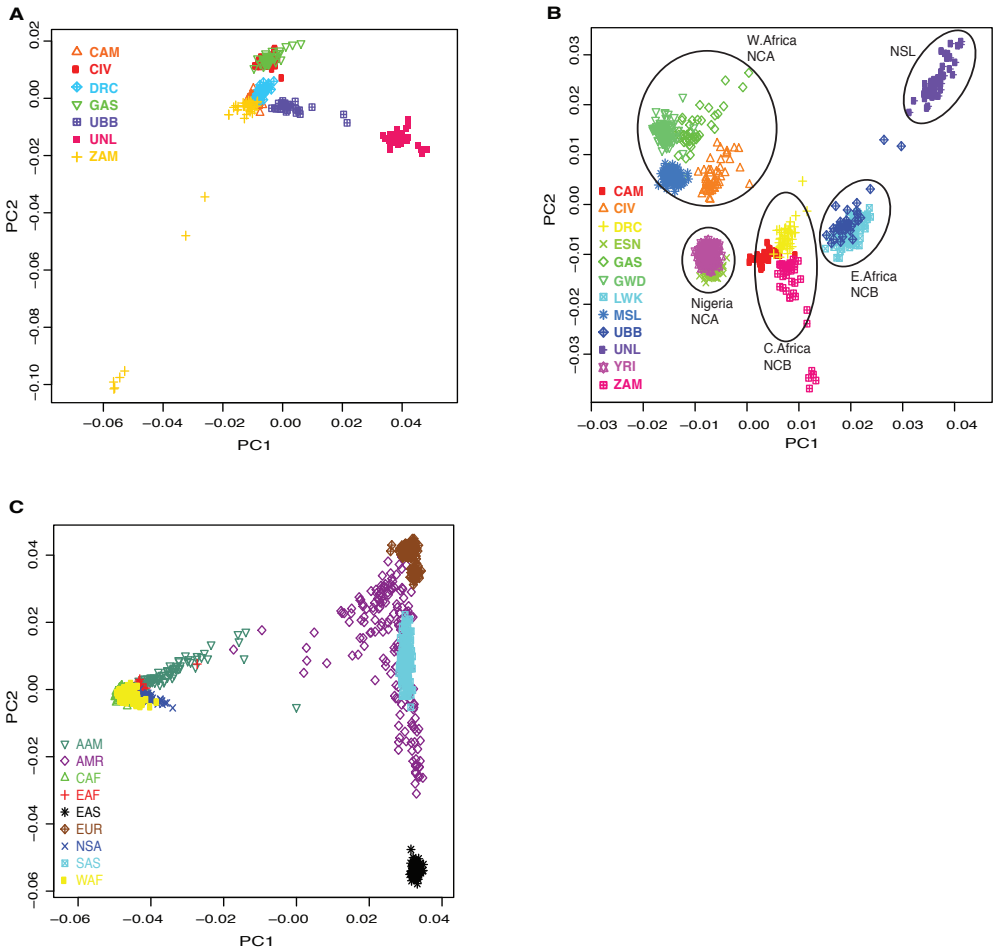
971 **Additional file 10**: Figure S7. Analysis of SNP loci in APOBEC3G gene
972 (Chr22:39480259-39485103). **A.** Plot of SNP minor allele frequency (MAF) against
973 chromosome position in the gene, the red dot is the position of rs112077004. **B.**

974 Haploview plot of Linkage Disequilibrium (LD) of haplotypes within the gene, **C.**
975 Zoomed LD plot of the SNPs upstream and downstream of the rs112077004 loci
976 which is unique to the UGN population. **D.** Ensemble gene analysis of *APOBEC3G*
977 gene locus on the GRCH37 build indicating the position of SNP rs112077004.

978 **Additional file 11**: Table S5.xlsx. Analysis of SNPs in the APOBEC3G gene
979 including the minor allele frequency (MAF) and linkage disequilibrium (LD)
980 statistics.

981 **Additional file 12**: Figure S8.pdf. Bar plots comparing the mean iHS *P*-values of the
982 SNP loci in *APOL1*, *HPR*, *HLAG*, and *IL6* genes within the TrypanoGEN
983 populations. These genes are associated with progression of Human African
984 Trypanosomiasis.

985 **Additional file 13**: Table S6.xlsx. Population iHS scores of SNPs in genes that play a
986 role in African trypanosomiasis pathology



987

988 Figure 1

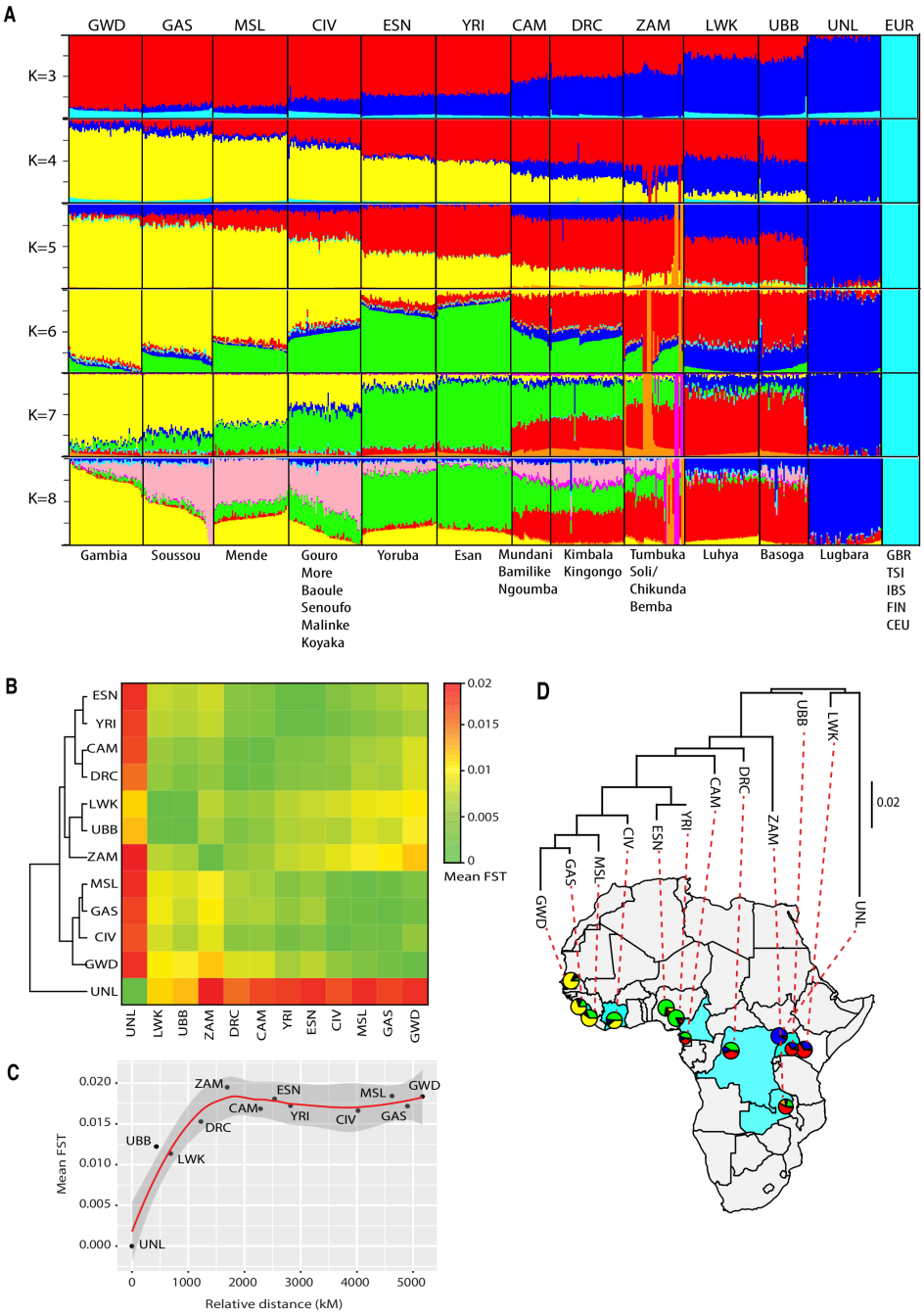

A
GWD
GAS
MSL
CIV
ESN
YRI
CAM
DRC
ZAM
LWK
UBB
UNL
EUR
K=3
K=4
K=5
K=6
K=7
K=8
Gambia
Soussou
Mende
Gouro
More
Baoule
Senoufo
Malinke
Koyaka
Yoruba
Esan
Mundani
Bamilike
Ngoumba
Kimbala
Kingongo
Tumbuka
Soli/
Chikunda
Bemba
Luhya
Basoga
Lugbara
GBR
TSI
IBS
FIN
CEU
B
Mean FST
C
Mean FST
Relative distance (kM)
D
0.02

Figure 2

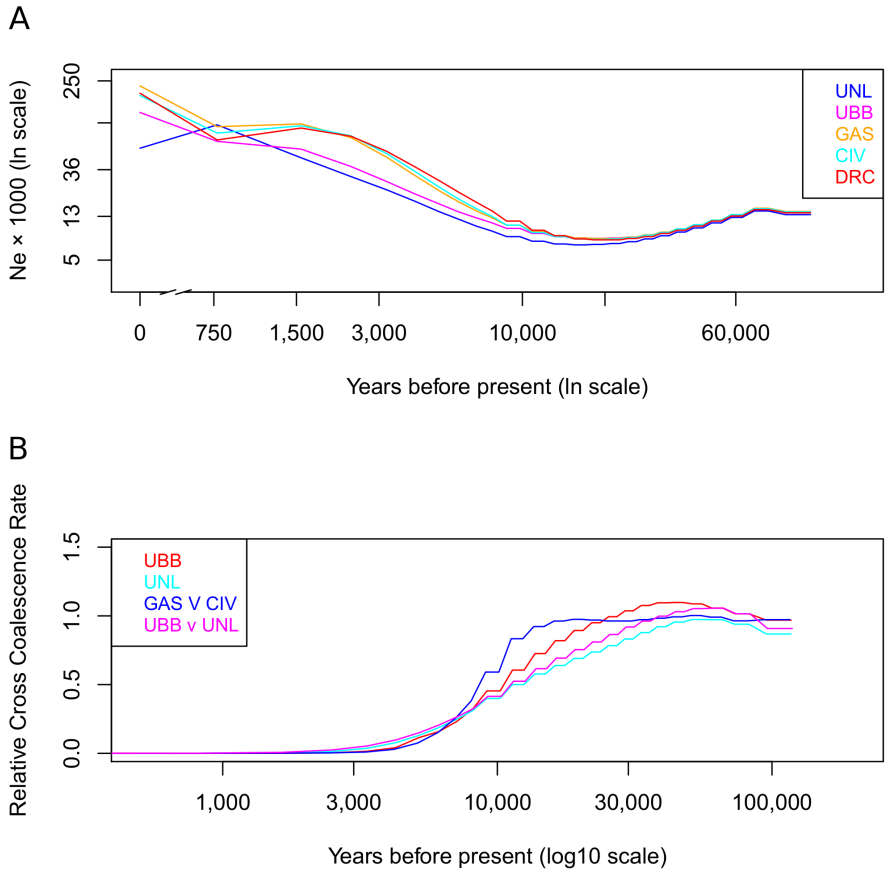

A
Ne × 1000 (ln scale)
250
36
13
5
UNL
UBB
GAS
CIV
DRC
0
750
1,500
3,000
10,000
60,000
Years before present (ln scale)
B
Relative Cross Coalescence Rate
1.5
1.0
0.5
0.0
UBB
UNL
GAS V CIV
UBB v UNL
1,000
3,000
10,000
30,000
100,000
Years before present (log10 scale)

Figure 3

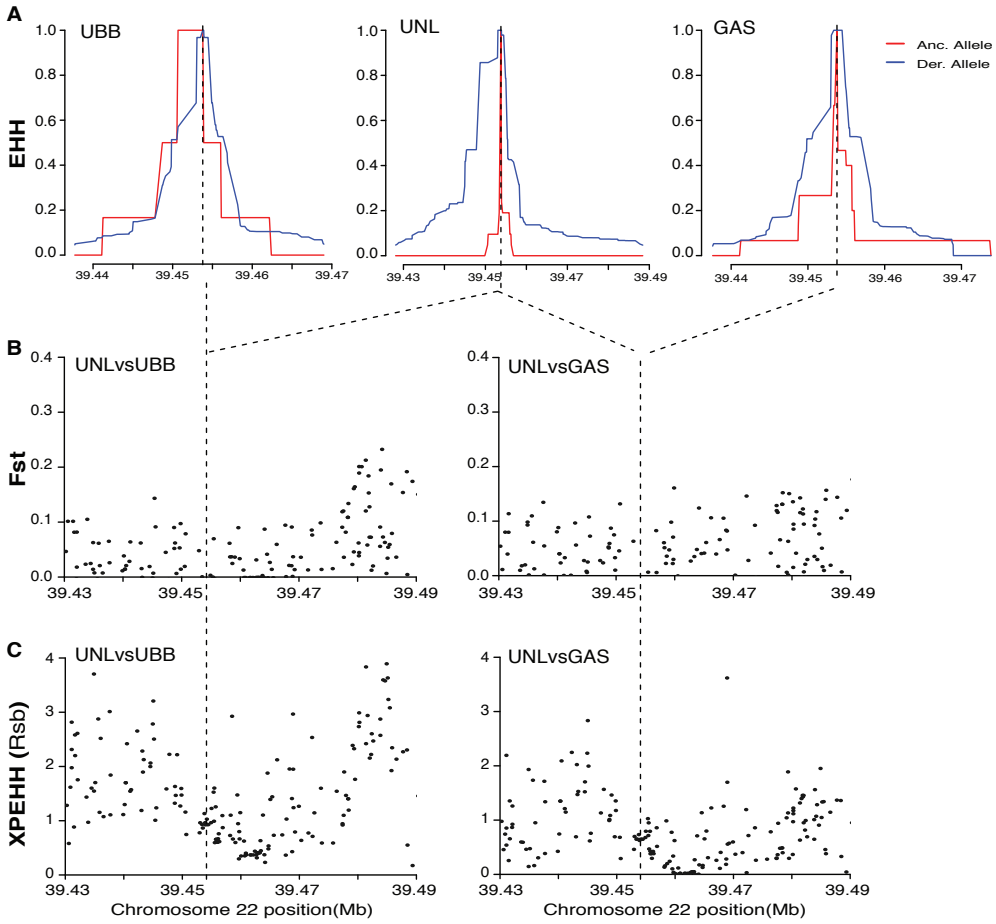

A
UBB
UNL
GAS
Anc. Allele
Der. Allele
EHH
B
UNLvsUBB
UNLvsGAS
Fst
C
UNLvsUBB
UNLvsGAS
XPEHH (Rsb)
Chromosome 22 position(Mb)
Chromosome 22 position(Mb)

Figure 4

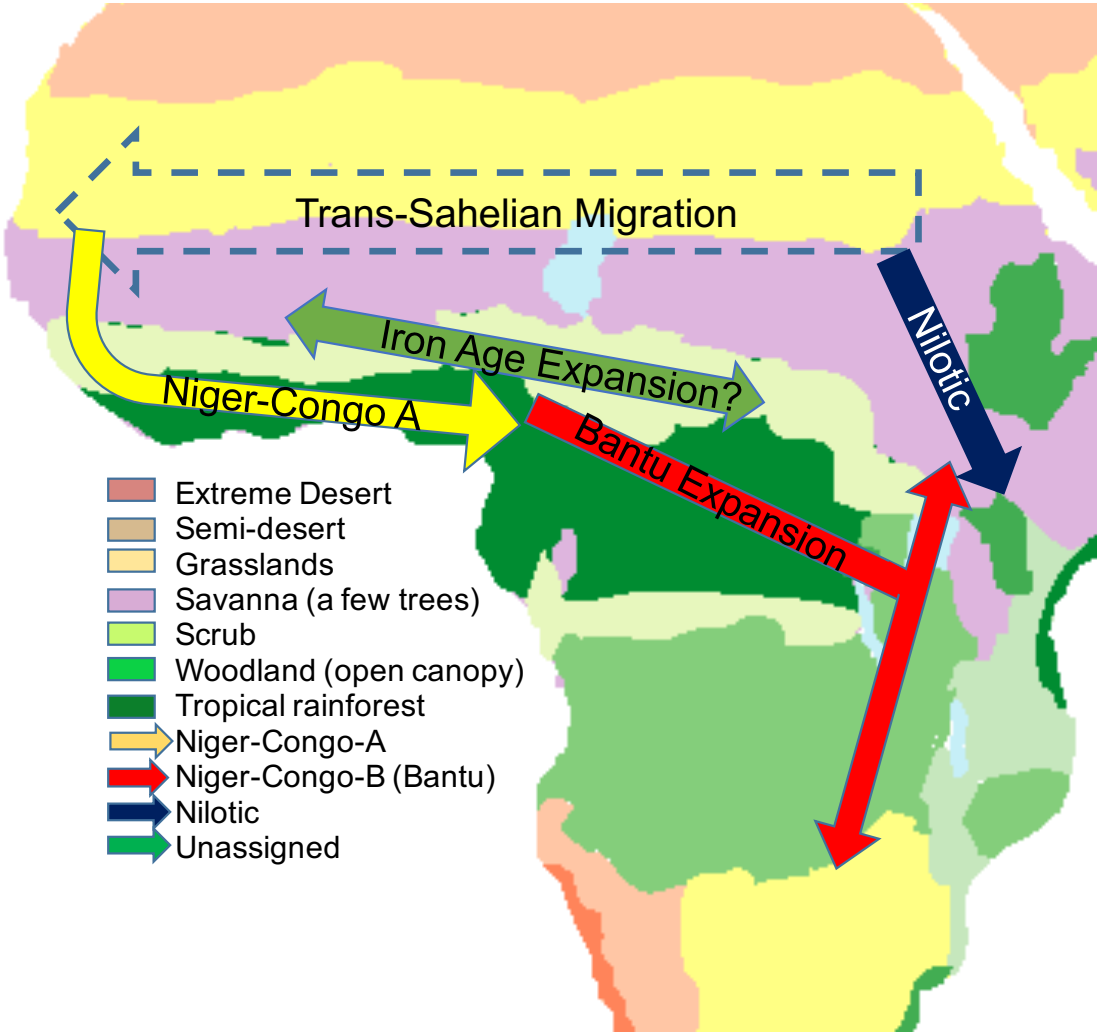

Trans-Sahelian Migration
Iron Age Expansion?
Niger-Congo A
Bantu Expansion
Nilotic
Extreme Desert
Semi-desert
Grasslands
Savanna (a few trees)
Scrub
Woodland (open canopy)
Tropical rainforest
Niger-Congo-A
Niger-Congo-B (Bantu)
Nilotic
Unassigned

Figure 5