PLOS GENETICS

RESEARCH ARTICLE

Genetic signatures of gene flow and malaria-driven natural selection in sub-Saharan populations of the "endemic Burkitt Lymphoma belt"

Mateus H. Gouveia[1,2,3☯], Andrew W. Bergen[4☯¤], Victor Borda[2☯], Kelly Nunes[5☯], Thiago P. Leal[2,6], Martin D. Ogwang[7], Edward D. Yeboah[8], James E. Mensah[8], Tobias Kinyera[7], Isaac Otim[7], Hadijah Nabalende[7], Ismail D. Legason[7], Sununguko Wata Mpoloka[9], Gaonyadiwe George Mokone[10], Patrick Kerchan[7], Kishor Bhatia[4], Steven J. Reynolds[11], Richard B. Birtwum[8], Andrew A. Adjei[8], Yao Tettey[8], Evelyn Tay[8], Robert Hoover[4], Ruth M. Pfeiffer[4], Robert J. Biggar[4], James J. Goedert[4], Ludmila Prokunina-Olsson[12], Michael Dean[12], Meredith Yeager[13], M. Fernanda Lima-Costa[1], Ann W. Hsing[14], Sarah A. Tishkoff[15], Stephen J. Chanock[4], Eduardo Tarazona-Santos[2‡], Sam M. Mbulaiteye[4‡*]

1 Instituto de Pesquisa René Rachou, Fundação Oswaldo Cruz, Belo Horizonte, Minas Gerais, Brazil, **2** Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **3** Center for Research on Genomics & Global Health, National Institutes of Health, US Department of Health and Human Services, Bethesda, Maryland, United States of America, **4** Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, US Department of Health and Human Services, Bethesda, Maryland, United States of America, **5** Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil, **6** Department of Statistics, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **7** EMBLEM Study, African Field Epidemiology Network, Kampala, Uganda, **8** University of Ghana Medical School, Accra, Ghana, **9** Department of Biological Sciences, University of Botswana, Gaborone, Botswana, **10** Department of Biomedical Sciences, University of Botswana School of Medicine, Gaborone, Botswana, **11** Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, US Department of Health and Human Services, Bethesda, Maryland, United States of America, **12** Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, US Department of Health and Human Services, Bethesda, Maryland, United States of America, **13** Cancer Genomics Research Laboratory, Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, US Department of Health and Human Services, Frederick, Maryland, United States of America, **14** Stanford Cancer Institute, Stanford University, Stanford, California, United States of America, **15** Department of Genetics and Biology, University of Pennsylvania, Philadelphia, United States of America

☯ These authors contributed equally to this work.
¤ Current address: Oregon Research Institute, Eugene, Oregon
‡ These authors also contributed equally to this work.
* mbulaits@mail.nih.gov

OPEN ACCESS

Citation: Gouveia MH, Bergen AW, Borda V, Nunes K, Leal TP, Ogwang MD, et al. (2019) Genetic signatures of gene flow and malaria-driven natural selection in sub-Saharan populations of the "endemic Burkitt Lymphoma belt" PLoS Genet 15(3): e1008027. https://doi.org/10.1371/journal.pgen.1008027

Editor: Jun Z Li, University of Michigan, UNITED STATES

Received: June 21, 2018

Accepted: February 17, 2019

Published: March 8, 2019

Data Availability Statement: The data reported in this paper will be deposited in dbGap at the following link for the EMBLEM Study: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001705.v1.p1 and for the Ghana data at: phs000838.v1.p1.

Funding: The work was funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI) (Contracts HHSN2612011000063C

Abstract

Populations in sub-Saharan Africa have historically been exposed to intense selection from chronic infection with *falciparum* malaria. Interestingly, populations with the highest malaria intensity can be identified by the increased occurrence of endemic Burkitt Lymphoma (eBL), a pediatric cancer that affects populations with intense malaria exposure, in the so called "eBL belt" in sub-Saharan Africa. However, the effects of intense malaria exposure and sub-Saharan populations' genetic histories remain poorly explored. To determine if historical migrations and intense malaria exposure have shaped the genetic composition of the eBL

and HHSN261201100007I), and the Intramural Research Program, National Institute of Allergy and Infectious Diseases (S.J.R.), National Institutes of Health, Department of Health and Human Services. M.H.G., E.T.-S., T.P.L. and M.F.L.-C. are supported by Brazilian National Research Council (CNPq) and Minas Gerais Research Agency (FAPEMIG). M.H.G. performed part of this study as CAPES-PDSE fellow (99999.007069/2015-04), V.B. is a PEC-PG fellow (88882.195664/2018-01) of CAPES and K.N. performed the initial part of this study as CAPES-PNPD fellow—Brazil (1645581) and the latter part as United States National Institutes of Health fellow (R01 GM075091). Bioinformatics support was provided by the Sagarana HPC cluster, CPAD-ICB-UFMG, Brazil. The work in the Tishkoff laboratory was funded by RO1 grants from the National Institutes of Health (1R01DK104339-0 and 1R01GM113657-01). The content of this manuscript does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. The content of this publication is the sole responsibility of the authors. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

belt populations, we genotyped ~4.3 million SNPs in 1,708 individuals from Ghana and Northern Uganda, located on opposite sides of eBL belt and with $\geq$ 7 months/year of intense malaria exposure and published evidence of high incidence of BL. Among 35 Ghanaian tribes, we showed a predominantly West-Central African ancestry and genomic footprints of gene flow from Gambian and East African populations. In Uganda, the North West population showed a predominantly Nilotic ancestry, and the North Central population was a mixture of Nilotic and Southern Bantu ancestry, while the Southwest Ugandan population showed a predominant Southern Bantu ancestry. Our results support the hypothesis of diverse ancestral origins of the Ugandan, Kenyan and Tanzanian Great Lakes African populations, reflecting a confluence of Nilotic, Cushitic and Bantu migrations in the last 3000 years. Natural selection analyses suggest, for the first time, a strong positive selection signal in the *ATP2B4* gene (rs10900588) in Northern Ugandan populations. These findings provide important baseline genomic data to facilitate disease association studies, including of eBL, in eBL belt populations.

Author summary

We present a genome-wide analyses of genetic structure, gene flow, and natural selection in Ghana and Northern Uganda populations, both residing in the Sub-Saharan eBL belt, a region with intense *falciparum* malaria transmission and high endemic Burkitt Lymphoma (eBL) incidence. These populations are from different ethnolinguistic groups and are located 2400 miles apart in sub-Saharan Africa. We characterized genetic composition of these populations in the context of 22 additional African populations and present evidence for gene flow events that occurred in the last 3000 years, possibly related to regional migrations in Western Africa and major migrations involving Nilotic, Cushitic, and Bantu groups. We identified in Northern Ugandans a strong signal of malaria-driven selection in the *ATP2B4* gene coding for a calcium transporter expressed in erythrocytes. Characterization of biological relationships between the *ATP2B4* gene and malaria may inform the investigation of complex genomic disease associations in eBL belt populations.

Introduction

The endemic Burkitt Lymphoma (eBL) belt is a geographic area spanning 10˚N-10˚S and altitudes below 1500m above sea level (Fig 1A) in sub-Saharan Africa, where there is a high geographical correlation between malaria and eBL (an aggressive pediatric B-cell non-Hodgkin lymphoma). This correlation has led to the identification of malaria infection as a major driver of eBL [1][2], which was confirmed by the evidence that the sickle cell trait that protects against severe malaria [3] also protects against eBL [4]. Because eBL occurs in areas of sub-Saharan Africa [5] with stable intense *Plasmodium falciparum (Pf)* malaria (for 7–12 months in the year), eBL burden provides a novel way to identify populations under strong malaria selective pressure. *Pf* malaria is one of the most important selective pressures that have shaped the African genetic diversity [6], but there are limited reports on the combined effects of malaria-related natural selection and the demographic history of populations in the eBL belt.

The eBL belt was the scenario of several human migrations over the last 3000 years and archaeological and linguistic evidence have described the following historical events: in West

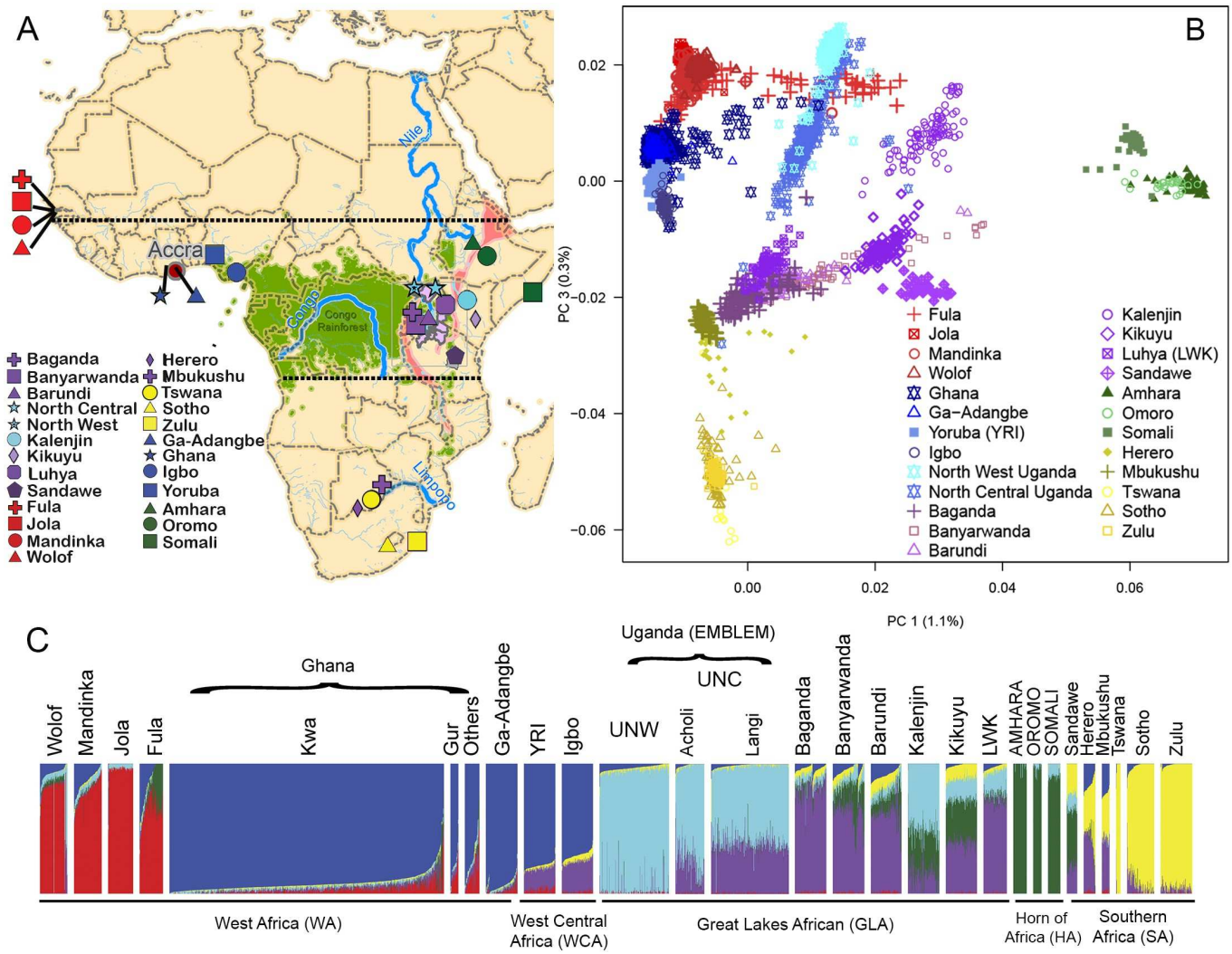

Fig 1. Populations studied in relation to major geographical barriers and analyses of population structure based on genotype data. (A) Map of Africa showing the geographical origin of the Pan-African populations used in the study comprising 22 populations from previous studies and three new populations in Ghana and the Uganda EMBLEM study (S1 Table). Horizontal dotted lines are the geographical extent of the endemic Burkitt lymphoma (eBL) belt (see S1 Fig and S1 Table for detailed information on Pan-African populations in the eBL belt). The map highlights major geographical features, such as the Congo rainforest (light green), major rivers and lakes, and the East and West African Rift valley systems (pink) that may have shaped migratory routes or constituted barriers to gene flow. (B) Principal Component Analysis (PCA) based on analysis of the genome-wide dataset of Uganda and Ghana integrated with 22 other pan-African populations. For better visualization, open symbols are used for the PCA plot and solid symbols are used in the map. The axes of the plot show the first and third principal components (See S10 Fig for other principal components). The PCA was repeated using similar number of individuals for each studied population (S9 Fig). (C) ADMIXTURE plot showing ancestry clusters in 28 populations from Africa (details in S1, S1A and S1B Table). The populations are listed left to right based on their geographical location in Africa from West to East and North to South. The colors represent different ancestral clusters, with K = 6 being the most likely number of clusters on admixture analysis (See S4 Fig). This ADMIXTURE analysis was repeated using similar number of individuals for each studied population (S6 Fig).

https://doi.org/10.1371/journal.pgen.1008027.g001

Africa: (i) interaction between West and West-Central Africa [7], (ii) cultural interaction between the local kingdoms of West-Central Africa [8,9], and (iii) migrations across the Sahel that include the westward Nilotic expansion [7,10,11]. In East Africa: (iv) Eastern Cushitics migrated from the Horn of Africa [7] into the Great Lakes region ~3000 years ago, maintaining (v) interactions with Nilotic groups that migrated from Southern Sudan [10,11], and subsequently, (vi) with Bantu speakers from West-Central Africa who reached the Great Lakes region ~2000 years ago [7,12–15]. Moreover, malaria imposed an important evolutionary

pressure well known for its effect on the genetic structure of affected populations, such as those that settled in the eBL belt.

Datasets representing African populations, such as those included in the 1000 Genomes Project [16], the African Genome Variation Project [17], the Tishkoff laboratory [18][11] and the H3Africa initiative [19][20], have provided an important baseline for genomic studies in Africa. However, due to the high genetic diversity among African populations, reference datasets should closely match populations in which specific scientific questions are explored. For example, the Nilotics in the Great Lakes region on Northern Uganda region, which experience high malaria intensity [21] and high eBL burden (S1 Table), have not been included in previous genomic studies [18].

To determine if the historical migrations described above (i-vi) and intense exposure to malaria have shaped the genetic composition in the eBL belt, we analyzed a new dataset of 945 Ghanaians and 568 Northern Ugandans in whom ~4.3 million single nucleotide polymorphisms (SNPs) were genotyped. These sub-Saharan Africa populations reside on opposite longitudes of the eBL belt (2400 miles apart) (Fig 1A), and are both exposed to high malaria pressure and have published evidence indicating a high eBL burden (S1 Table) [22].

Results

Study populations

Details of the study populations are given in S2 and S3 Tables. Briefly, the Ghanaian population included approximately 35 tribes, predominantly from the Kwa and Gur Niger-Congo language families (S2 Table). The Ugandan populations included approximately 17 tribes, predominantly of the Western Nilo-Saharan (Nilotic) language family (S3 Table). Because the Ugandan populations were recruited from opposite sides of the deep gorge of the East African Rift Valley, through which the Albertine Nile flows (Fig 1A) and this is a potential physical barrier to gene flow, we designated the populations descriptively as Uganda North West (UNW) for those recruited from the west side of the gorge and Uganda North Central (UNC) for those recruited from the east side of the gorge. We estimated the level of genetic relatedness of our dataset and excluded closely related individuals that may affect population-structure and natural selection analyses [23] (S1 Text and S2–S5 Figs).

Population structure and gene flow dynamics in the eBL belt

Population structure was evaluated using a Pan-African genome-wide dataset (PA dataset, Methods) that included 1.3M SNPs genotyped in 3,102 individuals, including 1,513 from the combined UNW, UNC, and National Cancer Institute (NCI) Ghana datasets, and 1,589 from 22 additional African populations [17,24,25] (S1 Table and S1 Fig). This Pan-African dataset is comprised of populations from five broad geographical regions: West Africa, West-Central Africa, Great Lakes Africa, Horn of Africa, and Southern Africa (Fig 1A and S1 Table). Specifically, the West African region includes Gambian and Ghanaian tribes [17], and the West-Central African region includes Nigerian tribes (Yoruba and Igbo). The Great Lakes African region includes our Northern Ugandan (UNW and UNC) populations and also Southwest Ugandan, Kenyan and Tanzanian populations.

Population structure and inferences of gene flow in West and West-Central Africa

Although our NCI Ghana set included individuals from approximately 35 tribes, ADMIXTURE results showed a homogeneous ancestry pattern (91% of the blue genomic ancestry

Fig 1C and S6, S7, S10 and S12 Figs), similar to the Ga-Adangbe tribe, with the blue genomic ancestry being predominant in West-Central Africa (Fig 1C and S6–S8 Figs). We observed similar ancestry composition of Ghanaians and Nigerians, who both share predominant West-Central Africa ancestry (blue). In accordance with their more Western location, Ghanaians shared a minor proportion of West African ancestry (red genomic ancestry in Fig 1C) related to Gambian tribes, while the Yoruba and Igbo shared a minor ancestry proportion (purple, Fig 1C) related to Eastern Bantu populations from the Great Lakes Africa region. This pattern of ancestry in Yoruba and Igbo has been seen in recent studies [17, 26–28]. Our Ghanaian population showed negligible Eurasian admixture (S9 Fig) with mean Eurasian ancestry of 0.4%.

Consistent with ADMIXTURE inferences, both GLOBETROTTER analysis and the three-population test (f_3 statistic) inferred episodes of gene flow from Gambian tribes, and also from Nilotics, to Ghana and Nigeria that occurred during the last 4000 years (Fig 2, S13 Fig and S4 Table). The pattern of genetic structure in Ghanaians and Nigerians, and the inferred episodes of gene flow into West-Central Africa show that historical cultural exchanges between West and West-Central Africa [8, 9] and migrations across the Sahel (historical events i-iii of the Introduction) involving populations from East Africa have shaped the genetic composition of West-Central African populations.

Population structure and inferences of gene flow in Uganda

The main feature of the genetic structure of Uganda shown by ADMIXTURE and PCA is the dichotomy between Northern Uganda populations, that show a predominantly Nilotic genomic ancestry (cyan ancestry in Figs 1 and 2 and S6–S9 Figs), and Southwest Uganda populations that have predominantly Eastern Bantu ancestry (purple ancestry, in Figs 1 and 2 and S6–S9 and S12 Figs). Within the predominantly Nilotic Northern Uganda populations, the UNW population is more homogeneous (93% Nilotic ancestry, Fig 1B and 1C and S7 Fig), while the UNC population is a mixture of Nilotic (64%) and Eastern Bantu genomic ancestry (Fig 1B and 1C and S7 and S12 Figs). Interestingly, Nilotic ancestry was detected in all Great Lakes African populations (Fig 1C and S7 Fig). In general, ADMIXTURE and PCA showed that the Great Lakes African region, which includes populations from Uganda, Kenya and Tanzania, was the most ancestry diverse region in sub-Saharan Africa (Fig 1B and 1C). Our Ugandan populations showed negligible Eurasian admixture (S9 Fig) with mean Eurasian ancestry of 0.02% in UNC and 0.015% in UNW.

GLOBETROTTER inferences suggest an episode of gene flow from West/West-Central Africa into UNW (849–936 years before present (YBP), 95% confidence interval, S13 Fig), although this was not confirmed by f_3 statistics (Fig 2 and S4 Table). In contrast to UNW, both f_3 (Fig 2 and S4 Table) and GLOBETROTTER (S13 Fig) consistently inferred several episodes of gene flow into the UNC and Southwest Uganda populations (Baganda, Barundi and Banyarwanda) from different sources: UNW (Nilotic), Southern Bantu, Horn of Africa (Cushitic), and also from West/West-Central African populations. GLOBETROTTER dates for these gene flow events (397–484 and 1499–2659 YBP, 95% confidence interval, S13 Fig) suggest two gene flow events that do not overlap with the inferred gene flow event into the UNW. We also inferred Nilotic-related (UNW and UNC) gene flow into Southwest Ugandan (Banyarwanda), Kenyan (Kikuyu and Kalenjin), Horn of Africa, West and West-Central African populations. Taken together, these results show that historical migrations (events iv-vi of the Introduction) of several human groups (Nilotic, Bantu and Cushitic) have shaped the current genetic composition in the Great Lakes region, and that Nilotic westward migration was accompanied by gene flow (historical event iii of the Introduction) (Fig 2 and S4 Table).

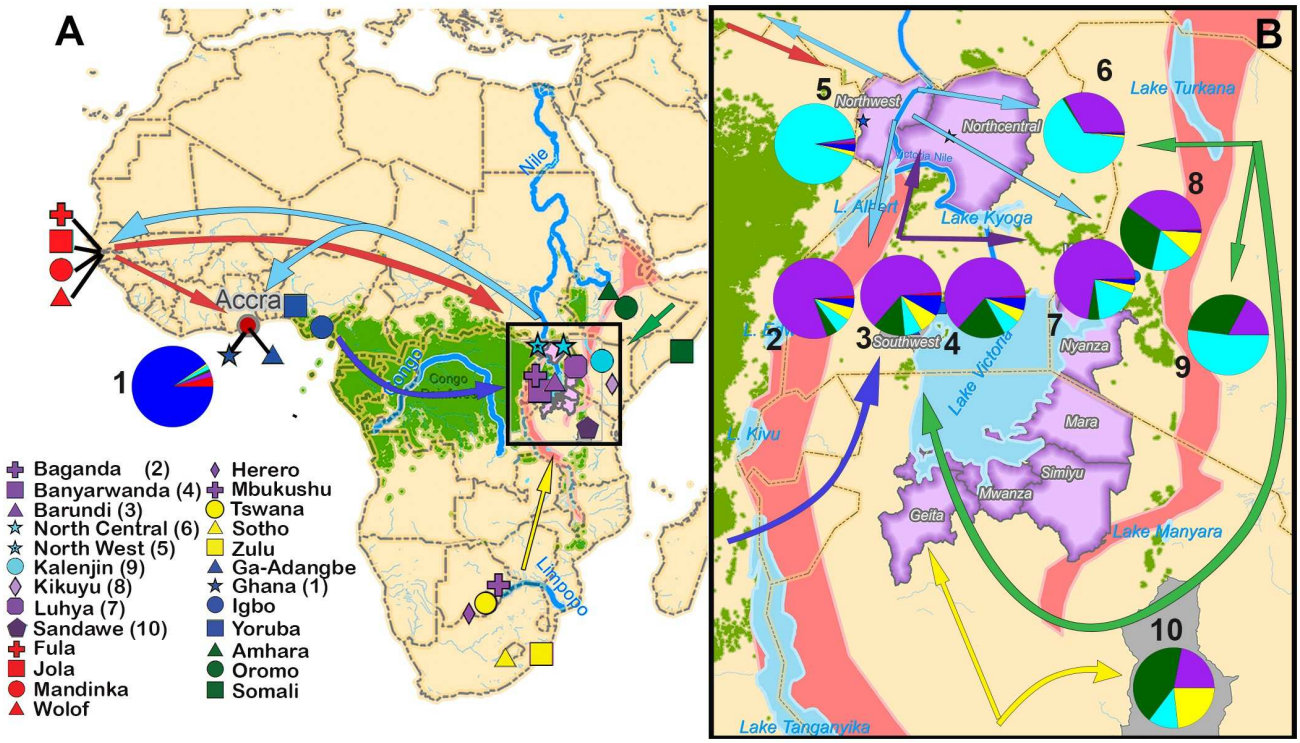

Fig 2. Populations movement routes in relation to major geographical barriers based on genetic inferences. (A) Map of Africa showing the geographical origin of the 22 previously reported Pan-African populations used herein, plus the three new populations from Ghana and the Uganda EMBLEM study (S1 Table). The arrows in the map indicate population gene flow based on significant f3 statistic tests (S2 Table). Colors of the arrows and the pie charts represent the ancestries inferred by ADMIXTURE (Fig 1C). Rather than shortest geographical distance, shapes of the arrows consider the major geographical barriers such as the Congo rainforest (light green), the East and West African Rift valley systems (pink), and the corresponding Rift Valley lakes and rivers (light blue). (B) The zoomed-out map shows the postulated dispersal routes in greater detail, as well as locations of the populations in the East African Rift Valley plateau sampled in previous studies (2 = Barundi, 3 = Banyarwanda, 4 = Baganda in Uganda; 7 = Luhya, 8 = Kalenjin, and 9 = Kikuyu in Kenya; and 10 = Sandawe in Tanzania), and the Uganda EMBLEM study populations [5 = Uganda North West (UNW), 6 = Uganda North Central (UNC)].

https://doi.org/10.1371/journal.pgen.1008027.g002

Natural selection in two distinct eBL belt populations subject to major malaria burden

While our studied populations (Ghana and Northern Uganda) share a high incidence of malaria and eBL burden (S1 Table and S1 Fig) [22], our population structure analyses showed that they have distinct patterns of genetic ancestry (Fig 1). In order to understand if they share common signals of natural selection despite their differential genetic history, we searched for genomic signatures of natural selection in Ghana and Northern Uganda populations. The eBL cases were excluded from this analysis to eliminate confounding of natural selection results with disease associations. We applied the population branch statistic (PBS) approach [29] to each of these as a focal population, using the Southern Bantu Sotho and Zulu populations as a sister group and Europeans as the reference population (S15 Fig and see Methods). We used Southern Bantu populations as a sister group because, after the Bantu expansion in the last 2000 years, they have occupied an area outside the eBL belt, where the climate is drier and cooler, and thus not conducive for malaria transmission [30], also supported by a low reported frequency of malaria-associated variants [31]. We compared the PBS outlier values (99.9th percentiles) against those generated by simulations of plausible neutral demographic models (Methods and S15–S17 Figs). In addition to the PBS statistic, we performed cross-population

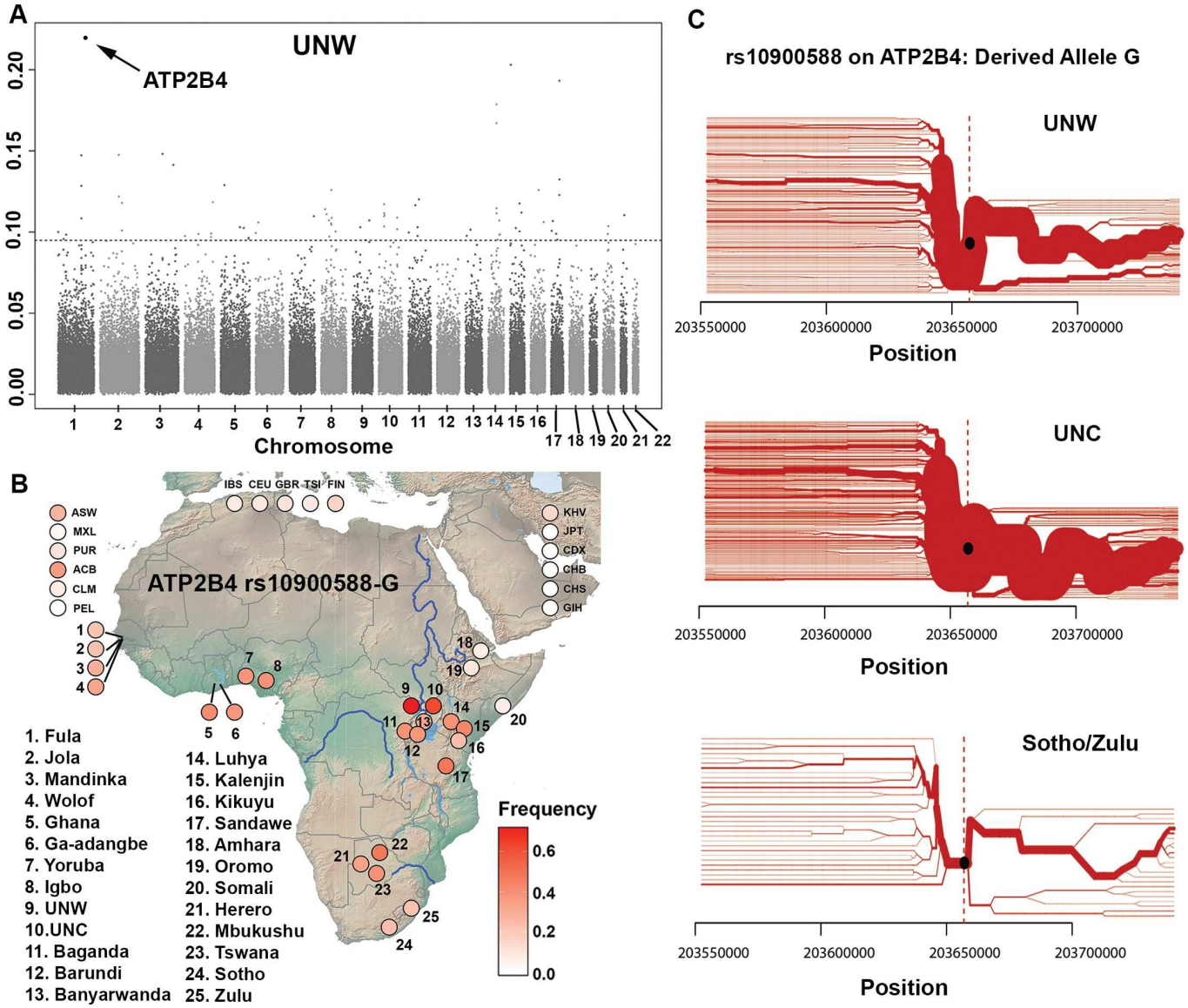

Fig 3. Malaria-driven natural selection analysis of rs10900588-G derived allele of gene *ATP2B4*. (A) Genome-wide population branch statistic (PBS). The mean PBS values (20 SNP windows) are represented by a Manhattan plot for Uganda North West (UNW). SNP rs10900588 on chromosome 1 showed the strongest signal of selection (See S15 Fig for LD graphs). (B) World-wide frequencies (red intensity) of the rs10900588-G derived allele, with map of the Pan-African plus UNW and UNC populations. (C) Haplotype bifurcation diagrams for the core haplotype at *ATP2B4* gene in Uganda North West (UNW), Uganda North Central (UNC), and Shoto/Zulu. ASW—Americans of African Ancestry in SW USA; MXL—Mexican Ancestry from Los Angeles USA; PUR—Puerto Ricans from Puerto Rico; ACB—African Caribbeans in Barbados; CLM—Colombians from Medellin, Colombia; PEL—Peruvians from Lima, Peru; KHV—Kinh in Ho Chi Minh City, Vietnam; JPT—Japanese in Tokyo, Japan; CDX—Chinese Dai in Xishuangbanna, China; CHB—Han Chinese in Beijing, China; CHS—Southern Han Chinese; GIH—Gujarati Indian from Houston, Texas.

https://doi.org/10.1371/journal.pgen.1008027.g003

haplotype-based approach (xpEHH) to identify genomic regions under positive selection. We report as candidate selection regions those that showed extreme signal in both PBS and the xpEHH approach (above the 99.9th percentiles for PBS and >2 for xpEHH).

We observed 14, 12 and 11 candidate genomic regions in the Ghanaian, UNW and UNC populations, respectively, (Fig 3, S16 Fig, and S5–S7 Tables), nominated by 32 index SNPs. While the Ghanaian sample yielded the largest number of candidate genomic regions, none of

them were significant in the demographic model performed (S5 Table). Of 32 candidate genomic regions, seven are found within/adjacent the same gene and shared between two populations: *RARB* found in Ghana and in UNC (different index SNPs), and six genomic regions within/adjacent to *KLHL20*, *ATP2B4*, *NIT2*, *TENM3*, *GPHN* and *HERC2*, are found in both UNW and UNC (five of six regions share the same index SNP) (S5–S7 Tables).

The extreme PBS values came from the genomic region at the *ATP2B4* gene in UNW (p-value = 0.0011) and UNC (p-value = 0.0021) (Fig 3A), but not in Ghana or other eBL belt populations evaluated (S5 and S8 Tables). Analysis using the xpEHH statistic (based on the pattern of extended haplotype homozygosity (EHH) between populations) corroborates the PBS signal in *ATP2B4* gene for both UNC and UNW (Fig 3 and S6–S8 Tables). *ATP2B4* encodes the plasma membrane Ca^{2+}-ATPase type 4 protein (PMCA4), the main calcium pump of the human erythrocyte [32]. Six SNPs within the genomic region in the *ATP2B4* gene (rs11240734-C, rs1541252-T, rs1419114-A, rs10900588-G, rs3851298-T, rs2228445-T) were detected as PBS outliers in both UNW and UNC. These six SNPs are located within two adjacent linkage disequilibrium (LD, $r^2 = 0.82$) blocks of 6 and 12 Kb (S18 Fig). The intronic SNP rs10900588-G derived allele exhibited the highest PBS values in both Northern Uganda populations (Fig 3A and 3B). This SNP is within a core haplotype observed with high frequency in both Northern Uganda populations (UNW and UNC) and much lower frequency in the South Bantu Sotho and Zulu populations (Fig 3C and S18 Fig). Consistently, the highest frequencies in Africa of the rs10900588-G were observed in UNW (0.72), followed by UNC (0.63), and the lowest frequencies in the Horn of Africa (0.064–0.096), followed by Fula (0.22), Zulu (0.23) and Sotho (0.24) (Fig 3C).

The other five shared signals of candidate selection in Northern Ugandans (UNW and UNC) for the following genes: *KLHL20* (p-values 0.0021 UNW, and 0.0043 UNC), *NIT2* (p-values 0.0027 UNW and 0.0035 UNC), *TENM3* (p-values 0.0022 UNW and 0.0041 UNC), *GPHN* (p-values 0.0036 UNW and 0.0018 UNC) and *HERC2* (p-values 0.0016 UNW and 0.0031 UNC). None of these genes have clear relationship with malaria pressure and are likely related to other selective pressures in Northern Uganda, which are not explored in the current study.

Discussion

Our study highlighted how the combined effects of demographic history and likely malaria-driven natural selection have shaped the genetic structure of populations in the eBL belt. We found evidence of gene flow events across the eBL belt in the last 3000 years, possibly related to regional migrations in Western Africa and major migrations involving Nilotic, Cushitic, and Bantu groups. Importantly, we identified for the first time in Africa a Northern Uganda-specific strong signal of malaria-driven selection in the *ATP2B4* gene.

Migrations in Africa and the genetic structure of eBL belt populations

Our results showed that historical migrations (denoted as i-vi in the Introduction) have left signals in the genome of eBL belt populations. The historical interactions of diverse linguistic groups (pastoral Nilotic, Cushitic and farming Bantu) along lush migratory corridors in the Lake Victoria basin plateau [33] is reflected in the current genomic composition of Uganda, Kenya and Tanzania populations (Figs 1 and 2 and S13 Fig). In the context of the six historical events highlighted in the Introduction (i-vi), the observed pattern of genetic structure is consistent with Nilotic dispersion southward into the Great Lakes region (event v, Figs 1C and 2, and S7 Fig) and westward across the Sahel region (event iii), which may have led to historical contacts with West African populations [11,26,34,35]. Our results showed that Nilotic

influence extends into the Great Lakes Africa region, and also to the Western African region, likely in the last 2000 years, as suggested by our GLOBETROTTER inferred dates (Fig 2,S7 and S13 Figs and S4 Table). The dichotomous pattern of ancestry between Northern Uganda (predominantly Nilotic) and Southern Uganda (predominantly Bantu) probably reflects the influence of the Nilotic migration into Northern Uganda, in contrast with the Bantu migration into Southern Uganda.

Our three dozen Ghanaian tribes showed high genetic homogeneity, but also evidence of gene flow from Gambian tribes in West Africa. Historically, the West and West-Central African regions have experienced extensive interactions between local kingdoms and tribes in the last 2000 years [8,9]. For Ghanaians, these interactions led some tribes to change their language due to social or economic motivation [7]. Local historical interactions such as these could explain the observed homogeneous genetic ancestry in Ghanaians. We inferred one gene flow event from Gambian tribes into Ghana and Yoruba about 1337–3022 YBP (S13 Fig). These inferred episodes of gene flow may be the signature of Mande migration into Ghana as part of trading networks [36], as well as of interactions of ancient populations along salt, gold, and slave trade routes [7,13].

Natural selection driven by malaria in the eBL belt

To search for natural selection driven by malaria in the eBL belt, we used eBL burden as an indicator of populations exposed to high sustained *falciparum* malaria transmission in the eBL belt (Fig 1, S1 Fig and S1 Table). By comparing the populations with the highest malaria pressures versus those with no malaria, we identified for the first time a candidate region for malaria-driven selection in the *ATP2B4* gene in African, specifically Northern Ugandan populations (Fig 3 and S6–S8 Tables). *ATP2B4* is ubiquitously expressed in human tissues, and encodes the plasma membrane Ca^{2+}-ATPase type 4 protein (PMCA4) [37], which is the most commonly expressed Ca^{2+} transporter in human erythrocytes [32]. We note that seven *ATP2B4* intronic SNPs (not present in our data) have been reported to be associated with multiple blood cell-related traits in African American, East Asian, European and Hispanic populations: mean corpuscular volume and hemoglobin concentration [38–41], lymphocyte counts, and red cell distribution width [38]. Furthermore, five of these seven *ATP2B4* SNPs (minor frequency alleles: rs10900585-G, rs2365860-C, rs10900589-A, rs2365858-G and rs4951074-A) were associated with resistance against severe *falciparum* malaria in Western African populations in Ghana and Gambia [42], and rs10900585 has been associated with reduced malarial placental infection and related maternal anemia in Ghana [43]. In an analysis of 11,890 cases of severe falciparum malaria and 17,441 controls from Africa, Asia and Oceania of 55 previously identified SNPs, rs10900585 was significantly associated with severe malaria over all African sites combined, and in the Ghanaian and Gambian samples [44]. Importantly, the protective minor alleles of these five SNPs above (not present in our data) are highly linked (mean $r^2 = 0.94$) with the minor allele (as defined in non-Nilotic populations) of our strongest signal of selection (*ATP2B4* rs10900588-G) in the Luhya population (LWK) from the 1000 Genomes Project.

While polymorphisms in the *ATP2B4* gene were described as protective against severe malaria in Ghana and Gambia [42], the outlier approach used in the present study did not identify *ATP2B4* as a candidate selection gene in Ghanaians and Nigerians (S8 Table). This result is in accordance with the absence of natural selection signals in the *ATP2B4* gene reported for previous studies using samples from Western Africa [17, 45–49]. The lack of concordance between association studies and natural selection analysis can be explained by the fact that the frequency of the protective haplotype observed in Ghanaians is sufficient to

identify significant disease association (a 6% difference between cases and controls across the protective haplotype) [42], but not sufficient to identify significant positive selection signal (an average 14% difference between West Central African and South African populations, compared to an average 45% difference between Northern Ugandan and South African populations, at rs10900588). In addition, when analyzing the *ATP2B4* association studies with cerebral malaria and severe malaria anemia in African, Asian and Oceanian populations, the Malaria Genomic Epidemiology Network [44] noted that the effect of the *ATP2B4* ancestral allele rs10900585-G on malaria might be heterogeneous across phenotypes and/or populations. The heterogeneity of effects may indicate presence of biological variation due to epistasis, gene-environment interactions, or that the analyzed SNP is in LD with an unknown causal allele associated with resistance to malaria. As LD patterns vary among populations, replication of the association would only be feasible if the causal SNP were genotyped.

The highest worldwide frequency of rs10900588-G allele and its related core haplotype observed in Northern Uganda populations (UNW and UNC, Fig 3 and S18 Fig) suggests a Northern Uganda- or Nilotic-specific selection in the *ATP2B4* gene, although the reasons for specificity are currently unclear to us. Consistent with this, our natural selection analyses using neighboring populations in Southern Uganda and Kenya did not identify signal of selection in the *ATP2B4* gene (S8 Table). The most likely explanation for this Northern Uganda-specific selection is that this region has historically experienced one of the highest levels of malaria infection worldwide (400–1,500 infectious mosquito bites per capita per year) [21]. A previous report has identified a signal of malaria-driven natural selection, at rs10900585 in the *ATP2B4* gene, by estimating the population-scaled selection coefficient in a time series of allele frequencies [50] in 92 ancient European samples from the Bronze Age (5000 bp) to the Post-Roman era [51], suggesting an ancient role of *ATP2B4* in malaria-driven selection.

The biological relationship between *ATP2B4* and malaria resistance is mediated by polymorphisms in *ATP2B4* changing PMCA4 structure or expression, which leads to a homeostatic disruption of intra-erythrocytic Ca^{2+} levels that are critical to the development of the *Plasmodium* parasite [42]. In an expression quantitative trait locus (eQTL) meta-analysis of whole blood gene expression [52], the allele rs10900588-G and linked SNPs were described as significant *cis*-eQTLs of *ATP2B4* (rs10900588-G with Z = -7.30, p-value = 2.91E-13, FDR = 0.00), i.e., the minor allele rs10900588-G is associated with significantly reduced *ATP2B4* expression. Recently, in a search for eQTLs enriched in human erythroblasts, Lessard *et al.* identified an erythroid-specific enhancer region just proximal to exon 2/alternate exon 1 of *ATP2B4* [53]. Lessard *et al.* demonstrated functional effects of the enhancer region through genome editing and *in vitro* cell culture, suggesting a Ca^{2+} homoeostasis defect as one possible pathway for the *ATP2B4* associations with malaria. The core haplotype we defined in the Northern Ugandan population extends from just proximal to exon 2/alternative exon 1 into intron 2/alternative exon 1. This haplotype overlaps with a minor *ATP2B4* haplotype in a European population (defined by the minor alleles in non-Nilotic populations of rs1541252, rs1541253, rs377342347, rs1419114, rs2228445, with mean $r^2 = 0.96$ with rs10900588 in the LWK population) that results in reduced erythrocyte PMC4A expression and reduced Ca^{2+} export [54]. Both Lessard *et al.* and Zámbó *et al.* have suggested mechanisms by which reduced Ca^{2+} export may be related to reductions in malaria risk: Lessard *et al.* suggests erythrocyte dehydration as a resistance factor, while Zámbó *et al.* suggests that reduced Ca^{2+} export into the invaginated extracellular membrane reduces Ca^{2+} concentration, which is required for *Pfa* maturation. Supporting the suggested mechanism, the most recent report [55] showed a significant association between low *falciparum* malaria parasitemia and the homozygous genotype for the *ATP2B4* rs1541255-G allele (not present in our data). Importantly, this allele is in perfect LD (R^2 and D' = 1) with our most important *ATP2B4* signal (rs10900588-G) in Kenya.

There are extensive reports in the literature regarding selection pressure driven by malaria in the *HBB*, *ABO*, *DARC* and *G6PD* genes [44, 56]. It should be noted that, in the present study, the tests used for the detection of positive selection are based on assumptions such as high differentiation between populations (PBS) and hard selective sweeps (xpEHH). Therefore, it is important to emphasize that this is not the case for *HBB* and *ABO*, that are evolving under a balancing selection regime [56], nor is this the case for *DARC*, that despite being under positive selection, is almost fixed and with low differentiation among African populations [57]. Also, as we did not examine the X chromosome, *G6PD*, found on the X chromosome, was not investigated in the present study.

Although malaria is the presumed major driver of natural selection in the eBL belt populations (S1 Table), we understand that other selection pressures, which were not investigated in our study, might be acting on our study populations. For example, we found significant signal of selection in Northern Ugandans for the *OCA2/HERC2* and *NIT2* genes (Fig 3, and S6 and S7 Tables). The first is significantly associated with skin, eyes and hair pigmentation [18] and the latter is a potential tumor suppressor [58].

Conclusions

After characterizing the genetic structure of the Ghanaian and Ugandan populations in the eBL belt, we showed that (i) historical interaction between West and West-Central Africa involved episodes of gene flow from West to West-Central Africa; (ii) the documented cultural interaction between the local kingdoms of West-Central Africa, specifically in Ghana, were accompanied by an homogenization of the gene pool of these populations, independently of their linguistic diversity; (iii) the pattern of genetic diversity of the eBL belt populations show the signature of migrations across the Sahel that include Nilotic expansion into West Africa; (iv) the genetic composition of Great Lakes African populations is the result of the interactions between Nilotics, Cushitics and Bantu groups in the last 3000 years; and, (v) the *ATP2B4* gene, which was previously associated with erythroid-related traits and malaria susceptibility, shows the signature of malaria-driven natural selection specific to Northern Uganda (UNW and UNC). These results provide important baseline genomic data to facilitate disease association studies, including of eBL, in eBL belt populations.

Methods

Ethics statement

Ethical approval for EMBLEM was obtained from the Uganda Virus Research Institute Research and Ethics Committee, the Uganda National Council for Science and Technology (H816), and the NCI Special Studies Institutional Review Boards (10-C-N133). The Ghana Prostate Health Survey was approved by the Noguchi Memorial Institute for Medical Research Institutional Review Board (001/01-02) and by the NCI SSIRB (02CN240). Participants in both the EMBLEM and Ghana Prostate Healthy Study gave informed written consent.

Ugandan and Ghanaian samples, genotyping and data curation

The NCI Ghana set included random samples of 964 healthy men from approximately 35 tribes (S2 Table) aged 50–74 years old enrolled for prostate cancer screening into the Prostate Healthy Survey [59]. The Ugandan samples were from 758 children aged 0–15 years old (including 197 eBL cases and 561 controls) from 13 tribes enrolled in the **E**pide**m**iology of **B**urkitt **L**ymphoma in **E**ast-African Children and **M**inors (EMBLEM) study in two regions of Northern Uganda (Uganda North West [UNW] and Uganda North Central [UNC]). The

healthy children were enrolled from 100 randomly selected villages in these regions (S3 Table) [60]. The samples were genotyped using the Illumina Infinium HumanOmni5-4v1 genotyping array in the Cancer Genomics Research Laboratory (CGR) at the National Cancer Institute (NCI); quality control was performed using PLINK 1.07 software [61] and in-house scripts [62].

Relatedness

We calculated the inbreeding (F) and the kinship coefficients (Φ_{ij}) using the PLINK 1.07 software [61] (S2 and S3 Figs). Following Kehdy et al. [24] a Φ_{ij} threshold ≥ 0.1 was used to create family networks (S2 and S3 Figs) and we excluded interactively individuals with the highest number of relatives, which allow us to reduce family structure, minimizing sample loss. Following this procedure, we created "unrelated" NCI Ghana and Ugandan datasets (S1 Table).

Merging genotyping data

We merged the NCI datasets (1,513 individuals with >48 tribal affiliations) with public African genome-wide datasets, creating a Pan-African dataset (PA dataset) of 1,287,642 SNPs for 3102 individuals, from 9 countries, and 11 ethnolinguistic groups in Sub-Saharan Africa (S1, S2 and S3 Tables). We also merged the PA dataset with all 1000 Genomes Project Phase 3 populations [24] creating the PA1KGP dataset, to test the extent of Eurasian admixture in the NCI datasets.

Population structure and demographic history

Since ADMIXTURE software [63] assumes independence among genetic markers, we used PLINK 1.07 to prune the SNPs in high linkage disequilibrium (LD) using a pairwise linkage disequilibrium maximum threshold of 0.4, a window size of 50, and a shift step of 10, creating the PA non-LD dataset with 727,834 SNPs. Then, we used the PA non-LD dataset to perform ADMIXTURE [63] and Principal Components Analysis (PCA) [64]. To verify possible sample size effects on ADMIXTURE and PCA analysis [65], we resampled the PA non-LD dataset to reach similar number of individuals for each studied population (S8 and S11 Figs).

We phased the PA dataset using SHAPEIT [66]. Using the phased dataset, we performed fineSTRUCTURE [67] analysis (10 million iterations of Markov chain Monte Carlo) to determine the genetically homogeneous groups and GLOBETROTTER [68] to infer historical admixture events.

We also estimated the $f3$ statistic to infer events of gene flow and their possible directions, as implemented in the software ADMIXTOOLS [69], for all possible combinations of three populations using the PA dataset. All $f3$ statistics with Z-score $\leq$ -3 were considered as highly significant evidence of gene flow. For the $f3$ statistic and GLOBETROTTER analysis of historical gene flow events, we described contributing ethnic groups or populations with the suffix "-like", representing present day surrogates of the real sources [67]. Masterscripts used for data curation and population structure analyses are available at the EPIGEN-Scientific Workflow (http://ldgh.com.br/scientificworkflow/, [62]).

Natural selection

To search for genomic footprints of selection in Ghana and Uganda, we explored allele frequency differentiation using Population Branch Statistic (PBS) using all the data, i.e., without LD pruning as done during the PCA and ADMIXTURE analysis [29], but excluding the eBL cases in Northern Uganda. PBS estimates were performed using NCI Ghanaians and Northern

Ugandan controls as study populations, the Southern Bantu populations (Sotho and Zulu) from the African Genome Variation Project [17] as a sister group, and the Europeans (CEU+TSI+FIN+GBR+IBS) from 1000 Genomes project [24] as reference population.

In addition to PBS, we performed Extended Haplotype Homozygosity (EHH) [70] analysis (SI) using the Cross-population Extended Haplotype Homozygosity (xpEHH) [71] in R package rehh v.2.0.2 [72]. To minimize spurious results of individual SNPs [73], all the selection analyses were performed on windows of 20 SNPs overlapping by 5 SNPs. For the density of SNPs used in the present study (~1,000,000), the average window size of 20 SNPs corresponds to an average ~ 50 Kb. We used ANNOVAR [74] to annotate SNPs found in candidate regions under selection. To consider a candidate region to be under selection, we adopted a conservative approach of filtering those regions that showed extreme signals in both PBS and xpEHH methods (S5–S7 Tables). For the intergenic natural selection signal, we represented the genetic distances from the closest genes (S5 and S6 Tables).

Simulations of the neutral coalescent model

Simulations were carried out using the demographic model [76] (S15 Fig), based on estimated divergence (thousands of years ago, kya) and effective population size (Ne) of African populations performed in Mallick *et al.* [75]. We used the Dinka population as a proxy for UNC and UNW, and the Luhya population as a proxy for Southern Bantu, with inferred divergence range of 9 and 25 kya (Mallick *et al.* high and low divergence inference), and current Dinka and Luhya Ne of $3x10^4$ and $3x10^4$, respectively [75]. We used the Yoruba population as a proxy of the Ghanaian population, and the estimated divergence from the Luhya of 5 and 10 kya and current Yoruba Ne of $7x10^4$. We also used the French population as a European proxy, 40 to 60 kya for an inferred divergence time and $3x10^4$ for current Ne. Considering that the study populations were involved in gene flow events, we introduced migration parameters between study populations and Southern Bantu considering the ancestry proportions inferred by ADMIXTURE (Fig 1C), as $4Nem_{ij}$, where 4Ne is the population effective size and m_{ij} the fraction of population *i* that is made up of migrants from population *j* (for more details see S15 Fig).

Additional Methods are presented in Supporting Information (S1 Text).

Supporting information

S1 Text. Additional information and methods.
(DOCX)

S1 Fig. All studied populations in relation to the endemic Burkitt lymphoma (eBL) belt as represented in the original paper by Haddow et al [31]. The eBL belt is shown in red shade and the incidence of eBL is denoted by the red color intensity.
(TIF)

S2 Fig. Kinship and inbreeding in the Uganda (EMBLEM) and Ghana datasets. (A) Kinship coefficients (Φ_{ij}) estimates by the probabilities of IBD = 0 estimates for all pairs of individuals. The colored dots are the theoretical relatedness degree probabilities of Φij and IBD = 0. (B) The distribution of individual inbreeding coefficients estimated for all individuals.
(TIF)

S3 Fig. Inbreeding in the EMBLEM Uganda sample separated by Burkitt lymphoma (Cases), pilot population controls (PPCs), matched population controls (MPCs) and health-center II controls (HCII).
(TIF)

S4 Fig. Representation of the virtual families (genomic inferences) in the Uganda dataset by complex networks. We represented in the same image the individual inbreeding coefficient and the pairwise kinship coefficient (Φ_{ij}) that represents the relatedness among the individuals. In this network, the nodes are the individuals and the edges are kinship relationships between individuals. Here, we linked only pairs of individuals with $\Phi_{ij} \geq 0.06$ (A) or ≥ 0.1 (B), which means we consider as related only individuals with relatedness $\geq$ third or second degree, respectively. The size of nodes is proportional to the absolute value of individual inbreeding and the shape of the node serves to signal whether inbreeding is positive (square) or negative (circle). The colors of the nodes represent the Uganda individual's tribe (S1A Table). We represented only the samples with proportion of identity by descent (Plink PI_HAT) > 0.05.
(TIF)

S5 Fig. Representation of the virtual families (genomic inferences) of Ghana dataset by complex networks with $\Phi_{ij} \geq 0.06$ (A) or ≥ 0.1 (B). In each network, the nodes are the individuals and the edges are kinship relationships between individuals. We represented only the samples with proportion of identity by descent (Plink PI_HAT) > 0.05.
(TIF)

S6 Fig. ADMIXTURE barplot representation of the individual ancestry proportions of the Pan-African populations. (Top) The proportions of individual ancestry values were calculated using ADMIXTURE unsupervised mode with the number of ancestral K = 2 to K = 15. (Bottom) ADMIXTURE cross-validation errors as a function of K.
(TIF)

S7 Fig. Mean ancestry composition inferred by ADMIXTURE for Ghanaian and Ugandan populations in relation to 19 Pan-African populations. The populations are: Ghana; Uganda North West; Uganda North Central; and three populations from Uganda South West (Baganda, Barundi, and Banyarwanda). The Pan-African populations are from West Africa (Fula, Mandika, Wolof, and Jola), West Central Africa (Ga-Adangbe, Yoruba, and Igbo), Southern Africa (Mbukushu, Herero, Tswana, Zulu, and Sotho), Horn of Africa (Amhara, Oromo, and Somali) and Great Lakes Africa (Luhya, Kalenjin, and Kikuyu).
(TIF)

S8 Fig. ADMIXTURE barplot representation of the individual ancestry proportions of the resampled PA non-LD dataset with similar number of individuals for each studied population.
(TIF)

S9 Fig. ADMIXTURE barplot representation of the individual ancestry proportion of the Pan-African populations combined with 1000 Genomes Phase 3 populations. We represented the unsupervised ADMIXTURE analysis with the number of ancestral clusters K = 10. This K captured the six African clusters represented in Fig 1C, and East Asian (pink), and South Asian (light green) Asian, European (black) and Native American (orange) ancestral clusters. ASW—Americans of African Ancestry in SW USA; MXL—Mexican Ancestry from Los Angeles USA; PUR—Puerto Ricans from Puerto Rico; ACB—African Caribbeans in Barbados; CLM—Colombians from Medellin, Colombia; PEL—Peruvians from Lima, Peru; KHV—Kinh in Ho Chi Minh City, Vietnam; JPT—Japanese in Tokyo, Japan; CDX—Chinese Dai in Xishuangbanna, China; CHB—Han Chinese in Beijing, China; CHS—Southern Han Chinese; GIH—Gujarati Indian from Houston, Texas.
(TIF)

S10 Fig. Principal component analysis (PCA) of the Pan-African populations. We compared the following PC combinations: PC1 vs PC2, PC3 vs PC4, PC1 vs PC3 and PC2 vs PC4.
(TIF)

S11 Fig. Principal component analysis of the resampled PA non-LD dataset with similar number of individuals for each studied population. We compared the following PC combinations: PC1 vs PC2, PC3 vs PC4, PC1 vs PC3 and PC2 vs PC4.
(TIF)

S12 Fig. Haplotype clustering analysis. (A) fineSTRUCTURE tree and (B) heatmap of the length of the chunks shared by individuals. Each row of the heatmap represent a copyvector of a recipient individual and each column represent the proportions of haplotypes that a donor shares with a recipient. Dark regions of the heatmap represent the long haplotype segments shared between individuals. Dark regions outside on the diagonal indicate more recent gene flow events. We highlight the inferred clusters using the colors of the ADMIXTURE ancestries.
(TIF)

S13 Fig. Admixture events inferred by GLOBETROTTER for West Central Africans (Ghana, Yoruba and Igbo) and Ugandan populations. Ancestry profiles and admixture dynamics were inferred using non-local donors. Donor populations were selected based on fineSTRUCTURE results. In the mixture model and event sources, the bars show the contribution of each African population to the recipient populations. The plot on the left represents the most likely estimated admixture dates inferred by GLOBETROTTER. The plot shows two admixture events, except in the West Africans and North West Uganda where only one admixture event was found. Inferred date(s) and 95% CIs are represented by the dots and horizontal lines in the graph. Bars corresponding to the event sources represent the inferred admixing sources for each estimated admixture event and the proportion of contribution of the African donor populations.
(TIF)

S14 Fig. Coancestry curves of GLOBETROTTER inferences for Ugandan and Ghanaian populations. These curves are informative for date estimation and the genetic composition of the sources of the admixture event. Each curve describes the probability to find two chunks of two donor populations along the genome of the target population. Curves with decreasing probability indicate that the two donor populations describe the genetic composition of one source population. Increasing probability indicates that the two donors could describe different sources. One date admixture (North West and Ghana) is characterized by a uniform curve that decreases or increases its probability. On the other hand, multiple date admixture (North Central) is characterized by a curve that changes its behavior, for example, from increasing (indicating different donor for the earlier event) to decreasing (both donor populations describe one source for the recent event).
(TIF)

S15 Fig. Neutral coalescent demographic model used in the PBS analysis. Ne = effective population size, kya = thousand years ago and *m* = migration rate. We used the migration rates following the current ancestry profile estimated by ADMIXTURE, as $4Nem_{ij}$, where 4Ne is the population effective size and m_{ij} the fraction of population *i* that is made up of migrants from population *j*. The simulated genotypes were obtained by the ms program [76] with the following command line: ms NTotalPop 10000 –s 1 –t 0.01 –I 3 NPop1 NPop2 NPop3 –eg 0.002 1 gPop1 at time1 –em 0.002 1 3 $NMigrants_{ij}$–em 0.002 3 1 $NMigrants_{ji}$ -ej 0.025 3 1 –en

0.06 1 gPop1 at time2 –en 0.06 1 gPop2 at time2 –ej 0.06 2 1 –en 0.1 1 gPop1 at time3.
(TIF)

S16 Fig. The mean PBS values (by 20 SNP windows) represented by the Manhattan plots for (A) GHANA, (B) Uganda North West (UNW) and (C) Uganda North Central (UNC). The dotted line demarcates the 99.9th percentile. The red dots represent genes that also have a selection signal by the xpEHH test (>2).
(TIF)

S17 Fig. *ATP2B4* PBS value observed against neutral distribution. PBS neutral values were generated by 10,000 simulations of plausible neutral demographic models (S15 Fig) for UNW and UNC populations respectively.
(TIF)

S18 Fig. Haploview LD table and core haplotype (rs10900588) of the ATP2B4 genomic region with the extreme PBS values (Fig 3A). We presented the LD table of the Uganda North West (UNW) population that showed the highest signal of selection.
(TIF)

S1 Table. Pan-African population samples used in the study and the malaria risk estimated by metrics of malaria prevalence and number of months in year with intense malaria transmission and eBL burden information for each population.
(XLSX)

S2 Table. Self-reported tribes of participants in the NCI Ghana Prostate Health Survey.
(XLSX)

S3 Table. Self-reported tribes of participants in the EMBLEM study in Northern Uganda.
(XLSX)

S4 Table. Admixture signal in Ghanaian, Ugandan, Kenyan and Horn of Africa populations using the three-population test (*f3* statistic). The three-population test (*f3*) statistic evaluates if the allele frequencies of a target population are intermediate of two sources which is interpreted as a result of admixture. Evidence of admixture is represented by negative *f3* values; significant evidence of admixture is inferred with a Z-score <-3. All populations evaluated in the global ADMIXTURE analysis were included. In addition, several populations were combined and evaluated as sources or target to assess broader groups (i.e., West_Central_Ghana, Banyarwand_Barundi). Only combinations that resulted in Z score ≤-3 are tabulated.
(XLSX)

S5 Table. Gene candidates for natural selection in the Ghana population based on the outliers PBS (99.9th percentile and p-value<0.05) and xpEHH (>2) tests. The SNP with the highest PBS value for each candidate gene is tabulated.
(XLSX)

S6 Table. Gene candidates for natural selection in the Uganda North West (UNW) population based on the outlier PBS (99.9th percentile and p-value<0.05) and xpEHH (>2) tests. The SNP with the highest PBS value for each candidate gene is tabulated.
(XLSX)

S7 Table. Gene candidates for natural selection in the Uganda North Central (UNC) population based on the outlier PBS (99.9th percentile and p-value<0.05) and xpEHH (>2) tests. The SNP with the highest PBS value for each candidate gene is tabulated.
(XLSX)

S8 Table. PBS and xpEHH test values for the *ATP2B4* gene in Ghana, Nigeria, Uganda and Kenya.
(XLSX)

Acknowledgments

We thank the study subjects for their participation. We thank Ms. Janet Lawler-Heavner at Westat Inc. (Rockville, MD, USA) and Mr. Erisa Sunday at the African Field Epidemiology Network (Kampala, Uganda) for managing the study. We thank Mr. Wilson Nyegenye at Uganda Bureau of Statistics (Kampala, Uganda) for survey support, and thank Ms. Laurie Buck, Dr. Carol Giffen, and Mr. Greg Rydzak at Information Management Services Inc. (Calverton, MD, USA) for preparing data analysis files, and Mr. Jeremy Lyman (IMS) for drawing the maps. We thank the staff of the NCI Cancer Genomics Research Laboratory for conducting genetic tests. We thank Garrett Hellenthal for help with haplotype-based methods.

Author Contributions

Conceptualization: Sam M. Mbulaiteye.

Data curation: Mateus H. Gouveia, Andrew W. Bergen, Victor Borda, Sam M. Mbulaiteye.

Formal analysis: Mateus H. Gouveia, Andrew W. Bergen, Victor Borda, Kelly Nunes, Thiago P. Leal.

Investigation: Mateus H. Gouveia, Andrew W. Bergen, Victor Borda, Kelly Nunes, Thiago P. Leal, Sam M. Mbulaiteye.

Methodology: Mateus H. Gouveia, Andrew W. Bergen, Victor Borda, Kelly Nunes, Thiago P. Leal, Martin D. Ogwang, Edward D. Yeboah, James E. Mensah, Tobias Kinyera, Isaac Otim, Hadijah Nabalende, Ismail D. Legason, Sununguko Wata Mpoloka, Gaonyadiwe George Mokone, Patrick Kerchan, Kishor Bhatia, Steven J. Reynolds, Richard B. Birtwum, Andrew A. Adjei, Yao Tettey, Evelyn Tay, Robert Hoover, Ruth M. Pfeiffer, Robert J. Biggar, James J. Goedert, Ludmila Prokunina-Olsson, Michael Dean, Meredith Yeager, M. Fernanda Lima-Costa, Ann W. Hsing, Sarah A. Tishkoff, Sam M. Mbulaiteye.

Project administration: Sam M. Mbulaiteye.

Software: Mateus H. Gouveia, Victor Borda, Kelly Nunes, Thiago P. Leal.

Supervision: Mateus H. Gouveia, Martin D. Ogwang, Sam M. Mbulaiteye.

Writing – original draft: Mateus H. Gouveia, Andrew W. Bergen, Victor Borda, Kelly Nunes, Thiago P. Leal, Ludmila Prokunina-Olsson, Eduardo Tarazona-Santos, Sam M. Mbulaiteye.

Writing – review & editing: Mateus H. Gouveia, Andrew W. Bergen, Victor Borda, Kelly Nunes, Ludmila Prokunina-Olsson, M. Fernanda Lima-Costa, Sarah A. Tishkoff, Stephen J. Chanock, Eduardo Tarazona-Santos, Sam M. Mbulaiteye.

References

1. Burkitt D. A children's cancer dependent on climatic factors. Nature. 1962; 194: 232–234.
2. Burkitt DP. Etiology of Burkitt's lymphoma–an alternative hypothesis to a vectored virus. J Natl Cancer Inst. 1969; 42: 19–28.
3. Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, Clark TG, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet. 2009; 41: 657–665. https://doi.org/10.1038/ng.388 PMID: 19465909

4. Legason ID, Pfeiffer RM, Udquim K-I, Bergen AW, Gouveia MH, Kirimunda S, et al. Evaluating the Causal Link Between Malaria Infection and Endemic Burkitt Lymphoma in Northern Uganda: A Mendelian Randomization Study. EBioMedicine. 2017; 25: 58–65. https://doi.org/10.1016/j.ebiom.2017.09.037 PMID: 29033373
5. Parkin DM, Sitas F, Chirenje M, Stein L, Abratt R, Wabinga H. Part I: Cancer in Indigenous Africans–burden, distribution, and trends. Lancet Oncol. 2008; 9: 683–692. https://doi.org/10.1016/S1470-2045(08)70175-X
6. Malaria Genomic Epidemiology Network. A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature. 2015; 526: 253–257. https://doi.org/10.1038/nature15390 PMID: 26416757
7. Ehret C, Posnansky M. The Archaeological and Linguistic Reconstruction of African History. Univ of California Press; 1982.
8. Gurstelle AW, Labiyi N, Agani S. Settlement history and chronology in the Savè area of central Bénin. Azania: Archaeological Research in Africa. Routledge; 2015; 50: 227–249.
9. Dakubu MEK. LINGUISTIC PRE-HISTORY AND HISTORICAL RECONSTRUCTION: THE GA-ADANGME MIGRATIONS. Transactions of the Historical Society of Ghana. Historical Society of Ghana; 1972; 13: 87–111.
10. Phillipson DW. The second millennium ad in sub-Saharan Africa. In: Phillipson DW, editor. African Archaeology. 3rd ed. Cambridge: Cambridge University Press; 2005. pp. 274–309.
11. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The genetic structure and history of Africans and African Americans. Science. 2009; 324: 1035–1044. https://doi.org/10.1126/science.1172257 PMID: 19407144
12. Ehret C. Bantu Expansions: Re-Envisioning a Central Problem of Early African History. Int J Afr Hist Stud. Boston University African Studies Center; 2001; 34: 5–41.
13. Scheinfeldt LB, Soi S, Tishkoff SA. Working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history. Proc Natl Acad Sci U S A. National Academy of Sciences; 2010; 107: 8931–8938.
14. Hanotte O, Bradley DG, Ochieng JW, Verjee Y, Hill EW, Rege JEO. African pastoralism: genetic imprints of origins and migrations. Science. 2002; 296: 336–339. https://doi.org/10.1126/science.1069878 PMID: 11951043
15. Grollemund R, Branford S, Bostoen K, Meade A, Venditti C, Pagel M. Bantu expansion shows that habitat alters the route and pace of human dispersals. Proc Natl Acad Sci U S A. 2015; 112: 13296–13301. https://doi.org/10.1073/pnas.1503793112 PMID: 26371302
16. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467: 1061–1073. https://doi.org/10.1038/nature09534 PMID: 20981092
17. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015; 517: 327–332. https://doi.org/10.1038/nature13997 PMID: 25470054
18. Crawford NG, Kelly DE, Hansen MEB, Beltrame MH, Fan S, Bowman SL, et al. Loci associated with skin pigmentation identified in African populations. Science. 2017; 358. https://doi.org/10.1126/science.aan8433 PMID: 29025994
19. Dandara C, Huzair F, Borda-Rodriguez A, Chirikure S, Okpechi I, Warnich L, et al. H3Africa and the African life sciences ecosystem: building sustainable innovation. OMICS. 2014; 18: 733–739. https://doi.org/10.1089/omi.2014.0145 PMID: 25454511
20. Retshabile G, Mlotshwa BC, Williams L, Mwesigwa S, Mboowa G, Huang Z, et al. Whole-Exome Sequencing Reveals Uncaptured Variation and Distinct Ancestry in the Southern African Population of Botswana. Am J Hum Genet. 2018; 102: 731–743. https://doi.org/10.1016/j.ajhg.2018.03.010 PMID: 29706352
21. Okello PE, Van Bortel W, Byaruhanga AM, Correwyn A, Roelants P, Talisuna A, et al. Variation in malaria transmission intensity in seven sites throughout Uganda. Am J Trop Med Hyg. 2006; 75: 219–225. PMID: 16896122
22. Emmanuel B, Kawira E, Ogwang MD, Wabinga H, Magatti J, Nkrumah F, et al. African Burkitt lymphoma: age-specific risk and correlations with malaria biomarkers. Am J Trop Med Hyg. 2011; 84: 397–401. https://doi.org/10.4269/ajtmh.2011.10-0450 PMID: 21363976
23. Kehdy FSG, Gouveia MH, Machado M, Magalhães WCS, Horimoto AR, Horta BL, et al. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. Proc Natl Acad Sci U S A. 2015; 112: 8696–8701. https://doi.org/10.1073/pnas.1504447112 PMID: 26124090

24. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015; 526: 68–74. https://doi.org/10.1038/nature15393 PMID: 26432245

25. Crawford NG, Kelly DE, Hansen MEB, Beltrame MH, Fan S, Bowman SL, et al. Loci associated with skin pigmentation identified in African populations. Science. 2017; Available: http://science.sciencemag.org/content/early/2017/10/11/science.aan8433.abstract

26. Triska P, Soares P, Patin E, Fernandes V, Cerny V, Pereira L. Extensive Admixture and Selective Pressure Across the Sahel Belt. Genome Biol Evol. 2015; 7: 3484–3495. https://doi.org/10.1093/gbe/evv236 PMID: 26614524

27. Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, et al. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science. 2017; 356: 543–546. https://doi.org/10.1126/science.aal1988 PMID: 28473590

28. Mulindwa J, Noyes HA, Ilboudo H, Nyangiri O, Koffi M, Mumba D, et al. Genomic evidence for population specific selection in Nilo-Saharan and Niger-Congo linguistic groups in Africa [Internet]. 2017. 10.1101/186700

29. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science. 2010; 329: 75–78. https://doi.org/10.1126/science.1190371 PMID: 20595611

30. Haddow AJ. AN IMPROVED MAP FOR THE STUDY OF BURKITT'S LYMPHOMA SYNDROME IN AFRICA. East Afr Med J. 1963; 40: 429–432.

31. Pule GD, Chimusa ER, Mnika K, Mhandire K, Kampira E, Dandara C, et al. Beta-globin gene haplotypes and selected Malaria-associated variants among black Southern African populations. Global Health, Epidemiology and Genomics. Cambridge University Press; 2017; 2. https://doi.org/10.1017/gheg.2017.14 PMID: 29868223

32. Stauffer TP, Guerini D, Carafoli E. Tissue distribution of the four gene products of the plasma membrane Ca2+ pump. A study using specific antibodies. J Biol Chem. 1995; 270: 12184–12190. PMID: 7538133

33. Newman JL. The Peopling of Africa: A Geographic Interpretation. Yale University Press; 1997.

34. Dobon B, Hassan HY, Laayouni H, Luisi P, Ricaño-Ponce I, Zhernakova A, et al. The genetics of East African populations: a Nilo-Saharan component in the African genetic landscape. Sci Rep. 2015; 5: 9996. https://doi.org/10.1038/srep09996 PMID: 26017457

35. Busby GB, Band G, Si Le Q, Jallow M, Bougama E, Mangano VD, et al. Admixture into and within sub-Saharan Africa. Elife. 2016; 5. https://doi.org/10.7554/eLife.15266 PMID: 27324836

36. Casey J. Holocene occupations of the forest and savanna. African archaeology. Blackwell Malden; 2005.

37. Adamo HP, Filoteo AG, Enyedi A, Penniston JT. Mutants in the putative nucleotide-binding region of the plasma membrane Ca(2+)-pump. A reduction in activity due to slow dephosphorylation. J Biol Chem. 1995; 270: 30111–30114. PMID: 8530416

38. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016; 167: 1415–1429.e19. https://doi.org/10.1016/j.cell.2016.10.042 PMID: 27863252

39. van Rooij FJA, Qayyum R, Smith AV, Zhou Y, Trompet S, Tanaka T, et al. Genome-wide Trans-ethnic Meta-analysis Identifies Seven Genetic Loci Influencing Erythrocyte Traits and a Role for RBPMS in Erythropoiesis. Am J Hum Genet. 2017; 100: 51–63. https://doi.org/10.1016/j.ajhg.2016.11.016 PMID: 28017375

40. Hodonsky CJ, Jain D, Schick UM, Morrison JV, Brown L, McHugh CP, et al. Genome-wide association study of red blood cell traits in Hispanics/Latinos: The Hispanic Community Health Study/Study of Latinos. PLoS Genet. 2017; 13: e1006760. https://doi.org/10.1371/journal.pgen.1006760 PMID: 28453575

41. Li J, Glessner JT, Zhang H, Hou C, Wei Z, Bradfield JP, et al. GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. Hum Mol Genet. 2013; 22: 1457–1464. https://doi.org/10.1093/hmg/dds534 PMID: 23263863

42. Timmann C, Thye T, Vens M, Evans J, May J, Ehmen C, et al. Genome-wide association study indicates two novel resistance loci for severe malaria. Nature. 2012; 489: 443–446. https://doi.org/10.1038/nature11334 PMID: 22895189

43. Bedu-Addo G, Meese S, Mockenhaupt FP. An ATP2B4 polymorphism protects against malaria in pregnancy. J Infect Dis. 2013; 207: 1600–1603. https://doi.org/10.1093/infdis/jit070 PMID: 23444010

44. Malaria Genomic Epidemiology Network, Malaria Genomic Epidemiology Network. Reappraisal of known malaria resistance loci in a large multicenter study. Nat Genet. 2014; 46: 1197–1204. https://doi.org/10.1038/ng.3107 PMID: 25261933

45. McManus KF, Taravella AM, Henn BM, Bustamante CD, Sikora M, Cornejo OE. Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. PLoS Genet. 2017; 13: e1006560. https://doi.org/10.1371/journal.pgen.1006560 PMID: 28282382

46. Ferreira Z, Hurle B, Rocha J, Seixas S. Differing evolutionary histories of WFDC8 (short-term balancing) in Europeans and SPINT4 (incomplete selective sweep) in Africans. Mol Biol Evol. 2011; 28: 2811–2822. https://doi.org/10.1093/molbev/msr106 PMID: 21536719

47. Sugden LA, Atkinson EG, Fischer AP, Rong S, Henn BM, Ramachandran S. Localization of adaptive variants in human genomes using averaged one-dependence estimation. Nat Commun. 2018; 9: 703. https://doi.org/10.1038/s41467-018-03100-7 PMID: 29459739

48. Reich D, Nalls MA, Kao WHL, Akylbekova EL, Tandon A, Patterson N, et al. Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. PLoS Genet. 2009; 5: e1000360. https://doi.org/10.1371/journal.pgen.1000360 PMID: 19180233

49. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002; 419: 832–837. https://doi.org/10.1038/nature01140 PMID: 12397357

50. Schraiber JG, Evans SN, Slatkin M. Bayesian Inference of Natural Selection from Allele Frequency Time Series. Genetics. 2016; 203: 493–511. https://doi.org/10.1534/genetics.116.187278 PMID: 27010022

51. Gelabert P, Olalde I, de-Dios T, Civit S, Lalueza-Fox C. Malaria was a weak selective force in ancient Europeans. Sci Rep. 2017; 7: 1377. https://doi.org/10.1038/s41598-017-01534-5 PMID: 28469196

52. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013; 45: 1238–1243. https://doi.org/10.1038/ng.2756 PMID: 24013639

53. Lessard S, Gatof ES, Beaudoin M, Schupp PG, Sher F, Ali A, et al. An erythroid-specific ATP2B4 enhancer mediates red blood cell hydration and malaria susceptibility. J Clin Invest. 2017; 127: 3065–3074. https://doi.org/10.1172/JCI94378 PMID: 28714864

54. Zámbó B, Várady G, Padányi R, Szabó E, Németh A, Langó T, et al. Decreased calcium pump expression in human erythrocytes is connected to a minor haplotype in the ATP2B4 gene. Cell Calcium. 2017; 65: 73–79. https://doi.org/10.1016/j.ceca.2017.02.001 PMID: 28216081

55. Ndila CM, Uyoga S, Macharia AW, Nyutu G, Peshu N, Ojal J, et al. Human candidate gene polymorphisms and risk of severe malaria in children in Kilifi, Kenya: a case-control association study. Lancet Haematol. 2018; 5: e333–e345. https://doi.org/10.1016/S2352-3026(18)30107-8 PMID: 30033078

56. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. Nat Rev Genet. 2014; 15: 379–393. https://doi.org/10.1038/nrg3734 PMID: 24776769

57. Hamblin MT, Di Rienzo A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet. 2000; 66: 1669–1679. https://doi.org/10.1086/302879 PMID: 10762551

58. Zheng B, Chai R, Yu X. Downregulation of NIT2 inhibits colon cancer cell proliferation and induces cell cycle arrest through the caspase-3 and PARP pathways. Int J Mol Med. 2015; 35: 1317–1322. https://doi.org/10.3892/ijmm.2015.2125 PMID: 25738796

59. Hsing AW, Yeboah E, Biritwum R, Tettey Y, De Marzo AM, Adjei A, et al. High prevalence of screen detected prostate cancer in West Africans: implications for racial disparity of prostate cancer. J Urol. 2014; 192: 730–735. https://doi.org/10.1016/j.juro.2014.04.017 PMID: 24747091

60. Maziarz M, Kinyera T, Otim I, Kagwa P, Nabalende H, Legason ID, et al. Age and geographic patterns of Plasmodium falciparum malaria infection in a representative sample of children living in Burkitt lymphoma-endemic areas of northern Uganda. Malar J. 2017; 16: 124. https://doi.org/10.1186/s12936-017-1778-z PMID: 28320389

61. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81: 559–575. https://doi.org/10.1086/519795 PMID: 17701901

62. Magalhães WCS, Araujo NM, Leal TP, Araujo GS, Viriato PJS, Kehdy FS, et al. EPIGEN-Brazil Initiative resources: a Latin American imputation panel and the Scientific Workflow. Genome Res. 2018; https://doi.org/10.1101/gr.225458.117 PMID: 29903722

63. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19: 1655–1664. https://doi.org/10.1101/gr.094052.109 PMID: 19648217

64. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38: 904–909. https://doi.org/10.1038/ng1847 PMID: 16862161

65. Lawson D, van Dorp L, Falush D. A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots [Internet]. bioRxiv. 2018. p. 066431. 10.1101/066431

66. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011; 9: 179–181. https://doi.org/10.1038/nmeth.1785 PMID: 22138821

67. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012; 8: e1002453. https://doi.org/10.1371/journal.pgen.1002453 PMID: 22291602

68. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science. 2014; 343: 747–751. https://doi.org/10.1126/science.1243518 PMID: 24531965

69. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012; 192: 1065–1093. https://doi.org/10.1534/genetics.112.145037 PMID: 22960212

70. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007; 449: 913–918. https://doi.org/10.1038/nature06250 PMID: 17943131

71. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006; 4: e72. https://doi.org/10.1371/journal.pbio.0040072 PMID: 16494531

72. Gautier M, Vitalis R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics. 2012; 28: 1176–1177. https://doi.org/10.1093/bioinformatics/bts115 PMID: 22402612

73. Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW. Limited evidence for classic selective sweeps in African populations. Genetics. 2012; 192: 1049–1064. https://doi.org/10.1534/genetics.112.144071 PMID: 22960214

74. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc. 2015; 10: 1556–1566. https://doi.org/10.1038/nprot.2015.105 PMID: 26379229

75. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016; 538: 201–206. https://doi.org/10.1038/nature18964 PMID: 27654912

76. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002; 18: 337–338. PMID: 11847089