Human Genetics (2019) 138:1123–1142
https://doi.org/10.1007/s00439-019-02045-1

ORIGINAL INVESTIGATION

Runs of homozygosity in sub-Saharan African populations provide insights into complex demographic histories

Francisco C. Ceballos[1] · Scott Hazelhurst[1,2] · Michèle Ramsay[1,3]

Received: 8 April 2019 / Accepted: 3 July 2019 / Published online: 16 July 2019

Abstract

The study of runs of homozygosity (ROH) can shed light on population demographic history and cultural practices. We present a fine-scale ROH analysis of 1679 individuals from 28 sub-Saharan African (SSA) populations along with 1384 individuals from 17 worldwide populations. Using high-density SNP coverage, we could accurately identify ROH > 300 kb using PLINK software. The genomic distribution of ROH was analysed through the identification of ROH islands and regions of heterozygosity (RHZ). The analyses showed a heterogeneous distribution of autozygosity across SSA, revealing complex demographic histories. They highlight differences between African groups and can differentiate the impact of consanguineous practices (e.g. among the Somali) from endogamy (e.g. among several Khoe and San groups). Homozygosity cold and hotspots were shown to harbour multiple protein coding genes. Studying ROH therefore not only sheds light on population history, but can also be used to study genetic variation related to adaptation and potentially to the health of extant populations.

Introduction

African human genetic diversity provides the ideal backdrop to reconstruct modern human origins, the genetic basis of adaptation to different environments and the development of more effective vaccines (Campbell and Tishkoff 2008). Boosted by efforts to increase the available data (H3 Africa Consortium et al. 2014; Gurdasani et al. 2015; Ramsay et al. 2016), studies on African populations have multiplied over the past years; however, this continent, and especially sub-Saharan Africa (SSA), is still being excluded in many genetics studies (Popejoy and Fullerton 2016; Martin et al. 2019). Due to the significant advances in genotyping and sampling of African populations, a study on runs of homozygosity (ROH) provides an interesting opportunity for a deep dive into the demographic history of Africans.

ROH are contiguous regions of the genome where an individual is homozygous (autozygous) across all sites (Ceballos et al. 2018b), and arise when two copies of an ancestral haplotype are brought together in an individual. The size of the ROH is inversely correlated with its age: longer ROH originate from recent common ancestors while shorter ROH come from distant ancestors because they have been broken down by recombination over many generations. Very short ROH, characterized by strong linkage disequilibrium (LD) among markers, are not always considered autozygous but nevertheless are due to the mating of distantly related individuals. Since their discovery in the late-1990s (Broman and Weber 1999) ROH have been found to be ubiquitous. We are all inbred to some degree and ROH capture this aspect of our demographic histories, with runs of homozygosity being the genomic footprint of the phenomenon known as pedigree collapse (Gunderson 1980). ROH are present in all populations, even in admixed or outbred populations and arise by two different processes: a limited effective population size (Ne) and by consanguineous unions (mating between relatives). Independently of how they were generated, ROH can be used to calculate the genomic inbreeding coefficient or F_{ROH} (McQuillan et al. 2008; Ceballos et al. 2018b).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00439-019-02045-1) contains supplementary material, which is available to authorized users.

✉ Francisco C. Ceballos
ceballoscamina@gmail.com

[1] Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[2] School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

[3] Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

F_{ROH} measures the actual proportion of the autosomal genome that is autozygous over and above a specific minimum length ROH threshold. When analysing ROH > 1.5 Mb, F_{ROH} correlates most strongly ($r = 0.86$) with the F obtained from an accurate six-generation pedigree (F_{PED}) (McQuillan et al. 2008). Using extended pedigrees of the royal European dynasties, with many complex inbreeding loops, it has been found that above the 10th generation the change in the inbreeding coefficient in less than 1% (Alvarez et al. 2009). Also, it has been found that individuals with no inbreeding loops in at least five generations (and probably 10) carried ROH up to 4 Mb in length but no longer (McQuillan et al. 2008). F_{ROH}, using a genomic approach, captures the total inbreeding coefficient of the individual independently of pedigree accuracy, or depth, within the resolution of the data available and the size of ROH that can be called (Ceballos et al. 2018a, b).

The ROH approach provides an opportunity to explore individual and demographic history (Ceballos et al. 2018b; Kirin et al. 2010) to understand the genetic architecture of traits and diseases (Joshi et al. 2015; McQuillan et al. 2012) and to study concepts in genome biology (Ceballos et al. 2018b; Gibson et al. 2006; Nothnagel et al. 2010; Pemberton et al. 2012). Different population histories give rise to divergent distributions of long and short ROH. These ROH have an uneven distribution along the genome, with a number of comparatively short regions with a high population-specific prevalence of ROH—known as ROH islands—on each chromosome, as well as cold-spots with a significant paucity of ROH (Curtis et al. 2008; Pemberton et al. 2012). ROH islands are prevalent in all populations and dominate the ROH in outbred groups; however, they are overshadowed by much larger ROH arising from recent pedigree loops that are randomly distributed across the genomes of some populations (Ceballos et al. 2018b). The origin of these islands is still a subject of debate. In some cases, the haplotypes segregating at high frequencies in the population may be due to positive selection; for example, a ROH island around the lactase persistence gene (LCT) on chromosome 2q21 was found in Europeans (Curtis et al. 2008). In addition, numerous genes that are targets of recent positive selection have been found in multiple ROH islands in populations around the globe (Pemberton et al. 2012). Another potential biological explanation is that ROH islands include small inversions that suppress recombination (Curtis et al. 2008).

Sub-Saharan Africa (SSA) is a sub-continent with a complex demographic history where a deep ROH analysis would provide interesting insights. Previous studies on ROH were hampered by small sample sizes and inadequate African population representation (Gibson et al. 2006; Kirin et al. 2010; Henn et al. 2011; Pemberton et al. 2012), genotype panels with low SNP coverage (Schlebusch et al. 2012; Patin et al. 2014; Ceballos et al. 2018b), non-optimized ROH calling conditions (Hollfelder et al. 2017) and in some cases preliminary and superficial ROH classification and analysis (Choudhury et al. 2017). In summary, these studies showed that African populations tend to have the lowest level of ROH (Pemberton et al. 2012; Choudhury et al. 2017; Ceballos et al. 2018b), especially for short ROH (Pemberton et al. 2012), and that within Africa the hunter-gatherer groups have a higher level of ROH in comparison to pastoralist groups (Henn et al. 2011; Schlebusch et al. 2012). Just one population, the Hadza (an isolated hunter-gatherer group in Tanzania), was found to have a ROH level similar to the most isolated population from Oceania and South America (Ceballos et al. 2018b).

The objective of this study was to perform fine-scale analysis of the ROH distribution in SSA, in a world context, in order to learn more about the demographic history of the continent and its populations. Public data from the African Genome Variation Project (AGVP) (Gurdasani et al. 2015), the 1000 Genomes Project (KGP) (Sudmant et al. 2015) and Schlebusch et al. (2012) were analysed and included 1679 individuals from 28 SSA populations (Fig. 1) and 1384 individuals from 17 worldwide populations. Even though the focus of this paper is SSA populations, we refer to other populations, present in the KGP dataset, to provide a global perspective and to highlight some key points. By analysing the proportion of the autosomal genome that is in ROH and deconstructing probable patterns of inbreeding, we present interpretations for the demographic histories of different SSA populations.

Results

Comparison of sum of ROH across different ROH sizes across African and world populations

Mean total lengths (sum of ROH) of different ROH length classes are plotted in Fig. 2. Figure 2a shows different scenarios for short (< 1 Mb) and long (> 4 Mb) ROH within African populations: short ROH, unlike the long ones, display differences between major regions and commonality among those within regions. The populations with the highest average sum of short ROH are from the Horn of Africa (Amhara, Oromo, Somali). Populations from Western Africa, Gulf of Guinea, Eastern Africa and Southern Africa, in this order and with slight differences, have intermediate levels of short ROH, and the recently admixed Coloured populations from South Africa have the lowest levels of short ROH. However, when long ROH (> 4 Mb) are considered, there are no apparent population similarities or geographic clustering. Three populations, Wolof and Fula, from western Africa, and Somali from

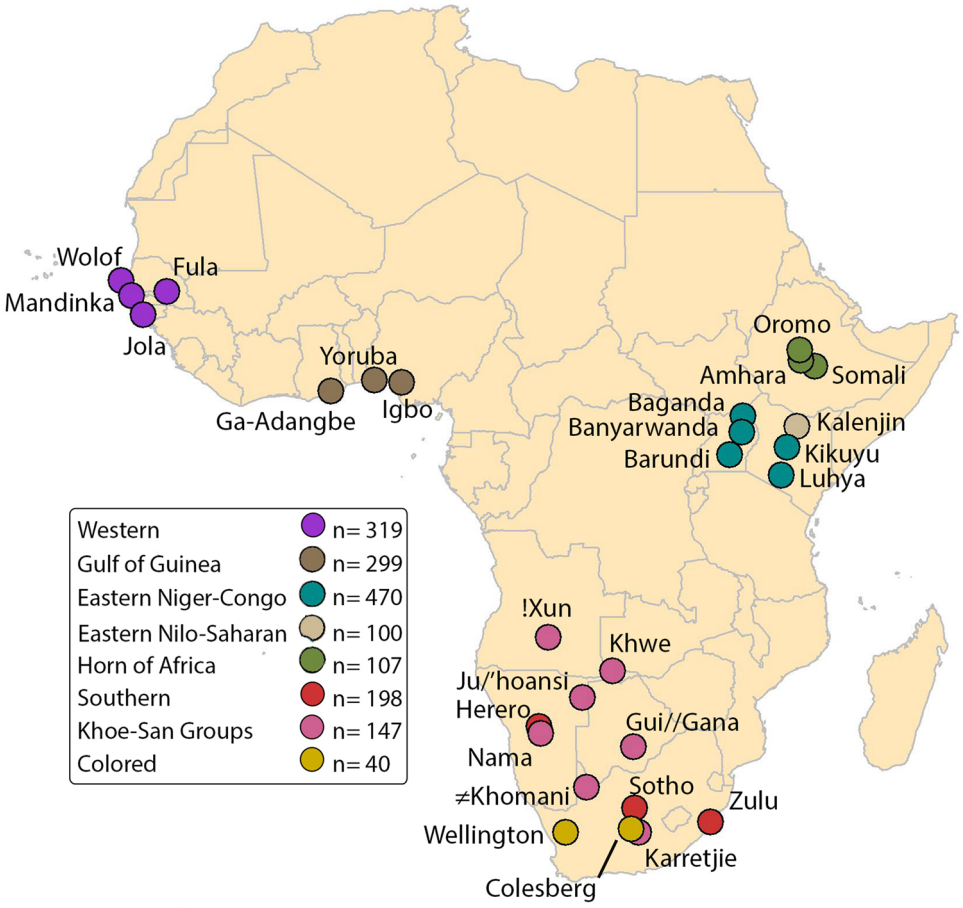

Fig. 1 Sub-Saharan African populations included in the study: 28 African populations in total including 16 from the African Genome Variation Project (AGVP), 2 from the 1000 Genomes Project (KGP) and 10 from Schlebusch et al. (2012). Populations were organized in eight groups according to their geographic, linguistic and/or admixture origins. Western Africa (shown in deep purple), Gulf of Guinea (shown in brown), Eastern Africa Niger–Congo populations (shown in light blue), Eastern Africa Nilo-Saharan population (shown in wheat), Horn of Africa (shown in dark green), Southern Africa (shown in red), Khoe and San populations (shown in pink) and Coloured admixed populations (shown in yellow). The number of individuals from each group is shown in Table 1. The term *Khoe–San* is often used in the literature, but is regarded by some as offensive as it conflates two distinct groups. The impact of colonialism had a very traumatic effect on population size and structure. We use the phrase *Khoe and San* to describe people who have either Khoe and/or San ancestry as a neutral term to describe people who live in similar regions and have had some shared history in the last centuries

the Horn of Africa, present the largest mean total length. Differences between long and short ROH are shown for populations around the world in Fig. 2b. African populations have the smallest mean total length of ROH, when considering short ROH, but for long ROH, only some African populations like the Wolof, Fula and Somali have mean total lengths larger than most of the KGP populations. An indigenous, but partially admixed population from Lima, Peru (PEL), had the largest mean total ROH length for both short and long ROH. Medium-size ROH (ROH between 1 and 4 Mb) (Fig. 2) reveals smaller differences. At a population level, the Khoe and San groups like Ju/'hoansi, !Xun and Khwe, have a higher mean total length for ROH from 2 to 8 Mb.

Violin plots: exploratory data analysis

The distribution of sum of ROH in two size categories (< 1.5 and > 1.5 Mb) is represented across global populations in Fig. 3 while the mean and median lengths of the ROH are given in Table 1. Geographic stratification is observed for mean sum of ROH < 1.5 Mb (Fig. 3a): SSA populations have the lowest medians (statistically significant using Whitney–Wilcoxon non-parametrical test Supplementary Figs. 1A and 2A). Within the continent, populations from the Horn of Africa have a significantly higher sum of ROH. Figure 3a and Table 1 show homogeneity among regions other than the Khoe–San and the Coloured populations. The kurtosis and skewness of the violin plots provide additional information. In general, the populations

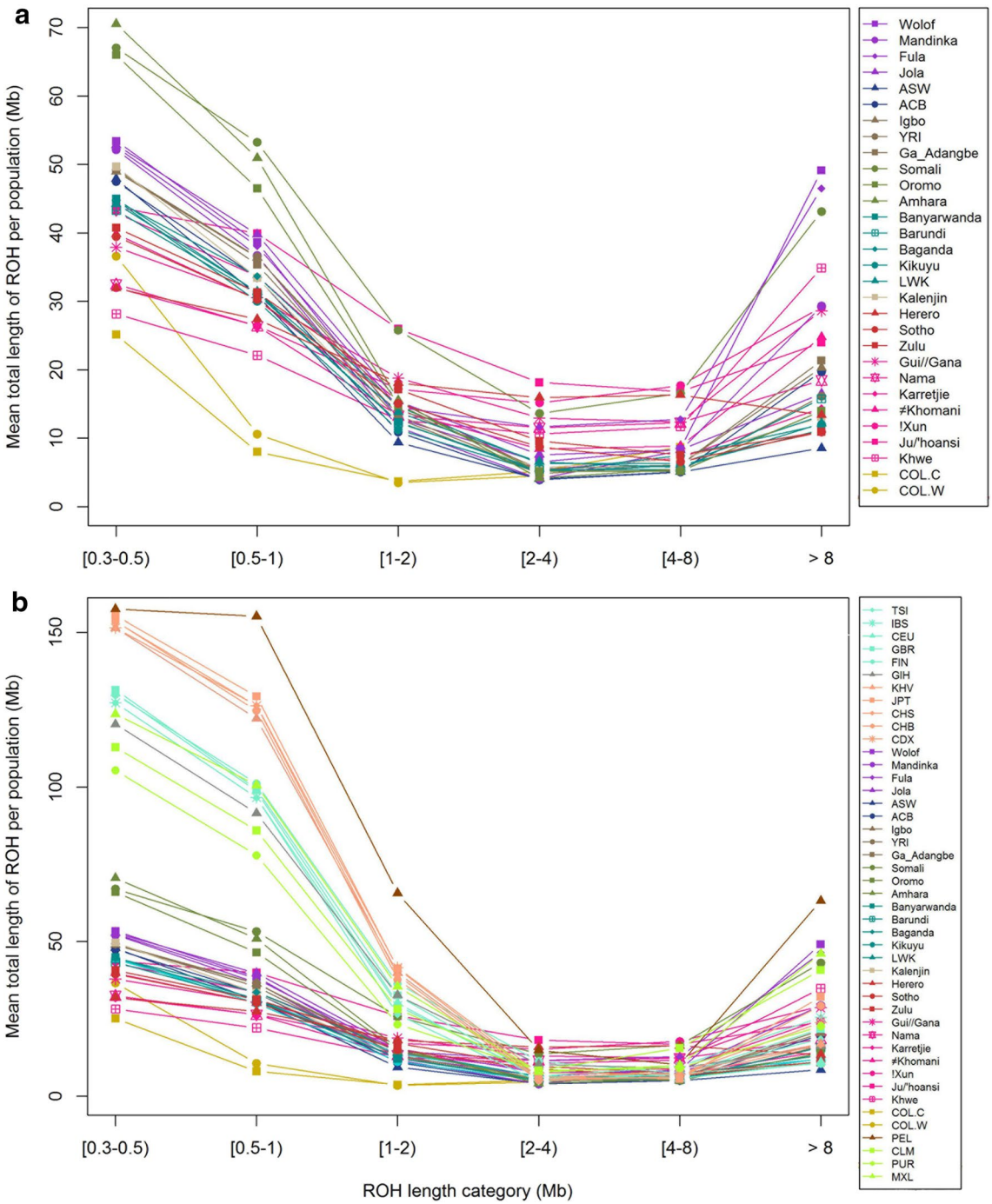

Fig. 2 Mean total length of ROH (Mb) over six classes of ROH tract lengths. ROH classes: $0.3 \leq \text{ROH} < 0.5$ Mb, $0.5 \leq \text{ROH} < 1$ Mb, $1 \leq \text{ROH} < 2$ Mb, $2 \leq \text{ROH} < 4$ Mb, $4 \leq \text{ROH} < 8$ Mb and $\text{ROH} \geq 8$ Mb. **a** Sub-Saharan African populations and admixed populations with African ancestry (ASW and ACB, shown in dark blue). Western Africa (shown in deep purple), Gulf of Guinea (shown in brown), Eastern Africa Niger–Congo populations (shown in light blue), Eastern Africa Nilo-Saharan population (shown in wheat), Horn of Africa (shown in dark green), Southern Africa (shown in red), Khoe and San populations (shown in pink) and Coloured admixed populations (shown in yellow). **b** All populations from the KGP, AGVP and Schlebusch et al. (2012). European populations are shown in aquamarine, Southern Asian population (GIH) is shown in grey, Eastern Asian populations are shown in light salmon, South America population (PEL) is shown in dark orange, admixed Hispanic-American populations are shown in light green. For KGP population identifiers see "Materials and methods" section

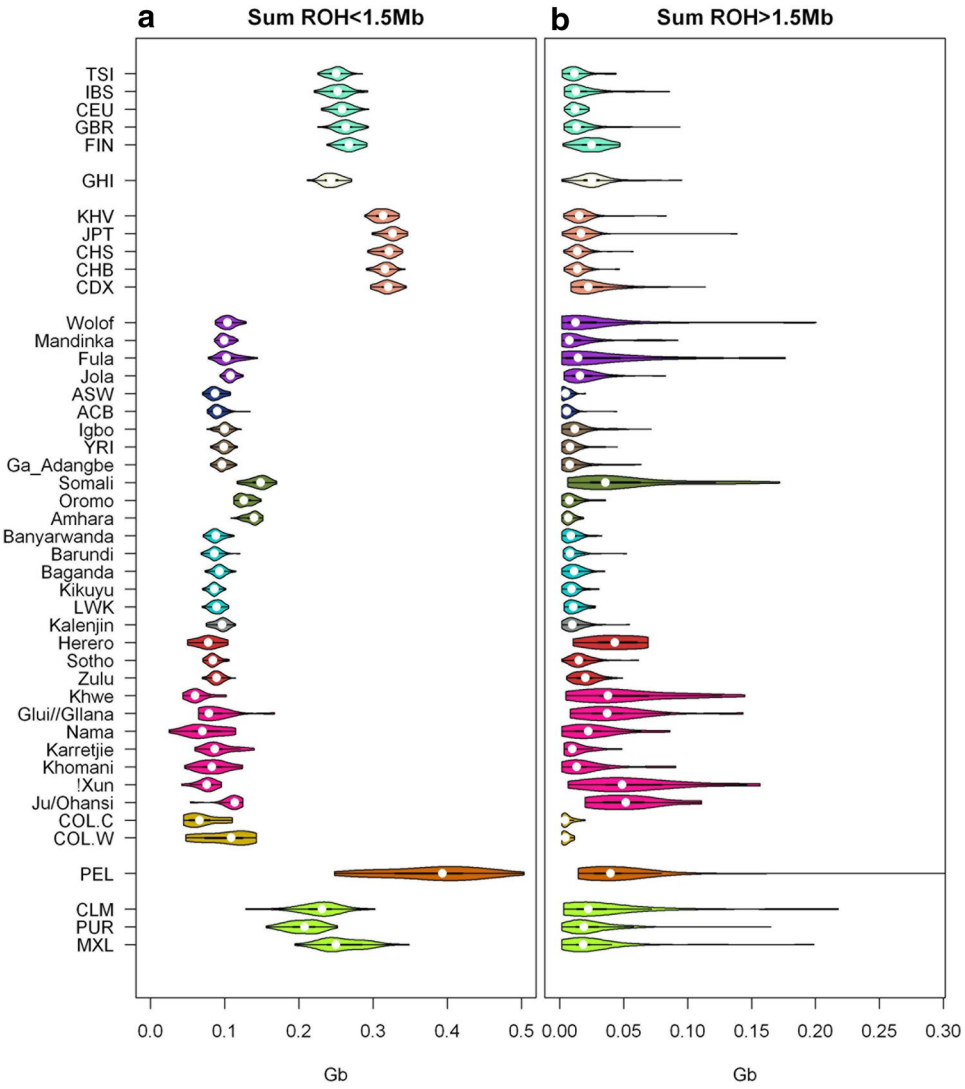

Fig. 3 Violin plots showing the distribution of ROH within populations for the mean total sum of ROH shorter than 1.5 Mb (**a**) and mean total length of ROH longer than 1.5 M (white dots) (**b**). The colours are coded according to the legend of Fig. 1. For KGP population identifiers see "Materials and methods" section

are relatively homogeneous, with very short tails and an almost normal distribution; however, the Khoe and San, Coloured and American populations present more variability with pronounced kurtosis (like in the Nama) along with positive and negative skewness (like in the Gui//Gana or Ju/'hoansi, respectively).

Distribution shapes are different for the sum of ROH > 1.5 Mb (Fig. 3b) and there is a greater variability in skewness and kurtosis across populations within and outside SSA. Wolof, and Somali show especially strong positive skewness, and just two populations outside SSA: PEL and CLM have longer tails (Table 1). Khoe and San populations form a heterogeneous group, but also show strong skewness and kurtosis; indeed two populations with the highest total sum of ROH are Khoe and San: the !Xun population from Angola and the Ju/`hoansi from Namibia (Table 1, Supplementary Figs. 1B and 2B).

Inbreeding coefficient from ROH: F_{ROH}

The genomic inbreeding coefficient from ROH (F_{ROH}) was obtained as the total sum of ROH > 1.5 Mb divided by the total length of the autosomal genome. Table 1 shows the mean F_{ROH}, the max F_{ROH}, the number and proportion (in %) of individuals with an F_{ROH} higher than second cousin ($F_{PED} = 0.015$). The highest average F_{ROH} among all populations was found in some of the Khoe and San populations, !Xun, Ju/'hoansi, Khwe and Gui//Gana, who have an average F_{ROH} ranging between 0.0204 and 0.0151 and the Somali population in the Horn of Africa ($F_{ROH} = 0.018$). Besides these the PEL population and the Gui//Gana Khoe–San have mean F_{ROH} higher than the equivalent of a second cousin kinship calculated from a pedigree. Within SSA, only the Wolof from Western Africa had individuals with genomic inbreeding coefficients higher than the inbreeding coefficient

Table 1 Sum of ROH (above and below 1.5 Mb) and summary statistics for the inbreeding coefficient calculated from ROH (F_{ROH}) across global regions and according to population

Population	*n*	Total sum ROH < 1.5 (Mb)		Total sum ROH > 1.5 (Mb)		Mean F_{ROH}		Max F_{ROH}	$F_{ROH} > 2C$
		Mean	SD	Mean	SD	Mean	SD		
Western Africa									
Wolof	78	104.90	9.74	27.07	40.92	0.0094	0.014	0.0696	20.5
Fula	74	105.09	13.76	33.84	42.75	0.0117	0.015	0.0612	29.7
Mandinka	88	100.44	7.30	15.12	19.34	0.0052	0.007	0.0321	8.0
Jola	79	107.98	6.82	18.87	12.47	0.0065	0.004	0.0287	7.6
Gulf of Guinea									
YRI	100	99.01	7.59	9.55	7.47	0.0033	0.003	0.0157	1.0
Ga_Adangbe	100	97.55	7.80	11.88	12.13	0.0041	0.004	0.0221	6.0
Igbo	99	99.67	7.97	14.33	11.97	0.0050	0.004	0.0248	7.1
Mix. Afr-Amer									
ACB	72	91.90	9.63	6.55	5.87	0.0023	0.002	0.0154	1.4
ASW	49	88.14	9.10	5.39	3.80	0.0019	0.001	0.0069	0.0
Horn of Africa									
Amhara	42	137.00	9.36	7.55	4.06	0.0026	0.001	0.0065	0.0
Oromo	26	127.30	9.75	9.70	7.89	0.0034	0.003	0.0124	0.0
Somali	39	146.03	12.59	52.28	42.80	0.0181	0.015	0.0597	48.7
Eastern Africa Niger–Congo									
Baganda	100	93.01	8.02	12.09	6.83	0.0042	0.002	0.0122	0.0
Banyarwanda	100	88.65	8.71	9.84	6.14	0.0034	0.002	0.0115	0.0
Barundi	97	86.60	8.43	9.88	6.58	0.0034	0.002	0.0181	1.0
Kikuyu	99	85.98	6.54	9.76	5.57	0.0034	0.002	0.0106	0.0
LWK	74	89.46	7.84	11.9	6.08	0.0041	0.002	0.0096	0.0
Eastern Africa Nilo-Saharan									
Kalenjin	100	95.27	9.11	11.38	7.87	0.0039	0.003	0.0189	1.0
Southern Africa									
Herero	12	77.40	16.07	43.25	19.63	0.0150	0.007	0.0239	50.0
Sotho	86	84.83	7.92	16.75	8.84	0.0058	0.003	0.0214	2.3
Zulu	100	89.17	8.01	20.51	8.45	0.0071	0.003	0.0170	4.0
Africa Khoe and San									
Ju/'hoansi	18	109.66	15.13	53.00	26.47	0.0184	0.009	0.0384	38.9
!Xun	19	75.59	12.70	58.86	38.66	0.0204	0.013	0.0543	73.7
Gui//Gana	15	87.22	25.72	42.85	32.28	0.0151	0.011	0.0497	60.0
ǂKhomani	39	84.08	18.92	22.22	22.63	0.0077	0.008	0.0314	15.4
Nama	20	73.91	24.63	25.92	21.38	0.0090	0.007	0.0298	20.0
Khwe	16	62.93	14.23	51.58	38.42	0.0179	0.013	0.0502	62.5
Karretjie	20	91.92	19.21	12.99	11.17	0.0045	0.003	0.0384	35.0
Africa Coloured									
Wellington	20	101.04	30.90	5.14	3.08	0.0011	0.001	0.0040	0.0
Colesberg	20	69.75	21.75	6.00	4.67	0.0021	0.002	0.0068	0.0
Europe									
CEU	95	259.12	12.36	12.78	4.88	0.0044	0.002	0.0079	0.0
FIN	97	267.43	12.27	25.49	10.78	0.0088	0.004	0.0163	16.5
GBR	91	263.46	13.06	16.55	12.41	0.0057	0.004	0.0326	5.5
IBS	99	253.15	14.65	18.09	14.96	0.0063	0.005	0.0298	9.1
TSI	92	250.29	11.04	12.94	8.96	0.0045	0.003	0.0153	4.3
Southern Asia									
GIH	95	244.41	12.30	26.50	13.77	0.0092	0.005	0.0331	13.7

Table 1 (continued)

Population	*n*	Total sum ROH < 1.5 (Mb)		Total sum ROH > 1.5 (Mb)		Mean F_{ROH}		Max F_{ROH}	$F_{ROH} > 2C$
		Mean	SD	Mean	SD	Mean	SD		
Eastern Asia									
CDX	83	319.44	11.45	27.35	18.01	0.0095	0.006	0.0396	20.5
CHB	98	316.12	10.07	14.37	6.92	0.0050	0.002	0.0161	2.0
CHS	86	318.98	11.23	15.25	7.40	0.0053	0.003	0.0199	2.3
KHV	96	313.42	11.04	17.10	10.39	0.0059	0.004	0.0289	4.2
JPT	96	326.15	11.05	17.51	13.76	0.0061	0.005	0.0481	1.0
South America									
PEL	50	378.33	64.76	46.54	54.82	0.0162	0.019	0.1400	58.0
Mix. Hispanic-American									
CLM	65	226.98	30.53	38.40	47.31	0.0133	0.016	0.0756	32.3
PUR	72	206.55	21.58	24.09	22.11	0.0084	0.008	0.0573	19.4
MXL	47	259.54	30.68	25.73	32.00	0.0089	0.011	0.0689	17.0

For KGP population identifiers see "Materials and methods" section

n number of individuals, *N ROH > 1.5* number of ROH > 1.5 Mb, *SD* standard deviation, *FROH > 2C* percentage of the population with an FROH higher than 0.015. Three letter population abbreviation are provided in the text

(F_{PED}) from a first-cousin union. Supplementary Fig. 3 plots the number of ROH (longer than 1.5 Mb) and the total sum of ROH > 1.5 Mb for each SSA individual from each population, and shows conservative limits for second- and first-cousin inbreeding coefficient as expected values of the proportion of the genome in autozygosity.

Discriminating between different sources of autozygosity: understanding population demographic history

F_{ROH} denotes the total inbreeding coefficient, but cannot discriminate whether the autozygosity was generated as a result of cultural practices favouring related unions, or because of a low effective population size and genetic drift.

In Fig. 4, the mean systematic inbreeding coefficient (F_{IS}) per population is plotted against the mean F_{ROH} for the same population. Three different regions can be considered in this plot, delineated by the diagonal, where $F_{IS} = F_{ROH}$, and the horizontal line $F_{IS} = 0$. (1) Populations close to the diagonal line, like the Somali, have a strong component of systematic inbreeding or F_{IS}, which means that the total inbreeding coefficient, F_{IT}, of this population is mainly produced by a deviation from panmixia, in other words, consanguinity. (2) Panmictic inbreeding, caused by genetic drift will be more relevant in populations positioned close to the line $F_{IS} = 0$; consanguinity, even though still present in the Fula and Wolof, is less relevant in these populations than for the Somali. (3) Low N_e, isolation and genetic drift become very relevant when populations have negative F_{IS}. Under this scenario of avoidance of consanguinity and excess of heterozygotes (compared to that expected under Hardy–Weinberg (H–W) proportions), the total inbreeding coefficient of populations like PEL, Khwe, Ju/'hoansi, !Xun or Herero will be provoked by genetic isolation and genetic drift: strong F_{ST}. Finding populations in the last region to be considered (where $F_{IS} > F_{ROH}$) does not make much sense under an inbreeding context and according to Wright *F* statistic. If a population presents with a larger F_{IS} than F_{ROH}, other phenomena must be taken into account. Besides inbreeding, natural selection pressure and the Wahlund effect can increase F_{IS}; however, natural selection is an evolutionary force that can change F_{IS} locally in specific genome regions, but not at a whole genome level. The only explanation is the Wahlund effect: a deficiency of heterozygotes and excess of homozygotes generated when subpopulations with different allele frequencies are lumped together (Hartl and Clark 2007). According to this explanation, the Karretjie, and particularly the Gui//Gana populations in Fig. 4 may indeed be the mixture of at least two different subpopulations with different alleles frequencies. This effect is better shown and described in Supplementary Fig. 4.

Genomic distribution of runs of homozygosity

ROH are not randomly distributed across the genome and there are regions with a high prevalence of ROH or complete absence (Ceballos et al. 2018b; Nothnagel et al. 2010; Pemberton et al. 2012). ROH islands, genomic regions with high prevalence of ROH, or regions of heterozygosity (RHZ)

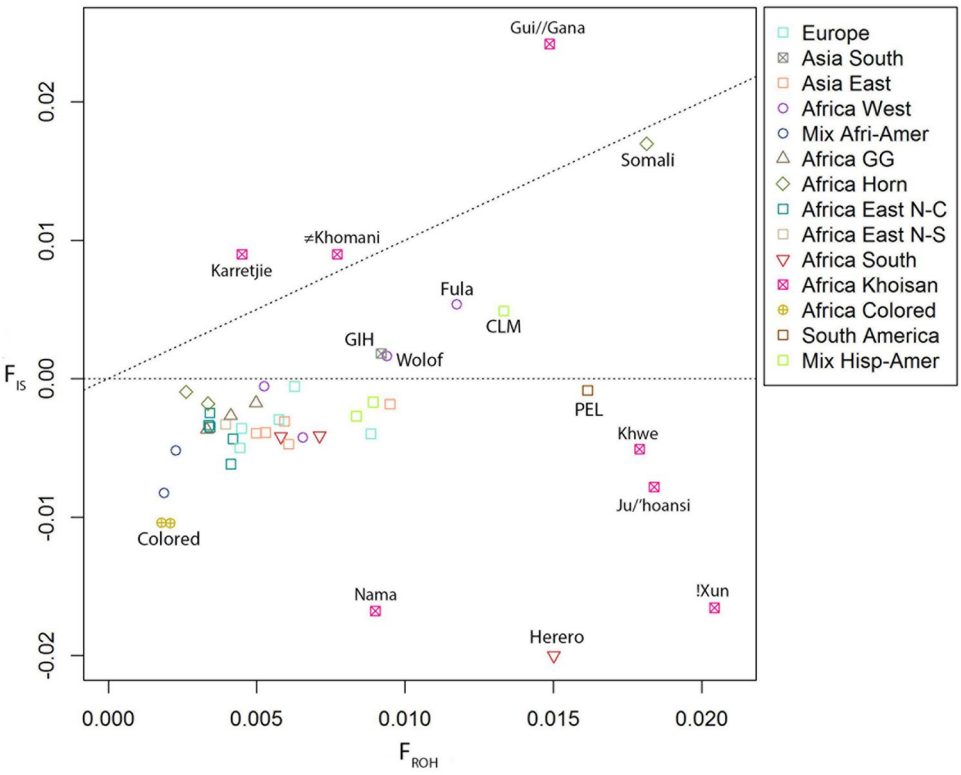

Fig. 4 Population analysis and components of the inbreeding coefficient. Systematic inbreeding coefficient (F_{IS}) versus the inbreeding coefficient obtained from ROH (F_{ROH}). In this context F_{IS} is the average SNP homozygosity within an individual relative to the expected homozygosity of alleles randomly drawn from the population and it was obtained using the—het function in PLINK. Diagonal broken line represents $F_{IS} = F_{ROH}$. Horizontal broken line represents $F_{IS} = 0$. For KGP population identifiers see "Materials and methods" section

were analysed by collapsing populations into their regional groups. From SSA: West, Gulf of Guinea, East, Horn of Africa, Southern Bantu and Khoe–San. From outside of SSA: Europe, East Asia, Hispanic-American admixed and African-American admixed. In Fig. 5 ROHi and RHZ are represented for the 22 autosomal chromosomes of the Khoe and San (A) and European groups (B).

Within SSA, the region of the Horn of Africa has the shortest (measured in Mb and cM) but a larger number of ROHi (544) (Table 2). The Khoe and San are the groups with the smallest number of ROHi (220), also they have an average size of ROHi when compared to other groups in Africa (0.5 Mb). Outside SSA, the Europeans form a group with the highest number of ROHi (795), 3.6 times more than the Khoe and San. Also, Europe is the group with larger ROHi, measured in Mb and cM, with 90 ROHi longer than 1 Mb. Interestingly the African-American admixed group has almost no ROH longer than 1 Mb, but is the group with the second highest number of ROHi. Surprisingly this group has longer ROHi with a mean size of 0.615 Mb or 0.25 cM, higher than most groups.

In order to appreciate differences between regional groups, three extremely long RHZ 0% (RHZ with no individual in autozygosity, see "Materials and methods"), shared by all groups, were removed before constructing Table 2. These three RHZ 0% are located in Chr1 (1253×10^5 to 1425×10^5; 17.3 Mb), Chr9 (457×10^5 to 664×10^5; 20.8 Mb) and Chr16 (384×10^5 to 463×10^5; 8 Mb). Table 2 shows bigger differences between regional groups when considering RHZ in comparison to ROHi, especially in number by size and mean length. This table also shows that for every group there are big differences between the number of RHZ 0% and 5%. These differences can be explained mainly by a drastic increase of short RHZ 5% regions (< 0.3 Mb) with the outcome of a reduction in the mean length (Mb and cM) of the RHZ 5% in comparison to RHZ 0%.

Table 3 shows the positions, lengths and presence of protein coding genes for the five most common ROHi per regional group in SSA. Almost every ROHi has at least one protein coding gene; just two ROHi from the African Khoe and San include no protein coding genes. Among the genes listed in Table 3 there are some already described to be under positive selection pressure. Hence, there are genes related to brain development: *GPHN*, *PCDH17*, *DARS*, *SCFD2*, *CPA6*, *DOCK3* or *CASC4* (Chen et al. 2010; Higasa et al. 2009; Liu et al. 2013; Lopez Herraez et al. 2009; Lopman and Gregson 2008; Mendizabal et al. 2012); involved in cancer or tumour processes: *ZCCHC11*, *SPOCK1*, *OLFML*, *EIF2S1*, *MPP5*, *CXCR4* (Lopman and Gregson 2008; Liu et al. 2013; Wagh et al. 2012); skin conditions: *NOMO1* (Oleksyk et al. 2008); Fanconi anaemia *FANCC* (Wang et al. 2006); pulmonary fibrosis: *PARN* (Grossman et al. 2013); Charcot–Marie–Tooth disease: *PLEK* (Grossman et al. 2013; Liu et al. 2013); and other metabolic and cellular processes (including *SH3RF* and *PC* (Lopman and Gregson 2008; Liu et al. 2013). Many of these ROHi with

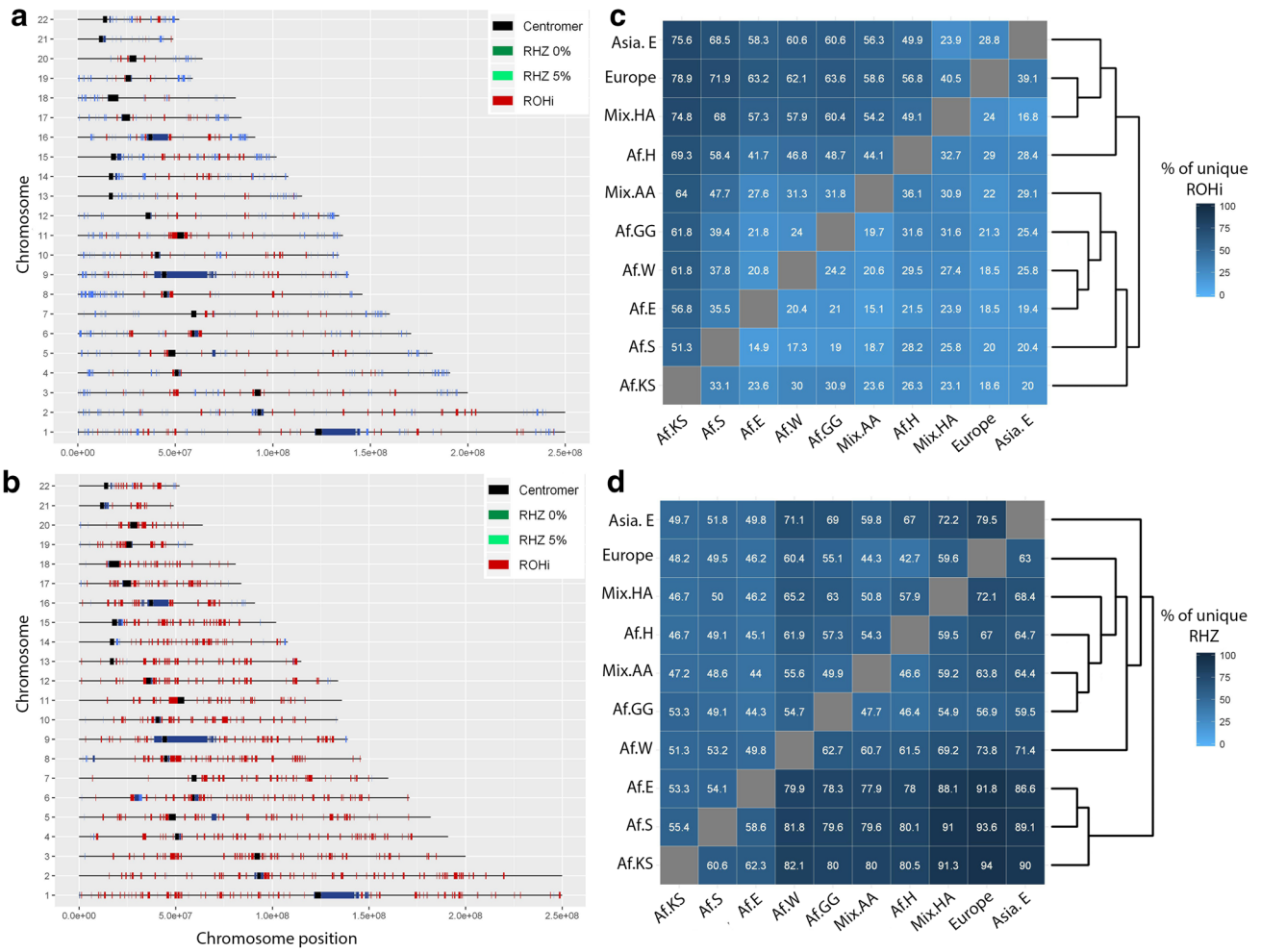

Fig. 5 Genomic distribution of ROH. Left side: genomic representation of the chromosomal location and size of runs of homozygosity islands (ROHi) and regions of heterozygosity (RHZ) for the Khoe and San (**a**) and European (**b**) regional groups. RHZ 0%: genomic regions where no individual in the group has a ROH. RHZ 5%: genomic regions where ≤5% of the population has ROH. Right side: heatmap and rooted dendrogram of the unique ROH islands (**c**) or RHZ (D) per geographical regional group and for admixed populations. The heatmap shows pairwise % of unique ROHi/RHZ between regional groups. The rooted dendrogram was obtained using optimal leaf ordering or OLO. *Af.KS* African Khoe and San populations, *Af.S* population from southern Africa, *Af.E* population from eastern Africa, *Af.W* population from western Africa, *Af.GG* population from the Gulf of Guinea, *Mix.AA* African-American admix populations, *Mix.HA* Hispanic-American admix populations, *Europe* European populations, *Asia.E* populations from eastern Asia

genes under positive selection are shared by more than one regional group. Without being exhaustive, the ROHi with the *FANCC* gene is present in all the SSA populations but not outside this region: 28.5% of the Western African populations has an ROH including this gene, 29.2% of the Gulf of Guinea populations, 19.5% of the Eastern Africa regional group, 23.6% of the people from the Horn of Africa, 17.3% of the populations from Southern Africa, 14.4 of the Khoe and San populations and 26.9% of the admixed African-American populations. Another example shared by all SSA, except the Khoe and San populations, is the ROHi with the *GPHN* gene: 21.7% of prevalence in Western Africa, 17.8% in the Gulf of Guinea, 22.2% in Eastern Africa, 26.1% in the Africa Horn, 14.9% in Southern Africa and 20.3% in the African-American admixed populations. ROHi with genes under positive selection were either present in all the populations like the *BCAS3* gene, or just present in one regional group like *HERC2* or *EDAR*, in Europe and East Asia, respectively (see Supplementary Table 1). Worthy of comment is the presence of an ROHi near the *LCT* gene; 38.8% of Europeans and 19.9% of East African individuals have a ROH across this gene, but it is not present in other SSA populations.

Table 4 shows the three longest RHZ 5%, with the presence of protein coding genes, for SSA regional population groups. Likewise Table 3 shows the three longest RHZ 0%, present in all regional groups, were removed. These three RHZ 0% have practically no protein coding genes, just two

Table 2 Summary statistics for the ROH islands (ROHi) and the regions of heterozygosity (RHZ) for populations combined from different geographic regions

Population	*n*	Number by size				Mean length (Mb)		Mean length (cM)		Max length		Mean number of SNP	
		> 1.0 Mb	0.5–1.0	0.3–0.5	< 0.3 Mb					Mb	cM		
						Mean	SD	Mean	SD			Mean	SD
Africa West													
ROHi	383	35	128	126	94	0.599	0.37	0.187	0.34	4.2	3.24	181.7	110.4
RHZ 0%	48	14	12	4	18	0.663	0.62	1.245	3.32	2.4	11.74	227.8	258.5
RHZ 5%	926	21	81	181	643	0.235	0.25	0.421	0.91	4.0	13.81	98.7	105.5
Africa GG													
ROHi	370	40	117	138	75	0.614	0.41	0.204	0.40	4.2	3.77	184.0	122.5
RHZ 0%	57	11	12	14	20	0.691	0.73	0.742	1.88	3.6	7.16	286.8	301.1
RHZ 5%	1295	21	130	258	886	0.259	0.31	0.467	0.95	4.1	13.81	107.3	126.1
Africa Horn													
ROHi	544	18	74	126	326	0.374	0.24	0.106	0.26	1.6	3.230	114.4	75.3
RHZ 0%	70	14	13	12	20	0.492	0.597	0.511	1.99	2.6	11.74	205.1	248.7
RHZ 5%	751	17	53	143	538	0.222	0.250	0.357	0.87	4	13.81	92.4	104.3
Africa East													
ROHi	371	47	134	134	56	0.647	0.37	0.209	0.39	3.3	3.779	210.0	120.5
RHZ 0%	57	11	13	15	18	0.731	0.740	0.885	2.01	3.6	7.16	298.8	300.5
RHZ 5%	1596	37	169	339	1051	0.279	0.336	0.526	1.12	4.0	14.24	114.2	137.4
Africa South													
ROHi	294	22	94	110	68	0.581	0.36	0.168	0.35	3.4	3.244	213.5	130.3
RHZ 0%	53	12	12	13	16	0.532	0.551	0.762	2.67	2.3	11.74	214.9	222.5
RHZ 5%	1300	32	152	342	774	0.261	0.261	0.467	0.95	2.6	13.81	105.5	105.3
Africa Khoe and San													
ROHi	220	19	61	79	61	0.565	0.37	0.099	0.19	3.3	2.622	210.6	137.1
RHZ 0%	49	9	12	10	18	0.650	0.69	1.105	3.22	3.6	11.30	262.3	281.6
RHZ 5%	1253	24	99	216	914	0.237	0.26	0.387	0.83	3.7	11.74	96.0	103.9
Mix. Afr-Amer													
ROHi	689	72	221	252	144	0.615	0.41	0.251	0.34	5.1	4.233	146.1	98.6
RHZ 0%	194	11	14	25	144	0.294	0.48	0.270	1.22	3.6	13.81	120.9	196.1
RHZ 5%	1859	39	217	404	1199	0.284	0.33	0.511	1.04	4.6	14.71	116.5	134.0
Europe													
ROHi	795	90	211	286	208	0.604	0.43	0.254	0.36	5.3	4.232	122.9	88.4
RHZ 0%	58	11	16	14	17	0.739	0.76	0.902	1.81	4.0	7.81	312.8	322.1
RHZ 5%	218	12	21	33	152	0.325	0.54	0.412	1.22	4.1	13.81	137.8	227.2
Asia. East													
ROHi	459	26	85	139	209	0.466	0.34	0.128	0.31	4.1	3.498	118.8	87.6
RHZ 0%	57	11	15	14	17	0.751	0.78	1.229	3.07	4	11.74	313.5	328.9
RHZ 5%	195	14	16	33	132	0.373	0.62	0.388	1.13	4.1	11.75	155.9	261.1
Mix. Hisp-Amer													
ROHi	645	56	171	205	213	0.561	0.40	0.202	0.34	5.3	4.232	144.9	104.6
RHZ 0%	59	11	16	14	18	0.726	0.76	0.846	1.77	4	7.16	302.4	316
RHZ 5%	273	12	22	48	191	0.304	0.49	0.403	1.10	4.1	13.81	126.6	203

For 1KG population identifiers see "Materials and methods" section

n number of ROHi and RHZ, *Mb* mega bases, *cM* centimorgans, *SD* standard deviation

members of the *SPATA31* subfamily (5 and 7) genes on Chr9 that are involved in spermatogenesis and are under positive selection (Bekpen et al. 2017). Table 4 shows that there are many protein coding genes present in regions with increased heterozygosity. The RHZ on Chr6 is shared by every regional group but the Khoe and San. It has a length of

Table 3 Location, length, percentage of individuals with ROH for the ROH islands (ROHi) and protein coding genes of the five most prevalent ROHi in the sub-Saharan African regional groups

Chr.	Start	End	Length (Mb)	% Indv	Protein coding genes
Africa W.					
7	64.9	66.5	1.6	32.97	*ZNF, ASL, CRCP, ERV3-1, GUSB, TPST1, VKORC1L1*
17	45.4	45.9	0.5	31.16	*ARHGAP27, CRHR1, PLEKHM1*
9	95.1	95.7	0.6	28.58	*FANCC, PTCH1*
1	114	114.5	0.5	26.39	*OLFML3, SYT6, TRIM33*
4	107	107.4	0.4	21.93	*DKK2*
Africa GG.					
9	95.1	95.7	0.6	29.26	*FANCC, PTCH1*
7	64	66.1	2.1	27.58	*ZNF680*
11	10	10.4	0.4	26.92	*SBF2*
16	14.6	15.6	1	26.62	*PARN, BFAR, NPIPA, NTAN, PDXDC1, NOMO1, MPV17L, PLA2G10, RRN3*
17	45.5	45.9	0.4	26.51	*CRHR1*
Africa E.					
16	18.3	19	0.7	28.27	*NPIPA8, NOMO2, RPS15A, SMG1, ARL6IP1*
7	64.4	66.5	2.1	24.62	*ZNF680*
14	66.8	67.9	1.1	22.22	*GPHN, ATP6V1D, EIF2S1, MPP5, PIGH, PLEK, RDH, TMEM229B, VTI1B, ARG2*
1	114	114.5	0.5	22.04	*OLFML3, SYT6, TRIM33*
4	107	107.4	0.4	21.93	*DKK2*
Africa H.					
2	135.9	136.8	0.9	36.24	*DARS, CXCR4*
8	67.6	68.4	0.8	35.63	*CPA6, PREX2*
1	52.5	53.1	0.6	33.96	*ZCCHC11, COA7, ECHDC2, GPX7, SCP2, SHISAL2A, ZYG*
11	66.9	67.5	0.6	33.33	*PC, ANKRD13D, CLCF1, GRK2, KDM2A, POLD4, PPP1CA, RAD9A, RHOD, SSH3, SYT12*
7	65.1	66	0.9	32.11	*ZNF680, ASL, CRCP, GUSB, TPST1, VKORC1L1, ZNF92*
Africa S.					
7	65	66.5	1.5	26.46	*ZNF680, ASL, CRCP, GUSB, TPST1, VKORC1L1, ZNF92*
17	45.4	45.9	0.5	22.10	*ARHGAP27, CRHR1, PLEKHM1*
13	57.7	58.3	0.6	21.46	*PCDH17*
3	50.7	51.9	1.2	20.66	*DOCK3, MANF, RBM15B, DCAF1, GRM2, IQCF6, RAD54L2, TEX264*
15	44.4	45	0.6	20.12	*CASC4, B2M, CTDSPL2, EIF3J, PATL2, SPG11, TRIM69*
Africa KS.					
3	75	75.4	0.4	28.04	
2	198.3	199	0.7	26.06	*PLCL1*
4	52.8	53.2	0.4	22.64	*RASL11B, SCFD2*
5	137.1	137.9	0.8	22.64	*SPOCK1, HNRNPA0, KLHL3*
12	60.4	60.9	0.5	21.89	

Locations of the different genes were obtained using the GRCh38.p12 genome assembly

Genes underlined have been previously reported to be under positive selection

Chr chromosome, *Start* position where the RHZ starts in Mb, *End* position where the RHZ finish in Mb, *%Ind in ROH* percentage of individuals in the population that has a ROH in this region

4 Mb, and has more than 140 protein coding genes including many members of the HLA complex family, olfactory receptor family, MHC class I genes, lymphocyte antigen 6 family, and the psoriasis susceptibility 1 candidate gene among others. As for ROHi, multiple RHZ are shared by different regional groups. Supplementary Table 2 shows the 3 longer RHZ for the non-African groups.

It is possible to use differences in ROHi and RHZ across regional groups to obtain a genetic distance that could provide an evolutionary perspective of the distribution of these homozygous and heterozygous genomic regions. Figure 5 shows a pairwise comparison of unique ROHi (C) and RHZ (D) in two heatmaps and a rooted dendrogram for each heatmap using the percentage of unique ROHi

Table 4 Location, length, percentage of individuals in ROH for each run of heterozygosity (RHZ) and protein coding genes for the three longest RHZ in the sub-Saharan African regional groups

Chr	Start	End	Length (Mb)	% Ind in ROH	Protein coding genes
Africa W.					
6	28.7	32.7	4	1.5	+ 140 genes
5	68.6	71.2	2.6	0.24	*SLC30A5, ANP32A, CORO2B, GLCE, KIF23, LARP6, NOX5, PAQR5, RPLP1, TLE3, UAUCA, SPESP1, THAP10, THSD4*
16	84.4	87	2.6	2.3	*FOXL1, FOXC2, COTL1, COX4I1, CRISPLD2, EMC8, FAM92B, FOX, GINS2, GSE1, IRF8, KIAA0513, KLHL36, MTHFSD, TDLC1, USP10, AZDHHC7*
Africa GG.					
6	28.7	32.7	4	0.35	+ 140 genes
12	128.6	131.3	2.7	2.8	*ADGRD1, FZD10, GLT1D1, PIWIL1, RAN, RIMBP2, STX2, TMEM, SLC15A5*
5	68.6	71.2	2.6	0	See Africa W. second RHZ
Africa E.					
6	28.7	32.7	4	0.51	+ 140 genes
9	39.3	42.9	3.6	0	*SPATA31A1, FOXD4L6, CBWD6, ANKRD20A2, CNTNAP3B*
12	127.8	131.3	3.5	2.1	See Africa GG. second RHZ
Africa H.					
6	28.7	32.7	4	0.42	+ 140 genes
12	128.6	131.3	2.7	2.8	See Africa GG. second RHZ
5	68.6	71.2	2.6	0	See Africa W. second RHZ
Africa S.					
6	28.7	32.7	4	0.54	+ 140 genes
9	38.5	42.6	4.1	0.2	See Africa E. second RHZ
16	84.4	87	2.6	2.3	See Africa W. third RHZ
Africa KS.					
9	39.2	42.9	3.7	0.2	See Africa E. second RHZ
8	6.4	8.2	1.8	3.3	+ 30 genes
5	69	70.7	1.7	0	See Africa W. second RHZ
Mix A.A.					
6	28.5	32.5	4	0.4	+ 140 genes
12	127.8	131.3	3.5	1.3	See Africa GG. second RHZ
16	84.3	87.4	3.1	1.4	See Africa W. third RHZ

Locations of the different genes were obtained using the GRCh38.p12 genome assembly

Genes underlined have been previously reported to be under positive selection

sChr chromosome, *Starts* position where the RHZ starts in Mb, *End* position where the RHZ finish in Mb, *%Ind in ROH* percentage of individuals in the population that has a ROH in that region

or RHZ as genetic distances. Both rooted dendrograms establish two main groups: SSA and out-of-Africa. Within SSA (with the exception of the Horn of Africa), both first split off the Khoe and San from the rest of groups and then split the Bantu-speaking populations from Southern Africa from the rest. Also, they include the mixed African-American group in the SSA branch. In the out-of-Africa branch both dendrograms group together European and admixed Hispanic-American populations. However, the ROHi dendrogram groups the Horn of Africa populations with the out-of-Africa branch, whereas the RHZ dendrogram groups them with the SSA branch.

Discussion

Previous ROH analyses of SSA populations have added limited value to understanding the demographic history of populations on the continent. Generally, they showed that Africa is the continent with the smallest level of ROH and that within Africa there is limited heterogeneity in ROH distribution, occurring essentially between the hunter-gatherers and the agro-pastoralists (Pemberton et al. 2012; Schlebusch et al. 2012; Ceballos et al. 2018b). Our study, however, shows that ROH distribution in SSA

is very heterogenous, with different scenarios for ROH shorter and longer than 1.5 Mb, reflecting a much more complex history of homozygosity in SSA. Although the vast majority of SSA populations have a low level of short ROH that is not the case for long ROH where we find SSA populations with a higher level in comparison to other worldwide populations present in the KGP dataset and that is a reflection of the unique demographic history of SSA populations. In contrast with previous studies, our fine-scale analysis has overcome some limitations: it has representation of populations from Western, Eastern and Southern Africa; it uses high-density SNP coverage (~ 1.2 M SNPs after QC), providing good resolution to accurately call ROH; the PLINK software parameters for ROH calling were optimized to accurately call short ROH; and analyses were developed to understand the ROH distribution and its demographic consequences.

Insights into the past: analysis of short ROH (ROH < 1.5 Mb)

The demographic history of SSA is characterized by large effective population sizes over many generations that have led to high genetic diversity, shorter LD structures and lower level of small ROH (Pemberton et al. 2012; Henn et al. 2016). Our study reports considerable structure in the distribution of short ROH in Africa with populations from the Horn of Africa (Somali, Oromo and Amhara) having the largest level of ROH < 1.5 Mb. In the absence of evidence to support a different evolutionary trajectory of the effective population size between these and other SSA populations, the most plausible explanation is that the short ROH were introduced through admixture of Semitic and Cushitic populations with others from the Arabian Peninsula. It has been found that Ethiopian individuals are characterized by a large (40–50%) non-African genetic component most likely originating mainly from Egypt, the Levant and Yemen in a migration that took place approximately 3000 years ago (kya) (Pagani et al. 2012; Pickrell et al. 2014; Gandini et al. 2016). This hypothesis is also supported by the ROHi profiling of populations in the Horn of Africa that have the highest number of short ROHi and the shortest mean ROHi length (Table 2). When compared with other regional groups (Fig. 5), the populations from the Horn of Africa share more ROHi with regional groups outside Africa. There is a reasonably homogeneous level of short ROH between Western, Gulf of Guinea, Eastern and Southern Bantu-speaking groups (Table 1 and Fig. 3), but the Khoe and San, having split from non-Khoe and San lineages at least by 100–150 kya (Kim et al. 2014), show heterogeneity (though of course recent work claims 10% Afroasiatic and East African ancestry introduced some 2 kya (Schlebusch and Jakobsson 2018).

The distribution of the ROH < 1.5 Mb is also highly informative. Admixed populations, originating from ancestral populations with different ROH level would have individuals with different sum of ROH < 1.5 Mb due to their distinct coalescent histories, as shown in Fig. 3 where most of the admixed populations present platykurtic and skewed distributions. Hispanic-American populations (CLM, PUR, MXL), with ROH < 1.5 Mb level similar to Europeans, had a small proportion of African ancestry (8%, 14% and 4%, respectively) but higher proportion of European (66.6%, 73.2% and 48.7%, respectively) and Native American (26%, 18% and 47%, respectively) ancestry (Montinaro et al. 2015; Martin et al. 2017). The PEL population had shorter ROH due to a greater Native American ancestry (2.5% African, 20.2 European and 77.3% Native American) (Montinaro et al. 2015; Martin et al. 2017). For these populations ROH < 1.5 Mb arose before the time of admixture; estimated as 14 generations for CLM, 7 for MXL and 16 for PUR. PEL population was found to have two different admixture pulses 12 and 5 generations ago, with the last one being 91% Native American (Montinaro et al. 2015; Martin et al. 2017). On the other hand, African-American admixed populations (ASW and ACB) had reasonably normal distributions with almost no skewness. These two populations seem to have a very tight distribution and small level of ROH < 1.5 Mb, similar to the Western Africans and Guinea Gulf populations. This could be explained by the elevated proportion of African ancestry (88% and 75.6%, respectively) and small proportions of European and Native American ancestry (ACB: 11.7% European, 0.3 Nat American; ASW: 21.3% European, 3.1% Nat American) (Montinaro et al. 2015; Martin et al. 2017). The South African Coloured populations, another example of a recently (150–300 years) highly admixed populations, have a ROH < 1.5 Mb level very similar to Khoe and San populations. This is consistent with studies reporting ancestry components for Coloured populations arising from Khoe, San, and Bantu speakers, as well as European, South Asian and Austronesian populations (Daya et al. 2013; Choudhury et al. 2017) giving insight into the complexity of these admixed populations. Finally, it is also possible to detect kurtosis and skewness in some Khoe and San populations which would indicate admixture. Also, Gui//Gana, Nama, Karretjie and ǂKhomani distributions for sum of ROH < 1.5 Mb reveal their admixture origins. In these four Khoe and San populations Bantu and even European ancestral components were found (Schuster et al. 2010; Schlebusch et al. 2012; Busby et al. 2016).

Consanguineous cultural practices and modern genetic isolation: analysis of long ROH (ROH > 1.5 Mb)

The study of ROH > 1.5 Mb is very useful to shed light on the role of cultural practices in genome homozygosity

levels. Different anthropological and human biology studies have identified African populations with a clear cultural preference for consanguineous marriages, and some that purposely avoid such unions (Schapera 1957; Tanner 1958; Scott-Emuakpor 1974; Lesthaeghe et al. 1989; Caldwell et al. 1992; Hampshire and Smith 2001; Bledsoe 2002). For example, one of the most recently published studies, which analysed 548 marriages over the period 1994–96 in the Fulani from Burkina Faso, found that 399 marriages (68.3%) were between relatives. The average inbreeding coefficient (α) was estimated as 0.0364 equivalent of first cousin once removed (Hampshire and Smith 2001). Similar inbreeding coefficients were found by other studies, for example an $\alpha = 0.0322$ in the Khartoum population from Sudan (Saha and El Sheikh 1988). Our study shows a very heterogeneous distribution of ROH > 1.5 Mb among SSA: populations with very little level of long ROH > 1.5 Mb, and complete absence of ROH > 4 Mb, for example in the Amhara from the Horn of Africa, the Yoruba from the Gulf of Guinea or the Kikuyu from Eastern Niger–Congo Africa, and populations with a high level of ROH > 1.5 Mb such as the Somali from the Horn of Africa, the Fula from Western Africa or the Khoe and San !Xun and Ju/'hoansi. A heterogeneous distribution of long ROH was found within SSA regions: Somali and Oromo populations, from the Horn of Africa, speak Cushitic languages, but Somalis are predominantly Sunni Muslims, with a preference for first-cousin unions, while Oromo people are predominantly Ethiopian Orthodox or follow traditional religions with no preference for consanguineous unions (Bittles 2012). Cultural differences among individuals within populations can be inferred from the skewness of the distributions in Fig. 3. Not surprisingly, populations with larger level of ROH > 1.5 Mb (in order: !Xun, Ju/'hoansi, Somali, Khwe, PEL, Gui//Gana, CLM, Fula, etc.) have the strongest positive skewness and the highest number of individuals with an inbreeding coefficient higher than $F = 0.0152$.

In order to sketch a more complete picture of genomic homozygosity in SSA populations, it is important to analyse the origins of this homozygosity. In summary, it is possible to establish a classification with 4 main groups characterized by demographic history: (1) very admixed populations with almost no ROH longer than 1.5 Mb like the South African Coloured populations or the African-Americans. (2) populations with different levels of cultural consanguinity practices like Somali, Fula, CLM, GIH and Wolof; (3) populations with low levels of inbreeding due to their large continental N_e, in this group we can find the bulk of European, Asian and SSA populations; and (4) populations with considerable genetic drift and recent genetic isolation like PEL, Khwe, Ju/'hoansi, !Xun and Herero. The representation of F_{IS} vs F_{ROH} is a useful approach to identify the origins of inbreeding since it provides information about the proportion of F_{ROH} due to deviation from panmixia or from genetic isolation. Furthermore, this representation is helpful to identify populations with an excess of homozygotes possibly due to the Wahlund effect, which may be expected for the Gui// Gana population, or, more surprisingly, with the Southern Tuu-speaking Khoe and San, and Karretjie people.

Genomic distribution of ROH and the identification of regions under selection

Examining ROH has been shown to be useful for studying genome biology and to identify regions under selection (Curtis 2007; Nothnagel et al. 2010; Pemberton et al. 2012). The existence of ROH islands (ROHi) and regions of heterozygosity (RHZ) could be explained in part as a consequence of stochastic processes across the genome, or by variation of the effects of demographic processes across the genome, influencing genetic diversity (Pemberton et al. 2012; Ceballos et al. 2018b). However, there is increasing evidence that ROHi may be a consequence of positive selection processes that reduce haplotype diversity and increase homozygosity around the target locus, increasing ROH frequencies in the regions under selection (Lencz et al. 2007; Pemberton et al. 2012). Besides the presence of specific protein coding genes, previously detected to be under positive selection (Table 3), we identified other genes that have been shown to be under positive selection specifically in African populations (Schlebusch et al. 2012; Chimusa et al. 2015; Gurdasani et al. 2015). Different loci associated with infectious disease susceptibility and severity, including *HP* (Gurdasani et al. 2015), *CLTA4* (Jacobs et al. 2002) and *PKLR* (Machado et al. 2010) for malaria, *IFIH1* and *OAS2* (Gurdasani et al. 2015) for Lassa fever, *FAS* (Martins et al. 2001) for trypanosomiasis and other genes involved in general immune response (e.g. *PRSS16* and *POM121L2* (Schlebusch et al. 2012) were found within ROHi in different geographical regions. For example, *CTLA4* was found in ROHi in every region, but *HP* and *PKLR* were found to be in ROHi just in Western and Eastern SSA and in the Horn of Africa. Other genes related to trypanosomiasis infection and kidney disease, like *APOL1* (Jacobs et al. 2002), or to different forms of hypertension, like *ATP1A1*, *AQP2* and *CSK* (Gurdasani et al. 2015), were found in ROHi in different regions from SSA. Within RHZ haplotypes (Table 4), we found multiple protein coding genes related to diverse biological functions like immune response (*HLA* complex or *IRF* gene family), cellular cycle (*ANP32A*) (Opal et al. 2003; Schafer et al. 2006), chromosomal aberrations (like different members of the *GOLGA* gene family (Silano et al. 2007)) cancer [*NOX5* (Fu et al. 2006)], brain development [*KIAA0513* (Lauriat et al. 2006)] and olfactory receptors (*OR* gene family), among others. These heterozygous regions might represent haplotypes enriched for variants that have a negative impact on fitness when homozygous, or regions that harbour

loci with heterozygote advantage (overdominance) under any form of balancing selection. Furthermore, this hypothesis is also supported by the fact that it was possible to establish differences and similarities between the locations of ROHi and RHZ between populations from different geographic regions.

We present the first fine-scale distribution analysis of ROH in SSA populations. By accurately calling long and short ROH along with their genomic distribution, we were able to shed new light on demographic history and the effects of modern cultural practices. Our analyses confirmed that Africa is the continent with the least level of ROH, but also showed that back-to-Africa gene flow has introduced short ROH in populations from the Horn of Africa. Also, the distribution of short ROH revealed admixture practices in the Khoe, San and Coloured populations. Long ROH, on the other hand, highlight a different scenario. These recent ROH revealed different cultural practices of consanguinity in populations like the Somali, Wolof and Fula or genetic isolation (endogamy) behaviour in the Khwe, Ju/'hoansi, !Xun Khoe and San populations. Hence, despite previous reports, we found African populations with mean genomic inbreeding coefficients (F_{ROH}) higher than several other isolated populations around the world, such as the PEL from Lima in Peru. These long ROH are useful to identify population structure caused by the Wahlund effect like in the Karretjie or Gui//Gana Koe and San populations. Finally, we showed that the majority (more than 75%) of ROHi and RHZ identified in this study included genomic regions that had previously been identified as sites of recent selection. These analyses raise the possibility that other regions in ROHi and RHZ may also harbour genes that have been subjected to positive or balancing selection. In the future, studying a better representation and larger sample size across different SSA populations will provide more nuanced interpretations of demographic histories. The H3Africa (Human Heredity and Health in Africa) initiative is generating genomic data including whole genome and exome sequences and genome-wide genotyping using an African tailored array that captures common genetic diversity in African genomes (H3 Africa Consortium et al. 2014; Ramsay et al. 2016). The added value of this resource lies in its rich phenotype and clinically relevant data that will enable biomedical research across the continent making it possible to study the distribution of ROH and RHZ in common complex traits.

Materials and methods

Description of the data

The study included a total of 3063 individuals from 45 populations from the 1000 Genomes Project—Phase 3 (KGP) (Sudmant et al. 2015), the African Genome Variation Project (AGVP) (Gurdasani et al. 2015) and Schlebusch et al. (2012) (Schlebusch et al. 2012). All individuals were genotyped using the Infinium Omni 2.5 array from Illumina, and all datasets were subjected to extensive QC procedures.

The KGP—Phase 3, includes a total of 1558 individuals from 19 populations (Sudmant et al. 2015). From Europe: FIN (Finish in Finland, $n=97$), GBR (British in England and Scotland, $n=91$), IBS (Iberian populations in Spain, $n=99$), TSI (Tuscany in Italy, $n=92$) and CEU (Utah residents with European ancestry = 95). From America: ASW (Americans of African ancestry in Houston, $n=49$), ACB (African Caribbean in Barbados, $n=72$), PUR (Puerto Rican in Puerto Rico with admixed ancestry, $n=72$), PEL (Peruvian in Lima, Peru with Amerindian ancestry, $n=50$), CLM (Colombian in Medellin, Colombia with admix ancestry, $n=65$) and MXL (Mexican with admixed ancestry in Los Angeles, USA, $n=47$). From South Asia: GIH (Gujarati Indian from Houston, Texas $n=95$). From East Asia: CDX (Chinese Han in Xishuangbanna, China, $n=83$), CHB (Chinese Han in Beijing, China, $n=98$), CHS (Southern Han Chinese, $n=86$), JPT (Japanese in Tokyo, Japan, $n=96$) and KHV (Kinh in Ho Chi Minh city, Vietnam $n=96$). From Africa Guinean Gulf: YRI (Yoruba in Ibadan, Nigeria, $n=100$), and from East Africa: LWK (Luhya in Webuye, Kenya, $n=74$).

The AGVP includes 1318 individuals from 17 populations from SSA (Gurdasani et al. 2015). Niger–Congo speakers from Western Africa: Wolof (Senegambian sub-group speakers from The Gambia, $n=78$), Fula (Senegambian from The Gambia, $n=74$), Mandinka (Mande sub-group speakers from The Gambia, $n=88$) and Jola (Bak sub-group speakers from The Gambia, $n=79$). Niger–Congo speakers from the Guinean Gulf: Ga-Adangbe (Kwa sub-group speakers from Ghana, $n=100$) and Igbo (Igboid sub-group speakers from Nigeria, $n=99$). Afro-Asiatic speakers from the Horn of Africa: Amhara (Semitic sub-group speakers from Ethiopia, $n=42$), Oromo (Cushitic sub-group speakers from Ethiopia, $n=26$) and Somali (Cushitic from Ethiopia and Somalia, $n=39$). Niger–Congo speakers from Eastern Africa: Baganda (Bantoid sub-group speakers from Uganda, $n=100$), Banyarwanda (Bantoid from Uganda, $n=100$), Barundi (Bantoid from Uganda, $n=97$) and Kikuyu (Bantoid from Kenya, $n=99$). Nilo-Saharan speakers from Eastern Africa: Kalenjin (Eastern Sudanic sub-group speakers from Kenya, $n=100$). Niger–Congo speakers from Southern Africa: Sotho (Bantoid from South Africa, $n=86$) and Zulu (Bantoid from South Africa, $n=100$).

In addition, 147 individuals from 7 different groups with Khoe and San ancestry, 40 South African Coloured individuals (20 from Colesberg and 20 from Wellington, both in South Africa) and 12 Herero Bantoid speakers from Namibia from the Schlebusch study (2012) were added. The term Khoe–San designates two groups of people: the

pastoralist Khoe and the hunter-gatherer San. The following were included in this study: Ju/'hoansi (San Ju speakers from Namibia, $n=18$), !Xun (San Ju speakers Angola, $n=19$), Gui//Gana (San Khoe–Kwadi speakers from Botswana, $n=15$), ǂKhomani (San Tuu speakers from South Africa, $n=39$), Nama (Khoe Khoe–Kwadi speakers from Namibia), Khwe (San Khoe–Kwadi speakers from the Caprivi strip: Namibia, Angola and Botswana) and Karretjie people (San Tuu speakers from South Africa, $n=20$).

SSA samples were grouped according to geographic region and principal components analysis into eight groups (Fig. 1): Western Africa ($n=319$), Gulf of Guinea ($n=299$), Eastern Africa Niger–Congo populations ($n=470$), Eastern Nilo-Saharan population ($n=100$), Horn of Africa ($n=107$), Southern Africa ($n=198$), Khoe and San groups ($n=147$) and Coloured South Africans ($n=40$). KGP populations from the rest of the world were grouped as follows: Mixed African-American populations ($n=121$), Europeans ($n=474$), Southern Asians ($n=95$), Eastern Asians ($n=459$), South Americans ($n=50$) and Mixed Hispanic-Americans ($n=184$). Since the three datasets used in this study were genotyped using the same SNP genotyping array, they could easily be merged (Joshi et al. 2015; Ceballos et al. 2018a). Only autosomal SNPs were included in this analysis. For each population, array data were filtered to remove SNPs with minor allele frequencies <0.05 and those that divert from H–W proportions with $p<0.001$. This filtering serves to limit the effects of ascertainment bias caused by the small number of individuals in the SNP discovery panel. After QC, there were 1.3 M SNPs on average in Western Africa populations, 1.4 M in Gulf of Guinea, 1.4 M in Eastern Africa Niger–Congo populations, 1.4 M in Eastern Nilo-Saharan population, 1.3 M in Horn of Africa populations, 1.3 M in Bantu-speaking Southern Africa populations, 1.4 M in Khoe and San populations from Southern Africa, 1.4 M in Coloured populations from Southern Africa, 1.4 M in African-American admixed populations, 1.2 M in European populations, 1.2 M in southern Asian populations, 1.1 M in Eastern Asian populations, 1.1 M in South American populations and 1.2 M in Hispanic-American admixed populations.

Merging with the human genome diversity project data

In an attempt to enrich the data further we merged the above data with the Human Genome Diversity Project Dataset (HGDP) since this dataset include isolates and urban populations from across the world. However, although the HGDP includes 1043 individuals from 51 populations from different parts of the world, these samples were genotyped using the Illumina (650 K SNP) making necessary to select the common SNP among both SNP arrays (Illumina BedStation and Infinium Omni 2.5). After merging and filtering for MAF and H-W proportions we obtained a dataset of 4106 individuals with genotypes for 382,840 SNPs. This dataset was not used in this study since due to its lower SNP coverage very large and very short ROH are systematically underestimated as can be seen in Supplementary Fig. 5.

Identification of runs of homozygosity

The observational approach implemented by PLINK v1.9 (Purcell et al. 2007) was used to call ROH. The simplicity of the approach used by PLINK allows efficient execution on data from large consortia and even different array platforms or sequencing technologies (Joshi et al. 2015; Ceballos et al. 2018b). Tests on simulated and real data showed that the approach used by PLINK outperformed its competitors in reliably detecting ROH (Howrigan et al. 2011).

The following PLINK conditions were applied to search for ROH:

- homozyg-snp 30. Minimum number of SNPs that a ROH is required to have,
- homozyg-kb 300. Length in kb of the sliding window,
- homozyg-density 30. Required minimum density to consider a ROH (1 SNP in 30 kb),
- homozyg-window-snp 30. Number of SNPs that the sliding window must have,
- homozyg-gap 1000. Length in kb between two SNPs in order to be considered in two different segments.
- homozyg-window-het 1. Number of heterozygous SNPs allowed in a window
- homozyg-window-missing 5. Number of missing calls allowed in a window
- homozyg-window-threshold 0.05. Proportion of overlapping window that must be called homozygous to define a given SNP as in a "homozygous" segment.

The objective of this study was to use autozygosity to learn more about demographic history in SSA populations. To achieve this goal short and long ROH were explored, since they provide different types of information (Ceballos et al. 2018a, b). The high SNP coverage of 1.2 M SNPs on average for all the populations included in the study, makes it possible to find a single SNP, on average, in a track of 2.4 kb. The Supplementary Methods and Figs. 6–10 demonstrate that this coverage allows accurate detection of ROH longer than 300 kb by considering 30 as a minimum number of SNPs per ROH and/or the required minimum SNP density to call ROH. To obtain a window with 30 SNPs, on average (assuming a homogeneous distribution of SNP along the genome), a tract of just 72 kb is needed. A threshold of 300 Kb was set for the minimum length in order to capture small ROH originating far in the past and also to ensure that these are true ROH that originated by genetic drift or

consanguinity. An alternative source of homozygosity originating from linkage disequilibrium (LD) typically produces tracts measuring up to about 100 kb, based on empirical studies (International HapMap C et al. 2007; Slatkin 2008). By using a minimum-length cutoff of 300 kb, most short ROH resulting from LD will be eliminated.

Analyses

Different variables were obtained and analyses performed in order to fully exploit the usefulness of the ROH in the understanding of demographic history and possible cultural practices of populations. First, we obtained the total sum of ROH for six ROH length classes: 0.3–0.5, 0.5–1, 1–2, 2–4, 4–8 and > 8 Mb. This exploratory data analysis allows us to delve into aspects of population history, since, due to recombination, the size of a ROH is inversely proportional to its age. Thus, plotting the total sum of ROH for these size classes will inform, for example, the relative change of the effective population size across generations.

For comparison purposes two variables were defined: total sum of ROH > 1.5 Mb as the population average total sum of ROH longer than 1.5 Mb; and total sum of ROH < 1.5 as the population average total sum of ROH shorter than 1.5 Mb. A cut-off of 1.5 Mb was chosen because the sum of ROH longer than 1.5 Mb has the best correlation with the inbreeding coefficient obtained from a pedigree of five complete generations (McQuillan et al. 2008). Preliminary results estimate that ROH length of 1.5 Mb may have a median age of approximately 30 generations (personal communication D.W. Clark) and ROH longer than 4 Mb may not be older than 10 generations[9]. Exploratory data analysis and data representation were illustrated using violin plots. These plots combine a box plot with a kernel density plot, where the interval width is obtained by the rule of thumb. The violin plot shows a coloured density trace with the interquartile range as a black line and median as a white dot. This representation is especially useful when dealing with asymmetric distributions where median is more informative than the mean. Statistical comparisons between total sum of ROH longer and shorter than 1.5 Mb between populations and geographic regions were performed using the Whitney–Wilcoxon non-parametrical test (MWW). All the analyses were performed using R (v.3.4.1).

Measuring different sources of inbreeding

Population geneticists use the word inbreeding to mean different things, as pointed out by Jacquard and Templeton in their respective classic articles (Jacquard 1975; Templeton and Read 1996). Inbreeding can be produced by a deviation from panmixia, in what G. Malecot called systematic inbreeding, or by genetic drift and low effective population size, also called panmictic inbreeding (Templeton and Read 1996). Systematic inbreeding has a direct effect on the H–W proportions of a population and can be measured using the Wright's fixation index or F_{IS} (Hartl and Clark 2007). In this study, this component of total inbreeding coefficient is measured using the het function in PLINK. In this context F_{IS} is the average SNP homozygosity within an individual relative to the expected homozygosity of alleles randomly drawn from the population. PLINK use the following expression:

$$F_{IS} = \frac{O(\mathrm{HOM}) - E(\mathrm{HOM})}{N - E(\mathrm{HOM})},$$

where *Observed Hom* is the observed number of homozygous SNPs, *Expected Hom* is the expected number of homozygous SNPs considering H–W proportions and N is the total number of non-missing genotyped SNPs. F_{IS} thus measures inbreeding in the current generation with $F_{IS} = 0$ indicating random mating, $F_{IS} > 0$ indicating consanguinity and $F_{IS} < 0$ indicating inbreeding avoidance.

The two different sources of inbreeding, namely, genetic drift (denoted by F_{ST}) and non-random mating (F_{IS}) are both components of the total inbreeding coefficient (F_{IT}), defined as the probability that an individual receives two alleles that are identical-by-descent. Sewall Wright developed an approach to consider these three different F coefficients in his F statistics $(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$ (Hartl and Clark 2007). First defined as correlations, Nei showed how these coefficients can be expressed in terms of allele frequencies and observed and expected genotype frequencies (Weir 2012). In this framework, F_{ST} can be considered a measure of the genetic differentiation of a subpopulation in comparison with an ideal population with a large N_e. F_{IT} is the total inbreeding coefficient, traditionally obtained using deep genealogies, and can be calculated using the F_{ROH} (ROH > 1.5 Mb):

$$F_{ROH} = \frac{\sum_{i=1}^{n} \mathrm{ROH} > 1.5\ \mathrm{Mb}}{3\ \mathrm{Gb}},$$

where the numerator is the sum of n ROH of length l_i (> 1.5 Mb) and the denominator is the total autosomal length.

Genomic distribution of ROH

The study of the genomic distribution of ROH can be used for different purposes. By identifying the regions where ROH are very prevalent, or completely absent in the population it is possible to identify candidate regions (including protein coding genes) under selection. Furthermore, the identification of common and unique ROHi in the different regional groups considered in this study can also shed light on population demographic history. In order to study

the spatial distribution of ROH across the genome two different variables were defined: islands of runs of homozygosity (ROHi) and regions of heterozygosity (RHZ) (see definitions below). In order to identify protein coding genes in these regions *biomartR* package for R was used. Differences in ROHi and RHZ between populations were used as genetic distances as a source to build a rooted dendrogram by using optimal leaf ordering (OLO) for hierarchical clustering available in the *heatmaply* R package. The OLO clusters similar groups (or leaves) taken from the UPGMA (unweighted pair grouping with arithmetic mean) algorithm and yields the leaf order that maximizes the sum of the similarities of adjacent leaves in the ordering (Brandes 2007).

Islands of runs of homozygosity (ROHi)

ROHi are defined as regions in the genome where the proportion of individuals of a population have ROH in a specific region that is more than expected by a binomial distribution. In order to search for ROHi, a sliding window of 100 kb was used. In every 100 kb genomic window the number of people with ROH was obtained; and to know if a specific genomic window has a significant enrichment of ROH across the population, a binomial test with $P < 2 \times 10^{-7}$ with Bonferroni correction for 2500 windows was applied. According to this procedure two variables could introduce bias when comparing populations across the globe: different population sizes and ROH background. In order to mitigate this source of bias the following steps were followed. Firstly, ROH of all the populations by geographical area and admixture were collapsed creating the following groups: Europe ($n = 474$ individuals), Eastern Asia ($n = 459$ individuals), admixed African-American ($n = 121$ individuals), Western Africa ($n = 319$ individuals), Africa Guinea Gulf ($n = 299$ individuals), Horn of Africa ($n = 107$ individuals), Eastern Africa ($n = 570$ individuals), Southern Africa ($n = 217$ individuals), Khoe and San ($n = 148$ individuals) and admixed Hispanic-American ($n = 184$ individuals). Secondly, ROH from 100 people in each group were resampled (with replacement) 100 times. Thirdly, statistically significant windows were obtained following the above methodology. Finally, consecutive windows found to be statistically significant in at least 50 resampling events were considered as part of the same ROHi.

In order to compare ROHi between populations it was considered that two ROHi from two different populations are indeed the same ROHi if they share at least 75% of their length.

Regions of heterozygosity (RHZ)

RHZ are defined as regions in the genome where < 5% of individuals in a population have ROH. In order to search for RHZ an extra step of QC consisting of removing the SNPs in LD using PLINK was performed before calling for ROH. For this analysis, ROH longer than 100 kb were called using 25 SNPs per window in PLINK. With this procedure all ROH longer than 100 kb, independent of their origin (LD or IBD), were detected with accuracy due to the SNP coverage available. Removing SNPs in LD, on average 1.1 M SNPs were still available for every population, enabling detection of ROH longer than 100 kb (2.8 kb per SNP, in 100 kb would be on average 35 SNPs, and a window of 25 SNPs is appropriate to cover genomic regions with less than the average number of SNPs). Once every ROH is called, it is straightforward to obtain regions outside ROH, and since SNPs in LD were pruned, these regions will be mostly heterozygous. In order to only identify informative heterozygous haplotypes, regions that have anomalous, unstructured, high signal/read counts in next generation sequence experiments were removed. These 226 regions, called ultra-high signal artefact regions, include high mapability islands, low mapability islands, satellite repeats, centromere regions, snRNA and telomeric regions (Consortium EP 2012). Regions not covered by the Human Omni Chip 2.5 were also removed from the analyses (e.g. *p* arms of chromosomes 13, 14, 15, 21 and 22). By moving a 100 kb window through the genome, two different cut-offs were considered to call RHZ in each window: no individual is homozygosity (RHZ 0%) or 5% or fewer of the individuals are homozygosity (RHZ 5%). Consecutive windows that fulfil this requirement were considered part of the same RHZ.

Acknowledgements FCC is a National Research Foundation of South Africa (NRF) postdoctoral fellow and MR holds a South African Research Chair in Genomics and Bioinformatics of African populations hosted by the University of the Witwatersrand, funded by the Department of Science and Technology and administered by the NRF. SH is partially supported by the NRF (IFR160214158079). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Compliance with ethical standards

Conflict of interest Authors declare that they have no competing interests.

Web resources 1000 Genomes database, http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/African Genome Variation Project dataset, http://www.ebi.ac.uk/ega/.

References

Alvarez G et al (2009) The role of inbreeding in the extinction of a European royal dynasty. PLoS One 4:e5174. https://doi.org/10.1371/journal.pone.0005174

Bekpen C et al (2017) Segmental duplications and evolutionary acquisition of UV damage response in the SPATA31 gene

family of primates and humans. BMC Genom 18:222. https://doi.org/10.1186/s12864-017-3595-8
Bittles AH (2012) Consanguinity in context. Cambridge University Press, Cambridge
Bledsoe C (2002) Contingent lives: fertility, time, and aging in West Africa. The University of Chicago Press, Chicago
Brandes U (2007) Optimal leaf ordering of complete binary trees. J Discret Algorithms 5:546–552. https://doi.org/10.1016/j.jda.2006.09.003
Broman KW, Weber JL (1999) Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am J Hum Genet 65:1493–1500. https://doi.org/10.1086/302661
Busby GB et al (2016) Admixture into and within sub-Saharan Africa. Elife. https://doi.org/10.7554/elife.15266
Caldwell JC et al (1992) The family and sexual networking in sub-Saharan Africa: historical regional differences and present-day implications. Popul Stud 46:385–410
Campbell MC, Tishkoff SA (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genom Hum Genet 9:403–433. https://doi.org/10.1146/annurev.genom.9.081307.164258
Ceballos FC et al (2018a) Assessing runs of homozygosity: a comparison of SNP Array and whole genome sequence low coverage data. BMC Genom 19:106. https://doi.org/10.1186/s12864-018-4489-0
Ceballos FC et al (2018b) Runs of homozygosity: windows into population history and trait architecture. Nat Rev Genet 19:220–234. https://doi.org/10.1038/nrg.2017.109
Chen H et al (2010) Population differentiation as a test for selective sweeps. Genome Res 20:393–402. https://doi.org/10.1101/gr.100545.109
Chimusa ER et al (2015) A genomic portrait of haplotype diversity and signatures of selection in indigenous southern African populations. PLoS Genet 11:e1005052. https://doi.org/10.1371/journal.pgen.1005052
Choudhury A et al (2017) Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. Nat Commun 8:2062. https://doi.org/10.1038/s41467-017-00663-9
Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. https://doi.org/10.1038/nature11247
Curtis D (2007) Extended homozygosity is not usually due to cytogenetic abnormality. BMC Genet 8:67. https://doi.org/10.1186/1471-2156-8-67
Curtis D et al (2008) Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. Ann Hum Genet 72:261–278. https://doi.org/10.1086/503875
Daya M et al (2013) A panel of ancestry informative markers for the complex five-way admixed South African coloured population. PLoS One 8:e82224. https://doi.org/10.1371/journal.pone.0082224
Fu X et al (2006) cAMP-response element-binding protein mediates acid-induced NADPH oxidase NOX5-S expression in Barrett esophageal adenocarcinoma cells. J Biol Chem 281:20368–20382. https://doi.org/10.1074/jbc.m603353200
Gandini F et al (2016) Mapping human dispersals into the Horn of Africa from Arabian Ice Age refugia using mitogenomes. Sci Rep 6:25472. https://doi.org/10.1038/srep25472
Gibson J et al (2006) Extended tracts of homozygosity in outbred human populations. Hum Mol Genet 15:789–795. https://doi.org/10.1093/hmg/ddi493
Grossman SR et al (2013) Identifying recent adaptations in large-scale genomic data. Cell 152:703–713. https://doi.org/10.1016/j.cell.2013.01.035
Gunderson RC (1980) Connecting your pedigree into royal, noble and medieval families. Genealogical Society of Utah, Salt Lake City
Gurdasani D et al (2015) The African genome variation project shapes medical genetics in Africa. Nature 517:327–332. https://doi.org/10.1038/nature13997
H3 Africa Consortium et al (2014) Research capacity. Enabling the genomic revolution in Africa. Science 344:1346–1348. https://doi.org/10.1126/science.1251546
Hampshire KR, Smith MT (2001) Consanguineous marriage among the fulani. Hum Biol 73:597–603
Hartl DL, Clark AG (2007) Principles of population genetics. Sinauer Associates, Sunderland
Henn BM et al (2011) Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci USA 108:5154–5162. https://doi.org/10.1073/pnas.1017511108
Henn BM et al (2016) Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc Natl Acad Sci USA 113:E440–E449. https://doi.org/10.1073/pnas.1510805112
Higasa K et al (2009) Evaluation of haplotype inference using definitive haplotype data obtained from complete hydatidiform moles, and its significance for the analyses of positively selected regions. PLoS Genet 5:e1000468. https://doi.org/10.1371/journal.pgen.1000468
Hollfelder N et al (2017) Northeast African genomic variation shaped by the continuity of indigenous groups and Eurasian migrations. PLoS Genet 13:e1006976. https://doi.org/10.1371/journal.pgen.1006976
Howrigan DP et al (2011) Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. BMC Genom 12:460. https://doi.org/10.1371/journal.pone.0028787
International HapMap C et al (2007) A second generation human haplotype map of over 31 million SNPs. Nature 449:851–861. https://doi.org/10.1038/nature06258
Jacobs T et al (2002) Murine malaria is exacerbated by CTLA-4 blockade. J Immunol 169:2323–2329. https://doi.org/10.4049/jimmunol.169.5.2323
Jacquard A (1975) Inbreeding—one word, several meanings. Theor Popul Biol 7:338–363. https://doi.org/10.1016/0040-5809(75)90024-6
Joshi PK et al (2015) Directional dominance on stature and cognition in diverse human populations. Nature 523:459–462. https://doi.org/10.1016/j.ajhg.2015.06.001
Kim HL et al (2014) Khoisan hunter-gatherers have been the largest population throughout most of modern-human demographic history. Nat Commun 5:5692. https://doi.org/10.1038/ncomms6692
Kirin M et al (2010) Genomic runs of homozygosity record population history and consanguinity. PLoS One 5:e13996. https://doi.org/10.1534/genetics.111.130922
Lauriat TL et al (2006) Characterization of KIAA0513, a novel signaling molecule that interacts with modulators of neuroplasticity, apoptosis, and the cytoskeleton. Brain Res 1121:1–11. https://doi.org/10.1016/j.brainres.2006.08.099
Lencz T et al (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc Natl Acad Sci USA 104:19942–19947. https://doi.org/10.1073/pnas.0710021104
Lesthaeghe R et al (1989) The nuptiality regimens in Sub-Saharan Africa. In: Lesthaeghe R (ed) REproduction and social organization in sub-Saharan Africa. University of California Press, Berkeley
Liu X et al (2013) Detecting and characterizing genomic signatures of positive selection in global populations. Am J Hum Genet 92:866–881. https://doi.org/10.1016/j.ajhg.2013.04.021
Lopez Herraez D et al (2009) Genetic variation and recent positive selection in worldwide human populations: evidence from nearly

1 million SNPs. PLoS One 4:e7888. https://doi.org/10.1371/journal.pone.0007888
Lopman B, Gregson S (2008) When did HIV incidence peak in Harare, Zimbabwe?. Back-calculation from mortality statistics. PLoS One 3:e1711. https://doi.org/10.1371/journal.pone.0001711
Machado P et al (2010) Malaria: looking for selection signatures in the human PKLR gene region. Br J Haematol 149:775–784. https://doi.org/10.1111/j.1365-2141.2010.08165.x
Martin AR et al (2017) Human demographic history impacts genetic risk prediction across diverse populations. Am J Hum Genet 100:635–649. https://doi.org/10.1016/j.ajhg.2017.03.004
Martin AR et al (2019) Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet 51:584–591. https://doi.org/10.1038/s41588-019-0379-x
Martins GA et al (2001) Fas–FasL interaction modulates nitric oxide production in *Trypanosoma cruzi*-infected mice. Immunology 103:122–129. https://doi.org/10.1046/j.1365-2567.2001.01216.x
McQuillan R et al (2008) Runs of homozygosity in European populations. Am J Hum Genet 83:359–372. https://doi.org/10.1007/s10048-009-0182-4
McQuillan R et al (2012) Evidence of inbreeding depression on human height. PLoS Genet 8:e1002655. https://doi.org/10.1371/journal.pgen.1002655
Mendizabal I et al (2012) Adaptive evolution of loci covarying with the human African pygmy phenotype. Hum Genet 131:1305–1317. https://doi.org/10.1007/s00439-012-1157-3
Montinaro F et al (2015) Unravelling the hidden ancestry of American admixed populations. Nature Commun 6:6596. https://doi.org/10.1038/ncomms7596
Nothnagel M et al (2010) Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Hum Mol Genet 19:2927–2935. https://doi.org/10.1093/hmg/ddq198
Oleksyk TK et al (2008) Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. PLoS One 3:e1712. https://doi.org/10.1371/journal.pone.0001712
Opal P et al (2003) Mapmodulin/leucine-rich acidic nuclear protein binds the light chain of microtubule-associated protein 1B and modulates neuritogenesis. J Biol Chem 278:34691–34699. https://doi.org/10.1074/jbc.m302785200
Pagani L et al (2012) Ethiopian genetic diversity reveals linguistic stratification and complex influences on the ethiopian gene pool. Am J Hum Genet 91:83–96. https://doi.org/10.1016/j.ajhg.2012.05.015
Patin E et al (2014) The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nat Commun 5:3163. https://doi.org/10.1038/ncomms4163
Pemberton TJ et al (2012) Genomic patterns of homozygosity in worldwide human populations. Am J Hum Genet 91:275–292. https://doi.org/10.1016/j.ajhg.2012.08.030
Pickrell JK et al (2014) Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci USA 111:2632–2637. https://doi.org/10.1073/pnas.1313787111
Popejoy AB, Fullerton SM (2016) Genomics is failing on diversity. Nature 538:161–164. https://doi.org/10.1038/538161a
Purcell S et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575. https://doi.org/10.1086/519795
Ramsay M et al (2016) H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. Global Health Epidemiol Genom 1:e20. https://doi.org/10.1017/gheg.2016.17
Saha N, El Sheikh FS (1988) Inbreeding levels in Khartoum. J Biosoc Sci 20:333–336
Schafer ZT et al (2006) Enhanced sensitivity to cytochrome c-induced apoptosis mediated by PHAPI in breast cancer cells. Cancer Res 66:2210–2218. https://doi.org/10.1158/0008-5472.can-05-3923
Schapera I (1957) Marriage of near Kin among the Tswana. J Int Afr Stud 27:139–159. https://doi.org/10.2307/1156807
Schlebusch CM, Jakobsson M (2018) Tales of human migration, admixture, and selection in Africa. Annu Rev Genom Hum Genet 19:405–428. https://doi.org/10.1146/annurev-genom-083117-021759
Schlebusch CM et al (2012) Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science 338:374–379. https://doi.org/10.1126/science.1227721
Schuster SC et al (2010) Complete Khoisan and Bantu genomes from southern Africa. Nature 463:943–947. https://doi.org/10.1038/nature08795
Scott-Emuakpor AB (1974) The mutation load in an African population. I. an analysis of consanguineous marriages in Nigeria. Am J Hum Genet 26:674–682
Silano M et al (2007) A decapeptide from durum wheat prevents celiac peripheral blood lymphocytes from activation by gliadin peptides. Pediatr Res 61:67–71. https://doi.org/10.1203/01.pdr.0000250173.88049.79
Slatkin M (2008) Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–485. https://doi.org/10.1038/nrg2361
Sudmant PH et al (2015) An integrated map of structural variation in 2504 human genomes. Nature 526:75–81. https://doi.org/10.1038/nature15394
Tanner RE (1958) Fertility and child mortality in cousin marriages. A study in a moslem community in east Africa. Eugen Rev 49:197–199
Templeton AR, Read B (1996) Inbreeding, one word, several meanings, much confusion. Biol Conserv 75:91–105
Wagh K et al (2012) Lactase persistence and lipid pathway selection in the Maasai. PLoS One 7:e44751. https://doi.org/10.1371/journal.pone.0044751
Wang ET et al (2006) Global landscape of recent inferred Darwinian selection for *Homo sapiens*. Proc Natl Acad Sci USA 103:135–140. https://doi.org/10.1073/pnas.0509691102
Weir BS (2012) Estimating *F*-statistics: a historical view. Br J Philos Sci 79:637–643. https://doi.org/10.1086/667904