

# Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries

Caitlin Uren,\* Minju Kim,<sup>†</sup> Alicia R. Martin,<sup>\*,§</sup> Dean Bobo,<sup>†</sup> Christopher R. Gignoux,\*\* Paul D. van Helden,\*  
Marlo Möller,\* Eileen G. Hoal,<sup>\*,1,2</sup> and Brenna M. Henn<sup>†,1,2</sup>

\*South African Medical Research Council Centre for Tuberculosis Research, Department of Science and Technology/National Research Foundation Centre of Excellence for Biomedical Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, 8000, South Africa, <sup>†</sup>Department of Ecology and Evolution, Stony Brook University, New York 11794, <sup>‡</sup>Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, <sup>§</sup>Program in Medical and Population Genetics, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, and \*\*Department of Genetics, Stanford University, California 94305

**ABSTRACT** Recent genetic studies have established that the KhoeSan populations of southern Africa are distinct from all other African populations and have remained largely isolated during human prehistory until ~2000 years ago. Dozens of different KhoeSan groups exist, belonging to three different language families, but very little is known about their population history. We examine new genome-wide polymorphism data and whole mitochondrial genomes for >100 South Africans from the ≠Khomani San and Nama populations of the Northern Cape, analyzed in conjunction with 19 additional southern African populations. Our analyses reveal fine-scale population structure in and around the Kalahari Desert. Surprisingly, this structure does not always correspond to linguistic or subsistence categories as previously suggested, but rather reflects the role of geographic barriers and the ecology of the greater Kalahari Basin. Regardless of subsistence strategy, the indigenous Khoe-speaking Nama pastoralists and the N!u-speaking ≠Khomani (formerly hunter-gatherers) share ancestry with other Khoe-speaking forager populations that form a rim around the Kalahari Desert. We reconstruct earlier migration patterns and estimate that the southern Kalahari populations were among the last to experience gene flow from Bantu speakers, ~14 generations ago. We conclude that local adoption of pastoralism, at least by the Nama, appears to have been primarily a cultural process with limited genetic impact from eastern Africa.

**KEYWORDS** ancestry; population structure; KhoeSan; pastoralism

The indigenous populations of southern Africa, referred to by the compound ethnicity “KhoeSan” (Schlebusch 2010), have received intense scientific interest. This interest is due both to the practice of hunter-gatherer subsistence among many groups—historically and to the present day—and genetic evidence suggesting that the ancestors of the KhoeSan diverged early on from all other African populations

(Behar *et al.* 2008; Tishkoff *et al.* 2009; Henn *et al.* 2011, 2012; Pickrell *et al.* 2012; Veeramah *et al.* 2012; Barbieri *et al.* 2013). Genetic data from KhoeSan groups have been extremely limited until very recently, and the primary focus has been on reconstructing early population divergence. Demographic events during the Holocene and the ancestry of the Khoekhoe-speaking pastoralists have received limited, mostly descriptive, attention in human evolutionary genetics. However, inference of past population history depends strongly on understanding recent population events and cultural transitions.

The KhoeSan comprise a widely distributed set of populations throughout southern Africa, speaking, at least historically, languages from one of three different linguistic families—all of which contain click consonants rarely found elsewhere. New genetic data indicate that there is deep population divergence even among KhoeSan groups (Pickrell *et al.* 2012; Schlebusch *et al.* 2012, 2013; Schlebusch and

Copyright © 2016 by the Genetics Society of America

doi: 10.1534/genetics.116.187369

Manuscript received January 20, 2016; accepted for publication July 7, 2016; published Early Online July 28, 2016.

Supplemental material is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187369/-/DC1>.

<sup>1</sup>These authors contributed equally to this work.

<sup>2</sup>Corresponding authors: Department of Ecology and Evolution, Life Sciences Bldg., Room 640, Stony Brook, NY 11794. E-mail: [brenna.henn@stonybrook.edu](mailto:brenna.henn@stonybrook.edu); and SA MRC Centre for TB Research, DST/NRF Centre of Excellence for Biomedical Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Tygerberg Campus, Parow, 7500, South Africa. E-mail: [egvh@sun.ac.za](mailto:egvh@sun.ac.za)

Soodyall 2012; Barbieri *et al.* 2013), with populations living in the northern Kalahari estimated to have split from southern groups 30,000–35,000 years ago (Pickrell *et al.* 2012; Schlebusch *et al.* 2012; Schlebusch and Soodyall 2012). Pickrell *et al.* (2012) estimate a time of divergence between the northwestern Kalahari and southeastern Kalahari population dating back to 30,000 years ago; “northwestern” refers to Juu-speaking groups like the !Xun and Ju|’hoansi, while “southeastern” refers to Taa speakers. In parallel, Schlebusch *et al.* (2012) also estimated an ancient time of divergence among the KhoeSan (dating back to 35,000 years ago), but here the southern groups include the ≠Khomani, Nama, Karretjie (multiple language families), and the northern populations refer again to the !Xun and Ju|’hoansi. Thus, KhoeSan populations are not only strikingly isolated from other African populations but they appear geographically structured among themselves. To contrast this with Europeans, the ≠Khomani and the Ju|’hoansi may have diverged >30,000 years ago but live only 1000 km apart, roughly the equivalent distance between Switzerland and Denmark whose populations have little genetic differentiation (Novembre *et al.* 2008). However, it is unclear how this ancient southern African divergence maps onto current linguistic and subsistence differences among populations, which may have emerged during the Holocene.

In particular, the genetic ancestry of the Khoe-speaking populations and specifically the Khoekhoe, (*e.g.*, Nama) who practice sheep, goat, and cattle pastoralism, remains a major open question. Archaeological data have been convened to argue for a demic migration of the Khoe from eastern African into southern Africa, but others have also argued that pastoralism represents cultural diffusion without significant population movement (Boonzaier 1996; MacDonald 2000; Robbins *et al.* 2005; Sadr 2008, 2015; Dunne *et al.* 2012; Pleurdeau *et al.* 2012; Jerardino *et al.* 2014). Lactase persistence alleles are present in KhoeSan groups, especially frequent in the Nama (20%), and clearly derive from eastern African pastoralist populations (Breton *et al.* 2014; Macholdt *et al.* 2014). This observation, in conjunction with other Y-chromosome and autosomal data (Henn *et al.* 2008; Pickrell *et al.* 2014), has been used to argue that pastoralism in southern Africa was another classic example of demic diffusion. However, the previous work is problematic in that it tended to focus on single loci (MCM6/LCT, Y chromosome), subject to drift or selection. Estimates of eastern African autosomal ancestry in the KhoeSan remain minimal (<10%) and the distribution of ancestry informative markers is dispersed between both pastoralist and hunter-gatherer populations. Here, we present a comprehensive study of recent population structure in southern Africa and clarify fine-scale structure beyond “northern” and “southern” geographic descriptors. We then specifically test whether the Khoe-speaking Nama pastoralists derive their ancestry from eastern Africa, the northeastern Kalahari Basin, or far southern Africa. Our results suggest that ecological features of southern Africa, broadly speaking, are better explanatory features

than either language, clinal geography, or subsistence on its own.

## Materials and Methods

### Sample collection and ethical approval

DNA samples from the Nama, ≠Khomani San, and South African Colored populations were collected with written informed consent and approval of the Human Research Ethics Committee of Stellenbosch University (N11/07/210), South Africa, and Stanford University (protocol 13829). Community level results were returned to the communities in 2015 prior to publication. A contract for this project was approved by the Working Group of Indigenous Minorities in Southern Africa (ongoing).

### Autosomal data and genotyping platforms

Two primary datasets were used: A) ~565,000 SNPs on the Affymetrix Axiom Genome-wide Human Origins Array derived from Pickrell *et al.* (2012), Lazaridis *et al.* (2014), with additional ≠Khomani San and Hadza individuals from our collections for a total of 33 populations and 396 individuals. B) ~320,000 SNPs from the intersection of HGDP (Illumina 650Y) (Li *et al.* 2008), HapMap3 (joint Illumina Human 1M and Affymetrix SNP 6.0), Illumina OmniExpressPlus and OmniExpress SNP array platforms generated here, as well as the dataset from Petersen *et al.* (2013) for a total of 21 populations and 852 individuals.

### Population structure

ADMIXTURE (Alexander *et al.* 2009) was used to estimate the ancestry proportions via a model-based approach. Iterations through various *k* values are necessary. The *k* value is an estimate of the number of original ancestral populations. Cross-validation (CV) was performed by ADMIXTURE and these values were plotted to acquire the *k* value that was the most stable. Depiction of the Q matrix was performed in R. Ten iterations were performed for each *k* value with 10 random seeds. Iterations were grouped according to admixture patterns to identify the major and minor modes by pong (Behr *et al.* 2015). These Q matrices from ADMIXTURE, as well as longitude and latitude coordinates for each population were adjusted to the required format for use in an R script supplied by Ryan Raaum to generate the surface maps (Figure 2).

### Estimating Effective Migration Surfaces (EEMs) analysis

Estimating Effective Migration Surfaces (EEMs) analyses (Petkova *et al.* 2016) were run on the Affymetrix Human Origins data set. Genetic dissimilarities were calculated using the bed2diffs script and EEMs was run using the run\_eems\_snps version of the program. A grid is constructed so as to house all demes in the data provided. Each individual is assigned to a specific deme. Using a stepping stone model, migration rates between demes are calculated. Genetic dissimilarities are calculated fitting an “isolation-by-distance

model.” In order for the MCMC iterations to converge, the number of MCMC iterations, burn iterations, and thin iterations were increased. The other parameters were optimized as per the manual’s recommendations, *i.e.*, diversity and migration parameters were adjusted so as to produce 20–30% acceptance rates. The PopGPlot R package was used to visualize the data.

### **Association between $F_{st}$ , geography, and language**

A Mantel test ( $F_{st}$  and geographic distance) and a partial Mantel test ( $F_{st}$  and language, accounting for geographic distance) were performed using the vegan package in R. Geographic distances (in kilometers) between populations were calculated using latitude and longitude values as tabulated in Supplemental Material, Table S1. Weir and Cockerham genetic distances ( $F_{st}$ ) were calculated from allele frequencies estimated with vcftools (Danecek *et al.* 2011). A Jaccard phonemic distance matrix was used as formulated in Creanza *et al.* (2015). Populations included in the analysis were the Nama, ≠Khomani, East Taa, West Taa, Naro, G|ui, G||ana, Shua, Kua, !Xuun, and Khwe.

### **Mitochondrial DNA network**

We utilized Network (ver. 4.6, copyrighted by Fluxus Technology), for a median-joining phylogenetic network analysis in order to produce Figure 5 and Figure S6. Network Publisher (ver. 2.0.0.1, copyrighted by Fluxus Technology) was then used to draw the phylogenetic relationships among individuals.

### **Data availability**

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article. Data files are freely available on GitHub: [https://github.com/bmhenn/khoesan\\_arraydata](https://github.com/bmhenn/khoesan_arraydata).

## **Results**

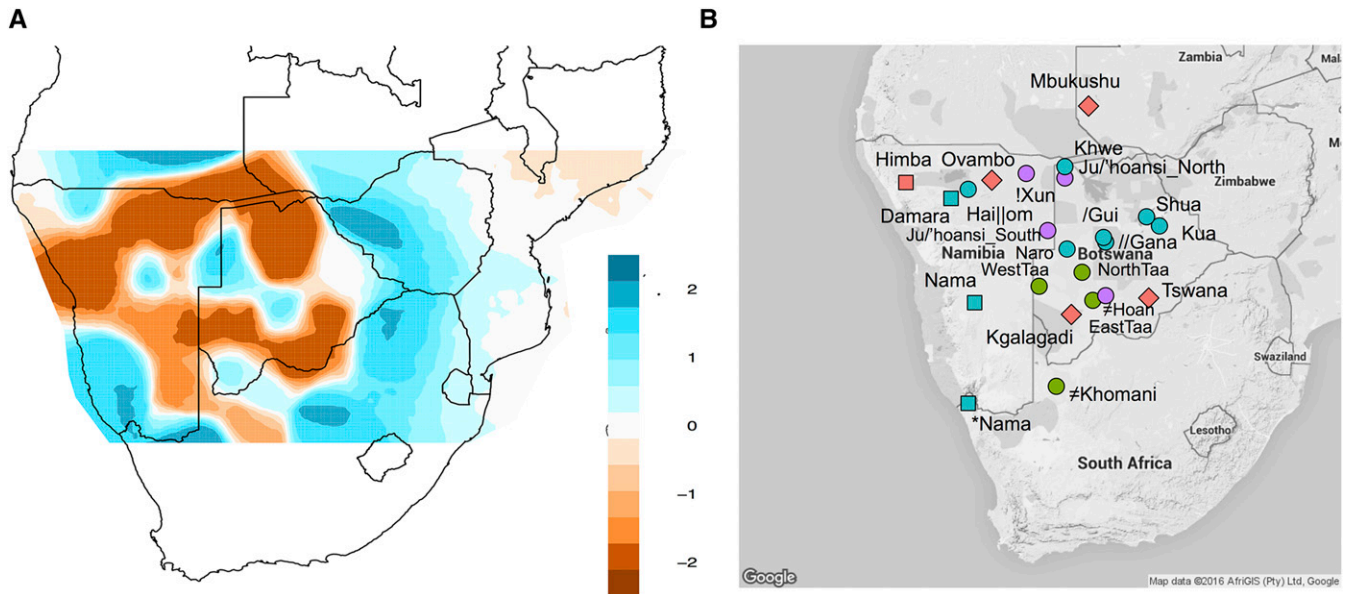
To resolve fine-scale population structure and migration events in southern Africa, we generated genome-wide data from three South African populations. We genotyped ≠Khomani San ( $n = 75$ ), Nama ( $n = 13$ ), and South African Colored (SAC) ( $n = 25$ ) individuals on the Illumina OmniExpress and OmniExpressPlus SNP array platforms. Sampling locations are listed in Table S1, in addition to language groupings and subsistence strategies. These data were merged with HapMap3 (joint Illumina Human1M and Affymetrix SNP 6.0) (International HapMap 3 Consortium *et al.* 2010), HGDP (Illumina 650Y) data (Li *et al.* 2008), and Illumina HumanOmni1-Quad (Petersen *et al.* 2013), resulting in an intersection of ~320,000 SNPs for 852 individuals from 21 populations. In addition, we used the Affymetrix Human Origins SNP Array generated as part of Pickrell *et al.* (2012) and Lazaridis *et al.* (2014), including  $n = 9$  ≠Khomani San individuals from our collection and encompassing >396 individuals from 33 populations. Whole mitochondrial ge-

nomes were generated from off-target reads from exome- and Y-chromosome capture short read Illumina sequencing. Reads were mapped to GRCh37, which uses the revised Cambridge reference sequence. Only individuals with >7× haploid coverage were included in the analysis: ≠Khomani San ( $n = 64$ ) and Nama ( $n = 31$ ); haplogroup frequencies were corrected for pedigree structure (Table S2). In this study, we address population structure among southern African Khoe-San, the genetic affinity of the Khoe, and how pastoralism diffused into southern Africa.

### **Population structure in southern African KhoeSan populations**

We first tested whether southern African populations conform to an isolation-by-distance model, or whether there is strong heterogeneity among populations relative to geographic distance. Using 22 southern African populations (with 560,000 SNPs from Affymetrix Human Origins array), we implemented the spatially explicit program EEMs (Petkova *et al.* 2016) to test for effective migration patterns across the region. We observe a higher effective migration rate ( $m$ ) in the central Kalahari Basin relative to a lower migration rate that forms a rim around the Kalahari Desert (Figure 1). A second resistance band stretches across northern Namibia, indicating higher gene flow above northern Namibia, Angola, and southern Zambia. Differences in effective migration rates can result from differences in effective population sizes. For example, a larger effective population size can result in higher effective migration rates, relative to neighboring demes, with smaller  $N_e$ ’s. The higher  $m$  in the central Kalahari Basin, relative to the rim, could result from either a larger  $N_e$  relative to Kalahari rim populations or simply higher migration among groups in a similar ecological area.

We then tested whether heterogeneity in population structure could be mapped to distinct genetic ancestries. Unsupervised population structure analysis identifies five distinct, spatially organized ancestries among the sampled 22 southern African populations. These ancestries were inferred from the Affymetrix Human Origins data set using ADMIXTURE (Figure S1) (Alexander *et al.* 2009). Multimodality per  $k$  value was assessed using pong (Behr *et al.* 2015) and results from  $k = 10$  are discussed below (6/10 runs assigned to the major mode, 3/10 other runs involved cluster switching only within East Africa). Visualization of these ancestries according to geographic sampling location specifically demonstrates fine-scale structure in and around the Kalahari Desert (Figure 2). While prior studies have argued for a northern vs. southern divergence of KhoeSan populations (Pickrell *et al.* 2012; Schlebusch *et al.* 2012; Schlebusch and Soodyall 2012; Barbieri *et al.* 2013, 2014), the structure inferred from our data set indicates a more geographically complex pattern of divergence and gene flow. Even recent migration events into southern Africa remain structured, consistent with ecological boundaries to gene flow (see below). The distribution of the five ancestries corresponds to: a northern Kalahari ancestry, central Kalahari ancestry, circum-Kalahari ancestry, a



**Figure 1** Effective migration rates among 22 southern African populations. (A) Using southern African samples from the Affymetrix HumanOrigins data set, we estimated effective migration rates among populations using EEMs. White indicates the mean expected migration rate across the data set, while blue indicates X-fold increase in migration among demes, and brown indicates decreased migration among demes (e.g., population structure). Effective migration rates,  $e_m$ , are plotted on a log scale as in Petkova *et al.* (2016). Hence,  $-1e_m$  would indicate 10-fold decrease in the migration rate relative to the expected rate among all demes accounting for geographic distance. These results demonstrate that southern Africa is a heterogeneous environment with barriers to gene flow in northwest Namibia and the Kalahari rim, but increased gene flow within the Kalahari Basin. The grid of plotted demes was restricted to prevent unwanted extrapolation to poorly sampled areas. (B) The topographic map indicates the subsistence strategy and language of each population sample. Colors represent language families: green, Tuu speakers; red, Niger-Congo speakers; blue, Khoe speakers; and purple, Kx'a speakers. Shapes represent subsistence strategies: circle, hunter-gatherers; square, pastoralists; and diamond, agropastoralists. \*Nama indicates a new, second Nama sample from South Africa, which was only included in Illumina SNP array analyses.

northwestern Namibian savannah ancestry, and ancestry from eastern Bantu speakers (Figure 2). This geographic patterning does not neatly correspond to linguistic or subsistence categories, in contrast to previous discussions (Pickrell *et al.* 2012; Schlebusch *et al.* 2012; Barbieri *et al.* 2014).

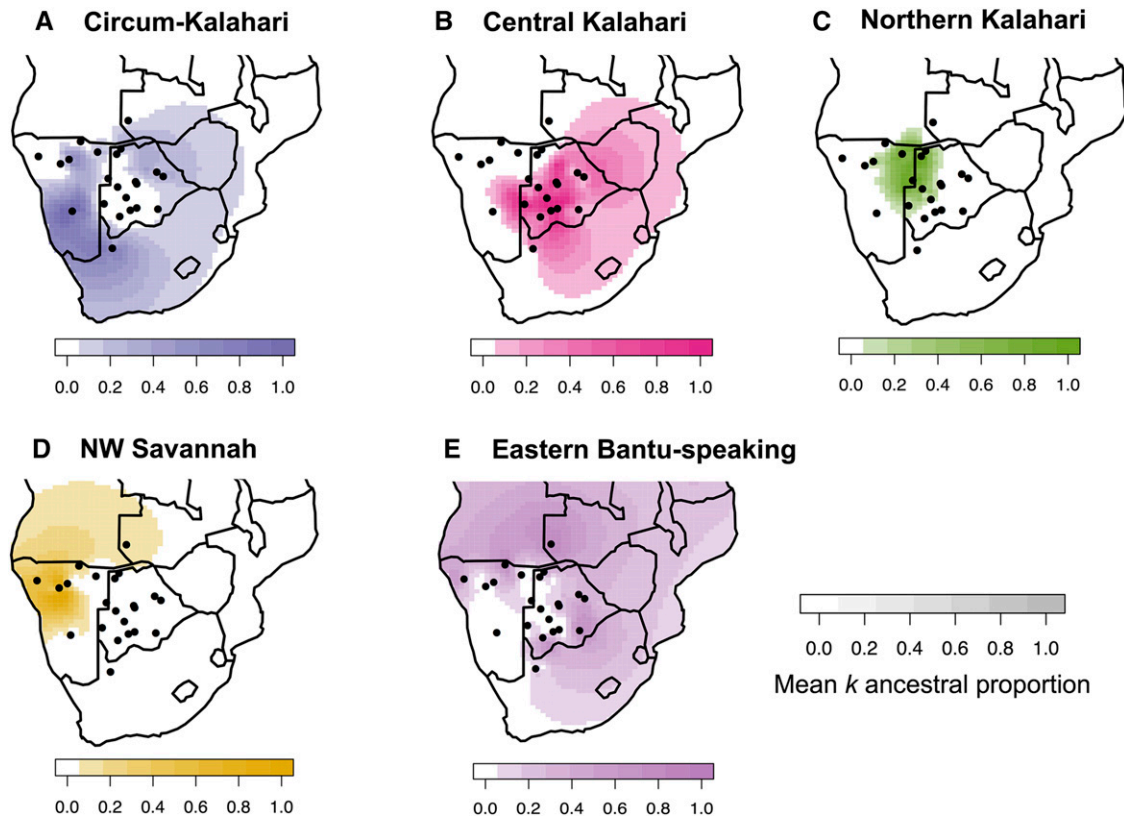
The northern Kalahari ancestry is the most defined of these ancestries, encompassing several forager populations such as the Ju/'hoansi, !Xun, Khwe, Naro, and to a lesser extent the Khoekhoe-speaking Hai||om. While these populations are among the best-studied Khoesan in anthropological texts with particular reference to cultural similarities (Dornan 1925; Bleek 1928; Schapera 1934; Barnard 1992), they represent only a fraction of the diversity among Khoisan-speaking populations. We note that this cluster includes Kx'a (Juu), Khoe-Kwadi, and Khoekhoe speakers, suggesting that language interacts in a complex fashion with other factors such as subsistence strategy and ecology. The Hai||om are thought to have shifted to speaking Khoekhoe from an ancestral Juu-based language (Barnard 1992). The second, central Kalahari ancestry, occupies a larger geographical area throughout the Kalahari Basin, with its highest frequency among the Taa speakers: G|ui, G||ana, ≠Hoan, and Naro. This ancestry spans all three Khoisan language families (Table S1), at considerable frequency in each; all are primarily foragers.

The third ancestry cluster is represented by southern Khoesan populations distributed along the rim of the Kalahari

Desert (Figure 2)—referred to here as the “circum-Kalahari ancestry.” The circum-Kalahari ancestry is at its highest frequency in the Nama and ≠Khomani (see also Figure S2), with significant representation in the Hai||om, Khwe, !Xun, and Shua. This ancestry spans all linguistic and subsistence strategies. We propose that the circum-Kalahari is better explained by ecology than alternative factors such as language or recent migration. Specifically, we find the Kalahari Desert is an ecological boundary to gene flow (Figure 1, Figure 2). The circum-Kalahari ancestry is not easily explained by a pastoralist Khoekhoe dispersal. This spatially distinct ancestry is common in both forager and pastoralist groups, indeed all of the circum-Kalahari populations were historically foragers (except for the Nama). Therefore, to support a Khoekhoe dispersal model, we would have to posit an adoption of pastoralism by a northeastern group, leading to demic expansion around the Kalahari, with subsequent reversion to foraging in the majority of the circum-Kalahari groups; this scenario seems unlikely (but see Smith 2014 for additional discussion).

Finally, our analysis reveals two additional ancestries outside of the greater Kalahari Basin: one ancestry composed of Bantu speakers, frequent to the north, east, and southeast of the Kalahari; and a second composed of Himba, Ovambo, and Damara ancestry in northwestern Namibia distributed throughout the mopane savannah. Interestingly, the Damara are a Khoekhoe-speaking population of former foragers (later





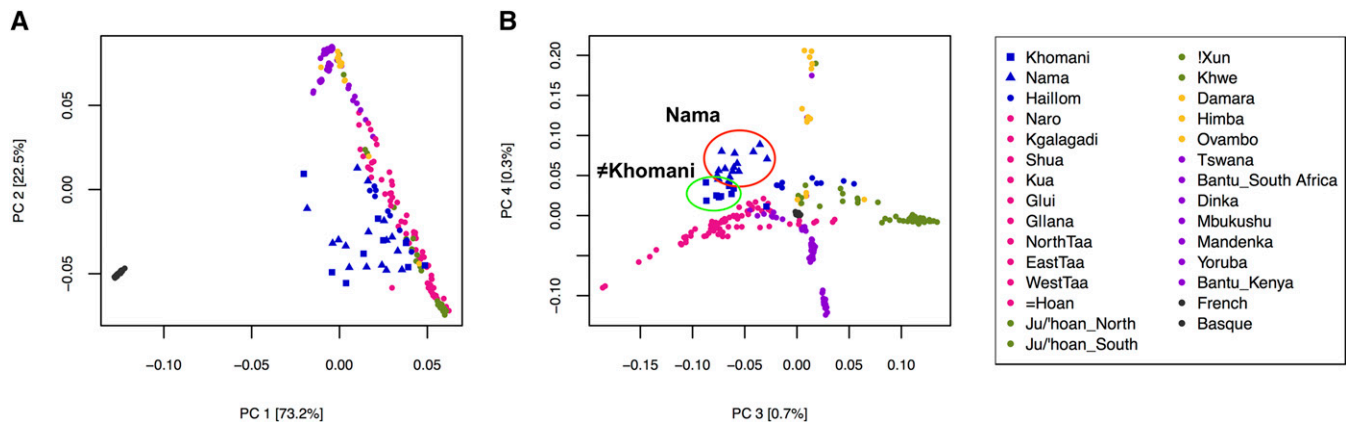
**Figure 2** Five spatially distinct ancestries indicate deep population structure in southern Africa. Using global ancestry proportions inferred from ADMIXTURE  $k = 10$ , we plot the mean ancestry for each population in southern Africa. The five most common ancestries in southern Africa, from the Affymetrix HumanOrigins data set, are shown separately in A–E. The x- and y-axes for each map correspond to latitude and longitude, respectively. Black dots represent the sampling location of populations in southern Africa. The third dimension in each map (depth of color) represents the mean ancestry proportion for each group for a given  $k$  ancestry, calculated from ADMIXTURE using unrelated individuals, and indicated in the color keys as 0–100% for five specific  $k$  ancestries. Surface plots of the ancestry proportions were interpolated across the African continent.

in servitude to the Nama pastoralists) whose ancestry has been unclear (see below).

We used our data and the Affymetrix HumanOrigins data set containing the greatest number of KhoeSan populations to date, to test whether language or geography better explains genetic distance (see language families and subsistence strategies in Table S1). The genetic data were compared to a phonemic distance matrix (Jaccard 1908) as well as geographic distances between each population (Table S3). In order to test whether genetic distance ( $F_{st}$ ) was associated with geography or language, we performed a partial Mantel test for the relationship between  $F_{st}$  and language (Creanza *et al.* 2015) accounting for geographic distance among 11 KhoeSan populations. This result was not significant ( $r = 0.06$ ,  $P = 0.30$ ). Although an association between  $F_{st}$  and geographic distance within Africa has been documented (Ramachandran *et al.* 2005; Tishkoff *et al.* 2009; Creanza *et al.* 2015), a Mantel test for the relationship between  $F_{st}$  and pairwise geographic distance in our data set was also null ( $r = 0.021$ ,  $P = 0.38$ ), reflecting the nonlinear aspect of shared ancestry in southern Africa as seen in Figure 1 and Figure 2.

Spatially distinct ancestries are also supported by principal components analysis (PCA) (Figure 3, Figure S3). The Khoe-

San anchor one end of PC1 opposite to Eurasians. PC2 separates other African populations from the KhoeSan, including western Africans, as well as central and eastern African hunter-gatherers. PC3 separates the Ju|'hoansi and !Xun (northern Kalahari) from ≠Hoan, Taa speakers and Khoe speakers, with other KhoeSan populations intermediate. PC3 and PC4 suggest that the present language distribution may reflect recent language transitions, as genetic ancestry and linguistic structure do not neatly map onto each other (Figure S4). For example, the ≠Hoan currently speak a Kx'a language but are genetically distinct from other northern Kalahari Kx'a speakers; rather, they appear to be more genetically similar to southern Kalahari Taa speakers who cluster together. We suggest that the patterns observed here are better explained by ecogeographic patterns than either language or subsistence alone (Figure S5). Specifically, PC3 discriminates northern vs. southern Kalahari ancestry (see below). PC4 discriminates western and eastern non-KhoeSan ancestry derived from Bantu speakers or other populations. Finally, the intermediate position of the Nama, ≠Khomani, and Hai|om on PC3 and PC4 is neither linguistic- nor subsistence based, but represents a nonlinear circum-Kalahari component featured in Figure 2.



**Figure 3** Clustering of KhoeSan populations and fine-scale population structure between the Nama and ≠Khomani San. A PCA of the Affymetrix Human Origins data set depicts the clustering of unrelated individuals based on the variation seen in the data set. Colors mimic similar major ancestry colors as shown in Figure 2. Yellow denotes populations with majority northwestern Namibian ancestry; purple denotes populations with majority Bantu-speaking ancestry; pink indicates southern Kalahari majority ancestry, green indicates northern Kalahari majority ancestry, and blue indicates circum-Kalahari ancestry. The red and green circles denote the fine-scale separation of the Nama and ≠Khomani populations (specified by triangles and squares, respectively). Note that these colored ancestries and the PCs do not map onto subsistence neatly (Figure S5).

### A divergent southern KhoeSan ancestry

This separation of northern (Ju|'hoansi) and southern (Taa and Khoe speakers) KhoeSan populations has been observed by Schlebusch *et al.* (2012) and Pickrell *et al.* (2012). We estimate that this trans-Kalahari genetic differentiation from the inferred ancestral allele frequencies (Figure S2) is substantial ( $F_{st} = 0.05$ ). We verify this divergence between the northern Kx'a speakers and the shared Nama and ≠Khomani ancestry in a new, second sample of Nama, from South Africa rather than central Namibia (Table S1, Figure S3). This southern KhoeSan ancestry is also present in admixed Bantu-speaking populations from South Africa (e.g., amaXhosa) as well as the admixed Western Cape SAC populations (de Wit *et al.* 2010), supporting a hypothesis of distinct southern-specific KhoeSan ancestry (Figure S1, Figure S2) shared between indigenous and admixed groups.

Mitochondrial data support this concept of a southern-specific KhoeSan ancestry (Schlebusch *et al.* 2013; Barbieri *et al.* 2013). Both mitochondrial DNA (mtDNA) haplogroups L0d and L0k are at high frequency in northern KhoeSan populations (Behar *et al.* 2008), but L0k is absent in our sample of the Nama ( $n = 31$ ) and there is only one ≠Khomani individual ( $n = 64$ ) with L0k (1.56%) (Table 1). L0d dominates the haplogroup distribution for both the Nama and ≠Khomani (84 and 91%, respectively), with L0d2a especially common in both. L0d2a, inferred to have originated in southern Africa, was also previously found at high frequencies in the Karretjie people further south in the central Karoo of South Africa, as well as the SAC population in the Western Cape (Quintana-Murci *et al.* 2010; Schlebusch *et al.* 2013). L0d2b is also common in the Nama (16%).

### Minimal population structure between the Nama and ≠Khomani

The ≠Khomani San are a N|u-speaking (!Ui classified language) former hunter-gatherer population that inhabit the

southern Kalahari Desert in South Africa, bordering on Botswana and Namibia. The Nama, currently a primarily caprid pastoralist population, live in the Richtersveld along the northwestern coast of South Africa and up into Namibia. The ancestral geographic origin of the Nama has been widely contested over a number of years (Nurse and Jenkins 1977; Barnard 1992; Boonzaier 1996), but a leading hypothesis suggests that they originated further north in Botswana/Zambia and migrated into South Africa and Namibia ~2000 years ago (Nurse and Jenkins 1977; Barnard 1992; Boonzaier 1996; Pickrell *et al.* 2012). The Nama and N|u languages are in distinct, separate Khoisan language families [Khoe and Tuu (!Ui-Taa), respectively] and these groups historically utilized different subsistence strategies. For this reason, we hypothesized that there would be strong population structure between the two populations.

Our global ancestry results, inferred from ADMIXTURE, show minimal population structure between the Nama and ≠Khomani San in terms of their southern KhoeSan ancestry. The ≠Khomani share ~10% of their ancestry with the Botswana KhoeSan populations (Figure S1, Figure S3), consistent with their closer proximity to the southern Botswana populations (Taa speakers !Xo and ≠Hoan). PCA reveals a degree of fine-scale population structure between the Nama and ≠Khomani, with each population forming its own distinct cluster at PC4, partly due to the increase in Damara ancestry in the Nama (Figure 3B, Figure S1), but the two groups are clearly proximal. This increase in Damara ancestry (as depicted from  $k = 9$  in all modes of Figure S1) is likely due to integration of the Damara people as clients of the Nama over multiple generations. However, our second sample of Nama from South Africa do not harbor significant western African ancestry, suggesting heterogeneity in the Damara component (Figure S2).

**Table 1 Mitochondrial DNA haplogroup frequencies of the Nama and ≠Khomani**

		≠Khomani San		Nama	
Haplogroup	n	Frequency		n	Frequency
L0d	L0d1a	8	12.50%	3	9.68%
	L0d1a1	2	3.13%	3	9.68%
	L0d1b	4	6.25%	3	9.68%
	L0d1b1	9	14.06%	2	6.45%
	L0d1b2	2	3.13%	0	
	L0d1c1	1	1.56%	1	3.23%
	L0d1c1a	1	1.56%	1	3.23%
	L0d2a	25	39.06%	3	9.68%
	L0d2a1	0		2	6.45%
	L0d2b	0		5	16.13%
	L0d2c	5	7.81%	2	6.45%
	L0d3	1	1.56%	1	3.23%
L0f1	1	1.56%		0	
L0k1	1	1.56%		0	
L3'4	0			1	3.23%
L3d3a1a	0			1	3.23%
L3e1a2	1	1.56%		0	
L4b2a2	1	1.56%		1	3.23%
L5c	0			1	3.23%
M36	M36	1	1.56%	0	
	M36d1	1	1.56%	0	
M7c3c	0			1	3.23%
<b>Total (n)</b>	<b>64</b>	<b>100%</b>		<b>31</b>	<b>100%</b>

### Recent patterns of admixture in South Africa

Two Bantu-speaking, spatially distinct ancestries are present in southern Africa. The first is rooted in the Ovambo and Himba in northwestern Namibia; the other reflects gene flow from Bantu-speaking ancestry present in the east (Figure 2). We estimated the time intervals for admixture events into the southern KhoeSan via analysis of the distribution of local ancestry segments using RFMix (Maples *et al.* 2013) and TRACTs (Gravel 2012) for the ≠Khomani OmniExpress data set ( $n = 59$  unrelated individuals) (Figure 4, Table S2). The highest likelihood model suggests that there were three gene flow events. Approximately 14 generations ago ( $\sim 443$ –473 years ago assuming a generation time of 30 years and accounting for the age of our sampled individuals), the ≠Khomani population received gene flow from a Bantu-speaking group, represented here by the Kenyan Luhya. Our results are consistent with Pickrell *et al.* (2012) who found that the southern Kalahari Taa speakers were the last to interact with the expanding Bantu speakers  $\sim 10$ –15 generations ago. Subsequently, this event was followed by admixture with Europeans between 6 and 7 generations ago ( $\sim 233$ –263 years ago), after the arrival of the Dutch in the Cape and the resulting migrations of “trekboers” (nomadic pastoralists of Dutch, French, and German descent) from the Cape into the South African interior. Lastly, we find a recent pulse of primarily KhoeSan ancestry 4–5 generations ago ( $\sim 173$ –203 years ago). This event could be explained by gene flow into the ≠Khomani from another KhoeSan group, potentially as groups shifted local ranges in response to the expansion of European farmers in the Northern Cape, or other population movements in southern Namibia or Botswana.

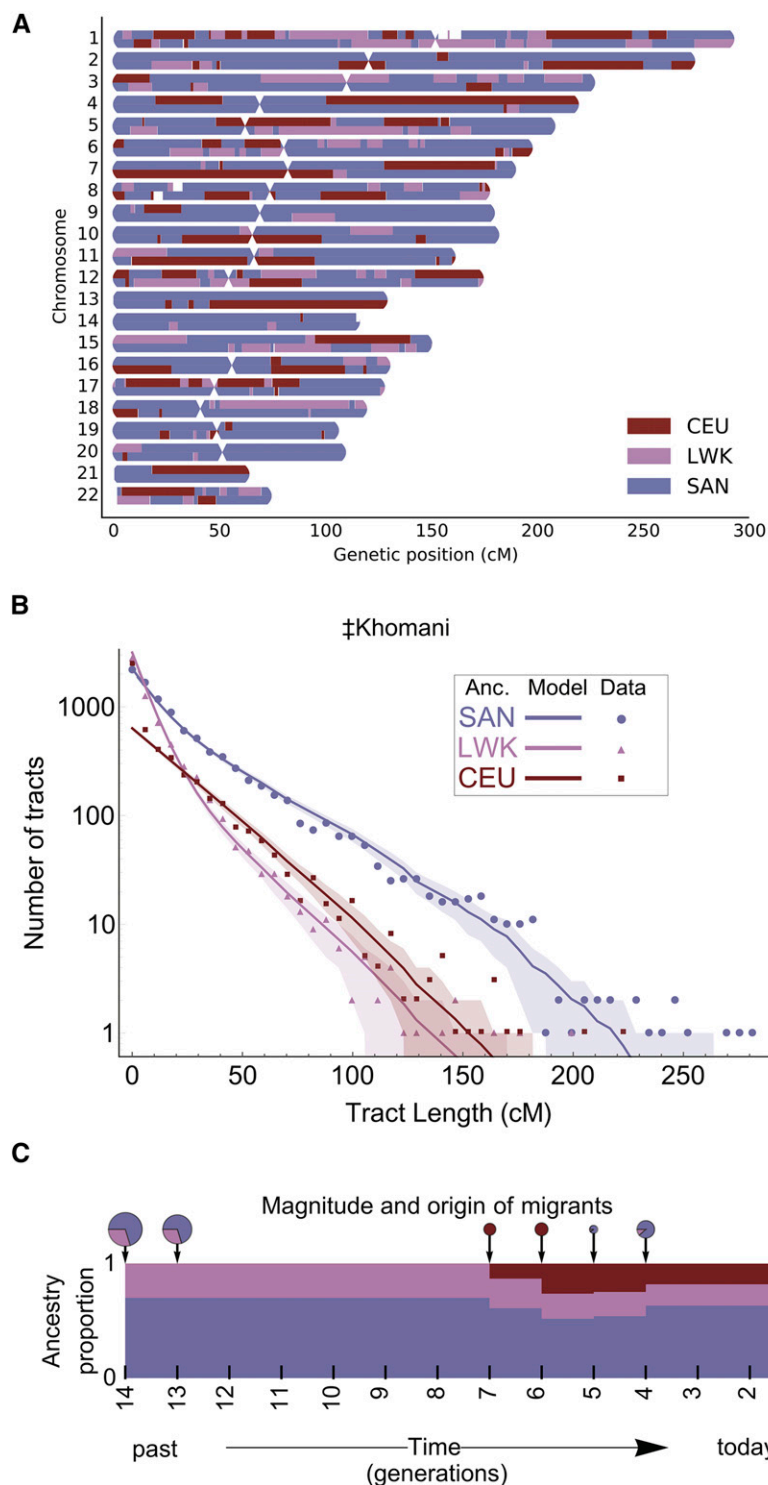
We also considered the impact of recent immigration into indigenous South Africans, derived from non-African source

populations. The SAC populations are a five-way admixed population, deriving ancestries from Europe, eastern African, KhoeSan, and Asian populations (de Wit *et al.* 2010). This unique, admixed ethnic population was founded by the Dutch who settled on the southern tip of South Africa by the 17th century and by the importation of slaves from Indonesia, Bengal, India, and Madagascar. However, within the SAC, strong differences in ancestry and admixture proportions are observed between different districts within Cape Town, the Eastern Cape, and the Northern Cape Provinces. SAC individuals from the Northern Cape, where historically there was a greater concentration of European settlement (Theal 1887), have higher European ancestry. The SAC individuals from the Eastern Cape, which is the homeland of the Bantu-speaking Xhosa populations, have relatively more ancestry from Bantu-speaking populations (Figure S2). The “ColouredD6” population is from an area in Cape Town called District 6. Historically, this was a district where the slaves and political exiles from present day Indonesia resided, as well as many who were from Madagascar and India based on written documentation (du Plessis 1947). The SAC D6 population consequently has a noticeable increase in south/eastern Asian ancestry represented by the Pathan and Han Chinese populations in our data set (Figure S2).

This south/eastern Asian ancestry is not confined to the SAC population, as attested by the presence of the M36 mitochondrial haplogroup. The M36 haplogroup (South Indian/Dravidian in origin) is present in two of 64 ≠Khomani San matrilineages (Table 1). The presence of M36 is likely derived from slaves of South Asian origin who escaped from Cape Town or the surrounding farms and dispersed into the northwestern region of South Africa. In addition, we observe one M7c3c lineage in the Nama (Table 1), which traces back to southeastern Asia but has been implicated in the Austronesian expansion of Polynesian speakers into Oceania (Kayser 2010; Delfin *et al.* 2012) and Madagascar (Poetsch *et al.* 2013). The importation of Malagasy slaves to Cape Town may best explain the observation of M7c3c in the Nama.

### Discussion

The KhoeSan are distinguished by their unique phenotype(s), genetic divergence, click languages, and hunter-gatherer subsistence strategy compared to other African populations; classifications of the many KhoeSan ethnic groups have primarily relied on language or subsistence strategy. Here, we generate additional genome-wide data from three South African populations and explore patterns of fine-scale population structure among 22 southern African groups. We find that complex geographic or “ecological” information is likely a better explanatory variable for genetic ancestry than language or subsistence. We identify five primary ancestries in southern Africans, each localized to a specific geographic region (Figure 2). In particular, we



**Figure 4** Demographic reconstruction of recent admixture in the ≠Khomani San using local ancestry. (A) Local ancestry karyogram for a representative three-way admixed ≠Khomani San individual was constructed using RFMix. Haplotypes for admixed individuals were assigned to one of three possible ancestries: SAN (Namibian San), LWK (Bantu-speaking Luhya from Kenya), or CEU (Central Europeans). UNK indicates unknown ancestry (*Materials and Methods*). (B) Markov models implemented in TRACTs to test multiple demographic models and assess the best fit to the observed ≠Khomani haplotype distributions. Local ancestry tract lengths were inferred as in A. (C) The tract length distribution for each ancestry across all individuals was used to estimate migration time (generations ago), volume of migrants, and ancestry proportions over time. Colored dots show the observed distribution of ancestry tracts for each ancestry, solid lines show the best fit from the most likely model, and shaded areas indicate confidence intervals corresponding to  $\pm 1$  SD.

examined the circum-Kalahari ancestry, which appears as a ring around the Kalahari Desert and accounts for the primary ancestry of the Nama, representative of the Khoekhoe-speaking pastoralists.

We observe striking ecogeographic population structure associated with the Kalahari Desert. There are two distinct ancestries segregating within the Kalahari Desert Khoesan populations, described here as northern Kalahari and central

Kalahari ancestries. Analyses of migration rates across the 22 populations indicate particularly high migration within the Kalahari Desert. This may indicate a larger effective population size for the two desert ancestries or extensive migration related to shifting ranges in response to climatic and ecological changes over time. It is worth noting that the northern Kalahari formerly supported an extensive lake (*i.e.*, Makgadikgadi) just before and after the Last Glacial



Maximum, as well as the presence of the Okavango Delta and associated river systems; archeological data may suggest high population density near the pans, although this likely predates the genetic structure we observe today (Burrough 2016; Robbins *et al.* 2016). Our lack of samples outside of Botswana, Namibia, and northern South Africa prevent precise inference of *m* in Zambia, Limpopo, and Mozambique; but Figure 2 indicates recent extensive gene flow in the east, consistent with the expansion of Bantu-speaking agriculturalists into eastern grasslands and coastal forests. Additionally, we find a separate ancestry segregating in the far western border of Namibia and Angola, particularly frequent in the Damara and Himba, and to a lesser extent in the Ovambo and Mbukushu. This intersection of steppe and savannah along the Kunene may have facilitated recent settlement of the area during the past 500 years by Bantu-speaking pastoralists, but it is noteworthy that little Kalahari KhoeSan ancestry persists in these populations. Rather, the Damara (currently Nama speaking) or related hunter-gatherers may have been formerly more widespread in this area and subsequently absorbed into the western Bantu-speaking pastoralists.

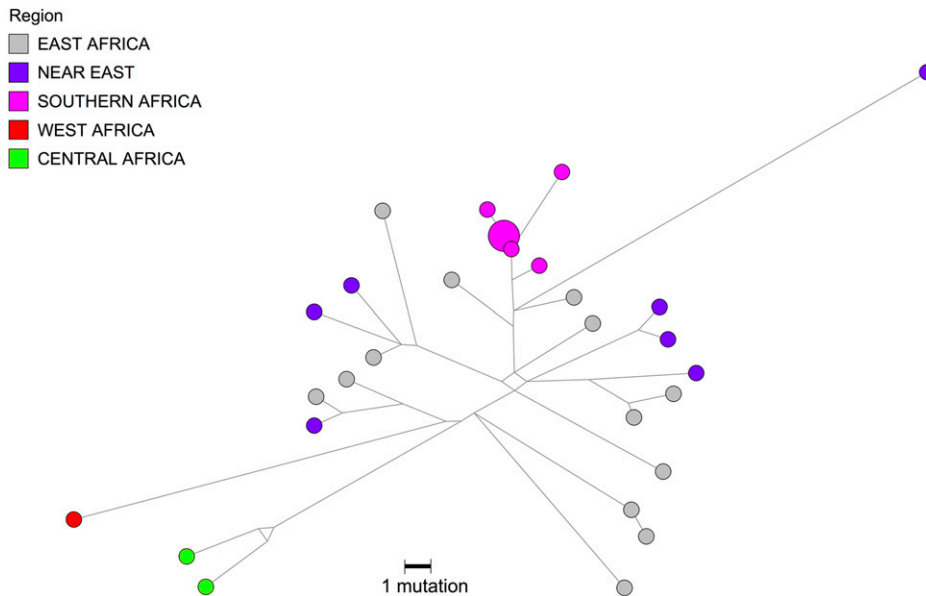
The practice of sheep, goat, and cattle pastoralism in Africa is widespread. Within KhoeSan populations, pastoralist communities are limited to the Khoekhoe-speaking populations. Earlier hypotheses proposed that the Khoekhoe-speaking pastoralists derived from a population originating outside of southern Africa. However, more recent genetic work supports a model of autochthonous Khoe ancestry influenced by either demic or cultural diffusion of pastoralism from East Africa ~2500 years ago (Pleurdeau *et al.* 2012; Pickrell *et al.* 2014). For example, the presence of lactase persistence alleles in southern Africa indicates contact between East African herders and populations in south-central Africa, with subsequent migration into Namibia (Breton *et al.* 2014). This scenario is also supported by Y-chromosomal analysis that indicates a direct interaction between eastern African populations and southern African populations ~2000 years ago (Henn *et al.* 2008). However, in both cases (*i.e.*, MCM6/LCT and Y-chromosome M293), the frequency of the eastern African alleles is low in southern Africa and occurs in both pastoralist and hunter-gatherer populations. A simple model of eastern African demic diffusion into south-central Africa, leading to the adoption of pastoralism and a Khoekhoe population expansion from this area cannot be inferred from the genetic data.

Our samples from the Khoekhoe-speaking Nama pastoralists demonstrate that their primary ancestry is shared with other far southern nonpastoralist KhoeSan, such as the ≠Khomani San and the Karretjie (see also Schlebusch *et al.* 2011). mtDNA also suggests that the Nama display a haplogroup frequency distribution more similar to KhoeSan south of the Kalahari than to any other population in south-central Africa. Our results indicate that the majority of the Nama ancestry has likely been present in far southern

Africa for longer than previously assumed, rather than resulting from a recent migration from further north in Botswana where other Khoe speakers live. The only other Khoekhoe-speaking population in our data set is the Hai||om who share ~50% of the circum-Kalahari ancestry with the Nama and ≠Khomani, but are foragers rather than pastoralists. We conclude that Khoekhoe-speaking populations share a circum-Kalahari genetic ancestry with a variety of other Khoekhoe-speaking forager populations in addition to the !Xun, Karretjie, and ≠Khomani (Figure 1, Figure 2). This ancestry is divergent from central and northern Kalahari ancestries, arguing against a major demic expansion of Khoekhoe pastoralists from northern Botswana into South Africa. Rather, in this region, cultural transfer likely played a more important role in the diffusion of pastoralism. Of course, a demic expansion of the Khoekhoe within a more limited region of Namibia and South Africa may still have occurred—but geneticists currently lack representative DNA samples from many of the now “Coloured” interior populations, which may carry Khoekhoe ancestry.

This is an unusual case of cultural transmission (Jerardino *et al.* 2014). Other prehistoric economic transitions have been shown to be largely driven by demic diffusion (Gignoux *et al.* 2011; Fort 2012; Lazaridis *et al.* 2014; Skoglund *et al.* 2014; Malmström *et al.* 2015). Recent analysis of Europe provides a case study of demic diffusion, which appears far more complex than initially hypothesized. The initial spread of Near Eastern agriculturalists into southern Europe clearly replaced or integrated many of the autochthonous hunter-gatherer communities. Even isolated populations such as the Basque have been shown to derive much of their ancestry from Near Eastern agriculturalists (Skoglund *et al.* 2014). The early demic diffusion of agriculture exhibits a strong south-to-north cline across Europe, reflecting the integration of hunter-gatherers into composite southern agriculturalist populations, which then expanded northward with mixed ancestry (Sikora *et al.* 2014). The cline of the early Near Eastern Neolithic ancestry becomes progressively diluted in far northern European populations. In contrast, we see little evidence of a clear eastern African ancestry cline within southern African KhoeSan; nor is the putative “Khoe” ancestry identified in the Nama of eastern African origin or even of clear origin from northeastern Botswana where initial pastoralist contact presumably occurred.

However, the transfer of pastoralism from eastern to southern Africa itself was not purely cultural (see above). We also report here the presence of mitochondrial L4b2 that supports limited gene flow from eastern Africa, approximately during the same time frame as the pastoralist diffusion. L4b2, formerly known as L3g or L4g, is a mtDNA haplogroup historically found at a high frequency in eastern Africa, in addition to the Arabian Peninsula. L4b2 is at high frequency specifically in click-speaking populations such as the Hadza and Sandawe in Tanzania (sometimes described as “Khoisan speaking”) (Knight *et al.* 2003). Nearly 60% of



**Figure 5** L4b2 mtDNA haplogroup network. New L4b2 mitochondrial genomes from ≠Khomani and Nama individuals, indicated in pink as Southern Africa, were analyzed together with publically available L4b2 mtDNA genomes from NCBI (as outlined in File S1). All individuals were assigned to mtDNA haplogroups using haplogrep and the haplotypes were plotted using Network Publisher.

the Hadza population and 48% of Sandawe belong to L4b2 (Tishkoff *et al.* 2007). Even though both Tanzanian click-speaking groups and the southern African KhoeSan share some linguistic similarities and a hunter-gatherer lifestyle, they have been isolated from each other over the past 35,000 years (Tishkoff *et al.* 2007). The L4b2a2 haplogroup is present at a low frequency in both the Nama and ≠Khomani San, observed in one matriline in each population (Table 1). L4b2 was also formerly reported in the SAC population (0.89%) (Quintana-Murci *et al.* 2010) but has not been discussed in the literature. We identified several additional southern L4b2 haplotypes from whole mtDNA genomes deposited in public databases (Behar *et al.* 2008; Barbieri *et al.* 2013) and analyzed these samples together with all L4b2 individuals available in National Center for Biotechnology Information (NCBI). Median-joining phylogenetic network analysis of the mtDNA haplogroup, L4b2, supports the hypothesis that there was gene flow from eastern Africans to southern African KhoeSan groups. As shown in Figure 5 (and in more detail in Figure S6), southern African individuals branch off in a single lineage from eastern African populations in this network (Salas *et al.* 2002; Tishkoff *et al.* 2007; Gonder *et al.* 2007). The mitochondrial network suggests a recent migratory scenario (estimated to be <5000 years before present), although the source of this gene flow, whether from eastern African click-speaking groups or others, remains unclear (Pickrell *et al.* 2014).

## Conclusions

Analysis of 22 southern African populations reveals that fine-scale population structure corresponds better with ecological rather than linguistic or subsistence categories. The Nama pastoralists are autochthonous to far southwestern Africa, rather than representing a recent population movement from further north. We find that the KhoeSan

ancestry remains highly structured across southern Africa and suggests that cultural diffusion likely played the key role in adoption of pastoralism.

## Acknowledgments

We thank Jeffrey Kidd for assisting with genotyping of samples, David Poznik for providing off-target mtDNA reads from a separate next-generation sequencing experiment, Aaron Behr and Sohini Ramachandran for prepublication use of pong, and Meng Lin for help with analyses. We thank Carlos Bustamante for his encouragement and support of this project and Marcus Feldman for a close reading of our manuscript. We thank Julie Granka, Justin Myrick, and Cedric Werely for assistance with the saliva sample collection and Ben Viljoen for DNA extractions. Guidance from Ryan Raaum with regards to formulating the surface plots is appreciated. We also thank the Working Group of Indigenous Minorities in Southern Africa and the South African San Institute for their encouragement and advice. Finally, we thank Richard Jacobs, Wilhelmina Mondzinger, Hans Padmaker, Willem de Klerk, Hendrik Kaiman, and the communities in which we have sampled; without their support, this study would not have been possible. Funding was provided by a Stanford University Center on the Demographics and Economics of Health and Aging CDEHA seed grant to B.M.H. (National Institutes of Health, National Institute of Aging, NIA P30 AG017253-12) as well as a Stanford University Computation, Evolutionary, and Human Genomics trainee research grant to A.R.M. C.U. was funded by the National Research Foundation of South Africa. C.R.G. was funded by Predoctoral Training Grant 32.

Author contributions: C.U., M.K. A.R.M., and D.B. performed analysis. C.R.G., M.M., A.R.M., C.U., and B.M.H. collected DNA samples. P.D.v.H., M.M., E.G.H., and B.M.H. conceived of the study. C.U., C.R.G., M.M., E.G.H., and

B.M.H. wrote the manuscript in collaboration with all coauthors. All authors read and approved of the manuscript.

## Literature Cited

- Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19: 1655–1664.
- Barbieri, C., M. Vicente, J. Rocha, S. W. Mpoloka, M. Stoneking *et al.*, 2013 Ancient substructure in early mtDNA lineages of Southern Africa. *Am. J. Hum. Genet.* 92: 285–292.
- Barbieri, C., T. Güldemann, C. Naumann, L. Gerlach, F. Berthold *et al.*, 2014 Unraveling the complex maternal history of Southern African Khoisan populations. *Am. J. Phys. Anthropol.* 153: 435–448.
- Barnard, A., 1992 *Hunters and Herders of Southern Africa: A Comparative Ethnography of the Khoisan Peoples*. Cambridge University Press, Cambridge, UK.
- Behar, D. M., R. Villemes, H. Soodyall, J. Blue-Smith, L. Pereira *et al.* Genographic Consortium, 2008 The dawn of human matrilineal diversity. *Am. J. Hum. Genet.* 82: 1130–1140.
- Behr, A. A., K. Z. Liu, G. Liu-Fang, P. Nakka, and S. Ramachandran, 2016 pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*: btw327.
- Bleek, D. F., 1928 *The Naron: A Bushman Tribe of the Central Kalahari*, Cambridge University Press Archive, Cambridge, UK.
- Blench R., and K. C. MacDonald, 2000 *The Origins and Development of African Livestock: Archaeology, Genetics, Linguistics, and Ethnography*. UCL Press. London.
- Boonzaier, E., 1996 *The Cape Herders: A History of the Khoikhoi of Southern Africa*, New Africa Books, Kaapstad, South Africa.
- Breton, G., C. M. Schlebusch, M. Lombard, P. Sjödin, H. Soodyall *et al.*, 2014 Lactase persistence alleles reveal partial East African ancestry of southern African Khoe pastoralists. *Curr. Biol.* CB 24: 852–858.
- Burrough, S. L., 2016 Late quaternary environmental change and human occupation of the Southern African interior, pp. 161–174 in *Africa from MIS 6–2, Vertebrate Paleobiology and Paleoanthropology*, edited by B. A. Stewart and S. C. Jones. Springer-Verlag, Berlin.
- Creanza, N., M. Ruhlen, T. J. Pemberton, N. A. Rosenberg, M. W. Feldman *et al.*, 2015 A comparison of worldwide phonemic and genetic variation in human populations. *Proc. Natl. Acad. Sci.* USA 112: 1265–1272.
- Danecek, P., A. Auton, G. Abecasis, C. A. Albers, and E. Banks; 1000 Genomes Project Analysis Group, 2011 The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158.
- Delfin, F., S. Myles, Y. Choi, D. Hughes, R. Illek *et al.*, 2012 Bridging near and remote Oceania: mtDNA and NRY variation in the Solomon Islands. *Mol. Biol. Evol.* 29: 545–564.
- Dornan, S. S., 1925 *Pygmies and Bushmen of the Kalahari: An Account of the Hunting Tribes Inhabiting the Great Arid Plateau of the Kalahari Desert*, Seeley, Service & Company, London.
- Dunne, J., R. P. Evershed, M. Salque, L. Cramp, S. Bruni *et al.*, 2012 First dairying in green Saharan Africa in the fifth millennium BC. *Nature* 486: 390–394.
- du Plessis, I. D. D., 1947 *The Cape Malays*. South African Institute of Race Relations, Johannesburg, South Africa.
- Fort, J., 2012 Synthesis between demic and cultural diffusion in the Neolithic transition in Europe. *Proc. Natl. Acad. Sci. USA* 109: 18669–18673.
- Gignoux, C. R., B. M. Henn, and J. L. Mountain, 2011 Rapid, global demographic expansions after the origins of agriculture. *Proc. Natl. Acad. Sci. USA* 108: 6044–6049.
- Gonder, M. K., H. M. Mortensen, F. A. Reed, A. de Sousa, and S. A. Tishkoff, 2007 Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol. Biol. Evol.* 24: 757–768.
- Gravel, S., 2012 Population genetics models of local ancestry. *Genetics* 191: 607–619.
- Henn, B. M., C. Gignoux, A. A. Lin, P. J. Oefner, P. Shen *et al.*, 2008 Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc. Natl. Acad. Sci. USA* 105: 10693–10698.
- Henn, B. M., C. R. Gignoux, M. Jobin, J. M. Granka, J. M. Macpherson *et al.*, 2011 Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc. Natl. Acad. Sci. USA* 108: 5154–5162.
- Henn, B. M., L. L. Cavalli-Sforza, and M. W. Feldman, 2012 The great human expansion. *Proc. Natl. Acad. Sci. USA* 109: 17758–17764.
- International HapMap 3 Consortium; D. M., Altshuler, R. A., Gibbs, L., Peltonen, D. M., Altshuler, *et al.*, 2010 Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52–58.
- Jaccard, P., 1908 Nouvelles Recherches Sur La Distribution Florale. *Bull. Soc. Vaud. Sci. Nat.* 44: 223–270.
- Jerardino, A., J. Fort, N. Isern, and B. Rondelli, 2014 Cultural diffusion was the main driving mechanism of the Neolithic transition in southern Africa. *PLoS One* 9: e113672.
- Kayser, M., 2010 The human genetic history of Oceania: near and remote views of dispersal. *Curr. Biol.* 20: R194–R201.
- Knight, A., P. A. Underhill, H. M. Mortensen, L. A. Zhivotovsky, A. A. Lin *et al.*, 2003 African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Curr. Biol.* 13: 464–473.
- Lazaridis, I., N. Patterson, A. Mittnik, G. Renaud, S. Mallick *et al.*, 2014 Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513: 409–413.
- Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319: 1100–1104.
- MacDonald, K. C. R. H. M., 2000 The origins and development of African livestock: archaeology, genetics, linguistics and ethnography. *Orig. Dev. Domest. Anim. Arid West Afr.*: 127–162.
- Macholdt, E., V. Lede, C. Barbieri, S. W. Mpoloka, H. Chen *et al.*, 2014 Tracing pastoralist migrations to southern Africa with lactase persistence alleles. *Curr. Biol.* CB 24: 875–879.
- Malmström, H., A. Linderholm, P. Skoglund, J. Storå, P. Sjödin *et al.*, 2015 Ancient mitochondrial DNA from the northern fringe of the Neolithic farming expansion in Europe sheds light on the dispersion process. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370: 20130373.
- Maples, B. K., S. Gravel, E. E. Kenny, and C. D. Bustamante, 2013 RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93: 278–288.
- Novembre J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko *et al.*, 2008 Genes mirror geography within Europe. *Nature* 456: 98–101.
- Nurse, G. T., and T. Jenkins, 1977 Health and the hunter-gatherer. Biomedical studies on the hunting and gathering populations of Southern Africa. *Monogr. Hum. Genet.* 8: 1–126.
- Petersen, D. C., O. Libiger, E. A. Tindall, R.-A. Hardie, and L. I. Hannick *et al.* Indian Genome Variation Consortium, 2013 Complex patterns of genomic admixture within southern Africa. *PLoS Genet.* 9: e1003309.
- Petkova, D., J. Novembre, and M. Stephens, 2016 Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* 48: 94–100.
- Pickrell, J. K., N. Patterson, C. Barbieri, F. Berthold, L. Gerlach *et al.*, 2012 The genetic prehistory of southern Africa. *Nat. Commun.* 3: 1143.

- Pickrell, J. K., N. Patterson, P.-R. Loh, M. Lipson, B. Berger *et al.*, 2014 Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad. Sci. USA* 111: 2632–2637.
- Pleurdeau, D., E. Imalwa, F. Déroit, J. Lesur, A. Veldman *et al.*, 2012 “Of sheep and men”: earliest direct evidence of caprine domestication in Southern Africa at Leopard Cave (Erongo, Namibia). *PLoS One* 7: e40340.
- Poetsch, M., A. Wiegand, M. Harder, R. Blöhm, N. Rakotomavo *et al.*, 2013 Determination of population origin: a comparison of autosomal SNPs, Y-chromosomal and mtDNA haplogroups using a Malagasy population as example. *Eur. J. Hum. Genet.* 21: 1423–1428.
- Quintana-Murci, L., C. Harmant, H. Quach, O. Balanovsky, V. Zaporozhchenko *et al.*, 2010 Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture. *Am. J. Hum. Genet.* 86: 611–620.
- Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman *et al.*, 2005 Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. USA* 102: 15942–15947.
- Robbins, L. H., A. C. Campbell, M. L. Murphy, G. A. Brook, P. Srivastava *et al.*, 2005 The advent of herding in Southern Africa: early AMS dates on domestic livestock from the Kalahari Desert. *Curr. Anthropol.* 46: 671–677.
- Robbins, L. H., G. A. Brook, M. L. Murphy, A. H. Ivester, and A. C. Campbell, 2016 The Kalahari during MIS 6–2 (190–12 ka): archaeology, paleoenvironment, and population dynamics, pp. 175–193 in *Africa from MIS 6–2, Vertebrate Paleobiology and Paleoanthropology*, edited by B. A. Stewart and S. C. Jones. Springer-Verlag, Berlin.
- Sadr, K., 2008 Invisible herders? The archaeology of Khoekhoe pastoralists. *South. Afr. Humanit.* 20: 179–203.
- Sadr, K., 2015 Livestock first reached southern Africa in two separate events. *PLoS One* 10: e0134215.
- Salas, A., M. Richards, T. De la Fe, M.-V. Lareu, B. Sobrino *et al.*, 2002 The making of the African mtDNA landscape. *Am. J. Hum. Genet.* 71: 1082–1111.
- Schapera, I., 1934 *The Khoisan Peoples of South Africa*. Routledge & Kegan Paul, London.
- Schlebusch, C., 2010 Issues raised by use of ethnic-group names in genome study. *Nature* 464: 487, author reply 487.
- Schlebusch, C. M., and H. Soodyall, 2012 Extensive population structure in San, Khoe, and mixed ancestry populations from southern Africa revealed by 44 short 5-SNP haplotypes. *Hum. Biol.* 84: 695–724.
- Schlebusch, C. M., M. de Jongh, and H. Soodyall, 2011 Different contributions of ancient mitochondrial and Y-chromosomal lineages in “Karretjie people” of the Great Karoo in South Africa. *J. Hum. Genet.* 56: 623–630.
- Schlebusch, C. M., P. Skoglund, P. Sjödin, L. M. Gattepaille, D. Hernandez *et al.*, 2012 Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* 338: 374–379.
- Schlebusch, C. M., M. Lombard, and H. Soodyall, 2013 MtDNA control region variation affirms diversity and deep substructure in populations from southern Africa. *BMC Evol. Biol.* 13: 56.
- Sikora, M., M. L. Carpenter, A. Moreno-Estrada, B. M. Henn, P. A. Underhill *et al.*, 2014 Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe. *PLoS Genet.* 10: e1004353.
- Skoglund, P., H. Malmström, A. Omrak, M. Raghavan, C. Valdiosera *et al.*, 2014 Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* 344: 747–750.
- Smith, A., 2014 *The Origins of Herding in Southern Africa: Debating the “Neolithic” model*. Lap Lambert Academic Publishing, Saarbrücken, Germany.
- Theal, G. M., 1887 History of the Boers in South Africa, or the wanderings and wars of the emigrant farmers [microform]: from their leaving the Cape colony to the acknowledgement of their independence by Great Britain. S. Sonnenschein, Lowrey, London.
- Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt *et al.*, 2007 Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 39: 31–40.
- Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro *et al.*, 2009 The genetic structure and history of Africans and African Americans. *Science* 324: 1035–1044.
- Veeramah, K. R., D. Wegmann, A. Woerner, F. L. Mendez, J. C. Watkins *et al.*, 2012 An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Mol. Biol. Evol.* 29: 617–630.
- de Wit, E., W. Delpont, C. E. Rugamika, A. Meintjes, M. Möller *et al.*, 2010 Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Hum. Genet.* 128: 145–153.

Communicating editor: L. B. Jorde



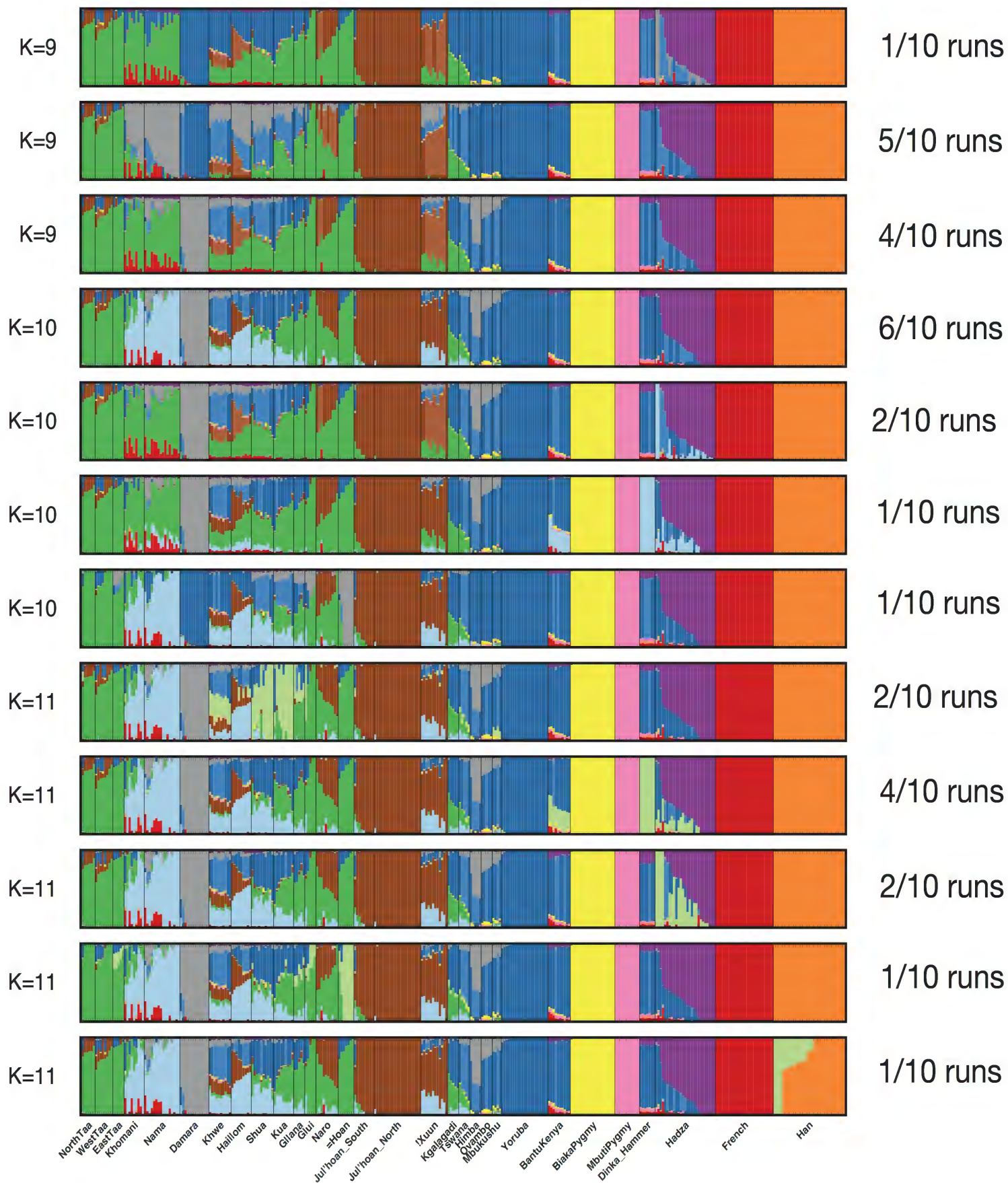
# GENETICS

**Supporting Information**

[http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187369 /-/DC1](http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.187369/-/DC1)

## **Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries**

**Caitlin Uren, Minju Kim, Alicia R. Martin, Dean Bobo, Christopher R. Gignoux, Paul D. van Helden,  
Marlo Möller, Eileen G. Hoal, and Brenna M. Henn**




10/10 runs

10/10 runs


5/10 runs

5/10 runs

7/10 runs



9/10 runs

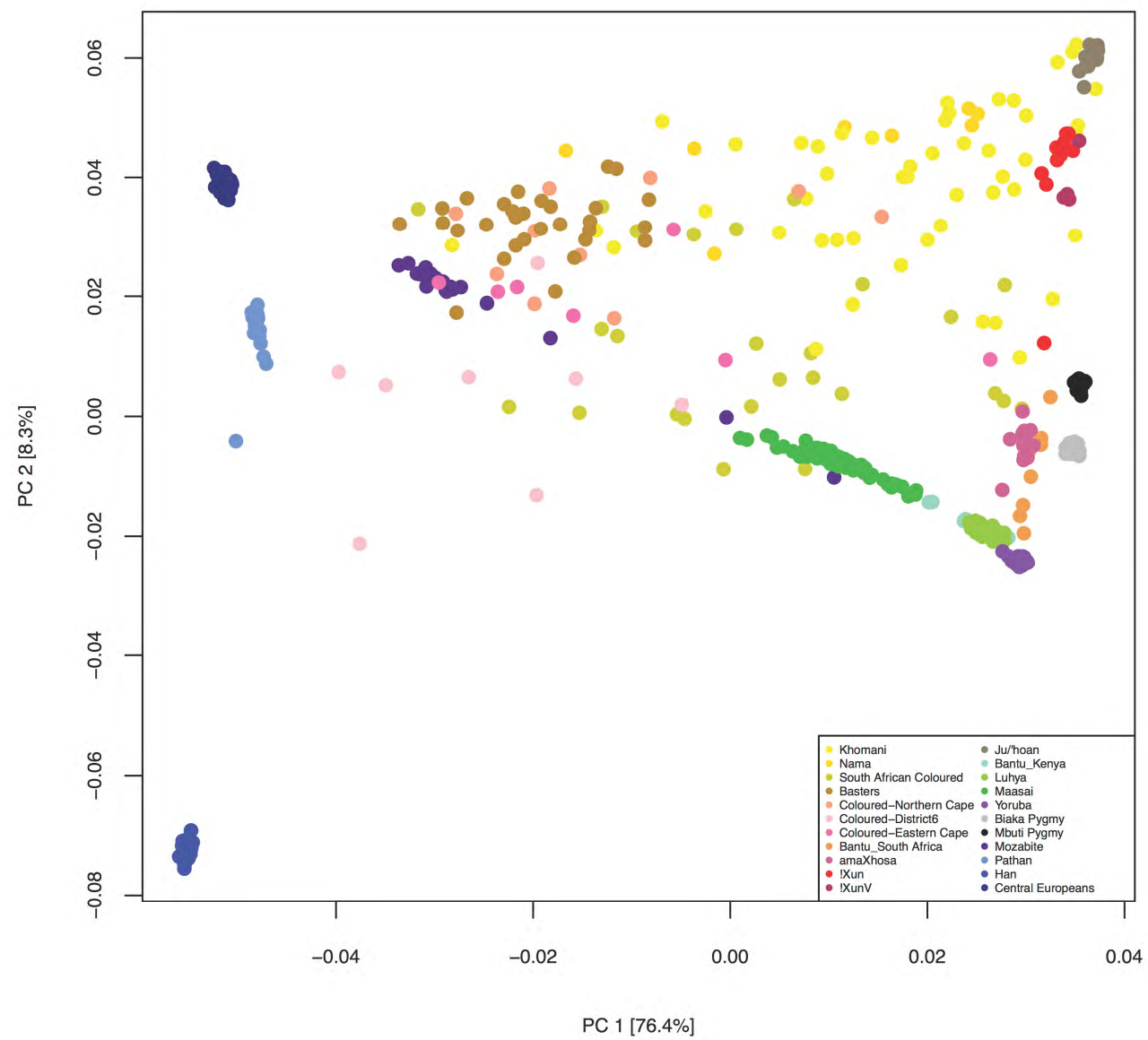


5/10 runs

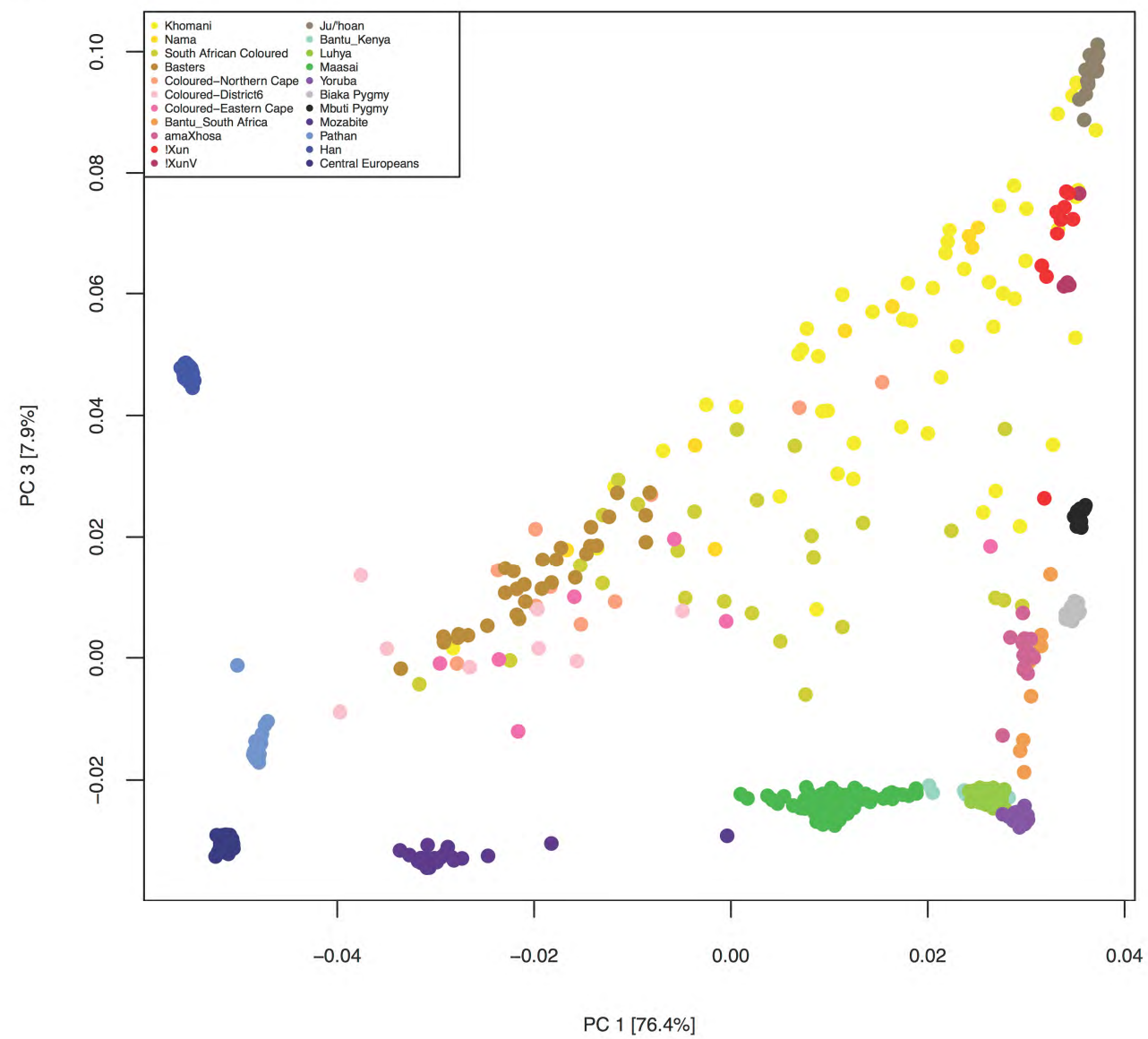
Juhoan Xun Nanya  
Xun Nanya  
Khomani  
Basters  
ColouredFEC  
ColouredNC  
SAC  
amaXhosa  
ColouredD6  
BantuKenya  
BantuUSA  
LWK  
YRI  
MKK  
Mozabite  
BiakaPygmy  
MbutiPygmy  
Pathan  
CEU  
Han



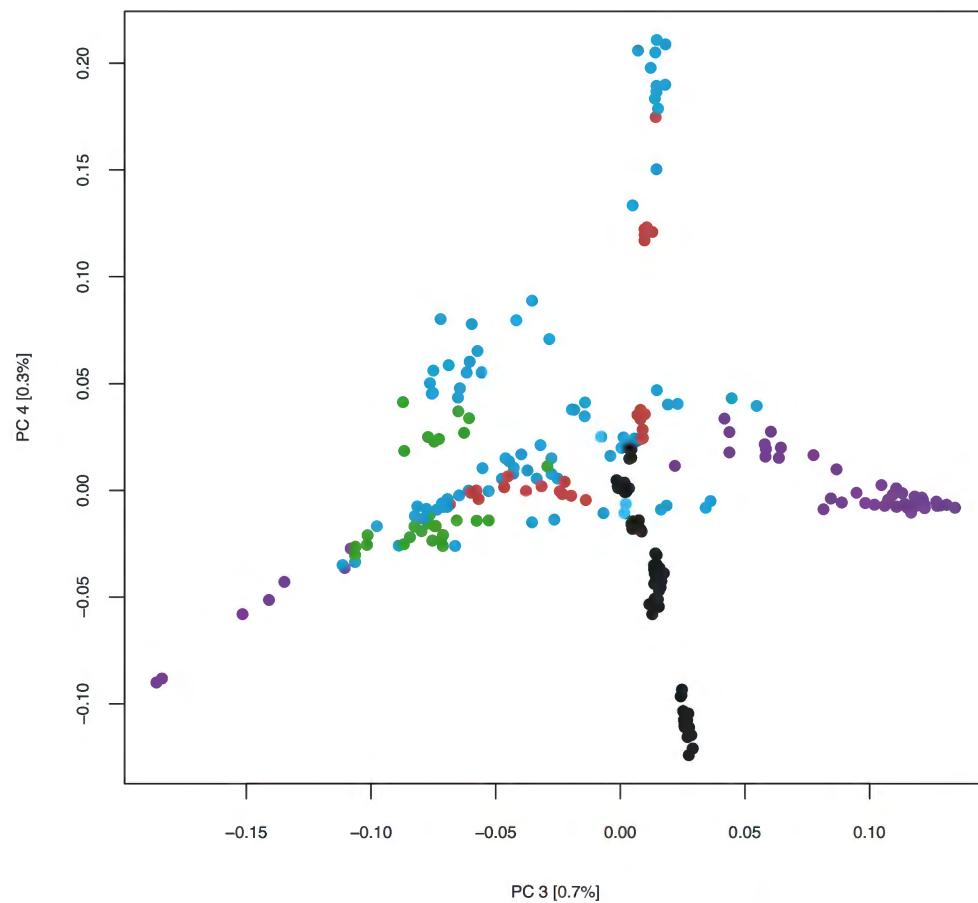
A

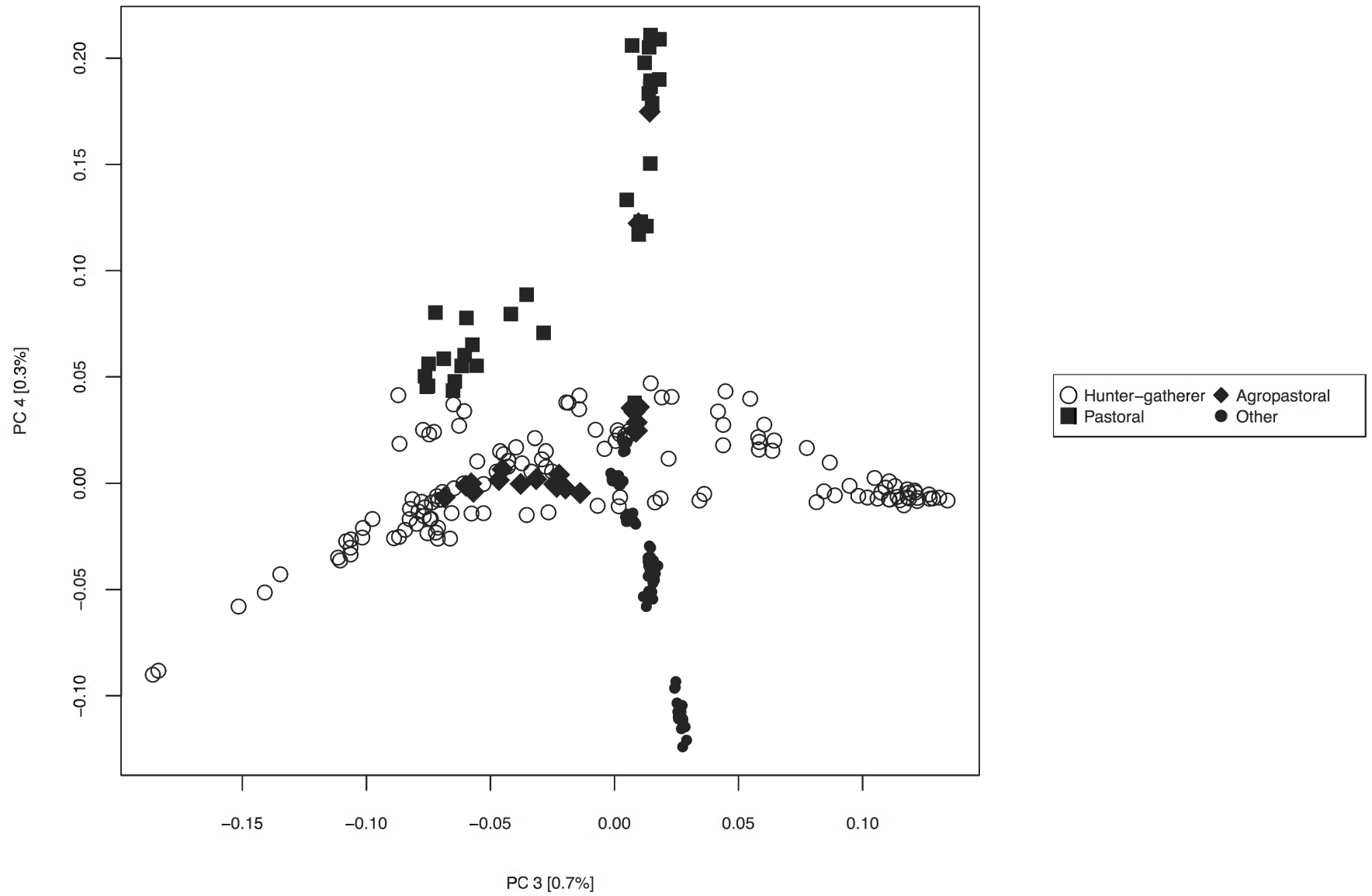


B



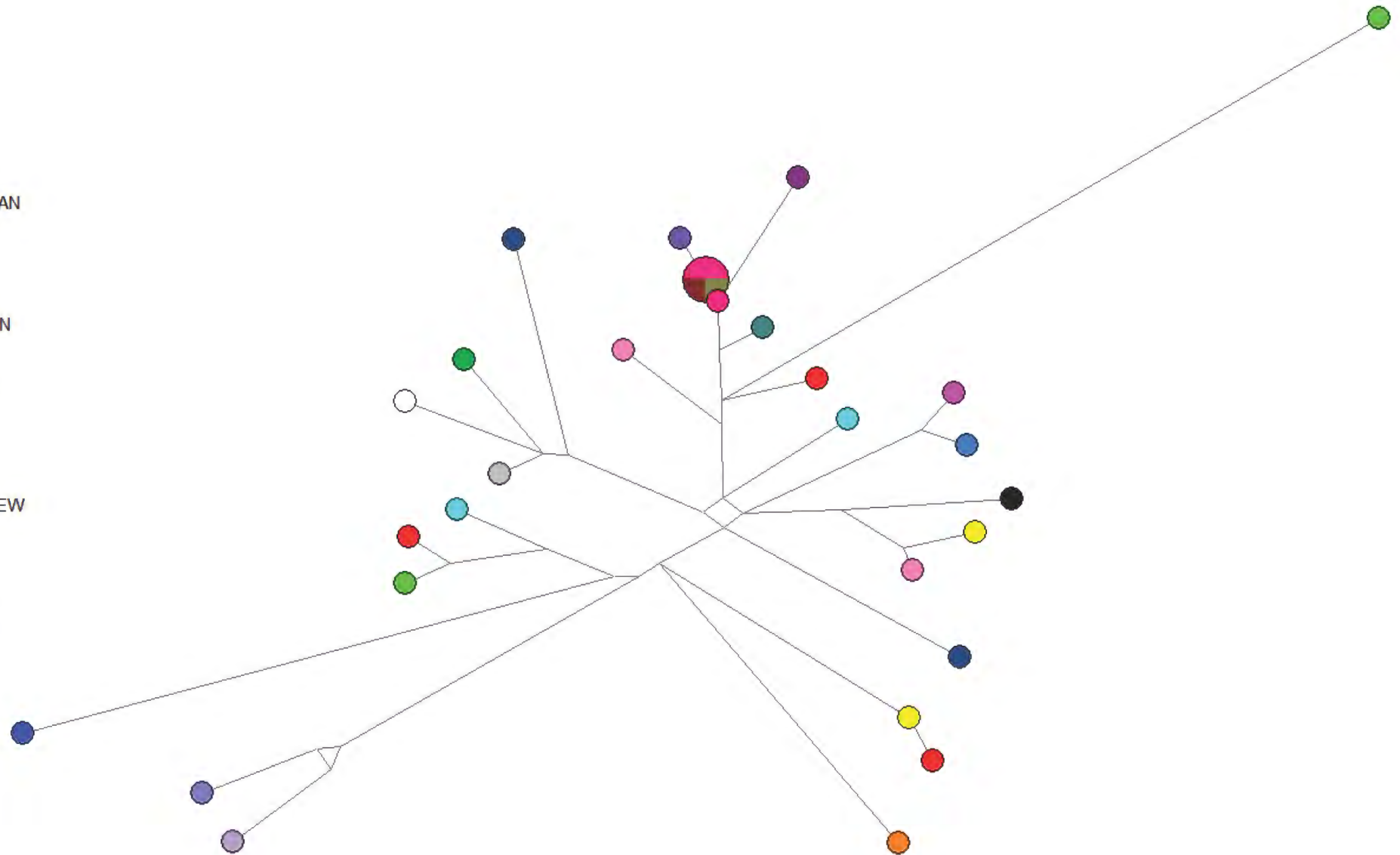






# Ethnicity

- KENYAN
- SOMALIAN
- BEMBA
- YEMENITE
- TANZANIAN
- SAUDI ARABIAN
- #KHOMANI
- SUDANESE
- ETHIOPIAN
- ARAB EMIRIAN
- JU'I'HOANSI
- KHOESAN
- SHUA
- NARO
- HAIJOM
- ETHIOPIAN JEW
- SYRIAN
- NUBIAN
- KUWAITI
- NIGERIAN
- CHADIAN
- BUDU



<b>Population Sample</b>	<b>Location of Sample</b>	<b>Latitude</b>	<b>Longitude</b>	<b>Language Family</b>	<b>Historical Subsistence</b>
!Xun	Namibia and Angola	<b>-18.7</b>	<b>19.7</b>	Kx'a	Hunter-gatherers
//Gana	Botswana (Central Kalahari)	<b>-21.7</b>	<b>23.4</b>	Khoe	Hunter-gatherers
/Gui	Botswana	<b>-21.5</b>	<b>23.3</b>	Khoe	Hunter-gatherers
#Hoan	Botswana	<b>-24.0</b>	<b>23.4</b>	Kx'a	Hunter-gatherers
#Khomani	South Africa (southern Kalahari)	<b>-27.8</b>	<b>21.1</b>	Tuu (!Ui-Taa)	Hunter-gatherers
amaXhosa	South Africa (Eastern Cape)	-31.5	28.3	Niger-Congo	Agropastoral
Bantu_Kenya	Kenyan Bantu-speakers	<b>-3.0</b>	<b>37.0</b>	Niger-Congo	Agropastoral
Bantu_SA	South African Bantu-speakers	-28.0	31.0	Niger-Congo	Agropastoral
Basque	France	43.0	0.0	Language isolate	Wage-based economy
Basters	South Africa (Northern Cape)	-23.3	17.1	Indo-European	Agropastoral
Biaka Pygmy	Southwestern Central African Republic	<b>4.0</b>	<b>17.0</b>	Niger-Congo	Hunter-gatherers
CEU	Europeans from Utah, USA	39.3	-111.1	Indo-European	Wage-based economy
ColouredD6	South Africa (District 6, Western Cape)	-33.9	18.4	Indo-European	Wage-based economy
ColouredEC	South Africa (Eastern Cape)	-34.0	25.6	Indo-European	Wage-based economy
ColouredNC	South Africa (Northern Cape)	-29.4	18.2	Indo-European	Wage-based economy
Damara	Northwest Namibia	<b>-19.8</b>	<b>16.2</b>	Khoe	Pastoral
Dinka	Southern Sudan	<b>8.8</b>	<b>27.4</b>	Nilo-Saharan	Agropastoral
EastTaa	Namibia, Botswana and South Africa	<b>-24.2</b>	<b>22.8</b>	Tuu (!Ui-Taa)	Hunter-gatherers
French	France	<b>46.0</b>	<b>2.0</b>	Indo-European	Wage-based economy
Hadza	North-Central Tanzania	<b>-3.6</b>	<b>35.1</b>	Language isolate	Hunter-gatherers
Hail om	Namibia (Etosha)	<b>-19.4</b>	<b>17.0</b>	Khoe	Hunter-gatherers
Han	China	<b>32.3</b>	<b>114.0</b>	Sino-Tibetan	Wage-based economy
Herero	Namibia, Botswana and Angola	-22	19.0	Niger-Congo	Pastoral
Himba	Northern Namibia (Kunene)	<b>-19.1</b>	<b>14.1</b>	Niger-Congo	Pastoral
Ju/'hoansi_North	Namibia, Angola	<b>-18.9</b>	<b>21.5</b>	Kx'a	Hunter-gatherers
Ju/'hoansi_South	Namibia, Botswana and Angola	<b>-21.2</b>	<b>20.7</b>	Kx'a	Hunter-gatherers
Kgalagadi	Botswana	<b>-24.8</b>	<b>21.8</b>	Niger-Congo	Agropastoral
Khwe	Namibia, Botswana and Angola	<b>-18.4</b>	<b>21.5</b>	Khoe	Hunter-gatherers
Kua	Botswana	<b>-21</b>	<b>25.9</b>	Khoe	Hunter-gatherers

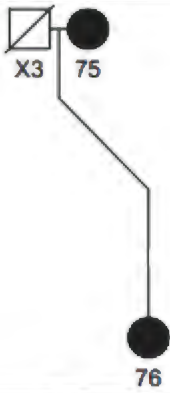




Luhya	Kenya	0.7	34.7	Niger-Congo	Agropastoral
Maasai	Southern Kenya and northern Tanzania	-1.8	36.6	Nilo-Saharan	Pastoral
Mandenka	Gambia	12.0	-12.0	Niger-Congo	Agropastoral
Mbukushu	Zambia	<b>-15.7</b>	<b>22.6</b>	Niger-Congo	Agropastoral
Mbuti Pygmy	Central Congo	<b>1.0</b>	<b>29.0</b>	Nilo-Saharan	Hunter-gatherers
Mozabite	Northern Algeria	32	3	Afro-Asiatic	Wage-based economy
Nama_AffyOrigins	Namibia	<b>-24.3</b>	<b>17.3</b>	Khoe	Pastoral
Nama_Illumina	South Africa	-28.5	17.0	Khoe	Pastoral
Naro	Namibia and Botswana (Ghanzi District)	<b>-22.0</b>	<b>21.6</b>	Khoe	Hunter-gatherers
NorthTaa	Namibia, Botswana and South Africa	<b>-23.0</b>	<b>22.3</b>	Tuu (!Ui-Taa)	Hunter-gatherers
Oroqen	China	50.4	126.5	Northern Tungusic	Wage-based economy
Ovambo	Namibia and Angola	<b>-19.0</b>	<b>18.1</b>	Niger-Congo	Agropastoral
Pathan	Pakistan	<b>33.5</b>	<b>70.5</b>	Indo-European	Wage-based economy
SAC	South Africa Coloured (Western Cape)	-33.9	18.4	Indo-European	Wage-based economy
Sandawe	Central Tanzania	-5.4	34.4	Language isolate	Hunter-gatherers
Shua	Botswana	<b>-20.6</b>	<b>25.3</b>	Khoe	Hunter-gatherers
Tswana	Botswana	<b>-24.1</b>	<b>25.4</b>	Niger-Congo	Agropastoral
WestTaa	Namibia, Botswana and South Africa	<b>-23.6</b>	<b>20.3</b>	Tuu (!Ui-Taa)	Hunter-gatherers
Yoruba	Southwestern Nigeria and southern Benin	<b>8.0</b>	<b>5.0</b>	Tonal Niger-Congo	Agropastoral

\*Black: unknown (35), Red: known (from the 91 NGS mtDNA) (38), Blue: inferred (41)

Family (total # of members)	Pedigree	Males	Females	# of matriline	Haplogroups
F1 (5)		x37 x35(L0d1c1)	45(L0d1c1) x36(L0d2a) 87(L0d2a)	2 matriline ① L0d1c1 (45 → x35(s)) ② L0d2a (x36→87(d))	(matriline) L0d1c1(1) L0d2a(1)  (individual) L0d1c1(2) L0d2a(2) Unknown(1)
total	5 ( 1 + 2 + 2 )	2 ( 1 + 1 )	3 ( 2 + 1 )	1	2 haplogroups

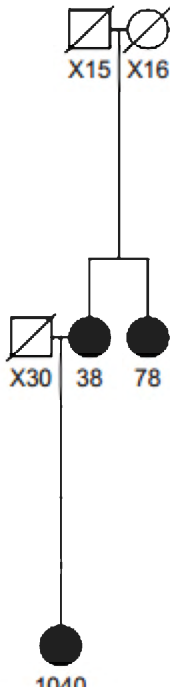
F2 (16)		x4 x51 x7 1032(L0d2a) x5 x11(L0d2a) 84(L0d2a)	47(L0d2a) 90(L0d2a) 1024(L0d2a) x6(L0d2c) x10(L0d2a) 1017(L0d2a) 1036(L0d2c) 1037(L0d2c) 1025(L0d2a)	<b>2 matriline</b> ① L0d2a (47→1024(d)→84(s)) (47→1032(s)) (47→1017(d)→1025(d)) ② L0d2c (x6→1036(d)&1037(d))	<b>(matriline)</b> L0d2a (1) L0d2c(1)  <b>(individual)</b> L0d2a(9) L0d2c(3) Unknown(4)
total	16 ( 4 + 4 + 8 )	7 ( 4 + 1 + 2 )	9 ( 3 + 6 )	2	<b>2 haplogroups</b>
F3 (3)		69(L0d1b1) 1002(L3e1a2)	70(L3e1a2)	<b>1 matriline</b> ① L3e1a2 (70→1002(s))	<b>(matriline)</b> L3e1a2(1)  <b>(individual)</b> L0d1b1(1) L3e1a2(2)
total	3 ( 2 + 1 )	2 ( 2 )	1(1)	1	<b>1 haplogroup</b>

F4 (3)		x3	75(L0d2a) 76(L0d2a)	1 matriline ① L0d2a (75→76(d))	(matriline) L0d2a(1)  (individual) L0d2a(2) Unknown(1)
total	3 ( 1 + 2 )	1 (1)	2 (2)	1	1 haplogroup
F5 (3)		x53	93 85	1 matriline ① *unknown (93→85(d)) (reason : hg of SA093 and SA085 are unknown)	(matriline) *unknown(1)  (individual) Unknown(3)
total	3 (3)	1 (1)	2 (2)	1 ( 1 )	?
F6 (3)		x23	1001(L0d2a) 79(L0d2a)	1 matriline ① L0d2a (1001→79(d))	(matriline) L0d2a(1)  (individual) L0d2a(2) Unknown(1)
total	3 ( 1 + 1 + 1 )	1 (1)	2 ( 1 + 1 )	1 ( 1 )	1 haplogroup

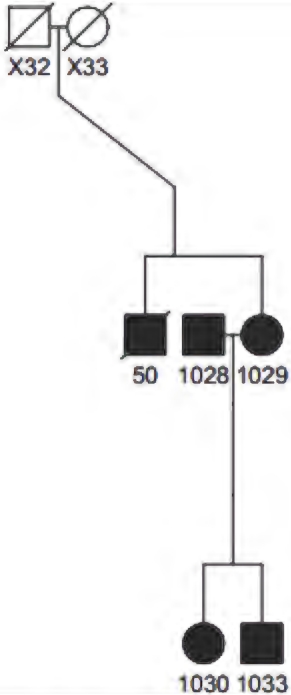
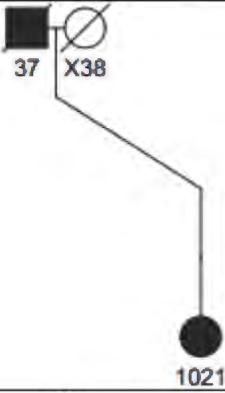


F7 (3)	<p>55 1022 95</p>	55(L0d2a)	1022(L0d1b1) 96(L0d1b1)	1 matriline ① L0d1b1 (1022→95(d))	(matriline) L0d1b1(1)  (individual) L0d1b1(2) L0d2a(1)
total	3 ( 2 + 1 )	1 ( 1 )	2 ( 1 + 1 )	1 ( 1 )	1 haplogroup
F8 (3)	<p>X52 1115 1117</p>	x52	1115 1117	1 matriline ① *unknown (1115→1117(d)) (reason : hg of SA1115 and SA1117 are unknown.)	(matriline) *unknown(1)  (individual) Unknown(3)
total	3 ( 3 )	1 ( 1 )	2 ( 2 )	1 ( 1 )	
F9 (7)	<p>X1 X2 67 68 80 1000 1003</p>	x1	x2(L0d2a) 67(L0d2a) 68(L0d2a) 80(L0d2a) 1000(L0d2a) 1003(L0d2a)	1 matriline ① L0d2a (x2→67(d),68(d),80(d),1000(d), and 1003(d))	(matriline) L0d2a(1)  (individual) L0d2a(6) Unknown(1)
total	7 ( 1 + 1 + 5 )	1 ( 1 )	6 ( 1 + 5 )	1 ( 1 )	1 haplogroup


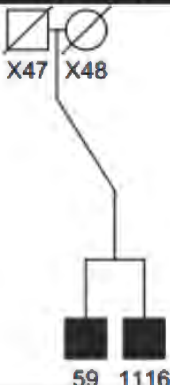

F10 (8)	<p>X8 X9 X34</p> <p>16 72 1009 1012 36</p>	x8 x34 1012(L0d2a)	x9(L0d2a) 16(L0d2a) 72 (L0d2a) 1009(L0d2a) 36(L0d2a)	<b>1 matriline</b> ① L0d2a (x9→16(d),72(d),1009(d),and 1012(s)) (x9→36(d))	(matriline) L0d2a(1)  (individual) L0d2a(6) Unknown(2)
total	8 ( 2 + 1 + 6 )	3 ( 2 + 1 )	5 ( 1 + 4 )	1 ( 1 )	<b>1 haplogroup</b>
F11 (5)	<p>X13 X12 X14</p> <p>52 54</p>	x12 52(L0d2c)	x13(L0d2c) x14(L0d2c) 54(L0d2c)	<b>2 matriline</b> ① L0d2c (x13→52) ② L0d2c (x14→54)	(matriline) L0d2c(2)  (individual) L0d2c(4) Unknown(1)
total	5 ( 1 + 2 + 2 )	2 ( 1 + 1 )	3 ( 1 + 2 )	2 ( 2 )	<b>2 haplogroups</b>

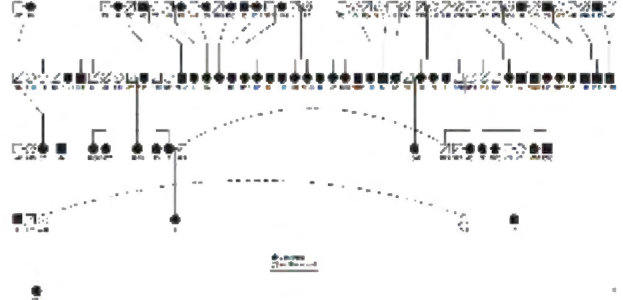
F12 (6)		x15 x30	x16(L0d1b1) 38(L0d1b1) 78(L0d1b1) 1040(L0d1b1)	<b>1 matriline</b> <b>①</b> L0d1b1 (x16→78(d)) (x16→38(d)→1040(d))	(matriline) L0d1b1(1)  (individual) L0d1b1(4) Unknown(2)
total	6 ( 2 + 3 + 1 )	2 (2)	4 ( 3 + 1 )	1 ( 1 )	<b>1 haplogroup</b>

<p>F13 (29)</p>		<p>x17 x27 x43(L0d2a) x20(L0d1b1) x29(L0d2a) x26 x22 x40 x45 91 x24</p>	<p>X18(L0d2a) x28(L0d1b1) 34(L0d2a) x44(L0d2a) x19(L0d2a) 17(L0d1b1) 43(L0d2a) 92(L0d2a) 1023(L0d2a) x21(L0d2a) 73(L0d2a) x41(L0d1a) 19(L0d1a) 9(L0d1a) 7(L0d1a) x46 x25(L0d2a) 1016(L0d2a)</p>	<p><b>5 matriline</b>  <b>①</b> L0d2a  (x18→34(d))  (x18→x19(d)→x21(d)→73(d))  <b>②</b> L0d2a  (x44→43(d)→x25(d)→1016(d))  <b>③</b> L0d1b1  (x28→17(d)&amp;x20(s))  <b>④</b> L0d1a  (x41→9(d)→7(d))  (x41→19(d))  <b>⑤</b> *unknown  (x46→91(s) &amp; x24(s))  (reason : hg of SA091 is unknown)</p>	<p>(matriline)  L0d2a(2)  L0d1b1(1)  L0d1a(1)  *unknown(1)    (individual)  L0d2a(13)  L0d1b1(3)  L0d1a(4)  Unknown(9)</p>
<p>total</p>	<p>29 ( 9 + 10 + 10 )</p>	<p>11 ( 8 + 3 )</p>	<p>18 ( 1 + 10 + 7 )</p>	<p>5 ( 4 + 1 )</p>	<p>4 haplogroups</p>

F14 (7)		x32 50(L0d1b) 1028(L0d1a) 1033(L0d1b)	x33(L0d1b) 1029(L0d1b) 1030(L0d1b)	<b>1 matriline</b> ① L0d1b (x33→1029(d)→1030(d)&1033(s)) (x33→50)	(matriline) L0d1b(1)  (individual) L0d1b(5) L0d1a(1) Unknown(1)
total	7 ( 1 + 4 + 2 )	4 ( 1 + 2 + 1 )	3 ( 2 + 1 )	1 ( 1 )	<b>1 haplogroup</b>
F15 (3)		37	x38(L0d1b) 1021(L0d1b)	<b>1 matriline</b> ① L0d1b (x38→1021(d))	(matriline) L0d1b(1)  (individual) L0d1b(2) Unknown(1)
total	3 ( 1 + 1 + 1 )	1 ( 1 )	2 ( 1 + 1 )	1 ( 1 )	<b>1 haplogroup</b>



F16 (3)		39(L0d2a)	x39(L0d2a) 1014(L0d2a)	1 matriline ① L0d2a (x39→1014(d))	(matriline) L0d2a(1)  (individual) L0d2a(3)
total	3 ( 2 + 1 )	1 ( 1 )	2 ( 1 + 1 )	1 ( 1 )	1 haplogroup
F17 (4)		x47 59 1116	x48	1 matriline ① *unknown (x48→59(s)&1116(s)) (reason : hg of SA059 and SA1116 are not known.)	(matriline) *unknown(1)  (individual) Unknown(4)
total	4 ( 4 )	3 (3)	1 (1)	1 ( 1 )	?
F18 (3)		1118(L0d2a) 1119	x49	1 matriline ① *unknown (x49→1119(s)) (reason : hg of SA1119 is unknown.)	(matriline) *unknown(1)  (individual) L0d2a(1) Unknown(2)
Total	3 ( 2 + 1 )	2 ( 1 + 1 )	1 (1)	1 ( 1 )	?

114 individuals from 18 families		46 (29 + 9 + 8)	68 (7 + 29 + 33)	25 independent matriline	7 haplogroups +unknown
--	---	-----------------	------------------	--------------------------	---------------------------

**A**

	Kung	Juhoan	Nama	Kua	Shua	Khwe	Gana	Gui	Naro	EastTaa	WestTaa	Khomani
Kung	0,000	0,013	0,021	0,026	0,032	0,032	0,020	0,016	0,011	0,020	0,011	0,018
Juhoan	0,013	0,000	0,038	0,044	0,054	0,055	0,036	0,025	0,012	0,028	0,021	0,034
Nama	0,021	0,038	0,000	0,019	0,020	0,021	0,016	0,019	0,024	0,025	0,024	0,003
Kua	0,026	0,044	0,019	0,000	0,014	0,016	0,010	0,019	0,031	0,029	0,029	0,018
Shua	0,032	0,054	0,020	0,014	0,000	0,013	0,017	0,028	0,039	0,037	0,039	0,021
Khwe	0,032	0,055	0,021	0,016	0,013	0,000	0,019	0,031	0,041	0,041	0,041	0,022
Gana	0,020	0,036	0,016	0,010	0,017	0,019	0,000	0,011	0,022	0,019	0,019	0,013
Gui	0,016	0,025	0,019	0,019	0,028	0,031	0,011	0,000	0,011	0,013	0,013	0,015
Naro	0,011	0,012	0,024	0,031	0,039	0,041	0,022	0,011	0,000	0,014	0,008	0,019
EastTaa	0,020	0,028	0,025	0,029	0,037	0,041	0,019	0,013	0,014	0,000	0,011	0,021
WestTaa	0,011	0,021	0,024	0,029	0,039	0,041	0,019	0,013	0,008	0,011	0,000	0,019
Khomani	0,018	0,034	0,003	0,018	0,021	0,022	0,013	0,015	0,019	0,021	0,019	0,000

**B**

	Kung	Juhoan	Nama	Kua	Shua	Khwe	Gana	Gui	Naro	EastTaa	WestTaa	Khomani
Kung	0	191	671	698	624	193	511	489	417	691	549	1023
Juhoan	191	0	742	516	441	56	369	345	345	605	538	992
Nama	671	742	0	957	920	788	689	690	509	558	315	544
Kua	698	516	957	0	77	544	271	275	459	478	645	900
Shua	624	441	920	77	0	468	232	231	414	476	614	908
Khwe	193	56	788	544	468	0	418	393	401	660	592	1047
Gana	511	369	689	271	232	418	0	25	189	285	382	718
Gui	489	345	690	275	231	393	25	0	184	305	387	736
Naro	417	345	509	459	414	401	189	184	0	274	223	648
EastTaa	691	605	558	478	476	660	285	305	274	0	263	435
WestTaa	549	538	315	645	614	592	382	387	223	263	0	474
Khomani	1023	992	544	900	908	1047	718	736	648	435	474	0

C

	Kung	Juhoan	Nama	Kua	Shua	Khwe	Gana	Gui	Naro	EastTaa	WestTaa	Khomani
Kung	0,000	0,407	0,673	0,588	0,650	0,564	0,518	0,513	0,532	0,558	0,676	0,602
Juhoan	0,407	0,000	0,633	0,598	0,661	0,598	0,540	0,535	0,514	0,568	0,612	0,623
Nama	0,673	0,633	0,000	0,548	0,567	0,528	0,471	0,486	0,470	0,515	0,733	0,590
Kua	0,588	0,598	0,548	0,000	0,225	0,505	0,388	0,365	0,383	0,511	0,651	0,586
Shua	0,650	0,661	0,567	0,225	0,000	0,500	0,392	0,367	0,408	0,560	0,683	0,632
Khwe	0,564	0,598	0,528	0,505	0,500	0,000	0,266	0,263	0,363	0,477	0,660	0,557
Gana	0,518	0,540	0,471	0,388	0,392	0,266	0,000	0,028	0,141	0,429	0,616	0,516
Gui	0,513	0,535	0,486	0,365	0,367	0,263	0,028	0,000	0,139	0,424	0,612	0,510
Naro	0,532	0,514	0,470	0,383	0,408	0,363	0,141	0,139	0,000	0,405	0,658	0,500
EastTaa	0,558	0,568	0,515	0,511	0,560	0,477	0,429	0,424	0,405	0,000	0,648	0,333
WestTaa	0,676	0,612	0,733	0,651	0,683	0,660	0,616	0,612	0,658	0,648	0,000	0,686
Khomani	0,602	0,623	0,590	0,586	0,632	0,557	0,516	0,510	0,500	0,333	0,686	0,000

## **Supplemental Methods**

### **Population structure:**

chromoPainter (Lawson *et al.* 2012) takes as input SNP data from a pre-defined recipient and donor populations as well as a genetic recombination map. The program ‘paints’ each recipient individual on the basis of every other individual in the dataset. fineSTRUCTURE (Lawson *et al.* 2012) places individuals into populations based on a model for “expected variability”. Software was freely available at [www.paintmychromosomes.com](http://www.paintmychromosomes.com).

Principle components analysis (PCA) was performed in R and the PC loadings were calculated from the ‘.chunkcounts.out’ file generated from chromoPainter. These were mean transformed and plotted in the R programming environment. Three different PCA’s were plotted. Figure 3 was colour and shape coded according to the majority ancestry in Figure 2. Populations in Figure S5 were plotted as different shapes according to their subsistence strategy. The language family of every population is used to colour population present in the PCA in Figure S6.

### **Local Ancestry Assignment and TRACTs:**

We merged all ≠Khomani individuals genotyped on the OmniExpress and OmniExpressPlus arrays, the Schuster *et al.*, (Schuster *et al.* 2010) Namibian genotypes, along with CEU and LWK individuals genotyped in 1000 Genomes. As reference panels, we defined separate classes for European, Bantu, and KhoeSan ancestries respectively using CEU, LWK, and ≠Khomani and Schuster *et al.*, (Schuster *et al.* 2010) individuals with >90% KhoeSan ancestry as inferred via ADMIXTURE. We phased individuals using SHAPEIT2 with the 1000 Genomes phase 3 as a reference panel. We inferred local ancestry using RFMix (Maples *et al.* 2013) with a node size of 5 to reduce bias resulting from unbalanced reference panels, a minimum window size of 0.2 cM, and 1 EM iteration to better inform the small amount of admixture in the KhoeSan reference samples. We assessed the fit of 7 different models in TRACTs (Gravel 2012), including several two-pulse and three-pulse models. Ordering the populations as KhoeSan, Bantu, and European, we tested the following models: ppp\_ppp, ppp\_pxp, ppp\_xxp, ppx\_xxp, ppx\_xxp\_ppx, ppx\_xxp\_pxx, and ppx\_xxp\_xxp, where the order of each letter corresponds with the order of population given above, an underscore indicates a distinct migration event with the first event corresponding with the most generations before present, p corresponding with a pulse of the ordered ancestries, and x corresponding with no input from the ordered ancestries. We tested all 7 models preliminarily 3 times, and for all models that converged and were within the top 3 models, we subsequently fit each model with 100 starting parameters randomizations. The log-likelihood of the best fit model was -342, which provided a substantially better fit than all other models tested (next best model achieved best log-likelihood = -402).

### **mtDNA haplogroup frequency and networks:**

#### **Haplogroup frequency:**

Coverage per individual was set at a minimum of 6.5x, therefore only 80 out of the 91 ≠Khomani and 36 Nama were used for further analysis (Table 1). To prevent oversampling of the same haplogroup in families, only one individual per matrilineage was included (Table S2). These individuals were then grouped with other publically available data. Haplotypes were assigned to haplogroups using *haplogrep* (Kloss-Brandstätter *et al.* 2011).

#### **mtDNA Network:**

We utilized Network (ver. 4.6, copy righted by Fluxus Technology Ltd.), for a median-joining phylogenetic network analysis in order to produce Figures 4 and S4. Network Publisher (ver. 2.0.0.1, copy righted by Fluxus Technology Ltd.) was then used to draw the phylogenetic relationships among individuals.



### Supplemental References:

Gravel S., 2012 Population Genetics Models of Local Ancestry. *Genetics* **191**: 607–619.

Kloss-Brandstätter A., Pacher D., Schönherr S., Weissensteiner H., Binna R., Specht G., Kronenberg F., 2011 HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.* **32**: 25–32.

Lawson D. J., Hellenthal G., Myers S., Falush D., 2012 Inference of population structure using dense haplotype data. *PLoS Genet.* **8**: e1002453.

Maples B. K., Gravel S., Kenny E. E., Bustamante C. D., 2013 RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**: 278–288.

Schuster S. C., Miller W., Ratan A., Tomsho L. P., Giardine B., Kasson L. R., Harris R. S., Petersen D. C., Zhao F., Qi J., Alkan C., Kidd J. M., Sun Y., Drautz D. I., Bouffard P., Muzny D. M., Reid J. G., Nazareth L. V., Wang Q., Burhans R., Riemer C., Wittekindt N. E., Moorjani P., Tindall E. A., Danko C. G., Teo W. S., Buboltz A. M., Zhang Z., Ma Q., Oosthuysen A., Steenkamp A. W., Oostuisen H., Venter P., Gajewski J., Zhang Y., Pugh B. F., Makova K. D., Nekrutenko A., Mardis E. R., Patterson N., Pringle T. H., Chiaromonte F., Mullikin J. C., Eichler E. E., Hardison R. C., Gibbs R. A., Harkins T. T., Hayes V. M., 2010 Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**: 943–947.

### Figure legends:

**Figure S1:** *Population structure in southern Africa and further evidence for a southern African specific KhoeSan ancestry.* These diagrams display the ancestral contributions as ascertain by an unsupervised ADMIXTURE analysis. Ancestral proportions are shown as varying degrees of each color i.e. each ancestry. This is displayed for a large number of KhoeSan populations in the Affymetrix Human Origins dataset. Every hypothesis of the number of ancestral populations is taken into account ( $k$  values). As seen here due to the hypothesis of structure, multiple  $k$  values were used. Every run utilized a different random seed and thus it was necessary to pool similar results as shown, by the use of *pong*.

**Figure S2:** *Population structure in southern Africa and further evidence for a southern African specific KhoeSan ancestry, utilizing more South African specific populations.* ADMIXTURE plots as generated from an unsupervised analysis of the 340k merged dataset. Each color represents a specific ancestry and every hypothesis of the number of ancestral populations are taken into account ( $k$  values). Multi-modularity was assessed using *pong* as in Figure S1, however only the major modes are displayed here. Each run utilized a different random seed and thus there were differing results. These results were grouped according to similarity using *pong*.

**Figure S3:** *Lack of clustering as well as structure related to the Nama and ≠Khomani.* A PCA of the merged 340k dataset depicts the clustering of unrelated individuals based on the variation seen in the dataset. PCA loadings were calculated from the *\*chunkcounts.out* file from chromopainter using the *precomp* function in R. PC 1 and 2 are depicted in A) and PC 1 and 3 are depicted in B).

**Figure S4:** *Color-coding of populations based on language family shows no association between language and genetic differences.* A PCA of the Affymetrix Human Origins dataset depicts the clustering of unrelated individuals based on the variation seen in the dataset. This PCA is identical to that in Figure 2 but is color-coded based on the language family of each population as tabulated in Table S1. Green are Tuu speaking populations. Blue are Khoe speaking populations. Purple are Kx'a speaking populations. Red are Niger-Congo speaking populations. Populations color-coded blacks were not included, as they did not form part of the analysis in Figure 2.

**Figure S5:** *Differentiation based on subsistence strategies shows some association between genetic distance and subsistence strategies.* A PCA of the Affymetrix Human Origins dataset depicts the clustering of unrelated individuals based on the variation seen in the dataset. This PCA is identical to that in Figure 3 but it is coded in different shapes based on the subsistence strategy of each population as tabulated in Table S1. Populations depicted by a grey circle were not included, as they did not form part of the analysis in Figure 2.

**Figure S6:** *L4b2 mtDNA haplogroup network- color coded per country.* ≠Khomani and Nama individuals were merged with publicly available data from NCBI (as outlined in the Supplementary Methods). All individuals were assigned mtDNA haplogroups using haplogrep and the haplotypes were plotted using Network Publisher.

**Table S1:** *The diversity associated with the geographical location of samples populations, their language family and subsistence strategy.* Populations in bold were used to plot Figure 2. Longitude and latitude values of sampled populations were taken from Lazaridis et al (2014).

**Table S2:** *Inferred Pedigree for ≠Khomani Samples*

**Table S3:** *Genetic (A), Geographic (B) and Phonemic (C) distance matrices per samples population*