JASs Journal of Anthropological Sciences
Open Access
Research articles Vol. 102 (2024), pp. 123 - 143
doi.10.4436/jass10204

Reconstructing micro-evolutionary dynamics shaping local variation in Southern African populations using genomics, metagenomics and personal metadata

Gonzalo Oteo-García[1,2,3], Giacomo Mutti[1], Matteo Caldon[1], Ockie Oosthuitzen[4], Matteo Manfredini[1] & Cristian Capelli[1]

1) Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Italy
e-mail: gonzalo.oteogarcia@uniroma1.it

2) Dipartimento di Biologia Ambientale, Sapienza Università di Roma, Rome, Italy

3) Centre for Palaeogenetics & Department of Archaeology and Classical Studies, Stockholm University, Stockholm, Sweden

4) School of Medicine, University of Namibia, Windhoek, Namibia

Summary - *Geography is a well-known factor shaping genetic variation in human populations. However, the potential role played by cultural variables remains much understudied. This study investigates the impact of socio-cultural variables on genomic similarity and the saliva microbiome, using data from populations in Lesotho and Namibia. Geographic distance within Lesotho increases genetic differentiation, while shared clan affiliation surprisingly increases it. In Namibia, ethnicity is the predominant factor influencing genetic affinity. Saliva metagenomic data shows a negative correlation between age and alpha diversity, with notable differences in host-interacting taxa and viral load. These findings highlight the role of geography in shaping genetic affinity even at small scales and the complex influences of cultural factors. The saliva microbiome appears primarily affected by unrecorded individual behaviors rather than geographic or cultural variables. At population-level these oral microbiomes reveal insights into some dietary habits, oral health, and also the communal viral load, which appears to have greater incidence in Lesotho possibly related to the long-term effects of the HIV epidemic in the country.*

Keywords - *Genomics, Metagenomics, Geography, Ethnicity, Clan, Southern Africa.*

Introduction

Patterns of human genetic and metagenomic variation are influenced by geography (Handley et al. 2007; Ruan et al. 2022). At global scale, physical distance tends to be the main contributing factor in shaping differentiation (Prugnole et al. 2005; Ramachandran et al. 2005; Li et al. 2008; Novembre et al. 2008; Tishkoff et al. 2009; de Filippo et al. 2012; González-Santos et al. 2022). Geographic patterns have also been observed at smaller scales (Leslie et al. 2015; Skoglund et al. 2016; Lipson et al. 2018; Bycroft et al. 2019; Raveane et al. 2019; Ioannidis et al. 2021; Willerslev and Meltzer 2021). At these small scales, there are instances of culturally defined dynamics that contribute to the distribution of genetic variation by regulating the pattern of gene flow across neighbouring communities (Oota et al. 2001; Wilder et al. 2004; Langergraber et al. 2007; Marks et al. 2012).

Human populations in Southern Africa are a paradigmatic crossroads of exceptional genetic diversity but what socio-cultural dynamics drive patterns of genetic affinity beyond geography has nor beet investigated thoroughly (1KGPC 2015; Choudhury et al. 2020; Oliveira et al. 2023). Modern African countries like Namibia, Botswana, South Africa and Lesotho are home to indigenous populations (Ellenberger 1912; Hann et al. 1966) that belong to diverse lifestyle

traditions and ethnolinguistic groups such as the Bantu-speakers (Van Warmelo 1962; Retshabile et al. 2018; Sengupta et al. 2021) and Khoisan-speakers (Barnard 1992; Schuster et al. 2010; Uren et al. 2016; Gronau et al. 2011; Veeramah et al. 2012). Although outside influences have impacted these societies, some populations in these regions maintain a degree of traditional ways of life (Barnard 1992; Tishkoff et al. 2007; Schlebusch 2010; Henn et al. 2011; Oliveira et al. 2023). In varying degrees, some were nomadic hunter-gatherers and pastoralists, others settled agriculturalists; each one with a potentially diverse set of oral and gut microbiomes (Leeming et al. 2019; Singh et al. 2017; De Angelis et al. 2020; Oduaran et al. 2020).

Lesotho comprises more than 2 million people, the Basotho, who speak Sesotho, a southern Bantu language (Huffman 2007; Hammond-Tooke 2004). This ethnic group has at least twice as many members in neighbouring South Africa and other countries. The roots of the Sotho people in Southern Africa trace back to Bantu-speaking migrants from the north (Van Warmelo 1962; Sengupta et al. 2021) who established themselves there in the first centuries of the Common Era as part of the "Bantu expansion" (Tishkoff et al. 2009, 2012; Pickrell et al. 2012; Veeramah et al. 2012; Marks et al. 2014; Marks et al. 2015; Gonzalez-Santos et al. 2015; Skoglund et al. 2017; Gonzalez-Santos et al. 2022). Historically the Basotho nation conformed itself as an amalgamation of Bantu-speaking clans of various Sotho origins and some other nations, Bantu-speaking or not (Eldredge 1993; Montinaro et al. 2016; Montinaro et al. 2017). What is more, the consolidation of Lesotho as a country in the 19th century was the result of the diplomatic and military efforts of king Moshoeshoe I that integrated diverse peoples into his own Sotho clans. Although patrilineal inheritance of clan affiliation remains as a system (Montinaro et al. 2016), the importance and power traditionally placed in the hands of chiefs of clans decayed with the birth of the country of Lesotho (Eldredge 1993).

Namibia represents the other side of the spectrum, a large and sparsely populated multi-ethnic country hosting both Bantu and Khoisan-speaking peoples (Petersen et al. 2013; Montinaro et al. 2017; Oliveira et al. 2023), where ethnic groups are largely structured by territory (Barnard 1992; Malan 1995).

The Bantu-speaking Ovambo are the largest ethnic group in Namibia and account for half of the population but are concentrated in a small region in the north-central lands. They consist of various culturally related peoples that also inhabit southern Angola (Davies 1994). Their traditional lifestyle is based on farming and raising cattle. The Himba and Herero are closely related (Oliveira et al. 2018) Bantu-speaking groups which represent 7% of Namibia's population. Their arrival to Namibian territory is attested in the 17th century. Once in Namibia they split and those that remained in the Northeast became the Himba, while those that ventured into the hinterland to settle became the Herero.

Namibia is also home to other non-Bantu-speaking groups such as the Damara (Barnard 1992) who speak Khoekhoe languages (Barbieri et al. 2014; Montinaro et al. 2017). The Damara are among the earliest known inhabitants of Namibia together with Khoe and San groups and make up around 8% of the population. However, despite speaking Khoekhoe, the Damara are genetically closer to other Bantu-speakers and may be descended from a population related to the Himba (Oliveira et al. 2018; Vicente et al. 2019).

The groups listed above will be the focus of this manuscript but Namibia alone harbours other minorities, and southern Africa as whole harbours many more ethnicities and languages (Oliveira et al. 2023).

Many ethnic groups inhabiting southern Africa today share ties in the form of ancient or ongoing admixture, shaping their genetic ancestry in one direction or another (Choudhury et al. 2021). Such events have been studied in recent years (Henn et al. 2011; Montinaro et al., 2017; Choudhury et al. 2021; Sengupta et al. 2021; Oliveira et al. 2023) improving our understanding of the complexity of African diversity. The

main groups (Bantu, San and Khoekhoe) have been shown to be highly divergent as the result of physical barriers and distances that enabled through time the emergence of genetic structure.

Other factors can explain population structure in the absence of physical barriers. These include assortative mating, philopatry, lifestyles, subsistence strategies and other cultural norms (Li et al. 2008; Marks et al. 2012; Robinson et al. 2017; Norries et al. 2019). These non-geographical factors are far more complex to investigate in human groups and the effects are so small that can only be detected at a micro-evolutionary scale. In fact, there is little empirical evidence on how much and for how long behaviours such as philopatry need to act on gene flow within populations in order to lead to the dissolution of the panmictic state of a sympatric population (Langergraber et al. 2007; Oota et al. 2001; Wilder et al. 2004).

How the interaction between cultural and geographical factors shapes human microbiomes is also difficult to investigate (Ruan et al. 2022). This is in part due to development of post-industrial societies and globalisation. These phenomena have transformed the traditional *modus vivendi* and social dynamics leading to greater levels of homogenization in many countries of the world. Globalisation has also impacted oral and gut microbiomes by imposing changes in traditional lifestyle and diet (Yatsunenko et al. 2012; Lasalle et al. 2018; Oduaran et al. 2020).

What factors exactly govern the microbial composition of the mouth and gut are still not perfectly understood, in particular the role of the host genome. The salivary microbiome in particular appears variable in measurements within the same individuals made at different points in time but it is still even more variable between different individuals (Armstrong et al. 2021). The core of the salivary microbiota therefore is both somewhat stable at the individual level (Gajer et al. 2012; Romero et al. 2014) and highly variable at population level with minimal population differentiation (Ruan et al. 2022). These differences have been reported to be driven by individuals' host behaviour. Geography has also been reported to correlate somewhat with small differences in salivary microbiomes (Ruan et al. 2022). However, other more tangible factors such as age, cohabitation, diet, tobacco and alcohol consumption can also play relevant roles (Yatsunenko et al. 2012; Shaw et al. 2017; Lasalle et al. 2018; Ruan et al. 2022).

To address some of these open questions, we report here new data in the form of imputed genomes and associated saliva microbiomes from 249 individuals from Lesotho and Namibia (Marks et al. 2012; Montinaro et al. 2016; Montinaro et al. 2017; Gonzalez-Santos et al. 2022). Along with the collection of the genetic material, metadata about each individual and their immediate ancestors was recorded in the form of variables such as ethnicity, birth location, clan affiliation and age. By combining the metadata with the genomic and metagenomic information we investigated the contribution these variables have on shaping local diversity, within and between individuals and groups.

Materials and Methods

Samples

Saliva samples from male individuals from Lesotho and Namibia (Fig. 1A) were collected between 2009 and 2010 with written consent from participants and approval by local ethics boards (Namibian Ministry of Health and Social Services, the Ministry of Health and Social Welfare of Lesotho and Oxford Tropical Research Ethics Committee (OxTREC)) (Marks et al. 2012; Marks et al. 2015; Montinaro et al. 2016; Montinaro et al. 2017; Gonzalez-Santos et al. 2022). The DNA used for sequencing had been previously extracted from saliva for previous works (Marks et al. 2012; Marks et al. 2015; Montinaro et al. 2016; Montinaro et al. 2017; Gonzalez-Santos et al. 2022).

The totality of the participants in Lesotho (n=103) identified as Basotho (SOT), a southern Sotho Bantu-speaking ethnic group. The sampling strategy in Lesotho was designed to

maximise spatial variation between Highland and Lowland ecozones. Three locations in the Lowlands (TB, MOR, ROM) and three locations in the Highlands (MEL, SEH, SLB) were visited for sample collection (see Fig. 1A). During the process of saliva collection, non-genetic, individual-level personal metadata about the participant was collected. The variables recorded, with varying degrees of missingness, were the following: ethnic group ("Ethnicity", variable only for Namibia), year of birth ("YoB"), place of birth ("Birth"), village of residence ("Village"), ecozone region ("Region", considered only for Lesotho) and clan affiliation ("Clan"). Furthermore, information about the birthplaces, clan and ethnicity of parents and grandparents of each participant were also recorded if known (Fig. 1B).

The participants from Namibia (n=146) self-identified as a member of four different groups: Himba (HIM), Herero (HER), Ovambo (OWA) and Damara (DAM) (Fig. 1A). Other ethnic groups are present in Namibia but we focused on these Bantu-speaking (HIM, HER, OWA) and KoeSan speaking (DAM) groups.

In South African societies, particularly among the Basotho, clans hold significant cultural and social importance. Theoretically, a clan is a group that traces descent from a common ancestor, often patrilineally, and can shape roles in identity, kinship, and social hierarchy. Clans can also be associated with totems, in the form of animals, plants, or natural elements. These totems may influence cultural practices, such as restrictions on eating or harming the totem animal. Clans often also have surnames strongly associated with them.

In Lesotho, two clans make up for more than half of the dataset (60% among participants, 57% at the parent and grandparent level). We labelled these as "large" clans. The remaining 11 clans were considered "small" (between 1-10 members among participants)

Information on clan affiliation in Namibia was either missing or virtually non-variable within groups, and therefore it was not further investigated. Similarly, Region and Ethnicity are overlapping in Namibia. We kept Ethnicity only as a more explicit descriptor of inter individual variation.

Sequencing and data analysis and validation

The genomic DNA samples extracted from the saliva of each individual were sequenced by Gencove to average depths of coverage below 1X, otherwise known as low-pass sequencing (Li et al. 2021; Martin et al. 2021), and then imputed as described in Li et al. 2021 using the 1000 Genomes Phase 3 haplotype reference panel. The resulting VCF files contain 37 million Single Nucleotide Polymorphisms (SNPs) of both true calls and imputed positions. We later filtered down the 37 millions to a set of 250k SNPs that we used for all later downstream analyses. This subset of markers was the result of the overlap between the imputed genotypes and Illumina genotyping chips (Infinium® Omni5-4 v1.2 BeadChip and Human610-Quad v1.0) that we leveraged for validating the imputation accuracy.

A subset of 23 imputed individuals included in the low-pass sequencing, had been previously genotyped with the platforms referred above and their genetic data was available (Montinaro et al. 2016, 2017). We used this subset of 23 individuals for direct and independent evaluation of the quality of the genotypes imputed (Li et al. 2021; Martin et al. 2021). We measured the quality of the imputation by checking SNP-by-SNP mismatch rate between the same individual (imputed vs genotyped) (SI Fig. 1). We achieved a mean imputation accuracy of 99.85% on this 250k subset of SNPs (247874 variants) (SI Tab. 1). We also confirmed these duplicates as homozygous twins. Despite the high accuracy achieved we acknowledge there might be limitations outside this subset of markers for diverse African groups following this imputation approach (eg. choice of reference panel).

Metagenomic classification of the non-human mapped reads of each individual sample was made using Kraken 2 and RefSeq database (including archaea, bacteria, eukarya and viruses reference sequences). This metagenomic data contained information about the relative abundance of bacteria and viruses up to the species level (Wood et al. 2019).

Population genetics tools

We used PLINK (v2.00a3) (Chang et al. 2015) to convert the individual VCFs containing 37 million variants into PED format and kept a subset of ~250k SNPs that overlap with Illumina Omni5. Once in PED format, we merged all individuals into one single file for downstream analyses. We used the PLINK --pca option to perform Principal Component Analysis (PCA) on the dataset. We also used the PLINK --homozyg function with default parameters to identify Runs of Homozygosity (RoH) in each individual genome. PLINK with default parameters was also used to obtain individual heterozygosity. We used ADMIXTURE (Alexander et al. 2009) with partitions from k=2 to k=5.

Relatedness measurement tools

We estimated pairwise kinship coefficients within and between ethnic groups in the dataset using KING-Robust algorithm (Manichaikul et al. 2010) in KING software (v2.2.7) using the --kinship option. The kinship coefficient estimation was calculated using only SNPs available for each given pair of individuals (SI Fig. 2). Positive values are predictive of the degree of relatedness between individuals. Coefficient values >0.354 indicate duplicates, values between 0.354-0.177 indicate first degree relationship, between 0.177-0.0884 indicate second degree and between 0.088-0.044 third degree. Values for unrelated pairs are typically around zero. Negative values of the kinship coefficient suggest substantial differentiation between the two individuals.

Following the geometric framework described in Oteo-Garcia and Oteo 2021, we used the L2 norm to measure the genomic distances between pairs of individuals, which estimates the Euclidean distance between individuals projected in a multidimensional space defined by the allele frequencies or genotypes (n) of the 250k SNPs used in the dataset. Given two points in a given n-dimensional space the Euclidean distance between them is given by $\sqrt{(X1-Y1)^2 + (X2-Y2\}^2 + ... + (Xn-Yn)^2}$. The unit for L2 distances is the same unit as the original elements that are being compared. In this case, we used multiple loci which have no units and therefore L2 distances presented here are not accompanied by a defining unit of measure, just a numeric value (Oteo-Garcia and Oteo 2021). L2 is correlated with both the proportion of markers with zero identical-by-state SNPs and KING kinship coefficients (SI Fig. 3). For kinship analysis, L2 values correspond to KING kinship coefficients as follows: L2 <330 indicates first degree, L2 >330 is second degree, L2 >360 is third degree. L2 between 375-385 indicates no close or discernable kinship but may be indicative or inbreeding. L2 >385 indicates clear unrelatedness. Related pairs of individuals up to the 3rd degree were removed from downstream analyses (SI Fig. 2).

Metagenomic diversity measurements

We calculated saliva metagenomic alpha diversity for each individual at phylum level. We also calculated alpha diversity at species level using a subset of potentially pathogenic bacteria species extracted from a table originally presented in Warinner et al. 2014 (Lewy et al. 2019). We chose Shannon diversity index (H') for the alpha diversity following the standard formula ($H' = -\sum pi \ln pi$), where pi is the relative abundance of each taxon within each individual.

Conversion of raw metadata into binary independent variables

We summarised and translated the metadata associated with each individual into binary variables suitable for linear regression models. We started by recording if birthplace locations coincided ("Birth" variable) between two individuals and turned that information into a binary independent variable (Yes vs No). We recorded the same for the "Clan", "Village", "Region" and "Ethnicity" variables. At the "Parent" and "Grandparent" level of metadata comparisons between pairs of participants, the presence of matches in relation to place of birth was summarised as either being present or not, independently of the number of matches. For example, for "Grandparents" we could find eight matches for the birthplace, zero or any intermediate combination. This created many categories but it was rare finding more than 2 matches so we summarised it as a binary.

Regression and calculation of beta coefficients

To compute the linear regression models and other statistical tests, we used the R function *lm* from *The R Stats Package* (version 3.6.2). The following independent binary (Yes=1 , No=0) variables were tested considering a genomic true distance between two individuals (L2) as the dependent variable. To provide an indication of the degree of association between variables, Phi coefficients between each pair of binary variables were calculated with the *sjt.xtab* function from the *sjPlot* package in R. The values of the phi coefficient can range from -1 to 1 and indicate the degree of association between the binary variables created here. The closer to zero the coefficient is, the more independent the two variables are between them (SI Tab. 3).

The following independent variables were tested when individual heterozygosity and saliva alpha diversity were tested as the dependent variables: age at the time of collection (based on "YoB"), "clan consistency" (defined as the fraction of immediate ancestors (parents and grandparents) who are reported to belong to the same self reported clan of the individual), sampling area, region, clan, and shared birthplaces in parents and grandparents.

To correlate genomic distance with geographical distance (geodesic units to correct for Earth's curvature) we used the distance between birthplaces of all individuals with available coordinates. However, this was only possible for the 103 individuals from Lesotho but none for Namibia. The Procrustes analysis comparing geographic and genetic distances was made in R using the Vegan package (version 2.6-4) with default parameters and iterating (n=10000).

Results

Insights from genomic profiles

We visualised the data collected from the individuals in Lesotho and Namibia (Fig. 1A-B) in its genomic and metagenomic form for initial exploration of the variation present in the dataset using PCA (Fig. 1C-D, SI Figure 4), ADMIXTURE (SI Fig. 5) and RoH (SI Fig. 6).

The degree of inter and intra population variation in the genomic PCA shows different trends in Lesotho and Namibia (Fig. 1B). The PCA indicates a high degree of genomic homogeneity among the Basotho people of Lesotho. Except for one outlier, no differences were present between Highland and Lowland regions. Genomic distances in Lesotho are not impacted by Khoe-San introgression (SI Fig. 7; Gonzalez-Santos et al. 2022).

In Namibia on the other hand, we observe population structure over three clusters. One cluster consists of the highly homogeneous Ovambo with only one outlier. The other cluster comprises Himba and Herero individuals, with some indication of admixture between the Ovambo and the Himba (Fig. 1C). There are two inbred Himba outliers (SI Fig. 6) on the top left corner of PCA (Fig. 1C). The third cluster is formed by the Damara, with some samples scattered as outliers (Fig. 1B) due to higher levels of Khoe-San-related ancestry (SI Fig. 4).

In contrast to Figure 1C, the PCA is based on phyla composition (Fig. 1D) shows complete overlap between ethnic groups. Structure appears to be mostly driven by the relative quantity of three phyla present in the saliva (SI Fig. 11). The relative quantity of Bacteroidetes in the saliva of each individual contributes to much of the variation along PC1, explaining 60% of total variation. The ratio between Firmicutes and Proteobacteria explains another 25% along the PC2 axis (Fig. 1D). Note that saliva samples were collected at different times during the day without recording further details on diet, drinking or smoking habits, known to affect bacteria diversity (Fan et al. 2018; Liao et al. 2022; Wirth et al. 2020; Al-Zyoud et al. 2020; Belstrøm 2020).

ADMIXTURE results indicate that K=2 is the best clustering in this dataset (SI Fig. 5). This partition separates the Namibian Bantu-speakers together with the Khoisan-speaking Damara from the Bantu-speakers from Basotho. This likely reflects an ancestral split between these two groups of Bantu-speaking populations (Choudhury et al. 2021) with the Damara being ultimately derived from an ancestral population

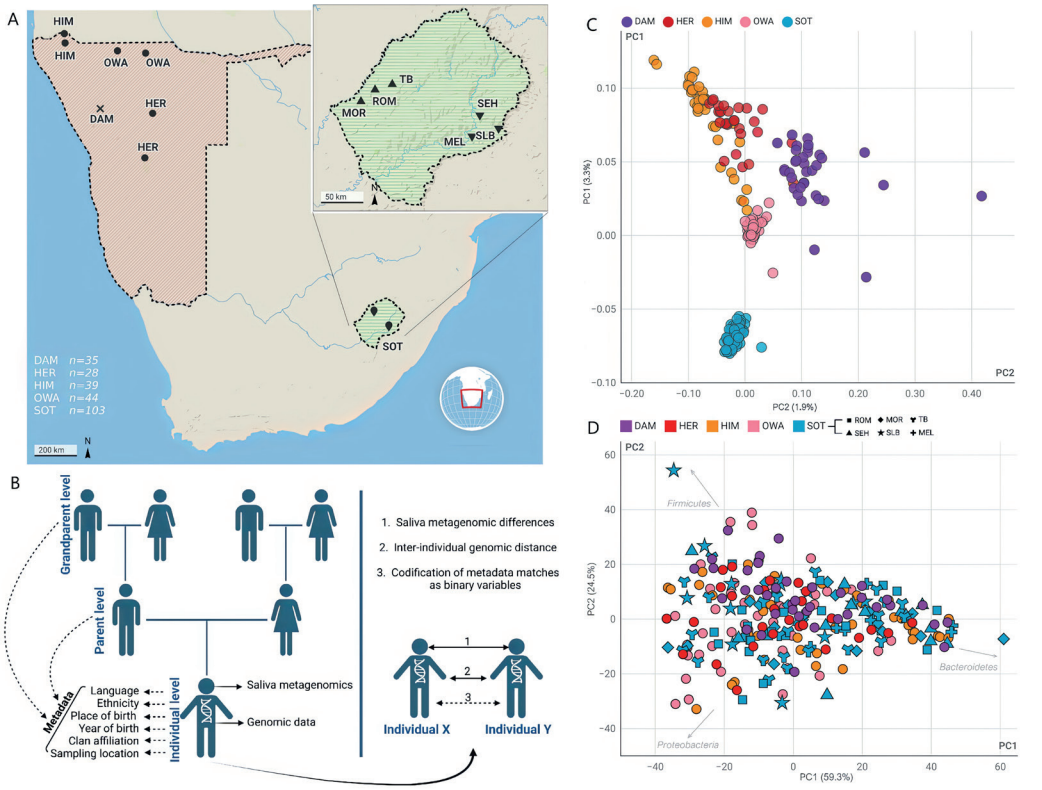

***Fig. 1** - A) Map with the populations sampled in Namibia (in red); Himba (HIM), Ovambo (OWA), Herero (HER), Damara (DAM), and sampling locations in the Lowlands (MOR, ROM, TB) and Highlands (SEH, SLB, MEL) of Lesotho (in green). B) Summary of metadata, genomic and metagenomic information collected (left) and processing of the genomic and metagenomic data and metadata for analyses (right). Made with Biorender. C) Principal component analysis coloured by population. D) Principal component analysis based on the composition of the main phyla. Arrows point towards the biplot direction of Bacteroidetes along the PC1 axis and Firmicutes-Proteobacteria along PC2 axis.*

related to the Himba (Oliveira et al. 2018). These individual ADMIXTURE profiles based on imputed data (SI Fig. 1) were compared with ADMIXTURE results of a subset of individuals for which SNP array data was available. We observed the same results using the imputed and genotyped data (SI Fig. 1).

Inter Individual genetic distances as a function of cultural and geographical variables

In order to explore how geography and culture influence inter-individual and intra-individual variation we estimated pairwise L2 genomic distances and individual heterozygosity for 103 and 146 individuals in Lesotho and Namibia respectively (Fig. 2A-B). The distribution of within-group variance of L2 inter-individual distances varied across populations but had similar means (SI Tab. 4). Only the Basotho and Ovambo had somewhat normal distributions confirming solid genomic homogeneity (Fig. 2A). The stretched L2 distributions in other groups indicate recent or ongoing admixture processes. Figure 2B suggests that the different distributions of L2 measurements are not totally conditional to the heterozygosity of

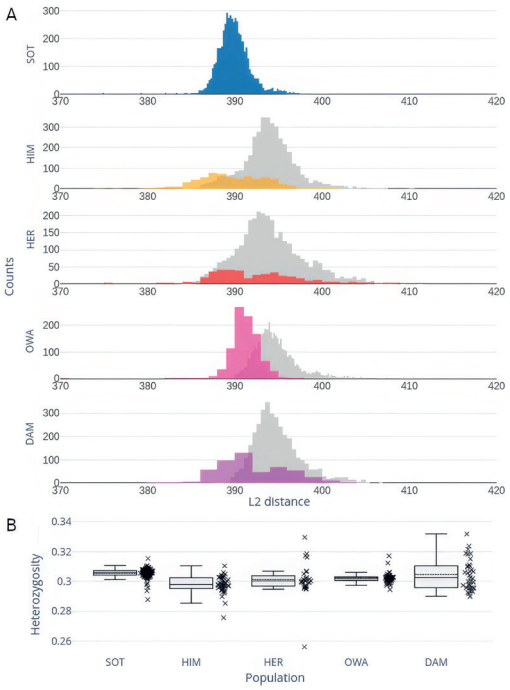

***Fig. 2** - A) Histograms of inter-individual L2 genomic distances within (in colour) and between ethnic groups (in grey, for Namibia only). The inter-population distances (in grey) refer to all the comparisons between the indicated population and the rest of the Namibian groups. Basotho distances were estimated only within the population of Lesotho. Mean values of within and between populations genetic distances are reported in SI Table 4. B) Boxplots displaying intra-individual heterozygosity estimates by population.*

each population. The heterozygosity of Bantu-speaking groups is similar across ethnicities (Fig. 2B). It is highest among the Basotho and lowest among the Himba. Damara have higher values for individual heterozygosity (Fig. 2B), likely linked with the higher levels of non-Bantu ancestry in some of the individuals (SI Fig. 4).

Given that previous investigations have shown that even low levels of genetic differentiation can be discriminant for geolocalization (Novembre et al. 2008), we tested to what extent this was reflected in our dataset. We focused on Lesotho since the geo-coordinates for birth locations of individuals could be retrieved with confidence. In Namibia, samples belonging to the same ethnic group were either sampled from the same place or were impossible to assign to a geographic location and therefore this test could not be applied.

For any two given individuals in Lesotho, we found a significant association between L2 genomic and geographic distances but the amount of variation explained by this simple model was very small and the regression β coefficient was close to zero ($\beta=6x10^{-6}$, R-squared=0.02, $F_{1,1483}=26.04$, p-value<0.001). The association disappears when considering distances only within the Lowlands or within the Highlands of Lesotho separately, implying the lack of further structure at smaller distances within regions. We ran Procrustes analysis independently to confirm this result. We found that Procrustes supports a significant correlation between geographic distance and genetic distance between pairs of individuals (SI Tab. 5).

Following this result, we explored whether the metadata collected could be correlated to the degree of genetic similarity between individuals. We initially ran simple regression analysis for each variable to identify those to be further explored. We tested the variables describing pairs of individuals in terms of their geographic provenance (place of "Birth", "Region" of residence and "Village" of residence) and cultural affiliation (ethnicity and clan affiliation, the latter only for Lesotho as in Namibia the records were too incomplete and/or ethnicity was strongly correlated with clan affiliation) (SI Fig. 8, SI Fig. 9). We also included information about birthplaces of parents and grandparents. We considered Namibia and Lesotho separately because they are geographically distant countries. All the variables were significantly associated with genetic distances between individuals, all with negative β coefficients except for "Clan" in Lesotho (SI Tab. 2). Of all the variables, "Ethnicity" in Namibia explained the largest fraction of the variance, almost 20%. Taken singularly, the other variables tested in Namibia explain 3-5%, while in Lesotho the largest contribution is below 2%.

*Tab. 1 - Summary of the results obtained for the estimated beta coefficients (β) and intercepts of the multivariate regression models estimated for Lesotho and Namibia. NA: not applicable, as the variable was not tested (see Materials and Methods). *Excluded from analysis due to missing data in Namibia (SI Fig. 9).*

	LESOTHO		NAMIBIA	
MULTIPLE REGRESSION VARIABLES	β COEFFICIENT	P-VALUE	β COEFFICIENT	P-VALUE
Intercept	390.21	<0.001	394.19	<0.001
Parents	-0.77	<0.001	+0.95	<0.001
Village	-0.32	<0.001	+1.12	<0.001
Region	-0.23	<0.001	*NA*	*NA*
Clan	+0.26	<0.001	*NA**	*NA**
Ethnicity	*NA*	*NA*	-3.03	<0.001
Birth	-0.15	0.43	-1.06	0.012
Grandparents	+0.07	0.68	-1.05	0.016

We then evaluated the degree of correlation between variables using the Phi coefficient (SI Tab. 3). In the Lesotho dataset the strongest signals of associations were by far those involving birthplaces of Parents-Grandparents and Parents-Birth. "Clan" did not show any evidence of association. On the contrary in Namibia variables were all associated to some extent (SI Tab. 3).

We implemented a multiple regression analysis that included all the explanatory variables (Tab. 1). We found that for Lesotho four out of the six tested variables in the multiple regression model were significant. Three of them have a negative coefficient and one has a positive modulating effect. The final model explains about 3% of the total variation (R-squared=0.03, $F_{6,3738}$=19.82, p-value<0.001), while the absolute contribution of the β coefficients was small, ranging between 0.20 and 0.77 points (Tab. 1). In Namibia the multiple regression model found three of the five tested variables statistically significant, explaining 35% of the variation in genetic distance between individuals (R-squared=0.35, p-value<0.001) (Tab. 1). The β coefficients were slightly larger than those in Lesotho, but "Ethnicity" was almost three times so. The sharing of birthplace locations at the individual and grandparent level did not contribute significantly to the model, possibly due to association with other variables (with "Parents" in Lesotho, with "Parents" and "Village" in Namibia; SI Tab. 3).

Of the four significant variables found in Lesotho, three are broadly related to geography ("Parents, Village, Region"), whereas the fourth ("Clan") is culturally defined and the only one generating an increase of L2 distances and differentiation when shared. It has been previously shown that clan affiliation is paternally transmitted with high fidelity in Lesotho (Montinaro et al. 2016). We confirmed this instance for our dataset (99% matching rate between participant and father's clan; 99% of matches between father and paternal grandfather clans). We also observed that 77 out of 103 (75%) parental marriages occurred between individuals belonging to different clans. In addition, more than 90% of the times at least one grandparent is from a clan different from the others, suggesting that clan membership does not represent a barrier for marriage and mating in Lesotho. Furthermore, average genetic distances within clans are larger than between clans (389.94 vs 389.82), the difference

being small but significant. To investigate possible elements shaping inter-clan marriage, we calculated an index of "clan consistency". In this manner, we obtained an average value of consistency for each clan, and, indirectly, of its permeability (measured as 1-consistency) (SI Fig. 10). We then ran multiple regression analysis explaining "Clan consistency" as a function of year of birth, region (Highlands/Lowlands), village of residency, and clan size (large/small; see Material and Methods). We found that being from a small clan and from the Lowlands region generated a statistically significant decrease in clan consistency, β values -0.14 and -0.1 respectively (R-squared=0.15, $F_{2,100}$=8.9, p-value<0.001). We reasoned that permeability might operate by increasing intra-clan genetic variation. We therefore tested the extent to which such dynamics might be reflected in the genetic profiles of individuals by comparing within clan L2 distances considering Lowlands and Highlands, small and large clans (SI Fig. 10). Unexpectedly, large clans and clans in the highlands showed significantly larger L2 distances than small clans and clans in the Lowlands despite both being less permeable (SI Fig. 10).

We finally explored if and how metadata variables affected individual variation, measured as the degree of heterozygosity, considering Namibia and Lesotho separately. Of all variables tested (age, locations, consistency of clan of affiliation in Lesotho only, birthplaces of immediate ancestors, ethnicity) we found that clan consistency in Lesotho is the only one that modulates genome heterozygosity when tested singularly. A higher index of clan consistency correlates negatively with heterozygosity (β=-0.004, R-squared=0.07,$F_{1,101}$=7.14, p-value=0.009) but the contribution disappears when considered together with other variables. No variable was significant for individual heterozygosity in Namibia.

The role of age and lactase persistence in shaping saliva microbiome composition

Ethnicity does not play a role in determining the composition of the saliva microbiome since no evident population structure can be observed at phylum level (Fig. 1D). We also found no outstanding differences in composition at the level of the subset of species we looked at (SI Fig. 12). However, some patterns arise when looking at particular taxa and viruses (SI Fig.s 17,19 and 20).

The pathogenic bacterial composition and pathogen load in the saliva of the populations investigated (Fig. 3A, SI Fig. 12) was similar. The mean and median content of pathogenic bacteria in the saliva in the dataset are 13.3% and 11.2% respectively. The highest mean load is among the Basotho (14.4%) and lower among the Namibian Bantu-speakers (Ovambo 12.5%, Himba 12.5%, Herero 11.3%), and intermediate in the Damara (13.7%) (Fig. 3A). ANOVA results show no statistically significant differences between the different populations (p-value = 0.423). Overall we found no evidence for differential pathogen load in the saliva between these groups.

However, we found striking differences between Lesotho and Namibia regarding the viral load in the saliva (Fig. 3C). The viral load among Namibian groups is very low, with many individuals below the threshold of detection (44%) compared to Lesotho (27%). In Lesotho we found more individuals with elevated viral loads and many outliers with outstanding loads (around 15% of the sample size).

HIV is not detectable in saliva. However, since its prevalence among the adult male population in Lesotho is very high (~25%) we decided to check whether traces of HIV in the saliva could be behind this observed difference between the two countries. As expected, we did not find any HIV sequences in the saliva metagenomic data. Even after mapping individuals with the highest viral loads specifically to the HIV-1 reference genome.

The viral load is almost entirely driven by the amount of human herpesvirus 4 (HHV-4), known as the Epstein-Barr virus (Fig. 3D). High amounts of HHV-4 can be associated with high prevalence of HIV, as the former can be a hitchhiking co-pathogen (Munawwar and Singh 2016). Highest viral load carriers are individuals born in Lesotho between 1959 and 1979 (Fig. 3E). We also observed an unexpected deficit of

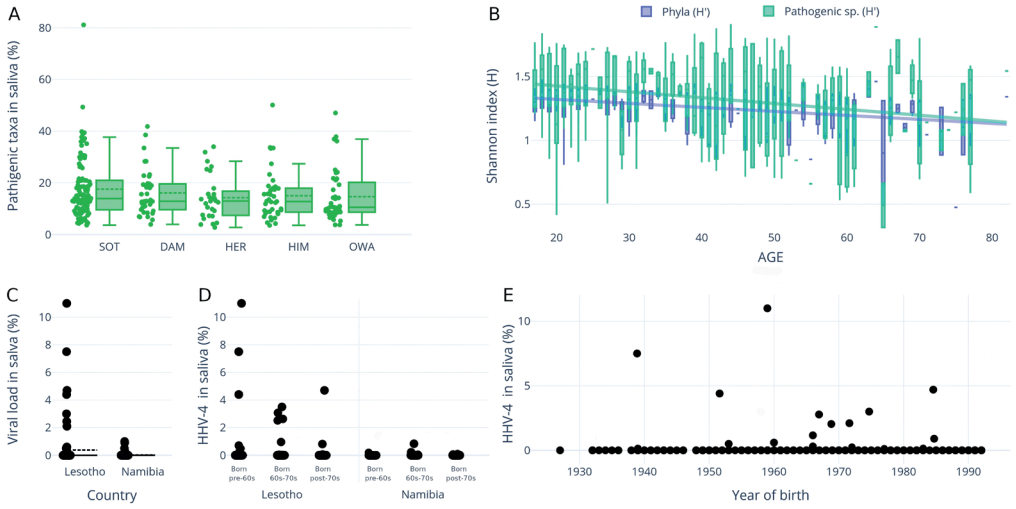

***Fig. 3** - A) Pathogenic species load in the saliva. B) Correlations between participants age at the time of collection and saliva alpha diversity (Shannon index) at phyla level and a subset of pathogenic species. C) Total viral load in saliva for Lesotho and Namibia. D) Viral load of HHV-4 in saliva by age group and country. E) Viral load of HHV-4 by sample by year of birth in Lesotho.*

people of that same age group (between the ages of 35 and 49 in 2009 when sampling took place) (SI Fig. 13), even after correcting for migration rate by age group (SI Fig. 14, SI Fig. 15) (United Nations 2023).

We wondered if the impact of the HIV pandemic could have affected the genetic profile of people on Lesotho by operating as a selective pressure on the frequency of alleles associated with protection against HIV. We tested if SNP rs9264942, known to confer certain protection against HIV effects (Herráiz-Nicuesa et al. 2017), had different allele frequencies in different cohorts (SI Fig. 16) but we did not detect significant shifts in the frequency of this variant, perhaps due to sample size limitations.

We then explored the possible relationship between host genotypes and microbiota composition by focusing on the metabolic ability of digesting lactose. It has been previously shown that a negative correlation exists between the occurrence of *Bifidobacterium* and the ability of the host in digesting lactose (Goodrich et al. 2016). Similar to *Bifidbactierum*, also the presence of *Lactobacillus* is related to the consumption of milk-related products (Vlasova et al. 2016). Interestingly, lactase persistence (LP) phenotype is variable in Namibia (Breton et al. 2014; Macholdt et al. 2014), and such variation offers the opportunity to explore the link between bacteria associated with dairy products consumption and LP alleles.

We evaluated the traces of *Bifidobacterium* and *Lactobacillus* in the saliva, two widely recognised genera with probiotic properties (Goodrich et al. 2016; Vlasova et al. 2016) whose presence is conditioned by the ability to digest lactose by the host. We found that *Lactobacillus* is ubiquitous in all populations with detection rates between 80-98%. Nonetheless, some populations (Ovambo, Himba, Basotho) display greater relative prevalence of *Lactobacillus* than others (Damara, Herero) where it is only detected at minimum threshold levels (Fig. 4). The detection rate of *Bifidobacterium* is different, it ranges between 45%-60% in the Ovambo, Herero, Himba and Basotho. The prevalence in these groups is always extremely low percentages just above threshold detection (Fig. 4). Remarkable traces of *Bifidobacterium* are found only in the

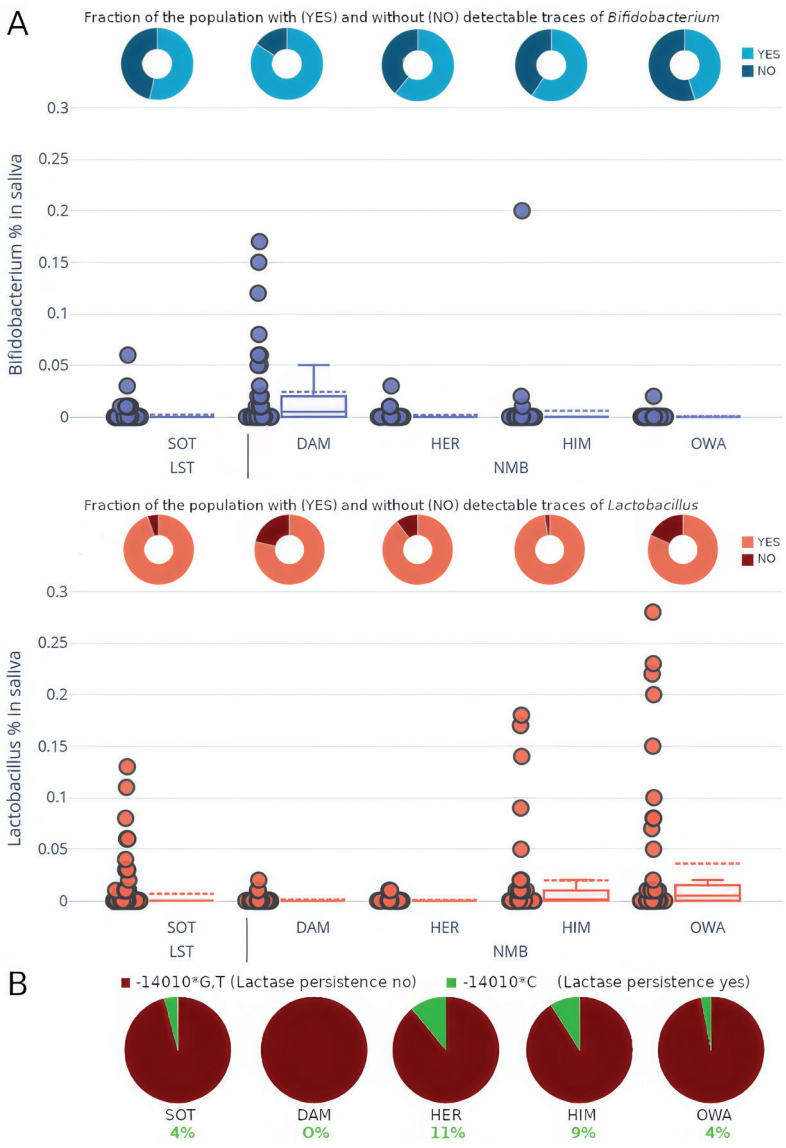

***Fig. 4** - A) Bifidobacterium and Lactobacillus rates of detection by population and prevalence in the saliva microbiome of each individual in Lesotho (LST) and Namibia (NMB). B) Allele frequency of SNP rs145946881 corresponding to each group at the bottom plot. The SNP is variant -14010*C in gene MCM6 (chr2:136608746 in Hg19 reference genome) which confers ability to digest lactose.*

Damara (Fig. 4). The rate of detection (>80%) is also considerably higher among the Damara compared to other groups (Fig. 4).

To further investigate correlation between the presence of *Bifidobacterium* in saliva and the ability to digest lactose at the population level (Ranciaro et al. 2014; Anguita-Ruiz et al. 2020; Campbell and Ranciaro 2021), we calculated the prevalence of the allele -14010*C in gene MCM6 that confers post-weaning LP phenotype in these populations

(Fig. 4) (Tishkoff et al. 2007b; Jensen et al. 2011; Macholdt et al. 2014; Macholdt et al. 2015). We found this variant in all Bantu-speaking groups of both countries, with 4% prevalence in Lesotho and around 8% frequency in Namibia (Fig. 4). This result is consistent with known allele frequencies for some of these groups (Breton et al. 2014; Macholdt et al. 2014; Anguita-Ruiz et al. 2020). No apparent imputation bias despite potential role of selection in driving loci frequency (Burger et al. 2020). The allele is not present in the Khoisan-speaking Damara despite being the population with highest admixture related to the Khoisan-speaking Nama which carry this allele in high frequencies (Breton et al. 2014; Macholdt et al. 2014). We observed that high proportions of *Bifidobacterium* and *Lactobacillus* are always found in individuals that are non-carriers of the LP allele. Carriers of the allele (23 individuals out of 249 in the dataset) do not display traces of the strains in the case of *Bifidobacterium* and with only a couple of exceptions in the case of *Lactobacillus* (SI Fig. 17).

We explored what metadata variables ("YoB", "Ethnicity", "Region", "Parents/Grandparents" and heterozygosity) might affect alpha diversity using both phyla and a subset of putative pathogenic bacteria (Warinner et al. 2014) (Fig. 3A). We found that saliva alpha diversity is independent of all variables tested except for age (Fig. 3B). We did not detect rarefaction bias in the alpha diversity since measurements were found to be independent from the number of metagenome reads of each sample (SI Fig. 18).

We found that age shows a weak but significant negative correlation with alpha diversity in the saliva. The Shannon index (H') for phyla diversity decreases with age (Tab. 2). This is also true when only putative pathogenic species (Warinner et al. 2014) are considered (Tab. 2) (Fig. 3B, SI Fig. 19).

The three most common pathogens were *P. gingivalis*, *V. parvula* and *T. forsythia*. While *P. gingivalis* and *V. parvula* seem to be in competitive exclusion, *P. gingivalis* and *T. forsythia* occurrence appear to be synergistic (SI Fig. 20).

***Tab. 2** - Simple regression model fits for the association between age and saliva alpha diversity calculated with phyla metagenomic composition and putative pathogenic species composition.*

SIMPLE REGRESSION - AGE & H'	PHYLA	PATHOGENIC SPECIES
β coefficient	-0.003	-0.005
R-squared	0.09	0.07
p-value	<0.001	<0.001

Discussion

In this work we attempted to capture which socio-cultural factors in one's life can shape inter and intra-individual genomic differentiation (Uren et al. 2016; Montinaro et al. 2016, 2017; Atkinson et al. 2022) and salivary metagenomic composition, using Lesotho and Namibia as case studies.

We found small scale geography-related factors with small effect, replicating patterns that are seen at continental level (Novembre et al. 2008; Prugnole et al. 2005; Ramachandran et al. 2005; Li et al. 2008; Vicente et al. 2019).

This observation was made in Lesotho. Our results indicate that physical distance between participants' place of origin is predictive of genomic affinity. For pairs of individuals within the same ecozone (Highlands or Lowlands) genomic distances are smaller compared to trans-ecozone measurements. The contribution of the effect is significant but small, both in terms of explained variation and coefficients. This pattern is not replicated when considering measurements within each ecozone.

We found complex interactions between the variables in the multiple regression models. Some variables capture the information provided by others that are significant in simple regressions. Our models were also sensitive to filtering of individuals by degree of missing data.

For example, sharing the same birthplace of at least two grandparents was found not significant in Lesotho and Namibia when together

with the other variables. The informativeness is captured by the sharing of birthplaces at parental level ("Parents" variable in Table 1, SI Fig. 2).

These results suggest that lack of information about the grandparents is not particularly concerning if the same information is already known about the parents. However, we caution that it might not hold true universally. In European contexts grandparents metadata has been shown to be informative of the genetic composition of individuals (Leslie et al. 2015; Bycroft 2018; Raveane et al. 2019).

Cultural variables like "Ethnicity" and "Clan" affiliation contribute significantly to inter-individual genetic distances. When the overlap in contributions by different variables was tested by removing one variable each time, we noted that only cultural-related variables remained unaffected, indicating their non-redundant contributions to the description of patterns of local diversity in opposition to geography-related ones (birthplaces of "Parents" and "Grandparents"). However "Ethnicity" in Namibia and "Clan" in Lesotho operate differently.

When not shared, "Ethnicity" is the major contributor to genetic differentiation in Namibia despite some shared origins and gene-flow between groups (Uren et al. 2016; Schlebusch et al. 2012; Pickrell et al. 2012; Oliveira et al. 2018).

In Lesotho, sharing "Clan" affiliation contributes to increased inter-individual differentiation. This effect appears to be driven by the influence of large clans that are genomically more heterogeneous than small ones. This is supported by evidence from the Y chromosome. Although paternal co-transmission of Y chromosome lineages and clan membership is well attested, clans are also very variable in Y chromosome haplogroup composition (Montinaro et al. 2016). Large clans have the lowest permeability for members of other clans marrying into them (in the last two generations). Large clans also have the largest variance of intra-clan genomic distances (SI Fig. 10).

Based on these findings, we infer fluid social dynamics that led these large clans to be more heterogeneous units today. Our results suggest an amalgamation of peoples for the origin of modern large clans in Lesotho. Interestingly, this is actually reflected in the socio-political events developing in Lesotho during the late 19th century under the leadership of Moshoeshoe I.

We also observed that inter-clan marriages are influenced by the sampling location (Lowlands or Highlands). Clans in the Highlands showed similar permeability but larger inter-individual distances than Lowlands, which reflect a more heterogeneous genomic composition in this region.

Overall, our results are in line with previous observations reporting geography as a predictor of genomic similarity between individuals, even at smaller scales in Lesotho. Despite Lesotho being a highly homogeneous country, we detected a subtle ongoing process of genomic differentiation between regions.

In Namibia, Ethnicity is the main force behind genomic distances. However, Namibia has very low population density and ethnic groups are highly structured territoriality, therefore an implicit correlation with geographic distances exists, despite signatures of past and ongoing gene flow between groups.

The importance of the parental birthplaces above other variables highlights the role of recent family history and the limitations of mobility that shape inter-individual genomic affinity. This is a clear example of the mechanisms behind isolation by distance, a common phenomenon in human populations across continents. It also highlights that signals observed at macro level, in part, arise from multiple local small contributions.

The broad composition of saliva microbiomes in our dataset is consistent with previous reports (Murugesan et al. 2020). We found no correlation between geography or ethnic divisions and microbiome composition in the saliva (Ruan et al. 2022). We reason that other factors must be responsible for differences in metagenomic composition, in accordance with Shaw et al. (2017).

However, we found a correlation between age and saliva microbiome alpha diversity (Takeshita et al. 2016; Lewy et al. 2019; Murugesan et al.

2020). We show that this correlation is negative and, although weak, significant (Fig. 3B). Overdominance of certain putative pathogens linked to diseases more typical in the old age, such as gingivitis and periodontal disease (e.g. *P. gingivalis, T. forsythia*), (SI Fig. 15, SI Fig. 16) seem to be the cause for this negative correlation. However, we did not find statistical support for any particular taxa. Although we did not find evidence that Bacteroidetes is an overrepresented taxa in the elderly, we show that Bacteroidetes is a key contributor to saliva microbiome variability (Fig. 1D).

Regarding host-microbiome dynamics, we focused on *Lactobacillus* and *Bifidobacterium* although saliva is likely not their preferred niche. However, we observe that *Lactobacillus* is easily found in the saliva of all groups (Fig. 4) regardless of the LP allele in the population. Interestingly, a noticeable percentage of the Ovambo, Himba and Basotho, all of them populations with LP phenotypes, display high quantities of *Lactobacillus.* Whereas the Damara and Herero show a minimum prevalence level despite similar high detection rates. On the other hand, *Bifidobacterium* appears to be detected in about half of the individuals in each of the Bantu-speaking groups and with low prevalence. It is only in the Damara (non LP) where high detection rate and high prevalence can be observed (Fig. 4).

We reason that these high traces of *Lactobacillus* might actually be related to consumption of fermented dairy products (e.g. *Mafi*, *Omashikwa*) typical in these countries (Gadaga et al. 2013; Misihairabgwi and Cheikhyoussef 2017). This would mean that the *Lactobacillus* presence in the saliva is not directly linked to the LP genotypes. *Bifidobacterium* presence in saliva appears to actually be conditioned by host genotype since it is only able to prevail significantly in the Damara, the only population in the dataset with no LP alleles. Notably, this replicates known interactions of host-*Bifidobacterium* in the gut (Goodrich et al. 2016). In addition, both taxa do not thrive within individuals with LP genotypes (SI Fig. 17).

The excess of HHV-4 viral load in the saliva of Lesotho (Fig. 3C) is striking but might have a simpler explanation linked to the HIV epidemic in the country. Due to the nature of the virus itself, we were unable to detect HIV in the saliva. However, since HHV-4 is commonly transmitted via saliva it was easily traceable in our samples. Since HHV-4 is correlated with HIV load (Talenti et al. 1993; Sachithanandham et al. 2009, 2014; Di Gennaro et al. 2023; Wan et al. 2023), we argue this might be a good proxy to infer HIV infection. This explains the contrast with Namibia and fits with the profile of the age group that is most affected (30 to 50-year-olds in 2009). This same age range of males has been shown to be the one more impacted by HIV infection (Schwitters et al. 2022). The deficit of people is particularly acute for those born between 1965 and 1969.

This missing percentage of the population is likely due to excess mortality caused by the HIV epidemic that spread rapidly among the youth of the 1990s decade (SI Fig. 15). Those born between 1959 and 1979 were likely to be more sexually active in the 1990s and therefore more exposed to HIV during the height of the epidemic that hit Lesotho extensively worse than Namibia.

These results validate saliva microbiome use to explore local dynamics relating to diet and health at population level through a low-pass sequencing strategy. Firstly, by means of illustrating how both culture and genes can condition dietary practices. Secondly, by proving differences between countries hit differently by a viral epidemic in the 1990s. Future studies should expand the sample size and focus on the impact on the genome of such elevated viral load in Lesotho. In particular, trying to uncover whether any natural selection has occurred in a given loci due to the great selective pressure exercised by the HIV epidemic. We caution however, that our sample size was small and we lacked true case and control HIV groups to systematically approach this question and a signal might have gone undetected in our analysis. Significant findings on this front could have important medical implications.

Acknowledgements

We would like to thank the people who volunteered DNA and made this study possible. We also thank Erika Oosthuizen for facilitating the collection of samples in Namibia and Ntjapeli Matlanyane, Nthontsi Qokolo and Chiara Batini for their contribution to the collection of the Lesotho samples. We thank Francesco Montinaro, Alessandro Raveane and Ryan Daniels for the support at different stages of the project. In addition we thank: the COMP-HUB Initiative, funded by the 'Departments of Excellence' program of the Italian Ministry for Education, University and Research (MIUR, 2018-2022); the HPC (High Performance Computing) facility of the University of Parma; The Doctorate program in Biotecnologie e Bioscienze (University of Parma); CC would like to acknowledge Gencove for generating the genetic data here analysed as part of the Gencove Initiative.

Data Availability

Requests for individual metadata should be addressed to C. Capelli (cristian.capelli@unipr.it). The processed genomic VCF data used in the analyses is available at https://ngdc.cncb.ac.cn/gvm/# under accession PRJCA019089. Raw data available at ENA under accession PRJEB70516.

Orcid

Gonzalo Oteo-García
https://orcid.org/0000-0002-0957-4014

Matteo Manfredini
https://orcid.org/0000-0001-7377-5921

Cristian Capelli
https://orcid.org/0000-0001-9348-9084

Giacomo Mutti
https://orcid.org/0000-0002-8687-3333

Matteo Caldon
https://orcid.org/0000-0002-5108-2021

References

1000 Genomes Project Consortium (1KGPC) (2015) A global reference for human genetic variation. Nature 526:68–74. https://doi.org/10.1038/nature15393

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

Anguita-Ruiz A, Aguilera CM, Gil Á (2020) Genetics of lactose intolerance: An updated review and online interactive world maps of phenotype and genotype frequencies. Nutrients 12:2689. https://doi.org/10.3390/nu12092689

Armstrong AJS, Parmar V, Blaser MJ (2021) Assessing saliva microbiome collection and processing methods. NPJ Biofilms Microbiomes 7:81. https://doi.org/10.1038/s41522-021-00254-z

Al-Zyoud W, Hajjo R, Abu-Siniyeh A, et al (2019) Salivary microbiome and cigarette smoking: a first of its kind investigation in Jordan. Int J Environ Res Public Health 17:256. https://doi.org/10.3390/ijerph17010256

Atkinson EG, Dalvie S, Pichkar Y, et al (2022) Genetic structure correlates with ethnolinguistic diversity in eastern and southern Africa. Am J Hum Genet 109:1667–1679. https://doi.org/10.1016/j.ajhg.2022.07.013

Barbieri C, Vicente M, Rocha J, et al (2013) Ancient substructure in early mtDNA lineages of southern Africa. Am J Hum Genet 92:285–292. https://doi.org/10.1016/j.ajhg.2012.12.010

Barbieri C, Güldemann T, Naumann C, et al (2014) Unraveling the complex maternal history of Southern African Khoisan populations. Am J Phys Anthropol 153:435–448. https://doi.org/10.1002/ajpa.22441

Barnard A (1992) Hunters and herders of Southern Africa: A comparative ethnography of the Khoisan peoples. Cambridge Studies in Social and Cultural Anthropology, Cambridge University Press.

Belstrøm D (2020) The salivary microbiota in health and disease. J Oral Microbiol

12:1723975. https://doi.org/10.1080/20002297.2020.1723975
Breton G, Schlebusch CM, Lombard M, et al (2014) Lactase persistence alleles reveal partial East African ancestry of Southern African Khoe Pastoralists. Curr Biol 24:852-858. https://doi.org/10.1016/j.cub.2014.03.027
Burger J, Link V, Blöcher J, et al (2020) Low prevalence of lactase persistence in Bronze Age Europe indicates ongoing strong selection over the last 3,000 years. Curr Biol 30:4307-4315.e13. https://doi.org/10.1016/j.cub.2020.08.033
Bycroft C, Fernandez-Rozadilla C, Ruiz-Ponte C, et al (2019) Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. Nat Comm 10:551. https://doi.org/10.1038/s41467-018-08272-w
Campbell MC, Ranciaro A (2021). Human adaptation, demography and cattle domestication: An overview of the complexity of lactase persistence in Africa. Hum Mol Genet 30:R98–R109. https://doi.org/10.1093/hmg/ddab027
Chang CC, Chow CC, Tellier LC, et al (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4:7. https://doi.org/10.1186/s13742-015-0047-8
Choudhury A, Aron S, Botigué LR, et al (2020) High-depth African genomes inform human migration and health. Nature 586:741–748. https://doi.org/10.1038/s41586-020-2859-7
Choudhury A, Sengupta D, Ramsay M, et al (2021) Bantu-speaker migration and admixture in southern Africa. Hum Mol Genet 30:R56–R63. https://doi.org/10.1093/hmg/ddaa274
Davies G (1994) Chapter One. The Ovambo in Context: The people, their land and European Colonisation in the medical culture of the Ovambo of Southern Angola and Northern Namibia [Doctoral dissertation] University of Kent at Canterbury.
De Angelis M, Ferrocino I, Calabrese FM, et al (2020) Diet influences the functions of the human intestinal microbiome. Sci Rep 10:4247. https://doi.org/10.1038/s41598-020-61192-y
de Filippo C, Bostoen K, Stoneking M, et al (2012) Bringing together linguistic and genetic evidence to test the Bantu expansion. Proc R Soc B 279:3256–3263. https://doi.org/10.1098/rspb.2012.0318
Di Gennaro F, Vergori A, Bavaro DF (2023) HIV and Co-Infections: Updates and Insights. Viruses 15:1097. https://doi.org/10.3390/v15051097
Eldredge EA (1993) A South African Kingdom: The pursuit of security in nineteenth-century Lesotho. Cambridge University Press.
Ellenberger DF (1912) History of the Basuto (Reprinted 1997). Morija Museum and Archives.
Fan X, Peters BA, Jacobs EJ, et al (2018) Drinking alcohol is associated with variation in the human oral microbiome in a large study of American adults. Microbiome 6:59. https://doi.org/10.1186/s40168-018-0448-x
Gadaga TH, Lehohla M, Ntuli V (2013) Traditional Fermented Foods of Lesotho. J Microbiol Biotechnol Food Sci 2:2387–2391. https://office2.jmbfs.org/index.php/JMBFS/article/view/7089
Gajer P, Brotman RM, Bai G, et al (2012) Temporal dynamics of the human vaginal microbiota. Sci Transl Med 4:132ra52. https://doi.org/10.1126/scitranslmed.3003605
Goodrich JK, Davenport ER, Waters JL, et al (2016) Cross-species comparisons of host genetic associations with the microbiome. Science 352:532–535. https://doi.org/10.1126/science.aad9379
González-Santos M, Montinaro F, Oosthuizen O, et al (2015) Genome-Wide SNP analysis of Southern African Populations provides new insights into the dispersal of Bantu-Speaking groups. Genome Biol Evol 7:2560–2568. https://doi.org/10.1093/gbe/evv164
Gonzalez-Santos M, Montinaro F, Grollemund R, et al (2022) Exploring the relationships between genetic linguistic and geographic distances in Bantu-speaking populations. Am J Biol Anthropol 179:104–117. https://doi.org/10.1002/ajpa.24589
Gronau I, Hubisz MJ, Gulko B, et al (2011) Bayesian inference of ancient human demography from individual genome sequences. Nat Genet 43:1031–1034. https://doi.org/10.1038/ng.937

Handley LJ, Manica A, Goudet J, et al (2007) Going the distance: human population genetics in a clinal world. Trends Genet 23:432–439. https://doi.org/10.1016/j.tig.2007.07.002

Hann CHL, Vedder H, Fourie L (1966) The native tribes of South West Africa. Frank Cass & Co.

Hammond-Tooke WD (2004) Southern Bantu origins: light from kinship terminology. Southern African Humanities 16:71–8. https://hdl.handle.net/10520/EJC84743

Henn BM, Gignoux CR, Jobin M, 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

Herráiz-Nicuesa L, Hernández-Flórez DC, Valor L, et al (2017) Impact of the Polymorphism rs9264942 near the HLA-C Gene on HIV-1 DNA reservoirs in asymptomatic chronically infected patients initiating antiviral therapy. J Immunol Res 2017:8689313. https://doi.org/10.1155/2017/8689313

Huffman TN (2007) Handbook to the Iron Age: The Archaeology of Pre-colonial farming societies in Southern Africa. University of KwaZulu-Natal Press.

Hui R, D'Atanasio E, Cassidy LM, et al (2020) Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. Sci Rep 10:18542. https://doi.org/10.1038/s41598-020-75387-w

Ioannidis AG, Blanco-Portillo J, Sandoval K, et al (2021) Paths and timings of the peopling of Polynesia inferred from genomic networks. Nature 597:522–526. https://doi.org/10.1038/s41586-021-03902-8

Jensen TGK, Liebert A, Lewinsky R, et al (2011) The 214010*C variant associated with lactase persistence is located between an Oct-1 and HNF1a binding site and increases lactase promoter activity. Hum Genet 130:483–493. https://doi.org/10.1007/s00439-011-0966-0

Langergraber KE, Siedel H, Mitani JC, et al (2007) The genetic signature of sex-biased migration in patrilocal chimpanzees and humans. PLoS One 2:e973. https://doi.org/10.1371/journal.pone.0000973

Lassalle F, Spagnoletti M, Fumagalli M, et al (2018) Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. Mol Ecol 27:182–195. https://doi.org/10.1111/mec.14435

Leeming ER, Johnson AJ, Spector TD, et al (2019) Effect of diet on the gut microbiota: rethinking intervention duration. Nutrients 11:2862. https://doi.org/10.3390/nu11122862

Leslie S, Winney B, Hellenthal G, et al (2015) The fine-scale genetic structure of the British population. Nature 519:309–314. https://doi.org/10.1038/nature14230

Lewy T, Hong BY, Weiser B, et al (2019) Oral microbiome in HIV-Infected women: shifts in the abundance of pathogenic and beneficial bacteria are associated with aging HIV load CD4 count and antiretroviral therapy. AIDS Res Hum Retroviruses 35:276–286. https://doi.org/10.1089/AID.2017.0200

Li JZ, Absher DM, Tang H, et al (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104. https://doi.org/10.1126/science.1153717

Li JH, Mazur CA, Berisa T, et al (2021) Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. Genome Res 3:529–537. https://doi.org/10.1101/gr.266486.120

Liao Y, Tong XT, Jia YJ, et al (2022) The effects of alcohol drinking on oral microbiota in the Chinese population. Int J Environ Res Public Health 19:5729. https://doi.org/10.3390/ijerph19095729

Lipson M, Skoglund P, Spriggs M, et al (2018) Population turnover in remote Oceania shortly after initial settlement. Curr Biol 28:1157–1165.e7. https://doi.org/10.1016/j.cub.2018.02.051

Macholdt E, Lede V, Barbieri C, et al (2014) Tracing pastoralist migrations to Southern Africa with lactase persistence alleles. Curr Biol 24:875–879. https://doi.org/10.1016/j.cub.2014.03.027

Macholdt E, Slatkin M, Pakendorf B, et al (2015) New insights into the history of the C-14010 lactase persistence variant in Eastern and Southern Africa. Am J Phys Anthropol 156:661–664. https://doi.org/10.1002/ajpa.22675

Malan JS (1995) Peoples of Namibia. Rhino Publishers. ISBN: 9781874946335

Marks SJ, Levy H, Martinez-Cadenas C, et al (2012) Migration distance rather than migration rate explains genetic diversity in human patrilocal groups. Mol Ecol 21:4958–4969. https://doi.org/10.1111/j.1365-294X.2012.05689.x

Marks SJ, Montinaro F, Levy H, et al (2015) Static and moving frontiers: the genetic landscape of Southern African Bantu-speaking populations. Mol Bio Evol 32:29–43. https://doi.org/10.1093/molbev/msu263

Martin AR, Atkinson EG, Chapman SB, et al (2021) Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. Am J Hum Genet 108:656-668. https://doi.org/10.1016/j.ajhg.2021.03.012

Manichaikul A, Mychaleckyj JC, Rich SS, et al (2010) Robust relationship inference in genome-wide association studies. Bioinformatics (Oxford England) 26:2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Misihairabgwi J, Cheikhyoussef A (2017) Traditional fermented foods and beverages of Namibia. Journal of Ethnic Foods 4:145–153. https://doi.org/10.1016/j.jef.2017.08.001

Montinaro F, Davies J, Capelli C (2016) Group membership geography and shared ancestry: Genetic variation in the Basotho of Lesotho. Am J Phys Anthropol 160:156–161. https://doi.org/10.1002/ajpa.22933

Montinaro F, Busby GB, Gonzalez-Santos M, et al (2017) Complex ancient genetic structure and cultural transitions in Southern African populations. Genetics 205:303–316. https://doi.org/10.1534/genetics.116.189209

Munawwar A, Singh S (2016) Human Herpesviruses as Copathogens of HIV infection, their role in HIV transmission, and disease progression. J Lab Physicians 8:5–18. https://doi.org/10.4103/0974-2727.176228

Murugesan S, Al Ahmad SF, Singh P, et al (2020) Profiling the salivary microbiome of the Qatari population. J Transl Med 18:127. https://doi.org/10.1186/s12967-020-02291-2

Norris ET, Rishishwar L, Wang L, et al (2019) Assortative mating on ancestry-variant traits in admixed Latin American Populations. Front Genet 10:359. https://doi.org/10.3389/fgene.2019.00359

Novembre J, Johnson T, Bryc K, et al (2008) Genes mirror geography within Europe. Nature 456:98–101. https://doi.org/10.1038/nature07331

Oduaran OH, Tamburini FB, Sahibdeen V, et al (2020) Gut microbiome profiling of a rural and urban South African cohort reveals biomarkers of a population in lifestyle transition. BMC Microbiol 20:330. https://doi.org/10.1186/s12866-020-02017-w

Oliveira S, Fehn AM, Aço T, et al (2018) Matriclans shape populations: Insights from the Angolan Namib Desert into the maternal genetic history of southern Africa. Am J Phys Anthropol 165:518–535. https://doi.org/10.1002/ajpa.23378

Oliveira S, Fehn AM, Amorim B, Stoneking M, et al (2023) Genome-wide variation in the Angolan Namib Desert reveals unique pre-Bantu ancestry. Sci Adv 9:eadh3822. https://doi.org/10.1126/sciadv.adh3822

Oteo-García G, Oteo JA (2021) A geometrical framework for f-Statistics. Bull Math Biol 83:14. https://doi.org/10.1007/s11538-020-00850-8

Oota H, Settheetham-Ishida W, Tiwawech D, et al (2001) Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet 29:20–21. https://doi.org/10.1038/ng711

Petersen DC, Libiger O, Tindall EA, et al (2013) Complex patterns of genomic admixture within Southern Africa. PLoS Genet 9:e1003309. https://doi.org/10.1371/journal.pgen.1003309

Pickrell JK, Patterson N, Barbieri C, et al (2012) The genetic prehistory of Southern Africa. Nat Comm 3:1143. https://doi.org/10.1038/ncomms2140

Prugnolle F, Manica A, Balloux F (2005) Geography predicts neutral genetic diversity of

human populations. Curr Biol 15:R160. https://doi.org/10.1016/j.cub.2005.02.038
Ramachandran S, Deshpande O, Roseman CC, 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. https://doi.org/10.1073/pnas.0507611102
Ranciaro A, Campbell MC, Hirbo JB, et al (2014) Genetic origins of lactase persistence and the spread of pastoralism in Africa. Am J Hum Genet 94:496–510. https://doi.org/10.1016/j.ajhg.2014.02.009
Raveane A, Aneli S, Montinaro F, et al (2019) Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. Sci Adv 5:eaaw3492. https://doi.org/10.1126/sciadv.aaw3492
Retshabile G, Mlotshwa BC, Williams L, et al (2018) Whole-exome sequencing reveals uncaptured variation and distinct ancestry in the Southern African Population of Botswana. Am J Hum Genet 102:731–743. https://doi.org/10.1016/j.ajhg.2018.03.010ù
Robinson M, Kleinman A, Graff M, et al (2017) Genetic evidence of assortative mating in humans. Nat Hum Behav1:0016. https://doi.org/10.1038/s41562-016-0016
Romero R, Hassan SS, Gajer P, et al (2014) The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. Microbiome 2:4. https://doi.org/10.1186/2049-2618-2-4
Ruan X, Luo J, Zhang P, et al (2022) The salivary microbiome shows a high prevalence of core bacterial members yet variability across human populations. NPJ Biofilms Microbiomes 8:85. https://doi.org/10.1038/s41522-022-00343-7
Sachithanandham J, Ramamurthy M, Kannangai R, et al (2009) Detection of opportunistic DNA viral infections by multiplex PCR among HIV infected individuals receiving care at a tertiary care hospital in South India. Indian J Med Microbiol 27:210–216. https://doi.org/10.4103/0255-0857.53202
Sachithanandham J, Kannangai R, Pulimood SA, et al (2014) Significance of Epstein-Barr virus (HHV-4) and CMV (HHV-5) infection among subtype-C human immunodeficiency virus-infected individuals. Indian J Med Microbiol 32:303–308. https://doi.org/10.4103/0255-0857.136558
Schlebusch C (2010) Issues raised by use of ethnic-group names in genome study. Nature 464:487. https://doi.org/10.1038/464487a
Schlebusch CM, Skoglund P, Sjödin P, 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, Miller W, Ratan A, et al (2010) Complete Khoisan and Bantu genomes from southern Africa. Nature 463:943–947. https://doi.org/10.1038/nature08795
Schwitters A, McCracken S, Frederix K, et al (2022) High HIV prevalence and associated factors in Lesotho: Results from a population-based survey. PLoS One 17:e0271431. https://doi.org/10.1371/journal.pone.0271431
Sengupta D, Choudhury A, Fortes-Lima C, et al (2021) Genetic substructure and complex demographic history of South African Bantu speakers. Nat Comm 12:2080. https://doi.org/10.1038/s41467-021-22207-y
Shaw L, Ribeiro ALR, Levine AP, et al (2017) The human salivary microbiome is shaped by shared environment rather than genetics: evidence from a large family of closely related individuals. mBio 8:10.1128/mbio.01237-17. https://doi.org/10.1128/mBio.01237-17
Singh RK, Chang HW, Yan D, et al (2017) Influence of diet on the gut microbiome and implications for human health. J Transl Med15:73. https://doi.org/10.1186/s12967-017-1175-y
Skoglund P, Posth C, Sirak K, et al (2016) Genomic insights into the peopling of the Southwest Pacific. Nature 538:510–513. https://doi.org/10.1038/nature19844
Skoglund P, Thompson JC, Prendergast ME, et al (2017) Reconstructing prehistoric African Population structure. Cell 171:59–71.e21. https://doi.org/10.1016/j.cell.2017.08.049
Takeshita T, Kageyama S, Furuta M, et al (2016) Bacterial diversity in saliva and oral health-related

conditions: the Hisayama Study. Sci Rep 6:22164. https://doi.org/10.1038/srep22164
Telenti SA, Uehlinger DE, Marchesi F, et al (1993) Epstein-Barr virus infection in HIV-positive patients. Eur J Clin Microbiol Infect Dis 12:601–609. https://doi.org/10.1007/BF01973638
Tishkoff SA, Gonder MK, Henn BM, et al (2007) History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol 24:2180–2195. https://doi.org/10.1093/molbev/msm155
Tishkoff SA, Reed FA, Ranciaro A, et al (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet 39:*31–40. https://doi.org/10.1038/ng1946
Tishkoff SA, Reed FA, Friedlaender FR, et al (2009) The genetic structure and history of Africans and African Americans. Science 324:1035–1044. https://doi.org/10.1126/science.1172257
United Nations (2023) International Migrant Stock 2020. Retrieved from https://www.un.org/development/desa/pd/content/international-migrant-stock
Uren C, Kim M, Martin AR, et al (2016) Fine-scale human population structure in Southern Africa reflects ecogeographic boundaries. Genetics 204:303–314. https://doi.org/10.1534/genetics.116.187369
Veeramah KR, Wegmann D, Woerner A, 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. https://doi.org/10.1093/molbev/msr212
Van Warmelo NJ (1962) Grouping and ethnic history. In: I. Schapera (ed) The Bantu-speaking tribes of South Africa (7th ed.), Maskew Miller Limited.
Vicente M, Jakobsson M, Ebbesen P, et al (2019) Genetic affinities among Southern Africa hunter-gatherers and the impact of admixing farmer and herder populations. Mol Biol Evol 36:1849–1861. https://doi.org/10.1093/molbev/msz089
Vlasova AN, Kandasamy S, Chattha KS, et al (2016) Comparison of probiotic lactobacilli and bifidobacteria effects, immune responses and rotavirus vaccines and infection in different host species. Vet Immunol Immunopathol 172:72–84. https://doi.org/10.1016/j.vetimm.2016.01.003
Wan Z, Chen Y, Hui J, et al (2023) Epstein-Barr virus variation in people living with human immunodeficiency virus in Southeastern China. Virol J 20:107. https://doi.org/10.1186/s12985-023-01743-3
Warinner C, Rodrigues JF, Vyas R, et al (2014) Pathogens and host immunity in the ancient human oral cavity. Nat Genet 46:336–344. https://doi.org/10.1038/ng.2906
Wilder JA, Kingan SB, Mobasher Z, et al (2004) Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. Nat Genet 36:1122–1125. https://doi.org/10.1038/ng1428
Willerslev E, Meltzer DJ (2021) Peopling of the Americas as inferred from ancient genomics. Nature 594:356–364. https://doi.org/10.1038/s41586-021-03499-y
Wirth R, Maróti G, Mihók R, et al (2020) A case study of salivary microbiome in smokers and non-smokers in Hungary: analysis by shotgun metagenome sequencing. J Oral Microbiol 12:1773067. https://doi.org/10.1080/20002297.2020.1773067
Wood DE, Lu J, Langmead B (2019) Improved metagenomic analysis with Kraken 2. Genome Biol 20:257. https://doi.org/10.1186/s13059-019-1891-0
Yatsunenko T, Rey FE, Manary MJ, et al (2012) Human gut microbiome viewed across age and geography. Nature 486:222–227. https://doi.org/10.1038/nature11053

Editor, Giovanni Destro Bisol